
بسم الله الرحمن الرحیم

دانشکده ریاضی

گروه ریاضی کاربردی

پایان نامه جهت اخذ مدرک کارشناسی ارشد

عنوان پایان نامه:

اسپلاین های تطبیقی و غیر تطبیقی در مدل های رگرسیونی نیمه پارامتری

دانشجو: **تکتم ولی زاده**

اساتید راهنما:

دکتر محمد آرشی

دکتر داود شاهسونی

آذرماه ۱۳۹۱

تقدیم به

پدر و مادرعزیزم

که سایه مهربانیشان سایه سار زندگیم می‌باشد، آنان‌که اسوه صبر و تحمل بوده و مشکلات مسیر را برایم تسهیل نمودند.

روح برادرم

که تا بود، صفایش شادی بخش و وجودش مایه دلگرمی من بود.

خانواده خوبم

که همیشه دوست و همراه من بودند.

با تمام وجود برایشان می گویم:

مراقب شادی توام

همچنان که تو پاسبان شادی منی

در آرامش نخواهم بود اگر در آرامش نباشید.

قدردانی

با سپاس فراوان به درگاه ایزد منان، که به این حقیر مجال جولان دادن در وادی لایضال علم هرچند کوتاه را عنایت فرمود. و از این‌که توانستم از این خرمن بیکران علم و دانش خوشه برگیرم و اجازت را از او و جسارت را در خویش یافتم که در بیکران دانستنیها گام نهم، آن هم به مدد اساتیدی شایسته و وارسته بسیار خرسند می‌باشم.

و به هر تقدیر:

آب دریا را گر نتوان کشید هم به قدر تشنگی باید چشید

اینک که خویش را در پایان یک مرحله از مراحل دانستن می‌یابم، برای وظیفه، ارادتم را نسبت به کلیه‌ی اساتید محترم و گرامی که استادانه آن‌چه را در توان داشتند، در کمال اخلاص به این حقیر ارزانی نمودند، ابراز می‌دارم.

در انتها شایسته است که مراتب سپاسگزاری و قدردانی خالصانه و صمیمانه خود را نسبت به اساتید عزیزم، جناب آقای دکتر محمد آرشی و جناب آقای دکتر داود شاهسونی که زحمات فراوانی را در جهت هدایت و راهنمایی این‌جانب در تهیه و تنظیم این پایان‌نامه کشیده‌اند، به‌جا آورم. امیدوارم آن-چه تحت عنوان پایان‌نامه تحصیلی خویش ارائه می‌نمایم، که ره توشه تلاش چندین ساله اساتید ارجمندم و همت کم‌رنگ خویش می‌دانم، مورد توجه و عنایتشان واقع گردد.

لازم می‌دانم کمال تشکر و احترام از پروفسور پیتر گرین[1] به‌دلیل راهنمایی‌های ارزنده‌شان در هدایت و راهنمایی این‌جانب از طریق پست الکترونیکی را به‌جا آورم.

و همچنین از اساتید داور محترم، جناب آقای دکتر مهدی روزبه (دانشگاه سمنان) و جناب آقای دکتر حسین باغیشنی (دانشگاه صنعتی شاهرود)، به‌دلیل حضور گرمشان و سعه صدرشان در تصحیح این پایان‌نامه، بسیار سپاسگزارم. و از مسئولین آموزش و دفتر دانشکده، آقای حسین‌پور و خانم خداوردی که با این‌جانب همکاری و همیاری لازم را داشتند، کمال تشکر را دارم.

[1] Professor Peter Green

و در پایان لازم می‌دانم از تمام دوستان عزیزم که در طول این کار کوچک من‌را یاری رسانند، و مورد لطف و محبت خود قرار دادند، قدردانی و تشکر نمایم و برایشان بهترین‌ها را آرزومندم.

تکتم ولی‌زاده- پاییز ۱۳۹۱

تعهد نامه

اینجانب **تکتم ولی زاده** دانشجوی دوره کارشناسی ارشد رشته **آمار**

دانشکده **ریاضی** دانشگاه صنعتی شاهرود نویسنده پایان نامه **اسپلاین‌های تطبیقی و غیرتطبیقی در مدل‌های رگرسیونی نیمه‌پارامتری** تحت راهنمائی آقایان **دکتر محمد آرشی** و **دکتر داود شاهسونی** متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

چکیده:

مدل‌های نیمه‌پارامتری اولین بار توسط انگل و همکارانش در سال ۱۹۸۶ معرفی شدند. انگل و همکارانش (۱۹۸۶) و چن و شیائو (۱۹۹۱) از روش کمترین توان‌های دوم جریمه‌ای (اسپلاین هموارساز)، فریدمن (۱۹۹۰) با استفاده از رگرسیون تطبیقی چند متغیره اسپلاین (MARS)، کیوزیک و همکارانش (۱۹۹۲) با استفاده از روش مانده‌های جزئی، سورینی و وانگ (۱۹۹۲) و کارول و همکارانش (۱۹۹۷) با استفاده از روش درستنمایی نیمرخ، پارامترهای این مدل‌ها را برآورد نمودند.

در این پایان‌نامه، هدف برآورد مدل‌های نیمه‌پارامتری به‌وسیله دو روش اسپلاین‌های تطبیقی (MARS) و اسپلاین‌های غیرتطبیقی (اسپلاین‌های هموارساز) است. در این راستا، ابتدا به توضیح هموارکننده‌های نمودار پراکنش و مفهوم اسپلاین‌ها پرداخته‌ایم. سپس دو روش ناپارامتری، اسپلاین-های هموارساز و MARS، را به‌تفصیل توضیح داده‌ایم. و در نهایت به شرح چگونگی استفاده از این دو روش در برآورد مدل‌های نیمه‌پارامتری پرداخته‌ایم.

به منظور مقایسه این دو روش در برآورد مدل‌های نیمه‌پارامتری، از یک مجموعه داده‌های شبیه‌سازی شده و چند مورد مطالعاتی استفاده کرده‌ایم. که بعد از بکارگیری هر دو روش بر روی این مجموعه داده‌ها، در تمامی موارد برتری روش اسپلاین‌های تطبیقی در برآورد مدل‌های نیمه‌پارامتری با داشتن ضریب تعیین بزرگتر و مجموع توان‌های دوم خطا کوچکتر به‌وضوح مشاهده می‌شود.

کلمات کلیدی: اسپلاین، هموارکننده نمودار پراکنش، MARS، اسپلاین هموارساز ، تطبیقی و غیر-تطبیقی، نیمه‌پارامتری.

مقاله مستخرج از پایان‌نامه:

Valizadeh. T., Shahsavani. D. and Arashi. M. (۲۰۱۲). An Application of Multivariate Adaptive Regression Splines in Housing Prices. ۱۱^{th} Iranian Statistical Conference. p ۱۷۱. Tehran, Iran.

فهرست

فصل اول

مقدمه

۱- ۱- مقدمه

سخت‌ترین بخش هر فرآیند آماری آغاز آن بوده و یکی از موضوعات مشکل در این خصوص انتخاب تحلیل درست آماری است. انتخاب درست به طبیعت داده‌ها و سوال خاصی که می‌خواهید به‌آن پاسخ دهید وابسته است.

آمار با یک مساله آغاز می‌شود سپس مجموعه‌ای از داده‌ها گردآوری می‌گردد و با تحلیل داده‌ها ادامه می‌یابد و سرانجام با نتیجه‌گیری‌هایی خاتمه پیدا می‌کند. آلبرت انیشتین می‌گوید: (نیرومند ۱۳۸۷)

" فرمول‌بندی یک مساله اغلب مهم‌تر از حل آن است که ممکن است صرفاً یک موضوع ریاضی یا مهارت تجربی باشد."

۱- ۲- رگرسیون چیست؟

در کتب آماری رگرسیون به‌عنوان روش تعیین و تحلیل روابط نادقیق بین متغیرهای آماری تعریف می‌شود. بنابراین در رگرسیون دو مطلب مورد توجه قرار می‌گیرد:

- تعیین روابط بین متغیرها
- تحلیل روابط بدست آمده

تعیین روابط بین متغیرهای آماری را رگرسیون توصیفی و تحلیل این روابط را رگرسیون استنباطی می‌نامند. به‌عنوان مثالی از رگرسیون، فرض کنید می‌خواهیم بدانیم که آیا مصرف سرانه سیگار با متغیرهای اجتماعی نظیر درآمد سرانه، سطح تحصیلات در جامعه، سهم دستمزد از درآمد کل و متوسط قیمت سیگار رابطه دارد یا خیر. یا در بسیاری از پژوهش‌های آزمایشی می‌خواهیم چگونگی اثر تغییرات یک متغیر بر متغیر دیگر را بررسی کنیم. رابطه بین مصرف سیگار با متغیرهای ذکر شده در فوق در قالب یک معادله یا الگویی است که متغیر وابسته (مصرف سیگار) را به یک یا چند متغیر پیش‌بین مربوط می‌کند. در این مثال، متغیر وابسته مصرف سیگار برحسب تعداد بسته‌های سیگار فروخته شده در یک استان بر مبنای سرانه در طول یک سال، محاسبه می‌شود و متغیرهای پیش‌بین، متغیرهای اقتصادی اجتماعی مختلف می‌باشند. متغیر پاسخ را با Y و p متغیر پیش‌بین را با $X_p, \dots, X_1$ نشان می‌دهیم. همچنین مقدار مشاهده شده پاسخ را با y و متغیرهای پیش‌بین را با $x_p, \dots, x_1$ نشان می‌دهیم.

هدف دیگر رگرسیون، در کنار دو هدف تعیین روابط بین متغیرها و تحلیل آن‌ها، استفاده از مدل نهایی در جهت پیش‌گویی است.

۱- ۳- رگرسیون خطی

یکی از قدیمی‌ترین و ساده‌ترین انواع رگرسیون، رگرسیون خطی ساده می‌باشد. فرض کنید n اندازه از متغیر پاسخ Y و متغیر پیش‌بین X داریم. در این‌صورت مدل رگرسیونی خطی ساده به‌صورت زیر می‌باشد

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

که در آن پارامترهای β_0 و β_1 مجهول و ϵ مولفه خطای تصادفی در مدل است. در حالتی کلی‌تر، مدل رگرسیونی چندگانه به صورت زیر تعریف می‌شود

$$Y = X\beta + \epsilon,$$

که درآن، $Y = (Y_1, \dots, Y_n)'$ یک بردار $n \times 1$ از متغیر پاسخ، $X = (X_1, \dots, X_n)'$ یک ماتریس $n \times p$ با بعد کامل ستونی p ، $\beta = (\beta_1, \dots, \beta_p)'$ بردار ضرایب مجهول رگرسیون و $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ بردار خطاهای تصادفی هستند. در رگرسیون خطی چندگانه فرض بر این است که رابطه بین متغیر پاسخ و متغیرهای پیش‌بین خطی است. این مدل پارامتری هنگامی مناسب است که پذیره خطی بودن رابطه میان متغیر پاسخ و متغیرهای پیش‌بین حداقل به‌صورت تقریبی برقرار باشد. اما هنگامی که رابطه میان متغیر پاسخ و متغیرهای پیش‌بین خطی نیست، استفاده از رگرسیون خطی کارا نمی‌باشد.

هنگامی که رابطه خطی میان متغیر پاسخ و متغیرهای پیش‌بین، وجود نداشته باشد، معمولاً با انجام تبدیلاتی روی متغیر پاسخ و پیش‌بین سعی در بهبود آن می‌شود. باکس و تیدول[۱] (۱۹۶۲) به مطالعه خانواده پارامتری تبدیلات روی متغیرهای پیش‌بین پرداختند. همچنین باکس و کاکس[۲] (۱۹۶۴) به معرفی یک خانواده از تبدیلات توانی که به تبدیل باکس و کاکس معروف است، پرداختند. هرچند روش باکس-کاکس در عمل بسیار استفاده می‌شود، اما این روش نیز دارای این محدودیت است که تنها به تبدیل روی متغیر پاسخ می‌پردازد (روزبه، ۱۳۹۰). اما در سال‌های اخیر با در دسترس

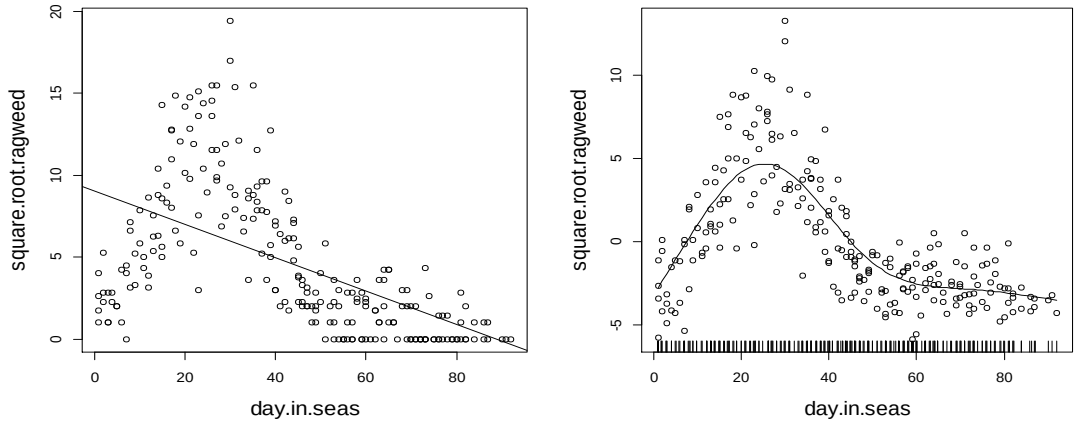
[۱] Box and Tidwell
[۲] Box and Cox

بودن برنامه‌های رایانه‌ای، تکنیک‌های ناپارامتری برازش منحنی از مقبولیت خاصی برخوردار شده است.

۱-۴- رگرسیون ناپارامتری[۱]

با توجه به اینکه در زندگی واقعی مدل داده‌ها اغلب غیرخطی‌اند، مدل‌های خطی مناسب نخواهند بود. بنابراین باید از روش‌های آماری انعطاف‌پذیر که برای شناسایی و تشخیص اثرات رگرسیون غیرخطی مفید می‌باشند، استفاده کرد.

فرض کنید که پراکنش داده‌ها شکل تقریبی یک تابع را نشان دهند. صورت تقریبی داده‌ها پیش-زمینه‌ای را برای هموارکننده‌های نمودار پراکنش[۲] ایجاد می‌کند. یک هموارکننده نمودار پراکنش روشی برای نشان دادن وابستگی تابعی بین داده‌ها بدون تحمیل کردن پذیره‌های پارامتری است. به منظور روشن شدن این مطلب، شکل ۱-۱ را در نظر بگیرید. در این شکل، نمودار پراکنش یک متغیر پاسخ ریشه دوم ابروسیا[۳] در مقابل متغیر پیش‌بین تعداد روزهای گرده‌افشانی در یک فصل مورد بررسی (مربوط به مورد مطالعاتی داده‌های ابروسیا فصل ۵) نشان داده شده است.



شکل۱- ۱: نمودار سمت چپ، نمودار پراکنش متغیر پاسخ ریشه دوم ابروسیا را در مقابل متغیر پیش‌بین تعداد روزهای گرده-افشانی در فصل مورد بررسی برای داده‌های ابروسیا همراه با خط رگرسیونی کمترین توان‌های دوم نشان می‌دهد. نمودار سمت راست، یک هموارسازی اسپلاین متغیر پاسخ ریشه دوم ابروسیا را در مقابل همان متغیر پیش‌بین به عنوان یک هموارکننده نمودار پراکنش نشان می‌دهد.

[۱] Nonparametric regression
[۲] Scatterplot smoothers
[۳] Ragweed

از روی شکل به وضوح دیده می شود که برازش یک خط مستقیم به داده‌ها، مناسب نمی‌باشد. (قاب سمت چپ) در صورتی‌که با بکاربردن یک هموارکننده نمودار پراکنش، برازش مناسبی برای داده‌ها بدست می‌آید (قاب سمت راست).

در عمل، مسائل زیادی نظیر مورد فوق وجود دارند، برای حل چنین مشکلی، از روش‌های انعطاف-پذیری که روابط غیرخطی را به طور موثری سازمان‌دهی می‌کنند، استفاده می‌شود. این روش‌ها را *رگرسیون ناپارامتری* می‌نامند. معادله یک رگرسیون ناپارامتری با متغیر پاسخ Y و تک متغیر پیش-بین X به صورت زیر می‌باشد

$$Y = f(X) + \epsilon,$$

که در آن f یک تابع هموار[۱] است که براساس داده‌های (x_i, y_i) ، $i = ۱, \dots, n$، با استفاده از یک هموارکننده نمودار پراکنش، برآورد می‌شود و $\epsilon = (\epsilon_۱, \dots, \epsilon_n)'$ خطاهای تصادفی مدل هستند. در حالت کلی‌تر، مدل ناپارامتری را به‌صورت زیر خواهیم داشت

$$Y = f\left(X_۱, X_۲, \dots, X_p\right) + \epsilon. \qquad (۱\text{-}۱)$$

معمولاً فرض می‌شود که ϵ دارای توزیع نرمال چندمتغیره با میانگین صفر و ماتریس کوواریانس $\sigma^۲ I_n$ است ($\epsilon \sim N_n(۰, \sigma^۲ I_n)$) که البته در عمل باید این پذیره با استفاده از روش‌های مناسب بررسی شود (روزبه ۱۳۹۰).

وقتی تعداد متغیرهای پیش‌بین در مدل رگرسیونی ناپارامتری افزایش می‌یابد، با مشکلاتی مواجه می‌شویم. برخی از این مشکلات عبارتند از:

۱- کم‌پشتی[۲] داده‌ها در این تنظیمات باعث می‌شود که واریانس برآوردها به طور غیرقابل قبولی بزرگ شود، که البته ممکن است در نمونه‌های خیلی بزرگ رخ ندهد. البته باید در نظر داشت که در حالت کلی با افزایش بعد، واریانس به سرعت افزایش می‌یابد. برای اطلاعات بیشتر به راهنمای نرم‌افزار SAS مراجعه کنید.

۲- تفسیر مدل رگرسیونی براساس روش‌های برآوردیابی ناپارامتری، مشکل است.

[۱] Smooth
[۲] Sparseness

۳- اغلب در درک اطلاعات این برآوردها که شامل ارتباط بین متغیرهای وابسته و پیش‌بین است، دچار مشکل می‌شویم.

برای رفع چنین مشکلاتی، مدل‌های جمعی[۱] در سال ۱۹۸۵توسط استون[۲] ارائه شد، که در بخش بعد به توضیح آن‌ها می‌پردازیم.

۱- ۵- مدل‌های جمعی

یک مدل جمعی به‌صورت زیر تعریف می‌شود

$$Y = f_{\cdot} + \sum_{j=1}^{p} f_j(X_j) + \epsilon\,, \quad (۱-۲)$$

که در آن $X_1, \dots, X_p$ متغیرهای پیش‌بین، Y متغیر پاسخ و f_jها توابع ناپارامتری هموار و مجهول هستند که از هموارکننده‌های نمودار پراکنش برای برآورد آن‌ها استفاده می‌شود. ϵ مولفه خطای تصادفی مدل است. در مدل (۱-۲) به ازای هر $j = 1, \dots, p$، فرض می‌شود $E\{f_j(X_j)\} = 0$.

مزایای این مدل‌ها نسبت به مدل‌های ناپارامتری کلی (۱-۱) عبارتند از:

۱) مشکل افزایش بعد که مدل‌های ناپارامتری کلی با آن مواجه بودند، وجود ندارد.

۲) برآوردهای هر یک از گزاره‌ها در مدل، به طور منحصر بفرد، توضیح می‌دهند که چطور متغیر وابسته با هریک از متغیرهای پیش‌بین متناظر، تغییر می‌کند.

در عمل، ممکن است که در مسئله مورد بررسی بعضی از متغیرهای پیش‌بین با متغیر پاسخ ارتباط خطی و بعضی ارتباط غیرخطی داشته باشند. بنابراین نیاز به *مدل‌های رگرسیونی نیمه‌پارامتری*[۳] که ترکیبی از مولفه‌های پارامتری و ناپارامتری هستند، احساس می‌شود. مدل‌های رگرسیونی نیمه-پارامتری در دو دهه اخیر کاربردهای زیادی را پیدا نموده و توجه بسیاری از محققان را به خود جلب کرده‌اند. مدل‌های رگرسیونی نیمه‌پارامتری نخستین بار توسط انگل[۴] و همکارانش در سال ۱۹۸۶ به‌کار برده شدند. هدف آن‌ها، بررسی رابطه بین مصرف ماهیانه برق Y با متغیرهای قیمت ماهیانه برق X_1، در آمد ماهیانه X_2 و دمای هوا t بود. آن‌ها حدس زدند که متغیر وابسته یعنی Y دارای یک رابطه خطی با متغیرهای X_1 و X_2 و یک رابطه غیرخطی با متغیر t است. اشمالنسی و استاکر[۵] (۱۹۹۹) از

[۱] Additive models
[۲] Stone
[۳] Semiparametric regression
[۴] Engle
[۵] Schmalensee and Stoker

این مدل‌ها برای بررسی رابطه بین مصرف خانگی گازوئیل با متغیرهای درآمد، سن تعداد افراد خانواده، محل اقامت و نوع خانه در ایالت متحده استفاده نمودند. آن‌ها دریافتند که لگاریتم مصرف گازوئیل بر حسب گالن دارای یک رابطه غیرخطی با لگاریتم متغیرهای درآمد و سن، و یک رابطه خطی با لگاریتم تعداد افرادی که رانندگی می‌کنند و تعداد افراد هر خانواده و سایر متغیرهاست.

تاکنون روش‌های متعددی جهت برآورد مدل‌های رگرسیونی نیمه‌پارامتری ارائه شده‌اند. به‌عنوان چند مثال می‌توان به موارد زیر اشاره کرد:

- روش کمترین توان‌های دوم جریمه‌ای[۱] مبتنی بر اسپلاین هموارساز[۲] (انگل و همکارانش، ۱۹۸۶ و چن و شیائو[۳]، ۱۹۹۱)
- روش رگرسیون تطبیقی چندگانه اسپلاین[۴](MARS) (فریدمن[۵]، ۱۹۹۰)
- روش مانده‌های جزئی[۶] (کیوزیک[۷]و همکارانش، ۱۹۹۲)
- روش درستنمایی نیمرخ[۸] (سورینی و وانگ[۹]، ۱۹۹۲ و کارول[۱۰] و همکارانش، ۱۹۹۷).

فن[۱۱] و همکارانش (۱۹۹۸) با فرض معلوم بودن قسمت خطی به برآورد تابع ناپارامتری پرداخته و سپس قسمت خطی را برآورد نمودند. در مورد روش‌های پیشنهاد شده برای حل قسمت غیرخطی می‌توان به موارد زیر اشاره نمود

- روش انتگرال حاشیه‌ای[۱۲] (جستیم و اوستد[۱۳]، ۱۹۹۴ و لینتن و نیلسن[۱۴]، ۱۹۹۵)
- الگوریتم پس‌برازش بوجا [۱۵] (هستی و تیبشیرانی[۱۶]،۱۹۹۰)
- رویکرد تقریب بوسیله سری فوریه (آماتو[۱۷]و همکارانش، ۲۰۰۲)
- روش تفاضلی (روزبه، ۱۳۹۰).

---

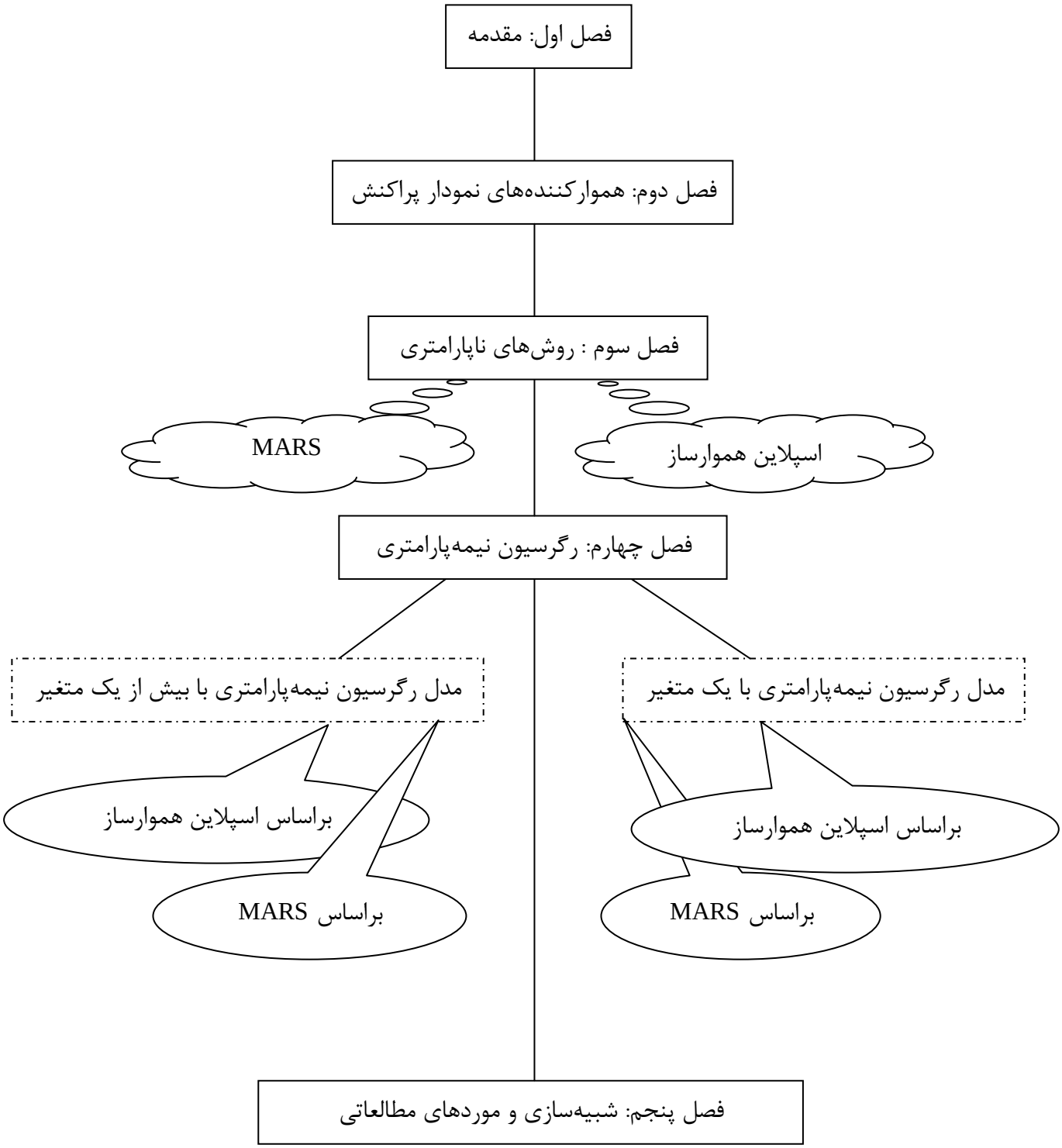
[۱] Penalized least square
[۲] Smoothing spline
[۳] Chen and Shiau
[۴] Multiple adaptive regression spline
[۵] Friedman
[۶] Partial residual method
[۷] Cuzik
[۸] Profile likelihood method
[۹] Severini and Wong
[۱۰] Carrol
[۱۱] Fan
[۱۲] Marginal integration
[۱۳] Jestheim and Auestad
[۱۴] Linton and Nielsen
[۱۵] Backfitting Buja algorithm
[۱۶] Hastie and Tibshirani
[۱۷] Amato

تمرکز این پایان‌نامه بر برآورد پارامترهای مدل‌های رگرسیونی نیمه‌پارامتری با استفاده از دو روش اسپلاین هموارساز و MARS است.

ادامه این مجموعه مبتنی بر ۴ فصل است که مطالب هر فصل به اختصار به‌صورت زیر می‌باشد:

- با توجه به اینکه برآورد در مدل‌های رگرسیونی ناپارامتری، به‌وسیله هموارکننده‌های نمودار پراکنش صورت می‌گیرد، از این‌رو در فصل دوم به توضیح هموارکننده‌های نمودار پراکنش می‌پردازیم و در ادامه فصل مفهوم اسپلاین‌ها را توضیح می‌دهیم.
- فصل سوم، شامل توضیح دو روش ناپارامتری برپایه اسپلاین، با عنوان اسپلاین‌های هموارساز و رگرسیون اسپلاین‌های تطبیقی چندگانه است.
- در فصل چهارم، به بیان مدل‌های رگرسیونی نیمه‌پارامتری براساس روش‌های ناپارامتری فصل ۳ می‌پردازیم.
- فصل پنجم نیز شامل شبیه‌سازی و موردهای مطالعاتی کاربردی است که در آن، روش‌های رگرسیونی ارائه شده در فصل‌های قبل بر روی داده‌های شبیه‌سازی شده و داده‌های واقعی اعمال و نتایج مقایسه می‌شوند.

نمایی کلی از پایان‌نامه



فصل دوم

هموارکننده‌های نمودار پراکنش

۲-۱- مقدمه

در آمار، هموارکردن یک مجموعه داده، استفاده از الگوهای مهم داده‌ها در جهت ایجاد یک تابع تقریبی است. یک هموارکننده[۱] تابعی برای خلاصه‌سازی متغیر پاسخ Y به عنوان تابعی از متغیرهای پیش‌بین $X_1, \dots X_p$ است. برای آگاهی بیشتر به هستی و تیبشیرانی (۱۹۹۰) مراجعه کنید.

یکی از خصوصیات مهم هموارکننده‌ها، ماهیت ناپارامتری آن‌ها است. هموارکننده‌ها همواره صورت تابعی ساده‌ای از وابستگی Y به $X_1, \dots X_p$ را ارائه می‌دهند، به همین دلیل اغلب در مدل رگرسیون، ساختار ناپارامتری دارند. برآورد تولید شده توسط یک هموارکننده را یک هموار[۲] می نامیم (هستی و تیبشیرانی، ۱۹۹۰).

به عنوان مثالی ساده از یک هموارکننده، می‌توان به میانگین متحرک[۳] اشاره کرد، یک خط رگرسیونی به دلیل داشتن صورت پارامتری نمی‌تواند به عنوان یک هموارکننده در نظر گرفته شود.

هموارکننده‌ها دو کاربرد عمده دارند: اولین کاربرد آن‌ها تجسم سازی است. به این معنا که یک هموارکننده نمودار پراکنش می‌تواند در جهت افزایش فهم چگونگی پراکنش Y در مقابل X، در جهت انتخاب گرایش در نمودار کمک کند. دومین کاربرد آن، برآورد میزان وابستگی میانگین Y به پیش-بین‌ها است که در ادامه توضیح داده می‌شود (هستی و تیبشیرانی، ۱۹۹۰).

۲- ۲- هموارکننده نمودار پراکنش

فرض کنید $y = (y_1, \dots, y_n)'$ مقادیر متغیر پاسخ Y و $x = (x_1, \dots, x_n)'$ مقادیر متغیر پیش-بین باشند. یک هموارکننده نمودار پراکنش، تابعی از x و y است که برازش مناسبی، در راستای خلاصه‌سازی، داده‌ها ایجاد می‌کند.

در دو دهه اخیر، تحقیقات زیادی در رابطه با هموارکننده‌های نمودار پراکنش، انجام شده‌اند. که بعضی از آن‌ها عبارتند از: واهبا[۴] (۱۹۹۰) ، گرین و سیلورمن[۵] (۱۹۹۴) ، دیرکس[۶] (۱۹۹۵)، گیو[۷] (۲۰۰۲) و هانسن و همکاران[۸](۲۰۰۳) از اسپلاین‌های هموارساز، مولر[۹](۱۹۸۸)، نادارایا[۱](۱۹۸۹)،

[۱] Smoother
[۲] Smooth
[۳] Running mean
[۴] Wahba
[۵] Green and Silverman
[۶] Dierckxe
[۷] Gu
[۸] Hansen et al.
[۹] Muller

هاردل[۲](۱۹۹۰و ۱۹۹۱)، وند و جونس[۳](۱۹۹۵)، فان و گیجبل[۴](۱۹۹۶)، سیمونوف[۵](۱۹۹۶)، بومن و آزالانی[۶](۱۹۹۷)، هارت[۷](۱۹۹۷)، لادر[۸](۱۹۹۹)، پاگان و یولاح[۹](۱۹۹۹) و فوکس[۱۰](۲۰۰۰) به چند-جمله‌ای موضعی[۱۱] و هسته‌ای[۱۲] به عنوان هموارکننده‌های نمودار پراکنش، اشاره کرده‌اند. همچنین در برخی منابع، روش‌های سری کلاسیک را به عنوان هموارکننده‌های نمودار پراکنش به‌کار برده‌اند از جمله تومپسون و تاپیا[۱۳] (۱۹۹۰)، تارتر و لوک[۱۴](۱۹۹۳) و افروموویچ[۱۵](۱۹۹۹) در نهایت روش‌های موجکی[۱۶] در اگدن[۱۷](۱۹۹۶)، لوییس[۱۸]،مس[۱۹] و ریدر[۲۰](۱۹۹۷)، هاردل[۲۱]و همکاران (۱۹۹۸)، مولر[۲۲] و ویداکویک[۲۳] (۱۹۹۹)، ویداکویک (۱۹۹۹)، ناسون[۲۴] و سیلورمن (۲۰۰۰) و والتر[۲۵] و شن[۲۶] (۲۰۰۱) به عنوان هموارکننده مورد بحث قرار گرفته اند.

با توجه به این‌که روش‌های ناپارامتری و به‌دنبال آن مدل‌های نیمه‌پارامتری، که در فصل‌های آینده ارائه شده‌اند، بر پایه اسپلاین‌هاست، در ادامه به توضیح آن‌ها می‌پردازیم.

۲- ۳- اسپلاین‌ها

اصطلاح اسپلاین در اصل، به وسیله‌ای در جهت رسم نمودارها اطلاق می‌شود. هموارکننده‌های بر پایه اسپلاین‌ها، شیوه‌هایی از رگرسیون ناپارامتری هستند. یکی از برجستگی‌های برازش براساس اسپلاین‌ها، برآورد راحت آن‌ها در مدل‌های نیمه‌پارامتری است (لوک کیل، ۲۰۰۸).

---

۱ Nadaraya
۲ Hardle
۳ Wand and Jones
۴ Fan and Gijbels
۵ Simonoff
۶ Bowman and Azzalini
۷ Hart
۸ Loader
۹ Pagan and Ullah
۱۰ Fox
۱۱ Local polynominal
۱۲ Kernel
۱۳ Thompson and Tapia
۱۴ Tarter and Lock
۱۵ Efromovich
۱۶ Wavelet
۱۷ Ogden
۱۸ Louis
۱۹ Maass
۲۰ Rieder
۲۱ Hardle
۲۲ Muller
۲۳ Vidakovic
۲۴ Nason
۲۵ Walter
۲۶ Shen

اسپلاین‌ها، تابع‌های رگرسیونی قطعه‌ای[۱] هستند که در نقاطی که گره نامیده می‌شوند، به هم متصل می‌گردند. در اصل، توابع رگرسیونی جداگانه‌ای در داخل ناحیه بین هر دو گره برازش داده می‌شود و این قطعات در گره‌ها به یکدیگر متصل می‌شوند. تحلیل‌گر باید به منظور برازش مدلی به‌وسیله اسپلاین‌ها مواردی از قبیل، تعیین درجه چندجمله‌ای ، تعداد گره‌ها و مکان گره‌ها را از قبل انتخاب نماید.

شاید گیج‌کننده‌ترین جنبه اسپلاین‌ها، وجود انواع بسیار مختلف آن‌ها باشد. برخی از آن‌ها عبارتند از، اسپلاین‌های رگرسیونی[۲]، اسپلاین‌های مکعبی[۳]، P-اسپلاین‌ها[۴]، اسپلاین‌های طبیعی[۵]، اسپلاین‌های صفحه نازک[۶] و اسپلاین های هموارساز[۷]. علاوه براین، از ترکیب آن‌ها نیز به عنوان یک اسپلاین استفاده می شود. به‌عنوان مثال اسپلاین‌های مکعبی طبیعی.[۸]

۲- ۳- ۱- تعیین گره‌ها در اسپلاین‌ها

انتخاب کران‌ها، ناحیه‌ها و فواصل در اسپلاین‌ها بسیار مهم است. یک گره، پایان یک ناحیه و آغاز ناحیه دیگر را مشخص می‌کند. معمولاً گره، آن جایی که رفتار تابع تغییر می‌کند ایجاد می‌شود. در انتخاب گره‌ها دو نکته اساسی زیر در نظر گرفته می‌شود:

۱- انتخاب درست ناحیه‌ها با توجه به کران‌هایشان
۲- تعیین درست تعداد فواصل لازم برای هر متغیر (یعنی اگر یک تابع در یک ناحیه خیلی ناهموار است، فواصل زیادی نیاز است. در عوض اگر در ناحیه‌ای تابع خطی باشد، فقط یک فاصله نیاز دارد).

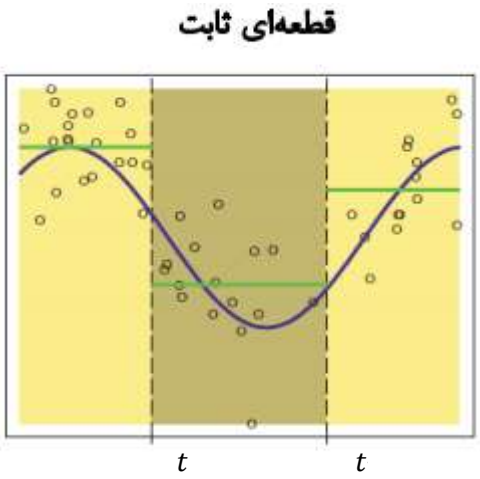
۲- ۳- ۲- چندجمله‌ای‌های قطعه‌ای و اسپلاین‌ها

فرض کنید X یک متغیر یک بعدی باشد که مشاهدات آن از توزیع گوسی (نرمال) با میانگین یک و واریانس صفر تولید شده باشند و یک تابع چندجمله‌ای قطعه‌ای $f(X)$ به‌وسیله تقسیم دامنه X به فواصل پیوسته به‌دست آمده و در هر فاصله، f نماینده یک چندجمله‌ای باشد. اولین تابعی که می-

[۱] Piecewise regression functions
[۲] Regression splines
[۳] Cubic splines
[۴] P-splines
[۵] Natural splines
[۶] Thin plate splines
[۷] Smoothing splines
[۸] Natural cubic splines

توان در هر فاصله برازش داد یک خط راست، نظیر خط سبز شکل ۲-۱ است. یعنی توابع قطعه‌ای ثابتی همانند توابع زیر

$$h_1(X) = I(X < t_1),\ \ h_2(X) = I(t_1 < X < t_2),\ \ h_3 = I(t_2 < X)$$

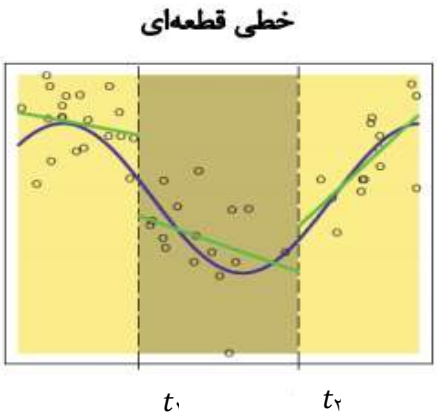


شکل۲-۱: برازش یک تابع ثابت قطعه‌ای به تعدادی داده ساختگی در هر فاصله (منحنی آبی تابع صحیح می‌باشد)

که در آن $I(A)$ تابع نشانگر و t_1 و t_2 نقاط گره هستند، زیرا تابع در آن نقاط تغییر وضعیت می‌دهد. منحنی آبی در نمودار نشان‌دهنده تابع واقعی می‌باشد. در این حالت، برآورد کمترین توان‌های دوم پارامترها در مدل $f(X) = \sum_{m=1}^{3} \beta_m h_m(X)$، $\bar{\beta}_m = \bar{y}_m$ است، که $\bar{y}_m$ میانگین m-امین ناحیه می‌باشد. بنابراین مدل برازش داده شده در این حالت به صورت زیر خواهد بود:

$$f(X) = \bar{y}_1 I(X < t_1) + \bar{y}_2 I(t_1 < X < t_2) + \bar{y}_3 I(t_2 < X).$$

دومین تابعی که می‌توان به این داده‌های ساختگی در هر فاصله برازش داد، تابع خطی قطعه‌ای نظیر شکل ۲-۲ می‌باشد.



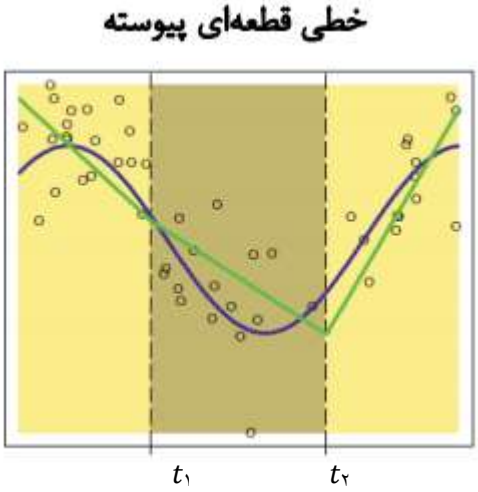
شکل۲-۲: برازش یک تابع خطی قطعه‌ای به تعدادی داده ساختگی در هر فاصله (منحنی آبی تابع صحیح می‌باشد)

در این حالت، سه تابع دیگر به‌صورت $h_{m+3} = h_m(X).X, \quad m = 1,2,3$، علاوه بر توابع قطعه‌ای ثابت، نیاز است. به عبارت دیگر، هر ناحیه علاوه بر تابع قطعه‌ای ثابت، شامل یک تابع به صورت $h(X).X$ می‌باشد. بنابراین در این حالت می‌توان مدل زیر را به این داده‌ها برازش داد:

$$f(X) = \hat{\beta}_1 I(X < t_1) + \hat{\beta}_2 XI(X < t_1) + \hat{\beta}_3 I(t_1 < X < t_2) + \hat{\beta}_4 XI(t_1 < X < t_2) + \hat{\beta}_5 I(t_3 < X) + \hat{\beta}_6 XI(t_3 < X),$$

که در آن ضرایب $\hat{\beta}_i$، $i = 1, \dots, 6$ برآوردگر کمترین توان‌های دوم پارامتر β_i است.

در مواردی خاص، برازش‌هایی همچون شکل ۲-۳ ترجیح داده می شوند، که البته آن نیز خطی قطعه-ای اما با محدودیت پیوستگی در دو گره، می‌باشد. این محدودیت‌های پیوستگی، منجر به محدودیت-های خطی روی پارامترها می‌شود.



شکل۲-۳: برازش یک تابع خطی قطعه‌ای به تعدادی داده ساختگی در هر فاصله با اعمال محدویت پیوستگی تابع در گره‌ها (منحنی آبی تابع صحیح می‌باشد)

با اعمال پیوستگی در گره اول و دوم ، باید $f(t_1^-) = f(t_1^+)$ و $f(t_2^-) = f(t_2^+)$ و در نتیجه به ترتیب محدودیت‌های خطی

$$\beta_1 + t_1\beta_4 = \beta_2 + t_2\beta_5$$

$$\beta_3 + t_2\beta_4 = \beta_5 + t_2\beta_6$$

بر روی پارامترها اعمال می‌شود. این دو محدودیت باعث می‌شود که بتوان دو پارامتر را براساس چهار پارامتر دیگر به‌دست آورد. در واقع تعداد شش پارامتر به چهار پارامتر کاهش یافته و توابع $h(X)$ با اعمال این محدودیت‌ها به صورت زیر خواهند بود

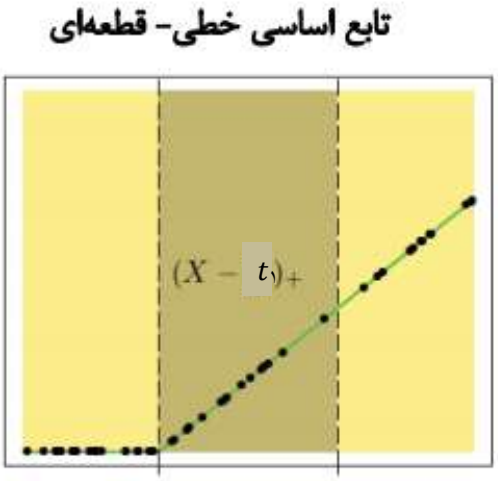
$$h_۱(\mathrm{X}) = ۱$$

$$h_۲(\mathrm{X}) = \mathrm{X}$$

$$h_۳(X) = (X - t_۱)_+$$

$$h_۴(X) = (X - t_۲)_+$$

که در آن "+"، نشان‌دهنده قسمت مثبت تابع می‌باشد. نمودار تابع $h_۳(\mathrm{X})$ در شکل ۲-۴ نشان داده شده است.



شکل۲-۴: نمایش تابع اساسی خطی قطعه‌ای $h_۳(X) = (X - t_۱)_+$ که در $t_۱$ پیوسته است. نقاط مشکی مقادیر یک نمونه nتایی $h_۳(x_i), i = ۱, \dots, n$ را مشخص می‌کنند.

در اغلب موارد، توابع هموارى ترجیح داده می‌شوند که با افزایش درجه چندجمله‌ای به‌دست می‌آیند. شکل ۲- ۵ یک مجموعه از برازش‌های چندجمله‌ای‌های قطعه‌ای مکعبی را برای یک مجموعه داده ساختگی با افزایش مرتبه پیوستگی نشان می‌دهد. در تمام نمودارهای این شکل، منحنی‌های سبز رنگ توابع برازش داده شده و منحنی آبی تابع واقعی می‌باشند. تابع قاب بالا، سمت چپ شکل ۲-۵، ارائه دهنده چندجمله‌ای قطعه‌ای مکعبی، ناپیوسته در مرز هر ناحیه است. همانطور که از شکل پیداست برازش مناسبی به داده‌ها صورت نگرفته است. نمودار سمت راست قاب بالا نیز هر چند دارای پیوستگی در تمام نقاط می‌باشد اما ارائه دهنده برازش خوبی به داده‌ها نیست. اما تابع قاب پایین، سمت راست شکل ۲-۵، پیوسته بوده و دارای مشتقات اول و دوم پیوسته در گره‌ها می‌باشد و ارائه دهنده برازش مناسبی به داده‌ها می‌باشد. چنین تابعی را اسپلاین مکعبی می‌نامند. با اعمال محدودیت‌های پیوستگی بر تابع چندجمله‌ای مکعبی و پیوستگی مشتقات اول و دوم این تابع در گره‌ها، می‌توان توابع پایه اسپلاین مکعبی، با گره‌هایی در $t_۱$ و $t_۲$ را به صورت زیر به‌دست آورد

$$h_۱(X) = ۱$$

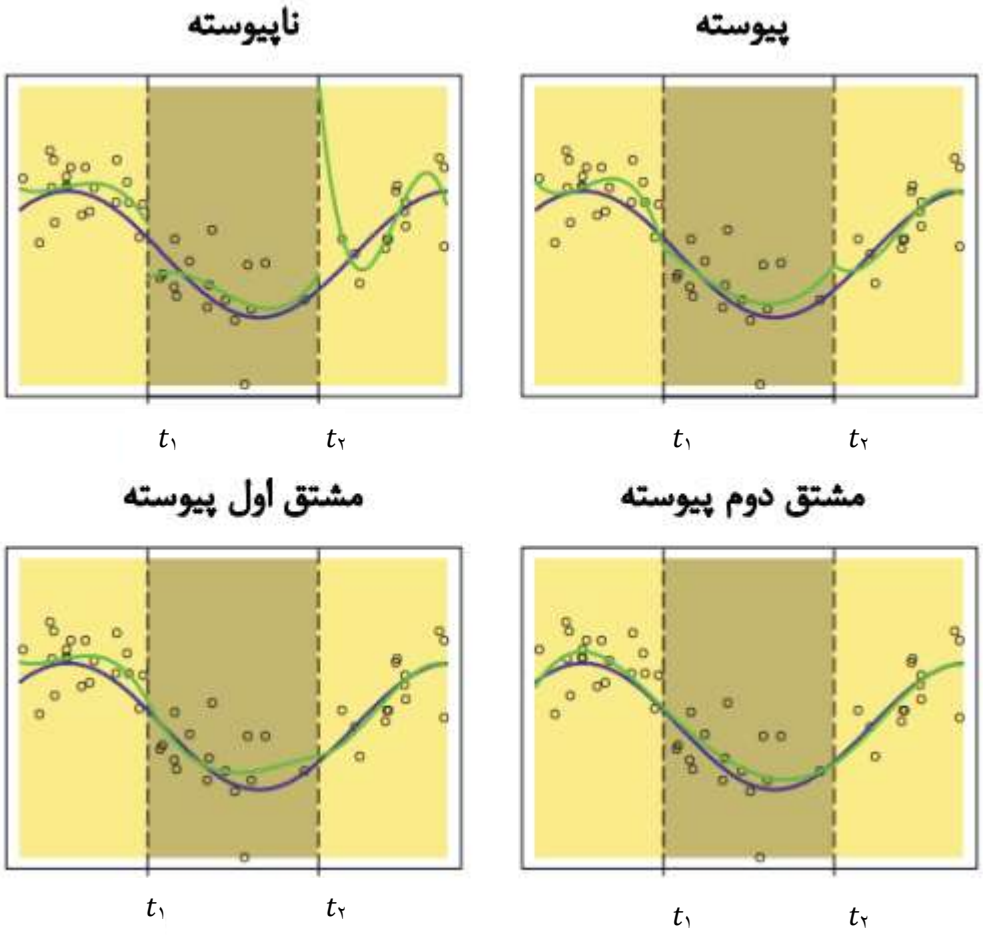
$$h_۲(X) = X$$

$$h_۳(X) = X^۲$$

$$h_۴(X) = X^۳$$

$$h_۵(X) = (X - \xi_۱)_+^۳$$

$$h_۶(X) = (X - \xi_۲)_+^۳.$$



شکل ۲- ۵: یک مجموعه از چندجمله‌ای‌های قطعه‌ای مکعبی، با افزایش مرتبه پیوستگی.

یک اسپلاین درجه M با گره‌های t_j به ازای $j = ۱, \ldots, K$ یک چندجمله‌ای قطعه‌ای درجه M است که دارای مشتقات پیوسته تا درجه $M - ۲$ باشد. در حقیقت تابع ثابت قطعه‌ای، یک اسپلاین درجه یک است، درحالی‌که تابع خطی قطعه‌ای پیوسته، یک اسپلاین درجه ۲ است. یک اسپلاین مکعبی، دارای $M = ۴$ و مشتقات اول و دوم پیوسته می‌باشد.

علاوه بر این، صورت کلی توابع پایه اسپلاین‌ها به صورت زیر می‌باشد:

$$f(X) = \sum_{j=۱}^{M+k} \beta_j h_j(X),$$

که در آن

$$h_{\mathrm{j}}(\mathrm{X}) = \mathrm{X}^{\mathrm{j}-۱} \ , \quad \mathrm{j} = ۱, \dots, \mathrm{M}, \quad h_{\mathrm{M+l}} = (X - \xi_l)_+^{M-۱} \ , \quad l = ۱, \dots, K.$$

به ندرت از اسپلاین‌هایی با درجات بیشتر از ۳ استفاده می‌شود، مگر آن‌که علاقه‌مند به هموارسازی در مشتقات باشیم. در عمل به طور گسترده از اسپلاین‌هایی با درجات $M = ۱, ۲, ۴$ استفاده می‌شود (هستی و همکاران، ۲۰۰۹).

فصل سوم

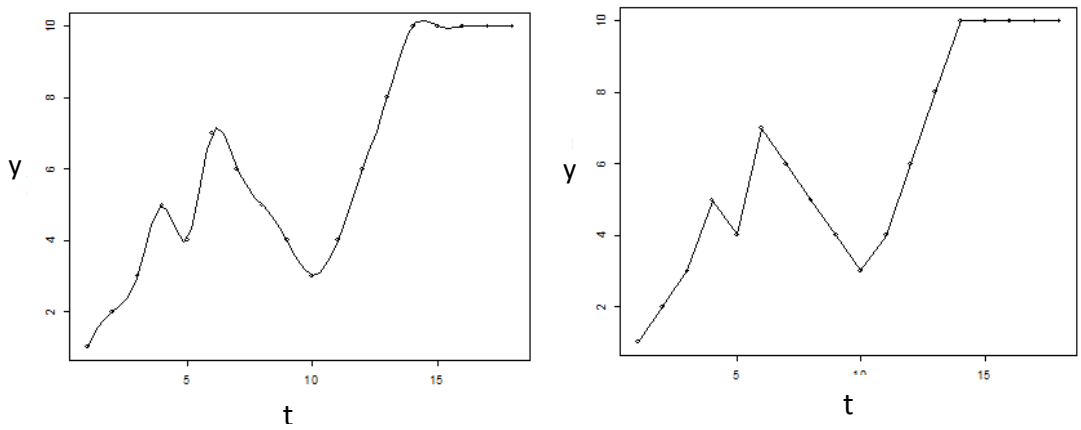
روش های ناپارامتری

۳- ۱- مقدمه

در دسته ای از اسپلاین‌ها، هموارسازی توسط یک پارامتر و در دسته‌ای دیگر، این مهم توسط تعداد گره‌های انتخاب می‌شود (لوک کیل، ۲۰۰۸). در این فصل روش "اسپلاین‌های هموارساز" و " MARS" که به ترتیب جزو دسته اول و دوم می‌باشند را بیان می‌کنیم.

۳- ۲- اسپلاین‌های هموارساز

در شکل ۳-۱ دو منحنی با استفاده از روش درون‌یابی[۱] برای یک مجموعه از داده‌های ساختگی رسم شده است. در نمودار سمت راست، تابع درون‌یاب f، از اتصال خطوط شکسته تشکیل یافته است و در نمودار سمت چپ تابع f دارای مشتق دوم پیوسته می‌باشد. لازم به ذکر است که در هر دو مورد منحنی تابع f از نقاط داده شده (t_i, y_i) عبور کرده و لذا خطای تابع درون‌یاب در این نقاط صفر است یعنی $f(t_i) = y_i$.



شکل ۳-۱ : نمایشی از منحنی های درون‌یابی شده داده‌های ساختگی

اگرچه که نمودار سمت راست در موارد خاصی که تغییرات پدیده تحت مطالعه سریع است، ممکن است برای بیان رفتار آن پدیده توجیه پذیر باشد اما در عمل، توابع همواری با نمایشی شبیه نمودار سمت چپ معقول‌تر به نظر می‌آیند. لذا "همواری" تابع، موضوعی است که بایستی در نظر گرفته شود.

باید توجه داشت که رگرسیون کلاسیک از دیدگاه دیگری موضوع را بررسی می‌کند به‌طوری‌که در آن کمینه‌سازی مجموع توان‌های دوم خطا ملاک است نه صفر بودن خطا. همچنین موضوع همواری در رگرسیون کلاسیک مدنظر نیست. حال سوال این است که چگونه می‌توان علاوه بر کمینه‌سازی

[۱] Interpolation

مجموع توان‌های دوم خطا، موضوع "همواری" تابع به طور همزمان مدنظر قرار گیرد. در بخش بعدی، پاسخ این سوال تحت عنوان " روش جریمه ناهمواری[۱] " داده خواهد شد که مبادله‌ای پایاپای بین " ناهمواری " و " مجموع توان‌های دوم خطا " را القا می‌کند.

قبل از ورود به این بحث لازم است در خصوص اندازه میزان ناهمواری تابعی چون f در بازه $[a,b]$ مطالبی بیان شود. با توجه به اینکه اندازه ناهمواری نباید تحت تاثیر اضافه کردن یک ثابت یا یک تابع خطی واقع شود، لذا می‌توان نتیجه گرفت که میزان ناهمواری تابع مورد نظر در هر نقطه می‌تواند به مقدار مشتق دوم درآن نقطه وابسته باشد. از این رو می‌توان $\int_a^b f''^2(t)dt$ را میزان ناهمواری f در کل بازه $[a,b]$ نامید. همچنین برای اندازه‌گیری ناهمواری می‌توان تعداد نقاط تغییر تابع f یا ماکزیمم $|f''|$ را نیز در نظرگرفت اما f''^2 معیار مناسبی از دیدگاه سادگی محاسباتی می‌باشد.

لازم به‌ذکر است که در زمان‌های گذشته، قبل از اینکه نمودارهای کامپیوتری بوجود آید برای رسم نموداری هموار، از یک چوب نازک انعطاف‌پذیر استفاده می‌کردند که به آن اسپلاین می‌گفتند. نکته جالب اینکه انرژی لازم برای خم کردن آن چوب جهت رسم یک نمودار خاص مثلاً f برابر با $\int f''^2$ بوده است.

۳- ۲- ۱- روش جریمه ناهمواری

با توجه به آن‌چه در قسمت قبل به عنوان اهداف برازش منحنی گفته شد، سعی در برازش منحنی است که علاوه بر دارا بودن کمترین توان دوم‌های خطا، دارای نوسانات خیلی شدیدی نباشد. از این‌رو از روش جریمه ناهمواری در جهت تحقق این مهم استفاده می‌کنیم.

فرض کنید تابع f دارای مشتق دوم در فاصله $[a,b]$ است. *مجموع توان‌های دوم جریمه شده* به صورت زیر تعریف می‌شود (گرین و سیلورمن، ۱۹۹۴)

$$S(f) = \sum_{i=1}^{n}\{Y_i - f(t_i)\}^2 + \alpha \int_a^b \{f''(x)\}^2\, dx. \quad (۱-۳)$$

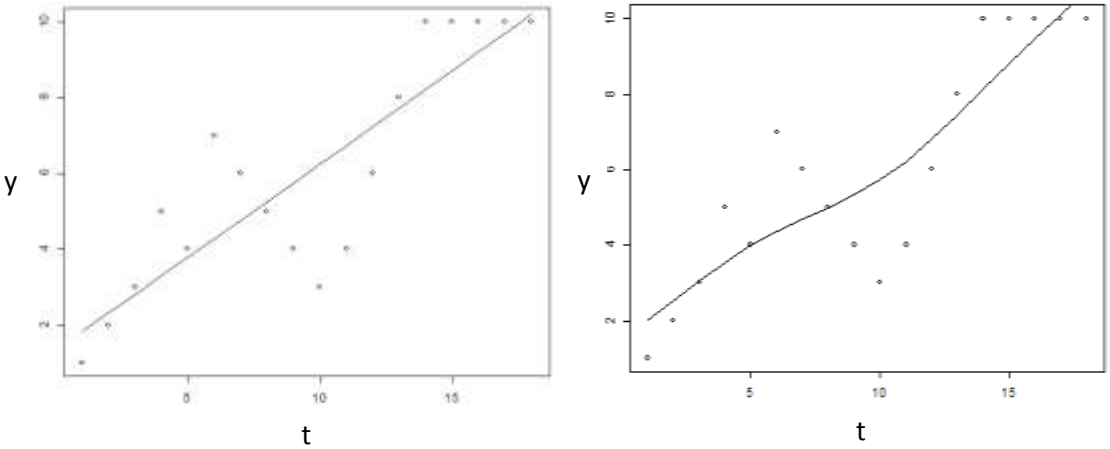
که در آن $\alpha > 0$ را پارامتر هموارسازی، جمله اول را مجموع توان‌های دوم خطا و جمله دوم را گزاره جریمه ناهمواری می‌نامند. تابعی که کمیت $S(f)$ را کمینه کند، برآوردگر کمترین توان‌های دوم جریمه شده نامیده می‌شود. به عبارت دیگر اگر آن تابع را با نماد $\hat{f}$ نمایش دهیم، آن‌گاه

$$\hat{f} = argmin_A\, S(f) \quad (۲-۳)$$

[۱] Roughness penalty method

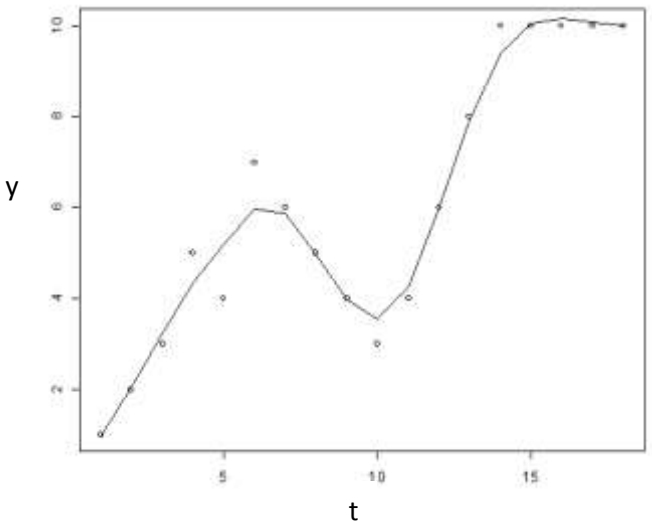
که در آن A مجموعه‌ای از توابعی است که مشتق دوم آن‌ها موجود است..

اگر α بزرگ باشد، آن‌گاه گزاره اصلی در $S(f)$ عبارت جریمه ناهمواری خواهد بود، از این‌رو مینیمم‌کننده $\hat{f}$ خمیدگی خیلی کمی، نظیر نمودار سمت راست شکل ۳-۲ خواهد داشت. همچنین اگر α به سمت بی‌نهایت میل‌کند، گزاره $\int \hat{f}''^{2}$ باید به اجبار صفر باشد و منحنی $\hat{f}$ به برازش رگرسیون خطی نزدیک می‌شود (نمودار سمت چپ شکل ۳-۲).



شکل ۳-۲: نمودار تابع f که $S(f)$ را برای مقدار بزرگ (سمت راست) و بی نهایت α (سمت چپ) مینیمم می‌کند.

از طرف دیگر اگر α به طور نسبی کوچک باشد آن‌گاه، بخش اصلی $S(f)$ را مجموع توان دوم‌های خطا تشکیل می‌دهد و لذا منحنی دارای ناهمواری زیادتری نظیر شکل ۳-۳ خواهد بود. همچنین اگر α به سمت صفر میل کند، منحنی $\hat{f}$ نزدیک به منحنی درون‌یاب نمودار سمت چپ شکل ۳-۱خواهد شد.



شکل ۳-۳: داده‌های ساختگی با منحنی که $S(f)$ را برای مقدار کوچک α مینیمم می‌کند.

از آنجا که ثابت می‌شود $\hat{f}$، گونه‌ای از توابع اسپلاین، موسوم به اسپلاین مکعبی طبیعی است لذا در بخش بعد به طور مفصل به بیان این نوع توابع می‌پردازیم.

۳- ۲- ۲- اسپلاین مکعبی و مکعبی طبیعی

جهت درک بهتر اسپلاین مکعبی، ابتدا تعریف چندجمله‌ای مرتبه n را مرور می‌کنیم.

یک چندجمله‌ای مرتبه n عبارت است از

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x^1 + a_0$$

که در آن $x \in \mathcal{R}$ و n یک عدد صحیح نامنفی و $a_n, \ldots, a_1, a_0$ ضرایب ثابت و مقادیری حقیقی هستند به‌طوری‌که $a_n \neq 0$. فرض کنید که $t_n, \ldots, t_2, t_1$ مقادیر واقع در بازه $[a, b]$ باشند به‌طوریکه $a < t_1 < \cdots < t_n < b$. تابع f روی $[a, b]$ را اسپلاین مکعبی نامند، اگر دارای دو شرط زیر باشد:

۱- تابع f در هر بازه $(a, t_1), (t_1, t_2), \ldots, (t_n, b)$، یک چندجمله‌ای درجه ۳ باشد.

۲- این چندجمله‌ای‌ها در نقاط t_i به‌گونه‌ای باشند که تابع f و مشتق اول و دوم آن در هر t_i پیوسته باشد، که در نتیجه پیوستگی روی کل بازه $[a, b]$ را نتیجه می‌شود.

به نقاط t_i گره می‌گویند. روش‌های زیادی برای ساخت یک اسپلاین مکعبی وجود دارد. که یکی از آن‌ها عبارتست از

$$f(t) = d_i(t - t_i)^3 + c_i(t - t_i)^2 + b_i(t - t_i) + a_i, \quad t_i < t < t_{i+1},$$

که در آن a_i، b_i، c_i و d_i به ازای $i = 0, \ldots, n$ ، ثابت‌های حقیقی هستند و $t_0 = a$ و $t_{n+1} = b$.

می‌دانیم که رفتار چندجمله‌ای برازش شده در نقاط انتهایی بازه $[a, b]$ یعنی نقاط داخل بازه‌های $[a, t_1]$ و $[t_n, b]$ نامنظم و غیرقابل پیش‌بینی است، بنابراین برون‌یابی[۱] در این نقاط غیرمعقول به نظر می‌رسد که می‌تواند با اسپلاین‌ها بدتر نیز گردد. لذا از آن‌جایی که نقاط انتهایی بر برازش اسپلاین تاثیر سوئی دارد می‌توانیم با گذاشتن شرط این‌که برازش مدل در نقاط انتهایی به‌صورت خط راست باشد این مشکل را برطرف کنیم، که این کار توسط اسپلاین طبیعی مکعبی امکان‌پذیر است. در حقیقت یک اسپلاین طبیعی مکعبی شرط اضافی خطی بودن اسپلاین در نقاط انتهایی را در مدل

[۱] Extrapolation

اضافه می‌کند. به عبارت دیگر، یک اسپلاین مکعبی بر فاصله $[a,b]$ را، یک *اسپلاین مکعبی طبیعی* گویند اگر و فقط اگر مشتق اول و دوم آن در a و b برابر صفر باشد.

۳- ۲- ۳- نمایش اسپلاین طبیعی مکعبی با استفاده از مقدارمشتق دوم[۱]

نمایشی که در بخش قبل برای اسپلاین مکعبی طبیعی ارائه شد، از نظر محاسباتی مناسب نمی-باشد. از این‌رو جهت ارائه فرم بهتری، با استفاده از مشتق دوم f در هر گره t_i ، نمایشی که به نمایش مشتق دومین مقدار مرسوم است، برای اسپلاین مکعبی طبیعی ارائه می‌دهیم.

فرض کنید f یک اسپلاین مکعبی طبیعی با گره‌های $t_1 < \cdots < t_n$ باشد. همچنین در نظر بگیرید

$$f_i = f(t_i), \quad \gamma_i = f''(t_i), \quad i = 1, \ldots, n.$$

بنا به تعریف اسپلاین مکعبی طبیعی، مشتق دوم f در t_1 و t_n باید صفر باشد، لذا: $\gamma_1 = \gamma_n = 0$.
حال فرض کنید $Q_{n\times(n-2)}$ با ورودی‌های q_{ij} به ازای $i = 1, \ldots, n$ و $j = 2, \ldots, n-1$[۲] به صورت زیر تعریف شود:

$$q_{j-1,j} = \frac{1}{h_{j-1}} \quad , \qquad q_{jj} = -\frac{1}{h_{j-1}} - \frac{1}{h_j} \quad , \quad q_{j+1,j} = \frac{1}{h_j}$$

برای $j = 2, \ldots, n-1$، که در آن $h_i = t_{i+1} - t_i, \ \ i = 1, \ldots, n-1$ و $|i-j| \geq 2$; $q_{ij} = 0$.
لذا ماتریس Q عبارت است از

$$Q = \begin{bmatrix} h_1^{-1} & 0 & \cdots & 0 \\ -h_1^{-1} - h_2^{-1} & h_2^{-1} & \cdots & 0 \\ h_2^{-1} & -h_2^{-1} - h_3^{-1} & \cdots & 0 \\ 0 & h_3^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{n-2}^{-1} \end{bmatrix}_{n\times(n-2)}$$

همچنین فرض کنید $R_{(n-2)\times(n-2)}$ یک ماتریس متقارن باشد که در آن مولفه‌های r_{ij} عبارتند از[۳]

---

[۱] Value- second derivative

[۲] در روشی غیراستاندارد جهت ساده سازی محاسبات j از ۲ آغاز می‌شود.

[۳] درایه های ماتریس R نیز در روشی غیراستاندار جهت سادگی محاسبات از ۲ آغاز می شود.

$$r_{ii} = \frac{۱}{۳}(h_{i-۱} + h_i) \quad ; \quad i = ۲, \dots, n-۱.$$

$$r_{i,i+۱} = r_{i+۱,i} = \frac{۱}{۶}h_i \quad ; \quad i = ۲, \dots, n-۲.$$

و $r_{ij} = ۰$ برای $|i-j| \geq ۲$. صورت ماتریسی R عبارت است از

$$R = \begin{bmatrix} \frac{1}{3}(h_1+h_3) & \frac{1}{6}h_2 & \cdots & 0 \\ \frac{1}{6}h_2 & \frac{1}{3}(h_2+h_3) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{3}(h_{n-2}+h_{n-1}) \end{bmatrix}_{(n-2)\times(n-2)}$$

ماتریس K را به‌صورت زیر در نظر بگیرید

$$K = QR^{-۱}Q^T$$

با استفاده از تعاریف فوق، می‌توان قضیه زیر را بیان کرد.

قضیه ۳- ۱: بردارهای f و مشتق‌های دوم یعنی γ یک اسپلاین مکعبی طبیعی را مشخص می‌کند اگر و فقط اگر

$$Q^{\mathrm{T}} f = R\gamma$$

برقرار باشد. که در آن $f=(f_۱, \dots, f_n)'$ و $\gamma = (\gamma_۲, \dots, \gamma_{n-۱})'$ اگر رابطه فوق برقرار باشد، آنگاه جریمه ناهمواری به صورت زیر خواهد بود (گرین و سیلورمن، ۱۹۹۴):

$$\int_a^{\mathrm{b}} f''(\mathrm{t})^۲ \, \mathrm{dt} = \gamma' \mathrm{R} \gamma = f' \mathrm{K} f.$$

۳- ۲- ۴- درون‌یابی اسپلاین‌ها

اگرچه تاکید اصلی در این پایان‌نامه بر مسائل هموارسازی است، اما در ادامه برای ساده‌سازی و یافتن مینیمم‌کننده مجموع توان‌های دوم خطا جریمه‌شده نیاز به مسائل مربوط به درون‌یابی داریم. از این‌رو به توضیح درون‌یابی با اسپلاین‌ها می‌پردازیم.

صفحه مختصات را در نظر بگیرید، فرض کنید مقادیر $y_1, \dots, y_n$ در نقاط $t_1, \dots, t_n$ داده شده باشند. هدف، یافتن یک منحنی هموار f است به‌طوریکه f از درون‌یابی در نقاط (t_i, y_i) بدست آید. به‌عبارتی داشته باشیم $f(t_i) = y_i \quad ; i = 1, \dots, n$.

به‌وضوح روش‌های زیادی برای ساختن یک تابع درون‌یاب، معقول f وجود دارد. ساده‌ترین و گسترده-ترین روش استفاده شده، متصل کردن نقاط (t_i, y_i) بوسیله یک خط مستقیم است. ضمن اینکه این روش بدون تردید برای اهداف زیادی کافی است، اما یک منحنی هموار فراهم نمی‌آورد، چون تابع t_i در نقاط f ، مشتق ناپیوسته دارد.

یک روش مناسب در جهت ساختن یک تابع درون‌یاب می‌تواند از تعریف جریمه ناهمواری که در بخش ۳-۲-۱ ارائه شد، نتیجه شود. مجموعه A در رابطه (۳-۲) را در نظر بگیرید. اگر به‌دنبال هموارترین منحنی درون‌یابی‌شده در میان نقاط داده شده باشیم، یک انتخاب بدیهی این است که منحنی درون‌یابی‌شده، مقدار $\int f''^2(t)dt$ را روی مجموعه A در میان تمام منحنی‌های هموار که به داده‌ها درون‌یابی شده، مینیمم کند. منحنی که این خاصیت را داشته باشد، یعنی در میان همه منحنی‌های f در A که درون‌یابی شده‌اند، مینیمم کننده $\int f''^2(t)dt$ در گزاره مجموع توان‌های دوم خطا جریمه‌شده باشد، یک اسپلاین مکعبی طبیعی با گره‌های t_i است.

بنابراین مسئله پیدا کردن تابع درون‌یاب با کمترین مقدار $\int f''^2(t)dt$، معادل پیدا کردن اسپلاین مکعبی طبیعی یکتایی با گره‌هایی در نقاط t_i و مقادیر $f(t_i) = y_i$ برای همه i هاست. این ادعا در قضایایی که در ادامه بیان می‌شود، اثبات می‌گردد.

نتیجه‌ای که در درون‌یابی یک مجموعه از مقادیر به‌وسیله یک اسپلاین طبیعی مکعبی در روشی منحصر بفرد حاصل می‌شود در قضیه زیر بیان شده است.

قضیه ۳- ۲: فرض کنید $n \geq 2$ ، $t_1 < \cdots < t_n$ و مقادیر $y_1, \dots, y_n$ داده شده باشد. در این صورت اسپلاین مکعبی طبیعی منحصربفردی مانند f، با گره‌هایی در نقاط t_i، وجود دارد به‌طوری‌که

$$f(t_i) = y_i \quad ; \quad \forall \ \ i = 2, \dots, n$$

برای اطلاعات بیشتر به گرین و سیلورمن (۱۹۹۴) مراجعه شود.

۳- ۲- ۵- خواص بهینگی درون‌یابی اسپلاین طبیعی مکعبی

در اینجا نشان می‌دهیم که درون‌یاب اسپلاین مکعبی طبیعی، روی کلاس بزرگتری از کلاس توابع هموار درون‌یاب، مینیمم مقدار $\int f''^{۲}(t)dt$ را دارد. مجموعه B را به صورت زیر در نظر بگیرید

$$B = \{f: f \text{ دارای مشتق اول مطلقاً پیوسته باشد} \}$$

به عبارتی f پیوسته بوده و روی فاصله $[a,b]$ دارای مشتق f' است، و یک تابع انتگرال‌پذیر f'' به ازای همه‌ی x ها در $[a,b]$ وجود دارد، به‌طوری‌که $\int_a^x f''(t)dt = f'(x) - f'(a)$.

این شرط مبین آن است که اگر f روی $[a,b]$ دارای مشتق دوم پیوسته باشد، B شامل همه توابع در A است.

قضیه ۳- ۳: فرض کنید $n \geq ۲$ و f یک درون‌یاب اسپلاین مکعبی طبیعی با مقادیر $y_۱, \dots, y_n$ در نقاط $t_۱, \dots, t_n$، $(a < t_۱ < \cdots < t_n < b)$ باشد. همچنین فرض کنید $\tilde{f}$ تابعی در B باشد به‌طوریکه به ازای $i = ۱, \dots, n$ داشته باشیم $\tilde{f}(t_i) = y_i$ آن‌گاه $\int \tilde{f}''^{۲} \geq \int f''^{۲}$ و تساوی برقرار است اگر و فقط اگر f و $\tilde{f}$ مساوی باشند.

برهان: فرض کنید $h = \tilde{f} - f$ تابعی روی B باشد که y_i به ازای $i = ۱, \dots, n$ نقاط درون‌یابی f و $\tilde{f}$ باشند. واضح است که h در همه نقاط t_i صفر است و لذا بنا به شرایط کرانداری، f'' در a و b صفر است. بنابراین با استفاده از انتگرال گیری جزء به جزء داریم

$$\begin{aligned}
\int_a^b f''(t)h''(t)dt &= -\int_a^b f'''h'(t)dt \\
&= -\sum_{j=۱}^{n-۱} f'''(t_j^+) \int_{t_j}^{t_{j+۱}} h'(t)dt \\
&= -\sum_{j=۱}^{n-۱} f'''(t_j^+)\{h(t_{j+۱}) - h(t_j)\} \\
&= ۰.
\end{aligned}$$

زیرا f''' روی هر فاصله $(a, t_۱)$ و (t_n, b) صفر، و روی هر فاصله $(t_j, t_{j+۱})$ ثابت و برابر با مقدار $f'''(t_j^+)$ است. لذا می‌توان نتیجه گرفت

$$\int_a^b \tilde{f}''^{۲} = \int_a^b (f'' + h'')^{۲} = \int_a^b f''^{۲} + ۲\int_a^b f''h'' + \int_a^b h''^{۲}$$

$$= \int_a^b f''^{۲} + \int_a^b h''^{۲} \geq \int_a^b f''^{۲}$$

در عبارت بالا، تساوی فقط و فقط وقتی برقرار است که $\int h''^{۲} = ۰$ و لذا $h = ۰$ و بنابراین $f = \tilde{f}$.

■

۳- ۲- ۶- وجود و یکتایی مینیمم کننده منحنی اسپلاین

در اینجا قصد داریم در بین همه توابع مجموعه B، تابعی که عبارت (۳-۱) را مینیمم می‌کند، به‌دست آوریم. همانند بخش‌های قبل، فرض کنید $t_۱, \dots, t_n$ نقاطی در بازه $[a, b]$ باشند به‌طوری‌که $a < t_۱ < \cdots < t_n < b$ و فرض کنید مشاهدات به‌صورت $Y_۱, \dots, Y_n$ باشند. به منظور اطمینان از برقراری شرایط قضیه فرض کنید که $n \geq ۳$ باشد.

قضیه ۳- ۴: فرض کنید گره‌های $t_۱, \dots, t_n$ به‌طوری‌که $a < t_۱ < \cdots < t_n < b$ مفروض هستند و $\alpha > ۰$ پارامتر هموارسازی است. همچنین فرض کنید که $\tilde{f}$ یک منحنی دلخواه با گره‌هایی در t_i باشد و f یک درون‌یاب اسپلاین مکعبی طبیعی با مقادیر $\tilde{f}(t_i)\ ;\ i = ۱, \dots, n$ باشد آن‌گاه

$$S(\hat{f}) \leq S(f)$$

که در آن $\hat{f}$ اسپلاین مکعبی طبیعی به‌صورت $f = (I + \alpha K)^{-۱} Y$ است. تساوی برقرار است اگر و فقط اگر $f = \hat{f}$.

برهان: بنا به فرض قضیه داریم $f(t_i) = \tilde{f}(t_i)$. بنابراین می‌توان نتیجه گرفت

$$\sum_{i=۱}^{n} \left(Y_i - \tilde{f}(t_i)\right)^{۲} = \sum_{i=۱}^{n} (Y_i - f(t_i))^{۲}$$

از طرفی بنا به قضیه ۳-۳ داریم $\int \tilde{f}''^{۲} \geq \int f''^{۲}$ ، بنابراین

$$S(f) \leq S(\tilde{f})$$

مجموع توان‌های دوم خطا برابرست با

$$\sum_{i=۱}^{n} \{Y_i - f(t_i)\}^{۲} = (Y - f)'(Y - f),$$

که در آن $Y = (Y_1, \dots, Y_n)'$ و $f(t) = (f(t_1), \dots, f(t_n))'$. با استفاده از قضیه ۳-۱ گزاره جریمه ناهمواری $\int f''^2$ را می‌توان به صورت $f'\mathrm{K}f$ نوشت. لذا داریم

$$S(f) = (Y - f)'(Y - f) + \alpha f'\mathrm{K}f$$

$$= Y'Y - Y'f - f'Y + f'f + \alpha f'\mathrm{K}f. \quad (۳\text{-}۳)$$

با استفاده از مشتق‌گیری از رابطه (۳-۳) نسبت به f، داریم:

$$\frac{1}{2}\frac{\partial S(f)}{\partial f} = -Y + f + \alpha K f = 0,$$

بنابراین

$$\hat{f} = (I + \alpha K)^{-1} Y.$$

و از طرفی

$$\frac{1}{2}\frac{\partial^2 S(f)}{\partial f^2} = \alpha K f > 0.$$

زیرا $\alpha > 0$ و K یک ماتریس معین مثبت است.

از قضیه ۳-۲ می‌دانیم که بردار f، اسپلاین f را به طور یکتا تعریف می‌کند. بنابراین روی فضای همه‌ی اسپلاین‌های مکعبی طبیعی با گره‌هایی در نقاط t_i، $S(f)$ مینیمم یکتایی را به صورت $\hat{f} = (I + \alpha K)^{-1} Y$ دارد. بنابراین خواهیم داشت

$$S(\hat{f}) \leq S(f)$$

■

همانطور که مشاهده شد، روش اسپلاین‌های هموارسازی مبتنی بر مینیمم کردن مجموع توان‌های دوم خطای جریمه‌شده است. با توجه به اینکه برآوردگر کمترین مجموع توان‌های دوم خطا جریمه-شده یک اسپلاین طبیعی مکعبی درون‌یاب می‌باشد و در ساختن آن از تمام مشاهدات به عنوان گره استفاده شده است، از این‌رو می‌توان روش اسپلاین‌های هموارساز را روش اسپلاین‌های غیرتطبیقی[1] نامید.

[1] Non- adaptive

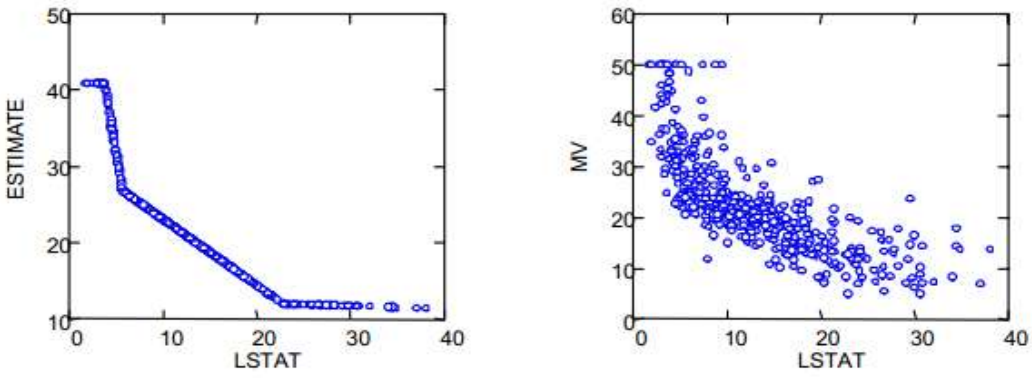
۳- ۳- مارس (MARS)

روش ناپارامتری MARS یک روش تطبیقی[۱] برای رگرسیون است و برای مسائلی با بعدهای بالا هنگامی که تعداد متغیرهای پیش‌بین زیاد است، به‌خوبی عمل می‌کند. این روش می‌تواند به عنوان تعمیمی از رگرسیون خطی گام به گام یا اصلاحی از روش درخت رگرسیونی (CART)[۲] در نظر گرفته شود (هستی و همکاران، ۲۰۰۹) به‌طوری که در روش MARS رویه پاسخ پیوسته است، اما در روش درخت رگرسیونی پیوسته نبوده و ناپیوستگی‌های آن در مرز نواحی افراز شده شکل می‌گیرد. در اینجا MARS را از نقطه نظر اول یعنی تعمیمی از رگرسیون خطی گام به گام معرفی می‌کنیم.

مدل‌سازی در این روش براساس برازش رگرسیون‌های خطی قطعه‌ای که ساده‌ترین نوع اسپلاین‌ها هستند، صورت می‌گیرد. در روش MARS، متغیرهای موثر و نقاط پایانی فواصل (گره‌ها) برای هر متغیر از طریق یک روش سریع اما بسیار فشرده تشخیص داده می‌شوند. علاوه بر جستجوی یک به یک متغیرها، MARS به جستجوی اثرات متقابل بین متغیرها تا هر مرتبه‌ای که مورد نظر باشد، می‌پردازد. مدل بهینه MARS در یک فرآیند دومرحله‌ای انتخاب می‌شود. در مرحله اول، MARS یک مدل بیش از حد بزرگ را از طریق یک مکانیزم رسمی می‌سازد. و در مرحله دوم توابع پایه‌ای که کمترین سهم را در مدل دارند، تا رسیدن به مدل بهینه از مدل حذف می‌شوند.

۳- ۳- ۱- یافتن گره‌ها در روش MARS

شکل ۳- ۴ را در نظر بگیرید. قاب سمت راست نشان‌دهنده داده‌های مسکن بوستن[۳] و قاب سمت چپ یک برآورد MARS را با سه گره برای این داده‌ها نمایش می‌دهد.



شکل ۳- ۴ : نمایش داده‌ها و برآورد MARS متغیر LSTAT داده‌های مسکن بوستون در مقابل متغیر پاسخ MV .

[۱] Adaptive
[۲] Classification and Regression Trees
[۳] Boston Housing

تفاوت اسپلاین‌ها و MARS در این است که در اسپلاین‌های معمولی، گره‌ها از پیش در فاصله‌های مساوی و از قبل تعیین می‌شوند در حالیکه در MARS، گره‌ها به‌وسیله یک روش جستجو تعیین می‌گردد. در این راستا، چنانچه معیار بهینگی برازش مدل رگرسیون، ضریب تعیین[۱] ($R^۲$) باشد در روش MARS با بررسی تعداد زیادی از گره‌های بالقوه، آن گره‌ای که بیشترین مقدار ضریب تعیین را دارد، انتخاب می‌شود. (سالفورد سیستم[۲]، ۲۰۰۱) اگرچه پیدا کردن بهترین جفت گره‌ها نیاز به محاسبات زیادی دارد و همچنین یافتن بهترین مجموعه گره‌ها، وقتی که تعداد گره‌های مورد نیاز مجهول می‌باشد، بسیار سخت و طاقت فرساست. MARS مکان و تعداد مورد نیاز گره‌ها را در یک روش پیش‌رو - پس‌رو[۳] می‌یابد. در مرحله پیش‌رو، مدلی که بیش‌برآورد دارد با گره‌های زیادی تولید می‌گردد. سپس آن گره‌هایی که کمترین سهم را در برازش کلی دارند، حذف می‌شوند. بنابراین انتخاب گره در مرحله پیش‌رو شامل مکان‌های نادرست زیادی برای گره‌ها خواهد بود، اما این گره‌های نادرست، احتمالاً در مرحله پس‌رو از مدل حذف می‌شوند. از این‌رو می‌توان روش MARS را یک روش اسپلاین تطبیقی نامید.

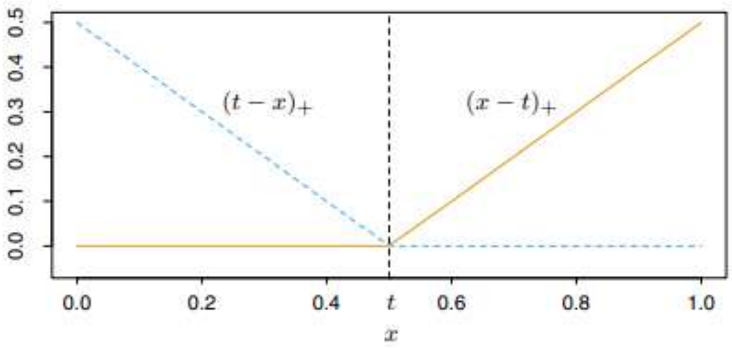
۳- ۳- ۲- توابع پایه MARS

اساس روش MARS متکی بر توابعی قطعه‌ای موسوم به توابع اسپلاین است که به صورت زیر تعریف می‌شوند

$$h_۱(x) = (t-x)_+ = \begin{cases} t-x & if\ \ t > x \\ ۰, & \text{درغیراینصورت} \end{cases}, \quad h_۲(x) = (x-t)_+ = \begin{cases} x-t & if\ \ x > t \\ ۰, & \text{درغیراینصورت} \end{cases}$$

که در آن t "گره" نامیده می‌شود و نماد "+"، نشان‌دهنده قسمت مثبت است. در این روش فرض بر آن است که گره t می‌تواند مقادیر مشاهدات متغیر x را اختیار کند، یعنی $t \in \{x_۱, x_۲, \dots, x_n\}$. در شکل ۳-۵، نمودار توابع $(x - ۰.۵)_+$ و $(۰.۵ - x)_+$ به عنوان مثال نشان داده شده است.

[۱] Coefficient of Determination
[۲] Salford systeme
[۳] Forward-Backward



شکل ۳-۵: نمایش نمودار توابع $(x-t)_+$ (خط تو پر) و $(t-x)_+$ (خط نقطه چین)

در ادامه بحث از این دو تابع به عنوان "جفت منعکس شده" یاد می‌شود (هستی و همکاران ۲۰۰۹).

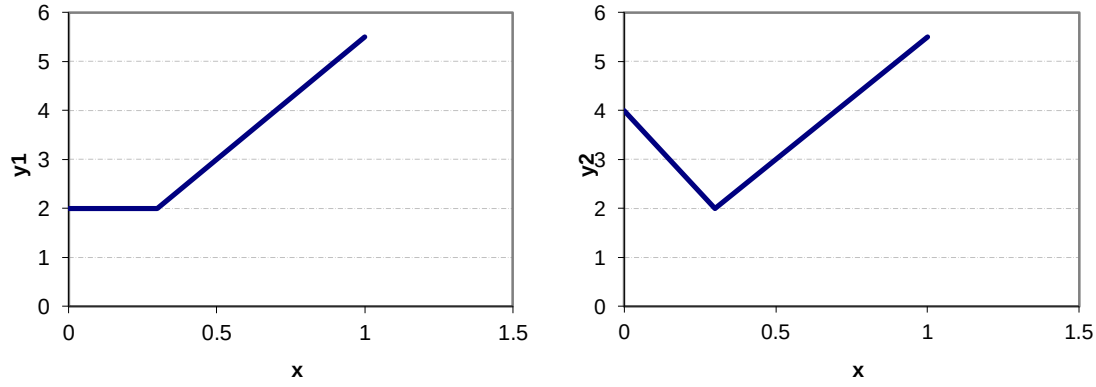
برای درک بهتر توابع پایه و ترکیب جفت‌های منعکس شده، چهار مدل زیر را که نمودار آن‌ها به‌ترتیب در شکل‌های ۳-۶ و ۳-۷ (چپ به راست) آمده است در نظر بگیرید

$$Y_۱ = ۲ + ۵(x - ۰.۳)_+$$

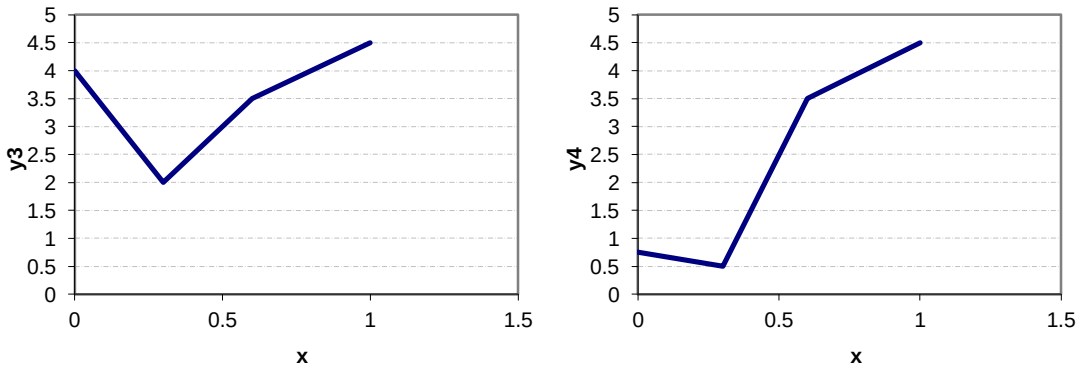
$$Y_۲ = ۲ + ۵(x - ۰.۳)_+ + ۶(۰.۳ - x)_+$$

$$Y_۳ = ۲ + ۵(x - ۰.۳)_+ + ٦(۰.۳ - x)_+ - ۳(x - ۰.٦)_+$$

$$Y_۴ = ۲ + ۵(x - ۰.۳)_+ + ۶(۰.۳ - x)_+ - ۳(x - ۰.۶)_+ - ۵(۰.۶ - x)_+$$



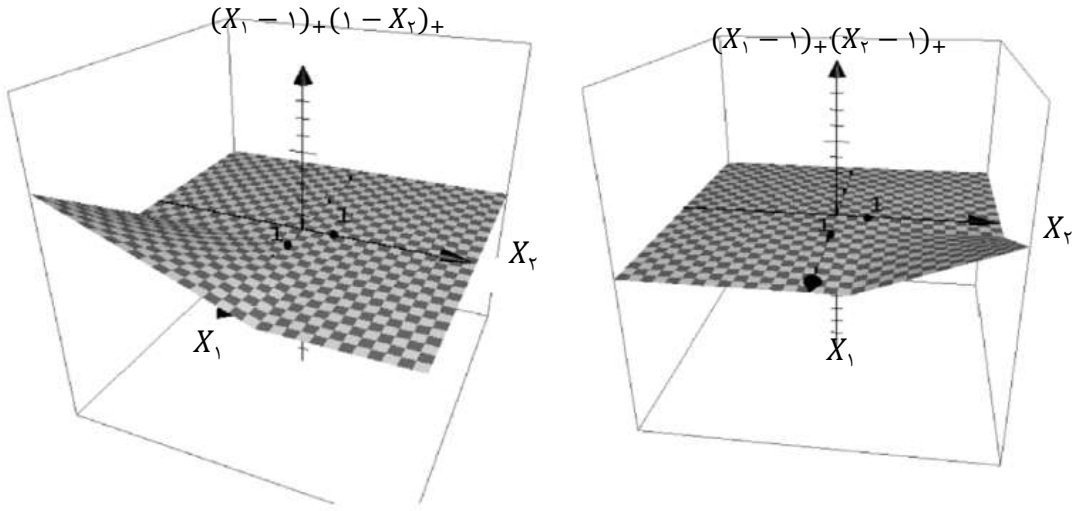
شکل ۳-۶: نمایش مدل‌های رگرسیونی $Y_۱ = ۲ + ۵(x - ۰.۳)_+$ (قاب سمت چپ) و $Y_۲ = ۲ + ۵(x - ۰.۳)_+ + ۶(۰.۳ - x)_+$ (قاب سمت راست)



شکل ۳-۷: نمایش مدل‌های رگرسیونی $Y_۳ = ۲ + ۵(x - ۰.۳)_+ + ۶(۰.۳ - x)_+ - ۳(x - ۰.۶)_+$ (قاب سمت چپ) و $Y_۴ = ۲ + ۵(x - ۰.۳)_+ + ۶(۰.۳ - x)_+ - ۳(x - ۰.۶)_+ - ۵(۰.۶ - x)_+$ (قاب سمت راست)

این نمودارها نشان می‌دهند که چگونه اضافه شدن یک تابع مبنا در مدل، می‌تواند نمودار حاصل را تحت تاثیر قرار دهد.

همانطور که گفته شد MARS می‌تواند به جستجوی اثرات متقابل بین متغیرها تا هر مرتبه‌ای که موردنظر باشد، بپردازد برای نشان دادن این توانایی MARS، فرض کنید که $p = ۲$ باشد و $X_۱$ و $X_۲$ دارای مشاهداتی با گره ۱ باشند. نمودارهای شکل ۳-۸ بیان‌کننده توانایی MARS در نشان دادن محل اثرات متقابل هستند.



شکل ۳-۸: حاصلضرب تابع پایه $(X_۱ - ۱)_+$ در توابع پایه $(۱ - X_۲)_+$ و $(X_۲ - ۱)_+$

روش MARS علاوه بر تعیین اثرات متقابل بین توابع پایه، توانایی نشان دادن محل این اثرات را در فضا دارد.

۳- ۳- ۳- روش MARS

استراتژی ساختن مدل شبیه یک رگرسیون خطی گام به‌گام است. اما به جای استفاده از متغیرهای ورودی، از توابع در مجموعه C و حاصلضرب‌هایشان (اثرات متقابل) استفاده می‌شود.

$$C = \left\{(X_j - t)_+, (t - X_j)_+\right\}, \quad t \in \{x_{1j}, x_{2j}, \dots, x_{nj}\}, \qquad j = 1, \dots, p.$$

اگر داده تکراری در مجموعه داده‌ها وجود نداشته باشد، مجموعه C شامل $2np$ تابع پایه خواهد بود. مدل MARS مورد نظر دارای ساختار تابعی زیر است

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X), \qquad (۳-۴)$$

که در آن هر $h_m(X)$ یک تابع در C یا حاصلضرب دو یا چند تا از این توابع است همچنین M تعداد توابع موجود در مدل است که پس از اعمال مرحله پس‌رو مشخص می‌گردد. با داشتن توابع h_m، ضرایب β_m بوسیله مینیمم کردن مجموع توان‌های دوم خطا، برآورد می‌شوند.

MARS یک متغیره

در ابتدا مدل MARS را فقط با یک متغیر ورودی X بیان کرده، سپس در بخش‌های بعد آن را به ورودی‌های چندگانه تعمیم می‌دهیم. در این حالت، داده‌های مشاهده شده را به‌صورت $\left((x_1, y_1), \dots, (x_n, y_n)\right)$ که (x_i, y_i) ، معرف i-امین مشاهده است، در نظر می‌گیریم.

مرحله پیش‌رو

در ابتدا مدل در نظر گرفته شده فقط شامل عرض از مبدا است، یعنی

$$\hat{f}_1(x) = \hat{\beta}_0,$$

که در آن $\hat{\beta}_0 = \bar{y}$. این مدل، مدل ۱ نامیده می‌شود. سپس برای ساختن مدل ۲، از بین مدل‌های زیر مدلی را که دارای کمترین مقدار مجموع توان‌های دوم خطا در میان سایر مدل‌ها می‌باشد، انتخاب می‌گردد

$$\hat{f}_{21}(X) = \hat{\beta}_0 + \hat{\beta}_1(X - x_1)_+ + \hat{\beta}_2(x_1 - X)_+,$$

$$\hat{f}_{22}(X) = \hat{\beta}_0 + \hat{\beta}_3(X - x_2)_+ + \hat{\beta}_4(x_2 - X)_+,$$

... .

$$\hat{f}_{۲n}(X) = \hat{\beta}_{۰} + \hat{\beta}_{۲n-۱}(X - x_n)_+ + \hat{\beta}_{۲n}(x_n - X)_+.$$

فرض کنید مدلی به‌صورت زیر انتخاب شود

$$\hat{f}_{۲}(X) = \hat{\beta}_{۰} + \hat{\beta}_{۳}(X - x_{۲})_+ + \hat{\beta}_{۴}(x_{۲} - X)_+,$$

که به صورت مختصرتر می‌توان به صورت زیر نمایش داد و آن را مدل ۲ نامید

$$\hat{f}_{۲}(X) = \hat{\beta}_{۰} + \hat{\beta}_{۳}h_{۳}(X) + \hat{\beta}_{۴}h_{۴}(X).$$

اکنون مدل ۳ را با افزودن سایر توابع پایه به مدل ۲ که افزودن آن توابع، بزرگترین کاهش را در مقدار مجموع توان دوم‌های خطا دارد، ساخته می‌شود. بنابراین مدل ۳ از بین مدل‌های زیر انتخاب خواهد شد.

$$\hat{f}_{۳۱}(X) = \hat{\beta}_{۰} + \hat{\beta}_{۱}^{*}h_{۳}(X) + \hat{\beta}_{۲}^{*}h_{۴}(X) + \hat{\beta}_{۳}^{*}h_{۱}(X) + \hat{\beta}_{۴}^{*}h_{۲}(X),$$

$$\hat{f}_{۳۲}(X) = \cdots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots + \hat{\beta}_{۳}^{*}h_{۵}(X) + \hat{\beta}_{۴}^{*}h_{۶}(X),$$

$$\ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots$$

$$\hat{f}_{۳(n-۱)}(X) = \cdots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots + \hat{\beta}_{۳}^{*}h_{۲n-۱}(X) + \hat{\beta}_{۴}^{*}h_{۲n}(X).$$

همانطور که گفته شد مدلی از بین مدل‌های فوق انتخاب می‌شود که بیشترین کاهش را در مقدار $SSE = \sum_{i=۱}^{n}\left(y_i - \hat{f}(x_i)\right)^{۲}$ داشته باشد. شایان ذکر است که در هر بار ساختن مدل جدید ضرایب β موجود در مدل جدید توسط مینیمم کردن مجموع توان‌های دوم مجدداً برآورد می‌شوند. این فرایند تا زمانی‌که K تابع پایه به مدل اضافه شود، ادامه می‌یابد. توجه شود که $K \leq ۲n$ یا توسط کاربر مشخص می‌شود.

مرحله پس‌رو

در پایان مرحله پیش‌رو، مدل تولید شده مدلی است که اگرچه برای داده‌های مدل‌ساز ممکن است مناسب باشد اما موجب بیش‌برازش داده‌های آزمون است. برای رفع این مشکل، روش MARS وارد مرحله دوم یعنی "حذف توابع پایه از مدل" می‌شود. بنابراین اکنون به حذف ممکن هر یک از توابع مدل f_k که دارای یک عرض از مبدا و $۲(K - ۱)$ تابع پایه است، توجه می‌شود. توابع پایه‌ای از مدل حذف خواهند شد که حذف آن‌ها کمترین افزایش را در مقدار مجموع توان‌های دوم خطا داشته باشد. این فرآیند تا زمانی‌که همه توابع از مدل، به‌جز عرض از مبدا حذف شوند، ادامه می‌یابد.

وقتی که فرآیند حذف توابع پایه به طور کامل انجام شد، تعداد $۲K - ۲$ مدل ساخته می‌شود (در هر مرحله یک مدل) که هر یک از آن‌ها، نامزد برای مدل نهایی هستند. برای هر کدام از این مدل‌ها، معیار اعتبار سنجی متقابل تعمیم‌یافته (GCV)، محاسبه می‌شود. اگر GCV_l مقدار GCV برای l-امین مدل در فرایند حذف پس‌رو به ازای $l = ۱, \dots, ۲K - ۲$ باشد داریم

$$GCV_l = \frac{SSE_l}{۱ - \left({(vm_l + ۱)}/{n} \right)}$$

که در آن m_l و SSE_l به ترتیب تعداد توابع پایه و مجموع توان‌های دوم خطا در مدل l-ام و v جریمه هر تابع پایه است. در واقع می‌توان گفت که v به‌عنوان یک پارامتر هموارسازی که کاربر تعریف می‌کند، عمل می‌نماید. درحقیقت v مبادله بین مدل‌های پیچیده و ساده را کنترل می‌نماید. در عمل v معمولاً بین ۲ و ۴ انتخاب می‌گردد. مدلی که در بین $۲K - ۲$ مدل دارای کمترین مقدار GCV به عنوان مدل نهایی و برآورد MARS انتخاب می‌شود.

MARS با بیش از دو متغیر پیش‌بین

اگر بیش از یک متغیر پیش‌بین وجود داشته باشد، MARS این توانایی را دارد که به‌جز توابع پایه، حاصلضرب آن‌ها یعنی اثر متقابل‌شان را در مدل لحاظ کند

$$h_m(X) = h_i(X)\left(X_j - t\right)_+, \quad h_{m+۱}(X) = h_i(X)(t - X_j)_+$$

که در آن $\{(X_j - t)_+, (t - X_j)_+\}$ جفت‌های منعکس‌شده از مجموعه C به ازای $j = ۱, \dots, p$ می‌باشند و $h_i(X)$ $(۱ \le i \le m - ۱)$ توابع پایه‌ای هستند که از قبل در مدل وجود داشته‌اند. انتخاب جفت توابع $\{h_m(X), h_{m+۱}(X)\}$ از طریق جستجوی حاصلضرب‌هایی از توابع پایه $h_i(X)$ از مدل و جفت‌های منعکس‌شده در مجموعه C انجام می‌شود، به طوری‌که $h_i(X)$ و جفت‌های منعکس‌شده هیچ یک در ساختارشان متغیر یکسان X_j را نداشته و وقتی به مدل اضافه می‌شوند، بیشترین کاهش را در مجموع توان‌های دوم خطا داشته باشند (دورماز [۱] و همکاران، ۲۰۱۰).

در ابتدا جهت روشن شدن موضوع، با فرض داشتن دو متغیر $X_۱$ و $X_۲$ و با مشاهدات $x_i = (x_{i۱}, x_{i۲}); \ i = ۱, \dots, n$ مدل MARS برازش شده را می‌توان به‌صورت زیر نوشت

[۱] Durmaz

$$f(x) = \beta_{\cdot} + \left(\sum_{i=1}^{n} \beta_{i1} h_{i1}(x_{i1}) + \sum_{i=1}^{n} \beta_{i2} h_{i2}(x_{i2})\right) + \sum_{j=1}^{n}\sum_{i=1}^{n} \beta_{(ij)12} h_{i1}(x_{i1}) h_{j2}(x_{j2})$$

اثرات دوتایی +(اثرات اصلی) + عرض از مبدأ = (۵-۳)

که در آن h_{i1} و h_{i2} به ترتیب توابع پایه براساس متغیرهای X_1 و X_2 می‌باشند.

به عنوان تعمیمی از مدل (۳-۵) در حالتی که p متغیر پیش‌بین $X_1, \dots, X_p$ داشته باشیم، مدل MARS به‌صورت زیر در می‌آید (استورلی[1] و همکاران، ۲۰۰۹)

$$\begin{aligned} f(x) = \beta_{\cdot} &+ \sum_{j=1}^{p}\left(\sum_{i=1}^{n} \beta_{ij} h_{ij}(x_j)\right) \\ &+ \sum_{j=1}^{p}\sum_{k>j}^{p}\left(\sum_{i=1}^{n}\sum_{l=1}^{n} \beta_{iljk} h_{ij}(x_j) h_{lk}(x_k)\right) + \cdots \\ &+ \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \dots \sum_{i_p=1}^{n} \beta_{i_1,\dots,i_p} h_{1,i_1}(x_1) h_{2,i_2}(x_2) \dots h_{2,i_p}(x_p) \end{aligned}$$

اثرات بالاتر + اثرات دوتایی (متقابل دوعاملی) + اثرات اصلی + عرض از مبدأ =

در عمل لازم نیست که مدل، شامل تمام اثرات متقابل باشد و لذا تعیین مرتبه آن دست کاربر است. اغلب یک مدل جمعی به‌صورت زیر (اثرات اصلی + عرض از مبدأ) کفایت می‌کند

$$f(x) = \beta_{\cdot} + \sum_{j=1}^{p}\left(\sum_{i=1}^{n} \beta_{ij} h_{ij}(x_j)\right).$$

در حالت کلی‌تر می‌توان به مدل اسپلاین شامل اثرات متقابل دوتایی نیز به‌عنوان یک تقریب برای f به صورت زیر توجه کرد

$$f(x) = \beta_{\cdot} + \sum_{j=1}^{p}\left(\sum_{i=1}^{n} \beta_{ij} h_{ij}(x_j)\right) + \sum_{j=1}^{p}\sum_{k>j}^{p}\left(\sum_{i=1}^{n}\sum_{l=1}^{n} \beta_{iljk} h_{ij}(x_j) h_{lk}(x_k)\right).$$

مدل‌هایی با اثرات متقابل سه تایی و یا حتی با اثرات متقابل بالاتر نیز می‌توانند مورد توجه باشند، اما در عمل عمومیت ندارند.

[1] Storlie

فصل چهارم

رگرسیون نیمه‌پارامتری

۴- ۱- مقدمه

رگرسیون نیمه‌پارامتری ترکیبی انعطاف‌پذیر از روابط غیرخطی در تحلیل‌های رگرسیونی است. در عمل این امکان وجود دارد که در مسئله مورد بررسی، بعضی از متغیرهای پیش‌بین با متغیر پاسخ ارتباط خطی و بعضی ارتباط غیرخطی داشته باشند. بنابراین نیاز به مدل‌های نیمه‌پارامتری که ترکیبی از مولفه‌های پارامتری و ناپارامتری هستند، احساس می‌شود. این مدل‌ها در دو دهه اخیر توجه شایانی را به‌خود جلب کرده‌اند. یک دلیل آن است که این مدل‌ها انعطاف بیشتری نسبت به مدل‌های خطی استاندارد دارند و دلیل دیگر می‌تواند تفسیر آسانتر آن‌ها در مقایسه با مدل‌های رگرسیون ناپارامتری کامل، باشد.

مدل‌های نیمه‌پارامتری حالت خاصی از مدل‌های جمعی هستند که در آن از p متغیر پیش‌بین موجود در مدل، q متغیر در قسمت خطی (پارامتری) و $p-q$ متغیر دیگر در قسمت غیرخطی (ناپارامتری) مدل قرار می‌گیرند. به عبارتی، صورت کلی یک مدل نیمه‌پارامتری عبارت است از

$$Y = X\beta + \sum_{j=q+1}^{p} f_j(X_j) + \varepsilon,$$

که در آن $Y = (y_1, y_2, \dots, y_n)'$ و $X = (x_1, x_2, \dots, x_n)'$ که در آن $x_i \; ; i = 1, \dots, n$ یک بردار q مولفه‌ای شامل متغیرهای پیش‌بین است. همچنین $\beta = (\beta_1, \beta_2, \dots, \beta_q)'$، $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ و f_jها به ازای $j = q+1, .. p$، توابع هموار مجهول هستند.

در مدل‌های نیمه‌پارامتری علاوه بر برآورد β، توابع f که توابع هموار مجهول هستند نیز باید برآورد شوند. روش‌های زیادی برای برآورد این توابع، نظیر، اسپلاین هموارساز، روش هسته، روش MARS، هموارکننده‌های بین[۱]، هموارکننده‌های میانگین متحرک و خط متحرک[۲] وجود دارد که هر یک به عنوان یک روش ناپارامتری مورد توجه است.

در این فصل، ابتدا مدل‌های نیمه‌پارامتری را در ساده‌ترین صورت آن توضیح داده و سپس به توضیح مدل‌های نیمه‌پارامتری پیچیده‌تر که در آن مولفه ناپارامتری مدل شامل دو یا بیشتر از دو متغیر پیش‌بین است، می‌پردازیم. در هر دو مورد یادشده، از روش‌های اسپلاین هموارساز و MARS

[۱] Bin smoothers
[۲] Running line

که در فصل قبل به تفصیل توضیح داده شدند، به عنوان روش‌های هموارسازی در جهت برآورد پارامترها استفاده خواهیم کرد.

۴- ۲- مدل‌های نیمه‌پارامتری ساده

مدل رگرسیونی زیر موسوم به مدل *نیمه‌پارامتری* یا *مدل خطی جزئی*[۱] را در نظر بگیرید

$$y_i = x_i'\beta + f(t_i) + \varepsilon_i\,, \quad i = ۱, \dots, n \qquad (۱-۴)$$

که در آن y_iها مشاهدات، $x_n, \dots, x_۱$ بردارهای p بعدی معلوم، $\beta = (\beta_۱, \beta_۲, \dots, \beta_p)'$ بردار p بعدی مجهول و همچنین ε_iها خطاهای تصادفی هم‌توزیع با توزیع $N(۰, \sigma^۲)$ می‌باشند. t_iها مقادیر متغیری هستند که برای مثال می‌توان آن‌ها را به عنوان زمان برداشت مشاهدات در نظر گرفت و f تابعی است که توسط روش‌های هموارسازی می‌توان آن را تقریب نمود. در مدل‌های نیمه‌پارامتری، هدف برآورد بردار پارامتر مجهول β و تابع ناپارامتری f است. مدل (۴-۱) ساده‌ترین مدل نیمه-پارامتری است که در آن مولفه ناپارامتری فقط شامل یک متغیر پیش‌بین است.

پس از انگل و همکارانش (۱۹۸۶) که برای اولین بار از این مدل استفاده کردند، توجه محققین زیادی به این مدل معطوف شد و روش‌های مختلفی برای برآورد این مدل از سوی آنان ارائه گردید. به عنوان مثال می‌توان به تحقیقات همکن[۲] (۱۹۸۶)، رایس[۳] (۱۹۸۶)، چن[۴] (۱۹۸۸)، رابینسن[۵] (۱۹۸۸)، اسپکمن[۶] (۱۹۸۸)، هانگ[۷] (۱۹۹۱)، گائو[۸] (۱۹۹۲)، لیانگ[۹] (۱۹۹۲)، گائو و ژائو[۱۰] (۱۹۹۳)، گائو و همکارانش (۱۹۹۵)، شک[۱۱] (۱۹۹۶) و باتاچاریا و ژائو[۱۲](۱۹۹۷) اشاره نمود.

در ادامه به برآورد مدل‌های نیمه‌پارامتری ساده با استفاده از دو روش اسپلاین‌های هموارساز و MARS می‌پردازیم.

---

۱ Partial linear model
۲ Heckman
۳ Rice
۴ Chen
۵ Robinson
۶ Espekman
۷ Hong
۸ Gao
۹ Liang
۱۰ Gao and Zhao
۱۱ Schick
۱۲ Battacharya and Zhao

۴- ۲- ۱- برآورد در مدل‌های نیمه‌پارامتری ساده با استفاده از اسپلاین هموارساز

در برآورد مدل‌های نیمه‌پارامتری به وسیله اسپلاین‌های هموارساز نیاز به تعریف کمترین توان‌های دوم جریمه شده برای این مدل‌ها داریم. بنابراین در ابتدا به تعریف این معیار در مدل‌های نیمه-پارامتری ساده پرداخته و سپس بر حسب نیاز به تعریف ماتریس وقوع[۱] می‌پردازیم. سپس از آن‌ها برای برآورد مدل‌های نیمه‌پارامتری ساده براساس اسپلاین هموارساز استفاده می‌کنیم.

کمترین توان‌های دوم جریمه‌شده برای مدل‌های نیمه‌پارامتری ساده

اگر اسپلاین هموارساز به‌عنوان هموارکننده نمودار پراکنش در جهت برآورد β و f، استفاده شود، همانند فصل قبل سعی در مینیمم سازی مجموع توان‌های دوم جریمه شده زیر می‌شود

$$S(\beta, f) = \sum_{i=1}^{n}\{y_i - x_i'\beta - f(t_i)\}^2 + \alpha \int f''^2(t)dt. \qquad (۴-۲)$$

که در آن α پارامتر هموارسازی و عبارت $\alpha \int f''^2(t)dt$ گزاره جریمه ناهمواری است.

ماتریس وقوع در مدل‌های نیمه‌پارامتری ساده

در فصل قبل فرض شد که t_iها مجزا و $t_1 < t_2 < \cdots < t_n$. در عمل این امکان وجود دارد که مقادیر t_iها تکراری باشد. ولذا نیاز به مرتب‌سازی آن‌هاست. با توجه به اینکه در این فصل با رگرسیون چندگانه مواجه‌ایم، ممکن است مرتب‌سازی همه متغیرهای مشاهده شده $\{y_i, x_i, t_i\}$ ساده نباشد. بنابراین بهتر است که یک روش مناسب برای آن در نظر بگیریم.

فرض کنید مقادیر مجزا و منظم $t_1, t_2, \ldots, t_n$ به‌وسیله $s_1, s_2, \ldots, s_q$ مشخص شده باشد. بنابراین بردار n بعدی t به بردار q بعدی s کاهش می‌یابد که بردار s شامل همان مقادیر بردار t، اما بدون تکرار می‌باشد. ارتباط بین $t_1, t_2, \ldots, t_n$ و $s_1, s_2, \ldots, s_q$ به‌وسیله ماتریس N با بعد $n \times q$ مشخص می‌شود به‌طوری‌که

$$\mathrm{N_{ij}} = \begin{cases} 1 & if\ \ t_i = s_j \\ 0 & \text{درغیراینصورت} \end{cases}, i = 1, \ldots, n, \quad j = 1, \ldots, q.$$

اگر بردار t شامل مقادیر تکراری نباشد، ماتریس N به یک ماتریس همانی تبدیل می‌شود.

[۱] Incidence

در ادامه با فرض تکراری بودن مقادیر t_iها سعی در بدست آوردن مینیمم $S(\beta, f)$ داریم. با این فرض $f(t_i),\ i = ۱, \dots, n$ در رابطه (۴-۲) به صورت حاصلضرب $Nf(s_j),\ j = ۱, \dots, q$ نوشته می‌شود.

برآوردگرهای کمترین توان‌های دوم

فرض کنید f بردار مقادیر $a_j = f(s_j),\ j = ۱, \dots, q$ باشد. بنابراین رابطه (۴-۲) را می‌توان به صورت زیر نوشت

$$S(\beta, f) = (Y - X\beta - Nf)'(Y - X\beta - Nf) + \alpha \int f''^{۲}(t)dt. \qquad (۴-۳)$$

مسئله مینیمم‌سازی $\int f''^{۲}(t)dt$ موضوع درون‌یابی f در نقاط $f(s_j) = a_j$ به‌طوری‌که $s_۱ < s_۲ < \cdots < s_q$ است، که در فصل ۳ با آن برخورد داشتیم. همانطور که درآن‌جا دیده شد، مینیمم‌سازی منحنی f توسط یک اسپلاین مکعبی طبیعی با گره‌های s_j صورت می‌گیرد. ماتریس‌های R و Q همانند فصل قبل تعریف می‌شوند با این تفاوت که $t_۱, t_۲, \dots, t_n$ جایگزین $s_۱, s_۲, \dots, s_q$ به‌عنوان گره می‌شوند و K به‌صورت $K = QR^{-۱}Q^T$ تعریف می‌گردد. با استفاده از قضیه ۳-۱ می‌توان نتیجه گرفت که $\int f''^{۲}(t)dt = f'\mathrm{K}f$. لذا می‌توان نوشت

$$\begin{aligned} S(\beta, f) &= (Y - X\beta - Nf)'(Y - X\beta - Nf) + \alpha f'\mathrm{K}f \\ &= Y'Y - Y'X\beta - Y'Nf - \beta'X'Y + \beta'X'X\beta + \beta'X'Nf \\ &\quad -f'N'Y + f'N'X\beta + f'N'Nf + \alpha f'Kf. \qquad (۴-۴) \end{aligned}$$

با مشتق‌گیری از رابطه (۴-۴) نسبت به f، داریم

$$\frac{\partial S}{\partial f} = -N'Y + N'X\beta - N'Y + N'X\beta + ۲N'Nf + ۲\alpha Kf.$$

بنابراین مساوی صفر قرار دادن $\frac{\partial S}{\partial f}$ و حل آن نتیجه می‌دهد

$$N'(Y - X\beta) = (N'N + \alpha K)f.$$

در نهایت برآورد f به صورت زیر حاصل می‌شود

$$\hat{f} = (N'N + \alpha K)^{-۱}N'(Y - X\beta).$$

با ضرب طرفین در ماتریس وقوع N خواهیم داشت:

$$N\hat{f} = N(N'N + \alpha K)^{-1}N'(Y - X\beta)$$
$$= S(Y - X\beta)$$

که در آن $S = N(N'N + \alpha K)^{-1}N'$ ماتریس هموار نامیده می‌شود.

اگر t_i ها مجزا و منظم شده باشند، آن‌گاه $N = I$ و ماتریس هموار S به صورت زیر ساده می‌شود.

$$S = (I + \alpha K)^{-1}.$$

در ادامه برای بدست آوردن برآورد β، دوباره از رابطه (۴-۴) نسبت به β مشتق می‌گیریم. لذا داریم

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta + 2X'Nf.$$

که با مساوی صفر قرار دادن $\frac{\partial S}{\partial \beta}$ نتیجه می‌شود

$$-X'(Y - X\beta) + X'Nf = 0. \qquad (۵-۴)$$

با جایگذاری مقدار برآورد Nf در رابطه (۴-۵) داریم:

$$-X'(Y - X\beta) + X'S(Y - X\beta) = 0$$
$$X'(I - S)(Y - X\beta) = 0$$
$$X'(I - S)Y - X'(I - S)X\beta = 0$$
$$X'(I - S)Y = X'(I - S)X\beta.$$

که در نهایت برآورد کمترین توان‌های دوم β به صورت زیر حاصل می‌شود

$$\hat{\beta} = [X'(I - S)X]^{-1}X'(I - S)Y.$$

۴-۲-۲- اعتبار سنجی متقابل تعمیم یافته

همواره یک مبادله پایاپای بین اریبی و واریانس در انتخاب پارامتر هموارسازی وجود دارد (روزبه ۱۳۹۰ را ببینید). چالش اصلی در مساله هموارسازی این است که برآوردگر ناپارامتری به چه میزان هموار باشد. چنانچه داده‌ها بیش هموار شوند، اریبی بزرگ شده و واریانس کوچک می‌شود و در صورتی‌که داده‌ها کم هموار شوند عکس حالت قبل اتفاق خواهد افتاد. لذا باید به‌دنبال معیاری باشیم که هر دو عامل را با هم در نظر بگیرد. یک معیار خوب برای انتخاب پارامتر هموارسازی می‌تواند مجموع توان دوم خطا و یا مخاطره تحت زیان توان دوم خطا می‌باشد، که مینیمم کردن چنین

معیاری عملاً امکان‌پذیر نیست زیرا، تابع f در آن معلوم نمی‌باشد. جهت حل این مشکل می‌توان از برآورد آن استفاده کرد. اولین برآوردی که به ذهن خطور می‌کند، میانگین توان دوم خطای مانده-هاست، این برآوردگر یک برآوردگر اریب می‌باشد، بنابراین مناسب نمی‌باشد. برای حل این مشکل می‌توان از معیار دیگری به‌نام اعتبار سنجی متقابل استفاده نمود که به اختصار با CV نشان داده شده به صورت زیر تعریف می‌شود

$$CV_\alpha = \sum_{i=۱}^{n} \left(\frac{y_i - \hat{y}_i}{۱ - A_{ii}} \right)^۲$$

که در آن

$$\hat{y} = \mathrm{Ay} = \mathrm{X}\hat{\beta} + N\hat{f}$$

به‌جای استفاده از معیار اعتبار سنجی متقابل برای تعیین پارامتر هموارسازی می‌توان از اعتبار سنجی متقابل تعمیم یافته (GCV)، به‌صورت زیر استفاده نمود (گرین و سیلورمن، ۱۹۹۴)

$$GCV_\alpha = \frac{\sum_{i=۱}^{n} (y_i - \hat{y}_i)^۲}{\left(۱ - \frac{۱}{n} trA \right)^۲}$$

معیار اعتبار سنجی متقابل تعمیم‌یافته، یک روش مناسب برای انتخاب پارامتر هموارسازی است. به-گونه‌ای که کمترین اعتبار سنجی متقابل تعمیم یافته، بهترین پارامتر هموارسازی را برای مدل ارائه می‌دهد. (کراوان[۱] و واهبا[۲]، ۱۹۷۹).

که در آن trA به‌صورت زیر تعریف می‌شوند

$$trA = trS + tr[X'(I - S)X]^{-۱} X'(I - S)^۲ X$$

۴- ۲- ۳- برآورد در مدل‌های نیمه‌پارامتری ساده با استفاده از روش MARS

مدل رگرسیونی زیر را در نظر بگیرید

$$Y = f(X) + \varepsilon$$

که در آن

[۱] Craven
[۲] Wahba

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \sum_{m=1}^{M} \beta_m h_m(X) \qquad (۶-۴)$$

قسمت خطی — قسمت ناپارامتری

که در آن از p متغیر پیش‌بین موجود در مدل، $p-1$ تای آن‌ها در قسمت خطی و یک متغیر در قسمت ناپارامتری مدل قرار می‌گیرد.[1] در این‌جا تابع $h_m(X)$ می‌تواند شامل موارد زیر باشد

۱) توابع پایه $(X_p - t_{ip})_+$ و $(t_{ip} - X_p)_+$ که در آن t_{ip}ها شامل مقادیر مشاهده شده متغیر X_p هستند.

۲) حاصلضرب این توابع پایه در یک یا چند متغیر قسمت خطی .

ضرایب β با استفاده از الگوریتم پیش‌رو- پس‌رو MARS، به روش مینیمم‌سازی کمترین توان‌های دوم برآورد می‌شوند. در همین راستا فریدمن (۱۹۹۱) با فرض این‌که قسمت خطی مدل ۴-۶ را می‌توان به‌صورت یک تابع کلی g نشان داد و صورت زیر را برای MARS نیمه‌پارامتری ارائه کرد

$$\hat{f}_{sp} = \sum_{j=1}^{J} c_j g_j(x) + \hat{f}(x),$$

که در آن $g_j(x)$ برای $j = 1, \ldots, J$ یک مجموعه از توابع هستند که هریک از این توابع می توانند به صورت هر ارتباط تابعی باشد و $f(x)$ بوسیله روش MARS برآورد می‌شود (فریدمن، ۱۹۹۱).

این کلاس از مدل‌های نیمه‌پارامتری ساده به عنوان مقدمه‌ای برای مدل‌های رگرسیونی نیمه-پارامتری پیچیده است، که در آن اثرات چندین متغیر پیش‌بین در قسمت ناپارامتری مدل لحاظ شده است و در بخش بعد به توضیح آن می‌پردازیم.

۴- ۶- مدل‌های نیمه‌پارامتری شامل دو یا چند مولفه ناپارامتری

هنگامی که بیش از یک متغیر پیش‌بین، ارتباط غیرخطی با متغیر پاسخ دارند، مدل‌های ساده نیمه‌پارامتری برای ارائه یک مدل رگرسیونی نیمه‌پارامتری مناسب نخواهند بود. بنابراین مجبور به استفاده از مدل‌های پیچیده‌تری همانند مدل زیر هستیم.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_q X_q + f_{q+1}(X_{q+1}) + \cdots + f_p(X_p) + \varepsilon, \qquad (۷-۴)$$

[1] در جهت همسان‌سازی مدل نیمه‌پارامتری (۴-۶) با فرم مدل نیمه‌پارامتری ساده، می‌توان فرض نمود متغیر X_p در قسمت ناپارامتری مدل همان متغیر t در مدل نیمه‌پارامتری ساده است و بقیه متغیرهای پیش‌بین، تشکیل ماتریس طرح قسمت خطی مدل را می‌دهند.

که در آن فرض شده است، از p متغیر پیش‌بین، q تا در قسمت پارامتری و $p-q$ تای دیگر در مولفه‌های ناپارامتری قرار گرفته و f_jها توابع همواری هستند که باید برآورد شوند. از این مدل‌ها با نام مدل‌های رگرسیونی نیمه‌پارامتری پیچیده استفاده می‌کنیم. مدل‌های نیمه‌پارامتری ساده، حالت خاصی از مدل (۴-۷) هستند که مولفه ناپارامتری فقط شامل یک متغیر پیش‌بین است.

۴- ۶- ۱- برآورد در مدل‌های رگرسیونی نیمه‌پارامتری پیچیده به عنوان یک مدل جمعی با استفاده از اسپلاین هموارساز

با توجه به اینکه مدل رگرسیونی نیمه‌پارامتری حالت خاصی از مدل‌های جمعی است، جهت برآورد این مدل‌ها، می‌توان از روش‌های برآورد مدل‌های جمعی استفاده کرد. عمومی‌ترین روش برای برآورد مدل‌های جمعی، برآورد هر تابع به‌وسیله یک هموارکننده دلخواه است. از بعضی از این هموارکننده‌ها می‌توان به اسپلاین هموارساز، هموارساز موزون موضعی[۱]، خط متحرک و هموارکننده‌های کرنل اشاره کرد.

در اینجا می‌خواهیم همانند مدل‌های نیمه‌پارامتری ساده، ابتدا اسپلاین هموارساز را مورد استفاده قرار دهیم. در این راستا نیاز به تعریف معیار کمترین توان‌های دوم برای این مدل‌ها داریم. قبل از تعریف این معیار در ابتدا به بیان چگونگی تعریف ماتریس‌های وقوع در این مدل‌ها می‌پردازیم.

ماتریس‌های وقوع در مدل‌های نیمه‌پارامتری پیچیده

فرض کنید بعضی از مقادیر مشاهده‌شده $X_j\ ;\ j=q+1,\dots,p$ در مدل نیمه‌پارامتری (۴-۷) تکراری باشند. همانند مدل‌های نیمه‌پارامتری ساده نیاز به ماتریس وقوع برای به‌دست آوردن روش مناسبی برای مرتب‌سازی مقادیر مشاهده‌شده متغیرهای X_j داریم. به این منظور فرض کنید مقادیر مجزا و منظم $x_{j1},\dots,x_{jn}$ برای j-امین متغیر پیش‌بین، به‌وسیله $s_{j1},\dots,s_{jq},\ j=q+1,\dots,p$ مشخص شده باشند. بنابراین مقادیر مجزا و منظم مشاهده شده هر متغیر X_j را می‌توان به صورت یک بردار S_j نمایش داد. ارتباط بین $x_{j1},\dots,x_{jn}$ و $s_{j1},\dots,s_{jq}$ بوسیله ماتریس N_j با بعد $n\times q$ مشخص می‌شود، به طوری‌که

$$N_j=\begin{cases}1 & x_{ji}=s_{jk}\\ 0 & \text{درغیراینصورت}\end{cases},\ i=1,\dots,n,\quad k=1,\dots,q,\quad j=q+1,\dots,p.$$

بنابراین مدل نیمه‌پارامتری (۴-۷) را می‌توان به صورت زیر نوشت

[۱] Locally -weighted

$$Y = X\beta + \sum_{j=q+1}^{p} N_j f_j(S_j) + \varepsilon, \qquad (۸-۴)$$

که درآن $X = (x_1, x_2, \dots, x_n)'$ و $x_i,\ i = 1, \dots, n$ها بردارهای q بعدی معلوم، $\beta = (\beta_1, \beta_2, \dots, \beta_q)'$ بردار q بعدی مجهول، $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ بردار خطای تصادفی و N_j ماتریس وقوع متناظر با هر متغیر X_j برای $j = q+1, \dots, p$ هستند. همچنین f_j ها توابع هموار مجهولی هستند که باید برآورد شوند و S_j نیز بردار مقادیر مجزا و منظم مشاهده‌شده متغیر X_j برای $j = q+1, \dots, p$ است.

کمترین توان‌های دوم جریمه‌شده

به منظور توسعه آن‌چه در معیار اسپلاین هموارساز برای مدل‌های نیمه‌پارامتری ساده گفته شد، برای مدل‌های نیمه‌پارامتری پیچیده، معیار کمترین توان‌های دوم جریمه‌شده را به‌صورت زیر تعمیم می‌دهیم.

$$S(\beta, f) = \left\{ Y - X\beta - \sum_{j=q+1}^{p} N_j f_j(S_j) \right\}^2 + \sum_{j=q+1}^{p} \alpha_j \int f_j''^2(t)dt.$$

توجه کنید که هر تابع در عبارت فوق بوسیله یک ثابت مجزای α_j جریمه شده است.

برآوردگرهای کمترین توان‌های دوم

در قضیه زیر برآوردهای f و β در مدل نیمه‌پارامتری (۸-۴) با استفاده از مینیمم سازی معیار کمترین توان‌های دوم جریمه‌شده $S(\beta, f)$ نشان داده شده است.

قضیه ۴- ۱: برآوردهای کمترین توان‌های دوم β و f_k به ازای $k = q+1, \dots, p$ برای مدل نیمه-پارامتری (۸-۴) از روابط زیر به‌دست می‌آید

$$N_k \hat{f}_k = S_k(Y - X\hat{\beta} - \sum_{\substack{j=q+1 \\ j \neq k}}^{p} N_j \hat{f}_j)$$

$$\hat{\beta} = (X'X)^{-1}X'(Y - \sum_{j=q+1}^{p} N_j \hat{f}_j)$$

که در آن $M_k = (I - S_k)$، $S_k = N_k(N_k'N_k + \alpha_k K_k)^{-1}N_k'$ و α_k پارامتر هموارسازی و K_k یک ماتریس متقارن مثبت است.

برهان: با توجه به معیار کمترین توان‌های دوم جریمه‌شده می‌توان نوشت

$$S(\beta, f) = \left(Y - X\beta - \sum_{j=q+1}^{p} N_j f_j\right)' \left(Y - X\beta - \sum_{j=q+1}^{p} N_j f_j\right) + \sum_{j=q+1}^{p} \alpha_j f_j' K_j f_j.$$

در این صورت داریم

$$S(\beta, f = \left(Y' - \beta' X' - \sum_{j=q+1}^{p} f_j' N_j'\right)\left(Y - X\beta - \sum_{j=q+1}^{p} N_j f_j\right) + \sum_{j=q+1}^{p} \alpha_j f_j' K_j f_j$$

$$= Y'Y - Y'X\beta - Y' \sum_{j=q+1}^{p} N_j f_j - \beta' X'Y$$

$$+\beta' X'X\beta + \beta' X' \sum_{j=q+1}^{p} N_j f_j - \sum_{j=q+1}^{p} f_j' N_j'Y + \sum_{j=q+1}^{p} f_j' N_j'X\beta + \sum_{j=q+1}^{p} f_j' N_j' \sum_{j=q+1}^{p} N_j f_j + \sum_{j=q+1}^{p} \alpha_j f_j' K_j f_j$$

$$= Y'Y - Y'X\beta - Y'N_k f_k - Y' \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j f_j - \beta' X'Y$$

$$+\beta' X'X\beta + \beta' X'N_k f_k + \beta' X' \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j f_j - f_k' N_k'Y$$

$$- \sum_{\substack{j=q+1 \\ j\neq k}}^{p} f_j' N_j' Y + f_k' N_k' X\beta + \sum_{\substack{j=q+1 \\ j\neq k}}^{p} f_j' N_j' X\beta + f_k' N_k' \sum_{j=q+1}^{p} N_j f_j$$

$$+ \sum_{\substack{j=q+1 \\ j\neq k}}^{p} f_j' N_j' \sum_{j=q+1}^{p} N_j f_j + N_k f_k \sum_{j=q+1}^{p} f_j' N_j' + \sum_{j=q+1}^{p} f_j' N_j' \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j f_j$$

$$+ \alpha_k f_k' K_k f_k + \sum_{j=q+1}^{p} \alpha_j f_j' K_j f_j.$$

حال در جهت بدست آوردن برآورد f، با مشتق‌گیری از عبارت $S(\beta, f)$ نسبت به f_{k} داریم

$$\frac{\partial S}{\partial f_k} = -N_k' Y + N_k' X\beta - N_k' Y \ + N_k' X\beta + N_k' \sum_{j=q+1}^{p} N_j f_j + N_k' \sum_{j=q+1}^{p} N_j f_j$$
$$+ ۲\alpha_k K_k f_k$$

با مساوی صفر قرار دادن $\frac{\partial S}{\partial f_k}$ داریم

$$-N_{\mathrm{k}}'(\mathrm{Y} - \mathrm{X}\beta) + N_k' N_k f_k + N_k' \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j f_j + \alpha_k K_k f_k = ۰$$

بنابراین

$$N_{\mathrm{k}}' \left(\mathrm{Y} - \mathrm{X}\beta - \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j f_j) = (\, N_k' N_k + \alpha_{\mathrm{k}} K_k \right) f_k.$$

که نتیجه می‌دهد

$$\hat{f}_k = (N_k' N_k + \alpha_k K_k)^{-1} N_k' (Y - \mathrm{X}\hat{\beta} - \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j \hat{f}_j).$$

با ضرب طرفین معادله فوق در N_k داریم

$$N_k\hat{f}_k = N_k(N_k'N_k + \alpha_k K_k)^{-1}N_k'(Y - X\hat{\beta} - \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j\hat{f}_j).$$

در نهایت می‌توان نوشت

$$N_k\hat{f}_k = S_k(Y - X\hat{\beta} - \sum_{\substack{j=q+1 \\ j\neq k}}^{p} N_j\hat{f}_j),$$

که در آن $S_k = N_k(N_k'N_k + \alpha_k K_k)^{-1}N_k'$ ماتریس هموارسازی می‌باشد.

حال برای بدست آوردن برآورد β از عبارت $S(\beta, f)$ نسبت به β مشتق می‌گیریم، داریم

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta + 2X'\sum_{j=q+1}^{p} N_j f_j$$

که با مساوی صفر قرار دادن $\frac{\partial S}{\partial \beta}$، نتیجه می‌گیریم

$$\beta = (X'X)^{-1}(X'Y - X'\sum_{j=q+1}^{p} N_j f_j)$$

بنابراین

$$\hat{\beta} = (X'X)^{-1}X'(Y - \sum_{j=q+1}^{p} N_j\hat{f}_j)$$

۴- ۶- ۲- برآورد در مدل‌های رگرسیونی نیمه‌پارامتری پیچیده با استفاده از روش MARS

در این‌جا یک مدل نیمه‌پارامتری پیچیده به‌وسیله روش MARS می‌پردازیم. به منظور تعریف یک مدل MARS نیمه‌پارامتری به‌وسیله توابع پایه که در فصل ۳ در توضیح روشMARS ارائه شد، مدل نیمه‌پارامتری (۴-۷) را با فرم کلی مدل MARS در جهت ساختن مدل زیر ترکیب می‌کنیم

$$Y = \beta_. + X'\beta + \sum_{m=1}^{M}\beta_m h_m(X) + \varepsilon, \qquad (۴-۱۰)$$

که در قسمت خطی X' شامل q متغیر پیش‌بین است و ضرایب β با استفاده از الگوریتم پیش‌رو - پس‌رو MARS (بخش ۳-۳-۳) برآورد می‌شوند. توابع $h_m(X)$ می‌توانند شامل موارد زیر باشند

۱) توابع پایه $(X_j - t)_+$ و $(t - X_j)_+$، به ازای $j = q + ۱, \ldots, p$، که در آن‌ها t ها از مجموعه $\{x_{ij},\ i = ۱, \ldots, n,\ j = q + ۱, \ldots, p\}$ انتخاب می‌شوند.

۲) حاصلضرب دو یا تعداد بیشتری از این توابع پایه.

۳)حاصلضرب توابع پایه در یک یا چند متغیر در قسمت خطی مدل.

به منظور یادآوری در مقایسه اسپلاین هموارساز و روش MARS از نقطه نظر تطبیقی و غیر-تطبیقی می‌توان گفت، چون گره‌ها نقشی کلیدی در این روش‌های ناپارامتری دارند. و به‌دنبال آن در مدل‌های رگرسیونی نیمه‌پارامتری که براساس این روش‌های ناپارامتری ساخته می‌شوند، موثرند. می-توان این دو روش را از این نقطه نظر از هم جدا نمود. همانطور که گفته شد در اسپلاین‌های هموارساز همه‌ی مشاهدات به‌عنوان گره‌ها در نظر گرفته می‌شوند در حالی‌که در روش MARS گره‌ها در مکان-هایی که رفتار تابع تغییر می‌کند براساس یک الگوریتم بهینه سازی تعیین می‌شوند. بنابراین می‌توان مدل‌های نیمه‌پارامتری که براساس روش MARS و اسپلاین‌های هموارساز گفته شد را به‌ترتیب به-عنوان مدل‌های رگرسیونی نیمه‌پارامتری تطبیقی و غیرتطبیقی در نظر گرفت.

فصل پنجم

شبیه‌سازی و موردهای مطالعاتی

۵- ۱- مقدمه

در این فصل، مدل‌های رگرسیونی ذکرشده در فصل‌های ۳ و ۴ را بر روی یک مجموعه از داده‌های شبیه‌سازی‌شده و چند مورد مطالعاتی با داده‌های واقعی به‌کار برده و مقایسه خواهیم کرد. داده‌های واقعی، شامل ۴ مجموعه داده در حیطه‌های مختلف است. دلیل استفاده از این موارد مطالعاتی، صورت مناسب آن‌ها در رابطه با مدل‌های بحث شده در فصول پیشین می‌باشد. برای هر مورد مطالعاتی، ابتدا توضیحات کاملی برای متغیرها داده شده و درادامه به برازش مدل‌های رگرسیونی پرداخته‌ایم.

۵- ۲- مطالعه شبیه‌سازی

در این قسمت جهت مقایسه عملکرد روش‌های بحث شده در فصل ۴ بر روی مدل‌های نیمه‌پارامتری، از شبیه سازی مونت کارلو برای تولید یک مدل نیمه‌پارامتری استفاده می‌کنیم. در جهت شبیه‌سازی به یک مدل نیمه‌پارامتری همچون مدل ۴-۷ توجه می‌کنیم. توابع هموار sin و log را برای قسمت ناپارامتری مدل در نظر می‌گیریم. از این‌رو مشاهدات را مدل زیر شبیه‌سازی می‌کنیم

$$Y = X_1 - X_2 + sin(T_1) + log(T_2) + \varepsilon \qquad (۱-۵)$$

که در آن X_1 و X_2 دارای توزیع یکنواخت $(۰,۱)$ و T_1 و T_2 از توزیع نمایی با میانگین ۱ پیروی می‌کنند. و بردار ε، خطاهای تصادفی از توزیع نرمال با میانگین صفر و واریانس σ^2 هستند. در این شبیه‌سازی سه مقدار ۰/۲، ۰/۵ و ۱ را برای σ^2 در نظر می‌گیریم. تنظیمات فوق را برای حجم نمونه ۳۰۰ با ۱۰۰ بار تکرار انجام می‌دهیم. در هر بار تکرار مقدار مجموع توان‌های دوم خطا را برای مدل نیمه‌پارامتری برای هر روش بدست می‌آوریم و در پایان تکرارها، میانگین مجموع توان‌های دوم خطا را به عنوان معیاری برای مقایسه دو روش برآورد مدل نیمه‌پارامتری شبیه‌سازی شده ۵-۱ استفاده می‌کنیم.

جدول ۵- ۱: برآورد میانگین مجموع توان‌های دوم خطا مدل شبیه‌سازی ۵-۱از دو روش اسپلاین‌های تطبیقی و غیر تطبیقی

حجم نمونه	**۳۰۰**		
σ^2	۰/۲	۰/۵	۱
مقدار برآورد میانگین مجموع توان دوم‌های خطا مدل ۵-۱ ازروش اسپلاین‌های تطبیقی	۱۱/۸۶	۶۷/۵۶۷	۲۴۰/۷۵۱
مقدار برآورد میانگین مجموع توان دوم‌های خطا مدل ۵-۱ از روش اسپلاین‌های غیر تطبیقی	۱۱/۶۵	۷۲/۰۲۶	۲۸۵/۲۴۵

از جدول ۵-۱ می‌توان دریافت که روش اسپلاین‌های تطبیقی با مقادیر برآورد میانگین مجموع توان-های دوم خطا کمتر، برآورد بهتری را نسبت به روش اسپلاین‌های غیرتطبیقی برای مدل نیمه‌پارامتری ۵-۱ داشته است. مقدار ضریب تعیین برآورد مدل نیمه‌پارامتری ۵-۱ از دو روش اسپلاین‌های تطبیقی و غیرتطبیقی برای زمانی که واریانس خطا ۰/۵ است، به‌ترتیب ۰/۸۹ و ۰/۸۸ می‌باشد. همچنین همانطور که از جدول ۵-۱ پیداست با افزایش مقدار واریانس خطا روش اسپلاین‌های تطبیقی برآورد بهتری نسبت به اسپلاین‌های غیرتطبیقی برای مدل نیمه‌پارامتری ۵-۱ ارائه داده است.

۵- ۳- مورد مطالعاتی ۱: داده‌های قیمت مسکن[۱]

این مثال شامل داده‌های قیمت ۹۲ خانه فروخته شده در سال ۱۹۸۷ در منطقه اتووا[۲] است که دارای ۷ متغیر پیش‌بین می‌باشد. در این مثال، متغیر پاسخ، قیمت فروش (Y) و متغیرهای پیش‌بین شامل متغیرهای کمی: فاصله تا بزرگراه (D)، میانگین درآمد محله (A)، متراژ خانه (S)، اندازه زمین (LT) و متغیرهای کیفی، وجود شومینه (F) و وجود گاراژ (G) می‌باشند. نمودار پراکنش متغیر پاسخ با هر یک از متغیرهای پیش‌بین در شکل ۵-۱ به صورت ماتریسی آمده است، به وضوح دیده می شود که متغیرهای پیش‌بین L و D رابطه غیرخطی با متغیر پاسخ دارند.

هدف، مقایسه عملکرد سه مدل پارامتری، نیمه‌پارامتری ساده غیرتطبیقی و مدل ناپارامتری MARS است. مدل پارامتری ساده برای این داده‌ها به‌صورت زیر می‌باشد.

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 F_i + \beta_3 G_i + \beta_4 A_i + \beta_5 S_i + \beta_6 L_i + \varepsilon_i$$

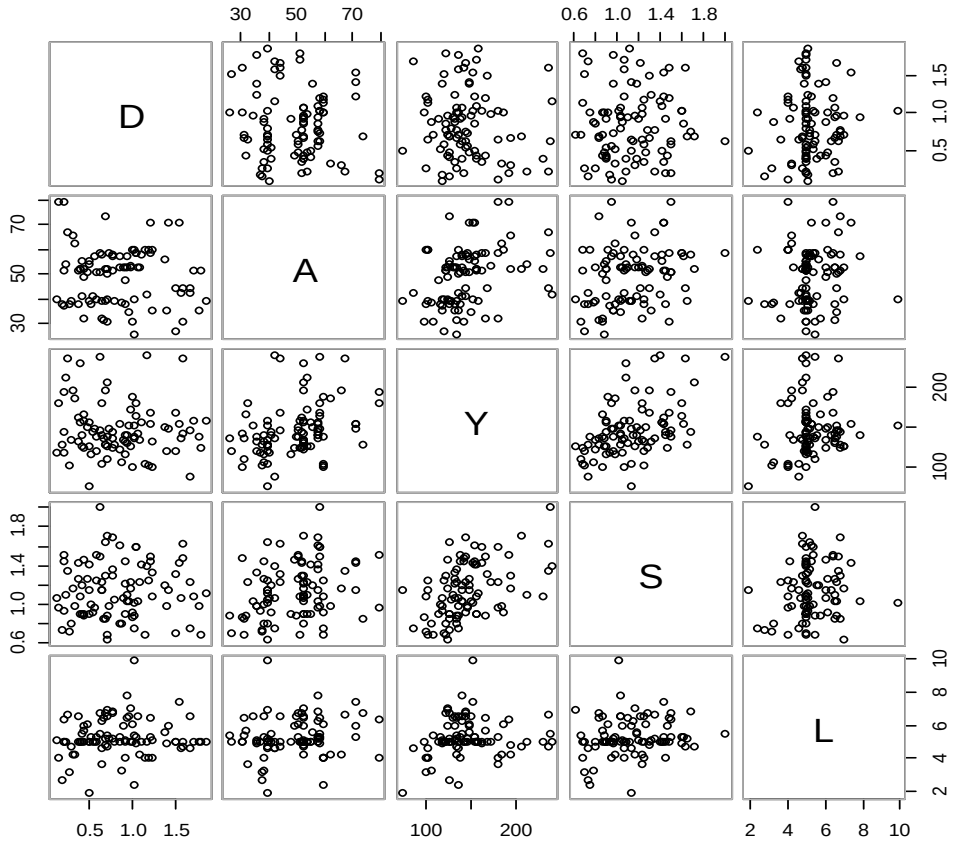
آکنادیز[۳] و تاباکان[۴] (۲۰۰۹) مدل نیمه‌پارامتری زیر را برای این داده‌ها پیشنهاد کردند.

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 F_i + \beta_3 G_i + \beta_4 A_i + \beta_5 S_i + f(L)_i + \varepsilon_i \quad (۵-۲)$$

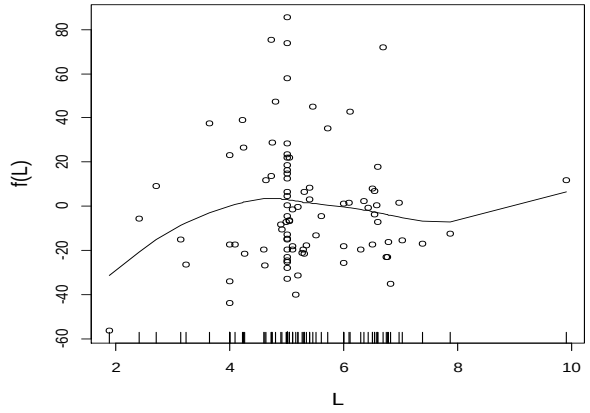
آن‌ها از اسپلاین هموارساز به‌عنوان هموارکننده نمودار پراکنش جهت برآورد تابع f استفاده کردند. در واقع می‌توان گفت که یک مدل نیمه‌پارامتری غیرتطبیقی برای این داده‌ها ارائه دادند. شکل ۵-۲ هموارسازی از طریق اسپلاین هموارساز برای متغیر L را نشان می‌دهد.

[۱] Housing prices data
[۲] Ottowa
[۳] Aknadiz
[۴] Tabakan

آن‌ها نشان دادند که مدل نیمه‌پارامتری (۵-۲) اجرای خیلی بهتری نسبت به مدل پارامتری دارد به این مفهوم که دارای مجموع توان‌های دوم خطای کوچکتر و ضریب تعیین بزرگتری نسبت به مدل پارامتری است.



شکل ۵- ۱: رابطه بین متغیر پاسخ Y و متغیرهای پیش‌بین (مورد مطالعاتی ۱)



شکل۵- ۲: نمودارهای هموار برپایه اسپلاین هموارساز برای مدل نیمه پارامتری(۵-۲)

برای برازش مدل MARS، همانطور که در توضیح بخش ۳-۳-۳ ذکر شد، دو پارامتر کلیدی این روش یعنی، تعداد توابع پایه در گام پیش‌رو (nk) و درجه اثرات متقابل (deg) باید بهینه شوند. با در نظر گرفتن $nk = ۱۱, ۱۲, \dots, ۲۰$ و $deg = ۱,۲,۳$، نتایج جدول ۵-۲ به‌دست آمده است.

جدول ۵-۲: مقادیر R-square، GCV و RSS برای ۳۰ مدل MARS براساس مقادیر مختلف nk و deg (مورد مطالعاتی ۱)

model	nk	Deg	R-square	GCV	RSS
۱	۱۱	۱	۰,٤٥٦٢٧١	٧٢١,٠٨٣٤	٥٥٣٠٣,٩٦
۲	۱۲	۱	۰,٤٥٦٢٧١	٧٢١,٠٨٣٤	٥٥٣٠٣,٩٦
۳	۱۳	۱	۰,٤٨٧٦٩	٧١٢,٩٦٢٧	٥٢١٠٨,٢٧
٤	١٤	۱	۰,٤٨٧٦٩	٧١٢,٩٦٢٧	٥٢١٠٨,٢٧
٥	١٥	۱	۰,٤٨٧٦٩	٧١٢,٩٦٢٧	٥٢١٠٨,٢٧
٦	١٦	۱	۰,٤٨٧٦٩	٧١٢,٩٦٢٧	٥٢١٠٨,٢٧
٧	١٧	۱	۰,٤٨٧٦٩	٧١٢,٩٦٢٧	٥٢١٠٨,٢٧
٨	١٨	۱	۰,٤٨٧٦٩	٧١٢,٩٦٢٧	٥٢١٠٨,٢٧
٩	١٩	۱	۰,٤٨٧٦٩	٧١٢,٩٦٢٧	٥٢١٠٨,٢٧
١٠	٢٠	۱	۰,٤٨٧٦٩	٧١٢,٩٦٢٧	٥٢١٠٨,٢٧
١١	١١	۲	۰,٤٨٩٤٨٨	٦٤٥,٩٠٧٦	٥١٩٢٥,٣٦
١٢	١٢	۲	۰,٤٨٩٤٨٨	٦٤٥,٩٠٧٦	٥١٩٢٥,٣٦
١٣	١٣	۲	۰,٥٣٣١٩٧	٦٠٤,٥٨٥١	٤٧٤٧٩,٦٤
١٤	١٤	۲	۰,٥٣٣١٩٧	٦٠٤,٥٨٥١	٤٧٤٧٩,٦٤
١٥	١٥	۲	۰,٥٣٣١٩٧	٦٠٤,٥٨٥١	٤٧٤٧٩,٦٤
١٦	١٦	۲	۰,٥٣٣١٩٧	٦٠٤,٥٨٥١	٤٧٤٧٩,٦٤
١٧	١٧	۲	۰,٥٣٣١٩٧	٦٠٤,٥٨٥١	٤٧٤٧٩,٦٤
١٨	١٨	۲	۰,٥٣٣١٩٧	٦٠٤,٥٨٥١	٤٧٤٧٩,٦٤
١٩	١٩	۲	۰,٥٥١٠٣٩	٥٩٥,٤٠٤١	٤٥٦٦٤,٩١
٢٠	٢٠	۲	۰,٥٥١٠٣٩	٥٩٥,٤٠٤١	٤٥٦٦٤,٩١
٢١	١١	۳	۰,٥١٨٨١١	٦٢٣,٢١٧٩	٤٨٩٤٢,٩٣
٢٢	١٢	۳	۰,٥١٨٨١١	٦٢٣,٢١٧٩	٤٨٩٤٢,٩٣
٢٣	١٣	۳	۰,٥٥٧٨٦٨	٥٨٦,٣٤٨	٤٤٩٧٠,٣٤
٢٤	١٤	۳	۰,٥٥٧٨٦٨	٥٨٦,٣٤٨	٤٤٩٧٠,٣٤
٢٥	١٥	۳	۰,٥٨٢٦٧٨	٥٦٦,٨٦٢	٤٢٤٤٦,٨٧
٢٦	١٦	۳	۰,٥٨٢٦٧٨	٥٦٦,٨٦٢	٤٢٤٤٦,٨٧
٢٧	١٧	۳	۰,٦١٩٨٨٤	٥٤٢,١٣٦٣	٣٨٦٦٢,٥٧
٢٨	١٨	۳	۰,٦١٩٨٨٤	٥٤٢,١٣٦٣	٣٨٦٦٢,٥٧
٢٩	١٩	۳	۰,٦٣٠٠٩٩	٥٤٠,٨٣٧٩	٣٧٦٢٣,٥
٣٠	٢٠	۳	۰,٦٣٠٠٩٩	٥٤٠,٨٣٧٩	٣٧٦٢٣,٥

همانطور که در جدول ۵-۲ مشاهده می‌شود، بهترین مدل که دارای بیشترین مقدار ضریب تعیین (R-square) و کمترین مقدار اعتبارسنجی متقابل تعمیم‌یافته (GCV) و مجموع توان‌های دوم خطا (RSS) است، با تنظیمات $nk = ۱۹$ و $deg = ۳$ حاصل می‌شود. مدل نهایی MARS بعد از طی مراحل پیش‌رو - پس‌رو به صورت زیر است:

$$\begin{aligned} SP = {} & ۱۴۲.۷۸ + ۱۵.۸۷G - ۴۰.۲h(۱.۵ - S) + ۷.۲۵h(۰.۷۷ - D)h(A - ۳۸.۱۲) \\ & + ۱۱۲۷.۳۷h(D - ۰.۷۸)h(S - ۱.۵) + ۲۵۶۰.۰۸h(۰.۷۸ - D) \\ & - ۲۸.۱۷h(S - ۱.۵)h(A - ۵۲.۳۵)h(S - ۱.۵) - ۳۶.۷۲h(۵۲.۳۵ - A)h(S \\ & - ۱.۵) + ۴.۷۸h(۳۸.۱۲ - A)h(۴.۸ - L) - ۲۵.۱۶h(۰.۷۷ - D)h(A \\ & - ۳۸.۱۲)h(S - ۱.۰۸) - ۳۷.۲۲h(۰.۷۷ - D)h(A - ۳۸.۱۲)h(۱.۰۸ - S) \\ & + ۲.۵۷h(۰.۷۷ - D)h(A - ۳۸.۱۲)h(L - ۴.۶) \end{aligned}$$

که در آن

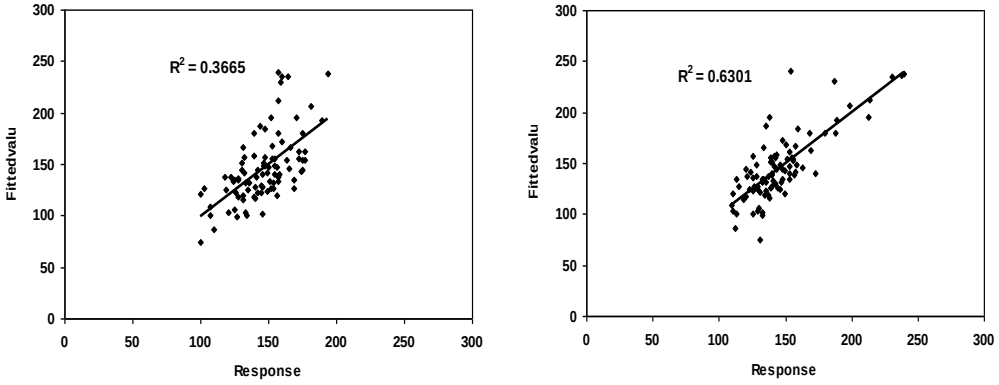
$$h(u) = u_+ = \begin{cases} u & \text{اگر } u > ۰ \\ ۰ & \text{درغیراینصورت} \end{cases}$$

اگرچه مدل فوق یک مدل ناپارامتری است، اما این مدل نیز می‌تواند به‌عنوان یک مدل نیمه‌پارامتری تطبیقی با وجود متغیر G در قسمت خطی و بقیه متغیرها در قسمت ناپارامتری مورد توجه باشد. جدول ۵-۳، مجموع توان‌های دوم خطا را برای سه مدل پارامتری، نیمه‌پارامتری و مدل MARS برازش داده‌شده به داده‌ها نشان می‌دهد که در آن مدل MARS دارای کمترین مقدار مجموع توان-های دوم خطا است.

جدول ۵-۳: مجموع توان‌های دوم خطا برای سه مدل (مورد مطلعاتی ۱)

مدل	مجموع توان دوم‌های خطا
پارامتری	۶۷۸۵۵/۴۲
نیمه‌پارامتری	۶۶۴۷۲
MARS	۳۷۶۲۳/۵

عملکرد مدل‌های رگرسیونی نیمه‌پارامتری و MARS را می‌توان در شکل ۵-۳ به‌وسیله نمودارهای مقادیر پیش‌بینی‌شده در مقابل مقادیر پاسخ، دید. مدل MARS برازش بهتری را به داده‌ها انجام داده است.

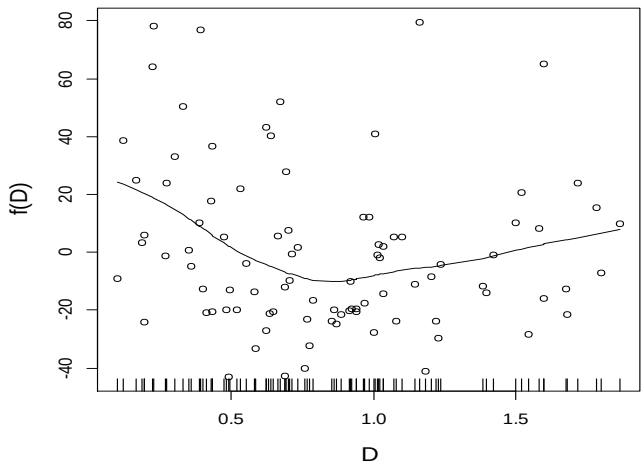


شکل ۵-۳: نمودار مقادیر پاسخ در مقابل مقادیر برازش داده شده برای مدل MARS (قاب سمت راست) و مدل نیمه-پارامتری ۵-۲ (قاب سمت چپ) در مورد مطالعاتی ۱

همانطور که گفته شد، متغیر D نیز با متغیر پاسخ ارتباط غیرخطی دارد (شکل ۵-۱). به عنوان یک جایگزین احتمالی برای مدل رگرسیونی نیمه‌پارامتری (۵-۲) که آکنادیز و تاباکان (۲۰۰۹) به داده‌ها برازش دادند، می‌توان مدل نیمه‌پارامتری زیر را برای این داده‌ها پیشنهاد کردیم

$$Y_i = \beta_0 + \beta_1 L_i + \beta_2 F_i + \beta_3 G_i + \beta_4 A_i + \beta_5 S_i + f(D)_i + \varepsilon_i \qquad (۳-۵)$$

شایان ذکر است که هر دو مدل‌های (۵-۲) و (۵-۳) نیمه‌پارامتری ساده است و در اینجا ما نیز همانند آکنادیز و تاباکان در روشی مشابه از اسپلاین‌هموارساز به عنوان هموارساز نمودار پراکنش استفاده کرده‌ایم. شکل ۵-۴ هموارسازی به‌وسیله اسپلاین هموارساز برای متغیر D را نشان می‌دهد.



شکل۵-۴: نمودارهای هموار برپایه اسپلاین هموارساز برای مدل نیمه پارامتری(۵-۳)

مقدار ضریب تعیین و مجموع توان‌های دوم خطا برای مدل نیمه‌پارامتری (۵-۲) که توسط آکنادیز و تاباکان برای این داده‌ها پیشنهاد شده بود به ترتیب ۰/۳۷ و ۶۶۴۷۲ است. در حالی که برای مدل نیمه‌پارامتری پیشنهادی (۵-۳) این مقادیر به‌ترتیب ۰/۴۱ و ۶۰۴۶۸/۴۹ می‌باشند.

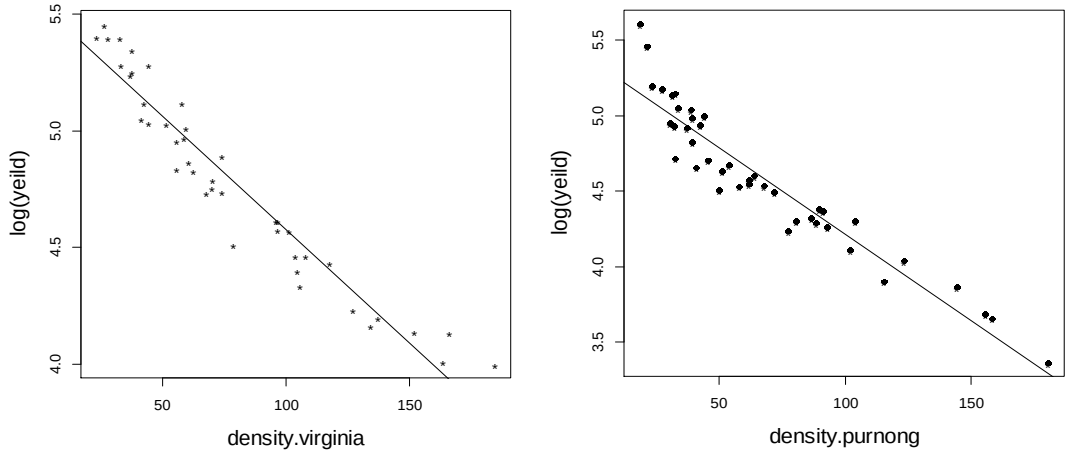
در ادامه می‌خواهیم مدل رگرسیونی نیمه‌پارامتری غیرتطبیقی و مدل رگرسیونی نیمه‌پارامتری تطبیقی را برای سه مجموعه داده متفاوت، برازش داده و با هم مقایسه کنیم.

۵- ۴- مورد مطالعاتی ۲: داده‌های پیاز

داده‌های پیاز شامل ۸۴ مشاهده از یک آزمایش تجربی مبنی بر تولید پیاز سفید اسپانیایی در دو منطقه لنگرگاه پرنونگ [۱] و ویرجینیا [۲] استرالیای جنوبی، می‌باشد. این مورد مطالعاتی یکی از مثال-های کتاب رگرسیون نیمه‌پارامتری روپرت[۳] (۲۰۰۳) است و مجموعه داده‌ها در آدرس الکترونیکی زیر در دسترس هستند.

http://www.uow.edu.au/~mwand/webspr/data.html

شکل ۵-۵ نمودار پراکنش لگاریتم تولید پیاز سفید اسپانیایی (متغیر پاسخ) را در مقابل تراکم گیاه (متغیر پیش‌بین) در دو منطقه نشان می‌دهد.



شکل ۵- ۵: نمودار پراکنش لگاریتم تولید در مقابل تراکم برای داده‌های پیاز همراه با برازش مدل خطی تعمیم یافته (نمودار سمت راست مربوط به منطقه لنگرگاه پرنونگ و نمودار سمت چپ مربوط به منطقه ویرجینیا می‌باشد)

[۱] Purnong Landing
[۲] Virginia
[۳] Ruppert

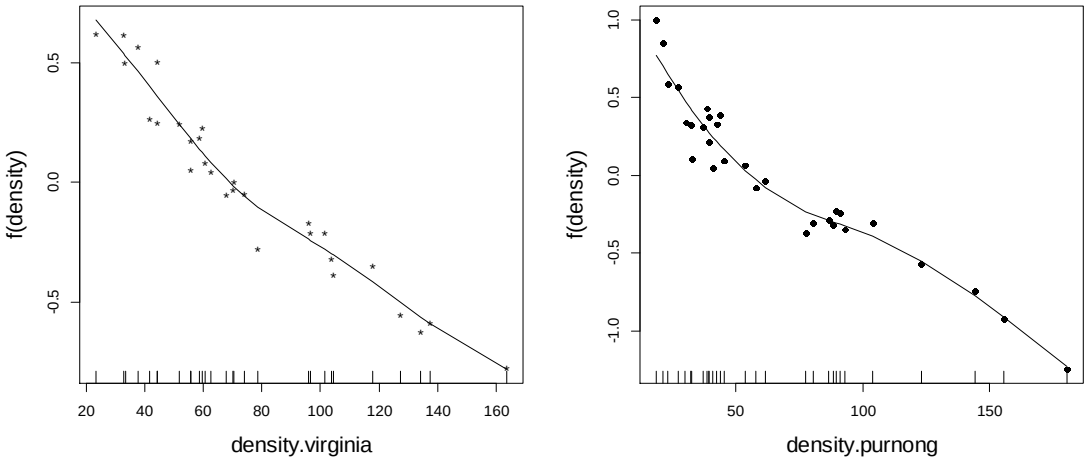
خط رسم شده در شکل ۵-۵ مربوط به برازش مدل خطی تعمیم‌یافته زیر است

$$\eta = \beta_0 + \beta_1 PL + \beta_2 density,$$

که در آن $\eta = logE(yield)$ و $PL = 0$ اگر اندازه مربوط به ویرجینیا باشد و $PL = 1$ است اگر اندازه متعلق به لنگرگاه پرنونگ باشد. یک بررسی بر روی شکل ۵-۵، مقداری انحنای مشهود را در نمودار پراکنش برای هر منطقه نشان می‌دهد. بنابراین می‌توان مدل نیمه‌پارامتری زیر را پیشنهاد کرد

$$\eta = \beta_0 + \beta_1 PL + f(density), \qquad (۴-۵)$$

که دارای مولفه ناپارامتری $f(density)$ و مولفه پارامتری $\beta_1 PL$ می‌باشد. متغیر دودویی PL رابطه بین تولید و تراکم مربوط به هر منطقه را نشان می‌دهد. برای بررسی مزایای مدل رگرسیونی نیمه-پارامتری تطبیقی نسبت به مدل رگرسیونی نیمه‌پارامتری غیرتطبیقی، مجموعه داده‌ها را به دو قسمت داده‌های مدل‌ساز و داده‌های آزمون، به ترتیب به نسبت ۷۰ به ۳۰ تقسیم می‌کنیم. یکی از مزایای این تقسیم‌بندی تعیین مدل بهتر برای پیش‌گویی است. ابتدا مدل رگرسیونی نیمه‌پارامتری غیرتطبیقی (۴-۵) را به داده‌های مدل‌ساز برازش می‌دهیم که نتیجه این برازش در شکل ۵-۶ نشان داده شده است.



شکل۵-۶: نمودار هموار برای متغیر تراکم براساس اسپلاین هموارساز در دو منطقه لنگرگاه پرنونگ (نمودار سمت راست) و ویرجینیا (نمودار سمت چپ)

شکل ۵-۶ برازش مناسبی را بر پایه اسپلاین هموارساز برای داده‌های مدل‌ساز نشان می‌دهد. برای برازش مدل نیمه‌پارامتری تطبیقی به داده‌های مدل‌ساز، همانطور که گفته شد دو پارامتر کلیدی

تعداد توابع پایه (nk) در مرحله پیش‌رو و درجه اثرات متقابل (deg)، باید بهینه شوند. نتیجه برازش ۸۰ مدل با $nk = ۱۱,۱۲, \dots ,۵۰$ و $deg = ۱,۲$ در جدول ۵-۴ ارائه شده است که از بین آن‌ها مدل ۴۵ با تنظیمات $nk = ۱۵$ و $deg = ۲$ یک مدل بهینه است، به این مفهوم که دارای بیشترین مقدار ضریب تعیین و کمترین مقدار GCV و RSS می‌باشد. این مدل علاوه بر اینکه یک مدل بهینه در مجموعه داده‌های مدل‌ساز می‌باشد، برای داده‌های آزمون نیز با مقدار ضریب تعیین ۰/۹۴۲۷، برازش خوبی داشته است.

جدول ۵-۴: مقادیر R-square، GCV و RSS برای ۸۰ مدل MARS نیمه‌پارامتری براساس مقادیر مختلف nk و deg (مورد مطالعاتی۲

			Training set			Test set
model	**nk**	**deg**	**R-square**	**GCV**	**RSS**	**R-square**
۱	۱۱	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۲	۱۲	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۳	۱۳	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
٤	۱٤	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
٥	۱٥	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
٦	۱٦	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۷	۱۷	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۸	۱۸	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۹	۱۹	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱۰	۲۰	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱۱	۲۱	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱۲	۲۲	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱۳	۲۳	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱٤	۲٤	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱٥	۲٥	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱٦	۲٦	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱۷	۲۷	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱۸	۲۸	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۱۹	۲۹	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۲۰	۳۰	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۲۱	۳۱	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۲۲	۳۲	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۲۳	۳۳	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۲٤	۳٤	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤
۲٥	۳٥	۱	۰,۹٤۸۷۷٤	۰,۰۱۲۸۰٤	۰,٦۰۹٥۹٥	۰,۹۳۸۰۰۱٤

۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	۳٦	۲٦
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	۳۷	۲۷
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	۳۸	۲۸
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	۳۹	۲۹
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤۰	۳۰
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤۱	۳۱
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤۲	۳۲
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤۳	۳۳
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤٤	۳٤
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤٥	۳٥
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤٦	۳٦
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤۷	۳۷
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤۸	۳۸
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٤۹	۳۹
۰,۹۳۸۰۰۱٤	۰,٦۰۹٥۹٥	۰,۰۱۲۸۰٤	۰,۹٤۸۷۷٤	۱	٥۰	٤۰
۰,۹٤۱۷۱٤٤	۰,٥۸٤۱۱٦	۰,۰۱۲۷٤٥	۰,۹٥۰۹۱٥	۲	۱۱	٤۱
۰,۹٤۱۷۱٤٤	۰,٥۸٤۱۱٦	۰,۰۱۲۷٤٥	۰,۹٥۰۹۱٥	۲	۱۲	٤۲
۰,۹٤۱۷۱٤٤	۰,٥۸٤۱۱٦	۰,۰۱۲۷٤٥	۰,۹٥۰۹۱٥	۲	۱۳	٤۳
۰,۹٤۱۷۱٤٤	۰,٥۸٤۱۱٦	۰,۰۱۲۷٤٥	۰,۹٥۰۹۱٥	۲	۱٤	٤٤
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۱٥	٤٥
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۱٦	٤٦
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۱۷	٤۷
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۱۸	٤۸
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۱۹	٤۹
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲۰	٥۰
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲۱	٥۱
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲۲	٥۲
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲۳	٥۳
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲٤	٥٤
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲٥	٥٥
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲٦	٥٦
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲۷	٥۷
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲۸	٥۸
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۲۹	٥۹
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۳۰	٦۰
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۳۱	٦۱
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۳۲	٦۲
۰,۹٤۲۷۱۰۱	۰,٥٥٦۹۸۱	۰,۰۱۲٦۳٤	۰,۹٥۳۱۹٥	۲	۳۳	٦۳

٦٤	٣٤	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٦٥	٣٥	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٦٦	٣٦	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٦٧	٣٧	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٦٨	٣٨	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٦٩	٣٩	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٠	٤٠	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧١	٤١	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٢	٤٢	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٣	٤٣	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٤	٤٤	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٥	٤٥	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٦	٤٦	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٧	٤٧	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٨	٤٨	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٧٩	٤٩	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١
٨٠	٥٠	٢	٠,٩٥٣١٩٥	٠,٠١٢٦٣٤	٠,٥٥٦٩٨١	٠,٩٤٢٧١٠١

مدل رگرسیونی نیمه‌پارامتری تطبیقی با تنظیمات فوق به‌صورت زیر است

$$log(yield) = ۵.۲ - ۰.۳۴PL - ۰.۰۱h(density - ۳۸.۵۵) + ۰.۰۲h(۳۸.۵۵ - density) + ۰.۰۰۶h(density - ۶۷.۷۱) + ۰.۰۱h(۳۸.۵۵ - density)PL + ۰.۰۱h(density - ۸۸.۴۶)PL - ۰.۰۱h(density - ۱۰۳.۷۸)PL. \quad (۵-۵)$$

نتایج برازش هر دو مدل در جدول۵-۵ خلاصه شده است. می‌توان دریافت که مدل (۵-۵) دارای مجموع توان‌های دوم خطای کوچکتری برای مجموعه داده‌های مدل‌ساز و ضریب تعیین بزرگتری برای هر دو مجموعه داده‌های مدل‌ساز و آزمون است. از این رو می‌تواند مدل بهتری نسبت به مدل نیمه‌پارامتری غیرتطبیقی (۵-۴) باشد.

جدول۵- ۵: مجموع توان‌های دوم خطا و ضریب تعیین مجموعه داده‌های مدل‌ساز و آزمون برای داده‌های پیاز

مدل	مجموعه داده‌های مدل‌ساز		مجموعه داده‌های آزمون
	مجموع توان-های دوم خطا	ضریب تعیین	ضریب تعیین
نیمه‌پارامتری تطبیقی (۵-۵)	۰/۵۵۶۹۸	۰/۹۵۳۲	۰/۹۴۲۷
نیمه‌پارامتری غیرتطبیقی (۵-۴)	۰/۶۱۳۰۸	۰/۹۴۸۵	۰/۹۳۸۵

۵- ۵- مورد مطالعاتی ۳: داده‌های ابروسیا[۱]

داده‌های ابروسیا[۲]، اطلاعاتی شامل سطوح ابروسیا و متغیرهای هواشناسی برای ۳۳۵ روز در منطقه کالامازوو[۳]میشیگان ایالات متحده آمریکا در سال ۱۹۹۳ می‌باشد. چون گیاه ابروسیا نقش بزرگی در آلرژی ناشی از گرده افشانی ایفا می‌کند یک موضوع مهم در هواشناسی، گسترش مدل‌های پیش‌بینی دقیق برای سطوح گرده‌افشانی روزانه این گیاه است. این مورد مطالعاتی نیز یکی از مثال‌های کتاب رگرسیون نیمه پارامتری روپرت (۲۰۰۳) می‌باشد و مجموعه داده‌ها در آدرس الکترونیکی زیر در دسترس است.

http://www.uow.edu.au/~mwand/webspr/data.html

مجموعه داده‌ها شامل متغیرهای زیر است:

ragweed: سطوح ابروسیا در هر روز (دانه در هر متر مکعب)

T (Temperature): دما در روز مورد بررسی (برحسب فارنهایت)

R (Rain): مقیاس معنی داری بارش باران در روز مورد بررسی ($R = ۱$ یعنی بارش باران کمتر از ۳ ساعت به‌طور یکنواخت یا کوتاه اما شدید و $R = ۰$ یعنی اگر بارش باران بیشتر از ۳ ساعت به طول انجامد)

S (Speed): سرعت پیش‌بینی باد برای روز مورد بررسی

day (day.in.seas): تعداد روزهای گرده‌افشانی ابروسیا در فصل جاری

متغیر ragweed (ابروسیا)، متغیر پاسخ و بقیه متغیرها، متغیرهای پیش‌بین هستند. رابطه بین متغیر پاسخ ابروسیا و متغیرهای پیش‌بین در شکل ۵-۷ نشان داده شده است. در این نمودارها چولگی متغیر ابروسیا به وضوح دیده می‌شود. لذا بهتر است ریشه دوم متغیر ابروسیا را به عنوان متغیر پاسخ در نظر بگیریم. با روشی همانند مورد مطالعاتی قبل، داده‌های این مثال را نیز به دو قسمت داده مدل‌ساز و آزمون تقسیم می‌کنیم. رابطه بین متغیر پاسخ و متغیرهای پیش‌بین در شکل ۵-۸ نشان داده شده

[۱] Ragweed data

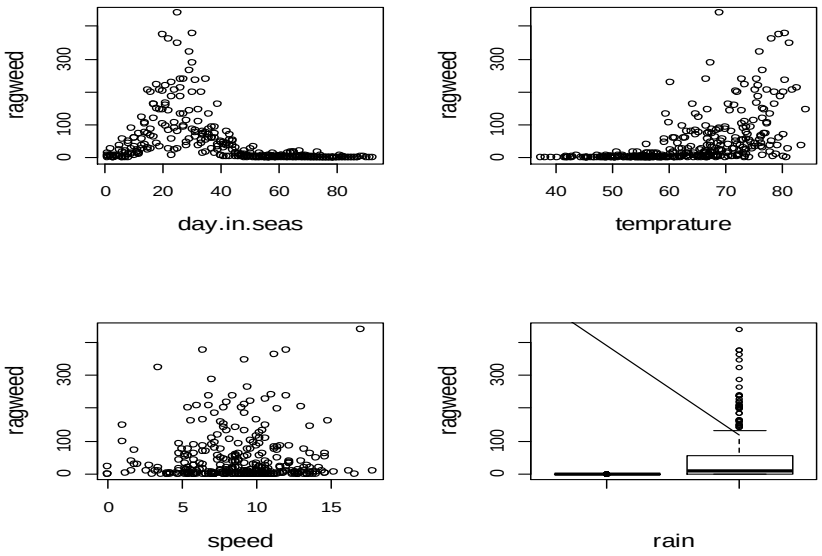
[۲] نوعی گیاه از خانواده گل آفتابگردان است که بیشتر در گرمترین مناطق نیمکره شمالی و همچنین آمریکای جنوبی یافت می‌شود و زمانی که بسیار حساسیت زا است به‌خصوص با قرار گرفتن گرده‌های این گیاه در معرض آلاینده‌های هوا از قبیل ازن در سطح زمین باعث می‌شود که افراد حساسیت بیشتری نسبت به گرده این گیاه بروز دهند.

[۳] Kalamazoo

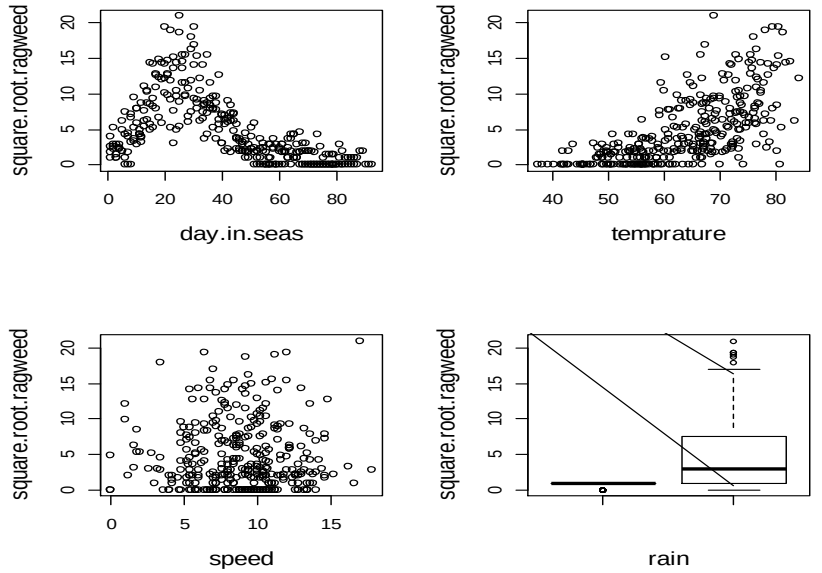
است. چون متغیر day یک رابطه غیرخطی با متغیر پاسخ دارد. بنابراین یک مدل رگرسیونی مفید برای این داده‌ها به‌صورت زیر می‌باشد

$$E\left(\sqrt{ragweed}\right) = \beta_{.} + \beta_{1}R + \beta_{2}T + \beta_{3}S + f(day), \qquad (۶-۵)$$

که در آن f را می‌توان براساس اسپلاین هموارساز برآورد کرد.

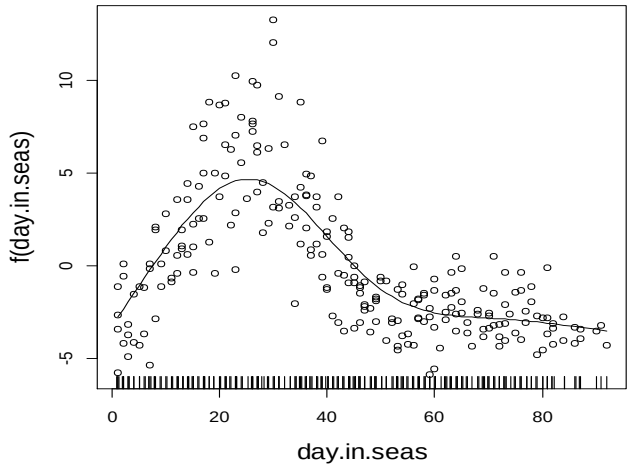


شکل ۵- ۷: رابطه بین متغیر ابروسیا و متغیرهای پیش‌بین در داده‌های ابروسیا



شکل ۵-۸: رابطه بین متغیر ریشه دوم ابروسیا و متغیرهای پیش‌بین در داده‌های ابروسیا

در شکل ۵-۹ نمودار برآورد f برای داده‌های مدل‌ساز نشان داده شده است که به نظر می‌آید برازش مناسبی باشد.



شکل ۵- ۹: نمودار هموار برای متغیر روز گرده افشانی براساس اسپلاین هموارساز

در ادامه یک مدل رگرسیونی نیمه‌پارامتری تطبیقی را با تنظیمات متفاوتی برازش داده که در نهایت مدلی با $nk = ۲۹$ و $deg = ۲$ به عنوان مدل بهینه با داشتن کمترین مقدار GCV و RSS و بیشترین مقدار R-square به داده‌های مدل‌ساز انتخاب شده است (جدول ۵-۶). همچنین این مدل با ضریب تعیین ۰/۸۲۸ برای داده‌های آزمون، یک مدل مناسب در میان سایر مدل‌ها می‌باشد.

جدول۵-۶: مقادیر R-square، GCVو RSS برای ۸۰ مدل MARS نیمه‌پارامتری براساس مقادیر مختلف nk و deg (مورد مطالعاتی ۳)

			Training set			Test set
model	**nk**	**deg**	**R-square**	**GCV**	**RSS**	**R-square**
۱	۱۱	۱	۰,۷۷۳۹٤٦	٤,۷۳۳۲۱۷	۱۰٤۷,۰۲۸	۰,۷۷۰٦٥٥
۲	۱۲	۱	۰,۷۷۳۹٤٦	٤,۷۳۳۲۱۷	۱۰٤۷,۰۲۸	۰,۷۷۰٦٥٥
۳	۱۳	۱	۰,۷۸۲۲۰۱	٤,٦۰۰٦٤۸	۱۰۰۸,۷۹٥	۰,۷۸۹۹۸٦
٤	۱٤	۱	۰,۷۸۲۲۰۱	٤,٦۰۰٦٤۸	۱۰۰۸,۷۹٥	۰,۷۸۹۹۸٦
٥	۱٥	۱	۰,۷۸٦٤۹٤	٤,٥۰۹۹٥۸	۹۸۸,۹۰۹	۰,۷۸۹۱٤۷
٦	۱٦	۱	۰,۷۸٦٤۹٤	٤,٥۰۹۹٥۸	۹۸۸,۹۰۹	۰,۷۸۹۱٤۷
۷	۱۷	۱	۰,۷۸٦٤۹٤	٤,٥۰۹۹٥۸	۹۸۸,۹۰۹	۰,۷۸۹۱٤۷
۸	۱۸	۱	۰,۷۸٦٤۹٤	٤,٥۰۹۹٥۸	۹۸۸,۹۰۹	۰,۷۸۹۱٤۷
۹	۱۹	۱	۰,۷۸٦٤۹٤	٤,٥۰۹۹٥۸	۹۸۸,۹۰۹	۰,۷۸۹۱٤۷
۱۰	۲۰	۱	۰,۷۸٦٤۹٤	٤,٥۰۹۹٥۸	۹۸۸,۹۰۹	۰,۷۸۹۱٤۷
۱۱	۲۱	۱	۰,۷۸٦٤۹٤	٤,٥۰۹۹٥۸	۹۸۸,۹۰۹	۰,۷۸۹۱٤۷

۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۲۲	۱۲
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۲۳	۱۳
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۲٤	۱٤
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۲٥	۱٥
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۲٦	۱٦
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۲۷	۱۷
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۲۸	۱۸
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۲۹	۱۹
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳۰	۲۰
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳۱	۲۱
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳۲	۲۲
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳۳	۲۳
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳٤	۲٤
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳٥	۲٥
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳٦	۲٦
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳۷	۲۷
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳۸	۲۸
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	۳۹	۲۹
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤۰	۳۰
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤۱	۳۱
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤۲	۳۲
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤۳	۳۳
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤٤	۳٤
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤٥	۳٥
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤٦	۳٦
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤۷	۳۷
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤۸	۳۸
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٤۹	۳۹
۰,۷۸۹۱۴۷	۹۸۸,۹۰۹	٤,٥۰۹۹٥۸	۰,۷۸٦٤۹٤	۱	٥۰	٤۰
۰,۷۷۰٦٥٥	۱۰٤۷,۰۲۸	٤,۷۳۳۲۱۷	۰,۷۷۳۹٤٦	۲	۱۱	٤۱
۰,۷۷۰٦٥٥	۱۰٤۷,۰۲۸	٤,۷۳۳۲۱۷	۰,۷۷۳۹٤٦	۲	۱۲	٤۲
۰,۷۸۷۳٤۱	۱۰۱۳,۲۳۳	٤,٥۸۰٤٤۲	۰,۷۸۱۲٤۳	۲	۱۳	٤۳
۰,۷۸۷۳٤۱	۱۰۱۳,۲۳۳	٤,٥۸۰٤٤۲	۰,۷۸۱۲٤۳	۲	۱٤	٤٤
۰,۷۸۷٤٥۷	۹۷۸,۰۷۱	٤,٥۰۰۰۹۲	۰,۷۸۸۸۳	۲	۱٥	٤٥
۰,۷۸۷٤٥۷	۹۷۸,۰۷۱	٤,٥۰۰۰۹۲	۰,۷۸۸۸۳	۲	۱٦	٤٦
۰,۷۹٥۹۹۱	۹۷۲,٦٥۹۸	٤,٤۳٥۸٥۳	۰,۷۹۰۰۰۲	۲	۱۷	٤۷
۰,۷۹٥۹۹۱	۹۷۲,٦٥۹۸	٤,٤۳٥۸٥۳	۰,۷۹۰۰۰۲	۲	۱۸	٤۸
۰,۸۱۱۲۷	۹۱۹,۹۳۰۳	٤,۳۰۸٥۰۷	۰,۸۰۱۳۸۷	۲	۱۹	٤۹

۵۰	۲۰	۲	۰,۸۰۱۳۸۷	٤,۳۰۸۵۰۷	۹۱۹,۹۳۰۳	۰,۸۱۱۲۷
۵۱	۲۱	۲	۰,۸۱٤۹۵۵	٤,۰۵۰۲٤۲	۸۵۷,۰۸۲۹	۰,۸۱۱۹۷٤
۵۲	۲۲	۲	۰,۸۱٤۹۵۵	٤,۰۵۰۲٤۲	۸۵۷,۰۸۲۹	۰,۸۱۱۹۷٤
۵۳	۲۳	۲	۰,۸۱۳۳٤٤	٤,۰٤۹۰۹٤	۸٦٤,۵٤۱۹	۰,۸۱۸۰۳۸
۵٤	۲٤	۲	۰,۸۱۳۳٤۵	٤,۰٤۹۰۹٤	۸٦٤,۵٤۱۹	۰,۸۱۸۰۳۸
۵۵	۲۵	۲	۰,۸۱۳۳٤۵	٤,۰٤۹۰۹٤	۸٦٤,۵٤۱۹	۰,۸۱۸۰۳۸
۵٦	۲٦	۲	۰,۸۱۳۳٤۵	٤,۰٤۹۰۹٤	۸٦٤,۵٤۱۹	۰,۸۱۸۰۳۸
۵۷	۲۷	۲	۰,۸۱۷۳۱٤	٤,۰۳٤۷۳	۸٤٦,۱٦۰۱	۰,۸۲۲٦۵۷
۵۸	۲۸	۲	۰,۸۱۷۳۱٤	٤,۰۳٤۷۳	۸٤٦,۱٦۰۱	۰,۸۲۲٦۵۷
۵۹	۲۹	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦۰	۳۰	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦۱	۳۱	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦۲	۳۲	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦۳	۳۳	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦٤	۳٤	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦۵	۳۵	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦٦	۳٦	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦۷	۳۷	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦۸	۳۸	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
٦۹	۳۹	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷۰	٤۰	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷۱	٤۱	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷۲	٤۲	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷۳	٤۳	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷٤	٤٤	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷۵	٤۵	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷٦	٤٦	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷۷	٤۷	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷۸	٤۸	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۷۹	٤۹	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸
۸۰	۵۰	۲	۰,۸۱۷۷۹۳	۳,۹۵۲٦۰۵	۸٤۳,۹٤۰۱	۰,۸۲۸۱۹۸

که مدل نهایی به صورت زیر است

$$\sqrt{ragweed} = ۱.۳ + ۳۹.۶۴R - ۰.۲۸S + ۰.۰۰۸TS + ۰.۰۰۵h(day - ۲۰)\ T - ۰.۰۰۹h(day - ۳۲)\ T + ۰.۰۰۳h(day - ۵۳)T - ۰.۹۴h(day - ۲۰)R + ۰.۴h(day - ۴۱)\ R + ۰.۴۸h(day - ۷۶)\ R - ۰.۵۶h(۷۶ - day)\ \ R. \quad (۷-۵)$$

نتیجه برازش هر دو مدل‌های رگرسیونی نیمه‌پارامتری غیرتطبیقی و تطبیقی در جدول ۵-۷ خلاصه شده است. همانند مثال قبل مدل رگرسیونی نیمه‌پارامتری تطبیقی برازش بهتری را به داده‌های مدل‌ساز داشته است.

جدول ۵-۷: مجموع توان‌های دوم خطا و ضریب تعیین مجموعه داده‌های مدل‌ساز و ضریب تعیین مجموعه داده‌های آزمون برای داده‌های ابروسیا

مدل	مجموعه داده‌های مدل‌ساز		مجموعه داده‌های آزمون
	مجموع توان دوم‌های خطا	ضریب تعیین	ضریب تعیین
نیمه‌پارامتری تطبیقی (۵-۷)	۸۴۳/۹۴۰۱	۰/۸۱۷۷۹۳	۰/۸۲۸۱۹
نیمه‌پارامتری غیر تطبیقی (۵-۶)	۱۱۰۷/۰۱۴	۰/۷۹۵۲۸	۰/۷۹۷۰۱۱

۵- ۶- مورد مطالعاتی ۴: داده‌های قیمت مسکن بوستن[۱]

مجموعه داده‌های مسکن بوستن (هاریسون و روبینفلد[۲] ۱۹۷۰) در رابطه با اطلاعات بازار خانه-سازی می‌باشد. این نوع اطلاعات اغلب در جهت مشخص کردن کیفیت اثرات فاکتورهای محیطی که قیمت دارائی را دگرگون می‌سازد، استفاده می‌شود. داده‌های مربوط به این مورد مطالعاتی در آدرس الکترونیکی زیر قابل دسترس است.

http://lib.stat.cmu.edu/datasets/boston_corrected.txt

این اطلاعات شامل ۵۰۶ مشاهده است که از منابع سرشماری ایالات متحده و کمیته‌های برنامه‌ریزی منطقه شهری بوستن گردآوری شده است. هر مشاهده از مجموعه داده‌ها، دارای ۱۴ صفت می‌باشد که عبارتند از:

۱) CRIM: نرخ بزه‌کاری در شهر

۲) ZN : نسبت زمین‌های مسکونی برای زمین‌های بالای ۲۵۰۰ مترمربع

۳) INDUS : نسبت شغل‌ها غیر از شغل‌های خرده‌فروشی در هر شهر

[۱] Boston Housing prices data
[۲] Harisson and Rubinfeld

۴) CHAS : متغیرساختگی رودخانه چارلز[۱] (۱= CHAS اگر مرزهای خانه مورد نظر رودخانه باشد و ۰= CHAS در غیراینصورت)

۵) NOX : غلظت اکسید نیتروژن (جز در[۲] ۱۰ میلیون)

۶) RM : میانگین تعداد اتاق‌ها در هر مسکن

۷) AGE : نسبت واحدهای ساختمانی مالک‌دار قبل ۱۹۴۰

۸) DIS : فاصله‌های وزنی به ۵ مرکز کاری بستون

۹) RAD : شاخص توانایی دسترسی به محور بزرگراه

۱۰) TAX : نرخ مالیات دارائی در هر ۱۰۰۰ دلار

۱۱) PTARTIO : نسبت معلم به شاگرد در هر شهر

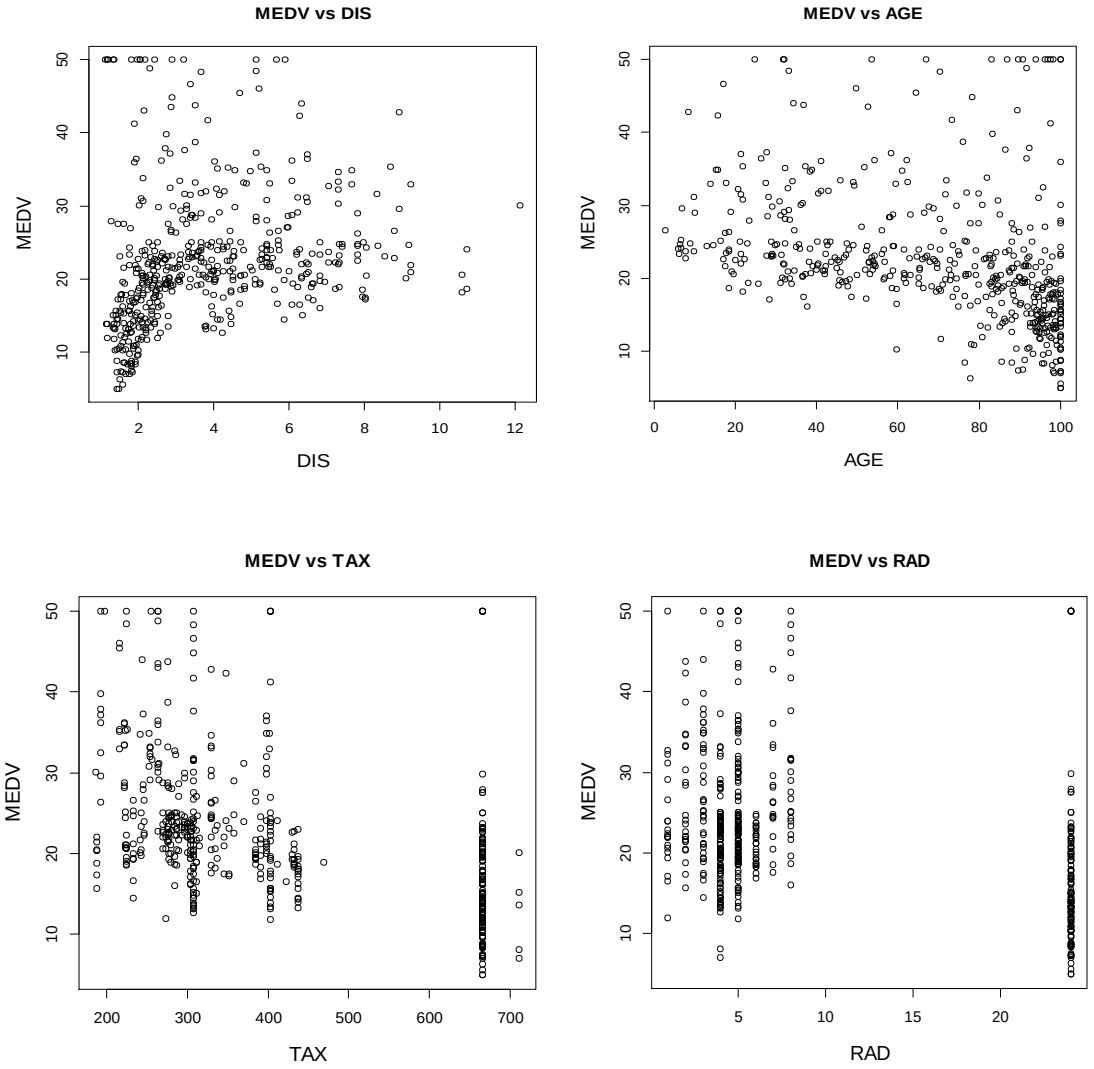
۱۲) B : $B = ۱۰۰۰(Bk - ۰.۶۳)^۲$ که BK در آن نسبت سیاه پوست‌ها در شهر است.

۱۳) LSTAT : وضعیت پایین جامعه[۳]

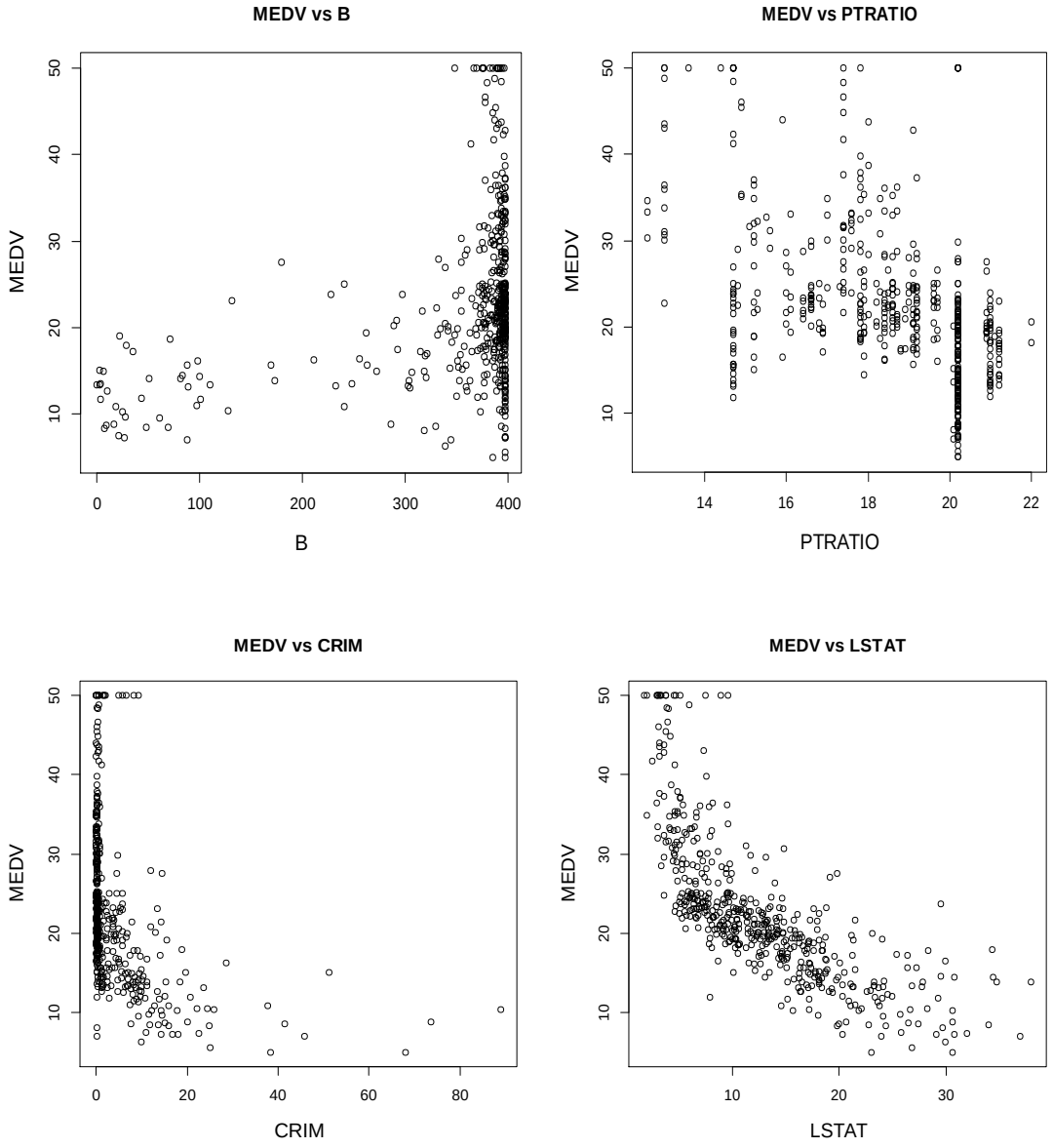
۱۴) MEDV : مقدار میانگین خانه‌های مالک‌دار

که درآن MEDV متغیر پاسخ و متغیرهای دیگر، متغیرهای پیش‌بین هستند. ارتباط بین متغیر پاسخ MEDV و متغیرهای پیش‌بین در شکل‌های ۵-۱۰ تا ۵-۱۳ برای تمام داده‌ها نشان داده شده است.

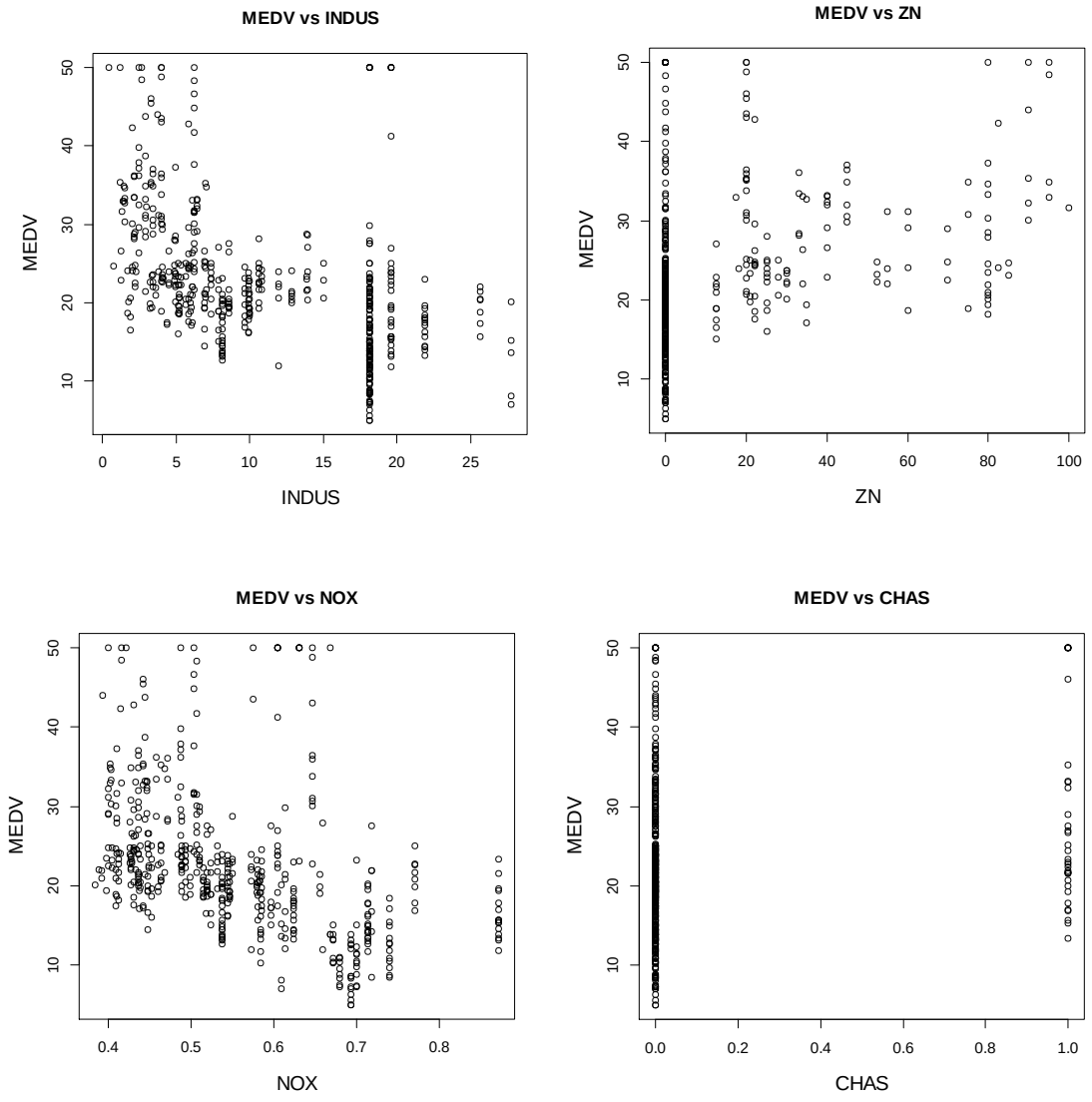
[۱] Charlz
[۲] Parts per
[۳] lower status of the population



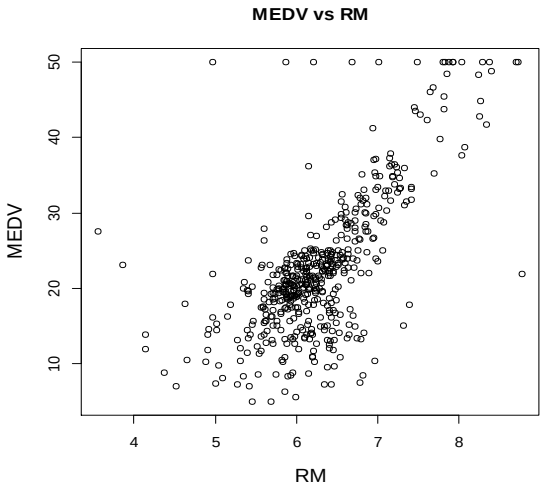
شکل ۵-۱۰: نمودار پراکنش متغیر پاسخ MEDV در مقابل متغیرهای AGE، DIS، RAD و TAX در داده‌های بوستن



شکل ۵-۱۱: نمودار پراکنش متغیر پاسخ MEDV در مقابل متغیرهای PTRATIO، B، LSTAT و CRIM در داده‌های بوستن



شکل ۵-۱۲: نمودار پراکنش متغیر پاسخ MEDV در مقابل متغیرهای ZN، INDUS، CHAS و NOX در داده‌های بوستن

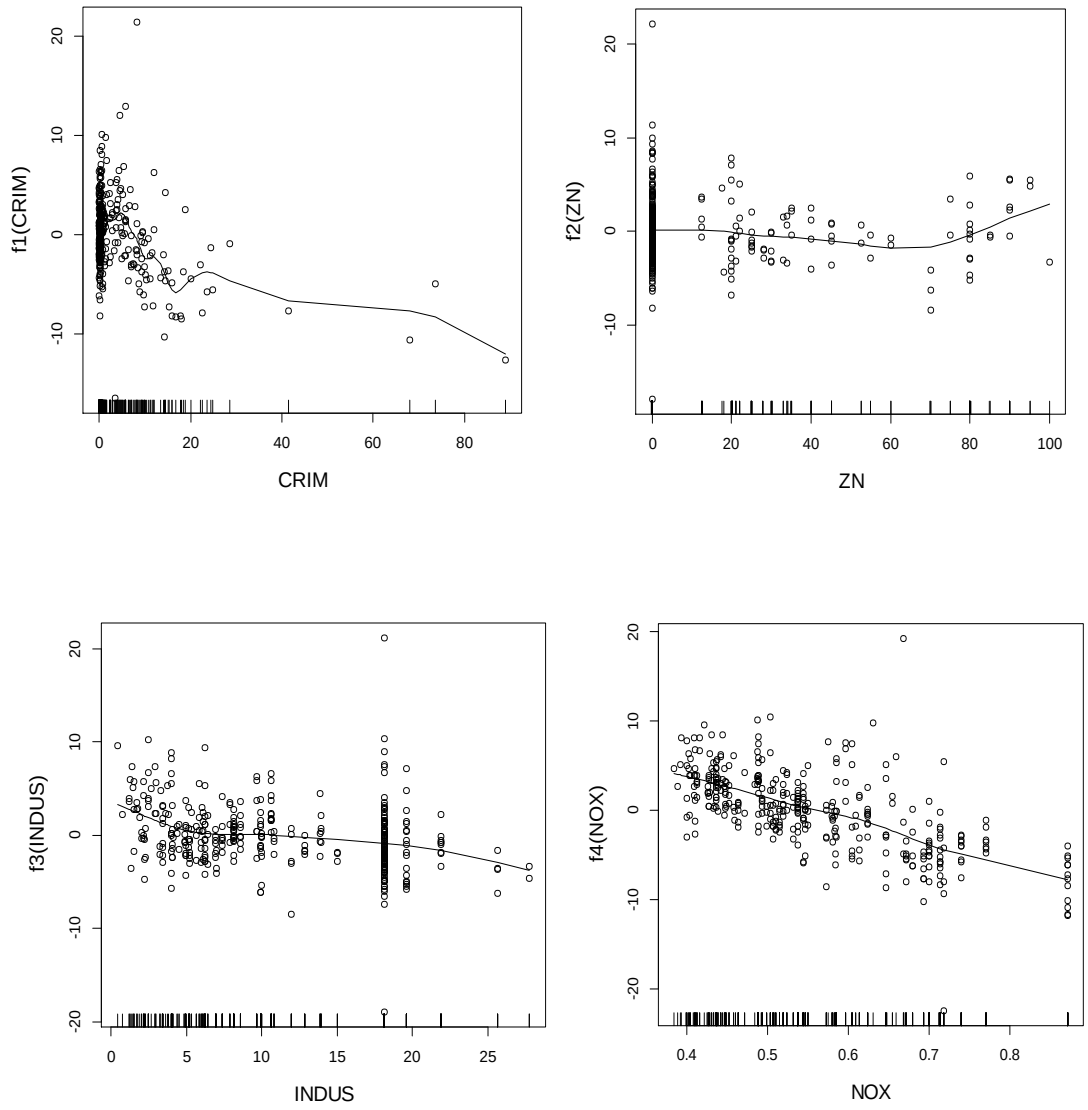


شکل ۵-۱۳: نمودار پراکنش متغیر پاسخ MEDV در مقابل متغیر RM در داده‌های بوستن

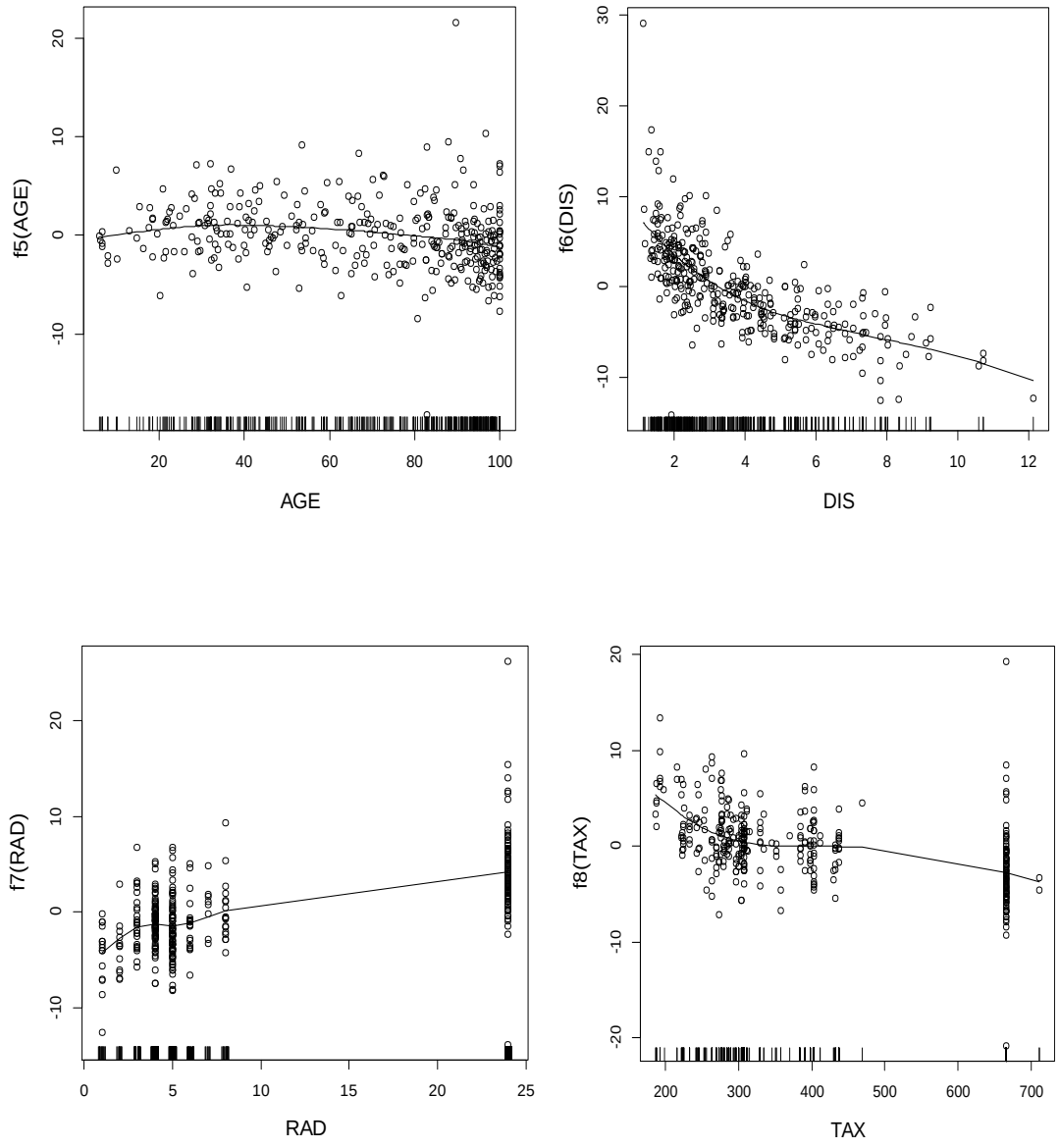
از شکل‌های ۵-۱۰ تا ۵-۱۳ می‌توان دریافت، کلیه متغیرها به جز متغیر دودویی CHAS و متغیر RM ارتباط غیرخطی با متغیر پاسخ MEDV دارند. در این مثال، همانند مثال‌های قبل، مشاهدات را به دو زیر مجموعه شامل داده‌های مدل‌ساز و آزمون تقسیم می‌کنیم. با توجه به نمودارهای پراکنش، یک مدل رگرسیونی نیمه‌پارامتری مفید برای این داده‌ها می‌تواند به صورت زیر باشد

$$E(MEDV) = \beta_0 + \beta_1 RM + \beta_2 CHAS + f_1(CRIM) + f_2(ZN) + f_3(INDUS) + f_4(NOX) + f_5(AGE) + f_6(DIS) + f_7(RAD) + f_8(TAX) + f_9(PTRATIO) + f_{10}(B) + f_{11}(LSTAT), \quad (۸-۵)$$

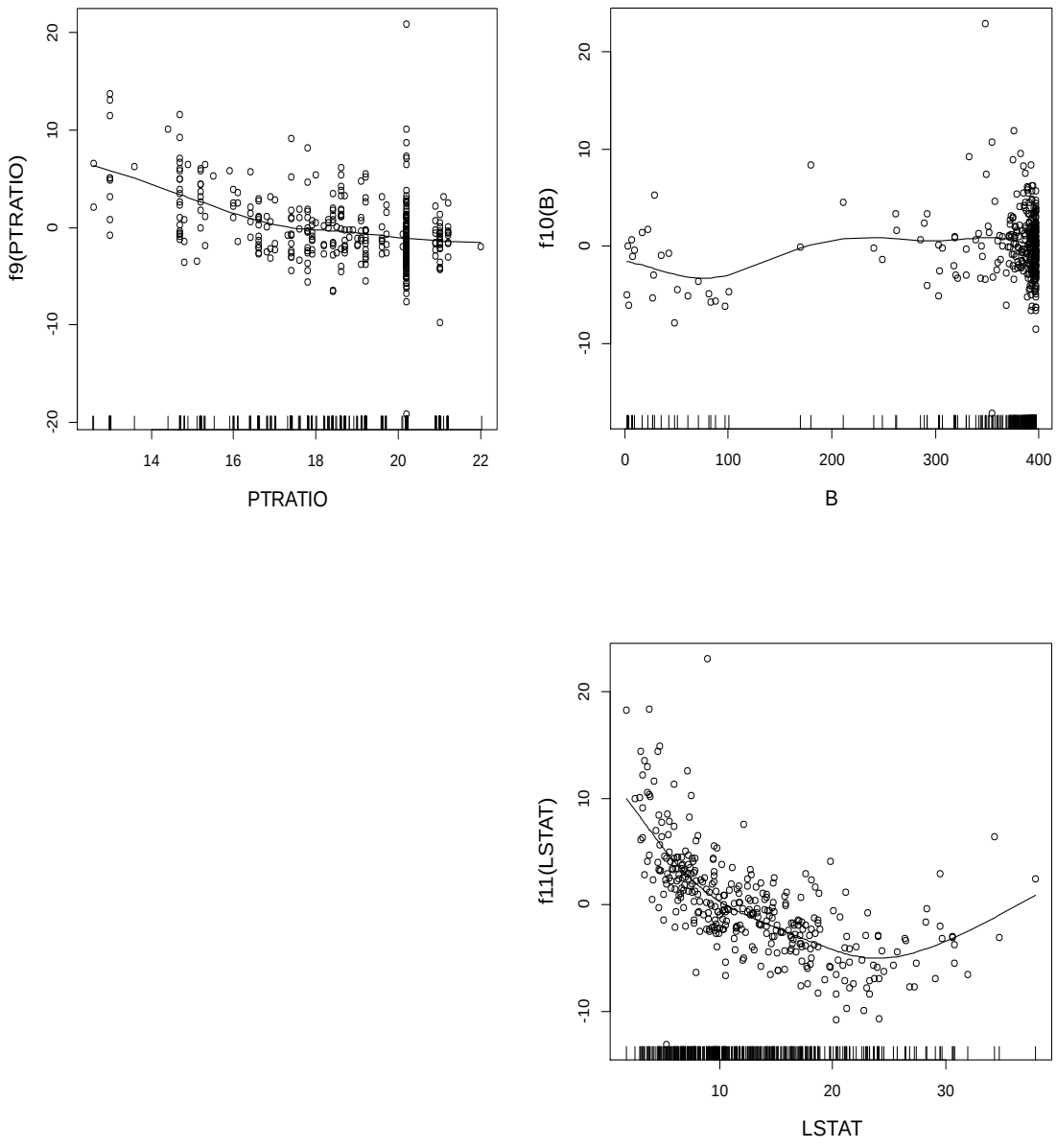
که در آن $f_1, \ldots, f_{11}$ توابع هموار هستند که هر یک به‌وسیله اسپلاین هموارساز برآورد می‌شوند. در اشکال ۵-۱۴ تا ۵-۱۶ برآورد توابع f برای داده‌های مدل‌ساز توسط اسپلاین هموارساز نشان داده شده است. از نمودارهای هموار رسم شده، می‌توان دریافت هموارسازی مناسبی برای این داده‌ها صورت گرفته است.



شکل ۵- ۱۴: نمودار های هموار برای متغیرهای CRIM، ZN، INDUS و NOX براساس اسپلاین هموارساز در داده‌های بوستن



شکل ۵-۱۵: نمودار های هموار برای متغیرهای AGE، DIS، RAD و TAX براساس اسپلاین هموارساز در داده‌های بوستن



شکل ۵-۱۶: نمودار های هموار برای متغیرهای PTRATIO، B و LSTAT براساس اسپلاین هموارساز در داده‌های بوستن

در ادامه جهت برازش مدل رگرسیونی نیمه‌پارامتری تطبیقی به داده‌های مدل‌ساز، سعی در بهینه-سازی دو پارامتر کلیدی nk و deg داریم. بنابراین با تغییر این دو پارامتر به طوری که در جدول ۵-۸ نشان داده شده است، بهترین مدل را که دارای بیشترین مقدار R-square و کمترین مقدار GCV و RSS است، از بین ۱۲۰ مدل تولید شده انتخاب می‌کنیم. این مدل، مدل شماره ۱۱۹ می‌باشد از طرفی چون ما در این موردهای مطالعاتی به دنبال انتخاب مدلی هستیم که علاوه بر کارایی مناسب

در داده‌های مدل‌ساز، بهترین برازش را برای داده‌های آزمون ارائه دهد، لذا مدل شماره ۸۱ را به عنوان مدل نهایی انتخاب می‌کنیم.

جدول۵-۸: مقادیر R-square، GCV و RSS برای ۱۲۰ مدل MARS نیمه‌پارامتری براساس مقادیر مختلف nk و deg (مورد مطالعاتی ۴)

			Training set			Test set
model	nk	deg	R-square	GCV	RSS	R-square
۱	۱۱	۱	۰,۷۶۵۷۵۷	۱۸,۵۲٤٦٦	٦۰۸٤,۹۳۱	۰,۷۸۵۳۷٤
۲	۱۲	۱	۰,۷٦۵۷۵۷	۱۸,۵۲٤٦٦	٦۰۸٤,۹۳۱	۰,۷۸۵۳۷٤
۳	۱۳	۱	۰,۷۷۸۰٦۸	۱۷,۷۰٦۵	۵۷٦۵,۱۲۹	۰,۷۸۱۳۷۹
٤	۱٤	۱	۰,۷۷۸۰٦۸	۱۷,۷۰٦۵	۵۷٦۵,۱۲۹	۰,۷۸۱۳۷۹
۵	۱۵	۱	۰,۷۸۹۸۱۳	۱۷,۲۲۲۹۷	۵٤٦۰,۰۲۲	۰,۷۹۷۳۱۲
٦	۱٦	۱	۰,۷۸۹۸۱۳	۱۷,۲۲۲۹۷	۵٤٦۰,۰۲۲	۰,۷۹۷۳۱۲
۷	۱۷	۱	۰,۷۹۸۳۵۵	۱٦,۳۷٦۰٤	۵۲۳۸,۱۲٤	۰,۸۱۱۹۸٤
۸	۱۸	۱	۰,۷۹۸۳۵۵	۱٦,۳۷٦۰٤	۵۲۳۸,۱۲٤	۰,۸۱۱۹۸٤
۹	۱۹	۱	۰,۸۰۸۷۲۷	۱۵,۸۱٤٤٤	٤۹٦۸,٦۹٤	۰,۸۲۱۱٤۸
۱۰	۲۰	۱	۰,۸۰۸۷۲۷	۱۵,۸۱٤٤٤	٤۹٦۸,٦۹٤	۰,۸۲۱۱٤۸
۱۱	۲۱	۱	۰,۸۱٤۹۰۱	۱۵,٤٤۲۵٦	٤۸۰۸,۳۰۷	۰,۸۲۰۸۳۱
۱۲	۲۲	۱	۰,۸۱٤۹۰۱	۱۵,٤٤۲۵٦	٤۸۰۸,۳۰۷	۰,۸۲۰۸۳۱
۱۳	۲۳	۱	۰,۸۱۹۱٤۵	۱۵,۲۲۵۷۵	٤٦۹۸,۰۵۸	۰,۸۱۷۸٤
۱٤	۲٤	۱	۰,۸۱۹۱٤۵	۱۵,۲۲۵۷۵	٤٦۹۸,۰۵۸	۰,۸۱۷۸٤
۱۵	۲۵	۱	۰,۸۲۲۸۷٤	۱۵,۰٤۸۱۱	٤٦۰۱,۱۹۳	۰,۸۱۰۸۲۵
۱٦	۲٦	۱	۰,۸۲۲۸۷٤	۱۵,۰٤۸۱۱	٤٦۰۱,۱۹۳	۰,۸۱۰۸۲۵
۱۷	۲۷	۱	۰,۸۲۲۸۷٤	۱۵,۰٤۸۱۱	٤٦۰۱,۱۹۳	۰,۸۱۰۸۲۵
۱۸	۲۸	۱	۰,۸۲۲۸۷٤	۱۵,۰٤۸۱۱	٤٦۰۱,۱۹۳	۰,۸۱۰۸۲۵
۱۹	۲۹	۱	۰,۸۲۸۹۸۸	۱٤,۹۳٤٤۲	٤٤٤۲,۳٦۷	۰,۸۲۲۹۳۳
۲۰	۳۰	۱	۰,۸۲۸۹۸۸	۱٤,۹۳٤٤۲	٤٤٤۲,۳٦۷	۰,۸۲۲۹۳۳
۲۱	۳۱	۱	۰,۸۳۱۷۱٦	۱٤,۸۳۲۹۸	٤۳۷۱,٤۹۸	۰,۸۱٦۷۵۱
۲۲	۳۲	۱	۰,۸۳۱۷۱٦	۱٤,۸۳۲۹۸	٤۳۷۱,٤۹۸	۰,۸۱٦۷۵۱
۲۳	۳۳	۱	۰,۸۳۱۷۱٦	۱٤,۸۳۲۹۸	٤۳۷۱,٤۹۸	۰,۸۱٦۷۵۱
۲٤	۳٤	۱	۰,۸۳۱۷۱٦	۱٤,۸۳۲۹۸	٤۳۷۱,٤۹۸	۰,۸۱٦۷۵۱
۲۵	۳۵	۱	۰,۸۳۳٤۳۲	۱٤,٦۸۱۷٦	٤۳۲٦,۹۳	۰,۸۱۵۸۲٤
۲٦	۳٦	۱	۰,۸۳۳٤۳۲	۱٤,٦۸۱۷٦	٤۳۲٦,۹۳	۰,۸۱۵۸۲٤
۲۷	۳۷	۱	۰,۸۳۵٤٤۲	۱٤,٦٤۰۲۳	٤۲۷٤,۷۱	۰,۸۰۸٤٦۷
۲۸	۳۸	۱	۰,۸۳۵٤٤۲	۱٤,٦٤۰۲۳	٤۲۷٤,۷۱	۰,۸۰۸٤٦۷
۲۹	۳۹	۱	۰,۸۳۷۷۱۹	۱٤,۵۷۳۳٤	٤۲۱۵,۵٦٤	۰,۸۱۳۳۷٦
۳۰	٤۰	۱	۰,۸۳۷۷۱۹	۱٤,۵۷۳۳٤	٤۲۱۵,۵٦٤	۰,۸۱۳۳۷٦

۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤۱	۳۱
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤۲	۳۲
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤۳	۳۳
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤٤	۳٤
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤٥	۳٥
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤٦	۳٦
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤۷	۳۷
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤۸	۳۸
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٤۹	۳۹
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥۰	٤۰
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥۱	٤۱
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥۲	٤۲
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥۳	٤۳
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥٤	٤٤
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥٥	٤٥
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥٦	٤٦
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥۷	٤۷
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥۸	٤۸
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٥۹	٤۹
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦۰	٥۰
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦۱	٥۱
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦۲	٥۲
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦۳	٥۳
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦٤	٥٤
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦٥	٥٥
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦٦	٥٦
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦۷	٥۷
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦۸	٥۸
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	٦۹	٥۹
۰,۸۱۳۳۷۶	٤۲۱٥,٥٦٤	۱٤,٥۷۳۳٤	۰,۸۳۷۷۱۹	۱	۷۰	٦۰
۰,۷۹۰۷٥	٥۸۱٤,۲۰۳	۱۷,٥٤٥۷۷	۰,۷۷٦۱۷۹	۲	۱۱	٦۱
۰,۷۹۰۷٥	٥۸۱٤,۲۰۳	۱۷,٥٤٥۷۷	۰,۷۷٦۱۷۹	۲	۱۲	٦۲
۰,۷۹۰۱٦	٥۲٥٤,۳۷۸	۱٦,۲۸۱۳۸	۰,۷۹۷۷۲۹	۲	۱۳	٦۳
۰,۷۹۰۱٦	٥۲٥٤,۳۷۸	۱٦,۲۸۱۳۸	۰,۷۹۷۷۲۹	۲	۱٤	٦٤
۰,۸۲۹۲٤٤	٤٥٥۱,۳۷۳	۱٤,۱۰۳۰۳	۰,۸۲٤۷۹۲	۲	۱٥	٦٥
۰,۸۲۹۲٤٤	٤٥٥۱,۳۷۳	۱٤,۱۰۳۰۳	۰,۸۲٤۷۹۲	۲	۱٦	٦٦
۰,۸۳۸۰٤۲	٤۱٥٦,۷۸۸	۱۳,۱۱۲۰۸	۰,۸۳۹۹۸۲	۲	۱۷	٦۷
۰,۸۳۸۰٤۲	٤۱٥٦,۷۸۸	۱۳,۱۱۲۰۸	۰,۸۳۹۹۸۲	۲	۱۸	٦۸

۶۹	۱۹	۲	۰,۸۵۱۶۴۷	۱۲,۳۷۶۹۳	۳۸۵۳,۷۷۱	۰,۸۴۹۰۲۵
۷۰	۲۰	۲	۰,۸۵۱۶۴۷	۱۲,۳۷۶۹۳	۳۸۵۳,۷۷۱	۰,۸۴۹۰۲۵
۷۱	۲۱	۲	۰,۸۷۰۱۳۲	۱۱,۱۳۴۵۶	۳۳۷۳,۵۹۲	۰,۸۴۶۶۱۲
۷۲	۲۲	۲	۰,۸۷۰۱۳۲	۱۱,۱۳۴۵۶	۳۳۷۳,۵۹۲	۰,۸۴۶۶۱۲
۷۳	۲۳	۲	۰,۸۸۰۸۲۵	۱۰,۲۱۷۷۶	۳۰۹۵,۸۱۴	۰,۸۴۹۳۵۶
۷۴	۲۴	۲	۰,۸۸۰۸۲۵	۱۰,۲۱۷۷۶	۳۰۹۵,۸۱۴	۰,۸۴۹۳۵۶
۷۵	۲۵	۲	۰,۸۹۰۳۰۳	۹,۵۷۹۸۰۸	۲۸۴۹,۵۹۴	۰,۸۵۸۵۲
۷۶	۲۶	۲	۰,۸۹۰۳۰۳	۹,۵۷۹۸۰۸	۲۸۴۹,۵۹۴	۰,۸۵۸۵۲
۷۷	۲۷	۲	۰,۸۹۵۲۶	۹,۱۴۶۹۱۱	۲۷۲۰,۸۲۵	۰,۸۶۸۴۰۸
۷۸	۲۸	۲	۰,۸۹۵۲۶	۹,۱۴۶۹۱۱	۲۷۲۰,۸۲۵	۰,۸۶۸۴۰۸
۷۹	۲۹	۲	۰,۹۰۳۱۰۵	۸,۷۰۱۵۲۷	۲۵۱۷,۰۵۲	۰,۸۶۸۰۰۸
۸۰	۳۰	۲	۰,۹۰۳۱۰۵	۸,۷۰۱۵۲۷	۲۵۱۷,۰۵۲	۰,۸۶۸۰۰۸
۸۱	۳۱	۲	۰,۹۰۸۹۲۹	۸,۲۵۵۶۷۳	۲۳۶۵,۷۴۶	۰,۸۸۸۲۷۹
۸۲	۳۲	۲	۰,۹۰۸۹۲۹	۸,۲۵۵۶۷۳	۲۳۶۵,۷۴۶	۰,۸۸۸۲۷۹
۸۳	۳۳	۲	۰,۹۱۳۵۷۴	۷,۹۰۸۹۱۹	۲۲۴۵,۰۸۳	۰,۸۸۷۷۹۵
۸۴	۳۴	۲	۰,۹۱۳۵۷۴	۷,۹۰۸۹۱۹	۲۲۴۵,۰۸۳	۰,۸۸۷۷۹۵
۸۵	۳۵	۲	۰,۹۲۰۰۹۸	۷,۵۲۴۰۴۲	۲۰۷۵,۶۲۱	۰,۸۸۳۳۳۱
۸۶	۳۶	۲	۰,۹۲۰۰۹۸	۷,۵۲۴۰۴۲	۲۰۷۵,۶۲۱	۰,۸۸۳۳۳۱
۸۷	۳۷	۲	۰,۹۲۴۰۷۷	۷,۲۱۸۴۹۸	۱۹۷۲,۲۶۱	۰,۸۸۵۳۶
۸۸	۳۸	۲	۰,۹۲۴۰۷۷	۷,۲۱۸۴۹۸	۱۹۷۲,۲۶۱	۰,۸۸۵۳۶
۸۹	۳۹	۲	۰,۹۲۸۴۰۳	۶,۸۷۳۲۷۹	۱۸۵۹,۸۶۸	۰,۸۸۰۶۱۲
۹۰	۴۰	۲	۰,۹۲۸۴۰۳	۶,۸۷۳۲۷۹	۱۸۵۹,۸۶۸	۰,۸۸۰۶۱۲
۹۱	۴۱	۲	۰,۹۳۲۲۴۸	۶,۶۳۲۱۴۳	۱۷۵۹,۹۹۵	۰,۸۸۳۴۰۵
۹۲	۴۲	۲	۰,۹۳۲۲۴۸	۶,۶۳۲۱۴۳	۱۷۵۹,۹۹۵	۰,۸۸۳۴۰۵
۹۳	۴۳	۲	۰,۹۳۵۹۱۹	۶,۳۹۷۳۷۹	۱۶۶۴,۶۲۴	۰,۸۸۳۸۳۶
۹۴	۴۴	۲	۰,۹۳۵۹۱۹	۶,۳۹۷۳۷۹	۱۶۶۴,۶۲۴	۰,۸۸۳۸۳۶
۹۵	۴۵	۲	۰,۹۳۷۷۱۷	۶,۲۷۹۸۴۳	۱۶۱۷,۹۲۹	۰,۸۸۲۸۸۹
۹۶	۴۶	۲	۰,۹۳۷۷۱۷	۶,۲۷۹۸۴۳	۱۶۱۷,۹۲۹	۰,۸۸۲۸۸۹
۹۷	۴۷	۲	۰,۹۳۹۷۶۳	۶,۱۳۴۲۹۱	۱۵۶۴,۷۶۸	۰,۸۸۳۲۶۶
۹۸	۴۸	۲	۰,۹۳۹۷۶۳	۶,۱۳۴۲۹۱	۱۵۶۴,۷۶۸	۰,۸۸۳۲۶۶
۹۹	۴۹	۲	۰,۹۴۳۱۴۸	۶,۰۲۷۸۶۷	۱۴۷۶,۸۳۲	۰,۸۵۳۱۸
۱۰۰	۵۰	۲	۰,۹۴۳۱۴۸	۶,۰۲۷۸۶۷	۱۴۷۶,۸۳۲	۰,۸۵۳۱۸
۱۰۱	۵۱	۲	۰,۹۴۴۷۴۵	۵,۹۱۸۷۰۳	۱۴۳۵,۳۵۲	۰,۸۵۲۲۰۸
۱۰۲	۵۲	۲	۰,۹۴۴۷۴۵	۵,۹۱۸۷۰۳	۱۴۳۵,۳۵۲	۰,۸۵۲۲۰۸
۱۰۳	۵۳	۲	۰,۹۴۶۸	۵,۸۱۷۱۵	۱۳۸۱,۹۸۴	۰,۸۳۷۳۰۵
۱۰۴	۵۴	۲	۰,۹۴۶۸	۵,۸۱۷۱۵	۱۳۸۱,۹۸۴	۰,۸۳۷۳۰۵
۱۰۵	۵۵	۲	۰,۹۴۷۱۲	۵,۷۲۲۷۶۶	۱۳۷۳,۶۶۲	۰,۸۳۴۱۱۳
۱۰۶	۵۶	۲	۰,۹۴۷۱۲	۵,۷۲۲۷۶۶	۱۳۷۳,۶۶۲	۰,۸۳۴۱۱۳

۱۰۷	۵۷	۲	۰,۹۴۸۴۴۷	۵,۵۲۲۱۵	۱۳۳۹,۱۸۴	۰,۸۲۳۶۴۴
۱۰۸	۵۸	۲	۰,۹۴۸۴۴۷	۵,۵۲۲۱۵	۱۳۳۹,۱۸۴	۰,۸۲۳۶۴۴
۱۰۹	۵۹	۲	۰,۹۵۳۶۰۶	۵,۲۳۴۰۹۲	۱۲۰۵,۱۷۵	۰,۸۱۴۵۸۸
۱۱۰	۶۰	۲	۰,۹۵۳۶۰۶	۵,۲۳۴۰۹۲	۱۲۰۵,۱۷۵	۰,۸۱۴۵۸۸
۱۱۱	۶۱	۲	۰,۹۵۴۹۰۳	۵,۱۴۱۶۷۸	۱۱۷۱,۴۸۹	۰,۸۱۷۴۴۳
۱۱۲	۶۲	۲	۰,۹۵۴۹۰۳	۵,۱۴۱۶۷۸	۱۱۷۱,۴۸۹	۰,۸۱۷۴۴۳
۱۱۳	۶۳	۲	۰,۹۵۶۵۰۱	۵,۰۱۲۲۳۷	۱۱۲۹,۹۶۶	۰,۸۲۲۷۷۶
۱۱۴	۶۴	۲	۰,۹۵۶۵۰۱	۵,۰۱۲۲۳۷	۱۱۲۹,۹۶۶	۰,۸۲۲۷۷۶
۱۱۵	۶۵	۲	۰,۹۵۸۳۲۱	۴,۹۵۹۳۲۶	۱۰۸۲,۷۰۲	۰,۸۲۲۱۸
۱۱۶	۶۶	۲	۰,۹۵۸۳۲۱	۴,۹۵۹۳۲۶	۱۰۸۲,۷۰۲	۰,۸۲۲۱۸
۱۱۷	۶۷	۲	۰,۹۵۹۳۶۴	۴,۷۸۳۴۴	۱۰۵۵,۶۰۳	۰,۸۱۶۰۷۲
۱۱۸	۶۸	۲	۰,۹۵۹۳۶۴	۴,۷۸۳۴۴	۱۰۵۵,۶۰۳	۰,۸۱۶۰۷۲
۱۱۹	۶۹	۲	۰,۹۶۰۴۸۵	۴,۷۰۱۸۱۸	۱۰۲۶,۴۸۴	۰,۸۲۸۲۴۱
۱۲۰	۷۰	۲	۰,۹۶۰۴۸۵	۴,۷۰۱۸۱۸	۱۰۲۶,۴۸۴	۰,۸۲۸۲۴۱

مدل نهایی با $nk = ۳۱$ و $deg = ۲$ به‌صورت زیر است

$$E(MEDV) = -۲۲.۸۷ + ۲.۲۶RM - ۱۰.۳h(NOX - ۰.۴۸۸) + ۰.۰۶h(TAX - ۳۱۱) - ۰.۲h(۳۱۱ - TAX) - ۰.۹۴h(PTRATIO - ۱۴.۷) + ۱.۶۸h(۱۴.۷ - PTRATIO) + ۱.۸۱h(LSTAT - ۶.۲۷) + ۰.۰۲RM\ h(۳۱۱ - TAX) - ۰.۰۲RM\ h(TAX - ۶۶۶) + ۰.۰۱RM\ h(۶۶۶ - TAX) + ۰.۴۶RM\ h(۶.۲۷ - LSTAT) - ۰.۲۶RM\ h(LSTAT - ۶.۲۷) - ۰.۶۱h(CRIM - ۵.۷)\ h(NOX - ۰.۴۸) - ۲۶۵.۵۴h(NOX - ۰.۶۴)h(۶.۲۷ - LSTAT) - ۱۰.۹۶h(۰.۶۴ - NOX)h(۶.۲۷ - LSTAT) + ۰.۴۶h(AGE - ۹۹.۳)h(PTRATIO - ۱۴.۷) + ۰.۰۱h(۹۹.۳ - AGE)h(PTRATIO - ۱۴.۷) - ۰.۰۰۵h(DIS - ۱.۶)h(TAX - ۳۱۱) + ۰.۰۹h(۱.۶ - DIS)h(TAX - ۳۱۱) - ۰.۰۹h(DIS - ۲.۱۱)h(LSTAT - ۶.۲۷) - ۰.۶h(۲.۱۱ - DIS)h(LSTAT - ۶.۲۷) - ۰.۰۰۱h(TAX - ۳۱۱)h(LSTAT - ۲۲.۷۴) + ۰.۰۰۳h(TAX - ۳۱۱)\ h(۲۲.۷۴ - LSTAT).$$ (۹-۵)

نتایج برازش هر دو مدل در جدول ۹-۵ آمده است. همانند دو مثال قبل مدل رگرسیونی نیمه-پارامتری تطبیقی با مجموع توان‌های دوم خطا کمتر و ضریب تعیین بیشتر برای هر دو مجموعه داده-های آموزشی و آزمون برازش بهتری را نسبت به مدل رگرسیونی نیمه‌پارامتری غیرتطبیقی دارد.

جدول ۹-۵ : مجموع توان‌های دوم خطا و ضریب تعیین مجموعه داده‌های مدل‌ساز و ضریب تعیین مجموعه داده‌های آزمون برای داده‌های قیمت مسکن بوستن

مدل	مجموعه داده‌های مدل‌ساز		مجموعه داده‌های آزمون
	مجموع توان-های دوم خطا	ضریب تعیین	ضریب تعیین
نیمه‌پارامتری تطبیقی (۵-۸)	۲۳۶۵/۷۴۶	۰/۹۰۸۹	۰/۸۸۸۲
نیمه‌پارامتری غیر تطبیقی (۵-۹)	۳۹۴۷/۷۴۵	۰/۸۷۵۱	۰/۸۵۱۶

نتیجه‌گیری و پیشنهادات

نتیجه‌گیری و پیشنهادات:

همانطور که در مقدمه گفته شد برای رفع مشکل مدل‌های ناپارامتری کلی، استون در سال ۱۹۸۵ مدل‌های جمعی را معرفی نمود. مدل‌های نیمه‌پارامتری حالت خاصی از مدل‌های جمعی هستند که از دو قسمت پارامتری و ناپارامتری تشکیل شده‌اند. در برآورد این مدل‌ها می‌توان از انواع هموارکننده-های نمودار پراکنش استفاده کرد. ما در این پایان‌نامه برآورد مدل‌های نیمه‌پارامتری را با استفاده از روش MARS انجام دادیم و با روش اسپلاین‌های هموارساز به عنوان یک هموارکننده نمودار پراکنش در برآورد مدل‌های نیمه‌پارامتری، مقایسه نمودیم.

با به‌کارگیری روش‌های اسپلاین‌های تطبیقی (MARS) و اسپلاین‌های غیرتطبیقی (اسپلاین‌های هموارساز) بر روی داده‌های شبیه‌سازی‌شده و داده‌های واقعی، دریافتیم که روش اسپلاین‌های تطبیقی می‌تواند جایگزین مناسبی برای اسپلاین‌های غیرتطبیقی در برآورد مدل‌های نیمه‌پارامتری باشد، به-خصوص در مواقعی که واریانس خطا رو به افزایش است.

با توجه به نتایج به‌دست آمده در این پایان‌نامه و موضوعات مطرح شد، می‌توان در آینده بر روی موارد زیر تحقیق کرد:

- در این پایان‌نامه توزیع خطا نرمال فرض شده است، می‌توان بررسی‌های انجام شده بر روی مدل‌های نیمه‌پارامتری را با توزیع دیگری برای خطا ادامه داد.
- هر چند در این پایان‌نامه صحبتی پیرامون چگونگی انتخاب برآوردگرها نشد، اما ممکن است در عمل به متغیرهایی برخورد کنیم که بین آن‌ها هم‌خطی وجود داشته باشد. آکنادیز و تاباکان (۲۰۰۹) در مدل‌های نیمه‌پارامتری ساده به بیان برآوردگر ریج[۱] در زمان وجود هم-خطی پرداخته‌اند. در جهت ادامه این راهکار می‌توان در مدل‌های نیمه‌پارامتری پیچیده به دنبال برآوردگری برای رفع مشکل هم‌خطی باشیم.

[۱] Ridge

کتاب نامه

[۱] روزبه، م. (۱۳۹۰). برآورد در مدل‌های خطی جزئی. رساله دکتری-دانشگاه فردوسی مشهد.

[۲] نیرومند، ح. ع. (۱۳۸۷). تحلیل رگرسیون خطی ابزاری برای تحقیق. انتشارات دانشگاه فردوسی مشهد.

[۳] Akdenz, F. and Tabakan, G. (۲۰۰۹). Restricted ridge estimators of the parameters in semiparametric regression model. *Communications in Statistics*, ۳۸, ۱۸۵۲-۱۸۶۹.

[٤] Amato, U., Antoniadis, A. and De Feis, I. (۲۰۰۲). Fourier series approximation of separable models. J. *Computation and Applied Mathematics*, ۱٤٦, ٤٥۹-٤۷۹.

[٥] Bhattacharya, P. K. and Zhao, P. L. (۱۹۹۷). Semiparametric inference in a partial linear model. *Ann. Statist*., ۲٥, ۲٤٤-۲٦۲.

[٦] Bowman, A. W. and Azzalini, A. (۱۹۹۷). *Applied Smoothing Techniques for Data Analysis*. Oxford, Clarendon.

[۷] Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (۱۹۹۷). Generalized partially single-index models. *J. Amer. Statist. Assoc*., ۹۲, ٤۷۷-٤۸۹.

[۸] Chen, H. and Shiau, J. H. (۱۹۹۱). A two-stage spline smoothing method for partially linear models. *J. Statist. Plann. Inference*, ۲۷, ۱۸۷-۲۰۱.

[۹] Chen, H. (۱۹۸۸). Convergence rates for parametric components in a partly linear model. *Ann. Statist*., ۱٦, ۱۳٦-۱٤٦.

[۱۰] Cravan, P. and Wahaba, G. (۱۹۷۹). Smoothing noisty data with spline functions. *Numer. Math*., ۳۱, ۳۷۷-۳۹۰.

[۱۱] Cuzick, J. (۱۹۹۲). Semiparametric additive regression. *J. Roy. Statist. Soc., Ser. B*, ٥٤, ۸۳۱-۸٤۳.

[۱۲] Dierckx, P. (۱۹۹٥). *Curve and Surface Fitting with Spline* (Monographs on Numerical Analysis). Oxford University Press.

[۱۳] Durmaz, M., Karsliglu, M. O. and Nohutcu, M. (۲۰۱۰). Regginal VTEC modeling with multivariate adaptive regression splines. *Advance in Space Research*, ٤٦, ۱۸۰-۱۸۹.

[۱٤] Efromovich, S. (۱۹۹۹). *Nonparametric Curve Estimation*. Now York, Springer-Verlag.

[۱٥] Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (۱۹۸٦). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc*., ۸۱, ۳۱۰-۳۲۰.

[۱٦] Fan, J., Haadle, W., and Mammen, E. (۱۹۹۸). Direct estimation of additive and linear components for high-dimensional data. *Ann. Statist*., ۲٦, ۹٤۳-۹۷۱.

[۱۷] Fan, J., and Gijbels, I. (۱۹۹۶). *Local Polynominal Modeling and Its Applications*. London, Chapman & Hall.

[۱۸] Fox, J. (۲۰۰۰). *Nonparametric Simple Regression: Smoothing Scatterplot*. Thousand Oaks, CA, Sage.

[۱۹] Friedman, J. H. (۱۹۹۰). Multivariate adaptive regression splines. *Ann. Statist*., ۱۹, ۱۱-٤۱.

[۲۰] Gao, J. T. (۱۹۹۲). *Large Sample Theory in Semiparametric Regression Models*. Ph. D. Thesis, Graduate School, University of Science and Technology of China, Hefei, P. R. China.

[۲۱] Gao, J. T. and Zhao, L. C. (۱۹۹۳). Adaptive estimation in partly linear models. *Sciences in China*, Ser. A, ۱٤, ۱٤-۲۷.

[۲۲] Gao, J. T., Hang, S. Y. and Liang, H. (۱۹۹٥). Convergence rates of a class of estimates in partly linear models. *Acta Mathematica Sinica*, ۳۸, ٦٥۸-٦٦۹.

[۲۳] Green, P. J. and Silverman, B. W. (۱۹۹٤). *Nonparametric Regression and Generalized Linear Models*. London, Chapman & Hall.

[۲٤] Guo, W. (۲۰۰۲). Functinal mixed effects models. *Biometrics*, ٥۸, ۱۲۱-۸.

[۲٥] Hansen, M. H., Huang, J. Z., Kooperberg, C., Ston, C. J. and Truong, Y. K. (۲۰۰۳). *Statistical Modeling with Spline Functions. Methology and Theory*, New York, Springer-Verlag.

[۲٦] Hardle, W. (۱۹۹۰). *Applied Non-parametric Regression* (Economic Society Monographs, vol, ۱۹). Cambridge University Press.

[۲۷] Hardle, W. (۱۹۹۱). *Smoothing Techniques, with Implementations in S*. New York, Springer-Verlag.

[۲۸] Hardle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (۱۹۹۸). *Wavelets, Approximation, and Statistical Applications* (Lecture Notes in Statistics, Vol, ۱۲۹). New York, Springer-verlag.

[۲۹] Hart, J. D. (۱۹۹۷). *Nonparametric Smoothing and Lack-of-Fit Tests*. New York, Springer-Verlag.

[۳۰] Hastie, T., Tibshirani, R. and Friedman, J. (۲۰۰۹). *The Elements of Statistical Learning Data Mining, Inference, and Prediction, ۲nd Ed.*, Springer, New York.

[۳۱] Hatie, T. and Tibshirani, R. J., (۱۹۹۰). *Generalized Additive Models*. Chapman & Hall, London, New York.

[۳۲] Hekman, N. E. (۱۹۸٦). *Spline smoothing in partly linear models*. J. Roy. Statist. Soc., Ser. B, ٤٨, ۲٤٤-۲٤٨.

[۳۳] Hong, S. Y. (۱۹۹۱). Estimation theory of a class of semiparametric regression models. *Sciences in China*, Ser. A, ۱۲,۱۲٥۸-۱۲۷۲.

[۳٤] Keele, L. (۲۰۰۸). *Semiparametric Regression for the Social Siences*. Wiley & Sons, Ltd.

[۳٥] Liang, H. (۱۹۹۲). *Asymptotic Efficiency in Semiparametric Models and Related Topics*. Ph. D. Thesis, Institute of Systems Science, Chinese Academy of Sciences, Beijing, P. R. China.

[۳٦] Linton, O. B., and Nielsen, J. P. (۱۹۹٥). A kernel method of estimating regression structured nonparametric regression based on marginal integration. *Biometrika*, ۸۲, ۹۳-۱۰۰.

[۳۷] Loader, C. (۱۹۹۹). *Local Regression and Likelihood*. New York, Springer-Verlag.

[۳۸] Louis, A. K., Maass, D. and Rieder, A. (۱۹۹۷). *Walvelets: Theory and Applications*. Now York, Wiley.

[۳۹] Muller, H. G. (۱۹۸۸). *Nonparametric Regression Analysis of Longitudinal Data* (Lecture Notes in Statistics, Vol, ٤٦). New York, Springer-Verlag.

[٤۰] Muller, P., and Vidakovic, B. (۱۹۹۹). *Bayesian Inference in Wavelet-based Models*. New York, Springer-verlag.

[٤۱] Nadaraya, E. A. (۱۹۸۹). *Nonparametric Estimation of Probability Densities and Regression Curves* [translate by S.Kotz]. Boston, Kluwer.

[٤۲] Nason, G. P., and Silverman, B. W. (۲۰۰۰). Wavelets for regression and other statistical problems. In M. G. Schimek (Ed.), Smoothing and Regression . Approaches, Computation and Application, pp. ۱٥۹-۹۱. New York, Wiley.

[٤۳] Ogden, R. T. (۱۹۹٦). *Essential Wavelets for Statistical Applications and Data Analysis*. Boston, Brikhauser.

[٤٤] Pagan, A. and Ullah, A. (۱۹۹۹). *Nonparametric Econometrics* (Themes in Modern Econometrics). Cambridge University Press.

[٤٥] Rice, J. (۱۹۸٦). Convergence rates for partially splined models. *Stat. Prob. Lett.*, ٤, ۲۰۳-۲۰۸.

[٤٦] Robinson, P. (۱۹۸۸). Root-n-consistent semiparametric regression. *Econometrica*, ٥٦, ۹۳۱-۹٥٤.

[٤٧] Ruppert, D., Wand, M. P. and Carroll, R. J. (٢٠٠٥). *Semiparametric Regression*. Cambridge: Cambridge University Press.

[٤٨] Schmalensee, R. and Stoker, T. M. (١٩٩٩). Household gasoline demand in the United States. *Econometrica*, ٦٧, ٦٤٥-٦٦٢.

[٤٩] Schick, A. (١٩٩٦a). Weighted least squares estimates in partly linear regression models. *Stat. Prob. Lett.*, ٢٧, ٢٨١-٢٨٧.

[٥٠] Schick, A. (١٩٩٦b). Root-n consistent estimation in partly linear regression models. *Stat. Prob. Lett.*, ٢٨, ٣٥٣-٣٥٨.

[٥٢] Salford systeme. (٢٠٠١). MARS User Guide.

[٥٣] Severini, T. A. and Wang, W.H. (١٩٩٢). Generalized profile likelihood and conditional parametric models. *Ann. Statist.*, ٢٠, ١٧٦٨-١٨٠٢.

[٥٤] Simonoff, J. S. (١٩٩٦). *Smoothing Methods in Statistics*. New York, Springer-Verlag.

[٥٥] Speckman, P. (١٩٨٨). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc., Ser. B*,٥٠, ٤١٣-٤٣٦.

[٥٦] Stone, C. J. (١٩٨٥). Additive regression and other nonparametric models. *Annals of Statistics*, ١٣, ٦٨٩ - ٧٠٥.

[٥٧] Storlie, C. B., Swiler, L. P., Helton, J. C. and Sallaberry, C. J. (٢٠٠٩). Implementation and evalution of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliability Engineering and System Safety*, ٩٤, ١٧٣٥- ١٧٦٣.

[٥٨] Tarter, M. E., and Lock, M. D. (١٩٩٣). *Model-free Curve Estimation*. New York, Chapman.

[٥٩] Thompson, J. R. and Tapia, R. A. (١٩٩٠*). Nonparametric Function Estimation, Modeling, and Simulation*. Philadelphia, SIAM.

[٦٠] Vidakovic, B. (١٩٩٩). *Statistical Modeling by Wavelets*. New York, Wiley.

[٦١] Wahba, G. (١٩٩٠). *Spline Models for Observational Data*. Philadelphia, SIAM.

[٦٢] Walter, G. G. and Shen, X. (٢٠٠١). *Wavelets and Other Orthogonal Systems, ٢nd ed*. Boca Raton, FL, Chapman & Hall/CRC press.

[٦٣] Wand, M. P. and Jones, M. C. (١٩٩٥). *Kernel Smoothing. London*, Chapman & Hall.

Abstract:

In this thesis, the aim is to estimate the nonparametric part of semiparametric models by multivariate adaptive regression splines (MARS) and smoothing splines, where the former is used as an adaptive spline while the latter is non adaptive. By the help of explaining scatterplot smoothers and concept of splines, the two aforementioned methods, MARS and smoothing spline, are described in details. And the use of these two methods is described in the estimation process of semiparametric models. Comparison is then done by simulation and applying to several real world examples. After utilization each of these two methods, it is clearly shown that adaptive splines performs better in the sense of having larger R-square and smaller residual sum of squares.

Keywords: Spline; Scatterplot Smoother; MARS; Smoothing spline; Adaptive; Semiparametric.

Shahrood University of Technology

Faculty Mathematics

Adaptive and Non-Adaptive Splines in Semiparametric Regression Models

Toktam Valizadeh

Supervisors:

Dr. Mohammad Arashi

Dr. Davood Shahsavani

Date: December ۲۰۱۲