# مقایسه‌ای بر دقت برآورد اسپلاین‌های رویهٔ نازک و کروی با رگرسیون چندگانه

روشنک علی محمدی، مهرنوش مختارپور؛ دانشگاه الزهرا، گروه آمار

دریافت ۹۴/۰۹/۱۱ پذیرش ۹۵/۰۹/۱۵

## چکیده

اسپلاین‌های رویهٔ نازک و کروی، روش‌های ناپارامتری مناسبی برای تحلیل داده‌های فضایی هستند. اسپلاین‌های رویهٔ نازک در درون‌یابی‌های فضایی راه حل‌های عملی کارا و دقت زیادی دارند. تحلیل‌ها به گسترش یک مدل آماری از تغییرات فضایی مشاهدات بستگی دارند که به‌عنوان برآورد خطا در نظر گرفته می‌شوند. ساختار خطای این مدل دو مؤلفه دارد که به‌صورت جداگانه، انحرافات فضایی داده‌ها و موقعیت‌های ناهمواری آن را در نظر می‌گیرد رگرسیون پارامتری رابطهٔ میان متغیرهای توضیحی و پاسخ را به‌صورت توابع تعیین شده مانند خطی، چندجمله‌ای، نمایی و غیره در نظر گرفته و خطای برازش را به‌صورت مجموع توان‌های دوم خطا محاسبه می‌کند. در این پژوهش با به‌کارگیری معیار خطای استاندارد باقی‌مانده‌ها، دقت روش‌های ناپارامتری اسپلاین رویهٔ نازک و اسپلاین کروی با روش پارامتری رگرسیون چندگانه به‌روش شبیه‌سازی، با به‌کارگیری نرم‌افزار R مقایسهٔ عددی می‌شود. برای این منظور داده‌های تصادفی در نرم‌افزار R تولید شده و برازش مدل به داده‌ها به سه روش مورد نظر انجام شده است. هم‌چنین به‌منظور بررسی عوامل مؤثر در برازش مدل و با این گمان که ممکن است یک روش همواره بر دیگری برتری نداشته باشد، با در نظر گرفتن هم‌بستگی‌های مختلف بین متغیرها و اندازه‌های نمونه‌ای متفاوت، تأثیر این موارد نیز بر نیکویی برازش مدل بررسی شد.

**واژه‌های کلیدی**: رگرسیون چندگانه، اسپلاین رویهٔ نازک، اسپلاین کروی، خطای استاندارد باقی‌مانده‌ها، کم‌ترین توان‌های دوم جریمه‌ای.

## مقدمه

یکی از ابزارهای مهم آماری برای تجزیه و تحلیل روابط بین متغیرها، رگرسیون است که با ایجاد الگویی میان متغیرهای توضیحی و پاسخ صورت می‌گیرد. پژوهش حاضر با تمرکز بر رگرسیون چندگانه انجام گرفته که برای به الگو درآوردن متغیر پاسخ از دو متغیر توضیحی استفاده می‌کند. رگرسیون پارامتری، رابطهٔ میان متغیرهای توضیحی و پاسخ را به‌صورت خاصی از توابع در نظر می‌گیرد، به‌قسمی که مجموع توان‌های دوم خطا را مینیمم کند. گاهی در روش‌های پارامتری با مواردی مواجه می‌شویم که مدل حاصل از این روش‌ها برازش چندان مناسبی ندارد و روند داده‌ها را به‌خوبی دنبال نمی‌کند، بنا براین کار را با یک روش ناپارامتری پیگیری می‌کنیم. به‌عنوان یک روش رگرسیون ناپارامتری، اسپلاین‌های رویهٔ نازک دقت زیادی دارند.

فرض کنید f تابعی از متغیر توضیحی چند متغیرهٔ $x=(x_1,\ldots,x_d)\in R^d$ باشد که $R^d$ فضای اقلیدسی d بعدی است. مدل رگرسیونی (۱) در نظر گرفته می‌شود:

*نویسندهٔ مسئول r_alimohammadi@alzahra.ac.ir

$$y_i = f(x_i) + \epsilon_i \quad (۱)$$

که در آن بردار $x_i = (x_{i1},\ldots,x_{id})$ و $\epsilon_i$ها خطاهای تصادفی مستقل با میانگین صفر و واریانس مشترک $\sigma^2$ هستند، اسپلاین رویهٔ نازک تعمیم روش اسپلاین هموارای است که برای حالت یک متغیر توضیحی (d=1) به‌کار می رود. در اسپلاین هموارای برای برآورد تابع f در رابطهٔ (۱) لازم است تابع مجموع توان‌های دوم جریمه‌ای به‌صورت (۲) مینیمم شود:

$$\sum(y_i - f(x_i))^2 + \lambda \int f''(x)dx \quad (۲)$$

در رابطهٔ (۲)، عبارت اول مجموع توان‌های دوم خطا و عبارت دوم، انتگرال مشتق دوم تابع f، مقدار ناهمواری مدل را نشان می‌دهد که توازن بین این دوبخش با پارامتر هموارای $\lambda$ کنترل می‌شود [۱۰].

در این تحقیق، اسپلاین رویهٔ نازک برای دو متغیر توضیحی (d=2) در نظر گرفته شده است. از این رو، می‌توان طول و عرض جغرافیایی را به‌عنوان متغیرهای توضیحی در نظر گرفت. این روش در پیش‌بینی مقادیر مورد نظر در زمینهٔ هواشناسی، مهندسی معدن و اپیدمیولوژی کاربرد دارد و امکان پیش‌بینی برای نقاطی که دسترسی به آن‌ها نداریم را فراهم می‌کنند. این نقاط در مورد داده‌های هواشناسی، می‌توانند مناطقی باشند که ایستگاه هواشناسی ندارند [۱]، [۲]. در بخش بعد، این روش به‌طور مبسوط بیان می‌شود.

اسپلاین‌های رویهٔ نازک به‌طور گسترده‌ای برای درون‌یابی فضایی متغیرهای سطوح آب و هوایی استفاده شده‌اند، که برخی از آن‌ها شامل میانگین بارش سالانه [۳]، میانگین ماهانهٔ بارش [۴]، مدل‌بندی داده‌های آب و هوا [۵] و درون‌یابی بارش روزانه [۶]، برازش اسپلاین‌های کروی به داده‌های ازون [۷] و داده‌های سالانهٔ دمای هوا [۸] است.

تحقیق پیشین، کاربردی از اسپلاین رویهٔ نازک برای داده‌های هواشناسی بوده است و تأثیر طول و عرض جغرافیایی به‌عنوان دو متغیر توضیحی بر میانگین رطوبت نسبی هوا، به‌عنوان متغیر پاسخ، با روش‌های اسپلاین رویهٔ نازک و کروی برازش و بررسی شد. این تحقیق، با ۱۴۸ نمونه از ایستگاه‌های هواشناسی سینوپتیکی در ایران، برتری اسپلاین رویهٔ نازک را در مقابل اسپلاین کروی نتیجه داد [۱۴].

هدف از این تحقیق، بررسی اسپلاین‌های رویهٔ نازک و کروی[1] با رگرسیون چندگانه در برازش مدلی مناسب در حالات مختلف و مقایسهٔ آن‌هاست. بدین‌منظور با استفاده از نرم‌افزار R و تکیه بر معیار خطای استاندرد باقی‌مانده[2] که از جذر مجموع توان‌های دوم خطا بر درجهٔ آزادی مدل به‌دست می‌آید، دقت برآورد این روش‌ها سنجیده و مقایسهٔ عددی می‌شود. نتایج به‌دست آمده از داده‌هایی است که از توزیعی نرمال با بردار میانگین صفر و ماتریس‌های کوواریانس متفاوت برای تعیین ساختار وابستگی بین متغیرها در حالت‌های مختلف ایجاد شده‌اند. در شبیه‌سازی این داده‌ها، فرض بر این شده که هم‌بستگی بین متغیرهای توضیحی صفر است این فرض به‌منظور جلوگیری از بروز هم خطی چندگانه است اما ارتباط میان هر یک از متغیرهای توضیحی با متغیر پاسخ که هدف کار در این پژوهش بررسی نوع این ارتباط است، با استفاده از تعیین ماتریس کوواریانس[3] در تولید داده‌های تصادفی قرار داده شده و کوواریانس‌های متفاوتی برای انجام مقایسه، در نظر گرفته شده است.

1. Spherical spline
2. Residual Standard Error

۳. رابطهٔ بین متغیرهای توضیحی با متغیر پاسخ در ماتریس کوواریانس مشخص می‌شود.

## اسپلاین‌های رویهٔ نازک

اسپلاین رویهٔ نازک تعمیمی از ایدهٔ مینیمم‌سازی معیار مجموع توان‌های دوم جریمه‌ای برای اسپلاین هموارى که در برازش مدل برای یک متغیر توضیحی به‌کار می‌رود، به بیش‌تر از یک بعد است. در این روش، علاوه بر معیار نیکویی برازش که در رگرسیون پارامتری بر اساس مینیم‌سازی مجموع توان‌های دوم خطا حاصل می‌شود، مقدار ناهمواری مدل حاصل نیز در نظر گرفته می‌شود و معیار مجموع توان‌های دوم جریمه‌ای به‌صورت (۳) تعریف می‌شود:

$$S(f) = ||y - f||^2 + \lambda J_m^d(f) \qquad (۳)$$

در رابطهٔ (۳)، $J_m^d(f)$ به‌صورت:

$$J_m^d(f)=\sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\ldots\alpha_d!} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \left(\frac{\partial^m f}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_d} x_d}\right)^2 \Pi_{j=1}^d dx_j$$

و پارامتر همواری $\lambda > 0$ مقدار نوسانات مدل را کنترل می‌کند. با تغییر پارامتر همواری در رابطهٔ (۲)، تابع اسپلاین رویهٔ نازک می‌تواند با گذر از تمامی نقاط و درون‌یابی دقیق داده‌ها، تا یک برازش کاملا هموار تغییر کند [۹].

یک مسئلهٔ مهم در اسپلاین رویهٔ نازک، انتخاب بهینهٔ پارامتر همواری است که از روش اعتبار متقابل تعمیم یافته[۴] برای این کار استفاده می‌شود. اعتبار متقابل تعمیم یافته، اندازه‌ای است از خطای پیش‌بینی سطح برازش شده که با حذف هر داده و تخمین آن بر اساس سایر داده‌ها و جمع بستن توان‌های دوم تفاوت هر مقدار داده با مقدار برآورد شده آن با وزن مناسب به‌دست می‌آید [۱۰]، [۱۱].

## اسپلاین‌های کروی

توابع اسپلاینی که داده‌ها را روی یک سطح کروی تخمین می‌زنند، اسپلاین کروی یا اسپلاین روی کره نامیده می‌شوند، دامنهٔ تابع در این نوع اسپلاین‌ها به‌صورت $\chi=S$ در نظر گرفته می‌شود که S کرهٔ واحد است. هر نقطهٔ x روی S با نماد $x = (\theta,\phi)$ نشان داده می‌شود که $\theta(0\leq\theta\leq2\pi)$ طول جغرافیایی و $\phi(-\frac{\pi}{2}\leq\phi\leq\frac{\pi}{2})$ عرض جغرافیایی است. در اسپلاین کروی، جریمهٔ ناهمواری (که معادل عبارت دوم سمت راست رابطهٔ (۳) است) بدین‌صورت تعریف می‌شود [۱۲]:

$$J(f) = \begin{cases} \int_0^{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (\Delta^{\frac{m}{2}} f)^2 \cos\phi \, d_\phi d_\theta & \text{زوج } m \\ \int_0^{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left\{ \frac{(\Delta^{\frac{m-1}{2}} f)_\theta^2}{(\cos\phi)^2} + (\Delta^{\frac{m-1}{2}} f)_\phi^2 \right\} \cos\phi \, d_\phi d_\theta & \text{فرد } m \end{cases}$$

که در آن $(g)_z$ نمایان‌گر مشتق g با توجه به z است و عمل‌گر $\Delta f$ لاپلاس سطح روی کرهٔ واحد را نشان می‌دهد که بدین‌صورت تعریف می‌شود:

$$\Delta f = \frac{1}{(\cos\phi)^2} f_{\theta\theta} + \frac{1}{\cos\phi} (\cos\phi \ f_\phi)_\phi$$

فضای مدل نیز عبارت است از:

$$W_2^m(S) = \{ f : \left| \int_S f \, dx \right| < \infty , J(f) < \infty \}.$$

4. Generalized Cross Validation (GCV)

برای بررسی بیش‌تر در مورد اسپلاین کروی می‌توان به [۱۳] مراجعه کرد.

در بخش بعدی با به‌کارگیری روش‌های ناپارامتری اسپلاین رویهٔ نازک و اسپلاین کروی و روش پارامتری رگرسیون چندگانه به برازش مدل به داده‌های فضایی شبیه‌سازی شده می‌پردازیم و دقت این روش‌ها با معیار خطای استاندارد باقی‌مانده مقایسهٔ عددی می‌شود.

## مقایسه شبیه سازی روش های اسپلاین رویهٔ نازک، کروی و رگرسیون چندگانه

طبق آن‌چه‌که در قسمت مقدمه برای تولید داده گفته شد، داده‌های تحقیق از یک توزیع نرمال با بردار میانگین صفر آمده‌اند که برای انجام مقایسه، کوواریانس‌های متفاوتی از آن‌ها در نظر گرفته شد. با هدف انجام مقایسه‌ای کاربردی میان روش پارامتری رگرسیون و روش‌های ناپارامتری اسپلاین رویهٔ نازک و اسپلاین کروی در برازش مدلی مناسب به داده‌ها، به‌کمک نرم‌افزار R و به‌صورت جداگانه، یک مدل رگرسیون پارامتری، یک اسپلاین رویهٔ نازک و یک اسپلاین کروی، به داده‌های شبیه‌سازی شده برازانده و برای بررسی تأثیر اندازهٔ نمونه (n) بر دقت روش‌های مورد نظر، از معیار خطای استاندارد باقی‌مانده استفاده شده است. برای محاسبهٔ این معیار مجموع توان‌های دوم خطا بر درجهٔ آزادی مدل تقسیم شده و سپس از مقدار حاصل جذر گرفته می‌شود. واضح است که مدلی با خطای استاندارد کم‌تر، دارای عملکرد بهتری است.

در این تحقیق هم‌چنین تأثیر هم‌بستگی‌های مختلف (ρ) بین متغیرها از طریق ماتریس‌های کوواریانس مختلف بررسی می‌شود، هم‌چنین برای از بین بردن اثر تصادفی بودن واحدهای نمونه‌ای، تکرارهای متفاوتی از تولید داده‌ها را در نظر گرفتیم.

در جدول۱، مقادیر خطای استاندارد حاصل از روش پارامتری رگرسیون، روش‌های ناپارامتری اسپلاین رویهٔ نازک و اسپلاین کروی برای اندازهٔ نمونه و هم‌بستگی‌های متفاوت به‌روش شبیه‌سازی محاسبه شده است.

برای همهٔ نمونه‌های موجود، بردار میانگین مقادیر به صورت $\mu = (0,0,0)$ در نظر گرفته شده است.

مقایسهٔ خطا برای هر n و به‌ازای ρهای مختلف نتایج واضح‌تری در بر دارد که در جدول ۲ آمده است.

به‌دلیل لزوم مثبت معین بودن ماتریس کوواریانس در نرم‌افزار تولید داده‌های شبیه‌سازی، در انتخاب مقادیر هم‌بستگی‌ها محدودیت داشته‌ایم از این رو، در جداول ۱ و ۲ مقادیر مشخصی برای هم‌بستگی در نظر گرفته شده است.

چنان‌که در جدول ۱ ملاحظه می‌شود، اسپلاین رویهٔ نازک به‌دلیل ایجاد خطای استاندارد کم‌تر، عمدتاً دارای عملکرد بهتری نسبت به سایر روش‌ها، در همهٔ اندازه‌های نمونه‌ای و مقادیر هم‌بستگی‌ها است. پس از آن روش رگرسیون چندگانه دارای خطای استاندارد، کم‌تری نسبت به اسپلاین کروی است و طبق معیار کم‌ترین خطای استاندارد، ضعیف‌ترین عملکرد را اسپلاین کروی نسبت به‌سایر روش‌ها داراست.

مقادیر جدول ۲ نیز نشان می‌دهد که بالا بودن هم‌بستگی‌های نمونه‌ای، روی عملکرد روش رگرسیون و اسپلاین رویهٔ نازک تأثیر مثبت داشته (پایین آوردن مقدار خطای استاندارد) اما در مورد برازش با اسپلاین کروی تقریباً بی‌تأثیر بوده است و گاهی حتی تأثیر معکوس داشته است. در این جدول، علاوه بر مقادیر هم‌بستگی برابر متغیرهای توضیحی با متغیر پاسخ، هم‌بستگی‌های نابرابر نیز در نظر گرفته شد و می‌توان ملاحظه کرد که وقتی حداقل یکی از متغیرهای

**جدول۱. مقادیر خطای استاندارد باقی‌مانده‌ها در مقابل هم‌بستگی‌های ثابت**

| هم‌بستگی | اندازهٔ نمونه | روش‌ها | | |
|---|---|---|---|---|
| | | رگرسیون | اسپلاین رویهٔ نازک | اسپلاین کروی |
| $\rho_{x_1,y} = ۰/۱۷$ $\rho_{x_2,y} = ۰/۱۷$ | ۳۰ | ۱۳/۲۷ | ۱۳/۰۳ | ۱۴/۶۴ |
| | ۵۰ | ۱۷/۰۸ | ۱۶/۷۱ | ۱۸/۵ |
| | ۱۰۰ | ۲۴/۴۶ | ۲۴/۴۱ | ۲۵/۵۳ |
| | ۲۰۰ | ۳۱/۹۵ | ۳۱/۳۷ | ۳۲/۷۴ |
| | ۵۰۰ | ۵۰/۷۱ | ۵۰/۷۱ | ۵۲/۳۵ |
| | ۱۰۰۰ | ۷۴/۳ | ۷۴/۱۲ | ۷۷/۶۳ |
| $\rho_{x_1,y} = ۰/۲۵$ $\rho_{x_2,y} = ۰/۲۵$ | ۳۰ | ۱۲/۷۹ | ۱۲/۱ | ۸/۳۴ |
| | ۵۰ | ۱۶/۲۶ | ۱۵/۷۵ | ۱۸/۶ |
| | ۱۰۰ | ۳۲/۴۶ | ۲۳/۳۷ | ۲۵/۵۲ |
| | ۲۰۰ | ۳۰/۷۲ | ۳۰/۰۵ | ۳۳/۱۴ |
| | ۵۰۰ | ۴۸/۶۳ | ۴۸/۶۳ | ۶۳/۲۷ |
| | ۱۰۰۰ | ۷۱/۲۴ | ۷۱/۱۵ | ۷۷/۸۳ |
| $\rho_{x_1,y} = ۰/۴۰$ $\rho_{x_2,y} = ۰/۴۰$ | ۳۰ | ۹/۹۸ | ۱۰/۰۹ | ۱۲/۳۴ |
| | ۵۰ | ۱۴/۷۶ | ۱۴/۴۲ | ۱۷/۶۹ |
| | ۱۰۰ | ۲۰/۱۷ | ۲۰/۰۴ | ۲۳/۲ |
| | ۲۰۰ | ۲۶/۷۲ | ۲۶/۳۵ | ۳۲/۱۷ |
| | ۵۰۰ | ۴۲/۴۸ | ۴۲/۴۹ | ۵۰/۷۲ |
| | ۱۰۰۰ | ۶۱/۵۸ | ۶۱/۵۵ | ۷۳/۰۵ |
| $\rho_{x_1,y} = ۰/۵۰$ $\rho_{x_2,y} = ۰/۵۰$ | ۳۰ | ۸/۹۸ | ۹/۱۶ | ۹/۷ |
| | ۵۰ | ۱۲/۰۰۳ | ۱۱/۷۸ | ۲/۲ |
| | ۱۰۰ | ۱۷/۰۰۱ | ۱۷/۰۰۱ | ۲۱/۴ |
| | ۲۰۰ | ۲۳/۰۵ | ۲۲/۹ | ۳۳/۴ |
| | ۵۰۰ | ۳۶/۳۵ | ۳۶/۱۵ | ۵۳/۳ |
| | ۱۰۰۰ | ۵۳/۲۷ | ۵۳/۲۷ | ۷۷/۴۶ |
| $\rho_{x_1,y} = ۰/۱۷$ $\rho_{x_2,y} = ۰/۹۰$ | ۳۰ | ۴/۱۱ | ۴ | ۱۰/۸۴ |
| | ۵۰ | ۵/۴۷ | ۵/۴۵ | ۱۹/۱۲ |
| | ۱۰۰ | ۷/۶۵ | ۷/۶۵ | ۲۴/۸۴ |
| | ۲۰۰ | ۱۰/۳۲ | ۱۰/۳۳ | ۳۵/۴۸ |
| | ۵۰۰ | ۱۶/۲۳ | ۱۶/۹۴ | ۵۵/۲۶ |
| | ۱۰۰۰ | ۲۳/۸ | ۲۴/۸۴ | ۷۹/۹۷ |
| $\rho_{x_1,y} = ۰/۵۰$ $\rho_{x_2,y} = ۰/۸۰$ | ۳۰ | ۳/۴۹ | ۳/۳۸ | ۵/۴ |
| | ۵۰ | ۴/۶۸ | ۴/۶۸ | ۱۸/۳ |
| | ۱۰۰ | ۶/۵۶ | ۶/۵۶ | ۲۴/۷۹ |
| | ۲۰۰ | ۹/۰۱ | ۹/۰۱ | ۳۵/۷۵ |
| | ۵۰۰ | ۱۴/۱ | ۱۴/۰۹ | ۵۵/۲۹ |
| | ۱۰۰۰ | ۲۰/۶۲ | ۲۰/۵۶ | ۷۹/۸۶ |
| $\rho_{x_1,y} = ۰/۷۰$ $\rho_{x_2,y} = ۰/۷۰$ | ۳۰ | ۱/۷۸ | ۱/۶۲ | ۱۵/۷۹ |
| | ۵۰ | ۲/۳۹ | ۲/۳۹ | ۰/۰۰۳ |
| | ۱۰۰ | ۳/۳۵ | ۳/۳۵ | ۲۵/۳۱ |
| | ۲۰۰ | ۴/۵۵ | ۴/۵۵ | ۳۶/۱۲ |
| | ۵۰۰ | ۷/۱۵ | ۷/۱۴ | ۵۵/۸ |
| | ۱۰۰۰ | ۱۰/۴۷ | ۱۰/۴۷ | ۸۰/۱۲ |

**جدول۲. مقادیر خطای استاندارد باقی‌مانده‌ها به‌ازای nهای ثابت**

| اندازهٔ نمونه | روش‌ها | همبستگی هر یک از متغیرهای توضیحی با پاسخ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | ۰/۱۷ و ۰/۱۷ | ۰/۲۵ و ۰/۲۵ | ۰/۴۰ و ۰/۴۰ | ۰/۵۰ و ۰/۵۰ | ۰/۱۷ و ۰/۹۰ | ۰/۸۰ و ۰/۵۰ | ۰/۷۰ و ۰/۷۰ |
| ۳۰ | رگرسیون | ۱۳/۲۷ | ۱۲/۷۹ | ۹/۹۸ | ۸/۹۸ | ۴/۱۱ | ۳/۴۹ | ۱/۷۸ |
| | اسپلاین رویه نازک | ۱۳/۰۳ | ۱۲/۱ | ۱۰/۰۹ | ۹/۱۶ | ۴ | ۳/۳۸ | ۱/۶۲ |
| | اسپلاین کروی | ۱۴/۶۴ | ۸/۳۴ | ۱۲/۳۴ | ۹/۷ | ۱۰/۸۴ | ۵/۴ | ۱۵/۷۹ |
| ۵۰ | رگرسیون | ۱۷/۰۸ | ۱۶/۲۶ | ۱۴/۷۶ | ۱۲/۰۰۳ | ۵/۴۷ | ۴/۶۸ | ۲/۳۹ |
| | اسپلاین رویه نازک | ۱۶/۷۱ | ۱۵/۷۵ | ۱۴/۴۲ | ۱۱/۷۸ | ۵/۴۵ | ۴/۶۸ | ۲/۳۹ |
| | اسپلاین کروی | ۱۸/۵ | ۱۸/۶ | ۱۷/۶۹ | ۲/۲ | ۱۹/۱۲ | ۱۸/۳ | ۰/۰۰۳ |
| ۱۰۰ | رگرسیون | ۲۴/۴۶ | ۳۲/۴۶ | ۲۰/۱۷ | ۱۷/۰۰۱ | ۷/۶۵ | ۶/۵۶ | ۳/۳۵ |
| | اسپلاین رویه نازک | ۲۴/۴۱ | ۲۳/۳۷ | ۲۰/۰۴ | ۱۷/۰۰۱ | ۷/۶۵ | ۶/۵۶ | ۳/۳۵ |
| | اسپلاین کروی | ۲۵/۵۳ | ۲۵/۵۲ | ۲۳/۲ | ۲۱/۴ | ۲۴/۸۴ | ۲۴/۷۹ | ۲۵/۳۱ |
| ۲۰۰ | رگرسیون | ۳۱/۹۵ | ۳۰/۷۲ | ۲۶/۷۲ | ۲۳/۰۵ | ۱۰/۳۲ | ۹/۰۱ | ۴/۵۵ |
| | اسپلاین رویه نازک | ۳۱/۳۷ | ۳۰/۰۵ | ۲۶/۳۵ | ۲۲/۹ | ۱۰/۳۳ | ۹/۰۱ | ۴/۵۵ |
| | اسپلاین کروی | ۳۲/۷۴ | ۳۳/۱۴ | ۳۲/۱۷ | ۳۳/۴ | ۳۵/۴۸ | ۳۵/۷۵ | ۳۶/۱۲ |
| ۵۰۰ | رگرسیون | ۵۰/۷۱ | ۴۸/۶۳ | ۴۲/۴۸ | ۳۶/۳۵ | ۱۶/۲۳ | ۱۴/۱ | ۷/۱۵ |
| | اسپلاین رویه نازک | ۵۰/۷۱ | ۴۸/۶۳ | ۴۲/۴۹ | ۳۶/۱۵ | ۱۶/۹۴ | ۱۴/۰۹ | ۷/۱۴ |
| | اسپلاین کروی | ۵۲/۳۵ | ۶۳/۲۷ | ۵۰/۷۲ | ۵۳/۳ | ۵۵/۲۶ | ۵۵/۲۹ | ۵۵/۸ |
| ۱۰۰۰ | رگرسیون | ۷۴/۳ | ۷۱/۲۴ | ۶۱/۵۸ | ۵۳/۲۷ | ۲۳/۸ | ۲۰/۶۲ | ۱۰/۴۷ |
| | اسپلاین رویه نازک | ۷۴/۱۲ | ۷۱/۱۵ | ۶۱/۵۵ | ۵۳/۲۷ | ۲۴/۸۴ | ۲۰/۵۶ | ۱۰/۴۷ |
| | اسپلاین کروی | ۷۷/۶۳ | ۷۷/۸۳ | ۷۳/۰۵ | ۷۷/۴۶ | ۷۹/۹۷ | ۷۹/۸۶ | ۸۰/۱۲ |

توضیحی هم‌بستگی بزرگ‌تر از ۰/۵ با متغیر پاسخ دارند، حتی اگر متغیر دوم دارای هم‌بستگی پایین با y باشد، برازش رگرسیون و اسپلاین رویهٔ نازک دقیق‌تر از حالتی است که هر دو متغیر هم‌بستگی متوسط ۰/۵۰ با متغیر پاسخ داشته باشند و این نکته در رابطه با اسپلاین کروی برقرار نیست. اما در صورتی که هر دو هم‌بستگی بالا باشد، صرف‌نظر از نوع روش، نتایج حاصل همان‌گونه که انتظار می‌رفت بیان‌گر کاهش قابل ملاحظه‌ای نسبت به سایر حالت‌هاست.

علاوه بر این، نتایج حاصل از جدول ۱ نشان می‌دهد که در هر هم‌بستگی، با زیاد شدن اندازهٔ نمونه، خطای استاندارد نیز افزایش پیدا کرده است. در توجیه این پدیده باید گفت که اگر چه تعداد نمونهٔ بیش‌تر، برازش دقیق‌تری ارائه داده است اما به‌دلیل وجود توان دوم در محاسبهٔ خطای استاندارد، با زیاد شدن تعداد جملات، انتظار چنین افزایشی نیز وجود داشت.

نکتهٔ دیگری که در این‌جا دریافت می‌شود این است که اسپلاین رویهٔ نازک در هم‌بستگی‌های پایین نمونه‌ای نیز هم‌چنان برتری خود در مقابل روش پارامتری رگرسیون را حفظ نموده است. حال بدون در نظر گرفتن تأثیر اندازهٔ نمونه و با احتساب هم‌بستگی‌های نمونه‌ای متفاوت، درستی این مطلب را بررسی می‌کنیم.

جدول ۳ تحت تأثیر هم‌بستگی‌های مختلف برای سه روش برازشی در نظر گرفته شده است.

جدول۳ نشان می‌دهد که در تمام هم‌بستگی‌ها (به جز در یک مورد با اختلاف اندک)، اسپلاین رویهٔ نازک با خطای استانداردی کم‌تر از رگرسیون ظاهر شده و اسپلاین کروی در برازش، به‌طور چشم‌گیر، بیش‌ترین خطا را در بین سه روش داشته است.

جدول ۳ هم‌چنین رابطه‌ای را بین مجموع دو مقدار هم‌بستگی و مقادیر خطای استاندارد آن‌ها نشان می‌دهد. به این صورت که هرچه مجموع هم‌بستگی‌ها بیش‌تر باشد با خطای استاندارد کم‌تری مواجه‌ایم.

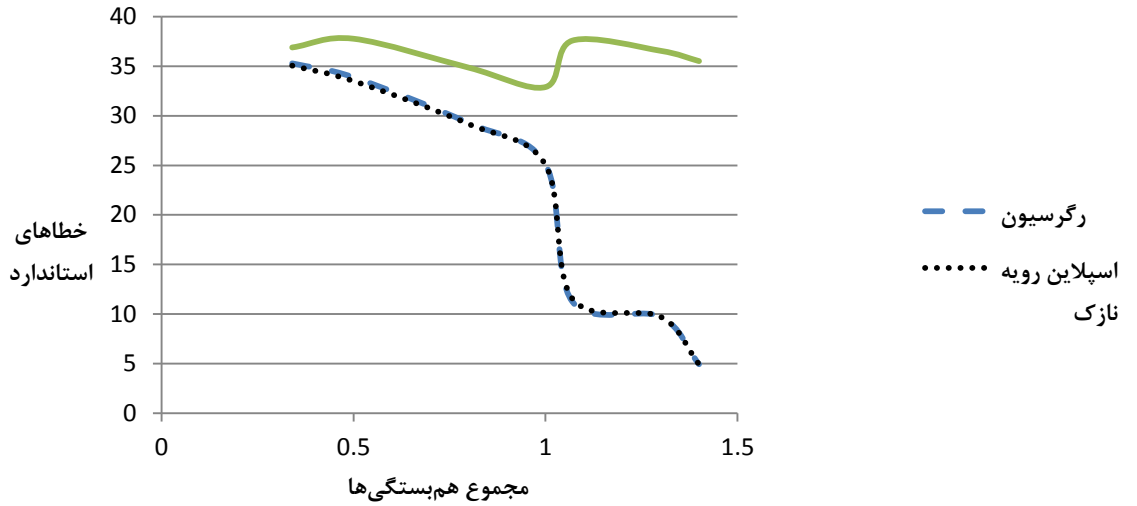
**جدول ۳. مقادیر خطای استاندارد سه روش در مقابل هم‌بستگی‌های مختلف**

| میزان هم‌بستگی | روش‌های برازشی | | |
|---|---|---|---|
| | رگرسیون | اسپلاین رویهٔ نازک | اسپلاین کروی |
| $\rho_{x_1,y} = 0/17$ $\rho_{x_2,y} = 0/17$ | ۳۵/۳ | ۳۵/۰۶ | ۳۶/۹ |
| $\rho_{x_1,y} = 0/25$ $\rho_{x_2,y} = 0/25$ | ۳۳/۹ | ۳۳/۵ | ۳۷/۷۸ |
| $\rho_{x_1,y} = 0/40$ $\rho_{x_2,y} = 0/40$ | ۲۹/۲۸ | ۲۹/۱۶ | ۳۴/۸۶ |
| $\rho_{x_1,y} = 0/50$ $\rho_{x_2,y} = 0/50$ | ۲۵/۱۱ | ۲۵/۰۴ | ۳۲/۹ |
| $\rho_{x_1,y} = 0/17$ $\rho_{x_2,y} = 0/90$ | ۱۱/۲۶ | ۱۱/۵۴ | ۳۷/۵۹ |
| $\rho_{x_1,y} = 0/50$ $\rho_{x_2,y} = 0/80$ | ۹/۷۴ | ۹/۷۱ | ۳۶/۵۶ |
| $\rho_{x_1,y} = 0/70$ $\rho_{x_2,y} = 0/70$ | ۴/۹۴ | ۴/۹۲ | ۳۵/۵۲ |

برای مشاهده عملکرد روش‌ها و چگونگی روابط میان سه روش برازشی، تمامی اطلاعات حاصل در نمودار ۱ آورده شده که در آن محور افقی هم‌بستگی هر یک از متغیرهای توضیحی با پاسخ و محور عمودی مقادیر خطای استاندارد را نشان می‌دهد و از آن‌جاکه مقادیر هم‌بستگی برای دو متغیر لزوما یکی نیست، بنا براین در محور افقی مجموع هم‌بستگی‌ها را در نظر می‌گیریم.

از نمودار ۱ مشخص است که خطای استاندارد حاصل از روش‌های رگرسیون (خط چین) و اسپلاین رویهٔ نازک (نقطه‌چین)، برهم منطبق هستند. این دو روش، چنان که انتظارمی‌رفت، با افزایش هم‌بستگی، خطای استاندارد کم‌تری ایجاد کردند. اما با توجه به نوسانات و ناهمواری‌های ایجاد شده برای اسپلاین کروی ( خط ممتد) و خطای استاندارد زیاد آن، تنها می‌توان به عملکرد ضعیف آن نسبت به دو روش دیگر پی برد.

اکنون برازش‌ها را با اندازه‌های متفاوت نمونه‌ای در نظر می‌گیریم. جدول ۴ مقادیر خطای استاندارد حاصل از هر یک از سه روش برازشی را برای اندازه‌های مختلف نمونه‌ای نشان می‌دهد:




**نمودار ۱. تأثیر هم‌بستگی بر خطاهای استاندارد برازشی**

**جدول ۴. مقادیر خطای استاندارد سه روش در مقابل اندازه‌های مختلف نمونه‌ای**

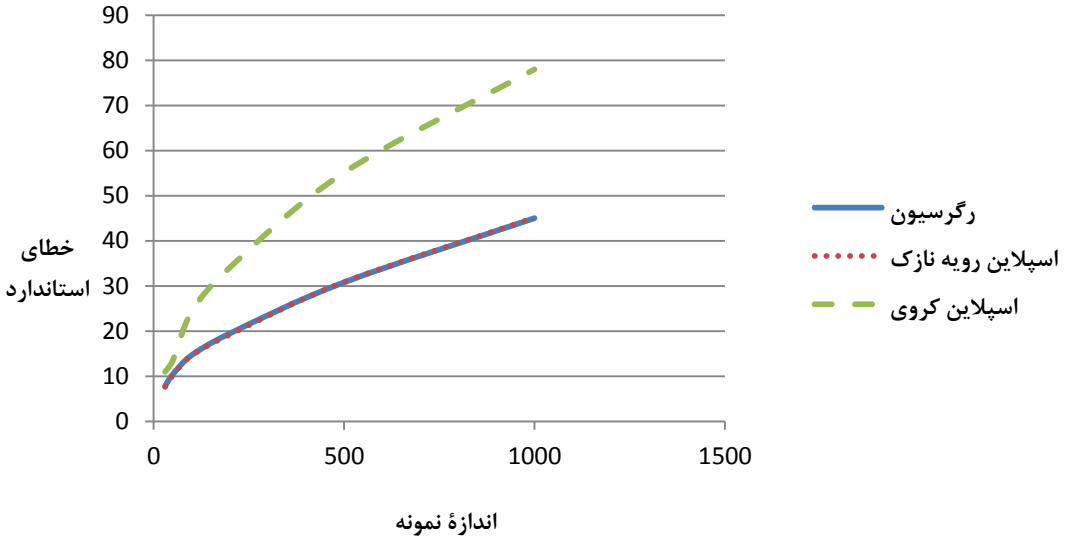
| اندازهٔ نمونه | روش‌های برازشی | | |
|---|---|---|---|
| | رگرسیون | اسپلاین رویهٔ نازک | اسپلاین کروی |
| ۳۰ | ۷/۷۷ | ۷/۶۳ | ۱۱/۰۰۷ |
| ۵۰ | ۱۰/۳۷ | ۱۰/۱۷ | ۱۳/۴۸ |
| ۱۰۰ | ۱۴/۶۶ | ۱۴/۶۳ | ۲۴/۳۷ |
| ۲۰۰ | ۱۹/۴۷ | ۱۹/۲۲ | ۳۴/۱۱ |
| ۵۰۰ | ۳۰/۸۱ | ۳۰/۸۸ | ۵۵/۱۴ |
| ۱۰۰۰ | ۴۵/۰۴ | ۴۵/۱۴ | ۷۷/۹۹ |

طبق مقادیر خطای استاندارد حاضر در جدول ۴، با صرف‌نظر از هم‌بستگی داده‌ها و در نظر گرفتن اندازهٔ نمونه، این بار نیز مقدار خطای استاندارد برازش با اسپلاین کروی همواره بیش از دو روش دیگر بررسی می‌شود.

با مقایسهٔ عملکرد دو روش دیگر، چنان‌که ملاحظه می‌شود، به‌ازای نمونه‌های بزرگ‌تر (<۲۰۰)، از کارایی اسپلاین رویهٔ نازک در مقابل رگرسیون پارامتری کاسته شده و می‌توان گفت که برای نمونه‌های بزرگ، رگرسیون پارامتری نیز روش مناسبی است.

نمودار ۲ به‌شدت تحت تأثیر اندازه‌های نمونه‌ای قرار گرفته است. محورهای افقی و عمودی این نمودار به‌ترتیب بیان‌گر اندازهٔ نمونه و خطای استاندارد هستند.

جدول۴، انطباق منحنی رگرسیون و اسپلاین رویهٔ نازک را نشان می‌دهد و نمودار۲ این تلاقی را به‌تصویر می‌کشد. نمودار ۲ هم‌چنین بیان‌گر فاصله‌ای است که اسپلاین کروی به‌ازای همهٔ اندازه‌های نمونه‌ای، با این دو روش ایجاد نموده. نکتهٔ آخر این که در هر سهٔ این برازش‌ها، افزایش اندازهٔ نمونه منجر به افزایش خطای استاندارد شده است.




**نمودار ۲. تأثیر اندازهٔ نمونه بر خطای استاندارد برازش سه روش**

نگاهی کلی به نمودارهای ۱ و ۲ می‌رساند که با تکیه بر معیار خطای استاندارد، روش پارامتری رگرسیون و ناپارامتری اسپلاین رویهٔ نازک در حالت کلی بهتر از روش اسپلاین کروی عمل کرده‌اند، البته شاهد استثناهایی نیز بوده‌ایم به این صورت که در دو مورد از اندازه‌های کوچک، یکی اندازهٔ ۳۰ و دیگری ۵۰، اسپلاین کروی خطای استاندارد کم‌تری نسبت به آن دو روش دیگر داشته و در چند نمونه از اندازه‌های بزرگ‌تر، اکثراً ۵۰۰ تایی، با وجود برتری رگرسیون و اسپلاین رویهٔ نازک در مقابل اسپلاین کروی، رگرسیون خطای استانداردی کم‌تر از اسپلاین رویهٔ نازک نشان داده است.

از این رو، با توجه به نتایج حاصل در این مقاله، هم‌چنین بر اساس جدول ۴، برای هم‌بستگی‌های مختلف همواره اسپلاین رویهٔ نازک و رگرسیون به‌روش دیگر برتری دارند. برای اندازه‌های کوچک و متوسط به‌کارگیری اسپلاین رویهٔ نازک و برای اندازه‌های نمونه‌ای بزرگ، روش پارامتری رگرسیون در برازش مدل به داده‌ها پیشنهاد می‌شود که با توجه به افزایش تعداد پارامترها و حجم محاسبات، به‌کارگیری این شیوه کاملا منطقی است.

## نتیجه‌گیری

در این تحقیق، رگرسیون پارامتری چندگانه و روش‌های ناپارامتری اسپلاین رویهٔ نازک و کروی با استفاده از معیار خطای استاندارد باقی‌مانده‌ها به‌روش شبیه‌سازی مقایسه‌ی عددی شدند، هم‌چنین تأثیر مقدار هم‌بستگی‌ها و اندازه‌های نمونه بر برازش مدل‌ها بررسی شد.

هنگامی که هم‌بستگی‌های متفاوتی برای داده‌ها قائل شدیم، اسپلاین کروی در مقابل این داده‌ها، روند برازشی خاصی از خود نشان نداد و در سطحی با خطای بیش‌تر از دو روش دیگر باقی‌ماند، در واقع میزان هم‌بستگی داده‌ها در مورد آن بی‌تأثیر بوده است. اما دو روش دیگر رگرسیون و اسپلاین رویهٔ نازک با قابلیتی تقریباً یک‌سان، به‌موازات بالا رفتن هم‌بستگی داده‌های نمونه‌ای، با خطای استاندارد کم‌تری در برازش همراه بودند، به این‌صورت که هرچه مجموع هم‌بستگی دو متغیر توضیحی با پاسخ مقدار بیش‌تری بود، خطای استاندارد کم‌تری در برازش آن‌ها به‌وجود آمده است.

## منابع

1. Wahba G., "Spline models for observational Data", University of Wisconsin at Madison, Madison Wisconsin (1990).
2. Gu C., "Smoothing spline Anova Models", Springer, New York, (2002).
3. Hutchinson M. F., "Interpolation of Rainfall Data with Thin Plate Smoothing Splines, Part1:Two Dimensional Smoothing of Data with short Range Correlation, center for Resource and Environmental studies", Australian National University, Canberra, ACT0200, (1995).
4. Hang Y., Nix H. A., Hutchinson M. F., Booth T. H., "Spatial Interpolation of Monthly Mean Climate Data for China", Chinese Academy of forestry, Beijing 100091 (2005).
5. Zheng X., Basher R., "Thin Plate Smoothing Spline modeling of Spatial Climate data" (1995-1998).
6. Tait A., Henderson R., Turner R., Zheng X., "Thin Plate Smoothing Spline Interpolation of Daily Rainfall for New Zealand using a climatological Rainfall surface", national Institute of water and atmospheric research, private Bag 14-901, Kilbirnie, Wellington New Zealand, (2006).
7. Jie M., Houghton D., Teebagy N., "Global Analysis of Ozone Data based on Spherical Spline", Bentley College, United States (1997).
8. Robeson S. M., "Spherical Methods for Spatial Interpolation Review and Evaluation", Cartography and Geographic information systems, Vol. 24, No.1 (1997) 3-20.
9. Tibshirani R., Wasserman L., Nonparametric Regression, Statistical Machine Learning, (2015).
10. Green P. J., Silverman B.W., "Nonparametric regression and Generalized linear models: A Roughness penalty approach", Chapman Hall (1994).
11. Keel L., "Semi parametric regression for the social sciences", Wiley (2008).
12. Wang Y., "Smoothing Splines methods and applications", University of California (2011).
13. Buss S., Fillmore J., "Spherical averages and applications to Spherical Splines", University of California, San Diego, pixel-cafe talk (2002) 4-5.
۱۴. علی‌محمدی روشنک، مختارپور مهرنوش، "بررسی دقت روش‌های اسپلاین رویه نازک و کروی"، مجموعه مقالات دوازدهمین کنفرانس آمار ایران، دانشگاه رازی، کرمانشاه، (تابستان۱۳۹۳) ۱۷۴.