مدل‌بندی داده‌های تابعی با رویکرد رگرسیون مؤلفه اصلی بر اساس معیار اعتبار سنجی متقابل تعمیم‌یافته

مهدی روزبه[۱]، آرتا روحی[۲]، فاطمه جهادی [۳]

تاریخ دریافت: ۱۴۰۰/۰۴/۱۲
تاریخ پذیرش: ۱۴۰۲/۰۲/۰۸

چکیده:
تحلیل داده‌های تابعی برای توسعه رویکردهای آماری در داده‌هایی مورداستفاده قرار می‌گیرد که دارای ماهیت تابعی و پیوسته هستند و چون این توابع به فضاهای با بعد بی‌نهایت تعلق دارند، استفاده از روش‌های متداول در آمار کلاسیک برای تحلیل آن‌ها، با چالش روبرو است. مشهورترین تکنیک تحلیل داده‌های آماری، رویکرد مؤلفه‌های اصلی تابعی می‌باشد که ابزاری مهم برای کاهش بعد است، در این مقاله با استفاده از روش رگرسیون مؤلفه اصلی تابعی بر اساس جریمه مشتق دوم، ریج و لاسو به تحلیل داده‌های تابعی آب‌وهوای کانادا و داده‌های تابعی طیف‌سنج پرداخته خواهد شد. بدین منظور برای تعیین مقدار بهینه پارامتر جریمه در روش‌های مورداستفاده از اعتبار سنجی متقابل تعمیم‌یافته، که معیاری معتبر و کارآمد است، استفاده می‌گردد.

واژه‌های کلیدی: اعتبار سنجی متقابل تعمیم‌یافته، رگرسیون تابعی، رگرسیون مؤلفه اصلی، تحلیل داده‌های تابعی.

۱ مقدمه

در داده‌های تابعی نخستین گام تبدیل داده‌های گسسته به پیوسته است که بدین منظور از روش‌های هموارسازی مانند پایه‌هایی فوریه برای داده‌های دارای دوره تناوب و اسپلاین بهره گرفته خواهد شد و سپس به برآورد ضرایب رگرسیونی در آن پرداخته می‌شود. در ادامه چون در عمل داده‌ها از قسمتی عددی و قسمتی تابعی تشکیل‌شده‌اند، تعدادی از مدل‌های رگرسیون تابعی بیان می‌شوند که به‌طور هم‌زمان بتوان پیش‌بینی‌های تابعی و غیرتابعی متعدد را کنترل و با جریمه مناسب، عامل‌های مهم خطا در پیش‌بینی را شناسایی کرد. به دلیل پرکاربرد بودن، روش‌های مختلفی ارائه‌شده است که هسته اصلی این روش‌ها، کاهش بعدهایی از پیش‌بینی‌کننده‌ها یا تنظیم منظم تابع ضریب در قالب یک جریمه ناهموار است که برای اطلاعات بیشتر می‌توان به هوروث و کوکوزکا [۵]، هزینگ و اوبانک [۶] و در اواخر به کوکوزکا و ریمهر [۷]، فبرو بانده و همکاران [۸] و ریس و همکاران [۱۱] مراجعه کرد. در این مقاله علاوه بر رگرسیون مؤلفه اصلی تابعی از جریمه‌های کمینه کردن طول منحنی مشتق دوم، جریمه‌ی لاسو و ریج استفاده شده است.

۱.۱ روش مؤلفه‌های اصلی

روش مؤلفه‌های اصلی، از روش‌های کاربردی با ایده‌ی کاهش ابعاد و حفظ بیشترین اطلاعات ممکن از متغیرهای توضیحی است. در این روش به دنبال ترکیب خطی از متغیرهای توضیحی هستیم که واریانس را به بیشترین مقدار خود برساند. در این روش از مقادیر ویژه و بردارهای ویژه ماتریس واریانس-کوواریانس یا ماتریس ٔ همبستگی استفاده می‌شود [۱۲]. از دیدگاه هندسی، این روش تبدیل خطی متعامدی است که داده‌ها را از یک دستگاه مختصات به دستگاه مختصات جدید می‌برد به‌نحوی‌که بزرگ‌ترین واریانس داده‌ها بر روی محور مختصات اول، دومین بزرگ‌ترین واریانس بر روی محور مختصات دوم قرار می‌گیرد و به همین ترتیب ادامه می‌یابد. در نتیجه با این کار متغیرها با کاهش بعد به مؤلفه‌هایی تبدیل می‌شوند که مؤلفه‌های اصلی نامیده می‌شوند. مؤلفه‌های اصلی ضمن ناهمبستگی به نحوی سازمان‌دهی می‌شوند که تعداد کمی از مؤلفه‌ها بتوانند درصد قابل‌توجهی از تغییرات متغیرهای توضیحی اولیه را توجیه کنند. انتخاب تعداد مناسب مؤلفه‌ها مورد توجه است و روش‌های مختلفی برای تعیین تعداد مناسب این مؤلفه‌ها پیشنهادشده که در زیر به برخی از آن‌ها اشاره می‌شود. یکی از روش‌های

[۱] هیئت‌علمی گروه آمار، دانشگاه سمنان، سمنان، ایران (نویسنده مسئول: mahdi.roozbeh@semnan.ac.ir).
[۲] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.
[۳] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.

تعیین تعداد مؤلفه‌های اصلی انتخاب شماری از مؤلفه‌هاست که بتوانند درصد قابل‌توجهی از واریانس (تغییرات کل) را توجیه کنند. برای مثال اگر ۱۰۰ متغیر توضیحی داشته باشیم و با اتخاذ ۲۰ مؤلفه ۷۰ درصد تغییرات واریانس کل و با اتخاذ ۴۰ مؤلفه ۷۵ درصد تغییرات واریانس کل توجیه شود، اضافه کردن ۲۰ مؤلفه دیگر برای دست یافتن به ۵ درصد تغییرات بیشتر به‌صرفه نخواهد بود. روش دیگر، برگزیدن تعداد مؤلفه‌هایی است که واریانس آن‌ها بزرگ‌تر یا مساوی متوسط واریانس کل است. توجه شود که اگر از ماتریس همبستگی استفاده می‌کنیم مؤلفه‌هایی که واریانس آن‌ها بزرگ‌تر یا مساوی ۱ است، برگزیده می‌شوند. از روش‌های شهودی نیز برای این امر استفاده از نمودار بازو [۴] می‌باشد. در این نمودار مقادیر ویژه هر مؤلفه (λ_i) در برابر i رسم می‌شود. با عمود کردن ناحیه‌ای از نمودار که شیب آن به‌طور ناگهانی کم می‌شود، تعداد مؤلفه‌ها تعیین می‌گردد. البته می‌توان لگاریتم (λ_i) در برابر i رسم کرد. استفاده از لگاریتم به‌این‌علت است که تابع لگاریتم مقیاس را کاهش می‌دهد. البته استفاده از این روش یا روش نمودار بازو تفاوت چندانی ندارد، گر چه نمودار بازو جهش‌های ناگهانی را بهتر نشان می‌دهد. توجه به این نکته ضروری است که ممکن است هرکدام از روش‌های مطرح‌شده پاسخ‌های متفاوتی ارائه دهند که محقق می‌تواند بنا بر کاربرد خود روش موردنظر خود را به کار گیرد. در استفاده از این روش باید به پایا بودن مقیاس متغیرهای توضیحی اولیه توجه نمود چراکه مؤلفه‌ی اصلی در مواجه با متغیر دارای واریانس بالا بسیار وابسته آن می‌شود که باعث خطای زیاد در کار می‌شود و در این حالت استفاده از ماتریس همبستگی به علت پایایی می‌تواند جلوی این مشکل را بگیرد. بعد از انتخاب تعداد متغیرها نوبت به برازش مدل رگرسیون مؤلفه‌های اصلی می‌شود. مدل رگرسیونی دیگر با مشکل هم خطی روبرو نیست و می‌توان از روش کمترین توان‌های دوم معمولی برای برآورد ضرایب آن استفاده کرد. از کاربردهای روش مؤلفه‌های اصلی می‌توان به تبدیل متغیرهای همبسته به متغیرهای ناهمبسته، یافتن ترکیبات خطی با واریانس نسبی بزرگ یا کوچک و کاهش در حجم داده‌ها نام برد.

۲.۱ روش رگرسیونی لاسو

رویکرد رگرسیون لاسو [۵] نخستین بار توسط لئو برایمن [۶] با نام یقه فلزی (برای خفه کردن محکوم) [۷] بیان شد [۱۳]. نخستین بار مسئله بهینه‌سازی نامنفی او به‌صورت زیر مطرح شد:

$$\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p}c_j\hat{\beta}_jX_{ji}\right)^2 \qquad c_j\geq 0,\quad \sum_j c_j\leq s \quad (1)$$

که در آن برآوردگر اولیه $\hat{\beta}_j$ با روش کمترین توان‌های دوم و s در آن پارامتر جریمه است که با کاهش آن گروتی محدود می‌شود. این روش در بین محققان به نام جریمه مرگ [۸] نیز معروف است. در این روش برخی از متغیرها حذف‌شده و بقیه آن‌ها منقبض می‌شود. از مزایای این روش این است که خطای آن در مقایسه با خطای رگرسیون بهترین مجموعه کمتر است و فقط هنگامی‌که ضرایب کوچک غیر صفر دارد، باید در استفاده از این روش احتیاط لازم انجام شود. با توجه به این مسئله، لاسو تابع هدفی ارائه می‌دهد که در آن از برآوردگرهای کمترین توان‌های دوم استفاده نمی‌کند که به‌صورت زیر نوشته می‌شود:

$$\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}\left(Y_i-\sum_{j=1}^{p}X_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\},$$
$$i=1,\ldots,n,\quad j=1,\ldots,p$$

از مزایای اثبات‌شده این روش این است که برآوردگرهای پایدار و پیوسته تولید می‌کند[۱۳]. همچنین این روش برخی از ضرایب را منقبض و بقیه را صفر می‌کند. از معایب روش برآورد لاسو می‌توان به اریبی و عملکرد ضعیف آن در حضور متغیرهای توضیحی همبسته اشاره کرد زیرا که بین متغیرهای همبسته به‌ویژه متغیرهایی که به‌صورت گروهی همبسته‌اند یکی از آن‌ها را انتخاب می‌کند و متغیر انتخاب‌شده لزومی ندارد که بهترین متغیر باشد. لاسو برای یک مدل رگرسیونی خطی با p متغیر توضیحی و n مشاهده، حداکثر n متغیر را انتخاب می‌کند، بنابراین هنگامی‌که متغیرهای بیشتری (بیشتر از n) در مدل معنی‌دار باشند، در این روش شانسی برای انتخاب ندارند [۲].

[4] Scree Diagram
[5] Lasso
[6] Leo Breiman
[7] Garrote
[8] Death Penalty
[9] Andrey Nikolayevich Tikhonov

۲ رگرسیون ریج

محققان در مسائل رگرسیونی باهم خطی متغیرها روبرو می‌شوند که معمولاً در مدل‌هایی با تعداد پارامتر بالا اتفاق می‌افتد. آندری نیکولایویچ تیخونوف [۹] که تحقیقات مهم او در توپولوژی، تجزیه‌وتحلیل عملکردی، فیزیک، ریاضیات شناخته شده است، منظم‌سازی تیخونوف را به‌عنوان راه‌حلی برای رویارویی با مسائل بدشرطی معرفی نمود. رگرسیون خطی می‌تواند نااریب باشد اما واریانس بالایی داشته باشد که ممکن است بعضی از اریبی‌ها را از دست بدهد تا واریانس کمتری داشته باشد. در داده‌های با بعد بالا چون ماتریس طرح وارون‌پذیر نیست، روش کمترین توان‌های دوم رگرسیون خطی نمی‌تواند مفید باشد. رگرسیون خطی زمانی که تعداد متغیرها بیشتر از داده‌ها باشد تعریف نمی‌شود. ذر این شرایط با استفاده از روش ریج می‌توان معکوس ماتریس را به دست آورد، بدین‌صورت که با افزودن مقدار λ به درایه‌های قطر اصلی ماتریس $\boldsymbol{X}^\top\boldsymbol{X}$ این ماتریس معکوس‌پذیر خواهد شد[۱۴]. اگر نوک کوه نقطه صفر در نظر گرفته شود، با افزایش انقباض متغیرها، ضرایب کوچک و کوچک‌تر شده که همانند کاهش عرض کوه و افزایش ارتفاع آن می‌باشد و ضرایب هیچ‌گاه دقیقاً مقدار صفر را نمی‌پذیرند. برآوردگر ریج دیگر نااریب نیست اما دارای واریانس کمتری می‌باشد و مقدار برآوردگر پایا است یعنی تغییرات جزئی در داده‌ها، بر برآوردگر تأثیری نخواهد داشت [۳]. برآورد $\boldsymbol{\beta}$ به روش ریج عبارت است از:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{Y}, \qquad k \geq 0 \qquad (۲)$$

که به k پارامتر ریج می‌گوییم و یافتن این پارامتر بسیار مهم می‌باشد. برای انتخاب k باید به‌گونه‌ای عمل کنیم که کاهش در واریانس برآوردگر اریب بیش از افزایش مربع اریبی باشد [۱۰]، در نتیجه MSE آن کمتر از واریانس برآوردگر نااریب خواهد بود.

طبق قضیه تجزیه مقادیر منفرد ماتریس‌ها که بر اساس آن یک ماتریس به دو ماتریس متعامد و یک ماتریس قطری تجزیه می‌شود، می‌توان گفت روش ریج همواره منجر به یافتن برآورد مدل رگرسیونی خواهد شد، به‌بیان‌دیگر می‌توان نوشت:

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$$

که در آن $\boldsymbol{D}$ ماتریس قطری با عناصر قطری مقادیر ویژه‌ی ماتریس $\boldsymbol{X}^\top\boldsymbol{X}$ بوده و $\boldsymbol{V}$ و $\boldsymbol{V}^\top$ ماتریس‌های متعامد هستند، ماتریس $\boldsymbol{X}^\top\boldsymbol{X}$ نیمه معین مثبت است زیرا:

$$\forall \boldsymbol{Z} \neq 0, \boldsymbol{Z}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{Z} = (\boldsymbol{X}\boldsymbol{Z})^\top(\boldsymbol{X}\boldsymbol{Z}) = \|\boldsymbol{X}\boldsymbol{Z}\| \geq 0$$

پس عناصر قطری آن بزرگ‌تر یا مساوی صفر هستند. حال در عبارت (۲) با جایگزینی ماتریس زیر بجای ماتریس $\boldsymbol{X}^\top\boldsymbol{X}$ می‌توان نوشت:

$$\boldsymbol{X}^\top\boldsymbol{X} + k\boldsymbol{I} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top + \boldsymbol{V}k\boldsymbol{I}\boldsymbol{V}^\top = \boldsymbol{V}(\boldsymbol{D} + k\boldsymbol{I})\boldsymbol{V}^\top$$

در عبارت فوق $\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}$ است. پس ماتریس مربوطه وارون‌پذیر است. برای k های مختلف داریم:

۱- اگر $k = 0$ باشد، برآوردگر ریج همان برآوردگر کمترین توان‌های دوم معمولی می‌شود.

۲-اگر $k = \infty$ آنگاه $\hat{\boldsymbol{\beta}} = 0$ است.

۳-اگر k بین صفر و بی‌نهایت باشد، مسئله در حال تعادل بین دو روش برازش کردن یک مدل خطی روی متغیرها و کوچک کردن ضرایب است.

۳ برآورد مدل تابعی

در ابتدا به برآورد یک منحنی یا خط ساده برای برازش روی داده‌ها نیاز است و چون داده‌ها به‌صورت گسسته می‌باشند، باید داده‌های گسسته را به داده‌های پیوسته تبدیل کرد. اگر داده‌ها بدون خطا باشند، از روش درون‌یابی استفاده می‌شود و اگر دارای خطا باشند، از روش هموارسازی می‌توان بهره گرفت. چون اکثراً داده‌ها دارای خطا هستند، پس باید از روش هموارسازی استفاده کرد.

۴ تبدیل داده‌های گسسته به پیوسته

با توجه به فرم قرارگیری داده‌ها، رگرسیون خطی به‌صورت زیر تعریف می‌شود:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \epsilon_i \qquad (۳)$$

و اگر به فرم منحنی باشند، به‌صورت زیر نوشته می‌شود:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \cdots + \epsilon_i \qquad (۴)$$

به‌طورکلی اگر متغیر وابسته به‌صورت یك ترکیب خطی از توابع پایه و ضرایب مربوطه باشد، آنکه مدل رگرسیونی به فرم زیر قابل‌نمایش است:

$$Y_i = \sum_{j=1}^{k} c_j\phi_j(t_i) + \epsilon_i = f(t_i) + \epsilon_i \qquad (۵)$$

به‌طوری‌که در معادله بالا $\phi_j(t)$ تابع پایه و وابسته به نوع داده‌ها می‌باشند و c_j ها ضرایب می‌باشند. در ادامه مثال‌هایی از توابع پایه آورده خواهند شد.

۱.۴ مثال‌هایی از توابع پایه

مرسوم‌ترین توابع پایه عبارت‌اند از:

۱. ثابت [10]: $\phi(t) = ۱$

۲. توانی [11]: $\phi(t) = t^{\lambda_۱}, t^{\lambda_۲}, t^{\lambda_۳}, \ldots$

۳. نمایی [12]: $\phi(t) = e^{\lambda_۱ t}, e^{\lambda_۲ t}, e^{\lambda_۳ t}, \ldots$

۲.۴ تقریب توابع پایه با استفاده از پایه‌های فوریه

تابع‌های تقریب‌زن پایه‌های فوریه اکثراً برای داده‌هایی کاربرد دارد که نوسانی بوده و دارای دوره تناوب باشند، همانند داده‌های مربوط به آب‌وهوا. پایه‌های فوریه به‌صورت زیر تعریف می‌شود [۱۵]:

$$\{۱, \sin(\omega t), \cos(\omega t), \sin(۲\omega t), \cos(۲\omega t), \ldots, \sin(m\omega t), \cos(m\omega t)\}$$

در تابع پایه بالا ω دوره‌ی نوسان نامیده شده و برابر است با:

$$\omega = \frac{۲\pi}{p}$$

که p دوره تکرارشونده است. برای مثال دوره تکرارشونده برای داده‌های آب‌وهوا برابر با $p = ۳۶۵$ روز است.

۳.۴ تقریب توابع پایه با استفاده از پایه‌های اسپلاین

پایه‌های اسپلاین [13] یا بی اسپلاین [14] هر دو نقطه‌ی مجاور را با یک تابع برآورد می‌کند. اسپلاین‌ها همانند تابع‌های چندجمله‌ای هستند که ابتدا داده‌های گسسته را به چند قسمت مساوی تقسیم کرده و سپس به دنبال بهترین منحنی برای برازش به هر قسمت می‌باشد که اگر درجه آن صفر باشد با یک خط عمودی و افقی به برآورد می‌پردازد و اگر درجه‌ی آن یک باشد به‌صورت خطی و درجه‌های بالاتر را به‌صورت منحنی برآورد می‌کند. در ضمن قسمت‌هایی از منحنی که در محل پیوستن به هم هستند را می‌توان هموار کرد، نقاطی را که در قسمت اتصال قرار می‌گیرند گره نامیده می‌شود.

برای انتخاب گره‌ها، اگر تعداد آن‌ها خیلی زیاد باشد اریبی بسیار کم و واریانس بسیار زیاد می‌شود و باعث ناهمواری نمودار می‌شود [۱۵].

نکته‌ای هم که باید در نظر داشت این است که باید در هر گره حداقل یک داده وجود داشته باشد.

۴.۴ برآورد ضرایب تابعی

تابع رگرسیونی به‌صورت

$$Y_i = f(t_i) + \epsilon_i$$

را در نظر می‌گیریم که در آن خطاها مستقل و دارای توزیع نرمال با میانگین صفر و واریانس $\sigma^۲$ و $i = ۱, \ldots, n$ می‌باشند. برای برآورد $f(t_i)$ بر اساس توابع پایه داریم:

$$\hat{f}(t_i) = \sum_{j=۱}^{p} c_j \phi_j(t_i)$$

برای برآورد ضرایب تابعی باید مجموع مربع خطاها را به‌صورت زیر حداقل کرد:

$$\boldsymbol{H}(\boldsymbol{c}) = \sum_{i=۱}^{n} (Y_i - f(t_i))^۲ = \sum_{i=۱}^{n} (Y_i - \sum_{j=۱}^{p} c_j \phi_j(t_i))^۲ \quad (۶)$$

همچنین می‌توان معادله‌ی فوق را به‌صورت زیر نوشت:

$$\boldsymbol{H}(\boldsymbol{c}) = (\boldsymbol{Y} - \boldsymbol{\Phi c})^\top (\boldsymbol{Y} - \boldsymbol{\Phi c}) \quad (۷)$$

اکنون با مشتق‌گیری و برابر صفر قرار دادن آن داریم:

$$\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-۱} \boldsymbol{\Phi}^\top \boldsymbol{Y} \quad (۸)$$

حال با توجه به برآورد بالا مقدار $\hat{f}(\boldsymbol{t})$ برابر است با:

$$\hat{f}(\boldsymbol{t}) = \hat{\boldsymbol{c}}^\top \boldsymbol{\Phi}$$

و همچنین $\hat{\boldsymbol{Y}}$ به‌صورت:

$$\hat{\boldsymbol{Y}} = \underbrace{\boldsymbol{\Phi}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-۱} \boldsymbol{\Phi}}_{\boldsymbol{S}} \boldsymbol{Y} = \boldsymbol{S}\boldsymbol{Y}$$

به دست می‌آید که به $\boldsymbol{S}$ ماتریس هموارساز می‌گویند. در عبارت فوق مقدار $\hat{\boldsymbol{Y}}$ ضریبی از مقدار $\boldsymbol{Y}$ است، یعنی با هموارکردن داده‌ها توسط $\boldsymbol{S}$ مقدار $\hat{\boldsymbol{Y}}$ به دست می‌آید. انتخاب تعداد توابع پایه بسیار اهمیت دارد، اگر تعداد این توابع کم در نظر گرفته شود منجر به اریبی زیاد و واریانس کم شده و زیاد در نظر گرفتن تعداد این توابع منجر به اریبی کم و واریانس زیاد می‌شود.

[10] Constant
[11] Power
[12] Exponential
[13] Spline
[14] B-Spline

۵ جریمه کمینه کردن طول مشتق دوم

با داشتن متغیر تابعی $X(t)$ و با تعریف مشتق‌های توابع می‌توان منحنی‌هایی هموار داشت. در ابتدا داریم:

$$DX(\boldsymbol{t}) = \frac{d}{dt}X(\boldsymbol{t})$$

و می‌توان مشتقات مراتب بالاتر را به‌صورت زیر تعریف کرد:

$$D^{۲}X(\boldsymbol{t}) = \frac{d^{۲}}{dt^{۲}}X(\boldsymbol{t}), \ldots, D^{k}X(\boldsymbol{t}) = \frac{d^{k}}{dt^{k}}X(\boldsymbol{t})$$

که $D^{۲}\boldsymbol{X}(t)$ را می‌توان طبق تعریف مشتق دوم به‌صورت زیر نوشت:

$$D^{۲}X(t_i) = \frac{Y_{i+۱} - Y_{i-۱} - ۲Y_i}{(\Delta t)^{۲}}$$

اکنون طول منحنی را با توجه به فرمول زیر می‌توان نوشت:

$$j_{۲}(X(\boldsymbol{t})) = \int \left(D^{۲}X(\boldsymbol{t})\right)^{۲} \mathrm{d}t \qquad (۹)$$

اگر طول منحنی کم باشد بدین معنی است که به نوسان‌های داده‌ها کمتر توجه شده و اگر اندازه منحنی زیاد باشد، پس باید جریمه افزایش یابد، به بیانی دیگر باید مشتق دوم ضرایب تابعی را هموار کرد تا منحنی این ضرایب هموار شود. با در نظر گرفتن مربع خطای جریمه‌شده به فرم زیر داریم:

$$PENSSE_{\lambda}(X(\boldsymbol{t})) = (\boldsymbol{Y} - X(\boldsymbol{t}))\,(\boldsymbol{Y} - X(\boldsymbol{t}))^{\top} + \lambda J_{۲}(X(\boldsymbol{t}))$$

به λ پارامتر هموارساز گفته می‌شود که کنترل بین مقادیر برازش شده و همواری منحنی بر داده‌ها می‌باشد. اگر مقدار λ افزایش یابد، هموارکننده بسیار جریمه می‌شود و $X(\boldsymbol{t})$ تبدیل به خط می‌شود و اگر λ کاهش یابد، جریمه هم کاهش می‌یابد و اجازه می‌دهد $X(\boldsymbol{t})$ روی داده‌ها برازش شود. باید مربع خطای جریمه‌شده را حداقل کرد و این امر در پایه‌های بی اسپلاین زمانی می‌افتد که مقدار درجه آن برابر ۴ باشد، زیرا دارای مشتق دوم بوده و می‌توان مشتق دوم را هموار کرد و منحنی‌های هموار مناسب‌تری ایجاد کرد. هنگامی‌که درجه بی‌اسپلاین برابر ۴ باشد به آن اسپلاین هموار مکعب گفته می‌شود. اکنون با تعریف $X(\boldsymbol{t}) = \boldsymbol{\Phi}^{\top}\boldsymbol{c}$ و با توجه به اینکه $X(\boldsymbol{t})$ یک ترکیب خطی می‌باشد داریم:

$$\int [D^{m}X(\boldsymbol{t})]^{۲}\,\mathrm{d}t = \boldsymbol{c}^{\top}\underbrace{\int D^{m}\boldsymbol{\Phi}D^{m}\boldsymbol{\Phi}^{\top}\mathrm{d}t}_{\boldsymbol{R}}\boldsymbol{c} = \boldsymbol{c}^{\top}\boldsymbol{R}\boldsymbol{c}$$

در معادله‌ی فوق $\boldsymbol{R}$ را به‌عنوان ماتریس جریمه می‌شناسیم. با حداقل کردن مربع خطاها و اندازه طول مشتق دوم، مقدار برآورد $\boldsymbol{c}$ به‌صورت زیر خواهد بود:

$$\hat{\boldsymbol{c}} = \left[\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi} + \lambda\boldsymbol{R}\right]^{-۱}\boldsymbol{\Phi}^{\top}\boldsymbol{Y}$$

و مقدار برآورد $\boldsymbol{Y}$ برابر است با:

$$\hat{\boldsymbol{Y}} = \underbrace{\boldsymbol{\Phi}\left[\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi} + \lambda\boldsymbol{R}\right]^{-۱}\boldsymbol{\Phi}^{\top}}_{\boldsymbol{S}}\boldsymbol{Y} = \boldsymbol{S}\boldsymbol{Y}$$

برای انتخاب پارامتر هموارساز می‌توان از مینیمم سازی معیارهای متداول زیر که برآوردهای مجانبا نااریب از تابع ریسک با زیان توان دوم هستند، استفاده کرد [۱۶]:

۱. اعتبار سنجی متقابل معمولی [۱۵] :

$$OCV(\lambda) = \frac{۱}{n}\sum_{i=۱}^{n}\frac{(Y_i - \hat{Y}_i)^{۲}}{(۱ - S_{ii})^{۲}} \qquad (۱۰)$$

که در آن S_{ii} مؤلفه iام روی قطر اصلی ماتریس $\boldsymbol{S}$ است.

۲. اعتبار سنجی متقابل تعمیم‌یافته [۱۶] :

$$GCV(\lambda) = \frac{۱}{n}\sum_{i=۱}^{n}\frac{(Y_i - \hat{Y}_i)^{۲}}{[n^{-۱}trace(\boldsymbol{I}_n - \boldsymbol{S})]^{۲}}$$

که در آن $\boldsymbol{I}_n$ ماتریس همانی از مرتبه n است. براي ديدن جزئيات بيشتر اين معيار مي‌توان به [۱، ۴، ۹] مراجعه كرد.

۶ تحلیل داده‌های واقعی

۱.۶ داده‌های طیف‌سنجی

در این قسمت کلیه تحلیل‌ها در نرم‌افزار R انجام‌شده است و برای انجام آن نیاز به فراخوانی کتابخانه‌های fda، fda.usc، goffda و refund می‌باشد.

در این بخش به بررسی داده‌هایی با متغیر مستقل تابعی و متغیر پاسخ عددی با بعد پایین و همچنین انجام رگرسیون مؤلفه‌های اصلی با جریمه‌های متفاوت و مقایسه آن‌ها پرداخته می‌شود. در ابتدا داده‌های tecator از کتابخانه fda.usc را فراخوانی می‌کنیم. این داده‌ها شامل ۱۰۰ طیف جذب مادون‌قرمز نزدیک (NIT)[۱۷] است که برای پیش‌بینی مقدار رطوبت، چربی و پروتئین برای تیکه‌های گوشت خرد شده استفاده

[15] Ordinary Cross Validation
[16] Generalized Cross Validation
[17] Near Infrared Transmission

می‌شود. این داده‌ها بر اساس آنالیز غذا و خوراک Tecator Infratec می‌باشد که در محدوده طول‌موج ۱۰۵۰ – ۸۵۰ نانومتر کار می‌کند و توسط اصل انتقال مادون‌قرمز نزدیک (NIT) ثبت می‌شود. هر نمونه حاوی گوشت خالص ریز خردشده با رطوبت، چربی و پروتئین متفاوت است. هدف پیش‌بینی محتوای چربی می‌باشد. این داده‌ها برای هر نمونه گوشت از ۱۰۰ کانال طیف تشکیل‌شده است که به اندازه‌گیری جذب و محتویات رطوبت، چربی و پروتئین خواهیم پرداخت. مدل تابعی برای داده‌های بالا به فرم زیر نوشته می‌شود:

$$Fat = \beta_0(t) + \sum_{i=1}^{100} \beta_i(t) X_i(t) \qquad (11)$$

در مدل فوق Fat میزان چربی موجود در هر قسمت گوشت خالص و $X_i(t)$ اطلاعات ثبت‌شده توسط امواج مادون‌قرمز با طول‌موج ۸۵۰–۱۰۵۰ می‌باشد و t برابر ۱۰۰ طیف کانال می‌باشد. با تقسیم داده‌ها و اختصاص دادن هشتاد درصد آن‌ها به داده‌های آموزش و بیست درصد داده‌ها به داده‌های آزمون به بررسی مدل روی داده‌های آموزش می‌پردازیم. اکنون با توجه به عددی بودن متغیر پاسخ و تابعی بودن متغیر مستقل می‌توان با استفاده از رگرسیون با سه مؤلفه اصلی مدل زیر را در نظر گرفت:

$$Fat = \alpha_1(t) PC_1(t) + \alpha_2(t) PC_2(t) + \alpha_3(t) PC_3(t) \qquad (12)$$

به‌طوری‌که در مدل فوق $PC_i, \quad i = 1,2,3$، مؤلفه‌های اصلی برای داده‌های مذکور می‌باشند. بدون هیچ‌گونه جریمه و بر اساس رگرسیون مؤلفه اصلی، مقدار مجذور همبستگی بین مقادیر برازش شده و مقادیر واقعی ۰٫۶۱ می‌باشد باقیمانده و باقیمانده استاندارد با مقادیر برازش شده را می‌توان در شکل ۱ مشاهده کرد. با استفاده از اعتبارسنجی متقابل تعداد ۳ مؤلفه اصلی نقش اساسی دارند. حال جریمه ریج و جریمه لاسو و جریمه مشتق دوم را در رگرسیون مؤلفه اصلی با انتخاب متغیرهای اول و سوم و چهارم و مقدار λ بهینه که توسط اعتبارسنجی متقابل به‌دست‌آمده است، با استفاده از جریمه مشتق دوم مقدار همبستگی بین مقادیر برازش شده و مقادیر واقعی عدد ۰٫۸۸ می‌باشد و نمایش نمودار باقیمانده و باقیمانده استاندارد با مقادیر برازش شده را می‌توان در شکل‌های ۲ و ۳ برای مدل‌های جریمه‌ای پیشنهادی مشاهده کرد. اکنون به بررسی برآورد ضرایب تابعی در مدل‌های فوق پرداخته می‌شود. با توجه به شکل ۴ نمودار بالا سمت چپ برآورد ضرایب تابعی مدل بدون جریمه می‌باشد، نمودار بالا سمت راست، نمودار برآورد ضرایب تابعی با جریمه ریج که با دخالت ۳ متغیر اصلی و نمودار پایین سمت چپ برآورد ضرایب تابعی با جریمه مشتق دوم، نمودار پایین سمت راست برآورد ضرایب تابعی برای مدل رگرسیون مؤلفه اصلی با جریمه لاسو می‌باشد.

اکنون داده‌های آزمون را در مدل‌ها بررسی و سپس به انتخاب یکی از سه مدل یعنی مدل بدون جریمه، مدل با جریمه مشتق دوم و مدل با جریمه ریج می‌پردازیم. می‌توان در شکل ۵ نمودارهای رسم شده مقادیر برازش شده در مقابل مقادیر واقعی برای داده‌های آزمون را مشاهده کرد. نتایج به‌دست‌آمده در جدول ۱ گویای آن است که با اختلاف خیلی کم مدل رگرسیون مؤلفه اصلی با جریمه لاسو مناسب‌تر نسبت به سایر جریمه‌ها عمل کرده است.

۲.۶ داده‌های آب‌وهوای کانادا

در این نوع داده دمای هوا و میزان بارندگی ۳۵ شهر کانادا در ۳۶۵ روز از سال مشخص است. حال باید به دنبال رگرسیون مؤلفه اصلی بین دمای این ۳۵ شهر با میزان بارندگی آن‌ها بود. با فراخوانی این داده‌ها از کتابخانه FDA می‌توان مدل تابعی زیر را در نظر گرفت:

$$Y_i = \beta_0(t) + \sum_{i=1}^{35} \beta_i(t) X_i(t) \qquad (13)$$

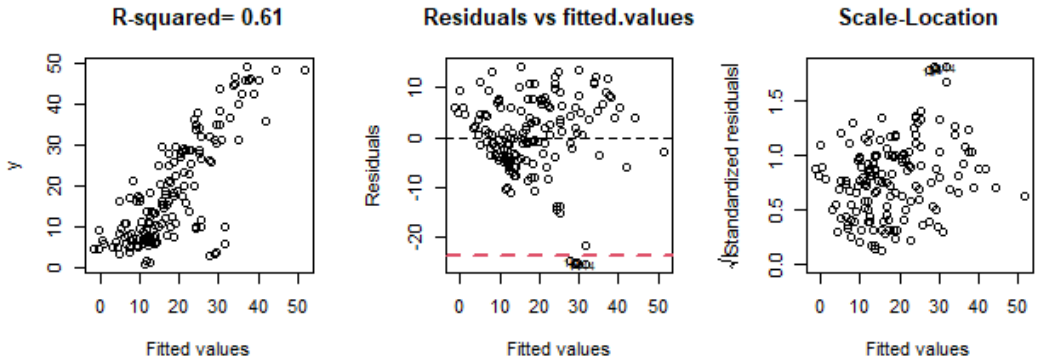
به‌طوری‌که در مدل بالا Y_i میزان بارندگی برای ۳۵ ایستگاه آب‌وهوایی می‌باشد و $X_i(t)$ میزان دمای ۳۵ ایستگاه در طول ۳۶۵ روز می‌باشد و $\beta(t)$ ضرایب تابعی مدل می‌باشند. با استفاده از توابع پایه فوریه متغیرهای پیش‌بین را می‌توان تبدیل به منحنی کرد:

$$X_i(t) = \sum_{i=1}^{n} c_i \phi_i(t)$$

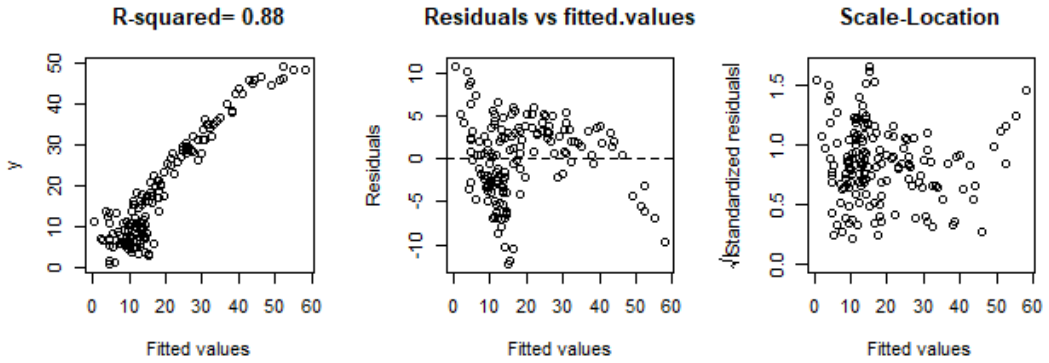
در ابتدا به تعریف ۳۵ تابع پایه برای تبدیل داده‌های گسسته دمای هوا به پیوسته توسط توابع پایه فوریه پرداخته می‌شود و دلیل استفاده از توابع فوریه تناوبی بودن داده‌ها می‌باشد که منحنی آن را می‌توان در نمودار ۶ مشاهده کرد. اکنون حالت‌های مختلف رگرسیون مؤلفه اصلی بدون جریمه، با جریمه مشتق دوم و جریمه لاسو را بررسی می‌شود. مدل رگرسیون مؤلفه اصلی بر اساس سه مؤلفه اول بدون در نظر گرفتن جریمه به‌صورت زیر می‌باشد:

$$Y_i = \alpha_1(t) PC_1(t) + \alpha_2(t) PC_2(t) + \alpha_3(t) PC_3(t) \qquad (14)$$

در معادله فوق Y_i میزان بارندگی برای ۳۵ ایستگاه آب‌وهوایی می‌باشد. پس از برازش، مقدار همبستگی بین مقادیر واقعی و مقادیر برازش شده عدد ۰٫۳۶ می‌باشد که حاکی از آن است که روش بدون جریمه مناسب عمل‌نکرده است. اکنون به دنبال برآوردی مناسب‌تر با استفاده از اعتبارسنجی متقابل برای انتخاب λ و انتخاب تعداد مؤلفه‌های



شکل ۱: مدل رگرسیون مؤلفه اصلی بدون جریمه، نمودار سمت چپ: پراکنش مقادیر واقعی در مقابل مقادیر برازش شده، نمودار وسط: باقیمانده‌ها در مقابل مقادیر برازش شده و نمودار سمت راست: ریشه قدرمطلق باقیمانده‌ها در مقابل مقادیر برازش شده.

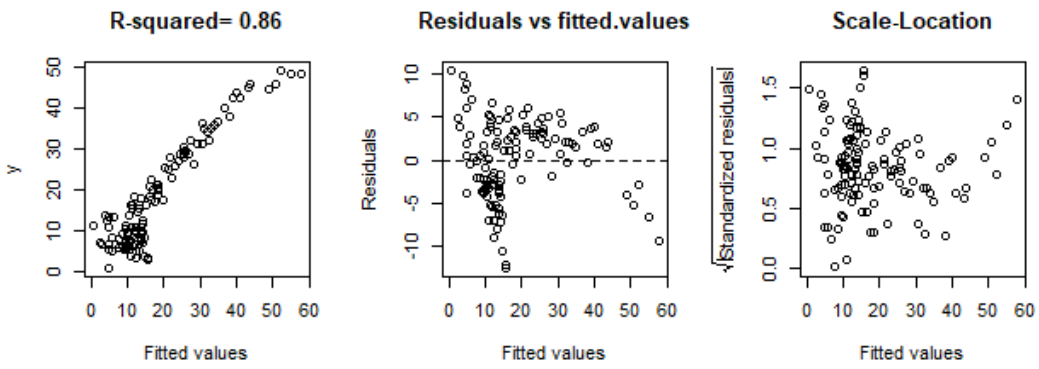


شکل ۲: مدل رگرسیون مؤلفه اصلی با جریمه مشتق دوم، نمودار سمت چپ: پراکنش مقادیر واقعی در مقابل مقادیر برازش شده، نمودار وسط: باقیمانده‌ها در مقابل مقادیر برازش شده و نمودار سمت راست: ریشه قدرمطلق باقیمانده‌ها در مقابل مقادیر برازش شده در مدل جریمه‌شده.

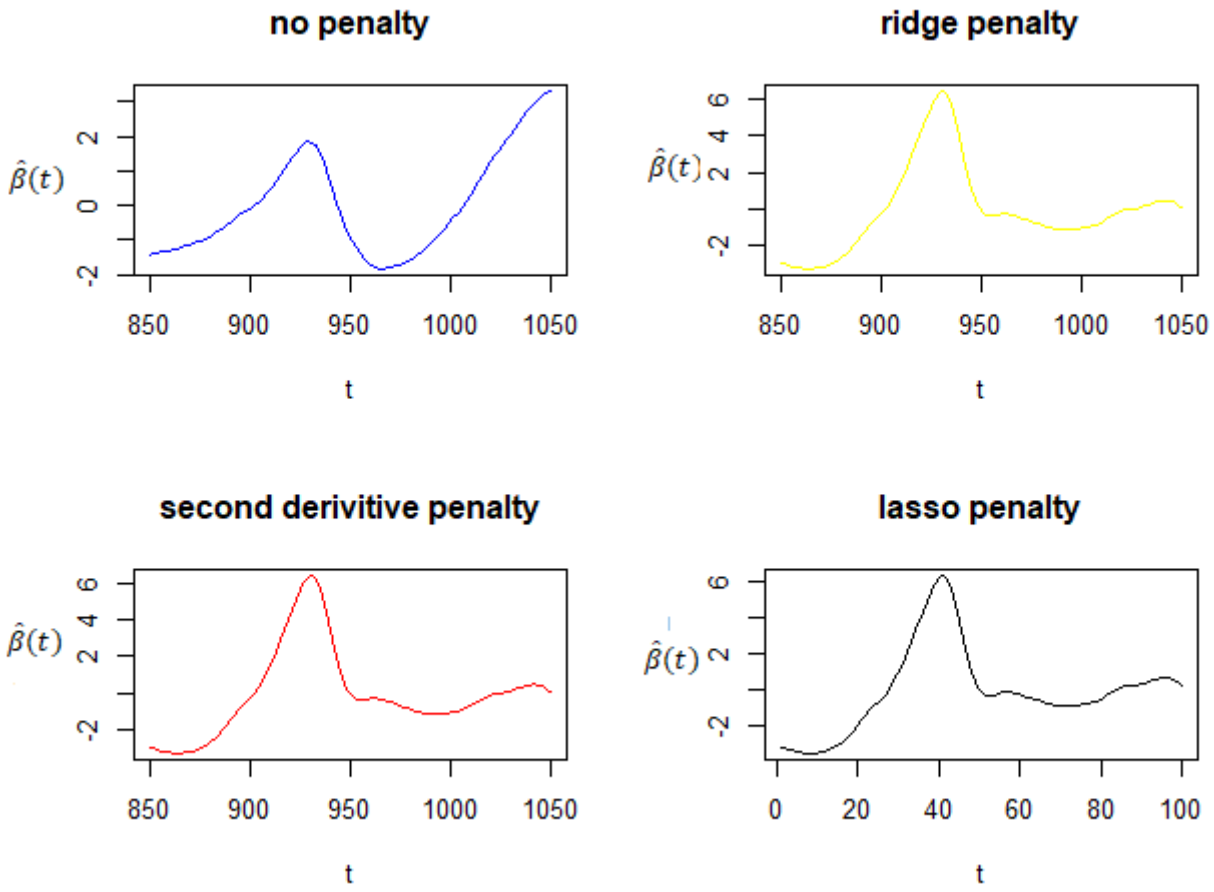
اصلی و جریمه کردن ضرایب تابعی می‌باشیم. با توجه به جریمه مشتق دوم صورت گرفته و عدد همبستگی ۰٫۸۶ مشخص است که مدل فوق مناسب‌تر شده است. اکنون با جریمه لاسو هم این مدل‌سازی انجام شد که مقدار λ را با استفاده از اعتبارسنجی متقابل به دست آورده و با توجه به عدد همبستگی بین مقادیر برازش شده و مقادیر واقعی می‌توان گفت این روش نسبتاً خوب بوده و می‌توان نمودار رسم شده بین داده‌های واقعی و مقادیر برازش شده را در تصویر ۷ مشاهده کرد. نمودار برآورد ضرایب تابعی را برای دو رگرسیون مؤلفه اصلی با جریمه مشتق دوم و جریمه لاسو می‌توان در تصویر ۸ مشاهده کرد. ضرایب مؤلفه‌های اصلی برای سه مدل فوق به‌صورت مشخص‌شده در جدول ۳ می‌باشد. با توجه به آزمون مجذور همبستگی به‌دست‌آمده در جدول ۲ مشخص است که مدل رگرسیون با جریمه مشتق دوم عملکرد مناسبی داشته است.

۷ بحث و نتیجه‌گیری

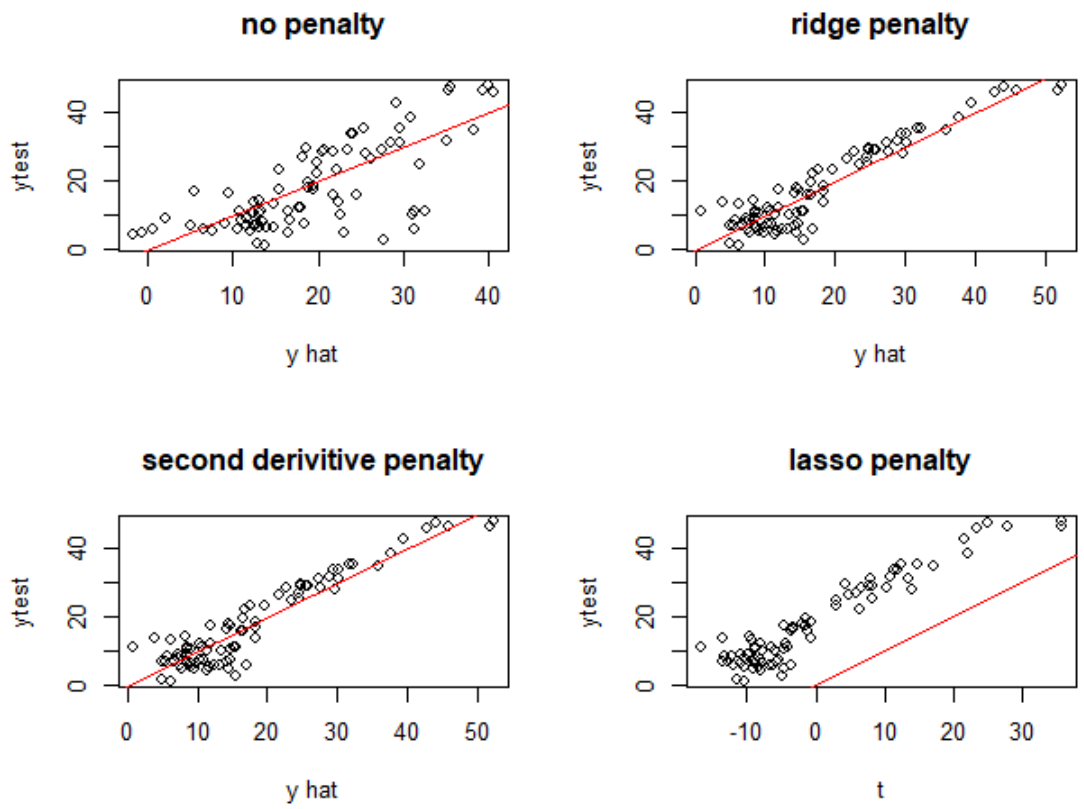
با بررسی داده‌های tecator با استفاده از رگرسیون مؤلفه اصلی و با جریمه‌های کمینه کردن طول منحنی مشتق دوم، جریمه لاسو و ریج به مدل‌سازی پرداخته شد که با توجه به نتایج به‌دست‌آمده رگرسیون مؤلفه اصلی با جریمه لاسو و ریج با میزان همبستگی بین مقادیر مشاهده‌شده و مقادیر برازش شده ۰٫۹۴۹۱۰۰۳ مدلی مناسب‌تر بود. در داده‌های آب‌وهوای کانادا دو جریمه ریج و مشتق دوم انجام‌شده است و نتایج با



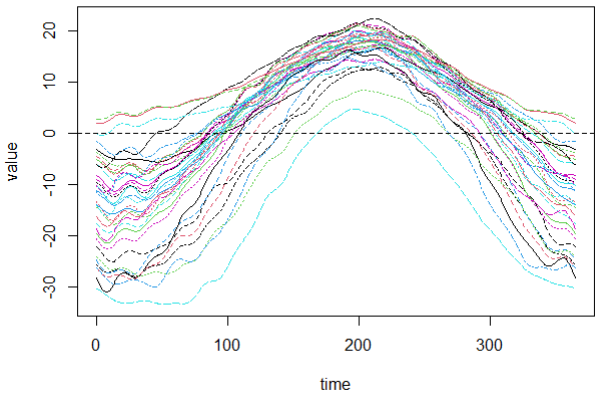
شکل ۳: مدل رگرسیون مؤلفه اصلی با جریمه ریج، نمودار سمت چپ: پراکنش مقادیر واقعی در مقابل مقادیر برازش شده، نمودار وسط: باقیمانده‌ها در مقابل مقادیر برازش شده و نمودار سمت راست: ریشه قدرمطلق باقیمانده‌ها در مقابل مقادیر برازش شده در مدل جریمه‌شده.



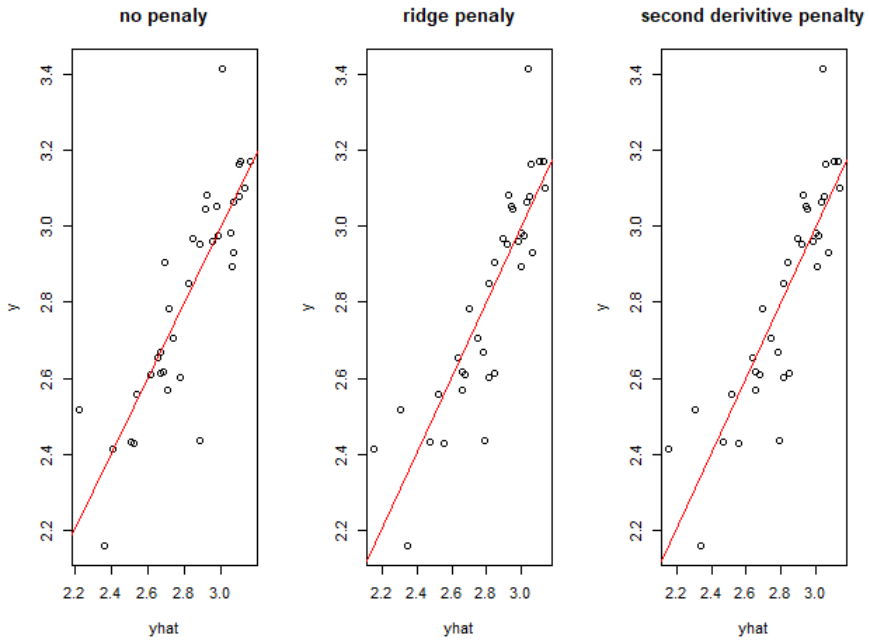
شکل ۴: نمودار برآورد ضرایب تابعی با جریمه‌های متفاوت



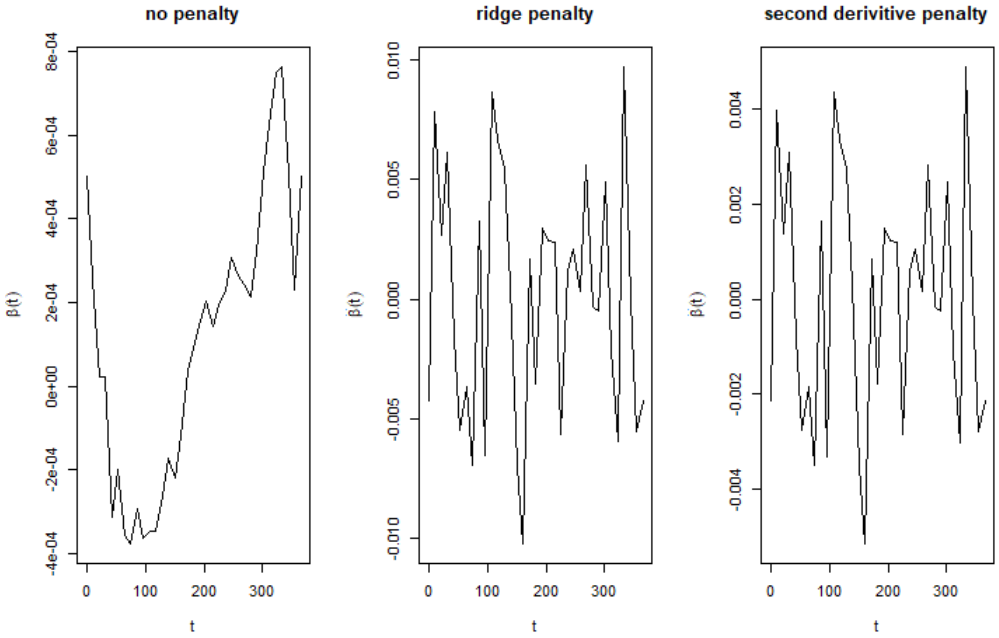
شکل ۵: نمودار مقادیر برازش شده در مقابل مقادیر واقعی برای داده‌های آزمون که نمودار بالا سمت چپ بدون جریمه، نمودار بالا سمت راست با جریمه ریج، نمودار پایین سمت چپ با جریمه مشتق دوم و نمودار پایین سمت راست با جریمه لاسو می‌باشد.



شکل ۶: نمایش منحنی دمای هوای ۳۵ ایستگاه آب‌وهوای کانادا



شکل ۷: مقادیر برازش شده و مقادیر واقعی در مدل بدون جریمه و با رگرسیون مؤلفه اصلی با جریمه لاسو و با جریمه مشتق دوم



شکل ۸: برآورد ضرایب تابعی در سه مدل متفاوت

جدول ۱: آزمون مجذور همبستگی بین مقادیر واقعی و مقادیر برازش شده در داده‌های طیف‌سنجی

نوع مدل	مجذور همبستگی
مدل رگرسیون مؤلفه اصلی بدون جریمه	۰٫۷۳۸۴۵۳۷
مدل رگرسیون مؤلفه اصلی با جریمه ریج	۰٫۹۳۹۱۷۱۳
مدل رگرسیون مؤلفه اصلی با جریمه مشتق دوم	۰٫۹۳۹۰۹۷۱
مدل رگرسیون مؤلفه اصلی با جریمه لاسو	۰٫۹۴۹۱۰۰۳

جدول ۲: آزمون مجذور همبستگی بین مقادیر واقعی و مقادیر برازش شده در داده‌های آب‌وهوای کانادا

نوع مدل	مجذور همبستگی
مدل رگرسیون مؤلفه اصلی بدون جریمه	۰٫۸۵۳۱۸۲
مدل رگرسیون مؤلفه اصلی با جریمه ریج	۰٫۸۶۸۳۹۴۲
مدل رگرسیون مؤلفه اصلی با جریمه مشتق دوم	۰٫۸۶۸۴۰۵۱

جدول ۳: در داده‌های آب‌وهوای کانادا ضرایب مؤلفه‌های اصلی

ضرایب	مؤلفه اصلی بدون جریمه	مؤلفه اصلی با جریمه مشتق دوم	مؤلفه اصلی با جریمه ریج
مؤلفه اول	−۰٫۰۰۱۹	۰٫۰۵۲۵	۰٫۰۴۸۳۵
مؤلفه دوم	−۰٫۰۰۱۸	−۰٫۰۰۵۳	−۰٫۰۰۴۹۳
مؤلفه سوم	−۰٫۰۰۶۲	۰٫۰۰۸۰	۰٫۰۰۷۳۷
عرض از مبدأ	۲٫۸۱۴۸	۲٫۸۱۴۸	۲٫۸۱۴۸

توجه به میزان همبستگی مقادیر برازش شده مدل رگرسیون مؤلفه اصلی با جریمه مشتق دوم در مقابل مقادیر مشاهده‌شده رضایت‌بخش بودند.

تقدیر و تشکر

نویسندگان مقاله کمال قدردانی و تشکر را از پیشنهادهای ارزنده داوران، سردبیر و ویراستار محترم مجله که باعث ارائه بهتر و افزایش سطح کیفی مقاله شده است، دارند.

مراجع

[۱] روزبه، م. و امینی، م. (۱۳۹۸)، برآوردگر استوار مرزبندی شده تعمیم‌یافته محتمل در مدل رگرسیون نیمه‌پارامتری، *مجله علوم آماری ایران*، ۱۳(۲)، ۴۴۱-۴۶۰.

[۲] معنوي، م. و روزبه، م. (۱۳۹۹)، روش های تحلیل رگرسیونی برای داده‌های بعد بالا، *اندیشه آماری*، ۲۵(۱)، ۶۹-۹۰.

[۳] روزبه، م.، ملك جعفریان، م. و معنوی، م. (۱۳۹۹)، کاربرد رگرسیون ستیغی کمترین توان‌های دوم پیراسته محدودشده تصادفی در مدل‌سازی مصرف آب، *اندیشه آماری*، ۲۶(۲)، ۹-۱۹.

[4] Amini, M. and Roozbeh, M. (2015), Optimal Partial Ridge Estimation in Restricted Semiparametric Regression Models, *Journal of Multivariate Analysis*,**136**, 26-40.

[5] Horveźth, L., and Kokoszka, P. (2012). *Inference for functional data with applications*, Springer Science and Business Media, New York.

[6] Hsing, T., and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, John Wiley and Sons, Hoboken.

[7] Dette, H., Kokot , K., and Aue, A. (2017). Functional data analysis in the Banach space of continuous functions, *arXiv preprint arXiv*, 1710.07781.

[8] Febrero Band, M., Galeano, P., and Gonzalez Manteiga, W. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study, *International Statistical Review* , **85(1)** , 61-83.

[9] Roozbeh, M. (2018), Optimal QR-Based Estimation in Partially Linear Regression Models with Correlated Errors Using GCV Criterion, *Computational Statistics & Data Analysis*, **117**, 45-61.

[10] Roozbeh, M. and Arashi, M. (2016),New ridge regression estimator in semiparametric regression models, *Communications in Statistics–Simulation and Computation*, **45**, 3683-3715.

[11] Reiss, P. T., Goldsmith, J., Shang, H. L. and Ogden, R. T. (2017). Methods for scalar-on-function regression, *International Statistical Review*, **85(2)** , 228-249.

[12] Jolliffe, I.T. (2002). *Principal Component Analysis*, Springer series in statistics, Aberdeen.

[13] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58(1)**, 267-288.

[14] Hoerl, A.E. and Kennard, R.W. (1975). Ridge regression some simulation, *Communication in Statistics*, **4**, 4105-4123.

[15] Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer-Verlag, New York.

[16] Wasserman, L. (2005), *All of Nonparametric Statistics*, Springer-Verlag, New York.

Modelling of functional data using principal component regression approach based on the generalized cross validation criterion

Mahdi Roozbeh[1] Arta Rouhi[2] Fatemeh Jahadi[3]

Abstract:

Functional data analysis is used to develop statistical approaches to the data sets that are functional and continuous essentially, and because these functions belong to the spaces with infinite dimensional, using conventional methods in classical statistics for analyzing such data sets is challenging. The most popular technique for statistical data analysis is the functional principal components approach, which is an important tool for dimensional reduction. In this research, using the method of functional principal component regression based on the second derivative penalty, ridge and lasso, the analysis of Canadian climate and spectrometric data sets is proceed. To do this, to obtain the optimum values of the penalized parameter in proposed methods, the generalized cross validation, which is a valid and efficient criterion, is applied.

Keywords: Functional Data Analysis, Functional Regression, Generalized Cross Validation, Principal Component Regression.

[1]**Faculty of mathematics, Semnan university, Semnan, Iran.**

[2]**Master's degree graduate, statistics and Computer science, Semnan university, Semnan, Iran.**

[3]**Master's degree graduate, statistics and Computer science, Semnan university, Semnan, Iran.**