
بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

رشته آمار، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

اسپلاین‌های جریمه‌ای و کاربردهای نوین آن

نگارنده: اکرم قائمی‌زاده

استاد راهنما

دکتر نگار اقبال

استاد مشاور

دکتر حسین باغیشنی

تیر ۱۳۹۷

شماره:
تاریخ:

باسمه‌تعالی

دانشگاه صنعتی شاهرود

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم **اکرم قائمی‌زاده** با شماره دانشجویی **۹۴۱۳۹۰۴** رشته آمار گرایش آمار ریاضی تحت عنوان **اسپلاین‌های جریمه‌ای و کاربردهای نوین آن**

که در تاریخ **۹۷/۰۴/۲۵** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

قبول (با درجه: خیلی خوب) ☑ مردود ☐

نوع تحقیق: نظری ■ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱- استادراهنمای اول	دکتر نگار اقبال	استادیار	
۳- استاد مشاور	دکتر حسین باغیشنی	استادیار	
۴- نماینده تحصیلات تکمیلی	دکتر عبدالله آل‌هوز	استادیار	A.A
۵- استاد ممتحن اول	دکتر داود شاهسونی	دانشیار	
۶- استاد ممتحن دوم	دکتر محمدرضا ربیعی	استادیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده: ۹۷/۰۴/۲۵

وزارت علوم، تحقیقات و فناوری

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می‌تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم به همسرم

که در سایه همیاری و همدلی او به این منظور نائل شدم.

تقدیم به پسرم

امید بخش جانم که آسایش او آرامش من است.

سپاس‌گزاری...

سپاس بی کران پروردگار یکتا را که هستی مان بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه چینی از علم و معرفت را روزیمان ساخت. اما از آنجایی که تجلیل از معلم، سپاس از انسانی است که هدف و غایت آفرینش را تامین می کند و سلامت امانت هایی را که به دستش سپرده اند، تضمین؛ بر حسب وظیفه و از باب ’’ من لم یشکر المنعم من المخلوقین لم یشکر اللّه عزّ و جلّ’’ :

از اساتید با کمالات و شایسته و دلسوز؛ سرکار خانم دکتر نگار اقبال و جناب آقای دکتر حسین باغیشنی که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این پایان‌نامه را بر عهده گرفتند نهایت تشکر و امتنان را دارم و برای آن بزرگواران آرزوی سلامتی و کامیابی از درگاه خداوند تبارک و تعالی می‌نمایم.

هم‌چنین برخود لازم می‌دانم از اساتید گرانقدر، جناب آقای دکتر داود شاهسونی و جناب آقای دکتر محمدرضا ربیعی که زحمت داوری این پایان‌نامه را متقبل شدند، تشکر و قدردانی نمایم.

و وظیفه می‌دانم از همسر عزیزم که همواره در طول تحصیل متحمل زحماتم بود و تکیه گاه من در مواجهه با مشکلات و وجودش مایه دلگرمی من می باشد صمیمانه سپاس‌گزاری کنم.

در پایان از دوستان عزیز و بسیار مهربانم، سرکار خانم مریم علی‌بیگی و خانم محبوبه محبی به پاس محبت های بی دریغشان که در تمامی لحظات رفیق راهم بودند کمال تشکر و قدردانی را دارم.

اکرم قائمی‌زاده
تیر ۱۳۹۷

تعهد نامه

اینجانب **اکرم قائمی‌زاده** دانشجوی کارشناسی ارشد رشته **آمار** دانشکده **علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **اسپلاین‌های جریمه‌ای و کاربردهای نوین آن** ، تحت راهنمایی **نگار اقبال** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود ’’ یا “ Shahrood University of Technology ” به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

اکرم قائمی‌زاده
تیر ۱۳۹۷

مالکیت نتایج و حق نشر

چکیده

تحلیل رگرسیونی در حقیقت پرکاربردترین روش در بین تکنیک‌های آماری و یک ابزار آماری برای تشخیص رابطه یک یا چند متغیر تبیینی با متغیر پاسخ است. زمانی‌که برای تحلیل یک مجموعه داده مدل رگرسیونی خطی کارآمد نباشد، می‌توان از روش‌های ناپارامتری رگرسیون استفاده نمود. اسپلاین‌ها به عنوان یکی از ابزارهای درونیابی، در راستاي رفع این مشکل، ابزاري ارزشمند و پرتوان به شمار مي‌روند و از جمله روش‌های ناپارامتری برای مدل‌بندی رگرسیون ناپارامتری هستند. اسپلاین‌ها منحنی‌هایی را می‌سازند که شامل چندجمله‌ای‌هایی با درجه یکسان روی زیربازه‌هایی از یک بازه مشخص هستند و با شرایط پیوستگی تعریف‌شده‌ای به هم می‌پیوندند و از گره‌های مشترک بین دو زیربازه عبور می‌کنند. در طی سال‌هایی که از معرفی اسپلاین‌ها می‌گذرد، مبانی نظری آن‌ها توسعه یافته‌اند و نسخه‌های تعمیم‌یافته متفاوتی از آن‌ها معرفی شده‌اند. اسپلاین‌های جریمه‌ای (P ـ اسپلاین‌ها) از جمله ابزار پرکاربرد برای مدل‌سازی ناپارامتری و مسایل هموارسازی محسوب می‌شوند. ظهور P–اسپلاین‌ها با تحولات زیادی در مسایل هموارسازی همراه و به یک زمینه پویا و فعال تحقیقاتی تبدیل شده است. با توجه به کاربرد وسیع و اهمیت شناخت اسپلاین‌ها، در این پایان‌نامه به معرفی اسپلاین‌ها و روند رشد و توسعه آن‌ها می‌پردازیم و برخی از ویژگی‌های آن‌ها را مطرح می‌کنیم.

کلمات کلیدی: اسپلاین جریمه شده، هموارسازی، مدل‌های خطی تعمیم‌یافته، مدل‌های جمعی تعمیم‌یافته، معیار اطلاع آکائیک.

لیست مقالات مستخرج از پایان‌نامه

۱. قائمی‌زاده ا. و اقبال ن، (۱۳۹۶) ، ’’ سیر تحول اسپلاین‌ها’’، دومین کنفرانس ملی محاسبات نرم، گیلان ـ رودسر، ایران، ۸۴۷-۸۳۹.

پیش‌گفتار

یکی از قدیمی‌ترین و پرکاربردترین روش‌های آماری، رگرسیون خطی می‌باشد. در مدل‌های کلاسیک آماری میانگین متغیر پاسخ به صورت تابع خطی از متغیرهای مستقل بیان می‌شود. مجموعه داده‌های زیادی وجود دارند که مدل‌های خطی، برازش خوبی برای آن‌ها ندارند. اسپلاین‌ها جایگزین مناسبی برای برازش به داده‌ها و توابع چندجمله‌ای تکه‌ای هستند که محدود به اتصال به یکدیگر در نقاط گره می‌باشند. با اتصال اولین تکه به تکه بعدی یک گره ایجاد می‌شود. بدین صورت که هر قسم از شکل تابع را در محدوده‌های موردنظر تکه تکه کرده و در هر تکه به هموارسازی تابع مشخص شده می‌پردازد. به طور معمول توابع چندجمله‌ای درجه سوم و اسپلاین‌های جریمه شده زیر مجموعه‌ی اسپلاین هموار شده است که مبتنی بر مدل رگرسیونی و برآورد ناپارامتریک می‌باشند. مدل رگرسیون ناپارامتریک اختصاص دارد به مدلی که در آن نتوان رابطه‌ی بین متغیر وابسته و متغیرهای مستقل را به فرم تابعی مشخص و معلوم بیان نمود. اسپلاین جریمه شده مدلی غیرخطی است که مقدار اثرات خطی را با در نظر گرفتن اثرات غیر خطی متغیرها بررسی می‌کند. زیرا ممکن است در صورت در نظر نگرفتن اثرات غیر خطی منجر به از دست دادن اطلاعات شود که بسته به اهداف مطالعه ممکن است تاثیرگذار باشد در نتیجه با حضور اثرات غیر خطی، به بررسی اثرات خطی می‌پردازد. وسعت کاربرد اسپلاین‌های جریمه‌ای بسیار گسترده است به عنوان مثال می‌توان به ساختار رگرسیون جمعی خطی و جمعی تعمیم‌یافته، مدل با اثرات متغیر(وابسته به زمان، مکان یا هر دو)، مدل‌هایی با متغیرهای تبیینی تابعی، برآورد توابع دو یا چند متغیره و مدل‌های رگرسیونی چندکی و گشتاوری اشاره کرد.

در این پایان‌نامه با توجه به ویژگی‌های نظری و کاربردی خوب اسپلاین‌ها، تمرکز بر معرفی و استفاده از آن‌ها در مدل‌های مختلف آماری است. در فصل اول تعاریفی از انواع مدل‌های خطی از قبیل مدل‌های خطی تعمیم‌یافته، مدل‌های جمعی و مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل (GAMLSS) آورده شده است. در فصل دوم به تفضیل اسپلاین‌ها و انواع آن می‌پردازیم. در فصل سوم انتخاب پارامتر هموارسازی و بهترین مدل برازشی را براساس معیارهای CV, BIC, AIC و Cp معرفی می‌کنیم و در نهایت فصل چهارم، مجموعه داده‌هایی را با استفاده از این مدل‌ها مورد تحلیل و بررسی قرار می‌دهیم.

فهرست مطالب

فهرست تصاویر ق

فهرست جداول ش

۱ مفاهیم اولیه و مورد نیاز ۱

۱.۱ مقدمه ۱

۲.۱ مدل‌های خطی ۱

۳.۱ مدل‌های خطی تعمیم‌یافته ۳

۱.۳.۱ ساختار مدل‌های خطی تعمیم‌یافته ۴

۴.۱ مدل جمعی تعمیم‌یافته ۵

۱.۴.۱ مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل ۶

۲ سیر تحول اسپلاین‌ها و انواع آن‌ها ۹

۱.۲ توابع پایه ۹

۲.۲ تابع توانی بریده‌شده ۱۰

۳.۲ اسپلاین‌ها ۱۱

۴.۲ اسپلاین مکعبی (CS) ۱۲

۱.۴.۲ درونیابی اسپلاین مکعبی ۱۳

۲.۴.۲ خواص اسپلاین‌های مکعبی ۱۳

۳.۴.۲ اسپلاین مکعبی طبیعی (NCS) ۱۶

۵.۲ B-اسپلاین‌ها ۱۷

۶.۲ هموارسازی اسپلاینی ۲۱

۷.۲ اسپلاین‌های جریمه‌ای (p-اسپلاین) ۲۳

۱.۷.۲ صورت ماتریسی و برآورد ضرایب اسپلاین جریمه‌ای با استفاده از کمترین توان‌های دوم ۲۴

۲.۷.۲ خواص اسپلاین جریمه‌ای ۲۶

۸.۲ اسپلاین حاصل‌ضرب تانسور ۲۷
۹.۲ اسپلاین صفحه نازک ۳۰
۱.۹.۲ مزایای اسپلاین صفحه نازک ۳۲

۳ برخی از مباحث نظری اسپلاین‌ها ۳۵
۱.۳ نحوه‌ی انتخاب پارامتر هموارسازی ۳۵
۱.۱.۳ معیار اطلاع آکائیک ۳۶
۲.۱.۳ اعتبار سنجی متقابل ۳۷
۳.۱.۳ اعتبارسنجي متقابل تعمیم‌یافته ۳۸
۲.۳ روش‌های انتخاب تعداد و موقعیت گره ۳۹
۳.۳ انتخاب مدل ۴۱
۱.۳.۳ معیار اطلاع بیزی (شوارتز) ۴۲
۲.۳.۳ آماره C_p مالو ۴۲

۴ برخی از کاربردهای اسپلاین‌ها ۴۷
۱.۴ مدل‌های جمعی ۴۸
۱.۱.۴ کاربرد اول:داده‌های اجاره مونیخ ۴۸
۲.۱.۴ کاربرد دوم: داده‌های گونه‌های ماهی ۵۵
۳.۱.۴ کاربرد سوم: داده‌های مدت بستری در بیمارستان ۵۹
۲.۴ اسپلاین دو بعدی ۶۱
۱.۲.۴ داده‌های فیلم ۶۱
۳.۴ مدل میدان‌های تصادفی گاوسی ۶۶
۱.۳.۴ داده‌های جرم و جنایت ۶۷
۴.۴ نتیجه‌گیری ۶۸

آ ۶۹
آ.۱ ضرب کرونکر ۶۹
آ.۲ مشتق تابع B-اسپلاین ۷۰
آ.۳ جدول توزیع‌های پیوسته و گسسته موجود در بسته‌ی gamlss.dist ۷۱
آ.۴ دستورات نرم‌افزار R ۷۲

مراجع ۸۹

فهرست تصاویر

۱.۲ توابع توانی بریده‌شده مرتبه یک ۱۱
۲.۲ نمایش اسپلاین‌های مکعبی و مکعبی طبیعی ۱۷
۳.۲ نمایش توابع پایه B-اسپلاینی مرتبه یک ۱۹
۴.۲ نمایش توابع پایه B-اسپلاین ۲۰
۵.۲ هموارسازی ۲۲
۶.۲ فلوچارت روش‌های هموارساز ۲۲
۷.۲ هموارسازی تابع رگرسیونی برای مقادیر مختلف λ ۲۶
۸.۲ حاصل‌ضرب دو تابع پایه‌ی حاشیه‌ای برای توابع هموار fx وfz ۲۸
۹.۲ نمایش اسپلاین صفحه نازک ۳۱
۱۰.۲ نمایش اسپلاین صفحه نازک هموارتر ۳۲
۱۱.۲ نمایش اسپلاین صفحه نازک ۳۳

۱.۳ پارامتر هموارسازی در مقابل معیار انتخاب برای تابع رگرسیونی P-اسپلاینی از مرتبه $m = ۴$ ۳۹
۲.۳ برازش P-اسپلاین مرتبه $k = ۴$ با تعداد گره کم ۴۰
۳.۳ برازش P-اسپلاین مرتبه $k = ۴$ با تعداد گره زیاد ۴۰
۴.۳ برازش اسپلاین مکعبی طبیعی با چهار درجه آزادی و قرار گرفتن موقعیت گره‌ها در چندک داده‌ها ۴۱
۵.۳ برازش چندجمله‌ای از مرتبه یک تا مرتبه هفت برای مجموعه داده‌های شبیه‌سازی‌شده ۴۳
۶.۳ مقایسه اثربخشی معیار اطلاع اکائیک و بیزی در انتخاب بهترین مدل ۴۴
۷.۳ انتخاب بهترین مدل با استفاده از معیار اعتبارسنجی متقابل ۴۴

۱.۴ نمودار متغیر پاسخ (میزان اجاره R) در مقابل هریک از متغیرهای توضیحی . . ۴۹
۲.۴ نمودار مانده‌های برازش مدل خطی ۵۰
۳.۴ نمودار مانده‌های برازش مدل گاما ۵۰
۴.۴ نمودار مقادیر برازشی برای مدل (GAM) ۵۱
۵.۴ نمودار مارپیچ برای مدل (GAM) ۵۲

۶.۴ نمودار مارپیچ برای مدل گامای دو پارامتری ۵۴
۷.۴ نمودار مارپیچ برای مدل سه پارامتری ۵۵
۸.۴ نمودار گونه‌های ماهی ۵۶
۹.۴ نمودار برازش مدل سیچل به داده‌های گونه‌های ماهی ۵۸
۱۰.۴ نمودار مارپیچ مدل برازشی سیچل به داده‌های گونه‌های ماهی ۵۸
۱۱.۴ نمودار نرخ نامناسب متغیر پاسخ در مقابل متغیرهای توضیحی در داده‌های مدت بستری ۵۹
۱۲.۴ توابع برازشی برای μ در مدل M4 . . . . . ۶۱
۱۳.۴ ترسیم نمودار سه بعدی داده‌های فیلم ۶۲
۱۴.۴ نمودار مارگون برای مدل M6 با برازش رویه برای پارامتر μ ۶۳
۱۵.۴ نمودار رویه‌ی برازشی برای مدل M6 . . . . . ۶۳
۱۶.۴ نمودار مارگون برای مدلM7 با برازش رویه برای پارامتر های μ و σ ۶۴
۱۷.۴ نمودار مارگون قاب سمت چپ مدل M8 و قاب سمت راست مدل M9 . . . . . ۶۵
۱۸.۴ نمودار مارگون برای مدل M9 با دو متغیر توضیحی ۶۶
۱۹.۴ برازش مقادیر برای داده‌های جرم جنایت با استفاده از تابع GMRF ۶۸

فهرست جداول

۱.۱ جدول توزیع‌های خانواده نمایی ۵

۱.۳ جدول مقادیر معیارهای AIC, CV, BIC, C_p برای انتخاب بهترین مدل ۴۵

۱.۴ تفسیر الگوهای مختلف در نمودار مارپیچ ۵۲
۲.۴ گزینش بهترین مدل با استفاده معیارهای AIC ۵۳
۳.۴ گزینش بهترین مدل با استفاده معیارهای AIC ۵۳
۴.۴ گزینش بهترین مدل با استفاده معیارهای AIC ۵۴
۵.۴ مقادیر به‌دست آمده با استفاده از AIC ۵۶
۶.۴ مقادیر به‌دست آمده با استفاده از معیار AIC ۵۷
۷.۴ مقادیر به دست آمده با استفاده از معیار AIC ۵۷
۸.۴ گزینش بهترین مدل با استفاده معیارهای AIC و BIC ۶۰
۹.۴ گزینش بهترین مدل با استفاده معیارهای AIC و BIC ۶۲
۱۰.۴ گزینش بهترین مدل با استفاده معیارهای AIC ۶۳
۱۱.۴ گزینش بهترین مدل با استفاده معیارهای AIC ۶۵
۱۲.۴ مدل برازشی به داده‌های columb با استفاده از تابع gmrf ۶۷

آ.۱ خلاصه‌ای از توزیع‌های پیوسته موجود در بسته‌ی gamlss.dist با تابع پیوند پیش‌فرض ۷۱

آ.۲ خلاصه‌ای از توزیع‌های گسسته موجود در بسته‌ی gamlss.dist با تابع پیوند پیش‌فرض ۷۲

فصل ۱

مفاهیم اولیه و مورد نیاز

۱.۱ مقدمه

در این فصل مفاهیم اساسی و اولیه که در فصل‌های آینده به آن نیاز داریم را معرفی می‌کنیم. ابتدا مدل خطی که یکی از قدیمی‌ترین و پرکاربردترین روش‌های آماری است را بیان کرده و در ادامه مدل‌های خطی تعمیم‌یافته، مدل‌های جمعی و مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل (GAMLSS) را شرح می‌دهیم.

۲.۱ مدل‌های خطی

مدل رگرسیون خطی[۱] یک مدل ساده اما کاربردی است که سالیان متمادی مبنای تجزیه و تحلیل داده‌ها را تشکیل می‌دهد. این تکنیک زمانی که رابطه بین متغیر وابسته و متغیرهای مستقل خطی است، ابزاری قدرتمند است و تکیه بر آن در بسیاری از کاربردها، محدود کننده است زیرا در بسیاری از پدیده‌ها روابط خطی قابل تعریف نیستند [۱۳]. یک مدل رگرسیون خطی بر پایه‌ی فرض‌های اساسی زیر ارائه می‌شود:

- استقلال مشاهدات

[۱] Linear regression model

- وجود رابطه خطی بین متغیرهای پاسخ و توضیحی
- ثابت بودن واریانس مشاهدات

در این مدل X به عنوان متغیر مستقل یا پیش‌گو عمل می‌نماید که مقادیر آن به وسیله آزمایشگر کنترل می‌شود این متغیر کنترل شده را متغیر پیش‌بین می‌نامند و Y که به متغیر X وابسته است را متغیر اثر یا پاسخ می‌نامند.
فرض کنید متغیر پاسخ Y_i با رابطه‌ی خطی زیر به متغیر مستقل X_i، در ارتباط باشد:

$$Y_i = \beta_۰ + \beta_۱ X_i + + \epsilon_i \qquad i = ۱, ۲, \cdots, n \tag{۱.۱}$$

که در آن $\beta_۰$ و $\beta_۱$ مجموعه پارامترهای نامعلوم‌اند که $\beta_۰$ عرض از مبداء و $\beta_۱$ شیب خط رگرسیونی است که ضرایب رگرسیونی نامیده می‌شوند و ϵ خطاست که متغیرهای غیرقابل مشاهده‌اند که دارای میانگین صفر و واریانس نامعلوم $\sigma^۲$ است. چنان‌چه بیش از یک متغیر مستقل در مدل داشته باشیم، مدل رگرسیون چندجمله‌ای حاصل می‌گردد که به‌صورت زیر است:

$$Y_i = \beta_۰ + \beta_۱ X_{i۱} + \cdots + \beta_r X_{ir} + \epsilon_i \qquad i = ۱, ۲, \cdots, n$$

در عمل، برآورد $\boldsymbol{\beta} = (\beta_۰, \beta_۱, \cdots, \beta_r)$ را به گونه‌ای برمی‌گزینیم که مجموع توان‌های دوم انحراف‌ها از خط حقیقی را کمینه کند. با به‌کارگیری روش کمترین توان‌های دوم، برآوردگرهای زیر حاصل می‌شود:

$$\hat{\boldsymbol{\beta}} = argmin_\beta (\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta})$$

در نتیجه

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-۱} \boldsymbol{X}^T \boldsymbol{Y} \tag{۲.۱}$$

هم‌چنین برآورد $\boldsymbol{\beta}$ از روش ماکزیمم درستنمایی(MLE) به‌صورت (۲.۱) است. برآورد پارامتر در روش درستنمایی ماکزیمم به توزیع متغیر پاسخ بستگی دارد، چون توزیع آن نرمال است بنابراین برآورد پارامتر از هر دو روش یکسان است.
با برآورد پارامترهای مدل، مقادیر برازشی $\hat{Y}$ از رابطه‌ی زیر حاصل می‌شود:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-۱} \boldsymbol{X}^T \boldsymbol{Y} = \boldsymbol{H}Y$$

H ماتریس هت نامیده می‌شود .
تفاوت بین مقادیر واقعی و مقادیر برازشی، مانده‌های رگرسیونی نامیده می‌شود.

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y}(I - H)$$

درجه آزادی یک مدل خطی، تعداد پارامترهای β که مستقل‌اند، به‌عبارتی تعداد ستون‌های مستقل $\boldsymbol{X}$ نشان دهنده اثر ماتریس هت است [۱] .

$$df = tr(H) = p$$

از اهداف اصلی رگرسیون، ارائه یک خلاصه یا کاهش داده‌های مشاهده شده به منظور کشف و ارائه‌ی ارتباط بین متغیرهای پاسخ و توضیحی است و یک هدف مهم دیگر آن، استفاده از مدل (۱.۱) و یافتن رابطه‌ای است که بتوان اثر تغییرات یک یا چند متغیر را بر روی متغیرهای دیگر پیش‌بینی کرد. پیش‌بینی یک جنبه بسیار مهم رگرسیون است. این روش دارای مزایا و معایبی است. یک مزیت این است که این روش را می‌توان به طیف گسترده‌ای از ساختارهای داده‌ها گسترش داد. برای مثال مدل‌های خطی تعمیم‌یافته توسط مک کولاغ و نلدر در سال ۱۹۸۹ معرفی شدند [۱۱].

۳.۱ مدل‌های خطی تعمیم‌یافته

رویکرد مدل‌های خطی تعمیم‌یافته[۲]، یک مدل پارامتری بوده و یک تعمیم انعطاف‌پذیری از مدل‌های خطی می‌باشد. هنگامی که متغیر پاسخ Y شمارشی باشد یا رابطه بین متغیر پاسخ و توضیحی خطی نباشد دیگر مدل‌های خطی کلاسیک جوابگوی این‌گونه مسائل نیستند و برای این منظور، مدل‌های خطی تعمیم‌یافته را به‌کار می‌بریم، که توسط جان نلدر و رابرت دربرن (۱۹۷۲) ارائه شدند و توسط مک کولاغ و نلدر در سال ۱۹۸۹ توسعه یافتند و برای انجام آن از بسته‌های محاسباتی آماری استفاده کردند. در این مدل متغیر پاسخ Y دارای توزیع نرمال است ولی در بسیاری از موقعیت‌های عملی، این فرض حتی به‌طور تقریبی برقرار نیست. مدل خطی تعمیم‌یافته به‌طور گسترده مدل‌های آماری پرکاربرد را در برمی‌گیرد از قبیل رگرسیون خطی برای پاسخ‌هایی که به‌صورت نرمال توزیع شده‌اند، مدل لجستیک برای داده‌های دوتایی (دودویی) و مدل‌های لگاریتم خطی برای داده‌های شمارشی (نوعی از داده آماری که در آن مشاهدات صرفا می‌توانند مقادیر اعداد صحیح غیرمنفی بگیرند). مدل‌های خطی تعمیم‌یافته این امکان را فراهم می‌سازد که متغیر پاسخ از خانواده توزیع نمایی تبعیت کند. خانواده نمایی شامل توزیع‌های نرمال،دوجمله‌ای، پواسن، هندسی، نمایی، گاما است.
در خانواده مدل‌های خطی تعمیم‌یافته دو مؤلفه مهم وجود دارد [۳، ۲۸، ۳۴] :

- توزیع متغیر پاسخ
- مدلی که میانگین متغیر پاسخ را به متغیرهای توضیحی X ارتباط دهد.

در مدل‌های خطی تعمیم‌یافته، واریانس مشاهدات ثابت فرض شده و تابعی از میانگین است.

[۲] Generalized linear model

۱.۳.۱ ساختار مدل‌های خطی تعمیم‌یافته

مدل‌های خطی تعمیم‌یافته دارای ویژگی‌های کلی زیر است:

- متغیرهای پاسخ $Y_1, Y_2, \cdots, Y_n$ مستقل با میانگین‌های $\mu_1, \cdots, \mu_n$ و واریانس‌های $\sigma_1^2, \cdots, \sigma_n^2$ می‌باشد و دارای توزیعی از خانواده توزیع‌های نمایی است.
- پیش‌گوی خطی مدل به‌صورت

$$\eta_i = \beta_0 + X_i^T\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ji}$$

تعریف می‌شود که شامل $\boldsymbol{X_i}$ یک بردار $(p \times 1)$ از متغیرهای توضیحی و $\boldsymbol{\beta}$ بردار $(p \times 1)$ از پارامترها می‌باشد.

- تابع پیوند g میانگین μ_i را با یک تابع معلوم و مشتق‌پذیر از میانگین مدل‌سازی می‌کند. این تابع، میانگین پاسخ را در پیش‌گوی خطی برای مقادیر $i = 1, \cdots, n$ به‌صورت $\eta_i = g(\mu_i)$ بیان می‌کند. هم‌چنین داریم:

$$\mu_i = E(Y_i) = g^{-1}(\eta_i)$$

تابع پیوند اغلب غیرخطی و بدون محدودیت است اما میانگین بعضی از توزیع‌ها مانند توزیع دو جمله‌ای محدودیت دارد. در مدل‌های خطی کلاسیک، تابع پیوند همان تابع همانی است، بنابراین $\mu = \eta$ و μ هر مقداری از اعداد حقیقی را می‌تواند اختیار کند.

تابع چگالی متغیر پاسخ Y با میانگین $E(Y) = \mu$ و پارامترهای θ و ϕ متعلق به خانواده‌ی توزیع‌های نمایی است اگر بتوان آن را به‌صورت زیر نوشت:

$$f_Y(y|\theta, \phi) = exp\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\} \qquad (۳.۱)$$

که در آن θ پارامتر طبیعی یا پارامتر کانونی و ϕ پارامتر مقیاس یا پراکندگی نامیده می‌شود که به‌طور معمول، معلوم فرض می‌شود و تابعی از میانگین متغیر پاسخ است. در نتیجه

$$E(Y) = b'(\theta) \qquad (۴.۱)$$
$$Var(Y) = \phi V(\mu)$$

با توجه به ویژگی $\int \mathrm{f(y|\theta, \phi)\, dy} = 1$ و مشتقات اول و دوم نسبت به θ از هر دو طرف معادله، نتایج (۴.۱) حاصل می‌شود.

$$\int \phi^{-1}(y - b'(\theta))\mathrm{f(y|\theta, \phi)\, dy} = 0$$

بنابراین

$$\mu = E(Y) = b'(\theta)$$

و با توجه به مشتق دوم داریم:

$$\int \{\phi^{-1}[y - b'(\theta)]^2 - b''(\theta)\} \mathrm{f}(\mathrm{y}|\theta, \phi)\, \mathrm{d}y = 0$$

درنتیجه

$$Var(Y) = \phi b''(\theta) = \phi b''(b'^{-1}(\mu)) = \phi V(\mu)$$

V تابع واریانس نامیده می‌شود و تابعی از μ است.
در جدول ۱.۱ برخی از توزیع‌ها که متعلق به خانواده‌ی توزیع‌های نمایی به همراه تابع پیوند مورد بررسی قرار گرفته اند [۳، ۲۸، ۳۴] :

جدول ۱.۱: جدول توزیع‌های خانواده نمایی

توزیع	پیوند	$\theta(\mu)$	ϕ	$c(y,\phi)$	$b(\theta)$	μ	$Var(Y)$
نرمال $N(\mu,\sigma^2)$	همانی	μ	σ^2	$\frac{-y^2}{2\phi} - log(\sqrt{2\pi\phi})$	$\frac{\theta^2}{2}$	θ	σ^2
برنولی $B(1,\mu)$	لوجیت	$log(\frac{\mu}{1-\mu})$	1	0	$log(1+e^\theta)$	$\frac{e^\theta}{1+e^\theta}$	$\frac{e^\theta}{(1+e^\theta)^2}$
پواسن $P(\mu)$	لگاریتمی	$log(\mu)$	1	$-log(y!)$	$exp(\theta)$	$exp(\theta)$	$exp(\theta)$

۴.۱ مدل جمعی تعمیم‌یافته

این مدل، یک مدل ناپارامتری بوده و بسط مدل‌های خطی تعمیم‌یافته است که خود بسط مدل‌های خطی می‌باشند. مدل جمعی تعمیم‌یافته[۳] روشی بسیار مناسب برای بررسی رابطه بین متغیر پاسخ و متغیرهای مستقل و هم‌چنین تحلیل داده‌ها ارائه می‌دهد و به وسیله‌ی بسته نرم افزاری mgcvR قابل اجرا می‌باشد. در مدل‌های جمعی تعمیم‌یافته بر خلاف مدل رگرسیون خطی، داده‌ها شکل منحنی پاسخ را مشخص می‌کنند. در این مدل فرض بر این است که متغیر پاسخ Y دارای توزیعی از خانواده‌ی توزیع‌های نمایی با میانگین $\mu = E(Y|x_1, \cdots, x_p)$ می‌باشد که از طریق تابع پیوند g به متغیرهای پیش‌گو $\boldsymbol{X}$ متصل می‌شود. ساختار مدل‌های جمعی تعمیم‌یافته به‌صورت زیر است:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \beta_0 + s_1(x_1) + \cdots + s_J(x_J)$$

[۳] Generalized additive models

که در آن s_j ها $j = ۱, \cdots, J$ توابع ناپارامتری هموار و نامعلوم می‌باشند. در مدل‌های جمعی تعمیم‌یافته $\sum_{j=۱}^{J} s_j(x_j)$ جانشین توابع خطی $\sum_{j=۱}^{p} \beta_j X_j$ در مدل‌های خطی تعمیم‌یافته می‌شود. دارا بودن هموارسازها (اسپلاین‌های مکعبی، اسپلاین‌های جریمه‌ای، لوییس...) یکی از مزایای مهم این مدل است که باعث توانایی این مدل در شناسایی روابط غیرخطی شده است. به‌عبارتی با اطلاعات بیشتری از روابط بین داده‌ها، کیفیت پیش‌بینی متغیر پاسخ را به حداکثر می‌رساند. هم‌چنین به خاطر انعطاف‌پذیر بودن در تعیین نوع و درجه ارتباط به یک مدل محبوب تبدیل شده است [۲۸].

۱.۴.۱ مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل

مدل‌های جمعی تعمیم‌یافته برای مکان، مقیاس و شکل[۴] (GAMLSS) دارای یک ساختار کلی برای برازش انواع مدل‌های رگرسیونی است که توسط ریگبی و استامینوپولوس[۵] (۲۰۰۱، ۲۰۰۵) معرفی شدند و بعدها برای غلبه بر برخی از محدودیت‌های مرتبط با مدل‌های خطی تعمیم‌یافته و مدل‌های جمعی تعمیم‌یافته گسترش یافت. این مدل در حقیقت قادر به ارائه یک مدل هموار و انعطاف‌پذیر است، فرض می‌شود که متغیر پاسخ از توزیعی پارامتری هموار است که ممکن دم کلفت یا با کشیدگی مثبت یا منفی باشد پیروی می‌کند، علاوه بر این تمام پارامترهای توزیع، مکان (میانگین) ، مقیاس (واریانس) و شکل توزیع (کشیدگی یا چولگی) را می‌توان به عنوان توابع خطی یا غیرخطی و یا هموارساز از متغیرهای توضیحی مدل‌سازی کرد [۲۴، ۱۷، ۲۸] .

ساختار مدل

در مدل جمعی تعمیم‌یافته برای مکان، مقیاس و شکل فرض می‌شود برای هر $i = ۱, ۲, \cdots, n$ متغیر پاسخ Y_i دارای تابع احتمال $f_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ مشروط بر $(\mu_i, \sigma_i, \nu_i, \tau_i)$ بردار چهار پارامتر توزیع که هر کدام می‌تواند تابعی برای متغیرهای توضیحی باشد و آن را با $Y_i \sim D(\mu, \sigma, \nu, \tau)$ نشان می‌دهند.

پارامترهای $(\mu_i, \sigma_i, \nu_i, \tau_i)$ پارامترهای توزیع می‌باشد. پارامترهای شرطی μ و σ به ترتیب برای مکان و مقیاس و همچنین ν و τ برای شکل توزیع (کشیدگی و چولگی) به کار می‌روند. این مدل می‌تواند برای پارامترهای هر توزیع به‌کار رود و هم‌چنین می‌تواند به بیش از چهار پارامتر توزیع تعمیم یابد.

با فرض این‌که $\boldsymbol{Y}^T = (Y_۱, Y_۲, \cdots, Y_n)$ برداری با طول n متغیر پاسخ باشد فرموله کردن GAMLSS به‌صورت زیر است:

---

[۴] Generalized additive model for location, scale and shape

[۵] Rigby and stasinopoulos

برای $k = ۱,۲,۳,۴$ و فرض $g_k(\cdot)$ تابع پیوند یکنوا پارامترهای توزیع

$$\boldsymbol{\theta}_k = (\theta_۱, \theta_۲, \theta_۳, \theta_۴) = (\mu_i, \sigma_i, \nu_i, \tau_i)$$

به متغیر پیش‌بینی η_k داریم:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=۱}^{J'_k} h_{jk}(x_{jk}), \quad k = ۱,۲,۳,۴$$

که در آن $\boldsymbol{\beta_k^T} = (\beta_{۱k}, \beta_{۲k}, \cdots, \beta_{J'_k k})$ بردار پارامترها به طول J'_k، X_k ماتریس طرح با بعد $n \times J'_k$ ، h_{jk} ماتریس طرح اثرات تصادفی با بعد $n \times q_{jk}$ و x_{jk} (پارامترهای اثرات تصادفی) متغیر تصادفی q_{jk} بعدی که دارای توزیع

$$x_{jk} \sim N(۰, G_{jk}^{-۱})$$

به‌طوری‌که $G_{jk}^{-۱}$ معکوس تعمیم‌یافته‌ی ماتریس متقارن $q_{jk} \times q_{jk}$ به شکل $G_{jk} = G_{jk}(\lambda_{jk})$ است که می‌تواند به بردار ابر پارامتر $\boldsymbol{\lambda_{jk}}$ وابسته باشد. به دلیل دستیابی به معادلات بسیار ساده در روند برازش مدل، متغیرهای تصادفی را به توزیع نرمال محدود می‌کنیم که در حالت کلی این محدودیت ممکن است حذف شود.

اگر عبارت جمعی در پارامترهای توزیع نباشد مدل GAMLSS پارامتری خطی به‌صورت زیر است:

$$\begin{aligned}
g_۱(\mu) &= \eta_۱ = X_۱\beta_۱ \\
g_۲(\sigma) &= \eta_۲ = X_۲\beta_۲ \\
g_۳(\nu) &= \eta_۳ = X_۳\beta_۳ \\
g_۴(\tau) &= \eta_۴ = X_۴\beta_۴
\end{aligned}$$

با قرار دادن تابع $h_k(X_k, \beta_k)$ که h_k که تابع غیرخطی است به جای $X_k\beta_k$ مدل GAMLSS پارامتری غیرخطی به صورت زیر به‌دست می‌آید:

$$\begin{aligned}
g_۱(\mu) &= \eta_۱ = h_۱(X_۱, \beta_۱) \\
g_۲(\sigma) &= \eta_۲ = h_۲(X_۲, \beta_۲) \\
g_۳(\nu) &= \eta_۳ = h_۳(X_۳, \beta_۳) \\
g_۴(\tau) &= \eta_۴ = h_۴(X_۴, \beta_۴)
\end{aligned}$$

پس از تعیین مدل، بردار پارامترهای $\boldsymbol{\beta}_k$ و پارامترهای اثرات تصادفی x_{jk} را می‌توان با ماکسیمم کردن تابع درستنمایی تاوانیده[۶] l_p که به‌صورت زیر است برآورد نمود:

$$\ell_p = \ell - \frac{۱}{۲}\sum_{k=۱}^{p}\sum_{j=۱}^{J_k} x'_{jk}G_{jk}(\lambda_{jk})x_{jk}$$

[۶] Penalized likelihood

که $\ell = \sum_{i=۱}^{n} log\{D(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)\}$ لگاریتم تابع درستنمایی است.

خواص توابع GAMLSS [۲۴، ۱۷، ۲۸] :

- چارچوبی بسیار انعطاف‌پذیر برای انواع مدل‌های رگرسیونی ایجاد می‌کند.
- اجرا و پیاده‌سازی آن‌ها برای توزیع‌های جدید و هم‌چنین عبارات جمعی جدید آسان است.
- در این تابع، متغیر پاسخ می‌تواند هر توزیعی داشته باشد و تمام پارامترهای توزیع می‌توانند به‌عنوان تابعی از متغیرهای توضیحی مدل شوند.

فصل ۲

سیر تحول اسپلاین‌ها و انواع آن‌ها

اسپلاین‌ها توابعی شامل چندجمله‌ای‌هایی با مرتبه m روی زیربازه‌هایی از یک بازه مشخص هستند و با شرایط پیوستگی تعریف‌شده‌ای به هم می‌پیوندند و از مرزهای مشترک بین دو زیربازه، موسوم به گره، عبور می‌کنند. اسپلاین‌ها ترکیب خطی از توابع پایه و وزن‌هایی برای هر تابع پایه، ساخته می‌شوند. توابع پایه توابعی بر حسب مشاهدات هستند و وزن‌های توابع پایه به عنوان پارامتر قلمداد می‌شوند که می‌توان با روش‌هایی هم‌چون روش کمترین توان‌های دوم خطا یا روش درستنمایی ماکسیمم آن‌ها را برآورد کرد. در طی سال‌هایی که از معرفی اسپلاین‌ها می‌گذرد، مبانی نظری آن‌ها توسعه یافته‌اند و نسخه‌های تعمیم‌یافته متفاوتی از آن‌ها معرفی شده‌اند. حوزه کاربردهای این ابزار قدرتمند ناپارامتری نیز به شدت گسترش یافته است.

اسپلاین‌ها انواع مختلفی هم‌چون اسپلاین‌های هموار، رگرسیون اسپلاینی، و B-اسپلاین‌ها دارند. برای اطلاعات جامع‌تر در این زمینه می‌توانید به هیستی و تیبشیرانی [۱۲]، دی بور [۶]، واهبا [۳۲]، گرین و سیلورمن [۱۱] مراجعه کنید.

۱.۲ توابع پایه

در جبر خطی منظور از یک پایه در یک فضای برداری، مجموعه‌ای از بردارهای موجود در آن فضا است به‌طوری که مستقل خطی باشند و هر بردار دیگر در فضای برداری را بتوان از

ترکیب خطی آن بردارها به‌دست آورد. برای مثال چندجمله‌ای درجه دو با ضرایب حقیقی را می‌توان به‌صورت

$$y = ۱a + bt + ct^۲$$

نوشت که در واقع از ترکیب خطی توابع پایه $t^۲, t, ۱$ شکل گرفته است [۳۰] .

۲.۲ تابع توانی بریده‌شده

یک منحنی اسپلاین از ترکیب خطی توابع پایه و پارامترهای اسپلاین تشکیل می‌شود که تابع توانی بریده‌شده عضوی از مجموعه توابع پایه است.
فرض کنید تعداد مشاهدات $x_n, \cdots, x_۱$ داشته باشیم و یک خط راست روی بازه $[t_۰, t_{m+۱}]$ که $t_۰ = min(x_i)$ و $t_{m+۱} = max(x_i)$ رسم شده باشد و خط را در نقطه $t_۱$ که $t_۰ < t_۱ < t_{m+۱}$ به دو قسمت تقسیم می‌کنیم به‌طوری که پیوستگی آن حفظ شود و می‌توان معادله خط که در $t_۱$ شکسته شده را به‌صورت ترکیب خطی از توابع پایه ۱ و x و $(x - t_۱)_+$ نوشت که $(x - t_۱)_+$ را تابع توانی بریده‌شده می‌نامند که به‌صورت زیر تعریف می‌شود.

$$(x - t_۱)_+ = max\Big\{۰, (x - t_۱)\Big\} = \begin{cases} x - t_۱ & x \geq t_۱ \\ ۰ & o.w. \end{cases}$$

برای به‌دست آوردن معادله خط، تابع f که در $t_۱$ شکسته شده در بازه $[t_۰, t_۱]$ دارای یک معادله خط و در بازه $[t_۱, t_{m+۱}]$ دارای معادله خط دیگری است که به‌صورت زیر داریم:

$$f(x) = \begin{cases} \beta_۰ + \beta_۱ x & x < t_۱ \\ \beta'_۰ + (\beta_۱ + \alpha_۱)x & x \geq t_۱ \end{cases}$$

با توجه به پیوستگی f در نقطه $t_۱$ داریم:

$$\beta_۰ + \beta_۱ t_۱ = \beta'_۰ + (\beta_۱ + \alpha_۱)t_۱$$

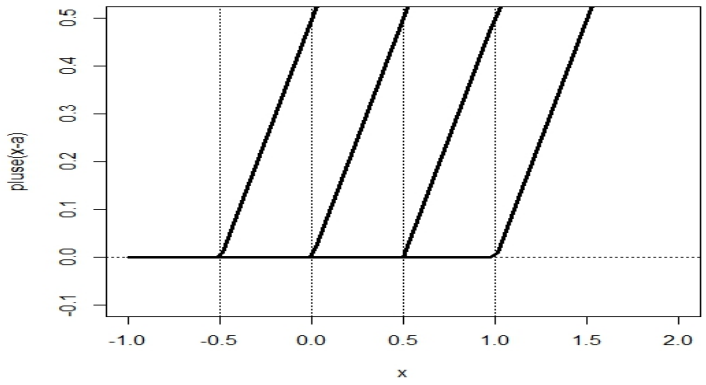
یا به‌طور معادل داریم:

$$\beta'_۰ = \beta_۰ - \alpha_۱ t_۱$$

بنابراین معادله کلی f به‌صورت زیر است.

$$\begin{aligned} f(x) &= (\beta_۰ + \beta_۱ x)I_{\{x<t_۱\}} + (\beta'_۰ + (\beta_۱ + \alpha_۱)x)I_{\{x\geq t_۱\}} \\ &= (\beta_۰ + \beta_۱ x)I_{\{x<t_۱\}} + (\beta_۰ + \beta_۱ x + \alpha_۱(x - t_۱))I_{\{x\geq t_۱\}} \\ &= \beta_۰ + \beta_۱ x + \alpha_۱(x - t_۱)I\{x \geq t_۱\} \\ &= \beta_۰ + \beta_۱ x + \alpha_۱(x - t_۱)_+ \end{aligned}$$

بنابراین تابع f را می‌توان به‌صورت ترکیب خطی از توابع پایه ۱ و x و $(x-t_1)_+$ نوشت [۱۴].
در شکل ۱.۲ تابع توانی بریده‌شده از مرتبه یک که در گره‌های $-0.5, 0, 0.5, 1$ بریده‌شده است، نشان داده شده‌اند.



شکل ۱.۲: توابع توانی بریده‌شده مرتبه یک

۳.۲ اسپلاین‌ها

اگر برای تحلیل یک مدل رگرسیونی، پذیره خطی بودن رابطه برقرار نباشد و نتوان از یک تابع پارامتری مشخص به‌سادگی برای مدل استفاده کرد، از روش‌های ناپارامتری رگرسیونی استفاده می‌شود. اسپلاین‌ها به‌عنوان یکی از ابزارهای درونیابی، از جمله روش‌های ناپارامتری برای مدل‌بندی رگرسیون ناپارامتری است. مکانیسم عملکرد روش برازش مبتنی بر اسپلاین‌ها به این صورت است که ابتدا کل بازه‌ای که مشاهدات در آن قرار دارند به تعدادی زیربازه تقسیم می‌شود. سپس در هر زیربازه یک منحنی از درجه $p \geq 1$ به مشاهدات برازش داده می‌شود به‌طوری که منحنی‌ها در نقاط مرزی بین زیربازه‌ها (یا همان نقاط گره) به هم متصل می‌شوند. فرض کنید تعدادی مشاهده داشته باشیم که در بازه $[t_0 = min(x_i), t_{m+1} = max(x_i)]$ قرار دارند. اگر طول بازه را به فواصل $[t_0, t_1], [t_1, t_2], \cdots, [t_m, t_{m+1}]$ تقسیم کنیم، آن‌گاه چنان‌چه منحنی‌های برازش‌شده f، در هر زیربازه به‌صورت یک چندجمله‌ای از مرتبه P باشد، یک تابع اسپلاین از مرتبه P با شرایط زیر داریم:

- تابع f در گره $i = 1, 2, \cdots, m$ t_j پیوسته است.
- مشتق‌های تابع f تا مرتبه P-1 وجود دارند و در گره‌ها پیوسته هستند .

ضابطه کلی توابع اسپلاین به‌صورت زیر است:

$$f(x) = \sum_{k=0}^{p} \beta_k x^k + \sum_{j=1}^{m} \beta_{p+j}(x - t_j)_+^p \qquad (۱.۲)$$

که در آن $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_{p+m})$ بردار ضرایب است. می‌توان اسپلاین‌های مرتبه p با $t_1, \cdots, t_m$ گره، که در هر قطعه یک چندجمله‌ای از مرتبه $p-1$ و همچنین دارای مشتقات پیوسته تا مرتبه $p-2$ زیر را به‌صورت

$$f(x) = \sum_{j=1}^{m} h_j(x)\beta$$

بازنویسی کرد که در آن $\boldsymbol{h_j(x)} = (1, x, \cdots, x^p, (x-t_1)_+^p, \cdots, (x-t_m)_+^p)$ توابع پایه هستند. با استفاده از (۱.۲) می‌توان مدل رگرسیون اسپلاینی را به فرم ماتریسی

$$Y = T\beta + \epsilon \qquad (۲.۲)$$

ارائه کرد که در آن $T = (h_j(t_1), \cdots, h_j(t_m))'$ ماتریس طرح $(n \times (p+m+1))$ بعدی است. با به‌کار بردن روش کمترین توان‌های دوم برآوردگر زیر حاصل می‌شود [۳۴، ۸، ۱۴].

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{T'T})^{-1}\boldsymbol{T'Y} \qquad (۳.۲)$$

یکی از جدی‌ترین معایب اسپلاین‌هایی که از توابع پایه ساخته می‌شوند، ناپایداری عددی است که با افزایش درجه چندجمله‌ای، هم‌خطی بین توابع بریده‌شده ایجاد می‌شود در این حالت معادله‌ی $T'T\beta = T'Y$ ناجور نامیده می‌شود و $\hat{\beta}$ ممکن است ناپایدار باشد، بدین مفهوم که تغییرات کوچکی در داده‌ها ممکن است موجب تغییراتی بزرگ در $\hat{\beta}$ شود. در بخش بعدی نوع دیگری از اسپلاین‌ها را معرفی می‌کنیم که از رابطه‌ي بازگشتی برای ساخت توابع پایه استفاده می‌شود تا ضمن رفع مشکل هم‌خطی، چندجمله‌ای‌های متعامدی را ایجاد می‌کنند [۲]. محققین بسیاری بر روی اسپلاین‌ها، مباحث نظری و کاربردهای آن‌ها پژوهش کرده‌اند. به‌عنوان چند نمونه می‌توان به هیستی و تیبشیرانی [۱۲]، وند و اورمرد [۳۳] و روپرت و همکارانش [۲۵] اشاره کرد.

۴.۲ اسپلاین مکعبی (CS)

اسپلاین مکعبی[1] یک ابزار قدرتمند برای تجزیه و تحلیل داده‌ها می‌باشد و کاربرد زیادی در هموارسازی دارد. یک اسپلاین درجه سوم با گره‌های $t_1, t_2, \cdots, t_m$ در بازه $[k_0, k_{m+1}]$ به‌طوری که $k_0 < t_1 < \cdots < t_m < k_{m+1}$ و مشتقات اول و دوم پیوسته به‌صورت زیر تعریف می‌گردد:

[1] Cubic splines

$$f(x) = \sum_{k=0}^{3} \beta_k x^k + \sum_{j=0}^{m} \beta_{3+j}(x - t_j)_+^3$$

قطعه‌های چندجمله‌ای به نحوی برازش می‌یابند که تابع $f(x)$ و مشتقات اول و دوم آن در هر گره پیوسته باشند در نتیجه روی تمام بازه پیوسته است.

۱.۴.۲ درونیابی اسپلاین مکعبی

اگر در هر زیربازه $[t_i, t_{i+1}]$ برای هر $i = 0, 1, \cdots, m+1$ ، t_i باشد آن را $s_i(t_i) = f(t_i)$ درونیاب اسپلاین مکعبی می‌نامند و به‌صورت زیر تعریف می‌شود:

$$S(x) = \begin{cases} s_0(x) & t_0 \le x \le t_1 \\ s_1(x) & t_1 \le x \le t_2 \\ s_2(x) & t_2 \le x \le t_3 \\ \vdots & \vdots \\ s_{m-1}(x) & t_{m-1} \le x \le t_m \\ s_m(x) & t_m \le x \le t_{m+1} \end{cases}$$

که هر یک از s_i ها یک چندجمله‌ای درجه سوم است که به‌صورت

$$s_i(x) = a_i(x - t_i)^3 + b_i(x - t_i)^2 + c_i(x - t_i) + d_i, \qquad i = 0, 1, 2, \cdots, m+1$$

است. یک روش معمول برای تعیین چندجمله‌ای درجه سوم در هر زیر بازه تعیین ضرایب چندجمله‌ای می‌باشد. ضرایب a_i, b_i, c_i, d_i مجهولات هر بازه را تشکیل می‌دهند. مشتقات اول و دوم برای $m+1$ معادله ضروری است و داریم:

$$s_i'(x) = 3a_i(x - t_i)^2 + 2b_i(x - t_i) + c_i \qquad (۴.۲)$$

$$s_i''(x) = 6a_i(x - t_i) + 2b_i. \qquad (۵.۲)$$

۲.۴.۲ خواص اسپلاین‌های مکعبی

اسپلاین‌های مکعبی در شرایط زیر صدق می‌کنند:

- تابع قطعه‌ای $s(x)$ تمام داده‌ها را در بر می‌گیرد.
- تابع $s(x)$ در بازه $[t_0, t_{m+1}]$ پیوسته است، به‌طوری که

$$s_i(t_i) = s_i(t_i) \quad i = ۱, ۲, \cdots, m$$

- تابع $s'(x)$ در بازه $[t_۰, t_{m+۱}]$ پیوسته است، به‌طوری که

$$s'_{i-۱}(t_i) = s'_i(t_i) \quad i = ۱, ۲, \cdots, m$$

- تابع $s''(x)$ در بازه $[t_۰, t_{m+۱}]$ پیوسته است، به‌طوری که

$$s''_{i-۱}(t_i) = s''_i(t_i) \quad i = ۱, ۲, \cdots, m$$

- شرط‌های زیر به‌صورت اختیاری هستند:

$$s''(t_۰) = s''(t_{m+۱}) = ۰$$
$$s'(t_۰) = f'(t_۰) \qquad s'(t_{m+۱}) = f'(t_{m+۱})$$

ضرایب چندجمله‌ای در هر زیربازه و چندجمله‌ای اسپلاین مکعبی با حل این دستگاه به‌دست می‌آیند:
چون تابع $S(x)$ تمام مشاهدات را درونیابی می‌کند، در نتیجه

$$\begin{aligned} s_i(t_i) =& f(t_i) \\ s_i(t_i) =& a_i(t_i - t_i)^۳ + b_i(t_i - t_i)^۲ + c_i(t_i - t_i) + d_i = d_i \\ f(t_i) =& d_i \end{aligned}$$

و با پیوستگی تابع $S(x)$ در تمام فواصل داریم

$$\begin{aligned} s_i(t_i) =& s_{i-۱}(t_i) \\ s_{i-۱}(t_i) =& a_{i-۱}(t_i - t_{i-۱})^۳ + b_{i-۱}(t_i - t_{i-۱})^۲ + c_{i-۱}(t_i - t_{i-۱}) + d_{i-۱}. \end{aligned}$$

بنابراین

$$d_i = a_{i-۱}(t_i - t_{i-۱})^۳ + b_{i-۱}(t_i - t_{i-۱})^۲ + c_{i-۱}(t_i - t_{i-۱}) + d_{i-۱}$$

با فرض $h = (t_i - t_{i-۱})$ رابطه‌ی زیر حاصل می‌گردد:

$$d_i = a_{i-۱}h^۳ + b_{i-۱}h^۲ + c_{i-۱}h + d_{i-۱} \qquad (۶.۲)$$

با تساوی مشتق اول در نقاط میانی با توجه به رابطه (۴.۲) روابط زیر به‌دست می‌آید:

$$\begin{aligned} s'_i(t_i) =& c_i \\ s'_i(t_i) =& s'_{i-۱}(t_i) \\ s'_i(t_i) =& ۳a_{i-۱}(t_i - t_{i-۱})^۲ + ۲b_{i-۱}(t_i - t_{i-۱}) + c_{i-۱}. \end{aligned}$$

درنتیجه

$$c_i = ۳a_{i-۱}h^۲ + ۲b_{i-۱}h + c_{i-۱} \qquad (۷.۲)$$

و با تساوی مشتق دوم در نقاط میانی و با توجه به رابطه‌ی (۵.۲) روابط زیر به‌دست می‌آید:

$$s''_{i+۱}(t_i) = ۲b_{i+۱},$$

$$\begin{aligned} s''_{i+۱}(t_i) &= s''(t_{i+۱}) \\ s''_i(t_{i+۱}) &= ۶a_i(t_{i+۱} - t_i) + ۲b_i \\ ۲b_{i+۱} &= ۶a_ih + ۲b_i. \end{aligned}$$

پس با جایگزین کردن $s''_i(t_i) = M_i$ رابطه‌ی زیر حاصل می‌شود:

$$\begin{aligned} s''_i(t_i) &= M_i \\ ۲b_i = M_i \Longrightarrow b_i &= \frac{M_i}{۲} \end{aligned}$$

اکنون a_i را می‌توان به‌دست آورد:

$$a_i = \frac{۲b_{i+۱} - ۲b_i}{۶h} = \frac{۲(\frac{M_{i+۱}}{۲}) - ۲(\frac{M_i}{۲})}{۶h} \qquad (۸.۲)$$

با استفاده از رابطه‌ي (۶.۲) و با جایگزین کردن a_i, b_i, d_i مقدار c_i حاصل می‌شود:

$$c_i = \frac{f_{i+۱} - f_i}{h} - (\frac{M_{i+۱} + ۲M_i}{۶})h \qquad (۹.۲)$$

اکنون که تمامی ضرایب اسپلاین درجه سه محاسبه شدند به‌راحتی می توان توابع اسپلاین را به‌دست آورد. ضرایب چندجمله‌ای به‌دست آمده به‌صورت زیر است:

$$\begin{cases} a_i = \frac{M_{i+۱} - M_i}{۶h} \\ b_i = \frac{M_i}{۲} \\ c_i = \frac{f_{i+۱} - f_i}{h} - (\frac{M_{i+۱} + ۲M_i}{۶})h \\ d_i = f_i \end{cases}$$

با قرار دادن ضرایب به‌دست آمده در رابطه‌ی $c_{i+۱} = ۳a_ih^۲ + ۲b_ih + c_i$ می‌توان فرم ماتریسی معادلات را به‌دست آورد:

$$\begin{aligned} ۳(\frac{M_{i+۱} - M_i}{۶h})h^۲ + ۲(\frac{M_i}{۲})h + \frac{f_{i+۱} - f_i}{h} - (\frac{M_{i+۱} + ۲M_i}{۶})h &= \frac{f_{i+۲} - f_{i+۱}}{h} - (\frac{M_{i+۲} + ۲M_{i+۱}}{۶})h \\ \frac{h}{۶}(M_i + ۴M_{i+۱} + M_{i+۲}) &= \frac{f_i - ۲f_{i+۱} + f_{i+۲}}{h} \\ M_i + ۴M_{i+۱} + M_{i+۲} &= ۶(\frac{f_i - ۲f_{i+۱} + f_{i+۲}}{h^۲}) \end{aligned}$$

برای $i = ۱,۲,\cdots,m-۲$ برقرار است.

$$\begin{bmatrix} ۱ & ۴ & ۱ & ۰ & \dots & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۱ & ۴ & ۱ & \dots & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۴ & \dots & ۰ & ۰ & ۰ & ۰ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ ۰ & ۰ & ۰ & ۰ & \dots & ۴ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۰ & ۰ & \dots & ۱ & ۴ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۰ & \dots & ۰ & ۱ & ۴ & ۱ \end{bmatrix} \begin{bmatrix} M_۱ \\ M_۲ \\ M_۳ \\ M_۴ \\ \vdots \\ M_{m-۳} \\ M_{m-۲} \\ M_{m-۱} \\ M_m \end{bmatrix} = \frac{۶}{h^۲} \begin{bmatrix} f_۱ - ۲f_۲ + f_۳ \\ f_۲ - ۲f_۳ + f_۴ \\ f_۳ - ۲f_۴ + f_۵ \\ \vdots \\ f_{m-۴} - ۲f_{m-۳} + f_{m-۲} \\ f_{m-۳} - ۲f_{m-۲} + f_{m-۱} \\ f_{m-۲} - ۲f_{m-۱} + f_m \end{bmatrix}$$

ذکر این نکته لازم است که برای هر بازه‌ی بین دو نقطه‌ی متوالی یک تابع اسپلاین به‌دست می‌آید که با رسم آن‌ها به‌صورت متوالی کل تابع تقریب در کل بازه به‌دست می‌آید [۱۹] .

۳.۴.۲ اسپلاین مکعبی طبیعی (NCS)

حالت خاصی از اسپلاین‌های مکعبی با اضافه کردن محدودیتی همراه است. اسپلاین‌های مکعبی طبیعی[۲] شامل قاعده‌ای است که باید مشتق دوم در بازه‌های انتهایی برابر صفر باشد.

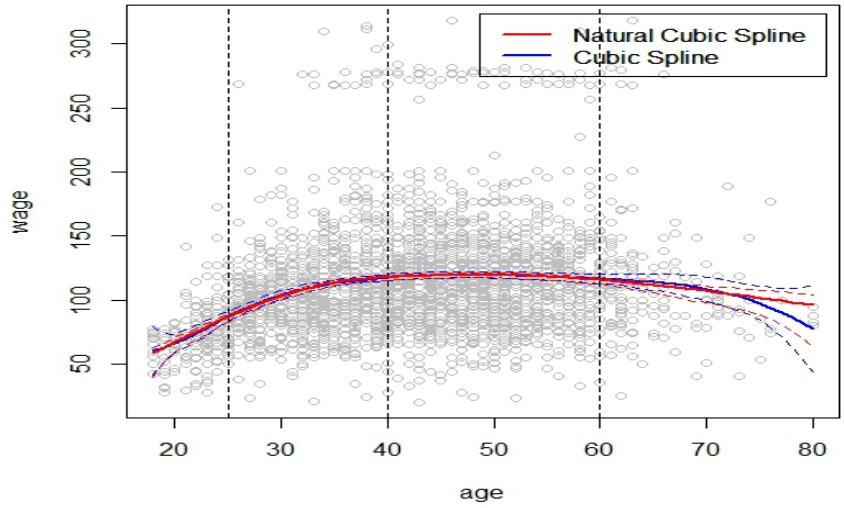
$$M_۱ = M_m = ۰$$

در نتیجه S یک اسپلاین مکعبی منحصربه فرد است اگر در شرایط مرزی آزاد $S''(a) = S''(b)$ صدق کند در نتیجه $a_۰ = b_۰ = c_۰ = d_۰ = ۰$ و منحنی برازشی در بازه‌های انتهایی به‌صورت خطی است. هم‌چنین فرم ماتریسی آن به‌صورت زیر است:

$$\begin{bmatrix} ۱ & ۰ & ۰ & ۰ & \dots & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۱ & ۴ & ۱ & \dots & ۰ & ۰ & ۰ & ۰ \\ ۰ & ۰ & ۱ & ۴ & \dots & ۰ & ۰ & ۰ & ۰ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ ۰ & ۰ & ۰ & ۰ & \dots & ۴ & ۱ & ۰ & ۰ \\ ۰ & ۰ & ۰ & ۰ & \dots & ۱ & ۴ & ۱ & ۰ \\ ۰ & ۰ & ۰ & ۰ & \dots & ۰ & ۰ & ۰ & ۱ \end{bmatrix} \begin{bmatrix} ۰ \\ M_۲ \\ M_۳ \\ M_۴ \\ \vdots \\ M_{m-۳} \\ M_{m-۲} \\ M_{m-۱} \\ ۰ \end{bmatrix} = \frac{۶}{h^۲} \begin{bmatrix} f_۱ - ۲f_۲ + f_۳ \\ f_۲ - ۲f_۳ + f_۴ \\ f_۳ - ۲f_۴ + f_۵ \\ \vdots \\ f_{m-۴} - ۲f_{m-۳} + f_{m-۲} \\ f_{m-۳} - ۲f_{m-۲} + f_{m-۱} \\ f_{m-۲} - ۲f_{m-۱} + f_m \end{bmatrix}$$

[۲] Natural cubic spline

برای سادگی محاسبات می‌توان ستون اول و اخر را که صفر می‌باشند حذف نمود [۱۹] .
در شکل ۲.۲ مثالی از توابع اسپلاین مکعبی و مکعبی طبیعی رسم شده است. طبق شکل زیر تفاوت زیادی بین اسپلاین مکعبی و مکعبی طبیعی دیده نمی‌شود، اسپلاین‌های مکعبی در کران‌ها انعطاف‌پذیرتر هستند و اسپلاین‌های مکعبی طبیعی دارای محدودیت‌هایی در کران‌ها می‌باشند که باعث ایجاد برآوردهای پایدارتری در کران‌ها می‌شود و فواصل اطمینان متناظر آن هم محدودتر هستند.



شکل ۲.۲: نمایش اسپلاین‌های مکعبی و مکعبی طبیعی

۵.۲ B-اسپلاین‌ها

توابع پایه B-اسپلاین، چندجمله ای‌هایی از مرتبه $k = p - ۱$ هستند که در گره‌ها پیوستگی دارند و هر کدام از این توابع پایه، نواحی کوچکی از کل مشاهدات را پوشش می‌دهند. مکانیزم کار در این نوع اسپلاین، مشابه اسپلاین‌های معمولی است، تنها تفاوت در نحوه ساختن توابع پایه است. اگر $\boldsymbol{t} = (t_{(k-۱)}, \cdots, t_{(m+k)})$ یک دنباله غیر نزولی از گره‌ها باشد، توابع پایه B-اسپلاین به‌صورت زیر

$$B_j^{k-۱}, \quad j = -(k-۱), \cdots, m$$

تعریف می‌شود. همچنین تعداد توابع پایه B-اسپلاینی $p + m + ۱$ است.
اولین تابع پایه B-اسپلاینی از مرتبه صفر ($k = ۱$) به‌صورت زیر تعریف می‌شود:

$$B_j^{\circ}(x) = \begin{cases} ۱ & t_j \le x \le t_{j+۱} \\ ۰ & o.w. \end{cases}$$

دی بور (۱۹۷۸) برای ساخت توابع پایه مراتب بالاتر از توابع پایه مرتبه‌های پایین‌تر استفاده کرد. بر اساس یک رابطه بازگشتی، می‌توان B-اسپلاین‌هایی از هر درجه دلخواه را به‌صورت زیر محاسبه کرد:

(۱۰.۲)
$$B_j^p(x) = B_j^{k-۱}(x) = \frac{x - t_j}{t_{j+k-۱} - t_j} B_j^{k-۲}(x) + \frac{t_{j+k} - x}{t_{j+k} - t_{j+۱}} B_{j+۱}^{k-۲}(x), j = -(k-۱), \cdots, m$$

و با توجه به معادله (۱۰.۲)، B-اسپلاین‌های مرتبه یک (k=2) را به‌دست می‌آوریم که بر اساس توابع پایه درجات پایین‌تر تعریف شده است.

$$B_j^{۱}(x) = \frac{x - t_j}{t_{j+۱} - t_j} B_j^{\circ}(x) + \frac{t_{j+۲} - x}{t_{j+۲} - t_{j+۱}} B_{j+۱}^{\circ}(x) \qquad (۱۱.۲)$$

$$= \begin{cases} \frac{x-t_j}{t_{j+۱}-t_j} & t_j \le x < t_{j+۱} \\ \frac{t_{j+۲}-x}{t_{j+۲}-t_{j+۱}} & t_{j+۱} \le x < t_{j+۲} \end{cases}$$

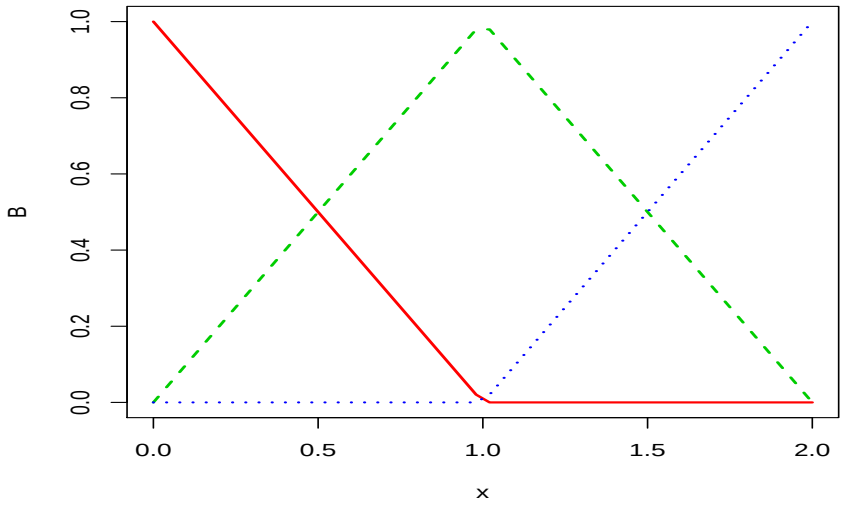
با جایگذاری گره‌ها در رابطه (۱۱.۲) تعدادی معادلات درجه یک به‌دست می‌آید.
به‌عنوان مثال در بازه‌ی $[۰, ۲]$ با در نظر گرفتن یک گره ۱ تعداد توابع پایه B-اسپلاینی p+m+1=3 برای j=-1,0,1 به‌صورت زیر به‌دست می‌آید:

$$B_{-۱}^{۱}(x) = \begin{cases} ۱ - x & ۰ \le x < ۱ \\ ۰ & o.w. \end{cases}$$

$$B_{۰}^{۱}(x) = \begin{cases} x & ۰ \le x < ۱ \\ ۲ - x & ۱ \le x < ۲ \\ ۰ & o.w. \end{cases}$$

$$B_{۱}^{۱}(x) = \begin{cases} x - ۱ & ۱ \le x < ۲ \\ ۰ & o.w \end{cases}$$

نمودار این توابع در شکل ۳.۲ رسم شده‌اند:



شکل ۳.۲: نمایش توابع پایه B-اسپلاینی مرتبه یک

توابع پایه مراتب دو و سه بین محققان محبوب‌تر واقع شدند، دلیل آن هم ویژگی‌ها و انعطاف‌پذیری خوبی است که دارا هستند. برای اطلاعات بیشتر در زمینه B- اسپلاین‌ها به دایرکس [۷] و دی بور [۶] مراجعه کنید.

برخی از ویژگی‌های مهم B-اسپلاین مرتبه $p = k - ۱$ به شرح زیر هستند [۸، ۱۴] :

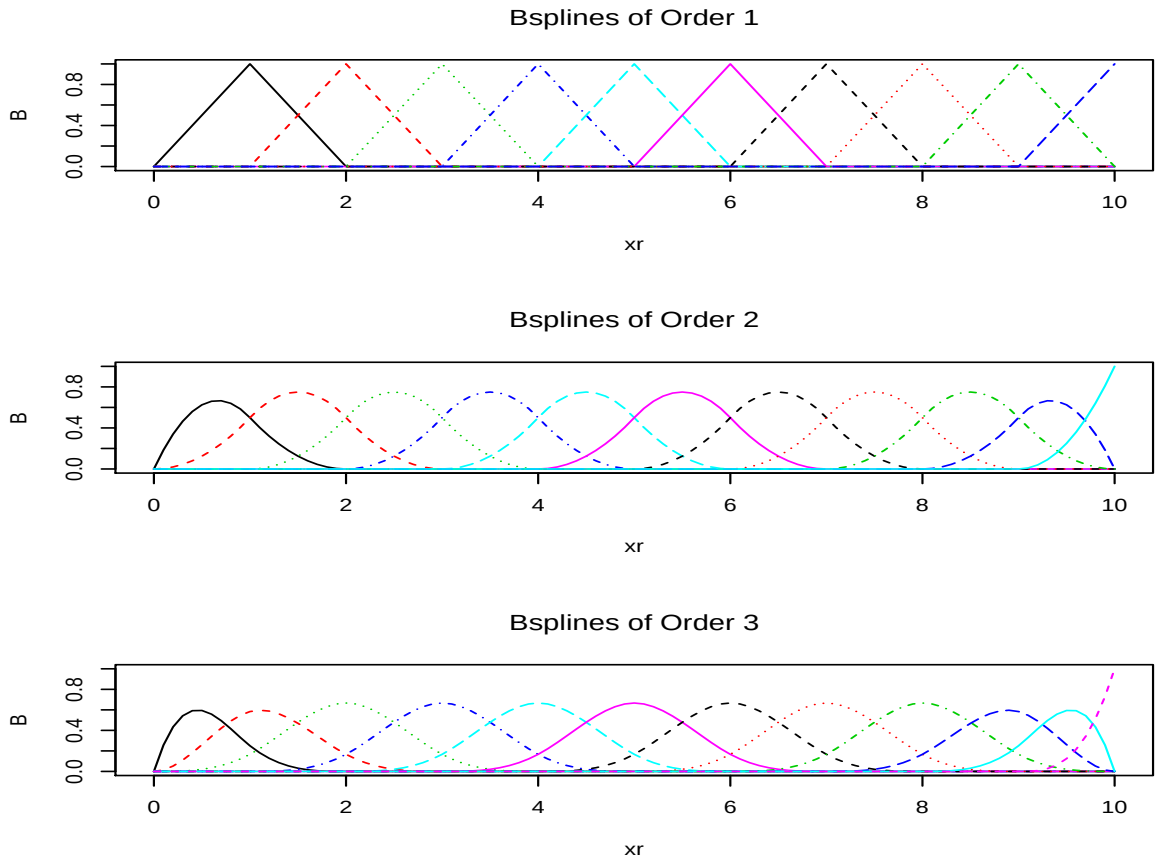
- هر تابع پایه با درجه p شامل p+1 تکه از درجه p است.
- تکه‌های چندجمله‌ای مجاور هم در گره‌ها پیوسته هستند.
- در گره‌ها از مرتبه ۱ تا مرتبه p-1 پیوسته و مشتق‌پذیر است.
- توابع پایه B-اسپلاین‌ها هم‌پوشانی دارند، به‌طور مثال اسپلاین‌ها با مرتبه یک با دو همسایگی هم‌چنین B-اسپلاین‌ها با مرتبه دو با چهار همسایگی هم‌پوشانی دارند و در قسمت انتهایی سمت راست و چپ اسپلاین‌ها دارای کمترین هم‌پوشانی است.
- توابع پایه روی تکیه‌گاه‌شان مثبت هستند، یعنی

$$B_j^{k-۱}(x) > ۰ \qquad x \in [t_j, t_{j+m}]$$

- جمع مقادیر این توابع برابر یک است.

$$\sum_{j=-(k-۱)}^{m} B_j^{k-۱}(x) = ۱$$

در شکل ۴.۲ اسپلاین‌های مرتبه یک، مرتبه دو و مرتبه سه نشان داده‌شده‌اند.



شکل ۴.۲: نمایش توابع پایه B-اسپلاین

۶.۲ هموارسازی اسپلاینی

یکی از کاربردهای اسپلاین‌ها برآورد روند موجود در داده‌ها است. برای این کار می‌توان از هموارسازهای اسپلاینی استفاده کرد. این نوع چندجمله‌ای‌ها در متون آماری به‌عنوان درونیاب یاد می‌شوند. اگر $(n+1)$، $x_0, x_1, \cdots, x_n$ نقطه دو به دو متمایز و f تابعی نامعلوم در این نقاط باشد چندجمله ای درونیاب f یک چندجمله‌ای حداکثر از درجهn است که از نقاط یا گره‌های فوق بگذرد به‌طوری که

$$s(x_k) = f(x_k) \quad k = 0, 1, \cdots, n$$

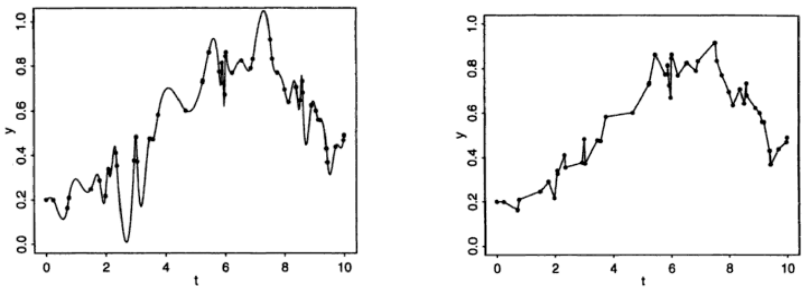
$$e(x_k) = f(x_k) - s(x_k) \qquad \text{تابع خطا}$$

ساده‌ترین درونیابی، درونیابی خطی است، نمودار این توابع خط شکسته است که تمام نقاط را به هم وصل می‌کند و در عمل مورد استفاده قرار نمی‌گیرد زیرا خطای آن بسیار زیاد است. اگر تابع درونیاب به‌صورت چندجمله‌ای باشد دارای مزیت‌های بسیاری است زیرا به‌راحتی می‌توان مشتق و انتگرال این توابع را محاسبه نمود. راه‌های زیادی برای به‌دست آوردن چندجمله‌ای‌های درونیاب وجود دارد مثل روش لاگرانژ، روش نیوتون،...یا درونیابی به‌صورت گویا که تابع در آن به‌صورت کسری است که در صورت و مخرجش چندجمله‌ای‌ها قرار دارند. درونیابی به روش‌های اخیر دارای دو مشکل بزرگ است:

- ممکن است نوسانات تابع درونیاب (به ویژه برای درونیابی چندجمله‌ای‌ها با درجه بالا) خیلی زیاد شود و خطای درونیابی به‌شدت افزایش یابد.
- اگر یک نقطه به نقاط اولیه اضافه کنیم فرم تابع کلا تغییر پیدا می‌کند.

استفاده از اسپلاین‌ها بهترین گزینه است زیرا این توابع اطلاعات را به‌طور موضعی درونیابی می‌کنند، به‌عبارتی بازه را به زیربازه‌هایی تقسیم کرده و تا جای ممکن درجه چندجمله‌ای درونیاب را کاهش می‌دهند، که این رهیافت تقریب قطعه به قطعه با چندجمله‌ای‌ها نامیده می‌شود. در هموارسازی اسپلاینی نوعی از اسپلاین ایجاد می‌شود که علاوه بر این که مجموع مربعات خطا را کمینه کند، باید تابعی هموار باشد یعنی دارای نوسانات سریع نباشد و بتواند از نزدیک داده‌ها عبور کند، نه فقط مشروط بر این‌که آن‌ها را درونیابی کند بلکه باید شرط مشتق‌پذیری تا مرحله معینی را نیز داشته باشد. بنابراین به برآوردی از یک منحنی دست خواهیم یافت که به‌طور تقریبا هموار از بین توده داده‌ها می‌گذرد که به این عمل هموارسازی اسپلاینی می‌گویند [۱۳، ۱۱].

در شکل ۵.۲ قاب سمت راست، داده‌ها با خط مستقیم به هم متصل شده‌اند که کاملا شکسته و بریده است و منحنی کاملا ناهمواری را ایجاد نموده است و در قاب سمت چپ، منحنی با مشتق دوم پیوسته از تمام نقاط عبور کرده و منحنی هموارتری را ایجاد نموده است. شکل

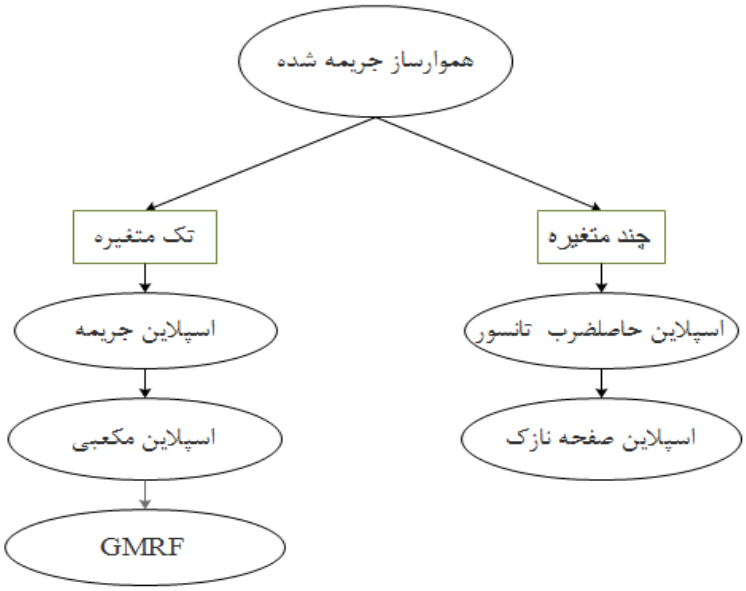


(آ) داده‌ها با خط مستقیم به هم متصل شده‌اند

(ب) داده‌ها با منحنی با مشتق دوم پیوسته به هم متصل شده‌اند

شکل ۵.۲: هموارسازی

۶.۲ روش‌های هموارسازی تک‌متغیره و چندمتغیره را نشان می‌دهد که در بخش‌های بعد این روش‌ها معرفی می‌شوند.



شکل ۶.۲: فلوچارت روش‌های هموارساز

۷.۲ اسپلاین‌های جریمه‌ای (p-اسپلاین)

فرض کنید تعداد n زوج داده داشته باشیم و بخواهیم روند برازشی این مشاهدات را با استفاده از B-اسپلاین‌ها برآورد کنیم. تابع B-اسپلاینی بهتر است که مقدار کمیت زیر را کمینه کند:

$$S = \sum_{i=۱}^{n} \left\{ y_i - \sum_{j=-(k-۱)}^{m} \alpha_j B_j^k(x) \right\}^۲ .$$

اگر تعداد گره‌ها را نسبتا زیاد در نظر بگیریم، منحنی برازشی بیشترین تغییرات تعدیل شده توسط داده‌ها را نشان خواهد داد، به‌عبارتی از تعداد مشاهدات بیشتری عبور خواهد کرد و در نتیجه مقدار مجموع توان‌های دوم خطا (S) کمتر خواهد شد. بنابراین تابعی که کمینه می‌شود با تعداد گره‌ها رابطه معکوس دارد. اما اگر تعداد گره‌ها افزایش یابد، منجر به زیربازه‌های بیشتری می‌شود و در نتیجه توابع پایه بیشتری خواهیم داشت و پیرو آن باید پارامترهای بیشتری را برآورد کنیم. از آن‌جایی که تعداد گره‌ها نامعلوم است، معمولا تعداد گره‌ها را برابر با تعداد مشاهدات در نظر می‌گیرند که در این صورت ممکن است با بیش‌برازشی مواجه شویم. به‌عبارتی نوسانات منحنی زیاد می‌شود و منحنی حالت هموار بودن خود را از دست می‌دهد. بنابراین باید بین برازش مناسب و کاهش پیچیدگی مدل (بر حسب تعداد گره‌های کافی) یک تعادل ایجاد شود. روبرت (۲۰۰۲) برای انتخاب تعداد و موقعیت گره‌ها روشی معرفی کرد که در آن تعداد گره‌ها برابر است با $min(\frac{N}{۴}, ۳۵)$. برای کنترل و جلوگیری از بیش‌برازشی اسیلوان (۱۹۸۶و۱۹۸۸) پیشنهاد داد که جریمه‌ای برای منحنی برازشی اضافه شود. او انتگرال توان دوم مشتق دوم منحنی برازش‌شده را به‌عنوان جریمه هموارسازی به تابع توان دوم خطای برازش به‌صورت زیر افزود:

$$S = \sum_{i=۱}^{n} \left\{ y_i - \sum_{j=-(k-۱)}^{m} \alpha_j B_j^k(x_i) \right\}^۲ + \lambda \int_{x_{\min}}^{x_{\max}} \left\{ \sum_{j=-(k-۱)}^{m} \alpha_j B_j''(x) \right\}^۲ \mathrm{d}x$$

انتگرال مشتق دوم به‌عنوان جریمه هموارساز رایج شد. هیچ تاکید خاصی برای مشتق دوم وجود ندارد در حقیقت می‌توان از مشتقات مراتب بالاتر یا پایین‌تر استفاده نمود در حالی که مشتق اول منجر به معادلات ساده و برازش خطی می‌شود و مشتق مراتب بالاتر منجر به محاسبات طولانی و پیچیده خواهد شد. مشتق تابع B-اسپلاین در پیوست آ.۲ شرح داده‌شده است. اسپلاین‌هایی که در برازش آن‌ها از جریمه هموارساز استفاده می‌شود به‌عنوان اسپلاین‌های جریمه‌ای معرفی شدند که اجزای اصلی آن B-اسپلاین‌ها و جریمه ناهمواری است. در واقع اسپلاین‌های جریمه‌ای ترکیبی از یک پایه‌ی B-اسپلاینی و یک جمله جریمه روی اختلاف ضرایب جملات مجموعی از توابع اسپلاین است. جمله جریمه بخش اساسی و نقطه قوت در اسپلاین‌های جریمه‌ای است.

ایلرز و مارکس (۱۹۹۶) پیشنهاد دادند که جریمه ناهمواری به‌صورت تفاضلات متناهی ضرایب مجاور هم قرار داده شوند. یعنی

$$S=\sum_{i=۱}^{n}\left\{y_i-\sum_{j=-(k-۱)}^{m}\alpha_j B_j^k(x)\right\}^۲+\lambda\sum_{j=-(k-۱)}^{m}(\Delta^k\alpha_j)^۲ \qquad (۱۲.۲)$$

این روش پیشنهادی موجب کاهش حجم محاسبات و همچنین نوشتن معادلات به شکل ماتریسی می‌شود.

۱.۷.۲ صورت ماتریسی و برآورد ضرایب اسپلاین جریمه‌ای با استفاده از کمترین توان‌های دوم

فرم ماتریسی جمله جریمه در معادله (۱۲.۲) به‌صورت زیر است:

$$\sum_{j=-(k-۱)}^{m}(\Delta^k\alpha_j)^۲=\|(\Delta^۲\alpha_{-(k-۱)}\dots\Delta^۲\alpha_m)^T\|^۲=\boldsymbol{\alpha}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\alpha}$$

D ماتریس جریمه k-امین مرتبه تفاضلات ضرایب مجاور است.
تفاضل مرتبه اول، دوم و به همین ترتیب kام به‌صورت زیر است:

$$\begin{aligned}
&\Delta\alpha_j=\alpha_j-\alpha_{j-۱}\\
&\Delta^۲\alpha_j=\Delta(\Delta\alpha_j)=\alpha_j-۲\alpha_{j-۱}+\alpha_{j-۲}\\
&\vdots\\
&\Delta^k\alpha_j=\Delta(\Delta^{k-۱}\alpha_j)
\end{aligned}$$

تفاضل مرتبه دوم به ازای هر j را می‌توان به‌صورت زیر نوشت:

$$\begin{aligned}
j=۲&\Rightarrow\alpha_۲-۲\alpha_۱+\alpha_۰=\Delta^۲\alpha_۲\\
j=۳&\Rightarrow\alpha_۳-۲\alpha_۲+\alpha_۱=\Delta^۲\alpha_۳\\
&\vdots
\end{aligned}$$

در نتیجه ماتریس $D_{(m+k-۲)(m+k)}$ حاصل می‌شود:

$$\boldsymbol{D}=\begin{pmatrix}
۱ & -۲ & ۱ & & & \\
 & ۱ & -۲ & ۱ & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & ۱ & -۲ & ۱
\end{pmatrix}$$

$\boldsymbol{D}^T\boldsymbol{D}$ ماتریسی با بعد $(m+k)\times(m+k)$ است را محاسبه می‌کنیم:

$$\boldsymbol{D}^T\boldsymbol{D} = \begin{pmatrix} ۱ & -۲ & ۱ & & & & & & \\ - & ۵ & -۴ & ۱ & & & & & \\ ۱ & -۴ & ۶ & -۴ & \cdot & & & & \\ & ۱ & -۴ & ۶ & \cdot & \cdot & & & \\ & & ۱ & -۴ & \cdot & \cdot & & & \\ & & & ۱ & \cdot & \cdot & -۴ & ۱ & \\ & & & & \cdot & \cdot & ۶ & -۴ & ۱ \\ & & & & & \cdot & -۴ & ۵ & -۲ \\ & & & & & & ۱ & -۲ & ۱ \end{pmatrix}$$

درنتیجه فرم ماتریسی رابطه‌ی (۱۲.۲) را می‌توان به‌صورت زیر نوشت:

$$\begin{aligned} S &= (y-B\boldsymbol{\alpha})^T(y-B\boldsymbol{\alpha}) + \lambda(\boldsymbol{\alpha}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\alpha}) \\ &= y^Ty - ۲y^TB\boldsymbol{\alpha} + \boldsymbol{\alpha}^TB^TB\alpha + \lambda(\boldsymbol{\alpha}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\alpha}) \end{aligned} \qquad (۱۳.۲)$$

برای مینمم کردن عبارت بالا نسبت به α مشتق می‌گیریم در نتیجه داریم:

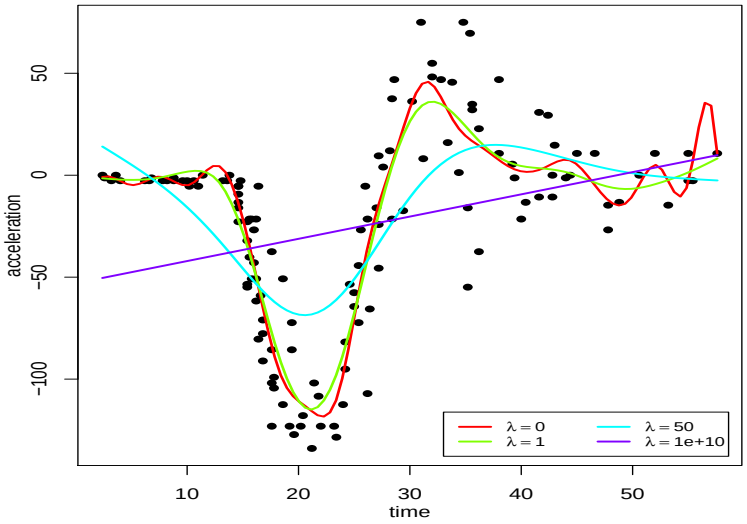
$$\begin{aligned} &\frac{\partial S}{\partial \alpha} = -۲\beta^Ty + ۲\beta^T\beta\hat{\boldsymbol{\alpha}} + ۲\lambda\boldsymbol{D}^T\boldsymbol{D}\hat{\boldsymbol{\alpha}} = ۰ \\ &\beta^Ty = \hat{\boldsymbol{\alpha}}(\beta^T\beta + \lambda\boldsymbol{D}^T\boldsymbol{D}) \end{aligned}$$

در نتیجه ضرایب برآورد شده به‌صورت

$$\hat{\boldsymbol{\alpha}} = (\beta^Ty)(\beta^T\beta + \lambda\boldsymbol{D}^T\boldsymbol{D})^{-۱} \qquad (۱۴.۲)$$

است.

پارامتر λ نیز پارامتر هموارساز نامیده می‌شود که مقداری مثبت است و برای کنترل پیوستگی منحنی مورد استفاده قرار می‌گیرد. اگر مقدار $\lambda = ۰$ باشد، منحنی برازش‌شده می‌تواند تابعی باشد که داده‌ها را به‌طور دقیق درونیابی می‌کند به‌عبارتی از روی تمام نقاط عبور می‌کند و مجموع مربعات خطا سهم اصلی را در S دارد، ، و اگر مقدار $\lambda = \infty$ در نظر گرفته شود، جمله جریمه به صفر میل می‌کند و منحنی برازشی یک خط رگرسیون خواهد بود. در شکل ۷.۲ کاملا مشهود می‌باشد.



شکل ۷.۲: هموارسازی تابع رگرسیونی برای مقادیر مختلف λ

۲.۷.۲ خواص اسپلاین جریمه‌ای

اسپلاین‌های جریمه‌ای دارای خواص خوبی هستند که تا حدودی از B-اسپلاین‌ها به ارث برده‌اند، برخی از خواص آن‌ها عبارتند از[۸] :

- اسپلاین‌های جریمه‌ای چندجمله‌ای‌هایی را به‌طور دقیق به داده‌ها برازش می‌دهند. فرض کنید (x_i, y_i) یک مجموعه داده باشد اگر y_i تابع چندجمله‌ای از x از مرتبه k باشد آنگاه B-اسپلاین‌ها از درجه k یا بالاتر به‌طور دقیق به داده‌ها برازش داده می‌شوند[۵]. اگر مرتبه جریمه $(k+1)$ یا دارای مرتبه‌های بالاتری باشد، صرف‌نظر از مقادیر مختلف λ (پارامتر هموارسازی) ، برای اسپلاین‌های جریمه‌ای برقرار می‌باشد. برای اثبات این مطلب به ایلرز و مارکس ۱۹۹۶ مراجعه شود.

- اسپلاین‌های جریمه‌ای گشتاور داده‌ها را حفظ می‌کنند. برای یک مدل خطی اسپلاین‌های تاوانیده از مرتبه $k+1$ و مرتبه جریمه $k+1$ یا بالاتر رابطه زیر برقرار است:

$$\sum_{i=1}^{m} x^k y_i = \sum_{i=1}^{m} x^k \hat{y}_i$$

برای همه مقادیر λ (پارامتر هموارسازی)، $\hat{y}_i = \sum_{i=1}^{n} b_{ij}\hat{\alpha}_j$ مقادیر برازش داده‌شده هستند و برای مدل‌های خطی تعمیم‌یافته به‌صورت زیر به‌دست می‌آیند:

$$\sum_{i=1}^{m} x^k y_i = \sum_{i=1}^{m} x^k \hat{\mu}_i.$$

این ویژگی خصوصا در هموارسازی تابع چگالی، مفید است زیرا میانگین و واریانس تابع چگالی برآورد شده، با میانگین و واریانس داده‌ها برابر است.

۸.۲ اسپلاین حاصل‌ضرب تانسور

اسپلاین حاصل‌ضرب تانسوری[۳] سالیان متمادی برای تقریب تک‌متغیره و روش‌های درونیابی دومتغیره مورد استفاده قرار گرفته‌اند و تعمیمی ساده و مستقیم از اسپلاین‌های n بعدی را فراهم می‌کنند. تقریب چندمتغیری از توابع توسط حاصل‌ضرب تانسوری دارای جهت‌گیری قوی در امتداد خطوط موازی با جهت محور می‌باشد. رویه‌های حاصل‌ضرب تانسوری بهشدت به توابع اسپلاینی وابسته می‌باشند. در این روش می‌توان برای ساختار توابع هموارگر از هر تعداد متغیر استفاده نمود. ساده‌ترین روش برای ساخت یک تابع هموارگر استفاده از سه متغیر x، z و ν است که هر یک از متغیرها دارای تابع هموارگر fx، fz و $f\nu$ با توابع پایه می‌باشند. که به‌صورت زیر قابل نمایش است:

$$f_x(X) = \sum_{i=1}^{I} \alpha_i a_i(x) \qquad f_z(Z) = \sum_{l=1}^{L} \delta_l d_l(z) \qquad f_\nu(\nu) = \sum_{k=1}^{K} \beta_k b_k(\nu)$$

که در آن α_i، δ_l و β_k پارامترها و $a_i(x)$، $d_l(z)$ و$b_k(\nu)$ توابع پایه معلوم می‌باشند و I، L و K می‌توانند بازه‌ای نسبتا بزرگ اختیار نمایند. تابع هموارگر $f_x(X)$ می‌تواند به یک تابع هموارگر$f(x,z)$ تبدیل شود، برای این کار لازم است که $f(x,z)$ به‌طور مساوی با z تغییر کند در نتیجه باید پارامتر α_i به‌طور مساوی با z تغییر کند و با به‌کار بردن توابع پایه تابع هموارگر z داریم:

$$\alpha_i(z) = \sum_{l=1}^{L} \delta_{il} d_l(z)$$

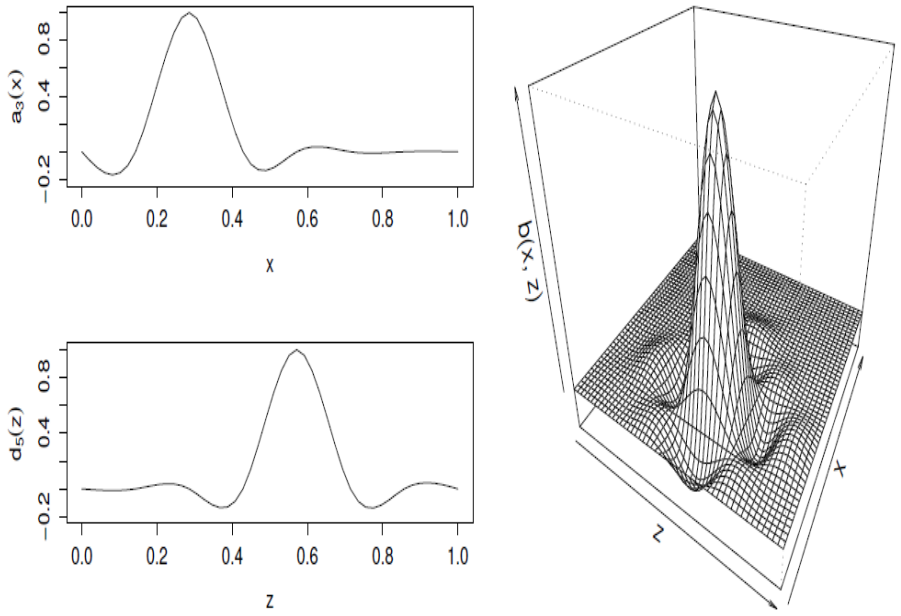
بنابراین ساختار پایه‌های حاصل‌ضرب تانسوری برای یک سطح به‌صورت زیر حاصل می‌شود:

$$f_{xz}(x,z) = \sum_{i=1}^{I} \alpha_i(z) a_i(x) = \sum_{i=1}^{I} \sum_{l=1}^{L} \delta_{il} d_l(z) a_i(x) \qquad (۱۵.۲)$$

شکل ۸.۲ ساختار معادله (۱۵.۲) را نشان می‌دهد. با ادامه این روند می‌توان تابع هموارگر x، z و ν را با تغییر دادن f_{xz} به‌طور یکسان با ν به‌دست آورد. پس با همان استدلال قبل داریم [۳۴، ۳۶]:

$$f_{xz\nu}(x,z,\nu) = \sum_{i=1}^{I} \sum_{l=1}^{L} \sum_{k=1}^{K} \beta_{ilk} b_k(\nu) d_l(z) a_i(x)$$

[۳] Tensor product splines



شکل ۸.۲: حاصل‌ضرب دو تابع پایه‌ی حاشیه‌ای برای توابع هموار fx وfz

این ساختار می‌تواند برای تعداد متغیرهای بیشتری تعمیم داده شود.
می‌توان میزان انعطاف‌پذیری این توابع را اندازه‌گیری نمود. فرض کنید که هر هموارگر حاشیه‌ای دارای تابعی مرتبط است که میزان انعطاف‌پذیری را اندازه‌گیری می‌کند و می‌تواند به فرم درجه دوم از پارامترهای حاشیه‌ای بیان شود که به‌صورت زیر می‌باشد:

$$J_x(f_x) = \int f''(x)^2 \mathrm{d}x = \boldsymbol{\alpha}^T \boldsymbol{S}_x \boldsymbol{\alpha}$$
$$J_z(f_z) = \int f''(z)^2 \mathrm{d}z = \boldsymbol{\delta}^T \boldsymbol{S}_z \boldsymbol{\delta}$$
$$J_\nu(f_\nu) = \int f''(\nu)^2 \mathrm{d}\nu = \boldsymbol{\beta}^T \boldsymbol{S}_\nu \boldsymbol{\beta}$$

$\boldsymbol{S}$ ماتریسی از ضرایب معلوم و $\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\alpha}$ بردار ضرایب از هموارگر حاشیه‌ای است. اکنون در نظر بگیرید $f_{x|z}(x)$ که به $f_{xz}(x,z)$ به‌عنوان تابعی از x با z ثابت تبدیل شده است. در نتیجه $J_x(f_{x|z})$ میزان انعطاف‌پذیری $f_{x|z}$ را اندازه‌گیری می‌کند و $\int J_x(f_{x|z})\mathrm{d}z$ با میانگین انعطاف‌پذیری در جهت x متناسب است. بنابراین یک جریمه مناسب به‌صورت معادله زیر می‌باشد:

$$J_x(f_{xz}) = \lambda_x \int_z J_x(f_{x|z})\mathrm{d}z + \lambda_z \int_x J_z(f_{z|x})\mathrm{d}x$$

به‌طور مشابه می‌توان برای توابع $f_{z|x\nu}(z)$ و $f_{\nu|xz}(\nu)$ تعریف نمود. بنابراین یک روش برای اندازه‌گیری انعطاف‌پذیری تابع $f_{xz\nu}$ به‌صورت معادله زیر می‌باشد.

$$J(f_{xz\nu}) = \lambda_x \int_{z\nu} J_x(f_{x|z\nu})\mathrm{d}z\mathrm{d}\nu + \lambda_z \int_{x\nu} J_z(f_{z|x\nu})\mathrm{d}x\mathrm{d}\nu + \lambda_\nu \int_{xz} J_\nu(f_{\nu|xz})\mathrm{d}x\mathrm{d}z$$

که در آن λ پارامتر هموارسازی است که انعطاف‌پذیری تابع را در جهت‌های مختلف کنترل می‌کند. به‌عنوان مثال با درنظر گرفتن تابع جریمه در جهت x تابع $f_{x|z\nu}(x)$ به‌صورت زیر بیان می‌شود:

$$f_{x|z\nu}(x) = \sum_{i=1}^{I} \alpha_i(z,\nu) a_i(x).$$

می‌توان ماتریس ضرایب $\boldsymbol{M}_{z,\nu}$ را تعریف نمود به‌طوری که $\boldsymbol{\alpha}(z,\nu) = \boldsymbol{M}_{z\nu}\boldsymbol{\beta}$ که $\boldsymbol{\beta}$ بردار پارامتر برای $f(x,z,\nu)$ است. از این رو

$$J_x(f_{x|z,\nu}) = \boldsymbol{\alpha}(z,\nu)^T \boldsymbol{S}_x \boldsymbol{\alpha}(z,\nu) = \boldsymbol{\beta}^T \boldsymbol{M}_{z\nu}^T \boldsymbol{S}_x \boldsymbol{M}_{z\nu}\boldsymbol{\beta}.$$

بنابراین

$$\int_{z,\nu} J_x(f_{x|z,\nu})\mathrm{d}z\mathrm{d}\nu = \boldsymbol{\beta}^T \int_{z,\nu} \boldsymbol{M}_{z\nu}^T \boldsymbol{S}_x \boldsymbol{M}_{z\nu}\mathrm{d}z\mathrm{d}\nu\boldsymbol{\beta}.$$

انتگرال فوق را می‌توان محاسبه نمود اما بسیار زمانبر است. به خصوص زمانی که تعداد متغیرها افزایش یابد. و این رویکرد برای تمام اجزای جریمه اعمال می‌شود. با این حال، با به‌کارگیری بازپارامتریدن ساده می‌توان تقریبی برای شرایط جریمه ارائه داد که به‌خوبی عمل می‌کند و از محاسبه انتگرال عددی جلوگیری می‌کند. برای نشان دادن عملکرد این رویکرد تابع حاشیه‌ای f_x را در نظر بگیرید و قرار دهید $\{x_i^* : i = 1,\cdots,I\}$، مجموعه‌ای از مقادیر x که به‌طور یکنواخت در دامنه مقادیر مشاهده شده x پخش شده‌اند. با بازنویسی پارامترهای تابع حاشیه‌ای f_x داریم:

$$\alpha_i'' = f_x(x_i^*).$$

تحت شرایط بازپارامتری کردن $\boldsymbol{\alpha}' = \boldsymbol{\Gamma}\boldsymbol{\alpha}$ و $\boldsymbol{\Gamma}_{ij} = a_i(x_j^*)$. بنابراین ماتریس مدل حاشیه‌ای به‌صورت $\boldsymbol{X}_x' = \boldsymbol{X}_x\boldsymbol{L}^{-1}$ و ماتریس ضرایب جریمه به $\boldsymbol{S}_x' = \boldsymbol{l}^{-T}\boldsymbol{S}_x\boldsymbol{l}^{-1}$ تبدیل می‌شود و داریم:

$$\int_{z\nu} J_x(f_{x|z\nu})\mathrm{d}z\mathrm{d}\nu \approx h\sum_{lk} J_x(f_{x|z_l^*\nu_k^*}),$$

که h مقداری ثابت از نسبت فاصله‌ی بین z_l^* و ν_k^* است. نشان دادن این تقریب در مجموع بالا سر راست است و داریم [۳۶، ۳۴]:

$$J_x^*(f_{xz\nu}) = \boldsymbol{\beta}^T\tilde{\boldsymbol{S}}_x\boldsymbol{\beta} \qquad \tilde{\boldsymbol{S}}_x = \boldsymbol{S}_x' \otimes \boldsymbol{I}_L \otimes \boldsymbol{I}_K.$$

با تعاریف مشابه برای دیگر اجزای جریمه معادلات زیر به‌دست آمده است:

$$J_z^*(f_{xz\nu}) = \boldsymbol{\beta}^T\tilde{\boldsymbol{S}}_z\boldsymbol{\beta} \qquad \tilde{\boldsymbol{S}}_z = \boldsymbol{I}_I \otimes \boldsymbol{S}_z' \otimes \boldsymbol{I}_K$$

و

$$J_\nu^*(f_{xz\nu}) = \boldsymbol{\beta}^T\tilde{\boldsymbol{S}}_\nu\boldsymbol{\beta} \qquad \tilde{\boldsymbol{S}}_\nu = \boldsymbol{I}_I \otimes \boldsymbol{I}_L \otimes \boldsymbol{S}_\nu'.$$

نماد $\otimes$ نمایانگر ضرب کرونکر است که در پیوست آ.۱ شرح داده‌شده است. بنابراین

$$J(f_{xz\nu}) \approx J^*(f_{xz\nu}) = \lambda_x J^*(f_{xz\nu}) + \lambda_z J^*(f_{xz\nu}) + \lambda_\nu J^*(f_{xz\nu})$$

۹.۲ اسپلاین صفحه نازک

اسپلاین صفحه نازک[۴] یک نوع از هموارسازی اسپلاینی است که برای تجسم روابط پیچیده بین متغیرهای پاسخ و پیش‌گوی پیوسته به‌کار می‌رود و به دلیل ظاهر چند بعدی برای بررسی اثر ترکیبی دو متغیر پیش‌گوی پیوسته در یک نتیجه واحد قابل استفاده است. اسپلاین صفحه نازک تکنیکی مبتنی بر اسپلاین برای درونیابی و هموارسازی مجموعه‌ای از نقاط کنترل است. در این روش یک سطح از میان نقاط کنترل عبور می‌کند. نام اسپلاین صفحه نازک اشاره‌ای به حالت فیزیکی صفحه نازک فلزی که قابل خم شدن است را دارد. داچون و ماینگت پایه و اساس اسپلاین صفحه نازک را طراحی کردند که سطحی قابل انعطاف است [۱۰، ۳۵].
اسپلاین صفحه نازک برای اولین بار در سال ۱۹۷۰ در علوم کامپیوتر در زمینه محاسبات هندسی مورد استفاده قرار گرفتند. این رویکرد مدل‌سازی امروزه در بسیاری از زمینه‌ها از جمله مهندسی (طراحی ساختاری)، اکولوژی (رشد جمعیت) و شناخت الگو (شناسایی اثر انگشت) کاربرد دارد. استفاده رایج آن در تحقیقات علمی در زمینه سلامت، شامل توسعه و تجزیه تحلیل تکنیک‌های تصویربرداری پزشکی است [۲۰].
با در نظر گرفتن n نقطه متمایز و متغیرهای ورودی $\{(x_i, y_i), i = ۱, \cdots, n\}$ و متغیر پاسخ z_i مطلوب‌ترین برازش، تعمیمی از اسپلاین‌هایی است که انحنای سطح برازشی را کمینه کند. در این حالت سطح $f(x, y)$ طوری تعیین شود که فاصله آن با نقاط کنترل کمینه شود، به‌عبارتی رابطه‌ی زیر مینمم شود.

$$E_{Tps} = \sum_{i=۱}^{n} (z_i - f(x_i, y_i))^۲ + \lambda j[f(x, y)] \qquad (۱۶.۲)$$

که λ پارامتر هموارساز است هنگامی که $\lambda = ۰$ باشد فقط داده‌ها بدون هموارسازی درونیابی می‌شوند و وقتی $\lambda \to \infty$ مساله یافتن سطحی است که فاصله آن از نقاط کنترل کمینه گردد. تابع جریمه را می‌توان با معیارهای مختلفی از جمله معیار اعتبارسنجی متقابل انتخاب نمود که فرم کلی آن به‌صورت زیر است:

$$j[f(x, y)] = \iint_{\mathbb{R}^۲} \left[\left(\frac{\partial^۲ f}{\partial x^۲}\right)^۲ + ۲\left(\frac{\partial^۲ f}{\partial x \partial y}\right)^۲ + \left(\frac{\partial^۲ f}{\partial y^۲}\right)^۲ \right] \mathrm{d}x\mathrm{d}y \qquad (۱۷.۲)$$

$f(x, y)$ یک نوع از توابع پایه شعاعی[۵] است. توابع پایه شعاعی برای درونیابی استفاده می‌شوند،

[۴] Thin plate spline

به‌صورت معادله‌ی زیر می‌باشد.

$$f(x,y) = a_0 + a_1 x + a_2 y + \sum_{i=1}^{n} W_i \varphi(\|x - x_i\|). \qquad (۱۸.۲)$$

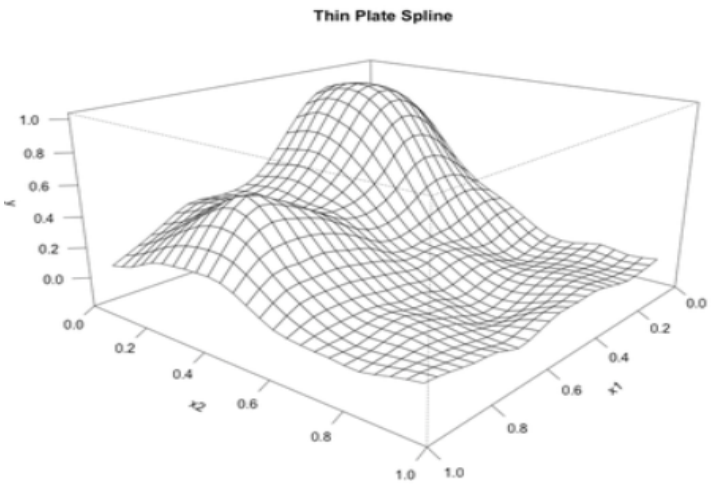
سه جمله اول مربوط به قسمت خطی است. صفحه‌ای تخت که به بهترین نحو بر نقاط کنترل منطبق باشد، تعریف می‌شود و عبارت آخر هم مربوط به نیروی خمش ناشی از نقاط کنترل است. $\|\cdot\|$ بیانگر فاصله اقلیدسی و w_i مجموعه‌ای از ضرایب نگاشت که هرکدام متعلق به یکی از مراکز داده‌ها است، $x_i, i = 1, \cdots, n$ مجموعه نقاط کنترل و φ تابع شعاعی کرنل که به‌صورت زیر تعریف می‌شود:

$$\varphi(r) = r^2 log(r) \qquad (۱۹.۲)$$

r فاصله اقلیدسی در فضای دو بعدی است و $\varphi(r)$ وابسته به فاصله اقلیدسی از مبدا x_i است.
برازش اسپلاین صفحه نازک با مدل جمعی تعمیم‌یافته (GAM) به‌صورت زیر است:

$$g(E(Y)) = \beta_0 + f(x) + \epsilon.$$

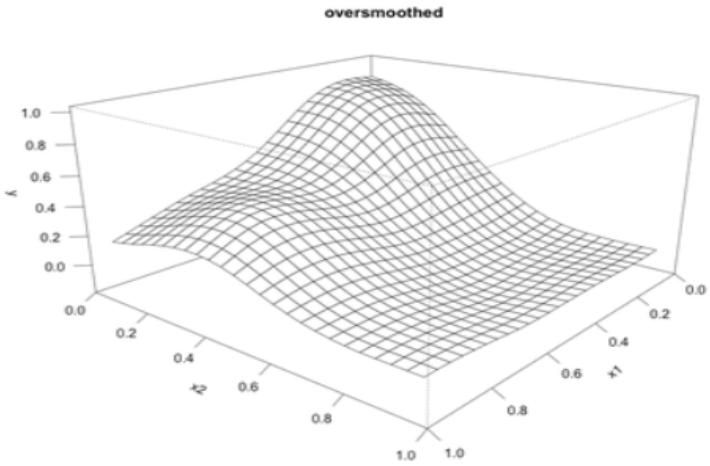
β_0 مقداری ثابت است، $f(x)$ تابعی انعطاف پذیر از x و ϵ عبارت خطا است و تابعی هموار بر اساس روش حداقل مربعات جریمه‌ای را فراهم می‌کند که با افزایش ϵ تابع هموارتری نسبت به توابع اسپلاین به‌دست می‌آید [۲۰، ۳۵] . شکل ۹.۲ نمایش اسپلاین صفحه نازک دو بعدی است که هر متغیر بر روی یک محور ترسیم شده است و یک سطح دو بعدی در سه بعد قابل نمایش است.



شکل ۹.۲: نمایش اسپلاین صفحه نازک

[۵] Radial basis function

عبارت خطا را می‌توان به عنوان کشش معرفی کرد یا میزان فشاری که برای خمیدگی یک ورق نازک فلزی لازم است. با کشش بالاتر مقاومت صفحات نازک برای خمیدگی بیشتر می‌شود و در نتیجه تاثیر متغیرهای پیش‌گو بر متغیر پاسخ افزایش یافته و در پی آن تابع هموارتری ظاهر خواهد شد.
در شکل ۱۰.۲ با افزایش ϵ تابع هموارتری نسبت به شکل ۹.۲ حاصل شده است.

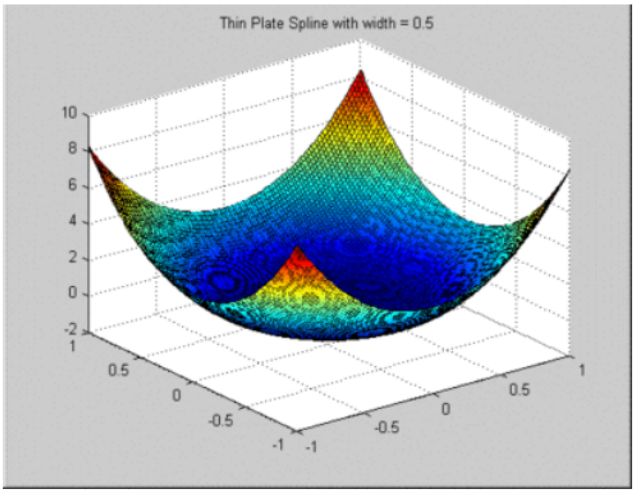


شکل ۱۰.۲: نمایش اسپلاین صفحه نازک هموارتر

۱.۹.۲ مزایای اسپلاین صفحه نازک

مزایای اسپلاین صفحه نازک مانند سایر اسپلاین‌های هموارساز است که در آن مدل‌های جمعی تعمیم‌یافته (GAM) نیازی به هیچ گونه اطلاعات پیشینی از فرم عملکرد داده‌ها یا روابط بین آن‌ها ندارند. تعیین تعداد گره و موقعیت آن‌ها چالشی در اسپلاین‌ها است که به‌طور موثر به‌عنوان بخشی از توابع هموارساز اسپلاین صفحه نازک است. ماهیت اسپلاین صفحه نازک آن‌ها را به یک ابزار قدرتمند و جذاب برای تجسم روابط پیچیده بین متغیرهای پیش‌گو و پاسخ تبدیل کرده است و با افزودن نقشه‌های گرمایی[۶] به اسپلاین صفحه نازک که بیشتر بر انحنای رویه تاکید دارند آن را به یک گرافیک بصری جذاب تبدیل می‌کنند، مانند شکل ۱۱.۲ [۱۱، ۱۶].

[۶] Heat maps



شکل ۱۱.۲: نمایش اسپلاین صفحه نازک

فصل ۳

برخی از مباحث نظری اسپلاین‌ها

در این فصل نحوه‌ی انتخاب پارامتر هموارسازی، روش‌های انتخاب تعداد و موقعیت گره‌ها و انتخاب بهترین مدل بر اساس معیارهای GCV, CV,AIC ,... شرح داده‌شده است.

۱.۳ نحوه‌ی انتخاب پارامتر هموارسازی

پارامتر هموارسازی برای کنترل همواری منحنی استفاده می‌شود. هیستی و تیبشیرانی (۱۹۹۰) نماد λ را برای نشان دادن پارامتر همواری به‌کار بردند. در هر حال λ باید به‌طور محافظه‌کارانه انتخاب شود تا مسئله بیش‌همواری یا کم‌همواری پیش نیاید و یک تابع هموار مناسب را نتیجه دهد. براي انتخاب بهینه‌ی پارامتر همواری روش‌های مختلفی وجود دارد. یک روش، انتخاب یک مقدار دلخواه برای پارامتر همواری است که با تغییر این مقدار و با توجه به نوع موضوع مورد بررسی یک مقدار برای آن انتخاب می‌گردد به‌طوری که برآورد حاصل، بیشترین تطبیق را با داده‌ها داشته باشد.

از دیگر روش‌ها براي انتخاب بهینه پارامتر هموارسازی استفاده از معیار اطلاع آکائیک[۱] ، روش اعتبارسنجی متقابل[۲] و اعتبارسنجی متقابل تعمیم‌یافته[۳] که در ذیل به تفصیل آن‌ها می‌پردازیم.

[۱] Akaike information criterion

[۲] Cross validation

[۳] Generalized cross-validation

۱.۱.۳ معیار اطلاع آکائیک

ایده‌ی اصلی معیار اطلاع آکائیک به این صورت است که لگاریتم تابع درستنمایی مدل برازشی را برای تعداد پارامترهای موثر بدست می‌آورد. در واقع این معیار تعادلی بین پیچیدگی مدل و دقت آن را برقرار می‌کند. زماني‌که براي مقايسه مدل‌های برازشی مختلف از این معیار استفاده می‌شود نشان می‌دهد که استفاده از یک مدل آماری به چه میزان باعث از دست رفتن اطلاعات می‌شود. بنابراین مدلي که کمترین مقدار آکائیک را داشته باشد به‌عنوان بهترین مدل انتخاب می‌شود. فرم کلي معیار آکائیک به‌صورت زیر است:

$$AIC = log\left(\frac{۱}{n}\sum_{i=۱}^{n}(y_i - \hat{y}_i)^۲\right) + \frac{۲ \times tr(H)}{n}$$

H ماتریس هت نامیده می‌شود.
ایلرز و مارکس (۱۹۹۶) براي به‌دست آوردن معیار آکائیک از انحراف به‌جای لگاریتم درستنمایی استفاده نمودند و معیار اطلاع آکائیک را به‌صورت زیر تعریف کردند:

$$AIC(\lambda) = dev(y, \alpha, \lambda) + ۲ \times dim(\alpha, \lambda)$$

در آن $dim(\alpha, \lambda)$ بعد (مؤثر) بردار پارامترها است و $dev(y, \alpha, \lambda)$ میزان انحراف مدل برازشی نسبت به داده‌ها است. محاسبه انحراف سرراست است اما تعیین بعد مؤثر پارامترها در برازش مدل‌های اسپلاین جریمه‌ایی پیچیده و مشکل است. هستای و تیبشیرانی (۱۹۹۰) براي این مساله روشی را پیشنهاد کردند به این صورت که اثر ماتریس هموارگر خطی را به‌عنوان یک تقریب براي ابعاد مؤثر مدل به‌کار بردند.

$$dim(\alpha) = tr(H)$$

مدل اسپلاین جریمه‌ای را در نظر بگیرید:

$$Y = \beta\alpha + \epsilon$$

با استفاده از

$$\hat{\alpha} = (\beta' Y)(\beta'\beta + \lambda D'D)^{-۱}$$

و مقادیر برازش داده‌شده به‌صورت زیر محاسبه می‌شوند:

$$\hat{Y} = \beta\hat{\alpha} = \underbrace{\beta(\beta^T W\beta + \lambda D^T D)^{-۱}\beta^T W}_{\text{H}} y = Hy$$

که ماتریس H را ماتریس hat می‌نامند و W ماتریس قطری از وزن‌ها است که به‌صورت زیر به‌دست می‌آید:

$$W_{ii} = \frac{1}{\nu_i}(\frac{\partial y_i}{\partial \mu})$$

ν_i واریانس y_i است. همچنین $tr(H) = \sum_{i=1}^{n} h_{ii}$ می‌توان اثر ماتریس را بدون در نظر گرفتن قطر اصلی محاسبه نمود که در زیر شرح داده‌شده است. فرض کنید n تا زوج داده به‌صورت (x_i, y_i) داریم اگر y_i يك چندجمله‌ای از درجه k در x باشد، آن‌گاه با افزایش λ اثر ماتریس H به k نزدیك خواهد شد اثبات به‌صورت زیر برقرار است. ابتدا یادآوری می‌شود که براي ماتریس‌هاي ضرب پذیر (سازگار) داریم:

$$tr(AB) = tr(BA).$$

فرض کنید

$$Q_\lambda = \lambda D^T D \qquad\qquad Q_\beta = \beta^T W \beta$$

$$\begin{aligned} \Longrightarrow tr(H) &= tr\{(Q_\beta + Q_\lambda)^{-1} Q_\beta\} \\ &= tr\{Q_\beta^{1/2}(Q_\beta + Q_\lambda)^{-1} Q_\beta^{1/2}\} \\ &= tr\{(I + Q_\beta^{-1/2} Q_\lambda Q_\beta^{-1/2})^{-1}\} \end{aligned}$$

اگر

$$L = Q_\beta^{-1/2} Q_\lambda Q_\beta^{-1/2}$$

آن‌گاه

$$tr(H) = tr\{(I + \lambda L)^{-1}\} = \sum_{j=1}^{n} \frac{1}{I + \lambda \gamma_j} \quad \gamma_j \qquad (۱.۳)$$

γ_j مقادیر ویژه L هستند. در رابطه (۱.۳) اگر مقدار λ بزرگ باشد کسر موردنظر صفر می‌شود و فقط جملاتي صفر نمی‌شوند که γ_j مربوط به آن‌ها برابر با صفر باشد در نتیجه کسر مربوط به آن‌ها یک می‌شود و در نتیجه مجموع برابر با k می‌شود.

۲.۱.۳ اعتبار سنجی متقابل

یکی دیگر از روش‌های انتخاب پارامتر همواری استفاده از معیار اعتبارسنجي متقابل است که به اختصار با CV نشان داده می‌شود. روشی ساده و بصری برای برآورد پیش‌گویی یک مدل است. در این روش مبنای انتخاب λ براساس مینیمم کردن $CV(\lambda)$ است.

در این روش داده‌ها را به دو قسمت نمونه آزمون و نمونه آموزشی (مجموعه داده‌هایی که برای برازش مدل استفاده می‌شود) تقسیم می‌شوند. فرض کنید n تا مشاهده $y_1, \cdots, y_n$

داریم ابتدا مشاهده y_i را حذف کرده و مدل را با $n-۱$ مشاهده‌ی باقیمانده برازش داده و سپس خطای پیش‌گویی $MSE_i = E(y_i - \hat{y}_i)^۲$ محاسبه می‌کنیم. این فرایند را تا زمانی‌که هر مشاهده یک بار خارج شود انجام می‌دهیم و با ادامه این روند برای سایر مشاهدات و میانگین گرفتن از آن‌ها برآورد خطای پیش‌گویی به‌صورت معادله‌ی زیر حاصل می‌شود.

$$CV_{(n)} = \frac{۱}{n}\sum_{i=۱}^{n} MSE_i. \qquad (۲.۳)$$

به‌طور کلی CV دارای پتانسیل بالایی برای پیاده‌سازی می‌باشد زیرا مدل n بار برازش داده می‌شود اما اگر n (تعداد مشاهدات) زیاد باشد محاسبه CV بسیار زمان‌بر و پرهزینه است. اما برای مدل‌های خطی یا رگرسیون چندجمله‌ای با استفاده از ماتریس hat محاسبه آن بسیار ساده و آسان است. اگر مقادیر روی قطر اصلی ماتریس H را با h_{ii} نشان دهیم، آماره اعتبارسنجی متقابل به‌صورت معادله‌ی زیر قابل محاسبه است:

$$CV(\lambda) = \sum_{i=۱}^{n}\left(\frac{Y_i - \hat{Y}_i}{۱ - h_{ii}}\right)^۲ \qquad (۳.۳)$$

و یک نتیجه بسیار شگفت آور این که برای محاسبه CV تنها یک‌بار مدل به کل داده‌ها برازش داده می‌شود [۴، ۱۳].

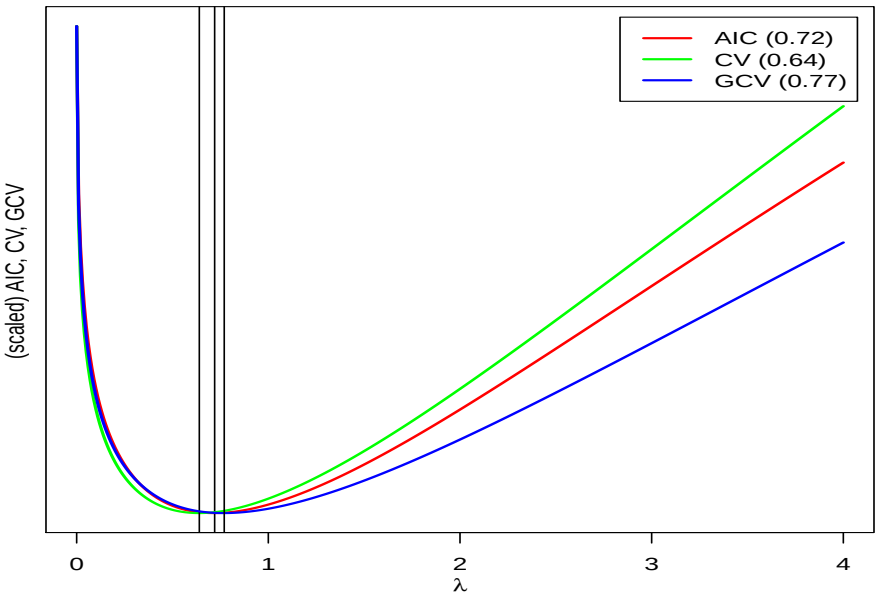
۳.۱.۳ اعتبارسنجي متقابل تعمیم‌یافته

روش دیگر برای انتخاب بهینه λ استفاده از معیار اعتبارسنجی متقابل تعمیم‌یافته که به‌صورت زیر است:

$$GCV(\lambda) = \sum_{i=۱}^{n}\frac{(Y_i - \hat{Y}_i)^۲}{(n - \sum_{i=۱}^{n} h_{ii})^۲} \qquad (۴.۳)$$

این معیار به گونه‌ای است که کمترین مقدار اعتبارسنجی متقابل تعمیم‌یافته، بهترین پارامتر هموارسازی را ارئه می‌دهد [۸، ۳۲].

در شکل ۱.۳ برای داده‌های موتور سیکلت شامل $n = ۱۳۳$ مشاهده برای آزمودن کلاه ایمنی که در آن متغیر وابسته، زمان پس از تاثیر (بر حسب میلی ثانیه) و متغیر مستقل شتاب سر (بر حسب گرم) است سیلورمن (۱۹۸۵)، نتایج مربوط به انتخاب پارامتر هموارسازی بر حسب معیارهای AIC، CV و GCV نشان داده‌شده است. مقادیر به‌دست آمده هر سه معیار، دقیقا یک مقدار را نشان نمی‌دهند اما نزدیک $\lambda = ۰/۷۵$ مقدار مینیمم را اختیار نموده‌اند.



شکل ۱.۳: پارامتر هموارسازی در مقابل معیار انتخاب برای تابع رگرسیونی P-اسپلاینی از مرتبه $m = ۴$

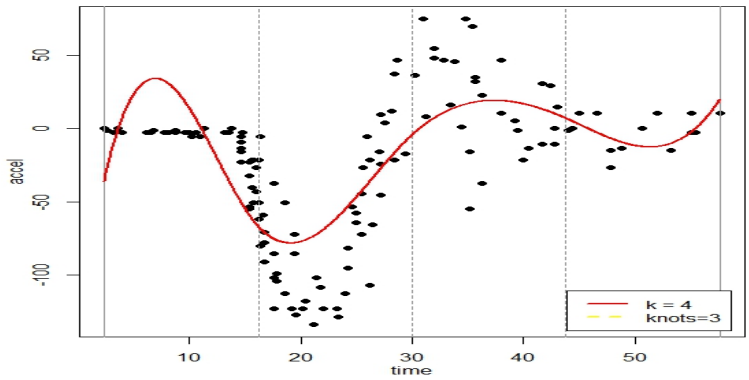
۲.۳ روش‌های انتخاب تعداد و موقعیت گره

یکی دیگر از چالش‌ها در رگرسیون اسپلاینی انتخاب تعداد و موقعیت گره‌ها می‌باشد. در رگرسیون اسپلاینی محدوده‌هایی انعطاف‌پذیری بیشتری دارند که تعداد گره‌ها در آن بیشتر باشد، زیرا در این محدوده‌ها تعداد ضرایب چندجمله‌ای به‌سرعت تغییر می‌کند و اثر زیادی روی برازش اسپلاینی دارد در نتیجه منحنی برازشی ناهموار خواهد بود. در مکان‌هایی که تعداد گره‌ها کمتر است به‌نظر می‌رسد منحنی پایدارتر است [۱۳].

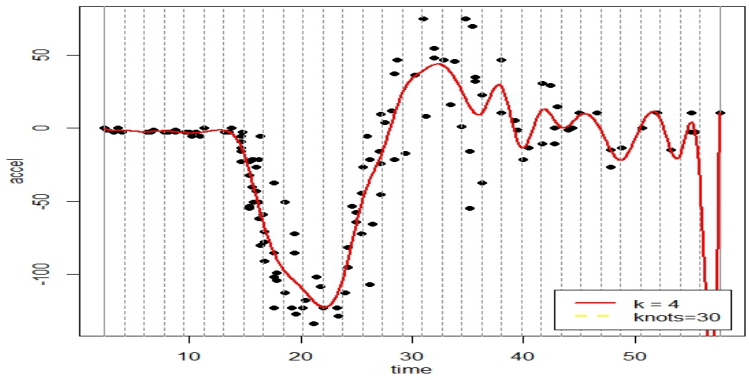
در شکل ۲.۳ و ۳.۳ داده‌های موتور سیکلت با استفاده از P-اسپلاین مرتبه $k = ۴$ با تعداد گره‌های مختلف برازش داده‌شده است که در شکل ۲.۳ با در نظر گرفتن تعداد کم گره، داده‌ها کم برازش شده‌اند به‌عبارتی اگر براساس تعداد پارامترهای کم، عمل برازش را انجام دهیم، مدل دچار کم‌برازشی خواهد شد و قدرت پیش‌بینی را از دست خواهد داد و به علت کمبود پارامترهای به‌کار رفته، خطای زیادی در مدل به‌وجود می‌آید. البته رفع مشکل کم‌برازشی با افزایش تعداد پارامترهای مدل قابل رفع است، ولی نکته‌ای که باید رعایت کرد، انتخاب تعداد پارامترهای مناسب در مدل است که آن را دچار بیش‌برازشی نکند.

در شکل ۳.۳ با در نظر گرفتن تعداد زیاد گره‌ها داده‌ها بیش برازش شده‌اند. مدل بیش‌برازش، مدلی بسیار پیچیده‌ای است به این معنی که در تحلیل رگرسیونی، مدلی با بیشترین پارامترها ایجاد می‌شود. در این گونه موارد، اگر مدل رگرسیون به‌دست آمده، برای پیش‌بینی نمونه دیگری به‌کار رود، مقدارهای پیش‌بینی شده اصلا مناسب به‌نظر نخواهند

رسید. انتظار ما از یک تحلیل رگرسیون مناسب، ایجاد مدلی است که نه تنها بتواند برای داده‌های مربوط به نمونه برازش مناسب را انجام دهد، بلکه برای داده‌هایی جدید نیز امکان برآورد مناسب وجود داشته باشد.

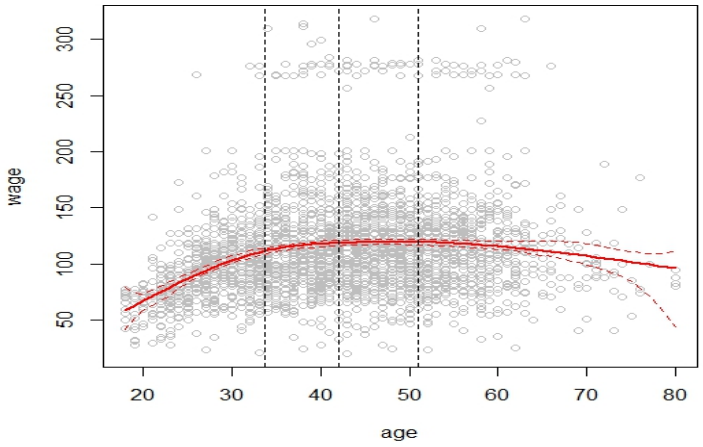


شکل ۲.۳: برازش P-اسپلاین مرتبه ۴ k با تعداد گره کم



شکل ۳.۳: برازش P-اسپلاین مرتبه ۴ k با تعداد گره زیاد

یک روش برای انتخاب تعداد گره‌های مناسب این است که تعداد گره‌های مختلف را به‌کار ببریم و بهترین منحنی تولید شده را انتخاب کنیم. در عمل بیشتر رایج است که چندک‌ها را به‌عنوان گره در نظر می‌گیرند و با مشخص کردن درجه آزادی، نرم افزار محل مربوط به تعداد گره‌ها را در چندک داده‌ها قرار می‌دهد. در شکل ۴.۳ یک اسپلاین مکعبی طبیعی با چهار درجه آزادی به داده‌ها برازش داده‌شده است و موقعیت گره‌ها به‌طور خودکار در ۲۵، ۵۰ و ۷۵ درصد از داده‌ها قرار گرفته است.



شکل ۴.۳: برازش اسپلاین مکعبی طبیعی با چهار درجه آزادی و قرار گرفتن موقعیت گره‌ها در چندک داده‌ها

اکنون یکی از چالش‌ها، انتخاب تعداد مطلوب گره و انتخاب درجه آزادی مناسب برای تابع اسپلاینی است. یک روش آن است که تعداد گره‌های مختلف را به‌کار ببریم و بهترین منحنی تولید شده متناظر با آن را مشاهده کنیم. روش دیگر استفاده از اعتبار سنجی متقابل است که معرفی شد، به این صورت که بخش‌هایی از داده‌ها را حذف کنیم (به‌طور مثال ۱۰٪) و یک تابع اسپلاینی با تعداد گره مشخص به داده‌های باقیمانده برازش داده و سپس مقدار خطای پیش‌گویی را به‌دست آوریم. این فرایند برای تعداد مختلف k گره می‌تواند تکرار شود تا زمانی‌که هر مشاهده یک‌بار خارج شود، سپس تعداد گره‌ها برابر کوچکترین مقدار خطای پیش‌گویی خواهد بود [۱۳] .
روش دیگر استفاده از الگوریتمی که توسط روپرت و کارول در سال ۲۰۰۰ ارائه شد که به‌صورت زیر است:

۱ـ تعداد گره $k = ۵$ قرار بده و مقدار $\hat{\alpha}_{۱GCV}$ را محاسبه کن.

۲ـ تعداد گره $k = ۱۰$ قرار بده و مقدار $\hat{\alpha}_{۲GCV}$ متناظر آن را محاسبه کن.

۳ـ اگر $\hat{\alpha}_{۲GCV} > ۰/۹۸\hat{\alpha}_{۱GCV}$ آن‌گاه $k = ۱۰$ نهایی شود. در غیر این صورت مقدار k افزایش می‌یابد.

۳.۳ انتخاب مدل

انتخاب مدل بخش مهمی از هر تحلیل آماری است. فرآیندی است که بهترین مدل را از مجموعه مدل‌های کاندید انتخاب می‌کند. هدف از انتخاب مدل، انتخاب یک مدل مناسب

است که به‌طور کامل داده‌ها را توضیح دهد و مقدار خطا را به حداقل برساند [۴]. برای گزینش بهترین مدل از معیارهای انتخاب مدل شامل AIC، BIC، CV و C_p ... استفاده می‌شود. مدلی که پایین‌ترین مقادیر معیارهای ذکر شده را به خود اختصاص دهد به‌عنوان مدل با دقت برازش بالا انتخاب می‌شود و مدلی که از نظر معیار مذکور بالاترین مقادیر را داشته باشد به‌عنوان مدل با عملکرد پایین معرفی می‌شود.

به دو معیار اطلاع اکائیک و اعتبارسنجی اشاره شد اکنون در مورد معیار اطلاع بیزی و آماره‌ی C_p بحث می‌کنیم.

۱.۳.۳ معیار اطلاع بیزی (شوارتز)

یکی دیگر از معیارهایی که می‌تواند برای بررسی و انتخاب بهترین مدل به‌کار رود، معیار اطلاع بیزی[۴] (BIC) است که توسط شوارتز در ۱۹۷۸ ارائه شد و براساس رابطه‌ی زیر قابل محاسبه می‌باشد:

$$BIC_p = nln(\frac{SSE_p}{n} + p(lnn))$$

که p تعداد پارامترهای قابل برآورد در مدل و n تعداد مشاهدات در نمونه تصادفی است.

۲.۳.۳ آماره C_p مالو

آماره‌ی C_p برای قضاوت در مورد یک مدل، باید به‌جای میانگین مربعات انحراف از مدل، میانگین مربعات خطای مقدار پیش‌بینی‌شده در نظر گرفته شود. میانگین مربعات خطای پیش‌بینی استاندارد شده j_p برای داده‌های مشاهده شده به‌صورت زیر محاسبه می‌شود.

$$j_p = \frac{1}{\sigma^2}\sum_{i=1}^{n} MSE(\hat{y}_i)$$

که در آن $MSE(\hat{y}_i)$ میانگین مربعات خطای iامین مقدار پیش‌بینی‌شده از یک معادله‌ی p جمله‌ای و σ واریانس مانده‌ها است. برای برآورد j_p مالو (۱۹۷۳) آماره‌ی زیر را به‌کار برد:

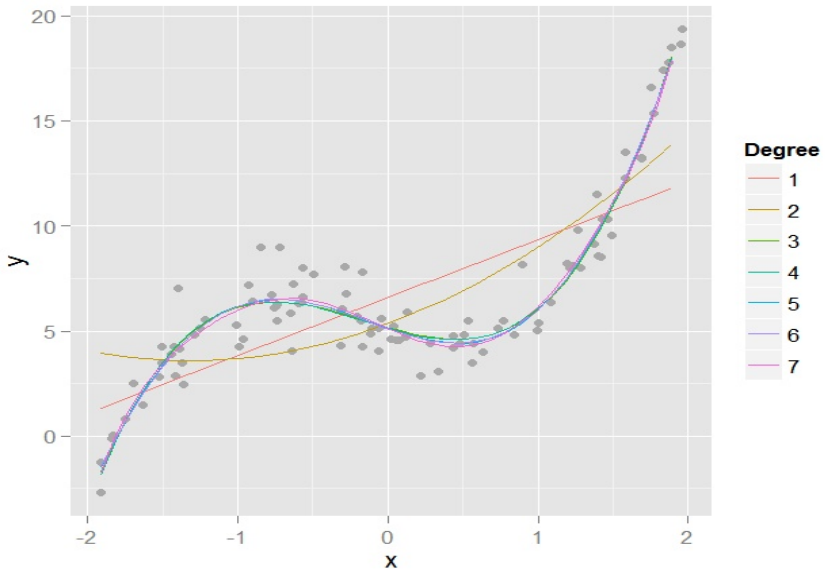
$$C_p = \frac{SSE_p}{\hat{\sigma^2}} + (2p - n)$$

که در آن $\hat{\sigma^2}$ برآورد σ^2 است. بر اساس این معیارها، مدلی که دارای کمترین مقدار معیار باشد به‌عنوان بهترین مدل انتخاب می‌شود.

[۴] Bayesian information criterion

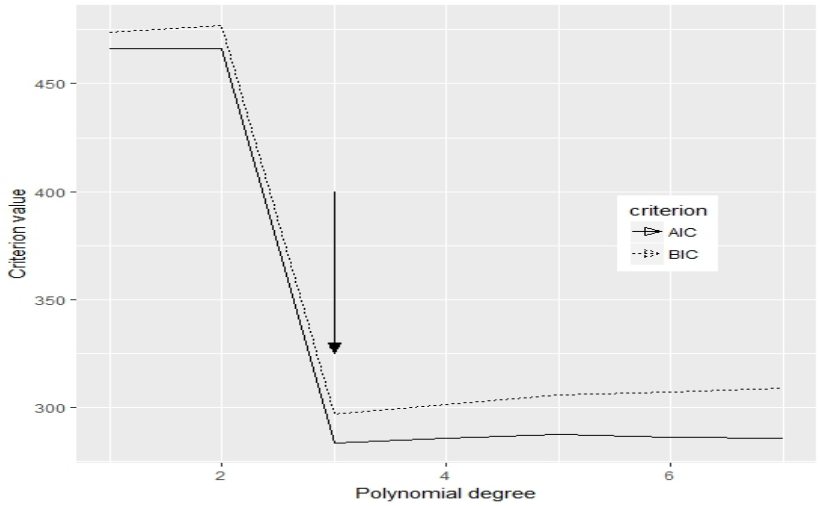
استون (۱۹۹۷) نشان داد که در مدل‌های خطی، معیار اطلاع آکائیک و اعتبارسنجی متقابل با هدف پیش‌گویی معادل هستند. برای بررسی این ایده، مجموعه داده‌ای از رگرسیون چندجمله‌ای شبیه‌سازی‌شده است و چندجمله‌ای از مرتبه یک تا هفت را به داده‌ها برازش داده و با استفاده از معیارهای فوق صحت آن را می‌آزماییم. هم‌چنین بر حسب علاقه مقدار آماره C_p و معیار اطلاع بیزی BIC را محاسبه نموده‌ایم.

شکل ۵.۳ مربوط به برازش مدل چندجمله‌ای از مرتبه یک تا هفت برای داده‌های شبیه‌سازی‌شده می‌باشد که در آن چندجمله‌ای‌های مرتبه‌ی سه تا هفت برازش بهتری نسبت به چندجمله‌ای با درجات یک و دو دارند و روند داده‌ها را بهتر توصیف می‌کنند. اکنون با استفاده از معیار اطلاع آکائیک، اعتبارسنجی و معیار اطلاع بیزی بهترین مدلی که می‌تواند به داده‌ها برازش یابد را مشخص می‌کنیم.



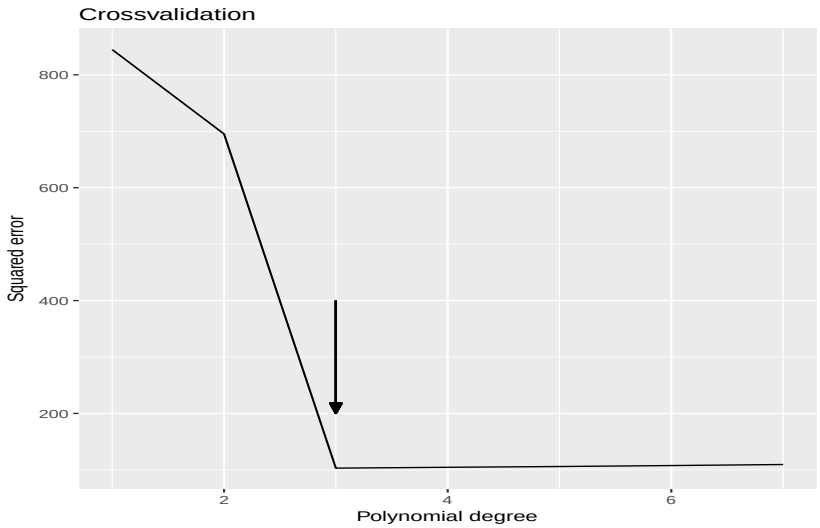
شکل ۵.۳: برازش چندجمله‌ای از مرتبه یک تا مرتبه هفت برای مجموعه داده‌های شبیه‌سازی‌شده

طبق شکل ۶.۳ مقدار معیار اطلاع آکائیک و بیزی برای چندجمله‌ای‌ها تا درجه هفت به دست آمده را رسم نموده و در نمودار رسم شده مشهود می‌باشد که چندجمله‌ای درجه سه دارای کمترین مقدار معیارهای فوق می‌باشد در نتیجه بهترین مدلی است که می‌تواند داده‌ها را توصیف کند.



شکل ۶.۳: مقایسه اثربخشی معیار اطلاع اکائیک و بیزی در انتخاب بهترین مدل

هم‌چنین با توجه به نمودار ۷.۳ مقدار CV برای چندجمله‌ای تا درجه هفت رسم شده گویای این است که مدل چندجمله‌ای با درجه سه به‌عنوان بهترین مدل برازشی است.



شکل ۷.۳: انتخاب بهترین مدل با استفاده از معیار اعتبارسنجی متقابل

در جدول ۱.۳ مقدار معیارهای معیار اطلاع آکائیک، اعتبارسنجی، معیار اطلاع بیزی و آماره‌ی C_p برای مدل چندجمله‌ای تا درجه هفت به دست آمده و نتایج حاصل حاکی از آن می‌باشد که مدل چندجمله‌ای درجه سه کمترین مقادیر معیارهای فوق را دارا می‌باشد، در نتیجه به‌عنوان مطلوب‌ترین مدل انتخاب می‌گردد.

جدول ۱.۳: جدول مقادیر معیارهای AIC, CV, BIC, C_p برای انتخاب بهترین مدل

درجه آزادی	مجموع توان دوم CV	BIC	Cp	AIC
۱	۶۱۸/۰۰۴۸	۴۷۳/۸۹۱۷	۵۲۵/۵۹۴۸	۴۶۶/۰۷۶۲
۲	۶۴۰/۲۸۹۵	۴۷۶/۸۰۳۲	۵۱۷/۱۵۵۵	۴۶۶/۳۸۲۵
۳	۱۰۰/۱۵۳۳	۲۹۶/۸۰۶۶	۴/۴۷۹۸	۲۸۳/۷۸۰۷
۴	۱۰۴/۴۸۸۹	۳۰۱/۴۱۰۷	۶/۴۷۸۸	۲۸۵/۷۷۹۷
۵	۱۰۸/۳۷۶۵	۳۰۵/۸۹۴۲	۸/۳۶۱۴	۲۸۷/۶۵۸۰
۶	۱۰۷/۱۴۳۴	۳۰۷/۰۸۶۶	۷/۱۲۸۳	۲۸۶/۲۴۵۳
۷	۱۱۰/۳۴۲۸	۳۰۹/۱۰۵۲	۶/۷۵۰۴	۲۸۵/۶۵۸۷

فصل ۴

برخی از کاربردهای اسپلاین‌ها

در این فصل کاربرد مدل‌های GAMLSS که چارچوبی بسیار عمومی برای تحلیل داده‌ها به منظور مدل‌سازی و یادگیری در زمینه‌های مختلف را فراهم می‌سازد ارائه شده است. این توابع در حال حاضر به‌طور گسترده‌ای در مدل‌سازی آماری مورد استفاده قرار می‌گیرند. از مزایای بسیار عالی این توابع فراهم کردن انتخاب گسترده‌ای از مدل‌های رگرسیونی و برازش مدلی انعطاف‌پذیر به داده‌ها می‌باشد. در این مدل متغیر پاسخ می‌تواند هر توزیعی داشته باشد و تمام پارامترهای توزیع می‌توانند مدل شوند. توابع GAMLSS در بسیاری از زمینه‌های کاربردی مورد استفاده قرار گرفته است از جمله علوم اجتماعی، زیست شناسی، انرژی، هواشناسی، بارش باران و ....مثال‌های استفاده شده زیر نمونه کوچکی برای نشان دادن قابلیت توابع GAMLSS با استفاده از نرم افزار R می‌باشد. این فصل برگرفته از کتاب استامینوپولوس و همکارانش [۲۸] می‌باشد.

مثال اول مربوط به مجموعه داده‌های اجاره که طریقه‌ی استفاده از توابع GAMLSS را نشان می‌دهد و با استفاده از این توابع یک مدل پیچیده سه پارامتری به داده‌ها برازش می‌یابد. مثال دوم مجموعه داده‌های گونه‌های مختلف ماهی که مثالی ساده از برازش توزیع‌های شمارشی مختلف با استفاده از توابع GAMLSS به داده‌ها است. مثال سوم مجموعه داده‌های مدت بستری که متغیر پاسخ دارای توزیع دوجمله‌ای است. این داده‌ها در توسعه توابع GAMLSS نقش زیادی داشته‌اند، چون اولین مجموعه داده بودند که محققان پی به وجود توابع GAMLSS بردند. توزیع‌های دو و سه پارامتری به‌کار رفته در مثال‌ها در پیوست در جدول آ.۱ و آ.۲ با

جزئیات بیشتری شرح داده‌شده است.

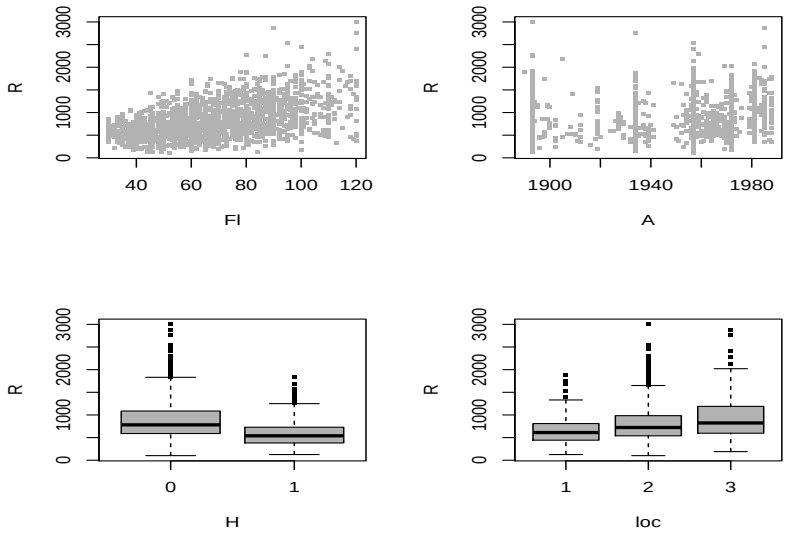
۱.۴ مدل‌های جمعی

- مدل جمعی تعمیم‌یافته (GAM)
- مدل جمعی تعمیم‌یافته برای مکان، مقیاس و شکل (GAMLSS)

مدل جمعی جایگزینی برای مدل‌های خطی می‌باشد و می‌تواند برای تعیین روابط بین متغیرها مورد استفاده قرار گیرد. این مدل می‌تواند کیفیت پیش‌بینی متغیر پاسخ را به حداکثر برساند و روابط غیرخطی بین متغیر پاسخ و مجموعه متغیرهای پیش‌گو را کشف کند. در شرایطی که پراکندگی داده‌ها زیاد باشد، برای ساخت مدلی دقیق‌تر بهتر است از مدل GAMLSS استفاده شود تا از خانواده‌ی توزیع‌های عمومی به‌جای خانواده‌ی توزیع‌های نمایی استفاده شود. مدل GAMLSS این قابلیت را دارد که می‌تواند داده‌های پراکنده را با خطای کمتری برازش نماید. از مزایای این مدل نسبت به مدل جمعی تعمیم‌یافته که فقط به مدل‌سازی میانگین شرطی متغیرهای پاسخ محدود هستند این است که برای هر پارامتر توزیع رابطه‌ی رگرسیونی ایجاد می‌کند. در فصل اول مدل‌های GAM و GAMLSS به‌طور کامل شرح داده‌شده است.

۱.۱.۴ کاربرد اول:داده‌های اجاره مونیخ

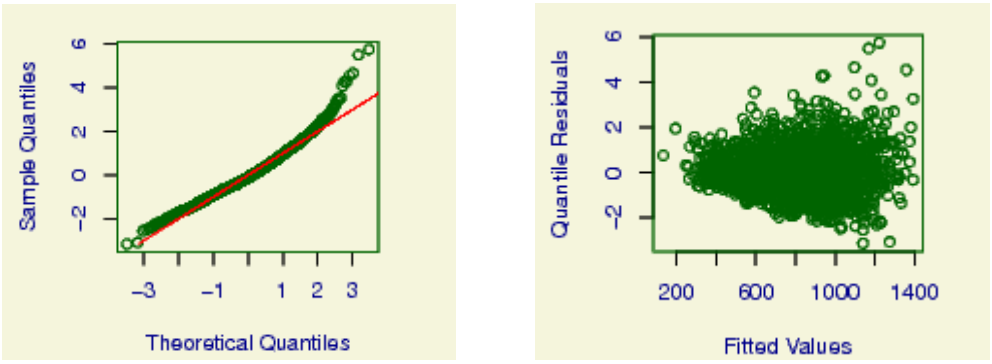
داده‌های مربوط به اجاره بر اساس یک نظرسنجی در اوریل ۱۹۹۳ توسط اینفراتست سوزیال جمع آوری شد که در آن یک نمونه تصادفی از محل اقامت با موافقت‌نامه اجاره جدید یا افزایش اجاره طی چهارسال گذشته در مونیخ انتخاب شد. این مجموعه داده‌ها شامل، $n = ۱۹۶۹$ مشاهده است که اجاره خالص ماهیانه نقش متغیر پاسخ و متغیرهای توضیحی متراژ (F1)، سال ساخت (A)، گرمایش مرکزی (H) شامل دو سطح، وجود دارد (۰) و عدم وجود (۱) و عاملی که نشان دهنده‌ی موقعیت ساختمان (Loc) شامل سه سطح متوسط رو به پایین (۱) ، متوسط (۲)، متوسط رو به بالا (۳).
شکل ۱.۴ نمودار متغیر پاسخ، میزان اجاره (R)، در مقابل هر یک از متغیرهای توضیحی را نشان می‌دهد.



شکل ۱.۴: نمودار متغیر پاسخ (میزان اجاره R) در مقابل هریک از متغیرهای توضیحی

نمودار ترسیم شده‌ی میزان اجاره (R) در مقابل متراژ(F1) نشان می‌دهد که با افزایش متراژ میزان اجاره نیز افزایش می‌یابد و رابطه‌ای مثبت دارند در نتیجه فرض همگنی واریانس نقض خواهد شد و ممکن است واریانس داده‌های اجاره به میانگین یا به متغیرهای توضیحی بستگی داشته باشد. هم‌چنین چولگی مثبتی هم در توزیع داده‌های اجاره نیز وجود دارد. نمودار ترسیم شده میزان اجاره در مقابل سال ساخت و ساز نشان می‌دهد که تا سال ۱۹۶۰ میزان اجاره تقریبا ثابت است اما با ساخت و سازهای بعد آن سال، روندی افزایشی در میزان اجاره وجود دارد. دو نمودار جعبه‌ای نشان می‌دهد که میزان اجاره طبق عامل‌های توضیحی متفاوت است یعني میزان متوسط اجاره اگر گرمایش مرکزی و موقعیت تغییر کند از متوسط رو به کم به متوسط و به متوسط رو به بالا افزایش می‌یابد.

اکنون باید مدل برازشی به داده‌های اجاره قادر به ارائه مدلی باشد که رابطه‌ی غیرخطی بین متغیر پاسخ و متغیرهای توضیحی، عدم همگنی واریانس داده‌ها و هم‌چنین چولگی مثبت توزیع دادهای اجاره را بیان کند. مدل‌هایی را به داده‌های اجاره برازش می‌دهیم و با استفاده از تابع GAMLSS در نرم افزار R به تحلیل داده‌های اجاره می‌پردازیم. برای ارزیابی کفایت یک مدل به یک مجموعه داده، می‌توان به بررسی مانده‌ها پرداخت. تابع GAMLSS از مانده‌های استاندارد شده استفاده می‌کند. مدل خطی $R \sim Fl+A+H+loc$ را به داده‌ها برازش می‌دهیم. شکل ۲.۴ نمودار مانده‌های حاصل از برازش مدل خطی است.

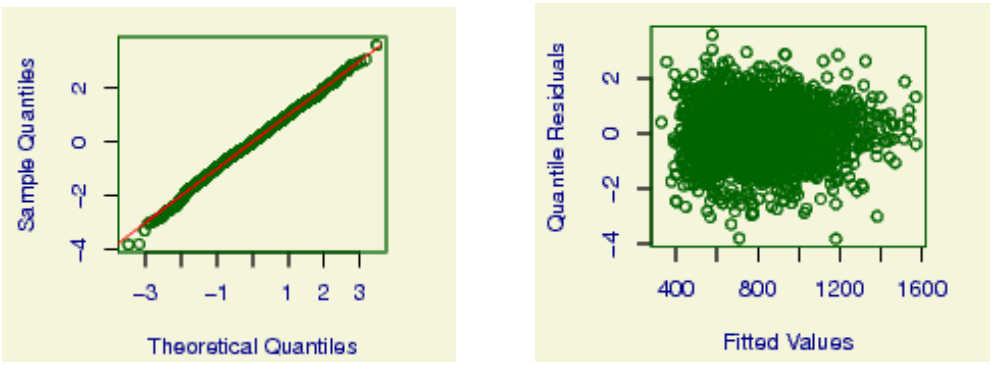


(آ) (ب)

شکل ۲.۴: نمودار مانده‌های برازش مدل خطی

باتوجه به نمودار چندک‌ـچندک شکل (ب)، هرچه نقاط نمودار به خط نزدیک‌تر باشند، توزیع داده‌ها به نرمال نزدیک‌تر است درنتیجه فرض نرمال بودن توزیع داده‌ها پذیرفته نمی‌شود. نمودار شکل (آ)، مانده‌ها در مقابل مقادیر برازشی بر روی خطی افقی در صفر به‌طور تصادفی پراکنده نشده‌اند و این نشان از ناهمگونی واریانس دارد یعنی واریانس با افزایش میانگین تغییر می‌کند. با توجه به بررسی‌های مدل، فرض نرمال بودن توزیع داده‌ها پذیرفته نمی‌شود.

فرض می‌کنیم مجموعه داده‌های اجاره دارای توزیع گاما که مدل خطی تعمیم‌یافته است و برای بررسی کفایت این مدل به مانده‌ها می‌پردازیم. در شکل ۳.۴ نمودارهای ترسیمی نمایش داده شده است.

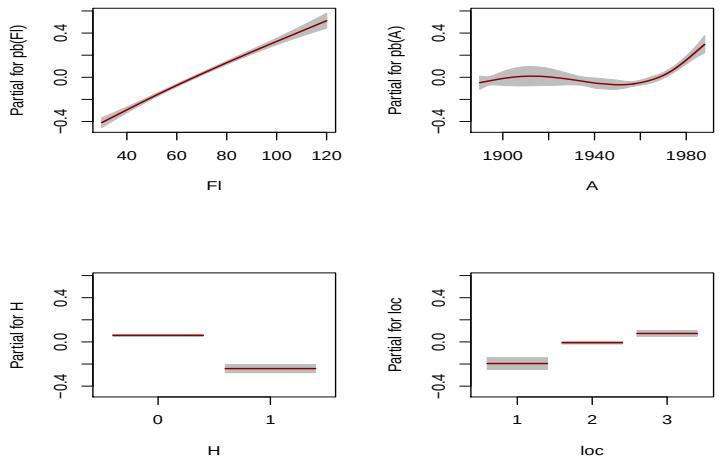


(آ) (ب)

شکل ۳.۴: نمودار مانده‌های برازش مدل گاما

مانده‌ها در شکل ۳.۴ وضعیت بسیار بهتری نسبت به شکل ۲.۴ دارند در آن ناهمگنی در مانده‌ها در مقابل مقادیر برازشی از بین رفته اند و همچنین خمیدگی نمودار چندک‌ـچندک به شدت کاهش یافته است، اگرچه ناهمگونی کمی در مانده‌ها وجود دارد.

در ادامه مدل‌های جمعی تعمیم‌یافته را برازش می‌دهیم که برای بررسی رابطه‌ی بین متغیرهای پیش‌گو و پاسخ توابع هموارساز (pb) را به‌کار می‌برد که مدلی انعطاف‌پذیر از هموارسازهای اسپلاین جریمه‌ای ارائه می‌دهد.



شکل ۴.۴: نمودار مقادیر برازشی برای مدل (GAM)

در نمودارهای ۴.۴ مشهود است که میزان اجاره تقریبا با افزایش متراژ (F1) به‌طور خطی افزایش می‌یابد و با سال ساخت (A) رابطه‌ای غیرخطی دارد و تا سال ۱۹۶۰ تقریبا ثابت است اما بعد آن رو به افزایش است و میزان تاثیر دو عامل H و loc همان مقدار مورد انتظار است، یعنی میزان اجاره کم است اگر عامل H=1 باشد و هم‌چنین نسبت به سطوح موقعیت، میزان اجاره از کم به متوسط و به زیاد افزایش پیدا می‌کند.

روش دیگر برای ارزیابی کفایت مدل، استفاده از نمودار مارپیچ[1] برای تحلیل باقیمانده‌ها است، می‌تواند به عنوان نمودار چندک‌ـچندک در نظر گرفته شود که توسط برن و فریدریک در سال ۲۰۰۱ معرفی شد . ابزاری تشخیصی است برای مقایسه مدل‌ها و مشخص نمودن مدل آماری متناسب با داده‌ها و پیدا کردن مکان‌هایی که مدل می‌تواند بهبود پیدا کند. دارای ویژگی‌هایی است:

- نقاط نمودار مارپیچ هر چه به خط افقی نزدیک‌تر باشد توزیع باقیمانده‌ها به توزیع نرمال استاندارد نزدیک‌تر است.
- در صورتی که مدل کفایت داشته باشد انتظار می‌رود که به‌طور تقریبی ۹۵٪ نقاط بین دو نمودار بیضوی و ۵٪ خارج از آن باشد.
- شکل (الگو) منحنی برازشی به نقاط نمودار مارپیچ، بازتابنده‌ی عدم کفایت مدل برازشی است که در جدول ۱.۴ شرح داده‌شده است. به‌عنوان مثال اگر سطح نقاط نمودار مارپیچ

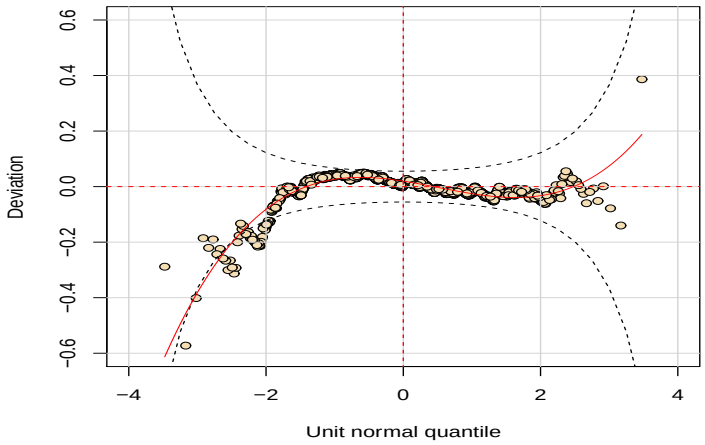
[1] Worm plot

بالاتر از خط افقی در مبدا باشد این نشان دهنده میانگین بسیار بالای مانده‌ها است، بدان معنی است که موقعیت توزیع برازشی بسیار پایین است که می‌توان با افزایش پارامتر μ اگر پارامتری مکانی است مدل را بهبود بخشید یا توزیع مدل را تغییر داد.

جدول ۱.۴: تفسیر الگوهای مختلف در نمودار مارپیچ

شکل منحنی برازشی	گشتاور	توزیع برازش داده‌شده
سطح نقاط اگر بالای مبدا باشد	میانگین بسیار بالا است	موقعیت برازش بسیار پایین است
سطح نقاط اگر زیر مبدا باشد	میانگین بسیار کم است	موقعیت برازش بسیار بالا است
خط دارای شیب مثبت باشد	واریانس بسیار بالا است	مقیاس برازش داده‌شده بسیار کم است
خط دارای شیب منفی باشد	واریانس بسیار کم است	مقیاس برازش داده‌شده بسیار بالا است
$\cup$ـ شکل باشد	چولگی مثبت دارد	توزیع به سمت چپ کشیده شده است
$\cap$ـ شکل باشد	چولگی منفی دارد	توزیع به سمت راست کشیده شده است
Sـ شکل باشد که به پایین خم شده است	کشیده	توزیع برازشی بسیار دم‌باریک است
Sـ شکل باشد که به بالا خم شده است	خم پخ	توزیع برازشی بسیار دم‌کلفت است

با توجه به نمودار ۵.۴ برای داشتن مدل برازشی مطلوب انتظار داریم که نقاط نزدیک به خط افقی میانه باشد و به‌طور تقریبی ۹۵٪ نقاط بین نمودار بیضوی بالا و پایین باشد و ۵٪ خارج از آن باشد.



شکل ۵.۴: نمودار مارپیچ برای مدل (GAM)

هم‌چنین برای ارزیابی عملکرد پیش‌بینی مدل برازشی می‌توان از معیار AIC استفاده نمود. دو مدل خطی تعمیم‌یافته و جمعی تعمیم‌یافته‌ی برازشی را با استفاده از معیارAIC مقایسه

می‌کنیم طبق جدول ۱۰.۴ کمترین AIC مربوط به مدل جمعی تعمیم‌یافته به میزان ۲۷۷۰۵/۶۵ است. در نتیجه در مقایسه با دو مدل دیگر مدل بهتری است.

جدول ۲.۴: گزینش بهترین مدل با استفاده معیارهای AIC

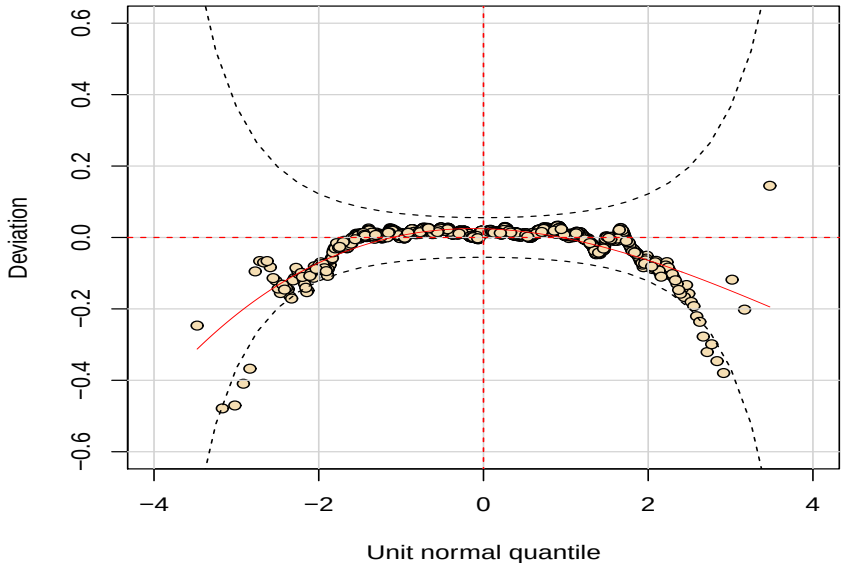
مدل	تابع پیوند		پیش‌گوی جمعی	AIC
	μ	σ		
نرمال	ident	log	Fl+A+H+loc	۲۸۱۷۳
گاما (مدل خطی تعمیم‌یافته)	log	log	Fl+A+H+loc	۲۷۷۷۸/۵۹
گاما (مدل جمعی تعمیم‌یافته)	log	log	pb(Fl)+pb(A)+H+loc	۲۷۷۰۵/۶۵

برای بهبود بخشیدن به مانده‌ها می‌توان پارامتر σ را مدل‌سازی کرد. داده‌های اجاره مونیخ را با استفاده از توزیع‌های گامای دو پارامتری و گاوسی وارون مدل می‌کنیم، مقدار AIC برای این دو مدل فوق و گاما که پارامتر مقیاس آن ثابت فرض شده است محاسبه شده است. با توجه به جدول ۳.۴ طبق معیار AIC مدلی که با استفاده از توزیع گاما دو پارامتری که پارامتر مقیاس آن مدل شده است، مدل مطلوب‌تری است.

جدول ۳.۴: گزینش بهترین مدل با استفاده معیارهای AIC

مدل	تابع پیوند		پیش‌گوی جمعی		AIC
	μ	σ			
گاما	log	log	pb(Fl)+pb(A)+H+loc	sigma.fo=1	۲۷۷۰۵/۶۵
گاما	log	log	pb(Fl)+pb(A)+H+loc	sigma.fo=pb(Fl)+pb(A)+H+loc	۲۷۶۱۴/۷۸
گاوسی وارون	log	log	pb(Fl)+pb(A)+H+loc	sigma.fo=pb(Fl)+pb(A)+H+loc	۲۷۷۱۶/۶۶

برای بررسی کفایت مدل گامای دو پارامتری، نمودار مارپیچ در زیر رسم شده است. با توجه به نمودار مارپیچ شکل ۶.۴ تعداد کمی از نقاط وجود دارد که خارج از فاصله اطمینان ۹۵٪ نمودار نقطه‌نقطه می‌باشد، که این نشان دهنده‌ی عدم کفایت مدل می‌باشد، علاوه بر این شکل U وارونه‌ی مانده‌ها اشاره‌ای به چولگی منفی مانده‌ها دارد که حاکی از این می‌باشد که توزیع گاما انعطاف‌پذیری کافی برای از بین بردن چولگی در داده‌ها را ندارد.



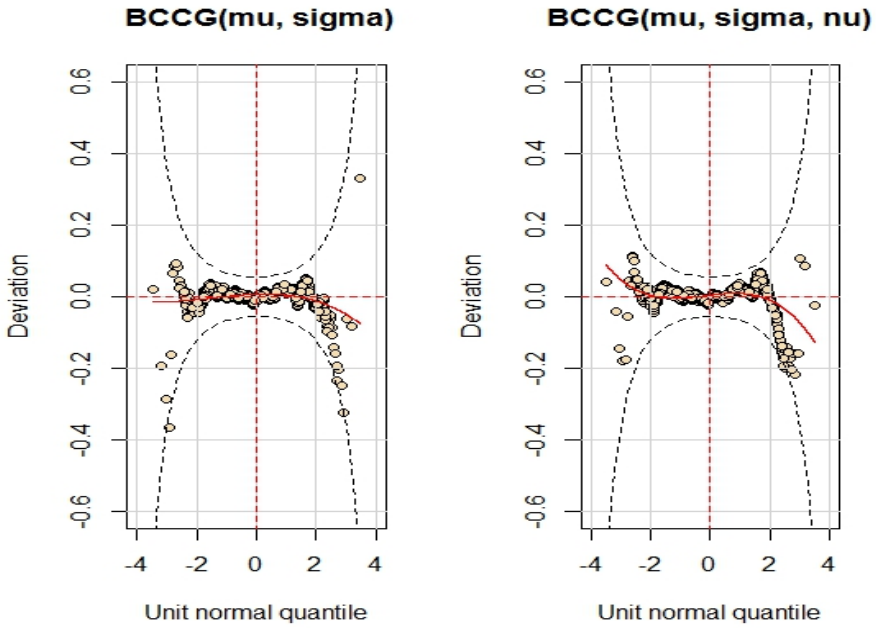
شکل ۶.۴: نمودار مارپیچ برای مدل گامای دو پارامتری

مدل باکس‌کاکس کول و گرین (BCCGO) که مدلی جمعی تعمیم‌یافته برای مکان، مقیاس و شکل (GAMLSS) را در نظر می‌گیریم که انعطاف‌پذیری بیشتری نسبت به مدل‌های پیشین دارد، توزیعی سه پارامتری است که هر پارامتر با استفاده از تابع هموارساز ناپارامتری مدل‌سازی شده است. دو مدل به داده‌های اجاره برازش داده، در یک مدل ν ثابت در نظر گرفته شده و در مدل دیگر ν مدل‌سازی شده و با گامای دو پارامتری مقایسه شده است. با توجه به جدول ۴.۴ طبق معیار AIC مدلی که پارامتر ν مدل‌سازی شده است بهبودی اندکی در برازش مدل ایجاد نموده است.

جدول ۴.۴: گزینش بهترین مدل با استفاده معیارهای AIC

مدل	تابع پیوند			پیش‌گوی جمعی	AIC
	μ	σ	ν		
باکس‌کاکس (BCCGO)	log	log	ident	pb(Fl)+pb(A)+H+loc sigma.fo=pb(Fl)+pb(A)+H+loc nu.fo=1	۲۷۶۱۱/۰۲
باکس‌کاکس (BCCGO)	log	log	ident	pb(Fl)+pb(A)+H+loc sigma.fo=pb(Fl)+pb(A)+H+loc nu.fo=pb(Fl)+pb(A)+H+loc	۲۷۶۰۸/۱۵

با استفاده از نمودار مارپیچ شکل ۷.۴ کفایت دو مدل برازشی را بررسی می‌کنیم.



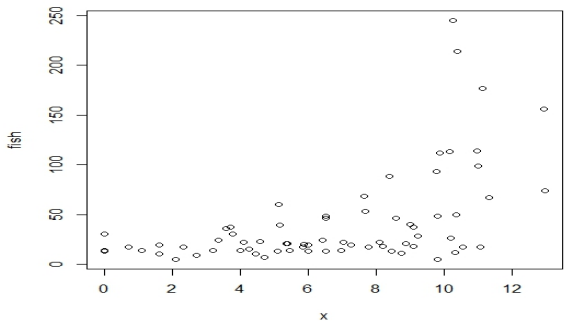
شکل ۷.۴: نمودار مارپیچ برای مدل سه پارامتری

دو نمودار ترسیمی حاکی از کفایت مدل توزیع سه پارامتری می‌باشند.

۲.۱.۴ کاربرد دوم: داده‌های گونه‌های ماهی

داده‌های گونه‌های ماهی توسط استین و جیرتیز [۲۹] با استفاده از توزیع معکوس پواسن گاوسی $\text{PIG}(\mu,\sigma)$ با در نظر گرفتن یک مدل خطی برای $\log(\mu)$ و یک مقدار ثابت برای σ تولید شده‌اند، ارائه شد. این داده‌ها در بسته نرم افزاری gamlss.data موجود می‌باشد که تعداد گونه‌های مختلف ماهی در ۷۰ دریاچه جهان (fish) به‌عنوان متغیر پاسخ و منطقه دریاچه (lake) متغیر توضیحی می‌باشد. در شکل ۸.۴ نمودار مربوط به تعداد گونه‌های ماهی در مقابل متغیر توضیحی x=lake نشان داده‌شده است.

داده‌ها را با استفاده از توزیع‌های گسسته مختلف مانند پواسن (PO)، پواسن دوگانه (DPO)، دوجمله‌ای منفی نوع یک و دو (NBI,NBII)، معکوس پواسن گاوسی(PIG)، دلاپورت (DEL)، سیچل (SICHEL) که در بسته‌ی gamlss.dist با تابع پیوندی پیش‌فرض موجود می‌باشد مدل‌سازی می‌کنیم. متغیر توضیحی x=lake را یک‌بار با مدل خطی و سپس مدل چندجمله‌ای درجه دوم در مقابل متغیر پاسخ برازش می‌دهیم و با استفاده از معیار AIC مدل‌های برازشی را مقایسه می‌کنیم.



شکل ۸.۴: نمودار گونه‌های ماهی

مقادیر به‌دست آمده، با استفاده از معیار AIC در جدول ۵.۴ آورده‌شده است.

جدول ۵.۴: مقادیر به‌دست آمده با استفاده از AIC

مقدار AIC برای برازش متغیر پاسخ در مقابل متغیر توضیحی با فرض خطی بودن						
PO	DPO	NBI	NBII	PIG	DEL	SICHEL
۱۹۰۰/۱۵۶۲	۶۵۴/۱۶۱۶	۶۲۵/۸۴۴۳	۶۴۷/۵۳۵۹	۶۲۳/۴۶۳۲	۴۶۲۶/۲۳۳۰	۶۲۵/۳۹۲۳
مقدار AIC برای برازش متغیر پاسخ در مقابل متغیر توضیحی با فرض چندجمله‌ای درجه دوم						
PO	DPO	NBI	NBII	PIG	DEL	SICHEL
۱۸۵۵/۲۹۶۵	۶۵۵/۲۵۲۰	۶۲۲/۳۱۷۳	۶۴۵/۰۱۲۹	۶۲۱/۳۴۵۹	۶۲۳/۵۸۱۶	۶۲۳/۰۹۹۵

با توجه به مقادیر به‌دست آمده در جدول فوق مقدار AIC برای مدل پواسن نسبت به سایر توزیع‌ها خیلی بزرگ است، بنابراین می‌توان نتیجه گرفت که داده‌ها بیش از حد پراکنده شده‌اند. نتایج نشان می‌دهد که متغیر توضیحی با مدل چندجمله‌ای درجه دوم نسبت به مدل خطی عملکرد بهتری دارد البته بجز توزیع پواسن دوگانه. تا این مرحله مدل معکوس پواسن گاوسی (PIG) با کمترین مقدار AIC به عنوان مدل مطلوب انتخاب می‌گردد.

اکنون در مدل معکوس پواسن گاوسی به‌جای برازش متغیر توضیحی با چندجمله‌ای درجه دوم، هموارگر اسپلاینی استفاده می‌کنیم و عملکرد این مدل را مورد ارزیابی قرار می‌دهیم. مقدار AIC برای این مدل به میزان۶۲۳/۴۶۳۷ است،که مقدار آن افزایش یافت درنتیجه به‌کار بردن هموارگر اسپلاینی موثر واقع نشده است. برای بهبود عملکرد مدل می‌توان $\log(\sigma)$ را مدل‌سازی نمود، اکنون $\log(\sigma)$ را به‌عنوان یک تابع خطی از x در توزیع‌های فوق بجز توزیع پواسن که پارامتر σ ندارد مدل کرد.

در جدول ۶.۴ مقادیر AIC برای توزیع‌های فوق به‌دست آمده نشان داده‌شده است.

جدول ۶.۴: مقادیر به‌دست آمده با استفاده از معیار AIC

مقدار AIC با مدل کردن پارامتر σ					
DPO	NBI	NBII	PIG	DEL	SICHEL
۶۲۶/۴۰۵۶	۶۱۴/۹۵۶۵	۶۱۵/۱۲۵۰	۶۱۲/۳۶۶۷	۶۱۴/۶۰۵۹	۶۱۳/۷۳۲۷

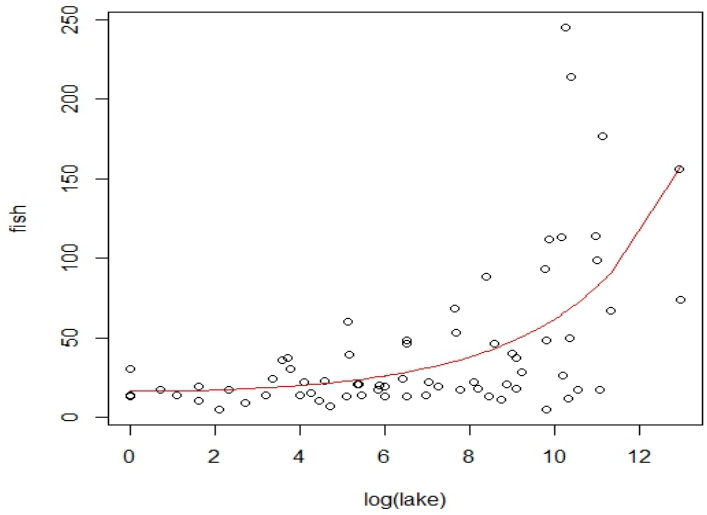
نتایج بیانگر این است که مدل کردن $\log(\sigma)$ به‌عنوان تابعی از x عملکرد تمام توزیع‌ها را بهبود بخشیده است و مدل معکوس پواسن گاوسی (PIG) با کمترین مقدار AIC به‌عنوان بهترین مدلی است که می‌تواند داده‌ها را توصیف کند. توزیع‌های سیچل و دلاپورت سه پارامتری می‌باشند که می‌توان برای بهبود بخشیدن به مدل، پارامتر سوم (ν) را به‌عنوان تابعی خطی از x مدل کرد. توزیع سیچل به‌عنوان پیش‌فرض تابع پیوندی از تابع همانی و توزیع دلاپورت از تابع لوجیت استفاده می‌کند.

در جدول ۷.۴ مقادیر AIC برای توزیع‌های سیچل و دلاپورت به‌دست آمده نشان داده‌شده است.

جدول ۷.۴: مقادیر به دست آمده با استفاده از معیار AIC

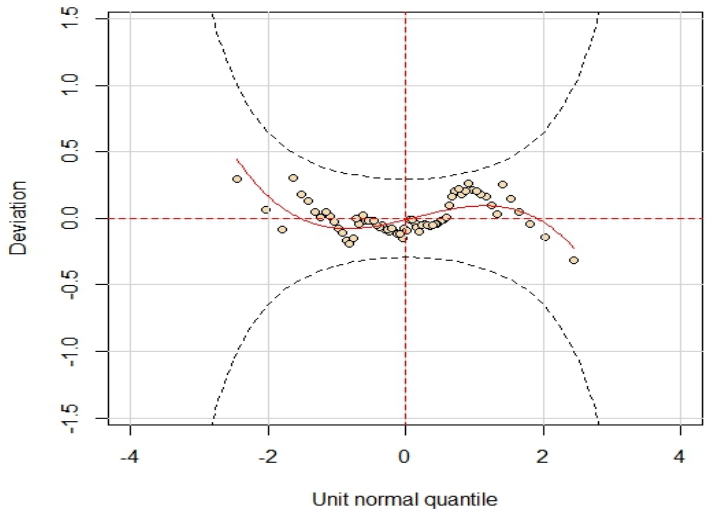
مقدار AIC برای توزیع‌های سیچل و دلاپورت با مدل کردن پارامتر (ν)	
DEL	۶۱۴/۷۳۷۶
SICHEL	۶۱۱/۶۳۴۶

با توجه به نتایج، مدل کردن پارامتر (ν) به‌عنوان تابعی از x، مدل سیچل (SIC) را نسبت به مدل معکوس پواسن گاوسی (PIG) بهبود بخشیده است و دارای مقدار AIC کمتری نسبت به آن شده است اما مدل دلاپورت (DEL) بهبودی نیافته است.
در شکل ۹.۴ برازش مدل سیچل به داده‌های گونه‌های ماهی نشان داده‌شده است که به‌عنوان بهترین مدلی است که داده‌ها را توصیف می‌کند.



شکل ۹.۴: نمودار برازش مدل سیچل به داده‌های گونه‌های ماهی

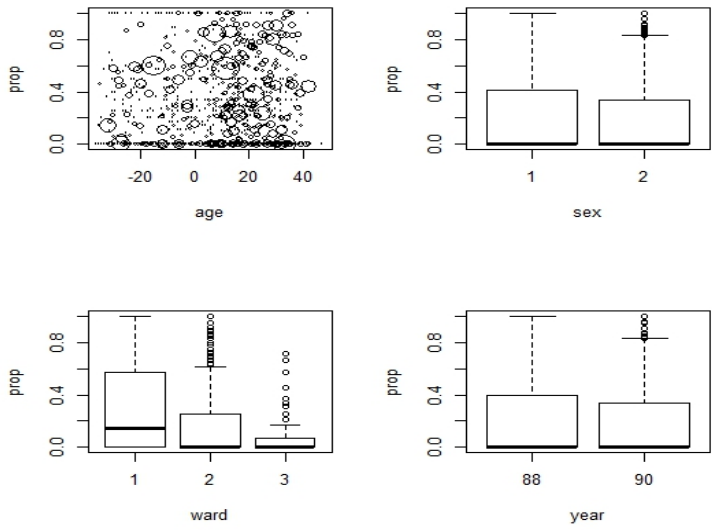
شکل ۱۰.۴ نمودار مارپیچ مدل برازشی سیچل را نشان می‌دهد که نمودار حاکی از کفایت مدل می‌باشد.



شکل ۱۰.۴: نمودار مارپیچ مدل برازشی سیچل به داده‌های گونه‌های ماهی

۳.۱.۴ کاربرد سوم: داده‌های مدت بستری در بیمارستان

داده‌های مدت بستری در بیمارستان aep در بسته gamlss.data موجود می‌باشد که دارای ۱۳۸۳ مشاهده، توسط گنگ و همکارانش در طول سال‌های ۱۹۸۸ و ۱۹۹۰ در بیمارستان دلمر بارسلونا جمع آوری شده است. متغیر پاسخ، تعداد روزهای غیرمفید از مجموع روزهایی که بیماران در بیمارستان بستری‌اند (noinap) و جنسیت بیمار (sex)، نوع بخش بیمارستان (ward)، سن بیمار (age)، سال ۱۹۸۸ یا ۱۹۹۰ (year)، لگاریتم (مجموع روزهای مدت بستری ÷ ۱۰) (loglos) به‌عنوان متغیرهای توضیحی می‌باشد. هر بیمار برای ماندن غیر مفید در بیمارستان توسط دو پزشک با استفاده از پروتکل ارزیابی مناسب[۲] (aep) مورد بررسی قرار گرفته‌اند.
در شکل ۱۱.۴ نمودار نرخ نامناسب متغیر پاسخ در مقابل age, sex, ward, yaer نشان داده‌شده است.



شکل ۱۱.۴: نمودار نرخ نامناسب متغیر پاسخ در مقابل متغیرهای توضیحی در داده‌های مدت بستری

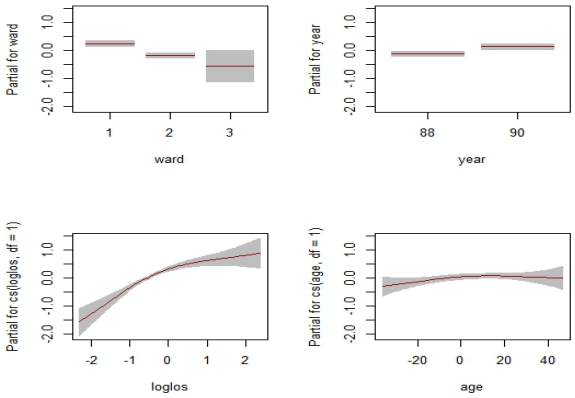
گنگ و همکارانش مدل رگرسیون دوجمله‌ای را برای تعداد روزهای غیرمفید و توزیع‌های دوجمله‌ای و بتا دوجمله‌ای را برای متغیر پاسخ به‌کار بردند و دریافتند که مدل بتا دوجمله‌ای با پارامترهای بازپارامتری شده برازش بهتری را به‌دست می‌آورد. آن‌ها هر دو پارامتر میانگین (μ) و مقیاس (σ) توزیع بتا دوجمله‌ای را با تابع پیوند لوجیت و پارامتر مقیاس را با تابع پیوند همانی به عنوان تابعی از متغیرهای توضیحی مدل کردند.
در جدول ۸.۴ چهار مدل به داده‌ها aep برازش داده شده که انعطاف‌پذیری مدل‌های gamlss را نشان می‌دهد و براساس معیارهای AIC و BIC بهترین مدل برازشی انتخاب می‌شود.

[۲] Appropriateness evaluation protoco

جدول ۸.۴: گزینش بهترین مدل با استفاده معیارهای AIC و BIC

مدل	تابع پیوند	پیش‌گوی جمعی	AIC	BIC
M1	$\text{logit}(\mu)$	ward+loglos+year	۴۵۳۳/۴	۴۵۷۰/۱
	$\log(\sigma)$	year		
M2	$\text{logit}(\mu)$	ward+loglos+year	۴۵۰۱/۰	۴۵۴۸/۱
	$\log(\sigma)$	year+ward		
M3	$\text{logit}(\mu)$	ward+cs(loglos,df=1)+year	۴۴۷۹/۴	۴۵۳۱/۷
	$\log(\sigma)$	year+ward		
M4	$\text{logit}(\mu)$	ward+cs(loglos,df=1)+year+cs(age,df=1)	۴۴۷۸/۴	۴۵۴۱/۱
	$\log(\sigma)$	year+ward		

نتایج چهار مدل برازشی به داده‌ها در جدول ۸.۴ نمایش داده‌شده است. مدل M1 مدل منتخب گنگ و همکارانش است. در مدل M2 متغیر ward برای σ وارد مدل شده است و در مدل M3 و M4 برای مدل کردن μ تابع اسپلاین مکعبی هموارساز با یک درجه آزادی استفاده شده است. با توجه به مقدار AIC مدل M4 بهترین مدل است و با اضافه کردن تابع هموارساز برای μ در مدل M4 مدل بهبود یافته و مقدار AIC کاهش یافته است. طبق معیار BIC بهترین مدل، مدل M3 است که با اضافه کردن تابع هموارساز برای age در مدل M4 مقدار BIC افزایش یافته در نتیجه مدل مطلوب نمی‌باشد. در زیر نمودار تمام توابع برازشی برای μ در مدل M4 نشان داده‌شده است.



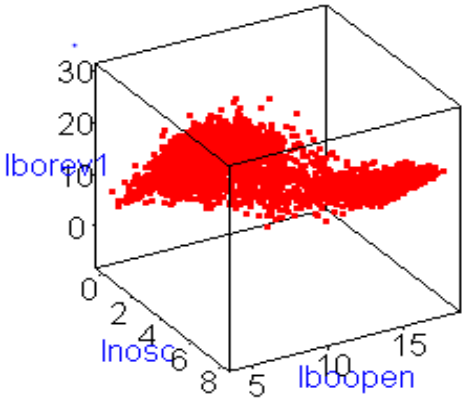
شکل ۱۲.۴: توابع برازشی برای μ در مدل M4

۲.۴ اسپلاین دو بعدی

روش‌های زیادی برای درونیابی و مدل‌سازی سطوح به‌کار می‌رود. زمانی‌که که بیش از یک متغیر مستقل موجود باشد، مساله‌ی برازش یک رویه‌ی چندبعدی پیش می‌آید و برای این که روند داده‌ها بهتر نشان داده شود، می‌توان روش‌های هموارسازی اسپلاین را تعمیم داد. اسپلاین حاصل‌ضرب تانسور و اسپلاین صفحات نازک از جمله روش‌هایی برای مدل‌سازی سطوح است. داده‌های فیلم را برای نشان دادن ویژگی‌های توابع gamlss و به‌عنوان مثالی برای برازش سطوح هموارساز دوبعدی برای یک متغیر پاسخ پیوسته به‌کار می‌بریم.

۱.۲.۴ داده‌های فیلم

داده‌های فیلم توسط ویدوریس و همکارانش [۳۷] برای نشان دادن بسیاری از ویژگی‌های تابع gamlss و بررسی اثرات متقابل هموارساز متغیرهای توضیحی به‌کار برده شدند. داده‌های فیلم شامل ۴۰۱۵ مشاهده و چهار متغیر می‌باشد. لگاریتم تعداد پرده‌های سینما که فیلم نمایش داده شده است (lnosc)، لگاریتم سود هفتگی صندوق (lboopen) و عاملی است که نشان می‌دهد آیا توزیع کننده‌ی فیلم مستقل یا عمده می‌باشد (dist)، به‌عنوان متغیرهای توضیحی و لگاریتم سود هفتگی بعد از اولین هفته (lborev1)، متغیر پاسخ می‌باشد. داده‌ها در دو یا سه بعد قابل نمایش می‌باشند. با استفاده از بسته rgl می‌توان داده‌ها را در فضای سه بعدی ترسیم نمود.



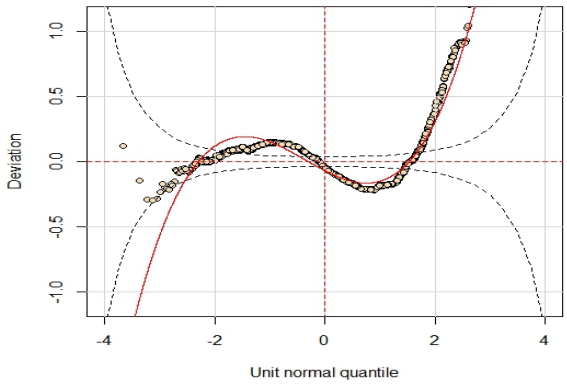
شکل ۱۳.۴: ترسیم نمودار سه بعدی داده‌های فیلم

بعد از تجزیه و تحلیل مقدماتی داده‌ها، با در نظر گرفتن توزیع نرمال برای متغیر پاسخ، مدل‌های برازشی به داده‌ها در جدول ۹.۴ نمایش داده‌شده است. در دو مدل اول برای هریک از متغیرهای توضیحی مدلی با اثر متقابل خطی، در دو مدل دوم هموارساز جمعی و در دو مدل سوم رویه‌ای هموار در نظر گرفته شده است. برای گزینش بهترین مدل از معیارهای AIC و BIC استفاده شده است.

جدول ۹.۴: گزینش بهترین مدل با استفاده معیارهای AIC و BIC

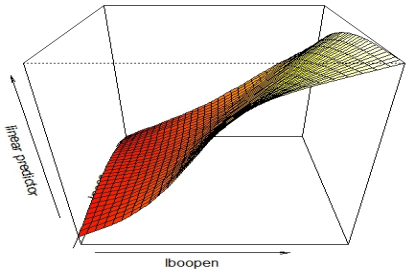
	مدل	توزیع متغیر پاسخ	پیش‌گوی جمعی	AIC	BIC
مدل با اثرات متقابل خطی	M1	normal	boopen∗lnosc	۱۲۲۲۶/۸۴	۱۲۲۵۸/۳۵
	M2	normal	lboopen∗lnosc+dist	۱۲۰۸۰/۹۹	۱۲۱۱۸/۸۰
مدل جمعی	M3	normal	pb(lnosc) +pb(lboopen)	۱۱۹۰۸/۷۳	۱۲۰۲۲/۹۶
	M4	normal	pb(lboopen)+ pb(lnosc)+dist	۱۱۸۴۳/۷۸	۱۱۹۴۵/۳۹
مدلی با برازش رویه	M5	normal	ga(te(lboopen,lnosc))	۱۱۸۲۸/۵۹	۱۱۹۴۴/۰۶
	M6	normal	ga(te(lboopen,lnosc))+dist	۱۱۷۷۹/۷۶	۱۱۸۸۰/۶۹

با توجه به مقادیر به‌دست آمده در جدول ۹.۴ مدل M6 که رویه‌ای را برای متغیرهای توضیحی برازش می‌دهد به‌عنوان بهترین مدل برازشی می‌باشد. نتایج حاصله حاکی از این می‌باشد که در نظر گرفتن توزیع نرمال مناسب نمی‌باشد که در نمودار مارگون رسم شده‌ی مدل M6 کاملا مشخص است. زیرا بیشتر نقاط خارج از بین دو نمودار بیضوی بالا و پایین می‌باشد، که این نشان دهنده‌ی عدم کفایت مدل می‌باشد.



شکل ۱۴.۴: نمودار مارگون برای مدل M6 با برازش رویه برای پارامتر μ

با استفاده از بسته mgcv می‌توان نمودار رویه‌ی برازشی برای مدل M6 را رسم نمود.



شکل ۱۵.۴: نمودار رویه‌ی برازشی برای مدل M6

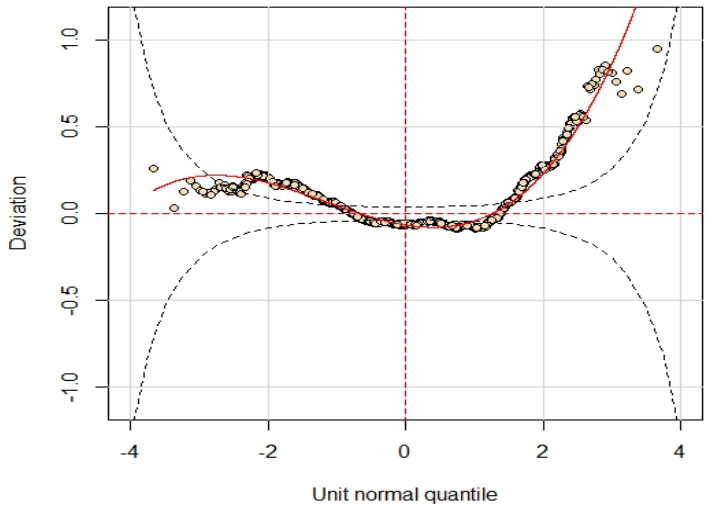
اکنون با مدل کردن پارامتر مقیاس به عنوان تابعی از متغیرهای توضیحی می‌توان بهبود یافتن مدل را بررسی نمود.

جدول ۱۰.۴: گزینش بهترین مدل با استفاده معیارهای AIC

	مدل	توزیع متغیر پاسخ	پیش‌گوی جمعی	AIC	BIC
مدلی با برازش رویه	M7	normal	ga(te(lboopen,lnosc))+dist sigma.fo=ga(te(lboopen,lnosc))+dist	۱۰۲۱۸/۱۲	۱۰۰۴۳/۸۹

میزان AIC مدل M7 نسبت به مدل M6 مقدار کمتری است، در نتیجه مدل بهتری نسبت به

مدل M6 می‌باشد. برای بررسی کفایت مدل، نمودار مارگون مدل M7 رسم شده است و مشهود می‌باشد که این مدل کفایت لازم را ندارد و نمی‌تواند متغیر پاسخ را به‌خوبی توضیح دهد.



شکل ۱۶.۴: نمودار مارگون برای مدلM7 با برازش رویه برای پارامتر های μ و σ

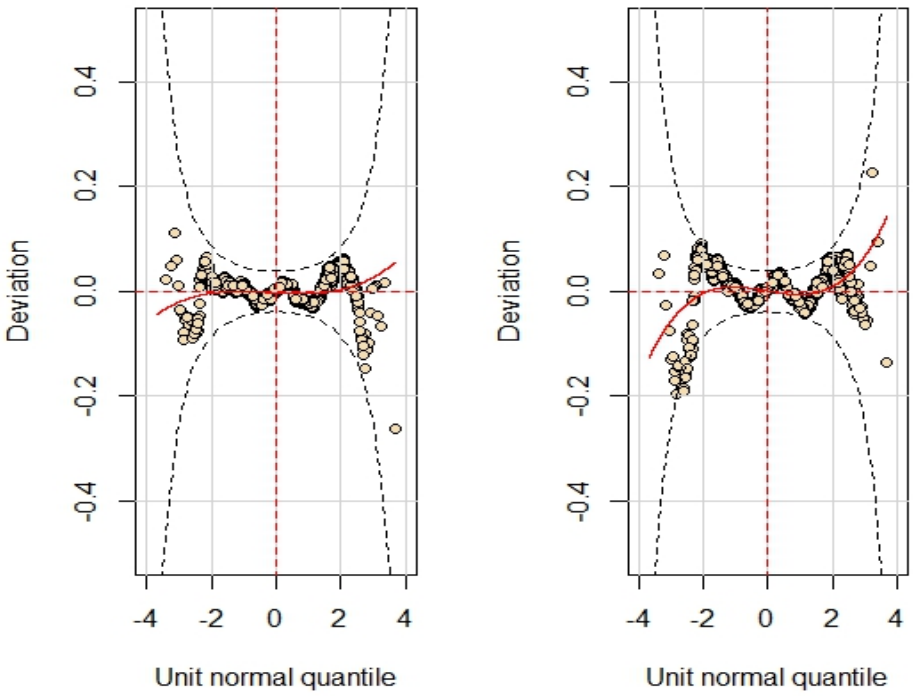
اکنون پارامتر پاسخ را با استفاده از توزیع BCPE [۳] که توزیعی از خانواده‌ی توزیع‌های موجود در بسته gamlss.dist می‌باشد برازش می‌دهیم. توزیعی چهار پارامتری است که پارامترهای μ و ν به‌طور پیش فرض دارای تابع پیوندی همانی و پارامترهای σ و τ دارای تابع پیوندی لگاریتمی می‌باشد. در مدل M8 به هر چهار پارامتر مدل جمعی و در مدل M9 رویه‌ی هموارساز برازش یافته و در جدول ۱۱.۴ نمایش داده‌شده است.

[۳] BoxCox power exponential

جدول ۱۱.۴: گزینش بهترین مدل با استفاده معیارهای AIC

مدل	تابع پیوند				پیش‌گوی جمعی	AIC
	μ	σ	ν	τ		
M8	ident	log	ident	log	pb(lboopen)+pb(lnosc) + dist sigma.fo = pb(lboopen)+pb(lnosc) + dist nu.fo = pb(lboopen)+pb(lnosc) + dist tau.fo =pb(lboopen)+pb(lnosc) + dist	۹۹۸۰/۹۴۸
M9	ident	log	ident	log	ga(te(lboopen,lnosc)) + dist sigma.fo =ga(te(lboopen,lnosc)) + dist nu.fo =ga(te(lboopen,lnosc)) + dist tau.fo =ga(te(lboopen,lnosc)) + dist	۹۹۸۰/۹۴۸

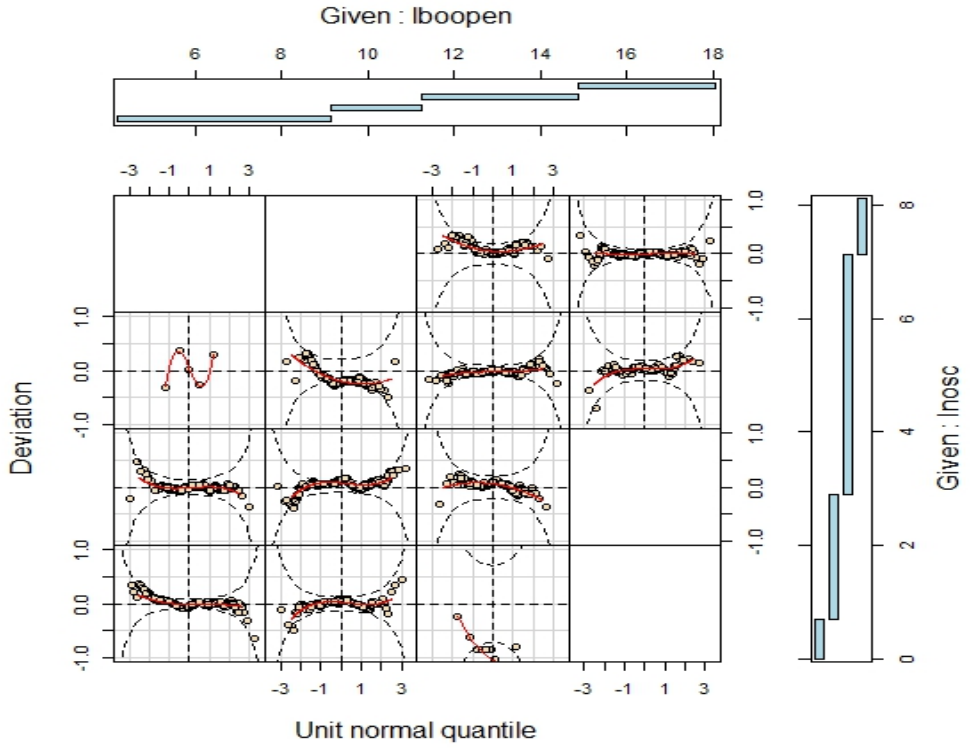
با توجه به مقادیر معیارها، عملکرد مدل M9 کمی بهتر است، با توجه به شکل ۱۷.۴ نمودار مارگون ترسیمی دو مدل فوق به‌نظر می‌رسد که مدل M8 بهتر از مدل M9 است.



شکل ۱۷.۴: نمودار مارگون قاب سمت چپ مدل M8 و قاب سمت راست مدل M9

تحت این شرایط انتخاب مدل بهتر، دشوار می‌باشد، لذا می‌توان در نواحی مشترک دو متغیر

lboopen و lnosc مدلی را برازش داد و نمودار مارگون برای هر دو متغیر توضیحی را به‌کار برد که در زیر رسم شده است.



شکل ۱۸.۴: نمودار مارگون برای مدل M9 با دو متغیر توضیحی

چهار ستون مربوط به چهار ناحیه از متغیر lboopen در بالای نمودار و چهار ردیف مربوط به چهار ناحیه از متغیر lnosc در سمت راست نمودار رسم شده است. در داخل نمودار، ۱۶ نمودار مارگون در ناحیه‌های پیوستگی دو متغیر وجود دارد. در بعضی از ناحیه‌ها مشاهده‌ای موجود نمی‌باشد. به‌طور کلی نمودارهای مارگون در ناحیه‌های پیوستگی نشان دهنده‌ی کفایت مدل برازشی است.

۳.۴ مدل میدان‌های تصادفی گاوسی

مدل میدان‌های تصادفی گاوسی (GMRF) به‌طور گسترده‌ای در آمار فضایی استفاده می‌شود. این مدل زمانی‌که یک عامل اطلاعات جغرافیایی را دربر دارد بسیار مناسب می‌باشند. به‌عنوان مثال اطلاعات مکان یا ناحیه‌ای مشخص، مناطق،.... این داده‌ها علاوه بر این که معرف اطلاعات ناحیه‌ای مشخص هستند، بین آن‌ها وابستگی وجود دارد. مشاهداتی که در یک همسایگی قرار دارند، مشابه‌تر هستند. لذا به‌کارگیری روش‌های معمول تحلیل که در آن‌ها فرض بر استقلال مشاهدات است، برای تحلیل این گونه داده‌ها که داده‌های فضایی مشبکه‌ای

نامیده می‌شوند، مناسب نمی‌باشد. عموما از یک مدل میدان تصادفی مارکوفی گاوسی که یک خانواده مهم از توزیع‌ها است برای مدل‌بندی چنین داده‌هایی استفاده می‌شود. یک مدل (GMRF)را می‌توان به‌عنوان یک مدل اسپلاین جریمه‌ای با توجه به معادلات (۱۳.۲) (۱۴.۲) در نظر گرفت که در آن ماتریس پایه B ماتریسی شاخص است که مشاهدات آن مربوط به منطقه فضایی می‌باشد و ماتریسD^TD حاوی اطلاعات همسایگی فضایی می‌باشد. در نرم افزار R با استفاده از بسته gamlss.spatia دستور (GMRF) برای برازش مدل‌های فضایی به‌کار گرفته می‌شود. برای نشان دادن کاربرد مدل‌های (GMRF) داده‌های جرم و جنایت (columb) در شهر کلمبوس ایالت اوهایو را به‌کار می‌بریم.

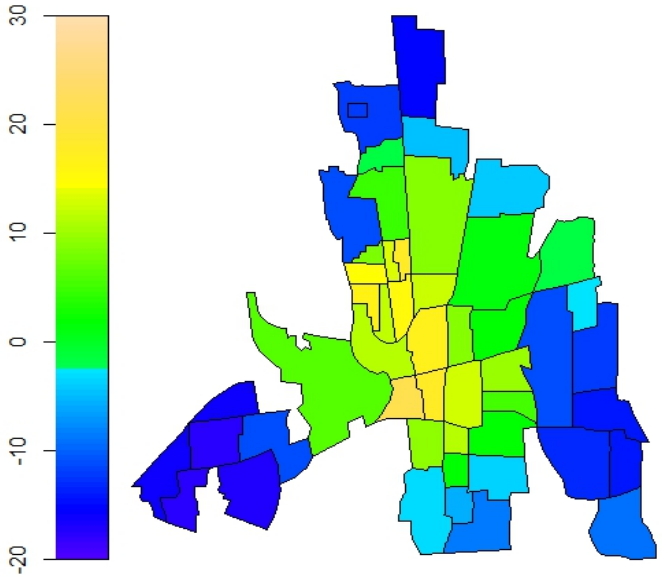
۱.۳.۴ داده‌های جرم و جنایت

داده‌های جرم تعداد ۴۹ مشاهده و دارای متغیرهای توضیحی، منطقه مسکونی (area)، قیمت مسکن (home value)، درآمد خانوار (income)، میزان فضای باز ناحیه (open.space)، کد شناسایی منطقه و مطابقت نام برای مجموعه داده‌ی columb.polys که حاوی چند ضلعی‌هایی برای مناطق می‌باشد (district) و سرقت های مسکونی و خودکار در هر ۱۰۰۰ خانوار (crime) به‌عنوان متغیر پاسخ می‌باشد. در جدول ۱۲.۴ مدل برازشی به داده‌های نمایش داده‌شده است.

جدول ۱۲.۴: مدل برازشی به داده‌های columb با استفاده از تابع gmrf

مدل	پیش‌گوی جمعی
مدل جمعی تعمیم‌یافته برای مکان، شکل و مقیاس	gmrf(district, polys=columb.polys)

در شکل ۱۹.۴ میزان جرم با توجه به طیف رنگی قابل تفسیر می‌باشد. نواحی زرد رنگ نشان دهنده‌ی منطقه‌ای است که بیشترین میزان جرم و جنایت در آن اتفاق افتاده است و نواحی آبی رنگ بیانگر کاهش میزان جرم و جنایت در آن منطقه می‌باشد.



شکل ۱۹.۴: برازش مقادیر برای داده‌های جرم جنایت با استفاده از تابع GMRF

۴.۴ نتیجه‌گیری

در این پایان‌نامه، به تفصیل، انواع مختلف اسپلاین‌ها را معرفی کردیم و روش‌های انتخاب پارامتر هموارساز را مطرح کردیم. یکی از جدی‌ترین معایب اسپلاین‌هایی که از توابع پایه ساخته می‌شوند، ناپایداری عددی است که با افزایش درجه چندجمله‌ای، هم‌خطی بین توابع بریده‌شده ایجاد می‌شود. بنابراین در این موارد، از رابطه بازگشتی برای ساختن توابع پایه استفاده می‌شود. در رابطه با اسپلاین‌های هموارساز یکی از معایب آن‌ها وابستگی به تعداد گره‌ها است که با افزایش تعداد گره‌ها با مشکلاتی از قبیل بیش‌برازشی و محاسبات سنگین مواجه می‌شویم. برای رفع این مشکل نیز از اسپلاین‌های جریمه‌ای استفاده می‌شود و هم‌چنین مدل‌های gamlss را معرفی کردیم که به منظور تحلیل و مدل‌سازی مورد استفاده قرار می‌گیرند که مدل‌های رگرسیونی و برازش‌های بسیار انعطاف‌پذیری را با کاربرد توابع ناپارامتری از جمله اسپلاین‌ها فراهم می‌کند با معیارهای انتخاب مدل هم‌چون AIC,CV, ... بهترین مدل برازشی را انتخاب می‌کنیم که بتواند داده‌ها را به‌خوبی توصیف نماید و دارای خطای کمتری باشد.

پیوست آ

آ.۱ ضرب کرونکر

ضرب کرونکر با نماد $\otimes$ نشان داده می‌شود، دارای خواص جالبی است که باعث شده در زمینه‌های مختلف به‌طور گسترده مورد استفاده قرار گیرد. اولین بار توسط یوهان گئورگ زفوس استفاده شد که به‌عنوان ضرب تانسوری نیز شناخته می‌شود.
فرض کنید $\mathbb{R}$ میدان اعداد مختلط و $\mathbb{R}^{m\times n}$ مجموعه ماتریس‌های شامل m سطر و n ستون با درایه‌های مختلط باشد [۱۵].

تعریف آ.۱.۱. ضرب کرونکر دو ماتریس $A \in \mathbb{R}^{m_A\times n_A}$ و $B \in \mathbb{R}^{m_B\times n_B}$ به‌صورت $A \otimes B$ نشان داده می‌شود و برابر است با

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n_A}B \\ \vdots & \ddots & \vdots \\ a_{m_A1}B & \cdots & a_{m_An_A}B \end{pmatrix}$$

هر $a_{ij}B$ یک بلوک در سایز $m_B \times n_A$ است بنابراین سایز $A\otimes B$ برابر $m_Am_B \times n_An_B$ است.

تعریف آ.۲.۱. جمع کرونکر ماتریس‌های مربعی $A \in \mathbb{R}^{n_A \times n_A}$ $B \in \mathbb{R}^{n_A \times n_A}$ به‌صورت $A \otimes B$ نمایش داده می‌شود و برابر است با

$$A \otimes B = A \otimes I_{n_B} + I_{n_A} \otimes B$$

خواص پایه:
شرکت‌پذیری

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C$$

توزیع‌پذیری ضرب به جمع

$$(A + B) \otimes (C + D) = A \otimes C + B \otimes C + A \otimes D + B \otimes D$$

سازگاری با معکوس ماتریس

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

سازگاری با ترانهاده‌ی ماتریس

$$(A \otimes B)^T = A^T \otimes B^T$$

یکی از مهمترین خواص ضرب کرونکر ارتباط آن با ضرب ماتریس است. برای ماتریس‌های $A \in \mathbb{R}^{m_A \times n_A}$ ، $B \in \mathbb{R}^{m_B \times n_B}$ ، $D \in \mathbb{R}^{m_D \times n_D}$ ، $C \in \mathbb{R}^{m_C \times n_C}$ و اگر $n_A = m_C$ ، $n_B = m_D$ آن‌گاه

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$$

آ.۲ مشتق تابع B-اسپلاین

دی بور (۱۹۷۸) فرمولی ساده برای مشتق توابع B-اسپلاین ارائه نمود.

$$h \sum_j \alpha_j \beta'_j(x,k) = \sum \alpha_j \beta_j(x,k-1) - \sum \alpha_{j-1} \beta_{j-1}(x,k-1) = \sum \Delta\alpha_j \beta_j(x,k-1)$$

که $\Delta\alpha_j = \alpha_j - \alpha_{j-1}$ تفاضل مرتبه اول و h فاصله‌ی بین گره‌ها است. با توجه به این مقدمات مشتق دوم را به‌دست می‌آوریم:

$$h^2 \sum_j \alpha_j \beta''_j(x,k) = (h \sum_j \alpha_j \beta'_j(x,k))' = (\sum \alpha_j \beta_j(x,k-1) - \sum \alpha_{j-1} \beta_{j-1}(x,k-1))' =$$
$$(\sum \alpha_j \beta_j(x,k-2) - \sum \alpha_{j-1} \beta_{j-1}(x,k-2)) - (\sum \alpha_{j-1} \beta_{j-1}(x,k-2) - \sum \alpha_{j-2} \beta_{j-2}(x,k-$$
$$2)) = \sum \alpha_j \beta_j(x,k-2) - 2 \sum \alpha_{j-1} \beta_{j-1}(x,k-2) + \sum \alpha_{j-2} \beta_{j-2}(x,k-2) =$$
$$\sum \alpha_j - 2\alpha_{j-1} + \alpha_{j-2} \beta_j(x,k-2) = \sum \Delta^2 \alpha_j \beta_j(x,k-2)$$

تفاضل مرتبه دوم $\Delta^۲\alpha_j = \alpha_j - ۲\alpha_{j-۱} + \alpha_{j-۲}$

۳.آ جدول توزیع‌های پیوسته و گسسته موجود در بسته‌ی gamlss.dist

جدول آ.۱: خلاصه‌ای از توزیع‌های پیوسته موجود در بسته‌ی gamlss.dist با تابع پیوند پیش‌فرض

توزیع	تابع چگالی احتمال	تابع پیوند			
		μ	σ	ν	τ
باکس‌کاکس کول و گرین (BCCGo)	$f(y\|\mu,\sigma,\nu) = (\frac{۱}{(\sqrt{(۲\pi)}\sigma)})(\frac{y^{(\nu-۱)}}{\mu^\nu})exp(\frac{-z^۲}{۲})$ $z = \begin{cases} \frac{(y\mu)^{(\nu)-۱}}{(\nu\sigma)} & \nu \neq ۰ \\ z = \frac{log(\frac{y}{\mu})}{\sigma} & \nu = ۰ \end{cases}$	*log*	*ident*	*log*	*log*
باکس‌کاکس توان نمایی (BCPE)	$f(y\|\mu,\sigma,\nu,\tau) = (\frac{y^{(\nu-۱)}}{\mu^\nu})(\frac{\tau}{\sigma})\frac{exp(-۱/۲\|\frac{z}{c}\|^\tau)}{(c^{۲(۱+\frac{۱}{\tau})})\Gamma(\frac{۱}{\tau})}$ $c = [\frac{۲^{(\frac{-۲}{\tau})}\Gamma(\frac{۱}{\tau})}{\Gamma(\frac{۳}{\tau})^{(\frac{۱}{۲})}}]$ $z = \frac{log(\frac{y}{\mu})}{\sigma}$	*log*	*ident*	*log*	*ident*
گاما (GA)	$f_z(z\|\mu,\sigma) = \frac{۱}{(\mu\sigma^۲)^{۱}/_{\sigma^۲}}\frac{z^{(\frac{۱}{\sigma^۲}-۱)}e^{-z}/_{(\mu\sigma)^۲}}{\Gamma(۱/_{(\sigma)^۲})}$	log	log	.	.
گاوسی وارون (IG)	$f(z\|\mu,\sigma) = \frac{۱}{(\sqrt{۲\pi\sigma^۲y^۳})exp(-(y-\mu)^۲/_{(}۲\mu^۲\sigma^۲y))}$	log	log	.	.

جدول آ.۲: خلاصه‌ای از توزیع‌های گسسته موجود در بسته‌ی gamlss.dist با تابع پیوند پیش‌فرض

توزیع	تابع جرم احتمال	تابع پیوند			
		μ	σ	ν	τ
پواسن (PO)	$f(y\|\mu)=\frac{e^{-\mu}\mu^y}{\Gamma(y+1)}$	log	.	.	.
پواسن دوگانه (DPO)	$f(y\|\mu,\sigma)=(\frac{1}{\sigma})^{\frac{1}{2}}[\frac{(e^{-y}y^y)}{y!}][\frac{(e\mu)}{y}]^{\frac{y}{\sigma}}C$	log	log	.	.
دوجمله‌ای منفی نوع یک (NBI)	$f(y\|\mu,\sigma)=\frac{\Gamma(y+\frac{1}{\sigma})}{\Gamma(y+1)\Gamma(\frac{1}{\sigma})}\frac{(\mu\sigma)^y}{(\mu\sigma+1)^{(y+(\frac{1}{\sigma}))}}$	log	log	.	.
دوجمله‌ای منفی نوع دو (NBII)	$f(y\|\mu,\sigma)=\frac{\Gamma(y+\frac{\mu}{\sigma})\sigma^y}{\Gamma(\frac{\mu}{\sigma})\Gamma(y+1)(1+\sigma)^{y+\frac{\mu}{\sigma}}}$	log	log	.	.
پواسن گاوسی وارون (PIG)	$f(y\|\mu,\sigma)=(\frac{2\alpha}{\pi})^{\frac{1}{2}}\mu^y e^{(\frac{1}{\sigma})}\frac{K(\alpha)}{(\alpha\sigma)^y y!}$ $\alpha^2=\frac{1}{\sigma^2}+\frac{2\mu}{\sigma}$ $K_\lambda(t)=\frac{1}{2}\int_0^\infty x^{\lambda-1}\exp\{-\frac{1}{2}t(x+x^{-1})\}dx$	log	log	.	.
دلاپورت (DEL)	$f(y\|\mu,\sigma,\nu)=(\frac{exp(-\mu\nu)}{\Gamma(\frac{1}{\sigma})})\{1+\mu\sigma(1-\nu)\}^{(\frac{-1}{\sigma})}S$ $S=\sum(P(y,j)((\mu^y)(\frac{\nu^{y-j}}{y!})\left\{1+(\frac{1}{(\sigma(1-\nu))})\right\}^j\Gamma((\frac{1}{\sigma})j)$	logit	log	log	.
سیچل (SI)	$f(y\|\mu,\sigma,\nu)=\frac{\mu^y K_{y+n}(\alpha)}{(\alpha\sigma)^{(y+v)}y!K\nu(\frac{1}{\sigma})}$ $\alpha^2=\frac{1}{\sigma^2}+\frac{2\mu}{\sigma}$ $K_\lambda(t)=\frac{1}{2}\int_0^\infty x^{\lambda-1}\exp\{-\frac{1}{2}t(x+x^{-1})\}dx$	logit	log	log	.

آ.۴ دستورات نرم‌افزار R

شکل 1.2

```
rm(list=ls())
plus <- function (x) {
  ifelse( x >= 0, x, 0 )
}
plus <- function (x) {
  ifelse( x >= 0, x, 0 )
}
op <- par(mfrow=c(1,2))
plot(0,type="n",xlim = c(-1, 2), ylim = c(-0.1, 0.5),xlab="x",ylab="pluse(x-a)")
for( a in c(-.5,1,0,.5)){
curve( plus(x - a),
         lwd = 3,
```

```
      type="l"    , add=T)
  abline(h=0, v=c(-0.5,0,0.5,1), lty=3)
}
col ="grey60"
plot(0,type="n",xlim = c(-1, 2), ylim = c(-0.1, 0.5),xlab="x",ylab="pluse(x-a)^3")
for( a in c(-.5,1,0,.5)){
curve( plus(x - a)^5,
         lwd = 3,
         col = "grey30", add=T)
  abline(h=0, v=c(-0.5,0,0.5,1), lty=3)
}
```

شکل 2.2

```
library ( ISLR)
attach (Wage )
agelims = range ( age )
age.grid =seq ( from= agelims [1], to= agelims [2])
library(splines)
plot(age ,wage , col =" gray ")
fit =lm(wage~ bs(age , knots =c(25 ,40 ,60) ),data= Wage)
pred= predict(fit , newdata = list(age =age.grid),se=T)
plot(age ,wage , col =" gray ")
lines (age.grid ,pred$fit ,col ="blue ",lwd =2)
lines (age.grid , pred$fit +2* pred$se,col ="blue" ,lty =2)
lines (age.grid ,pred$fit -2* pred$se ,col ="blue",lty =2)
abline(v=c(25 ,40, 60 ),lty=2)

fit2=lm( wage~ ns(age ,df =4) ,data= Wage)
 pred2 = predict (fit2 , newdata =list (age =age.grid),se=T)
lines (age.grid , pred2$fit ,col =" red ",lwd =2)
lines (age.grid , pred2$fit + pred2$se,col =" red ",lty=2)
lines (age.grid ,pred2$fit -2* pred$se ,col =" red ",lty=2)

legend("topright", inset=0.02, legend=c("Natural Cubic Spline ","Cubic Spline"),
```

```
lwd=2, col=c("red", "blue"))
```

شکل 3.2

```
library(splines)
x <- seq(0, 2, length=50)
B=bs(x, knots=1,Boundary.knots=c(0,2), intercept=TRUE, degree=1)
matplot(x, B,type = "l", col =2:5, lwd = 2)
matplot(x, B,type = "l", col =c("black","red","blue"), lwd = 2)
```

شکل 4.2

```
n=10
xr=seq(0,n,by=0.1)
yr=sin(xr/2)+rnorm(length(xr))/2
db=data.frame(x=xr,y=yr)
plot(db)
par(mfrow=c(3,1))
B=bs(xr,knots=1:9 ,Boundary.knots=c(0,10),degre=1)
matplot(xr,B,type="l")
title(main="Bsplines of Order 1",font.main=1)
B=bs(xr,knots=1:9,Boundary.knots=c(0,10),degre=2)
matplot(xr,B,type="l")
title(main="Bsplines of Order 2",font.main=1)
B=bs(xr,knots=1:9,Boundary.knots=c(0,10),degre=3)
matplot(xr,B,type="l")
title(main="Bsplines of Order 3",font.main=1)
```

شکل 5.2

```
library(monreg)
op <- par(mfrow=c(1,2))
y <- c(1:3,5,9,7:3,2*(2:5),rep(15,4))
plot(y, col.main = 1)
x <- seq(1,length(y), len=201)
s02 <- smooth.spline(y, spar = 0.1)
```

```
lines(predict(s02, x), col = 2)
title(main="spar=0.1")
y <- c(1:3,5,9,7:3,2*(2:5),rep(15,4))
plot(y, col.main = 2)
x <- seq(1,length(y), len=201)
s02 <-smooth.spline(y, spar = 0.8)
lines(predict(s02, x), col = 2)
title(main="spar=0.8")
```

شکل 6.2

```
library(MASS) # data sets
library(splines) # B-splines
library(mgcv)
knots_eq <- function(x, k, m)
{
c(min(x) - ((k-1):0) * (max(x)-min(x))/(m+1),
seq(from=min(x), to=max(x), length.out=m+2)[-c(1,m+2)],
max(x) + (0:(k-1)) * (max(x)-min(x))/(m+1))
}
# functions for polynomial regression
polynom <- function(x, p)
{
if (p==0) { return("1") }
if (p >0) {
return(paste(polynom(x=x, p=p-1), " + I(", x, "^",
p, ")", sep="")) }
}
polynom_m <- function(x, p)
{
if (p==0) { return(as.matrix(rep(1,length(x)))) }
if (p >0) { return(cbind(polynom_m(x=x,p=p-1),x^p)) }
}
# functions for information criteria
AIC <- function(res, H)
```

```
{
log(sum(res^2)/length(res)) + 2*sum(diag(H))/length(res)
}
SIC <- function(res, H)
{
log(sum(res^2)/length(res)) +sum(diag(H))*log(length(res))/length(res)
}
CV <- function(res, H)
{
sum((res/(1-diag(H)))^2)
}
GCV <- function(res, H)
{
sum(res^2)/(1-sum(diag(H))/length(res))^2
}
normed <- function(x) { (x-min(x))/(max(x) - min(x)) }
times <- mcycle$times
accel <- mcycle$accel

k <- 4
m <- 20
D <- matrix(0, nrow=m+k-2, ncol=m+k)
for (j in 1:(m+k-2))
{
d_j <- c(rep(0,j-1),1,-2,1,rep(0,(m+k)-3-(j-1)))
e_j <- c(rep(0,j-1), 1 ,rep(0,(m+k)-3-(j-1)))
D <- D + e_j%*%t(d_j)
}
D
y_star <- c(accel, rep(0,m+k-2))
y_star
X_spl <- splineDesign(x=times, knots=knots_eq(times,k,m), ord=k)
X_spl
ls <-c(0,1,50,10^10)
```

```
cols_l <- rainbow(length(ls))
cols_l
par(mai=c(0.65,0.6,0.1,0.1), mgp=c(2,1,0))
plot(times, accel, xlab="time", ylab="acceleration", pch=16)
for (ll in 1:length(ls))
{
l <- ls[ll]
X_star <- rbind(X_spl, sqrt(l)*D)
est <- lm(y_star ~ -1 + X_star)
plot(function(x) splineDesign(x=x, knots=knots_eq(times, k, m), ord=k) %*%
est$coef,
from=min(times), to=max(times), add=TRUE, lwd=2, col=cols_l[ll])
}
legend("bottomright", inset=0.02, col=cols_l, lwd=2, ncol=2, cex=0.8,
legend=parse(text=paste("lambda==", ls, sep="")))
```

شکل 1.3

```
lambdas <- c(10^{-5}, 1:400/100)
AICs <- rep(NA, length(lambdas))
SICs <- AICs
GCVs <- AICs
CVs <- AICs
for (ll in 1:length(lambdas))
{
l <- lambdas[ll]
X_star <- rbind(X_spl, sqrt(l)*D)
est <- lm(y_star ~ -1 + X_star)
Hat <- X_spl %*% solve(t(X_star)%*%X_star) %*% t(X_spl)
Res <- accel - X_spl%*% est$coef
AICs[ll] <- AIC(res=Res, H=Hat)
SICs[ll] <- SIC(res=Res, H=Hat)
GCVs[ll] <- GCV(res=Res, H=Hat)
CVs[ll] <- CV(res=Res, H=Hat)
}
```

```
par(mai=c(0.65,0.4,0.1,0.1), mgp=c(2,1,0))
plot(range(lambdas), 0:1, type="n", yaxt="n", xlab=expression(lambda),ylab="")
axis(2, at=0.5, label="(scaled) AIC, SIC, CV, GCV", tick=FALSE, line=-0.5)
SCs <- c("AICs", "SICs", "CVs", "GCVs")
for (SC in 1:length(SCs))
{
points(lambdas, normed(get(SCs[SC])),
type="l", lwd=2, col=rainbow(4)[SC])
abline(v=lambdas[which(get(SCs[SC])==min(get(SCs[SC])))],col=1)
}
legend("topright", inset=0.02,
legend=c(paste("AIC (", lambdas[which(AICs==min(AICs))], ")", sep=""),
paste("SIC (", lambdas[which(SICs==min(SICs))], ")", sep=""),
paste("CV (", lambdas[which(CVs ==min(CVs ))], ")", sep=""),
paste("GCV (", lambdas[which(GCVs==min(GCVs))], ")", sep="")),
col=rainbow(4), lwd=2, bg="white")
```

شکل 2.3 و 3.3

```
library(MASS) # data sets
library(splines) # B-splines
library(mgcv)
knots_eq <- function(x, k, m)
{
c(min(x) - ((k-1):0) * (max(x)-min(x))/(m+1),
seq(from=min(x), to=max(x), length.out=m+2)[-c(1,m+2)],
max(x) + (0:(k-1)) * (max(x)-min(x))/(m+1))
}
# functions for polynomial regression
polynom <- function(x, p)
{
if (p==0) { return("1") }
if (p >0) {
return(paste(polynom(x=x, p=p-1), " + I(", x, "^",
p, ")", sep="")) }
```

```
}
polynom_m <- function(x, p)
{
if (p==0) { return(as.matrix(rep(1,length(x)))) }
if (p >0) { return(cbind(polynom_m(x=x,p=p-1),x^p)) }
}
# functions for information criteria
AIC <- function(res, H)
{
log(sum(res^2)/length(res)) + 2*sum(diag(H))/length(res)
}
SIC <- function(res, H)
{
log(sum(res^2)/length(res)) +sum(diag(H))*log(length(res))/length(res)
}
CV <- function(res, H)
{
sum((res/(1-diag(H)))^2)
}
GCV <- function(res, H)
{
sum(res^2)/(1-sum(diag(H))/length(res))^2
}
normed <- function(x) { (x-min(x))/(max(x) - min(x)) }
times <- mcycle$times
accel <- mcycle$accel
m <- 3
ks <-4
#cols_k <- rainbow(length(ks))
par(mai=c(0.65,0.6,0.1,0.1), mgp=c(2,1,0))
plot(times, accel, xlab="time", ylab="acceleration", pch=16)
abline(v=knots_eq(times, 1, m), lty=c(1,rep(2,m),1), col="grey60")
for (kk in 1:length(ks))
{
```

```
k <- ks[kk]
est <- lm(accel ~ -1 +
splineDesign(x=times, knots=knots_eq(times, k, m), ord=k))
plot(function(x) splineDesign(x=x, knots=knots_eq(times, k, m), ord=k) %*%
est$coef,
from=min(times), to=max(times), n=1001, lwd=2, col=cols_k[kk], add=TRUE)
}
legend("bottomright", inset=0.02, legend=c(paste("k = ",ks,sep=""), "knots=3"),
lwd=2, lty=c(rep(1,length(ks)),2), col=c(cols_k, "grey60"), bg="white")
m <-30
ks <- 4
#cols_k <- rainbow(length(ks))
par(mai=c(0.65,0.6,0.1,0.1), mgp=c(2,1,0))
plot(times, accel, xlab="time", ylab="acceleration", pch=16)
abline(v=knots_eq(times, 1, m), lty=c(1,rep(2,m),1), col="grey60")
for (kk in 1:length(ks))
{
k <- ks[kk]
est <- lm(accel ~ -1 +
splineDesign(x=times, knots=knots_eq(times, k, m), ord=k))
plot(function(x) splineDesign(x=x, knots=knots_eq(times, k, m), ord=k) %*%
est$coef,
from=min(times), to=max(times), n=1001, lwd=2, col=cols_k[kk], add=TRUE)
}
legend("bottomright", inset=0.02, legend=c(paste("k = ",ks,sep=""), "knots=30"),
lwd=2, lty=c(rep(1,length(ks)),2), col=c(cols_k, "grey60"), bg="white")
```

شکل 4.3

```
library ( ISLR)
attach (Wage )
agelims = range ( age )
age.grid =seq ( from= agelims [1], to= agelims [2])
library(splines)
plot(age ,wage , col =" gray ")
```

```
fit2=lm( wage~ ns(age ,df =4) ,data= Wage)
 pred2 = predict (fit2 , newdata =list (age =age.grid),se=T)
lines (age.grid , pred2$fit ,col =" red ",lwd =2)
abline(v=c(33.75 ,42.00, 51.00 ),lty=2)
lines (age.grid , pred$fit +2* pred$se,col =" red ",lty=2)
lines (age.grid ,pred$fit -2* pred$se ,col =" red ",lty=2)
dim (ns(age , knots =c(25 ,40 ,60) ))
 dim (ns(age ,df =4))
attr(ns(age,df =4) ,"knots")
```

شکل 5.3

```
set.seed(11)
x <- runif(100,-2,2)
# generate the y response
y <- 2*x^3 + x^2 - 2*x +5 + rnorm(100)
plot(x,y)
xy <- data.frame(x=x, y=y)
library(ggplot2)
# specify the maximum polynomial degree that will be explored
max.poly <- 7
# cretaing data.frame which will store model predictions
# that will be used for the smooth curves in Fig. 1
x.new <- seq(min(x), max(x), by=0.1)
degree <- rep(1:max.poly, each=length(x.new))
predicted <- numeric(length(x.new)*max.poly)
new.dat <- data.frame(x=rep(x.new, times=max.poly),
                      degree,
                      predicted)
# fitting lm() polynomials of increasing complexity
# (up to max.degree) and storing their predictions
# in the new.dat data.frame
for(i in 1:max.poly)
{
  sub.dat <- new.dat[new.dat$degree==i,]
```

```
  new.dat[new.dat$degree==i,3] <- predict(lm(y~poly(x, i)),
                                          newdata=data.frame(x=x.new))
}
# plotting the data and the fitted models
p <- ggplot()
p + geom_point(aes(x, y), xy, colour="darkgrey")#darkgrey
p + geom_line(aes(x, predicted,colour=as.character(degree)) , new.dat)
p + scale_colour_discrete(name = "Degree")
p
```

شکل 6.3

```
# creating empty data.frame that will store
# AIC and BIC values of all of the models
AIC.BIC <- data.frame(criterion=c(rep("AIC",max.poly),
                                  rep("BIC",max.poly)),
                      value=numeric(max.poly*2),
                      degree=rep(1:max.poly, times=2))
# calculating AIC and BIC values of each model
for(i in 1:max.poly)
{
  AIC.BIC[i,2] <- AIC(lm(y~poly(x,i)))
  AIC.BIC[i+max.poly,2] <- BIC(lm(y~poly(x,i)))
}
a <- data.frame(cross=numeric(max.poly))
for(i in 1:max.poly)
{
  a[i,1] <- crossvalidate(x, y, degree=i)
}
# plotting AIC and BIC against model complexity
# (which is the polynomial degree)
AIC.plot <- qplot(degree, value, data=AIC.BIC,
                  geom="line", linetype=criterion) +
                  xlab("Polynomial degree") +
                  ylab("Criterion value") +
```

```
                labs(title="Information theory & Bayes")+
                geom_segment(aes(x=3, y=400,
                                 xend=3, yend=325),
                arrow = arrow(length = unit(0.3, "cm"),
                angle=20, type="closed")) +
                theme(legend.position=c(0.8,0.5))
AIC.plot
```

شکل 7.3

```
# function that will perform the "leave one out"
# crossvalidation for a y~poly(x, degree) polynomial
crossvalidate <- function(x, y, degree)
{
  preds <- numeric(length(x))
    for(i in 1:length(x))
    {
        x.in <- x[-i]
        x.out <- x[i]
        y.in <- y[-i]
        y.out <- x[i]
        m <- lm(y.in ~ poly(x.in, degree=degree) )
        new <- data.frame(x.in = seq(-3, 3, by=0.1))
        preds[i]<- predict(m, newdata=data.frame(x.in=x.out))
    }
  # the squared error:
  return(sum((y-preds)^2))
}
# crossvalidating all of the polynomial models
# and storing their squared errors in
# the "a" object
# plotting crossvalidated squared errors agains
# model complexity
cross.plot <- qplot(1:max.poly,cross, data=a, geom=c("line"))+
```

```
                    xlab("Polynomial degree") +
                    ylab("Squared error") +
                    geom_segment(aes(x=3, y=400,
                                     xend=3, yend=200),
                    arrow = arrow(length = unit(0.3, "cm"),
                    angle=20, type="closed")) +
                    labs(title="Crossvalidation")
cross.plot
```

کد مربوط به مثال داده‌های اجاره

```
library(gamlss)
PPP <- par(mfrow=c(2,2))
plot(R~Fl, data=rent, col=gray(0.7), pch=15, cex=0.5)
plot(R~A, data=rent, col=gray(0.7), pch=15, cex=0.5)
plot(R~H, data=rent, col=gray(0.7), pch=15, cex=0.5)
plot(R~loc, data=rent, col=gray(0.7), pch=15, cex=0.5)
par(PPP)
r1 <- gamlss(R ~ Fl+A+H+loc, family=NO, data=rent, trace=FALSE)
plot(r1)
r2 <- gamlss(R ~ Fl+A+H+loc, family=GA, data=rent)
plot(r2)
r3 <- gamlss(R ~ pb(Fl)+pb(A)+H+loc, family=GA, data=rent,trace=FALSE)
drop1(r3)
wp(r3, ylim.all=.6)
r4 <- gamlss(R ~ pb(Fl)+pb(A)+H+loc, sigma.fo=~pb(Fl)+pb(A)+H+loc,
family=GA, data=rent, trace=FALSE)
wp(r4, ylim.all=.6)
r6 <- gamlss(R ~ pb(Fl)+pb(A)+H+loc, sigma.fo=~pb(Fl)+pb(A)+H+loc,
nu.fo=~1, family=BCCGo, data=rent, trace=FALSE)
r7 <- gamlss(R ~ pb(Fl)+pb(A)+H+loc,sigma.fo=~pb(Fl)+pb(A)+
H+loc,nu.fo=~pb(Fl)+pb(A)+H+loc, family=BCCGo, data=rent,trace=FALSE)
wp(r6, ylim.all=.6) ; title("r6: BCCG(mu, sigma)")
wp(r7, ylim.all=.6) ; title("r7: BCCG(mu, sigma, nu)")
```

کد مربوط به مثال داده‌های گونه‌های ماهی

```
library(gamlss)
data(species)
# creating the log(lake)
species <- transform(species, x=log(lake))
plot(fish~x,data=species)
mSI<-gamlss(fish~poly(x,2),data=species, sigma.fo=~1, nu.fo=~x,
family=SICHEL, n.cyc=60, trace=FALSE)
plot(fish~log(lake), data=species)
lines(species$x[order(species$lake)], fitted(mSI)[order(
species$lake)], col="red")
m9 <- gamlss(fish~poly(x,2), nu.fo=~x, data=species, family=SICHEL,
trace=FALSE)
wp(m9)
```

کد مربوط به داده‌های مدت بستری

```
data(aep)
prop<-with(aep, noinap/los)
par(mfrow = c(2, 2))
plot(prop~age, data=aep, cex=los/30)
plot(prop~sex,data=aep)
plot(prop~ward,data=aep)
plot(prop~year,data=aep)
```

کد مربوط به داده‌های فیلم

```
library(gamlss)
data(film90)
#attach(film90)
names(film90)
## [1] "lnosc" "lboopen" "lborev1" "dist"
par(mfcol = c(1,2))
with(film90, plot(lnosc,lborev1,pch=c(21,24)[unclass(dist)],
bg=c("red","lightgray")[unclass(dist)],
xlab="log no of screens", ylab="log extra revenue", main="(a)"))
legend("bottomright",legend=c("Independent","Major"),pch=c(21,24),
```

```
pt.bg=c("red","lightgray"),cex=0.7)
with(film90, plot(lboopen,lborev1,pch=c(21,24)[unclass(dist)],
bg=c("red","lightgray")[unclass(dist)],
xlab="log opening revenue", ylab="log extra revenue", main="(b)"))
legend("bottomright",legend=c("Independent","Major"),pch=c(21,24),
pt.bg=c("red","lightgray"),cex=0.7)
install.packages("rgl")
library(rgl)
with(film90, plot3d(lboopen, lnosc, lborev1,
col=c("red","green3")[unclass(dist)]))
install.packages("gamlss.add")
library(gamlss.add)
m6 <- gamlss(lborev1~ga(~te(lboopen,lnosc))+dist, data=film90,trace=FALSE)
wp(m6, ylim.all=1.1)
library(mgcv)
plot(getSmo(m6))
vis.gam(getSmo(m6),theta = 0, phi = 30)
m7<- gamlss(lborev1~ga(~te(lboopen,lnosc))+dist,
sigma.fo=~ga(~te(lboopen,lnosc))+dist,
data=film90, trace=FALSE)
wp(m7, ylim.all=1.1)
m8 <- gamlss(lborev1 ~ pb(lboopen)+pb(lnosc) + dist,
sigma.fo = ~ pb(lboopen)+pb(lnosc) + dist,
nu.fo = ~ pb(lboopen)+pb(lnosc) + dist,
tau.fo = ~ pb(lboopen)+pb(lnosc) + dist,
family = BCPE, data = film90, trace=FALSE)
m9 <- gamlss(lborev1 ~ ga(~te(lboopen,lnosc)) + dist,
sigma.fo = ~ ga(~te(lboopen,lnosc)) + dist,
nu.fo = ~ ga(~te(lboopen,lnosc)) + dist,
tau.fo = ~ ga(~te(lboopen,lnosc)) + dist,
family = BCPE, data = film90, n.cyc=20, trace=FALSE)
par(mfrow=c(1,2))
wp(m8, ylim.all=0.5)
wp(m9, ylim.all=0.5)
```

```
wp(m9, xvar=~lboopen+lnosc, ylim.worm=1)
```

کد مربوط به داده‌های جرم و جنایت

```
library(gamlss.spatial)
data(columb)
data(columb.polys)
m1 <- gamlss(crime~ gmrf(district, polys=columb.polys), data=columb)
draw.polys(columb.polys, getSmo(m1), scheme="topo")
```

مراجع

[۱] بازرگان لاری ع.ر ،(۱۳۸۵)، **"رگرسیون خطی کاربردی"** چاپ دوم، انتشارات دانشگاه شیراز،

[۲] نیرومند ح.ع ،(۱۳۸۴)، **"الگوهای خطی تعمیم‌یافته: با کاربردهای آن در علوم و مهندسی"** انتشارات دانشگاه فردوسی مشهد،

[3] Agresti, A. (2015), " **Foundations of linear and generalized linear models**", John Wiley, Sons.

[4] Burnham, K. P and Anderson, D. R. (2002), "Model selection and multimodel inference: A practical information-theoretic approach",**Springer**.

[5] De Boor, C. (1977), "Package for calculating with B-splines", **SIAM Journal on Numerical Analysis**, 14(3), 441-472.

[6] De Boor, C. (1978), "**A Practical Guide to Splines** ",Springer, Berlin.

[7] Dierckx, P. (1993), "**Curve and Surface Fitting with Splines**", Oxford University Press.

[8] Eilers, P.H and Marx, B.D. (1996), " Flexible smoothing with B-splines and penalties", **Statistical science**, 11, 89-102.

[9] Eilers, P.H., Marx, B.D. and Durbán, M. (2015), "Twenty years of P-splines", **statistics and operations research transactions**, 39(2), 149-186

[10] Ghosh, A. and Kindermann, D.D.S. (2010), "**Efficient thin plate spline interpolation and its application to adaptive optics**".

[11] Green, P.J. and Silverman, B.W. (1994), "**Nonparametric regression and generalized linear models**" , CRC Press.

[12] Hastie, T.J and Tibshirani, R.J. (1990), "**Generalized additive models**". Chapman Hall, London.

[13] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), ”**An introduction to statistical learning**”, springer.

[14] Kagerer, K. (2013), ”’**A short introduction to splines in least squares regression analysis**”, University of Regensburg, Faculty of Business, Economics and Management Information Systems.

[15] Langville, A.N. and Stewart, W.J. (2004), ”The Kronecker product and stochastic automata networks”, **computational and applied mathematics**, 167(2), 429-447.

[16] Lambaert, H. (2006), ”**Manual registration with thin plates**”.

[17] Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012), ”Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting”, **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, 61(3), 403-427.

[18] McCullagh, P. and Nelder, J.A. (1989), ”**Generalized linear models**”, Chapman Hall, London.

[19] McKinley, S. and Levine, M. (1998), ”Cubic spline interpolation”, **College of the Redwoods**, 45(1), 1049-1060.

[20] Marra, G. and Radice, R. (2010),” Penalised regression splines: theory and application to medical research”,**Statistical Methods in Medical Research**, 19(2), 107-125.

[21] Nelder, J.A. (1972), ”Generalized linear models”, **Statistical Society Series A** , 135, 370-384.

[22] O’Sullivan, F. (1986), ”A statistical perspective on ill-posed inverse problems”, **Statistical science**, 1, 502-518.

[23] O’Sullivan, F. (1988), ”Fast computation of fully automated log-density and log-hazard estimators”, **SIAM Journal on scientific and statistical computing**, 9(2), 363-379.

[24] Rigby, R.A. and Stasinopoulos, D.M. (2005), ”Generalized additive models for location, scale and shape”, **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, 54(3), 507-554.

[25] Ruppert, D., Wand, M.P. and Carroll, R.J. (2009), ”Semiparametric regression during 2003–2007”, ’**Electronic journal of statistics**, 3, 1193-1256.

[26] Ruppert, D. (2002), ”Selecting the number of knots for penalized splines”, **Journal of computational and graphical statistics**, 11(4), 735-757.

[27] Ruppert, D. and Carroll, R.J. (2000), ”Theory Methods: Spatially adaptive Penalties for Spline Fitting”, **Australian New Zealand Journal of Statistics**, 42(2), 205-223.

[28] Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V. and De Bastiani, F. (2017), ”**Flexible regression and smoothing: using gamlss in r**”, CRC Press.

[29] Stein, G.Z. and Juritz, J.M. (1988), ”Linear models with an inverse gaussian poisson error distribution”, **Communications in Statistics-Theory and Methods**, 17(2), 557-571.

[30] Strang, G. and Aarikka, K. (1986), ”**Introduction to applied mathematics**”, Wellesley-Cambridge Press Wellesley, MA.

[31] Stoer, J. and Bulirsch, R. (2013),” **Introduction to numerical analysis**”, Springer Science Business Media.

[32] Wahba, G. (1990), **Spline models for observational data**, SIAM, Philadelphia.

[33] Wand, M.P. and Ormerod, J.T. (2008), ”On semiparametric regression with O’Sullivan penalized splines”, **Australian New Zealand Journal of Statistics**, 50(2), 179-198.

[34] Wood, S.N. (2006), **Generalized additive models: an introduction with R**, CRC Press.

[35] Wood, S.N. (2003) ”Thin plate regression splines”, **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, 65(1), 95-114.

[36] Wassermann, L. (2006),” All of nonparametric statistics”, **New York**.

[37] Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J. and Stasinopoulos, D. (2012),” Modelling skewness and kurtosis with the BCPE density in GAMLSS”, **Journal of Applied Statistics**, 39(6), 1279-1293.

Aabstract

Regression analysis is, in fact, the most widely used method amongst statistical techniques and a statistical tool for identifying the relationship between one or more covariates with a response variable. When a linear regression model is not efficient for analyzing a data set, we can use non-parametric regression methods. Splines as one of the interpolation tools are a powerful non-parametric method for modeling nonparametric regression. Splines construct curves that include polynomials of the same degree on substructures of a specified interval and join together with defined contiguous conditions, and pass through common knots between the two infrastructures. During the years since the introduction of Splines, their theoretical foundations have been developed, and different versions have been introduced. Penalized Splines (P-Splines) are among the most widely used tools for nonparametric modeling and smoothing problems. The appearance of P-Splines with many changes in smoothing problems has become a running and active field of research. Due to the broad applications and importance of Splines, we are going to introduce them and their growth as well as the development process and some of their features.

Keywords: Penalized Splines, Smoothing, Generalized Linear Models, Generalized Additive Models, Akaike Information Criterion.

Shahrood University of Technology

Faculty Of Mathematical Sciences

MSc Thesis in: Statistics

the spline penalized and new application

By: Akram Ghaemi zadeh

Supervisor

Dr. Negar Eghbal

Advisor

Dr. Hossein Baghishani

July 2018