رگرسیون خطی در حضور داده‌های بد تراز

عادله عصاره [۱] فیروزه ریواز [۲]

چکیده:

در این مقاله چهار رویکرد به مسئلهٔ برازش یک مدل رگرسیون خطی در حضور داده‌های بد تراز فضایی ارائه می‌شود. این رویکردها عبارتند از روش باجایگذاری، شبیه‌سازی، رگرسیون کالیبدنی و ماکسیمم درستنمایی. در دو رویکرد اول، با مدل‌بندی همبستگی موجود در متغیر توضیحی، پیشگویی آن در موقعیت‌های متناظر با متغیر پاسخ تعیین می‌شود. سپس باجایگذاری پیشگوهای به‌دست آمده به جای مقادیر واقعی در مدل رگرسیونی، برازش مدل انجام می‌شود. نشان داده می‌شود این کار باعث ایجاد خطای برکسن شده و این خطا نیز منجر به ایجاد اریبی در برآورد شیب مدل رگرسیونی می‌شود. برای تعدیل این اریبی، رویکرد رگرسیون کالیبدنی ارائه می‌شود. در رویکرد ماکسیمم درستنمایی مستقیماً از داده‌های بد تراز استفاده شده و پارامترهای مدل رگرسیونی برآورد می‌شوند. در واقع، دیگر نیازی به پیشگویی متغیر توضیحی در مکان‌های متناظر با متغیر پاسخ نیست. اما متأسفانه بررسی دقیق خواص برآوردگر ماکسیمم درستنمایی به‌دلیل نداشتن فرم تحلیلی، امکان‌پذیر نیست. در یک مطالعهٔ شبیه‌سازی، عملکرد کلیه رویکردها تحت چندین مدل فضایی برای متغیر توضیحی مورد بررسی قرار می‌گیرد. مشاهده می‌شود رگسیون کالیبدنی می‌تواند به میزان قابل توجهی اریبی برآوردگر شیب خط رگرسیونی را نسبت به روش‌های دیگر کاهش دهد. به‌علاوه، میزان پوشش اسمی بازهٔ اطمینان شیب خط رگرسیونی توسط این روش قابل توجه است.

واژه‌های کلیدی: داده‌های بد تراز فضایی، رویکرد باجایگذاری، رگرسیون کالیبدنی، خطای برکسن.

۱ مقدمه

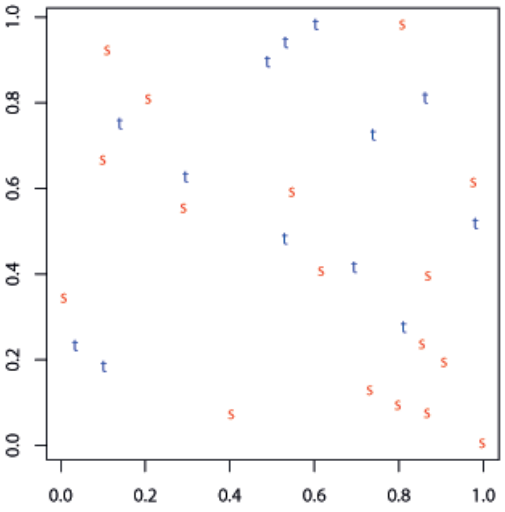
پیشرفت‌های متعدد در تکنیک‌های جمع‌آوری داده، موجب وفور متغیرهای پیشگوی بالقوه برای توضیح یک متغیر پاسخ فضایی شده است. با این وجود هنگامی که داده‌ها از منابع مختلف می‌آیند، مکان‌ها و مقیاس‌های فضایی به‌ندرت باهم متناظر هستند. به‌عبارت دیگر، گاهی مجموعه مکان‌های مشاهده شده برای متغیرهای توضیحی با مجموعه مکان‌های مشاهده شده برای متغیر پاسخ متفاوت است. به بیان دیگر متغیر پاسخ در موقعیت‌هایی مانند s_1، s_2، ...، s_n مشاهده شده است، در حالی که متغیرهای توضیحی در موقعیت‌های دیگری مانند t_1، t_2، ...، t_m مشاهده شده اند. این نوع داده‌ها، به *داده‌های بد تراز فضایی* معروف هستند.

برای مثال فرض کنید ارتباط میان میزان نوعی آلایندهٔ هوا در شهر تهران و تعداد بیماران قلبی در بیمارستان‌های این شهر در مقطعی از زمان مورد توجه است. اما از آنجا که مکان ایستگاه‌های هواشناسی متفاوت با مکان بیمارستان‌ها است، مشاهدات مربوط به این دو متغیر از نوع داده‌های بد تراز فضایی هستند. برای درک بهتر موضوع، اگر در یک فضای مورد مطالعهٔ مفروض، موقعیت‌های مربوط به مشاهدات متغیر توضیحی را با t و موقعیت‌های مربوط به مشاهدات متغیر پاسخ را با s نشان دهیم، تحققی

[۱] گروه آمار، دانشگاه شهید بهشتی

[۲] گروه آمار، دانشگاه شهید بهشتی

از داده‌های بد تراز فضایی را می‌توان در شکل ۱ مشاهده کرد.



شکل۱: داده‌های بد تراز فضایی

نخستین‌بار ژو و همکاران (۲۰۰۳) مسئلهٔ تحلیل رگرسیونی را برای این نوع داده‌ها مطرح نمودند. پس از آن مادسن و همکاران (۲۰۰۸) با در نظر گرفتن یک مدل رگرسیون خطی ساده، با استفاده از روش کریگیدن، متغیرهای توضیحی متناظر با پاسخ‌های مشاهده شده را پیشگویی و با دو روش کمترین توان های دوم و ماکسیمم درستنمایی برازش مدل را انجام دادند. همچنین گری‌پاریس و همکاران (۲۰۰۹) رویکردهای مختلفی برای تحلیل رگرسیونی داده‌های بد تراز فضایی با تکیه‌گاه نقطه‌ای ارائه دادند و به تعدیل خطای اندازه‌گیری ناشی از پیشگویی در موقعیت‌های بد تراز فضایی پدست زدند.

سپس یانگ و همکاران (۲۰۰۹) برای تصحیح این نوع خطا در مواجهه با داده‌های بد تراز فضایی پرداختند و در یک مطالعهٔ شبیه‌سازی این رویکردها را باهم مقایسه کردند. اشپیرو و همکاران (۲۰۱۱) به منظور تصحیح خطای اندازه‌گیری ناشی از پیشگویی در موقعیت‌های بد تراز فضایی، از روش‌های بوت استرپی استفاده کردند.

در این مقاله به مسئلهٔ برازش یک مدل خطی به داده‌های بد تراز فضایی می‌پردازیم. برای حل مسئلهٔ تحلیل رگرسیونی داده‌های بد تراز فضایی، در مکان‌هایی که متغیر پاسخ مشاهده شده است، متغیر توضیحی پیشگویی شده و از مقادیر حاصل به‌عنوان متغیر توضیحی در مدل رگرسیونی استفاده می‌شود. همچنین می‌توان بدون پیشگویی مقادیر متغیر توضیحی در مکان‌های متناظر با پاسخ، ضرایب رگرسیون را به روش ماکسیمم درستنمایی برآورد کرد. بنابراین در بخش ۲ ابتدا مدل‌بندی و نمادگذاری‌ها معرفی می‌شوند. سپس رویکردهای مختلف به مسئلهٔ برازش مدل رگرسیون خطی در حالت بد ترازی فضایی، برای متغیرهای کمّی شامل رویکردهای باجایگذاری، شبیه سازی، رگرسیون کالیبدنی و ماکسیمم درستنمایی به ترتیب در بخش‌های ۱.۲، ۲.۲، ۳.۲ و ۴.۲ ارائه می‌شوند. در بخش ۳ نیز، بر اساس یک مطالعهٔ شبیه‌سازی، عملکرد روش‌های مختلف مورد بررسی قرار می‌گیرد.

۲ مدل آماری

فرض کنید $\boldsymbol{Y} = (Y(s_1), Y(s_2), ..., Y(s_{n_y}))^T$ و $\boldsymbol{X} = (X(s_1), X(s_2), ..., X(s_{n_y}))^T$ به ترتیب بردار مشاهدات متغیر پاسخ و بردار مشاهدات متغیر توضیحی در موقعیت‌های s_1، s_2، ...، s_{n_y} باشند. همچنین $\boldsymbol{Z}$ ماتریسی از مقادیر دیگر پیشگوهای اندازه‌گیری‌شدهٔ بدون خطا در نظر گرفته می‌شود. مدل رگرسیون خطی چندگانهٔ

$$Y(s_i) = \beta_0 + \beta_1 X(s_i) + \boldsymbol{Z}^T(s_i)\boldsymbol{\beta}_z + \epsilon(s_i),\ i = 1, 2, ..., n_y \qquad (1)$$

را در نظر بگیرید، که در آن $\epsilon(s_i) \sim N(0, \sigma_\epsilon^2)$. همچنین β_1 و $\boldsymbol{\beta}_z$ پارامترهای نامعلوم، $X(s_i)$ متغیر توضیحی در مکان s_i و $\boldsymbol{Z}(s_i)_{(q\times 1)}$ بردار پیشگوهای اندازه‌گیری شدهٔ بدون خطا در مکان s_i است.

هنگامی که هر دو بردار $\boldsymbol{Y}$ و $\boldsymbol{X}$ مشاهده شده‌اند، می‌توان از روش‌های معمول برای برآورد پارامترهای مدل استفاده کرد. اما در مسائل بد ترازی، متغیر توضیحی X در مکان‌هایی متفاوت با مشاهدات متغیر پاسخ، مثلاً در مکان‌های t_1، t_2، ...، t_{n_w}، مشاهده شده است و با $\boldsymbol{W} = (X(t_1), X(t_2), ..., X(t_{n_w}))^T$ نشان داده می‌شود. از این رو می‌توان $\boldsymbol{X} = (X(s_1), X(s_2), ..., X(s_{n_y}))^T$ را با استفاده از بردار $\boldsymbol{W}$ پیشگویی کرده و آن را با $\hat{\boldsymbol{X}} = (\hat{X}(s_1), \hat{X}(s_2), ..., \hat{X}(s_{n_y}))^T$ نشان داد.

اگر متغیر پاسخ Y در n_y مکان و متغیر توضیحی X در n_w مکان مشاهده شده باشد، این اطلاعات در جدول ۱ خلاصه شده است.

در ادامه چهار رویکرد برای برازش مدل رگرسیون خطی در مواجهه با داده‌های بد تراز فضایی ارائه می‌شود.

جدول ۱: جدول داده‌های بد تراز فضایی

موقعیت	متغیر پاسخ	متغیرهای توضیحی	
s_i or t_j	Y	X	Z
s_1	$Y(s_1)$	–	$Z(s_1)$
s_2	$Y(s_2)$	–	$Z(s_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
s_{n_y}	$Y(s_{n_y})$	–	$Z(s_{n_y})$
t_1	–	$X(t_1)$	$Z(t_1)$
t_2	–	$X(t_2)$	$Z(t_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{n_w}	–	$X(t_{n_w})$	$Z(t_{n_w})$

۱.۲ رویکرد باجایگذاری

فرض کنید $Y(s_i)$ در مدل (۱) نرمال باشد. برای برازش مدل رگرسیونی به روش باجایگذاری[3] ابتدا $\boldsymbol{X}$ را با استفاده از یکی از روش‌های پیشگویی فضایی و مبتنی بر بردار $\boldsymbol{W}$ پیشگویی کرده و آن را با $\hat{\boldsymbol{X}}$ نشان می‌دهیم. سپس $\hat{\boldsymbol{X}}$ به عنوان متغیر توضیحی در مدل رگرسیون به کار برده می‌شود. در نهایت برآوردگرهای کمترین توان‌های دوم معمولی برای $\boldsymbol{\beta}$ و واریانس آن به شکل زیر محاسبه می شوند:

$$\hat{\boldsymbol{\beta}}_{plug-in} = (\hat{\mathbb{X}}^T\hat{\mathbb{X}})^{-1}\hat{\mathbb{X}}^T\boldsymbol{Y} \qquad (۲)$$

$$Var(\hat{\boldsymbol{\beta}}_{plug-in}) = \sigma_\epsilon^2(\hat{\mathbb{X}}^T\hat{\mathbb{X}})^{-1} \qquad (۳)$$

که در آن $\hat{\mathbb{X}} = [1_{n_y\times 1} \;\; \hat{\boldsymbol{X}} \;\; \boldsymbol{Z}]$. این روش برآوردگرهای ناااریبی را برای پارامترهای خط رگرسیون نتیجه می‌دهد، اما استفاده از $\hat{\boldsymbol{X}}$ به جای $\boldsymbol{X}$ در مدل رگرسیونی، عدم حتمیتی را وارد مدل می‌کند و موجب ایجاد همبستگی در ساختار خطا می‌شود. برای توضیح این مطلب، توجه کنید که مقادیر حاصل از روش‌های پیشگویی فضایی تغییرپذیری کمتری نسبت به مقادیر واقعی دارند، لذا می‌توان نوشت:

$$\boldsymbol{X} = \hat{\boldsymbol{X}} + \boldsymbol{V} \qquad (۴)$$

که در آن $\boldsymbol{V} = \boldsymbol{X} - \hat{\boldsymbol{X}}$ خطای مربوط به پیشگویی $\boldsymbol{X}$ و مستقل از $\boldsymbol{\epsilon}$ است. به‌علاوه $\boldsymbol{V} \sim N(0, \Sigma_V)$. استفاده از $\hat{\boldsymbol{X}}$ به جای $\boldsymbol{X}$ در مدل (۱) منجر به خطای برکسن[4] می‌شود (گری‌پاریس و همکاران، ۲۰۰۹). با جایگذاری (۴) در مدل رگرسیون داریم:

$$\begin{aligned}\boldsymbol{Y} &= \beta_0 1 + \beta_1\boldsymbol{X} + \boldsymbol{\beta}_Z\boldsymbol{Z} + \boldsymbol{\epsilon}\\ &= \beta_0 1 + \beta_1(\hat{\boldsymbol{X}} + \boldsymbol{V}) + \boldsymbol{\beta}_Z\boldsymbol{Z} + \boldsymbol{\epsilon}\\ &= \beta_0 1 + \beta_1\hat{\boldsymbol{X}} + \boldsymbol{\beta}_Z\boldsymbol{Z} +\end{aligned}$$

که در آن $\boldsymbol{\eta} = \beta_1\boldsymbol{V} + \boldsymbol{\epsilon}$. ماتریس کوواریانس خطای جدید، $\boldsymbol{\eta}$، دیگر قطری نیست. زیرا واریانس مانده‌ها برای مدل رگرسیون عبارت از:

$$var(\boldsymbol{\eta}) = \beta_1^2\Sigma_V + \sigma_\epsilon^2 I_{n_y}.$$

است. بنابراین اگرچه برآوردگر OLS برای β_1 ناااریب است (با این فرض که $\hat{\boldsymbol{X}}$ برای $\boldsymbol{X}$ ناااریب است)، اما واریانس آن دیگر واریانس واقعی $\hat{\beta}_1$ نخواهد بود. در این‌گونه مواقع از روش کمترین توان‌های دوم وزنی (WLS) استفاده می‌شود. برای این منظور عناصر قطری ماتریس کوواریانس خطا را به عنوان وزن‌هایی برای مدل رگرسیون در نظر می‌گیرند. اما چون در $var(\boldsymbol{\eta}) = \beta_1^2\Sigma_V + \sigma_\epsilon^2 I_{n_y}$ پارامترهای β_1 و σ^2 نامعلوم هستند، لذا این ماتریس و در نتیجه عناصر قطری آن نامعلوم هستند و بنابراین نمی‌توان از روش WLS برای برآورد ضرایب رگرسیونی استفاده کرد.

لازم به ذکر است که حالتی خاص از روش باجایگذاری به روش کریگ و رگرس [5] معروف است که در آن $\hat{\boldsymbol{X}}$ پیشگوی کریگی در نظر گرفته می‌شود و به عنوان متغیر توضیحی در مدل رگرسیون به کار می‌رود (مادسن و همکاران، ۲۰۰۸).

۲.۲ رویکرد شبیه‌سازی

روش‌های پیشگویی فضایی معمولاً رویه‌ای هموار از فرایند مورد مطالعه ارائه می‌دهند. لذا تغییرپذیری مقادیر پیشگویی بسیار کمتر از تغییرپذیری واقعی متغیر موردنظر است. بنابراین در زمین آمار اغلب به‌جای روش‌های هموارسازی فضایی از شبیه‌سازی زمین‌آماری استفاده می‌شود. گری‌پاریس و همکاران (۲۰۰۹) رویکرد شبیه‌سازی را به این صورت ارائه دادند که از برآورد توزیع متغیر توضیحی به شرط داده‌ها، M نمونهٔ $\boldsymbol{X}_{(k)} = \hat{\boldsymbol{X}} + \boldsymbol{V}_{(k)}$ ، $k = 1, 2, ..., M$، در موقعیت‌های s_1، s_2،

[3] Plug-in Approach

[4] Berkson error

[5] Krige-and-Regress approach

s_{n_y}، ... تولید می‌شود و هریک از این M نمونه به عنوان متغیر توضیحی در مدل رگرسیون (۱) مورد استفاده قرار می‌گیرد. در نتیجه M برآورد $\hat{\beta}_{۱(k)}$، $k = ۱, ۲, ..., M$، حاصل می‌شود و سپس برای به‌دست آوردن یک برآورد کلی از آن‌ها میانگین گرفته می‌شود. در این صورت واریانس $\hat{\beta}_۱$ برابر

$$Var(\hat{\beta}_۱) = Var(E(\hat{\beta}_{۱(k)}|\boldsymbol{X}_{(k)})) + E(Var(\hat{\beta}_{۱(k)}|\boldsymbol{X}_{(k)})).$$

است (گری‌پاریس و همکاران، ۲۰۰۹). یانگ و همکاران (۲۰۰۹) از روش تجزیهٔ چولسکی برای تولید تحقق‌هایی از متغیر توضیحی که خواص فضایی یکسانی با داده‌های اصلی دارند، استفاده کردند. برای توضیح این روش فرض کنید $\begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{X}_{(k)} \end{pmatrix}$ دارای توزیع نرمال با ماتریس کوواریانس

$$\Sigma = \begin{pmatrix} \Sigma_W & \Sigma_{XW} \\ \Sigma_{XW}^T & \Sigma_X \end{pmatrix}$$

باشد. Σ را به صورت

$$\Sigma = LL^T$$

تجزیه می‌کنیم که در آن $L = \begin{pmatrix} L_{۱۱} & ۰ \\ L_{۲۱} & L_{۲۲} \end{pmatrix}$. فرض کنید $\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_۱ \\ \boldsymbol{V}_{۲(k)} \end{pmatrix}$ برداری از تحقق‌های مستقل و هم‌توزیع $N(۰, ۱)$ باشد. آنگاه با استفاده از

$$\begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{X}_{(k)} \end{pmatrix} = \begin{pmatrix} L_{۱۱} & ۰ \\ L_{۲۱} & L_{۲۲} \end{pmatrix} \begin{pmatrix} L_{۱۱}^{-۱}\boldsymbol{W} \\ \boldsymbol{V}_{۲(k)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W} \\ L_{۲۱}L_{۱۱}^{-۱}\boldsymbol{W} + L_{۲۲}\boldsymbol{V}_{۲(k)} \end{pmatrix},$$
$$k = ۱, ..., M, \qquad (۵)$$

می‌توان M تحقق از فرایند X در موقعیت‌های $s_۱$، $s_۲$، ...، s_{n_y} تولید کرد. شایان ذکر است که مقادیر شبیه‌سازی‌شده بر اساس این روش در موقعیت‌هایی که مشاهده موجود است، مقادیر یکسانی با مشاهدات دارند. لذا روش ارائه شده در (۵) نسبت به روش گری‌پاریس و همکاران (۲۰۰۹)، با آنچه در عمل اتفاق می‌افتد، سازگارتر است.

اکنون با به کار بردن هریک از $\boldsymbol{X}_{(k)}$های تولیدشده به عنوان متغیر توضیحی در (۱)، مدل را برازش می‌دهیم و در نتیجه M برآورد برای $\beta_۱$ و واریانس درون شبیه‌سازی [۶] حاصل می‌شود که آن‌ها را به ترتیب با $\hat{\beta}_{۱(k)}$ و $W_{(k)}$، $k = ۱, ..., M$، نشان می‌دهیم. سپس برای به‌دست آوردن یک برآورد کلی از $\hat{\beta}_{۱(k)}$ها میانگین گرفته می‌شود:

$$\hat{\beta}_{۱M} = \frac{۱}{M}\sum_{k=۱}^{M}\hat{\beta}_{۱(k)}.$$

تغییرپذیری این برآوردگر دو جزء دارد: میانگین واریانس درون شبیه‌سازی $(\bar{W}_M)$ و واریانس بین شبیه‌سازی [۷] (B_M)، که به ترتیب عبارت از

$$\bar{W}_M = \frac{۱}{M}\sum_{k=۱}^{M}W_{(k)},$$

و

$$B_M = \frac{۱}{M-۱}\sum_{k=۱}^{M}(\hat{\beta}_{۱(k)} - \hat{\beta}_{۱M})^۲,$$

هستند. بنابراین تغییرپذیری کل به صورت

$$\widehat{Var}(\hat{\beta}_{۱M}) = \bar{W}_M + \frac{M+۱}{M}B_M,$$

خواهد بود که $(\frac{۱}{M} + ۱)$ تعدیلی برای M تعداد متناهی شبیه‌سازی است (یانگ و همکاران، ۲۰۰۹).

مقادیر $\boldsymbol{X}_{(k)}$ تولیدشده از شبیه‌سازی، تغییرپذیری یکسانی با مقادیر واقعی $\boldsymbol{X}$ دارند. اختلاف این دو را باید به عنوان خطای اندازه‌گیری کلاسیک در نظر گرفت (گری‌پاریس و همکاران، ۲۰۰۹). بنابراین استفاده از $\boldsymbol{X}_{(k)}$ به جای $\boldsymbol{X}$ در مدل (۱) را می‌توان در قالب یک مدل خطای اندازه‌گیری کلاسیک بررسی کرد و در نتیجه، برآوردگرهای حاصل به دلیل وجود این نوع خطا اریب خواهند بود.

۳.۲ رویکرد رگرسیون کالیبدنی

رگرسیون کالیبدنی [۸] (RC) روشی آماری برای تعدیل آثار خطای اندازه‌گیری در برآورد پارامترهای یک مدل است. لذا برای تعدیل آثار این نوع خطا در برآوردهای حاصل از رویکردهای باجایگذاری و

[6] within-simulation variance

[7] between-simulation variance

[8] Regression Calibration Approach

شبیه‌سازی که در بخش‌های قبل ارائه شدند، می‌توان از رویکرد رگرسیون کالیبدنی که توسط گری‌پاریس و همکاران (۲۰۰۹) ارائه شده است، استفاده نمود.

در این روش، ابتدا بردار مشاهدات متغیر توضیحی یعنی $\boldsymbol{W} = (X(t_1), X(t_2), ..., X(t_{n_w}))^T$، که در مکان‌هایی متفاوت با مشاهدات متغیر پاسخ مشاهده شده‌اند، به دو بردار $\boldsymbol{W}_1$ با طول k و $\boldsymbol{W}_2$ با طول $n_w - k$ افراز می‌شود. سپس $\boldsymbol{W}_1$ با استفاده از بردار $\boldsymbol{W}_2$ پیشگویی شده و با $\hat{\boldsymbol{W}}_1$ نشان داده می‌شود. همچنین ماتریس $\boldsymbol{Z}$ را به‌طور متناظر با $\boldsymbol{W}$، به دو ماتریس $\boldsymbol{Z}_1$ و $\boldsymbol{Z}_2$ افراز می‌کنیم. اکنون مبتنی بر $\boldsymbol{W}_1$، رگرسیون $\boldsymbol{W}_1$ روی $\hat{\boldsymbol{W}}_1$ برازش داده می‌شود. برای این منظور، مدل

$$W_1(t_i) = \gamma_0 + \gamma_1 \hat{W}_1(t_i) + \boldsymbol{\gamma}_z^T Z_1(t_i) + \delta(t_i),\ i = 1, ..., k, \qquad (۶)$$

را با $E(\delta(t_i)) = 0$ و $var(\delta(t_i)) = \sigma_\delta^2$ در نظر بگیرید. با برازش مدل (۶) با روش کمترین توان‌های دوم، برآوردهای $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\boldsymbol{\gamma}}_z^T)^T$، به دست می‌آید که از آن‌ها برای کالیبره کردن مقادیر پیشگویی $\hat{\boldsymbol{X}}$ استفاده می‌شود. با توجه به رابطهٔ (۶) داریم:

$$E(X(s_i)|\hat{X}(s_i)) = \gamma_0 + \gamma_1 \hat{X}(s_i) + \boldsymbol{\gamma}_z^T Z(s_i),\ i = 1, ..., n_y,$$

با جایگذاری $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\boldsymbol{\gamma}}_z^T)$ به جای $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \boldsymbol{\gamma}_z^T)$ در عبارت فوق، مقادیر کالیبره شدهٔ $\hat{\boldsymbol{X}}$ به‌صورت

$$\hat{X}_{cal}(s_i) = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{X}(s_i) + \hat{\boldsymbol{\gamma}}_z^T Z(s_i), i = 1, ..., n_y,$$

به‌دست می‌آید و لذا می‌توان نوشت

$$\hat{\boldsymbol{X}}_{cal} = \hat{\mathbb{X}}\hat{\gamma},$$

که در آن $\hat{\mathbb{X}} = [1_{n_y \times 1} \quad \hat{\boldsymbol{X}} \quad \boldsymbol{Z}]$ و 1 برداری از یک‌هاست. ماتریس تبدیل Γ را به صورت زیر تعریف می‌کنیم:

$$\Gamma = \begin{bmatrix} 1 & \gamma_0 & 0_{1\times q} \\ 0 & \gamma_1 & 0_{1\times q} \\ 0_{q\times 1} & \boldsymbol{\gamma}_z & I_{q\times q} \end{bmatrix}.$$

لازم به ذکر است که برآورد ماتریس Γ، ماتریس $[1_{n_y\times 1}|\hat{\boldsymbol{X}}|\boldsymbol{Z}]$ را به ماتریس $[1_{n_y\times 1}|\hat{\boldsymbol{X}}_{cal}|\boldsymbol{Z}]$ تبدیل می‌کند. حال با جایگذاری $\hat{\boldsymbol{X}}_{cal}$ در مدل رگرسیون (۱) خواهیم داشت:

$$\begin{aligned} \boldsymbol{Y} &= \beta_0 1_{n_y\times 1} + \beta_1 \hat{\boldsymbol{X}}_{cal} + \boldsymbol{\beta}_Z \boldsymbol{Z} + \boldsymbol{\epsilon} \\ &= \beta_0 1_{n_y\times 1} + \beta_1 \hat{\mathbb{X}}\hat{\boldsymbol{\gamma}} + \boldsymbol{\beta}_Z \boldsymbol{Z} + \boldsymbol{\epsilon} \end{aligned} \qquad (۷)$$

و لذا برآوردگر کالیبدنی $\boldsymbol{\beta}$ به صورت

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{cal} &= \hat{\Gamma}^{-1}\hat{\boldsymbol{\beta}}_{plug-in} \\ &= \hat{\Gamma}^{-1}(\hat{\mathbb{X}}^T\hat{\mathbb{X}})^{-1}\hat{\mathbb{X}}^T\boldsymbol{Y} \end{aligned} \qquad (۸)$$

به‌دست می‌آید که $\hat{\boldsymbol{\beta}}_{plug-in}$ در رابطهٔ (۲) ارائه شده است. همچنین واریانس $\hat{\boldsymbol{\beta}}_{cal}$ با استفاده از بسط تیلور بهشکل

$$Var(\hat{\boldsymbol{\beta}}_{cal}) = \hat{\Gamma}^{-1}[\beta_1^2\sigma_\delta^2(D^TD)^{-1} + \sigma_\epsilon^2(\hat{\mathbb{X}}^T\hat{\mathbb{X}})^{-1}](\hat{\Gamma}^{-1})^T$$

به‌دست می‌آید که در آن $D = [1 \quad \hat{\boldsymbol{W}}_1 \quad \boldsymbol{Z}_1]$.

۴.۲ رویکرد ماکسیمم درستنمایی

تا اینجا روش‌هایی که ارائه شد همگی مبتنی بر پیشگویی مقادیر مشاهده نشدهٔ متغیر توضیحی در موقعیت‌های متناظر با متغیر پاسخ و استفاده از مقادیر حاصل در مدل رگرسیونی بودند. اکنون روش ماکسیمم درستنمایی ارائه می‌شود که در آن برای برآورد ضرایب رگرسیونی، مستقیماً از بردار $\boldsymbol{W}$ استفاده می‌شود.

مدل (۱) با توابع میانگین و کوواریانس

$$E\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \\ \boldsymbol{W} \end{bmatrix} = \begin{bmatrix} [\beta_0 + \beta_1\mu_X + Z^T(s_i)\boldsymbol{\beta}_Z]_{i=1}^{n_y} \\ \mu_X 1_{(n_y+n_w)\times 1}, \end{bmatrix} \qquad (۹)$$

و

$$cov\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \\ \boldsymbol{W} \end{bmatrix} = \begin{bmatrix} \beta_1^2\Sigma_X + \Sigma_\epsilon & \beta_1\Sigma_X & \beta_1\Sigma_{XW} \\ \beta_1\Sigma_X & \Sigma_X & \Sigma_{XW} \\ \beta_1\Sigma_{XW}^T & \Sigma_{XW}^T & \Sigma_W, \end{bmatrix} \qquad (۱۰)$$

را در نظر بگیرید. با توجه به مفروضات مدل داریم:

$$\begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{W} \end{bmatrix} \sim N\left(\begin{bmatrix} [\beta_0 + \beta_1\mu_X + Z^T(s_i)\boldsymbol{\beta}_Z]_{i=1}^{n_y} \\ \mu_X 1_{n_w\times 1} \end{bmatrix}, \Sigma\right),$$

که در آن

$$\Sigma = \begin{bmatrix} \beta_1^2\Sigma_X + \Sigma_\epsilon & \beta_1\Sigma_{XW} \\ \beta_1\Sigma_{XW}^T & \Sigma_W \end{bmatrix}.$$

بنابراین چگالی توأم $\boldsymbol{Y}$ و $\boldsymbol{W}$ عبارت از

$$f_{YW}(y,\ w) = \frac{1}{\sqrt{(2\pi|\Sigma|)}} \times exp\left(-\frac{1}{2}\boldsymbol{V}^T\Sigma^{-1}\boldsymbol{V}\right),$$

است که در آن

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_Y \\ \boldsymbol{V}_X \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y} - [\beta_0 + \beta_1\mu_X + Z^T(s_i)\boldsymbol{\beta}_Z]_{i=1}^{n_y} \\ \boldsymbol{W} - \mu_X 1_{m\times 1} \end{bmatrix}.$$

فرض کنید $\boldsymbol{\theta}_X$ و $\boldsymbol{\theta}_\epsilon$ به ترتیب بردارهای پارامترهای نیم‌تغییرنگار فرایندهای پیشگو و خطا باشند. برآورد ماکسیمم درستنمایی $\boldsymbol{\beta}$ با مینیمم کردن مقدار

$$l(\boldsymbol{\phi}) = -log(f_{YW}) = \frac{1}{2}log|\Sigma| + \frac{1}{2}\boldsymbol{V}^T\Sigma^{-1}\boldsymbol{V}, \quad (11)$$

نسبت به پارامترهای $\boldsymbol{\phi} = [\boldsymbol{\beta}^T\ \mu_X\ \boldsymbol{\theta}_X^T\ \boldsymbol{\theta}_\epsilon^T]^T$ به‌دست می آید.

معمولاً $\boldsymbol{\theta}_X$ و $\boldsymbol{\theta}_\epsilon$ بردارهایی به طول ۳ هستند (زیرا شامل پارامترهای دامنه، اثر قطعه‌ای و ازاره هستند) و $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \boldsymbol{\beta}_Z]^T$ نیز برداری به طول $q+2$ است. بنابراین مینیمم کردن (۱۱) یک مسئله در فضای $(q+9)$ بعدی است که به صورت مستقیم تحت قید مثبت بودن پارامترهای نیم‌تغییرنگار تعیین می‌شوند.

نتیجهٔ بهینه سازی عددی، برداری از برآوردهای ML برای تمامی این پارامترهاست:

$$\hat{\boldsymbol{\phi}}_{ML} = [\hat{\boldsymbol{\beta}}_{ML}^T \quad \hat{\mu}_{X,ML} \quad \hat{\boldsymbol{\theta}}_{X,ML}^T \quad \hat{\boldsymbol{\theta}}_{\epsilon,ML}^T]^T,$$

و برآورد ML پارامتر شیب β_1 ، عنصر دوم $\hat{\boldsymbol{\phi}}$ خواهد بود که با $\hat{\beta}_{1,ML} = \hat{\phi}_{2,ML}$ نشان داده می‌شود.

برآورد واریانس $\hat{\beta}_{1,ML}$

اگر توزیع $[\boldsymbol{Y}\ \boldsymbol{W}]^T$ شرایط نظم را داشته باشد، که در حالت گاوسی بستگی به شکل توابع نیم‌تغییرنگار دارد، آنگاه برآوردگر ماکسیمم درستنمایی، نااریب و به‌طور مجانبی نرمال است (مادسن و همکاران، ۲۰۰۸). در این حالت ماتریس مجانبی کوواریانس $\boldsymbol{\phi}_{ML}$ معکوس ماتریس اطلاع فیشر $(I(\boldsymbol{\phi})^{-1})$ است که (i,j)-امین عنصر آن عبارت از

$$I(\boldsymbol{\phi})_{ij} = -E\left(\frac{\partial l^2}{\partial\phi_i\phi_j}l(\boldsymbol{\phi};\boldsymbol{Y},\boldsymbol{W})\right) \quad (12)$$

است. واریانس $\hat{\beta}_{1,ML}$ را می‌توان به شکل زیر

$$\widehat{var}(\hat{\beta}_{1,ML}) = I^{-1}(\boldsymbol{\phi})_{22}\Big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}_{ML}} \quad (13)$$

برآورد کرد. لازم به ذکر است واریانس مجانبی رابطهٔ (۱۳)، به تحقق‌های مستقل و هم‌توزیع $[\boldsymbol{Y}\ \ \boldsymbol{W}]^T$ وابسته است. اما در اغلب کاربردها، فقط یک بردار $[\boldsymbol{Y}\ \boldsymbol{W}]^T$ مشاهده شده است. بنابراین برآوردگر واریانس (۱۳) ممکن است خیلی خوب عمل نکند.

۳ مطالعهٔ شبیه‌سازی

به منظور مقایسهٔ رویکردهای ارائه شده در این مقاله برای رگرسیون داده‌های بد تراز فضایی، از یک مطالعهٔ شبیه‌سازی استفاده می‌کنیم. در این مطالعه چندین سناریوی مختلف با $N = 500$ مجموعه دادهٔ شبیه سازی شده برای هر سناریو استفاده می‌شود. مقادیر متغیر توضیحی بنابر رابطهٔ

$$\boldsymbol{X} = \boldsymbol{g} + \boldsymbol{\delta}$$

تولید می‌شود که در آن

$$\boldsymbol{g} \sim N(\mu 1, R(\phi,\nu)),$$

و برای R تابع همبستگی مترن به صورت

$$\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{2\sqrt{\nu}t}{\phi\pi}\right)^{\nu}K_\nu\left(\frac{2\sqrt{\nu}t}{\phi\pi}\right),$$

در نظر گرفته می‌شود که در آن t فاصله، ϕ دامنهٔ فضایی، $\nu > 0$ پارامتر همواری و K_ν تابع بسل اصلاح شدهٔ نوع دوم از مرتبهٔ ν است. همچنین $\boldsymbol{\delta}$ تغییرات مقیاس کوچک است و فرض می‌شود

$$\boldsymbol{\delta} \sim N(\boldsymbol{o}, \sigma_\delta^2 I_{n_w}),$$

برای بردار متغیر پاسخ فرض می‌کنیم

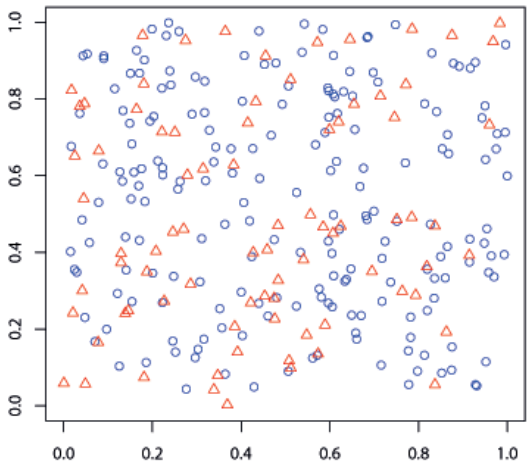
$$\boldsymbol{Y} \sim N(\beta_0 1 + \beta_1\boldsymbol{X}, \sigma_\epsilon^2 I),$$

و در کلیه سناریوها $\beta_0 = 0$ و $\beta_1 = 1$ قرار می‌دهیم. فرض استقلال خطاهای متغیر پاسخ بیان می‌کند که تنها عنصر مولد خودهمبستگی فضایی در متغیر پاسخ، متغیر توضیحی است.

معمولاً تعداد موقعیت‌هایی که متغیر پاسخ مشاهده می‌شود بیشتر از تعداد موقعیت‌هایی است که متغیر توضیحی مشاهده می‌شود. بنابراین برای تمام مجموعه داده‌ها فرض می‌کنیم $n_w = 80$ و $n_y = 200$ است. لازم به ذکر است که بردارهای $\boldsymbol{X}$ و $\boldsymbol{Y}$ را ابتدا در $n_w + n_y = 280$ موقعیت به طور کامل شبیه‌سازی و سپس آن‌ها را بد تراز می‌کنیم. به این صورت که در هر مجموعه داده ۸۰ مقدار اول بردار $\boldsymbol{Y}$ و ۲۰۰ مقدار

آخر بردار $\boldsymbol{X}$ را حذف می‌کنیم و به این ترتیب داده‌های فضایی بد تراز حاصل می‌شود.

فضای مطالعه را مربع واحد در نظر می‌گیریم و به تصادف ۲۸۰ موقعیت نقطه‌ای در آن انتخاب و فرض می‌کنیم در ۸۰ نقطه از این موقعیت‌ها، متغیر توضیحی و در ۲۰۰ نقطهٔ دیگر، متغیر پاسخ مشاهده شده است. شکل ۲ فضای مطالعه و موقعیت‌های داده‌های بد تراز فضایی را ارائه می‌دهد که در آن ۸۰ موقعیت مربوط به مشاهدات متغیر توضیحی با $\triangle$ و ۲۰۰ موقعیت مربوط به مشاهدات متغیر پاسخ با $\circ$ نشان داده شده است.



شکل ۲: فضای مطالعه و موقعیت‌های داده‌های بد تراز فضایی در مطالعهٔ شبیه‌سازی

اکنون پنج سناریو با پارامترهای هموواری و دامنه‌های فضایی مختلف به صورت زیر در نظر می‌گیریم:

- سناریوی A: یک فرایند فضایی با هموواری بالا و دامنهٔ ۱/۶

$$g \sim N(\circ, R(۱/۶, ۱)), \quad \delta \sim N(\circ, \sigma_\delta^۲ I_{۸\circ}),$$
$$\sigma_\delta^۲ = \circ/۱^۲, \quad \sigma_\epsilon^۲ = \circ/۸^۲$$

- سناریوی B: یک فرایند فضایی با هموواری بالا و دامنهٔ۱

$$g \sim N(\circ, R(۱, ۱)), \quad \delta \sim N(\circ, \sigma_\delta^۲ I_{۸\circ}),$$
$$\sigma_\delta^۲ = \circ/۲^۲, \quad \sigma_\epsilon^۲ = \circ/۸^۲$$

- سناریوی C: یک فرایند فضایی با هموواری کم و دامنهٔ ۱

$$g \sim N(\circ, R(۱, \circ/۵)), \quad \delta \sim N(\circ, \sigma_\delta^۲ I_{۸\circ}),$$
$$\sigma_\delta^۲ = \circ/۲^۲, \quad \sigma_\epsilon^۲ = \circ/۸^۲$$

- سناریوی D: یک فرایند فضایی با هموواری کم و دامنهٔ ۱ و اثر قطعه‌ای بالا

$$g \sim N(\circ, R(۱, \circ/۵)), \quad \delta \sim N(\circ, \sigma_\delta^۲ I_{۸\circ}),$$
$$\sigma_\delta^۲ = \circ/۴^۲, \quad \sigma_\epsilon^۲ = \circ/۸^۲$$

- سناریوی E: یک فرایند فضایی با هموواری کم و دامنهٔ ۰/۳

$$g \sim N(\circ, R(\circ/۳, \circ/۵)), \quad \delta \sim N(\circ, \sigma_\delta^۲ I_{۸\circ}),$$
$$\sigma_\delta^۲ = \circ/۲^۲, \quad \sigma_\epsilon^۲ = \circ/۸^۲$$

بدیهی است هرچه فرایند ناهموارتر باشد، برآورد پارامترها چالش برانگیزتر خواهد بود. شکل ۳ تحقق‌هایی از فرایند پیشگو را در این پنج سناریو نشان می‌دهد.

ابتدا در هر سناریو با استفاده از مجموعه داده‌های کامل (تراز)، مدل رگرسیونی خطی ساده برازش و پارامترهای مدل برآورد می‌شوند. سپس داده‌های بد تراز در نظر گرفته می‌شوند و مدل رگرسیونی، با روش‌هایی که ذکر شد، برازش داده می‌شود.

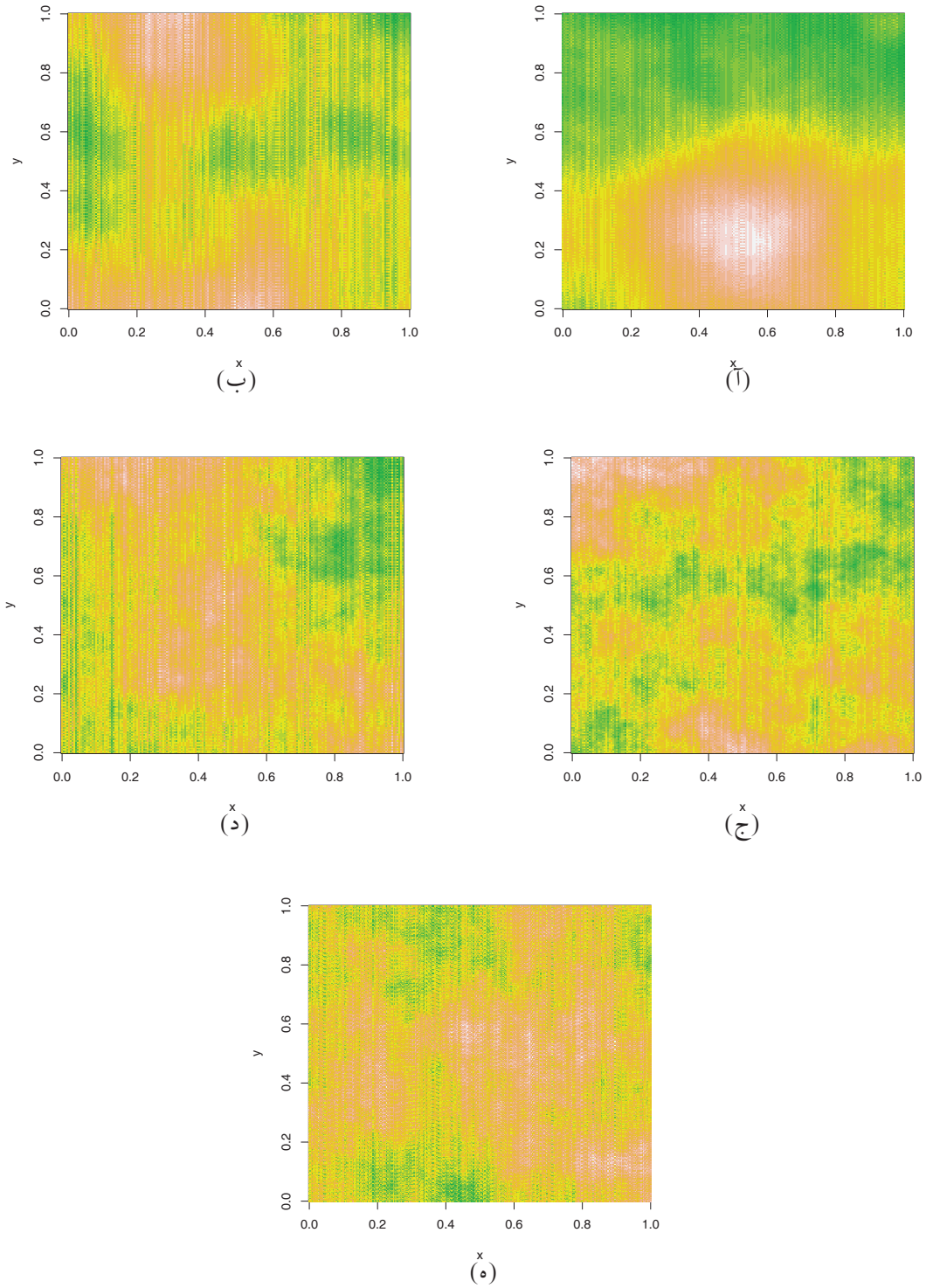
در روش باجایگذاری، برای هر مجموعه داده ابتدا بهترین پیشگوی خطی نااریب (کریگی) متغیر توضیحی در مکان‌های متناظر با متغیر پاسخ را با استفاده از تابع krige.conv در بستهٔ geoR در R محاسبه و سپس پارامترهای مدل رگرسیون خطی $\boldsymbol{Y}$ روی مقادیر کریگی حاصل برآورد می‌شوند.

در رویکرد شبیه‌سازی، ابتدا ماتریس هم‌تغییرنگار مانا را با استفاده از تابع cov.spatial در بستهٔ geoR و تجزیهٔ چولسکی این ماتریس را با استفاده از تابع chol محاسبه می‌کنیم. سپس برای هر مجموعه داده به تعداد $M = ۱۰۰$ تکرار عمل شبیه‌سازی متغیر توضیحی در موقعیت‌های متناظر با متغیر پاسخ را، همانگونه که در بخش ۲.۲ شرح داده شد، انجام می‌دهیم و در نتیجه ۱۰۰ مقدار برای پارامتر شیب خط رگرسیون حاصل می‌شود. میانگین این ۱۰۰ مقدار، برآورد $\beta_۱$ برای آن مجموعه داده است.

به‌علاوه، داده‌های شبیه‌سازی‌شده را با عمل کالبیدن، کالیبره کرده و سپس با استفاده از مقادیر کالیبره شده، رگرسیون را انجام می‌دهیم.

در رویکرد رگرسیون کالبیدنی برای هر مجموعه داده تمامی عناصر بردار $\boldsymbol{W}$ را به نوبت با استفاده از $n_w - ۱$ عنصر دیگر این بردار پیشگویی و سپس بردار $\boldsymbol{W}$ را روی بردار مقادیر کریگی آن مدل و بردار پارامترهای γ برآورد می‌کنیم. سپس با استفاده از بردار $\hat{\gamma}$، مقادیر کریگیدن در رویکرد

باجایگذاری را، همانگونه که در بخش ۳.۲ شرح داده شد، کالیبره و در نهایت رگرسیون خطی Y روی مقادیر کالیبره شده برازش داده می‌شوند. سپس، برآوردهای ماکسیمم درستنمایی پارامترهای مدل رگرسیونی محاسبه می‌شوند.

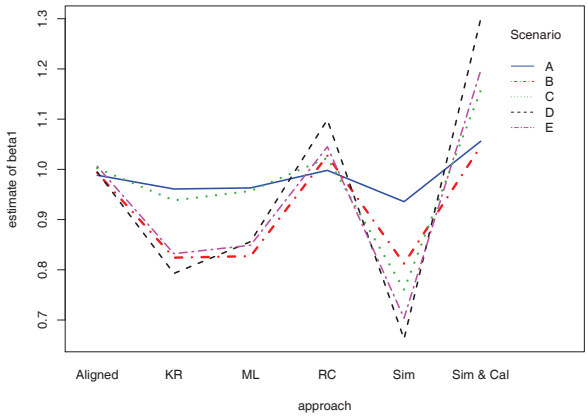


شکل ۳. تحقق‌هایی از فرایند پیشگو (آ) سناریوی A، (ب) سناریوی B، (ج) سناریوی C، (د) سناریوی D، (ه) سناریوی E.

نتایج حاصل از ۵۰۰ شبیه‌سازی در جدول ۶ خلاصه شده است. در این جدول، برآورد شیب خط رگرسیون، برآورد اریبی، متوسط انحراف معیار، میانگین توان‌های دوم خطا و میزان پوشش بازه‌های اطمینان ۹۵% گزارش شده است.

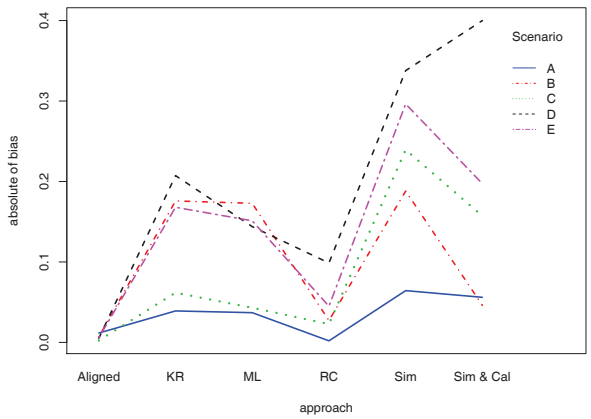
اکنون برای بررسی این نتایج از نمودارهای زیر استفاده می‌کنیم. توجه داریم که در این نمودارها مقادیر مربوط به رگرسیون خطی مجموعه داده‌های کامل (تراز) را با $Aligned$، رویکرد کریگ و رگرس را با KR، رویکرد رگرسیون کالیبدنی را با RC، رویکرد ماکسیمم درستنمایی را با ML، رویکرد شبیه‌سازی متغیر توضیحی را با Sim و رویکرد شبیه‌سازی و کالیبدن را با $Sim\&Cal$ نشان داده‌ایم.

شکل ۴ برآوردهای β_1 را تحت رویکردهای مختلف نشان می‌دهد. ملاحظه می‌شود که روش‌های رگرسیون کالیبدنی و شبیه‌سازی و کالیبدن، برای β_1 بیش‌برآوردی و رویکردهای کریگ و رگرس، ماکسیمم درستنمایی و شبیه‌سازی کم‌برآوردی دارند. از طرفی هرچه همواری فرایند کمتر می‌شود، انحراف برآوردها از مقدار واقعی پارامتر یعنی ۱ بیشتر می‌شود.



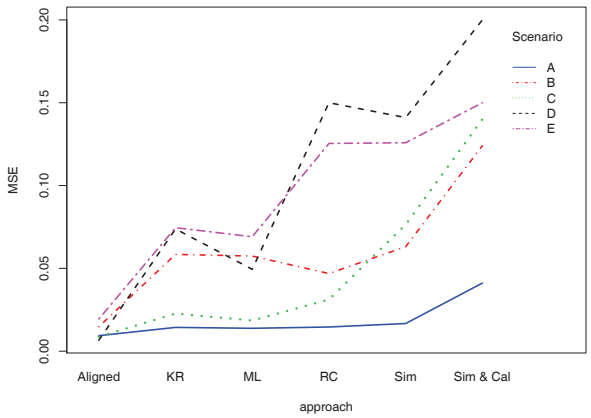
شکل ۴: برآورد شیب خط رگرسیون تحت رویکردهای مختلف

برای بررسی بهتر این موضوع، شکل ۵ را ببینید. در این شکل، قدر مطلق اریبی موجود در برآورد β_1 تحت رویکردهای مختلف ارائه شده است. مشاهده می‌شود به ویژه هنگامی که فرایند ناهموار می‌شود، اریبی روش‌های KR، ML و Sim به شدت افزایش می‌یابد. در مقایسه، روش RC در تمامی سناریوها اریبی بسیار کمی ایجاد کرده است و به‌عبارت دیگر اریبی موجود در روش KR را به خوبی تصحیح کرده است. همچنین مشاهده می‌شود که روش $Sim\&Cal$ اریبی روش Sim را به مقدار قابل توجهی تعدیل کرده است.



شکل ۵: قدر مطلق اریبی موجود در برآورد پارامتر شیب خط رگرسیون در رویکردهای مختلف

شکل ۶ مقدار MSE شیب خط رگرسیون را برای رویکردهای مختلف نشان می‌دهد. ملاحظه می‌شود که در سناریوی A، MSE تمامی رویکردها به جز رویکرد $Sim\&Cal$ تقریباً یکسان هستند. اما در فرایندهای ناهموارتر، MSE روش‌های مختلف با یکدیگر متفاوت هستند.



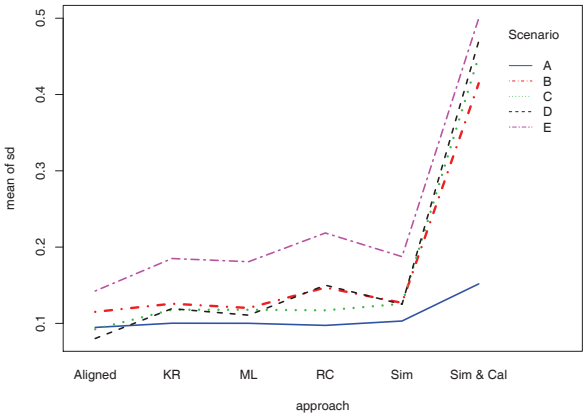
شکل ۶: $MSE(\hat{\beta}_1)$ در رویکردهای مختلف

به‌طور کلی، روش $Sim\&Cal$ بیشترین MSE و روش ML کمترین MSE را دارد. به‌علاوه مشاهده می‌شود که MSE روش RC بیشتر از روش KR است. اما همانطور که در بخش‌های قبل توضیح داده شد و در این مطالعهٔ شبیه‌سازی نیز مشاهده شد، روش KR برآوردهایی اریب ایجاد می‌کند و از این رو روش RC نسبت به این روش برتری دارد.

جدول ۲: نتایج مطالعهٔ شبیه‌سازی برای $\hat{\beta}_1$

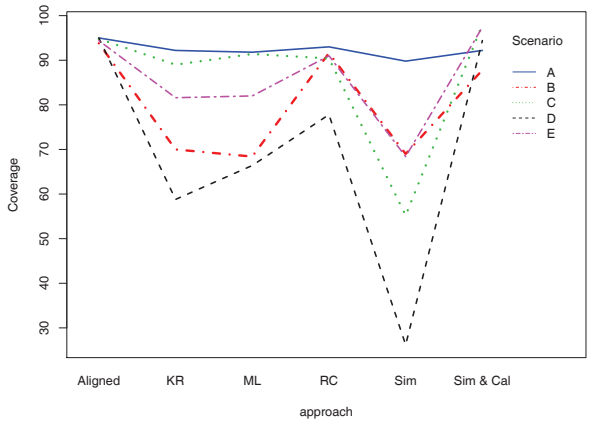
سناریو	رویکرد	$\hat{\beta}_1$	اریبی	انحراف معیار	MSE	پوشش (%)
A	داده‌های تراز	۰/۹۸۸	−۰/۰۱۲	۰/۰۹۵	۰/۰۰۹	۹۵/۰
	کریگ و رگرس	۰/۹۶۱	−۰/۰۳۹	۰/۱۰۰	۰/۰۱۴	۹۲/۲
	رگرسیون کالیبدنی	۰/۹۹۸	−۰/۰۰۲	۰/۰۹۷	۰/۰۱۵	۹۳/۰
	ماکسیمم درستنمایی	۰/۹۶۳	−۰/۰۳۷	۰/۱۰۰	۰/۰۱۴	۹۱/۸
	شبیه‌سازی	۰/۹۳۶	−۰/۰۶۴	۰/۱۰۳	۰/۰۱۷	۸۹/۸
	شبیه‌سازی و کالیبدن	۱/۰۵۶	۰/۰۵۶	۰/۱۵۲	۰/۰۴۱	۹۲/۲
B	داده‌های تراز	۰/۹۹۳	−۰/۰۰۷	۰/۱۱۵	۰/۰۱۵	۹۳/۸
	کریگ و رگرس	۰/۸۲۴	−۰/۱۷۶	۰/۱۲۶	۰/۰۵۸	۷۰/۰
	رگرسیون کالیبدنی	۱/۰۲۷	۰/۰۲۷	۰/۱۴۵	۰/۰۴۶	۹۱/۶
	ماکسیمم درستنمایی	۰/۸۲۷	−۰/۱۷۳	۰/۱۲۰	۰/۰۵۷	۶۸/۴
	شبیه‌سازی	۰/۸۱۲	−۰/۱۸۸	۰/۱۲۷	۰/۰۶۳	۶۹/۰
	شبیه‌سازی و کالیبدن	۱/۰۴۶	۰/۰۴۶	۰/۴۱۸	۰/۱۲۴	۸۷/۸
C	داده‌های تراز	۱/۰۰۲	۰/۰۰۲	۰/۰۹۲	۰/۰۰۹	۹۴/۸
	کریگ و رگرس	۰/۹۳۸	−۰/۰۶۱۸	۰/۱۱۸	۰/۰۲۲۸	۸۹/۰
	رگرسیون کالیبدنی	۱/۰۲۲	۰/۰۲۲	۰/۱۱۷	۰/۰۳۱	۹۰/۴
	ماکسیمم درستنمایی	۰/۹۵۶	−۰/۰۴۳	۰/۱۱۷	۰/۰۱۹	۹۱/۴
	شبیه‌سازی	۰/۷۶۱	−۰/۲۳۹	۰/۱۲۶	۰/۰۷۶	۵۵/۲
	شبیه‌سازی و کالیبدن	۱/۱۵۷	۰/۱۵۷	۱/۷۵۱	۱/۱۷۹	۹۸/۰
D	داده‌های تراز	۰/۹۹۵	−۰/۰۰۵	۰/۰۸۰	۰/۰۰۷	۹۵/۰
	کریگ و رگرس	۰/۷۹۳	−۰/۲۰۷	۰/۱۱۹	۰/۰۷۴	۵۸/۸
	رگرسیون کالیبدنی	۱/۰۹۹	۰/۰۹۹	۰/۱۵۰	۰/۴۵۹	۷۷/۸
	ماکسیمم درستنمایی	۰/۸۵۶	−۰/۱۴۴	۰/۱۱۱	۰/۰۴۹	۶۶/۴
	شبیه‌سازی	۰/۶۶۲	−۰/۳۳۸	۰/۱۲۵	۰/۱۴۱	۲۶/۲
	شبیه‌سازی و کالیبدن	۱/۴۶۹	۰/۴۶۹	۵/۰۰۹	۴۱/۰	۹۴/۴
E	داده‌های تراز	۱/۰۰۶	۰/۰۰۶	۰/۱۴۳	۰/۰۱۹	۹۴/۴
	کریگ و رگرس	۰/۸۴۷	−۰/۱۵۳	۰/۱۸۲	۰/۰۷۰	۸۵/۸
	رگرسیون کالیبدنی	۱/۰۶۸	۰/۰۶۸	۰/۲۲۱	۰/۱۰۹	۹۱/۸
	ماکسیمم درستنمایی	۰/۶۹۶	−۰/۳۰۴	۰/۱۴۳	۰/۱۱۸	۴۶/۶
	شبیه‌سازی	۰/۷۱۸	−۰/۲۸۲	۰/۱۸۵	۰/۱۱۹	۶۹/۶
	شبیه‌سازی و کالیبدن	۱/۱۷۱	۰/۱۷۱	۲/۲۲۱	۴/۲۰۳	۹۶/۸

شکل ۷ میانگین انحراف معیار $\hat{\beta}_1$ را در هر سناریو برای هر رویکرد نشان می‌دهد. ملاحظه می‌شود که در تمامی سناریوها رویکرد $Sim\&Cal$ و پس از آن رویکرد RC بیشترین انحراف معیار را دارند. توجه داریم که میانگین انحراف معیار در روش $Sim\&Cal$ بسیار زیاد است اما میانگین انحراف معیار در روش RC تفاوت چندانی با سایر روش‌ها ندارد.



شکل ۷: میانگین انحراف معیار در رویکردهای مختلف

شکل ۸ درصد پوشش بازه‌های اطمینان ۹۵% را نشان می‌دهد. این درصد برای هر سناریو به این صورت محاسبه شده است که برای هر مجموعه داده، بازه اطمینانی به صورت $\hat{\beta}_1 \pm 1/96 sd(\hat{\beta}_1)$ محاسبه شد و در نهایت ۵۰۰ بازه اطمینان به‌دست آمد. سپس بررسی شد که چند درصد از این بازه‌ها مقدار واقعی β_1 یعنی ۱ را در برمی‌گیرند. شکل ۸ نشان می‌دهد هنگامی که فرایند ناهموار می‌شود، درصد پوشش بازه‌های اطمینان ۹۵% در رویکردهای KR، ML و Sim به شدت کاهش می‌یابد. این درصد در رویکرد $Sim\&Cal$ بالا است و بدیهی است که دلیل آن بزرگی انحراف معیار این روش است. بازه‌های اطمینان ۹۵% در روش RC در تمامی سناریوها پوشش بالایی برای مقدار واقعی β_1 فراهم کرده‌اند و از این نظر نسبت به سایر روش‌های برآورد برتری دارد.



شکل ۸: درصد پوشش بازه‌های اطمینان ۹۵% در رویکردهای مختلف

بحث و نتیجه‌گیری

مسئلهٔ رگرسیون داده‌های بد تراز فضایی یکی از موضوعات مورد توجه محققان و تحلیل‌گران در بسیاری از زمینه های علوم است. در این مقاله رویکردهای مختلفی برای رگرسیون داده‌های بد تراز فضایی شامل رویکرد باجایگذاری، شبیه‌سازی، رگرسیون کالیبدنی و ماکسیمم درستنمایی ارائه شدند. در دو رویکرد باجایگذاری و شبیه‌سازی وجود خطای اندازه‌گیری منجر به ایجاد اریبی در برآورد پارامترهای مدل رگرسیونی شدند. لذا رویکرد رگرسیون کالیبدنی برای تعدیل این اریبی پیشنهاد شد. نتایج مطالعهٔ شبیه‌سازی نشان داد این رویکرد اگرچه MSE را مقداری افزایش می‌دهد اما اریبی موجود در روش‌های دیگر را تا حد زیادی تعدیل می‌کند. همچنین در این روش، میزان پوشش اسمی بازه‌های اطمینان پارامتر شیب رگرسیون قابل توجه است.

لازم به ذکر است که این مقاله در راستای معرفی روش‌هایی برای کاهش اریبی موجود در برآورد ضرایب مدل رگرسیونی نگارش یافته است. ارزیابی پیشگویی توسط مدل‌های حاصل از این رویکردها از نکات شایان توجه است که توسط نگارندگان در حال بررسی است.

مراجع

[1] Gryparis, A., C. J. Paciorek, A. Zeka, J. Schwartz, and B. Coull (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10**, 258–274.

[2] Madsen, L., D. Ruppert, and N. S. Altman (2008). Regression with spatially misaligned data. *Environmetrics* **19**, 453–467.

[3] Szpiro, A. A., L. Sheppard, and T. Lumley (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics* **0**, 1–14.

[4] Young, L. J., C. A. Gotway, G. Kearney, and C. Duclos (2009). Assessing uncertainty in support-adjusted spatial misalignment problems. *Communications in Statistics— Theory and Methods* **38**, 3249–3264.

[5] Zhu, L., B. Carlin, and A. Gelfand (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma er visits in atlanta. *Environmetrics* **14**, 537–557.