

## رگرسیون خطی در حضور داده‌های بد تراز

عادلۀ عصاره<sup>۱</sup> فیروزه ریواز<sup>۲</sup>

چکیده:

در این مقاله چهار رویکرد به مسئله برازش یک مدل رگرسیون خطی در حضور داده‌های بد تراز فضایی ارائه می‌شود. این رویکردها عبارتند از روش باجایگذاری، شبیه‌سازی، رگرسیون کالبدنی 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 برگسن شده و این خطا نیز منجر به ایجاد اریبی در برآورد شیب مدل رگرسیونی می‌شود. برای تعدیل این اریبی، رویکرد رگرسیون کالبدنی 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 به دلیل نداشتن فرم تحلیلی، امکان‌پذیر نیست. در یک مطالعه شبیه‌سازی، عملکرد کلیه رویکردها تحت چندین مدل فضایی برای متغیر توضیحی مورد بررسی قرار می‌گیرد. مشاهده می‌شود رگرسیون کالبدنی می‌تواند به میزان قابل توجهی اریبی برآوردگر شیب خط رگرسیونی را نسبت به روش‌های دیگر کاهش دهد. به علاوه، میزان پوشش اسمی بازه اطمینان شیب خط رگرسیونی توسط این روش قابل توجه است.

**واژه‌های کلیدی:** داده‌های بد تراز فضایی، رویکرد باجایگذاری، رگرسیون کالبدنی، خطای برگسن.

### ۱ مقدمه

...،  $t_m$  مشاهده شده اند. این نوع داده‌ها، به داده‌های بد تراز فضایی معروف هستند.

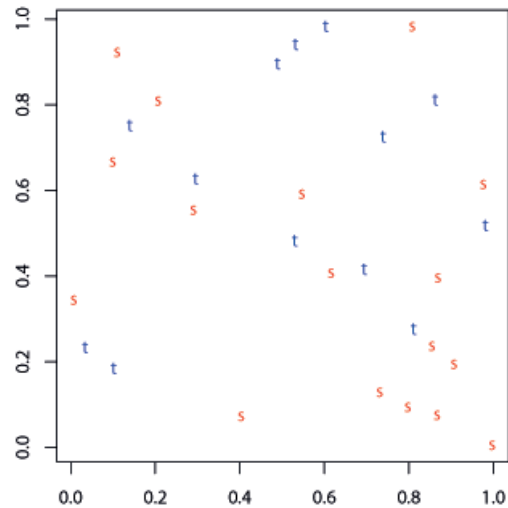
برای مثال فرض کنید ارتباط میان میزان نوعی آلاینده هوا در شهر تهران و تعداد بیماران قلبی در بیمارستان‌های این شهر در مقطعی از زمان مورد توجه است. اما از آنجا که مکان ایستگاه‌های هواشناسی متفاوت با مکان بیمارستان‌ها است، مشاهدات مربوط به این دو متغیر از نوع داده‌های بد تراز فضایی هستند. برای درک بهتر موضوع،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

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. به عبارت 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  $s_1, s_2, \dots, s_n$  مشاهده شده است، در حالی که متغیرهای توضیحی در موقعیت‌های دیگری مانند  $t_1, t_2, \dots, t_n$

<sup>۱</sup>گروه آمار، دانشگاه شهید بهشتی

<sup>۲</sup>گروه آمار، دانشگاه شهید بهشتی

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



شکل ۱: داده‌های بد تراز فضایی

نخستین بار ژو و همکاران (۲۰۰۳) مسئله 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 توان‌های 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 مطالعه 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 مسئله برازش یک مدل خطی به داده‌های بد تراز فضایی می‌پردازیم. برای حل مسئله 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 مسئله برازش مدل رگرسیون خطی در حالت بد تراز فضایی، برای متغیرهای کمی شامل رویکردهای باجایگذاری، شبیه‌سازی، رگرسیون کالبدینی 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 مطالعه شبیه‌سازی، عملکرد روش‌های مختلف مورد بررسی قرار می‌گیرد.

## ۲ مدل آماری

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

$$Y(s_i) = \beta_0 + \beta_1 X(s_i) + Z^T(s_i) \beta_z + \epsilon(s_i), \quad i = 1, 2, \dots, n_y \quad (1)$$

را در نظر بگیرید، که در آن  $\epsilon(s_i) \sim N(0, \sigma_\epsilon^2)$ . همچنین  $\beta_z$  و  $\beta_1$  پارامترهای نامعلوم،  $X(s_i)$  متغیر توضیحی در مکان  $s_i$  و  $Z(s_i)_{(q \times 1)}$  بردار پیشگوهای اندازه‌گیری شده بدون خطا در مکان  $s_i$  است.

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

در مدل (۱) منجر به خطای برکسن<sup>۴</sup> می شود (گری پاریس و همکاران، ۲۰۰۹). با جایگذاری (۴) در مدل رگرسیون داریم:

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

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

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

است. بنابراین اگرچه برآوردگر OLS برای  $\beta_1$  ناریب است (با این فرض که  $\hat{X}$  برای  $X$  ناریب است)، اما واریانس آن دیگر واریانس واقعی  $\hat{\beta}_1$  نخواهد بود. در این گونه مواقع از روش کمترین توان های دوم وزنی (WLS) استفاده می شود. برای این منظور عناصر قطری ماتریس کوواریانس خطا را به عنوان وزن هایی برای مدل رگرسیون در نظر می گیرند. اما چون در  $\text{var}(\eta) = \beta_1^2 \Sigma_V + \sigma_\epsilon^2 I_{n_y}$  پارامترهای  $\beta_1$  و  $\sigma^2$  نامعلوم هستند، لذا این ماتریس و در نتیجه عناصر قطری آن نامعلوم هستند و بنابراین نمی توان از روش WLS برای برآورد ضرایب رگرسیونی استفاده کرد.

لازم به ذکر است که حالتی خاص از روش باجایگذاری به روش کریگ و رگرس<sup>۵</sup> معروف است که در آن  $\hat{X}$  پیشگوی کریگی در نظر گرفته می شود و به عنوان متغیر توضیحی در مدل رگرسیون به کار می رود (مادسن و همکاران، ۲۰۰۸).

## ۲.۲ رویکرد شبیه سازی

روش های پیشگویی فضایی معمولاً رویه ای هموار از فرایند مورد مطالعه ارائه می دهند. لذا تغییرپذیری مقادیر پیشگویی بسیار کمتر از تغییرپذیری واقعی متغیر موردنظر است. بنابراین در زمین آمار اغلب به جای روش های هموارسازی فضایی از شبیه سازی زمین آماری استفاده می شود. گری پاریس و همکاران (۲۰۰۹) رویکرد شبیه سازی را به این صورت ارائه دادند که از برآورد توزیع متغیر توضیحی به شرط داده ها،  $M$  نمونه  $\epsilon$  است. به علاوه  $V \sim N(0, \Sigma_V)$ . استفاده از  $\hat{X}$  به جای  $X$  از

جدول ۱: جدول داده های بد تراز فضایی

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

## ۱.۲ رویکرد باجایگذاری

فرض کنید  $Y(s_i)$  در مدل (۱) نرمال باشد. برای برازش مدل رگرسیونی به روش باجایگذاری<sup>۳</sup> ابتدا  $X$  را با استفاده از یکی از روش های پیشگویی فضایی و مبتنی بر بردار  $W$  پیشگویی کرده و آن را با  $\hat{X}$  نشان می دهیم. سپس  $\hat{X}$  به عنوان متغیر توضیحی در مدل رگرسیون به کار برده می شود. در نهایت برآوردگرهای کمترین توان های دوم معمولی برای  $\beta$  و واریانس آن به شکل زیر محاسبه می شوند:

$$\hat{\beta}_{\text{plug-in}} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y \quad (2)$$

$$\text{Var}(\hat{\beta}_{\text{plug-in}}) = \sigma_\epsilon^2 (\hat{X}^T \hat{X})^{-1} \quad (3)$$

که در آن  $\hat{X} = [1_{n_y \times 1} \quad \hat{X} \quad Z]$ . این روش برآوردگرهای نارایی را برای پارامترهای خط رگرسیون نتیجه می دهد، اما استفاده از  $\hat{X}$  به جای  $X$  در مدل رگرسیونی، عدم حتمیتی را وارد مدل می کند و موجب ایجاد همبستگی در ساختار خطا می شود. برای توضیح این مطلب، توجه کنید که مقادیر حاصل از روش های پیشگویی فضایی تغییرپذیری کمتری نسبت به مقادیر واقعی دارند، لذا می توان نوشت:

$$X = \hat{X} + V \quad (4)$$

که در آن  $V = X - \hat{X}$  خطای مربوط به پیشگویی  $X$  و مستقل از  $\epsilon$  است. به علاوه  $V \sim N(0, \Sigma_V)$ . استفاده از  $\hat{X}$  به جای  $X$

<sup>3</sup> Plug-in Approach

<sup>4</sup> Berkson error

<sup>5</sup> Krige-and-Regress approach

دارند. لذا روش ارائه شده در (۵) نسبت به روش گری پاریس و همکاران (۲۰۰۹)، با آنچه در عمل اتفاق می‌افتد، سازگارتر است.

اکنون با به کار بردن هریک از  $X_{(k)}$  های تولیدشده به عنوان متغیر توضیحی در (۱)، مدل را برازش می‌دهیم و در نتیجه  $M$  برآورد برای  $\beta_1$  و واریانس درون شبیه‌سازی  $^6$  حاصل می‌شود که آن‌ها را به ترتیب با  $\hat{\beta}_{1(k)}$  و  $W_{(k)}$ ،  $k = 1, \dots, M$ ، نشان می‌دهیم. سپس برای به دست آوردن یک برآورد کلی از  $\hat{\beta}_{1(k)}$  ها میانگین گرفته می‌شود:

$$\hat{\beta}_{1M} = \frac{1}{M} \sum_{k=1}^M \hat{\beta}_{1(k)}.$$

تغییرپذیری این برآوردگر دو جزء دارد: میانگین واریانس درون شبیه‌سازی ( $\bar{W}_M$ ) و واریانس بین شبیه‌سازی  $(B_M)^7$ ، که به ترتیب عبارت از

$$\bar{W}_M = \frac{1}{M} \sum_{k=1}^M W_{(k)},$$

و

$$B_M = \frac{1}{M-1} \sum_{k=1}^M (\hat{\beta}_{1(k)} - \hat{\beta}_{1M})^2,$$

هستند. بنابراین تغییرپذیری کل به صورت

$$\widehat{Var}(\hat{\beta}_{1M}) = \bar{W}_M + \frac{M+1}{M} B_M,$$

خواهد بود که  $(1 + \frac{1}{M})$  تعدیلی برای  $M$  تعداد متناهی شبیه‌سازی است (یانگ و همکاران، ۲۰۰۹).

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

### ۳.۲ رویکرد رگرسیون کالبدنی

رگرسیون کالبدنی  $^8$  ( $RC$ ) 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

$s_{n_y}, \dots$  تولید می‌شود و هریک از این  $M$  نمونه به عنوان متغیر توضیحی در مدل رگرسیون (۱) مورد استفاده قرار می‌گیرد. در نتیجه  $M$  برآورد  $\hat{\beta}_{1(k)}$ ،  $k = 1, 2, \dots, M$ ، حاصل می‌شود و سپس برای به دست آوردن یک برآورد کلی از آن‌ها میانگین گرفته می‌شود. در این صورت واریانس  $\hat{\beta}_1$  برابر

$$Var(\hat{\beta}_1) = Var(E(\hat{\beta}_{1(k)} | X_{(k)})) + E(Var(\hat{\beta}_{1(k)} | X_{(k)})).$$

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

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

باشد.  $\Sigma$  را به صورت

$$\Sigma = LL^T$$

تجزیه می‌کنیم که در آن  $L = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}$  فرض کنید

$V = \begin{pmatrix} V_1 \\ V_{2(k)} \end{pmatrix}$  برداری از تحقق‌های مستقل و هم‌توزیع  $N(0, 1)$  باشد. آنگاه با استفاده از

$$\begin{aligned} \begin{pmatrix} W \\ X_{(k)} \end{pmatrix} &= \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} L_{11}^{-1} W \\ V_{2(k)} \end{pmatrix} \\ &= \begin{pmatrix} W \\ L_{21} L_{11}^{-1} W + L_{22} V_{2(k)} \end{pmatrix}, \\ k &= 1, \dots, M, \end{aligned} \quad (5)$$

می‌توان  $M$  تحقق از فرایند  $X$  در موقعیت‌های  $s_1, s_2, \dots, s_{n_y}$  تولید کرد. شایان ذکر است که مقادیر شبیه‌سازی‌شده بر اساس این روش در موقعیت‌هایی که مشاهده موجود است، مقادیر یکسانی با مشاهدات

<sup>6</sup>within-simulation variance

<sup>7</sup>between-simulation variance

<sup>8</sup>Regression Calibration Approach

در مدل رگرسیون (۱) خواهیم داشت:

$$\begin{aligned} Y &= \beta_0 \mathbf{1}_{n_y \times 1} + \beta_1 \hat{X}_{cal} + \beta_Z Z + \epsilon \\ &= \beta_0 \mathbf{1}_{n_y \times 1} + \beta_1 \hat{X} \hat{\gamma} + \beta_Z Z + \epsilon \end{aligned} \quad (7)$$

و لذا برآوردگر کالبدنی  $\beta$  به صورت

$$\begin{aligned} \hat{\beta}_{cal} &= \hat{\Gamma}^{-1} \hat{\beta}_{plug-in} \\ &= \hat{\Gamma}^{-1} (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y \end{aligned} \quad (8)$$

به دست می آید که  $\hat{\beta}_{plug-in}$  در رابطه (۲) ارائه شده است. همچنین واریانس  $\hat{\beta}_{cal}$  با استفاده از بسط تیلور به شکل

$$Var(\hat{\beta}_{cal}) = \hat{\Gamma}^{-1} [\beta_0^T \sigma_\epsilon^2 (D^T D)^{-1} + \sigma_\epsilon^2 (\hat{X}^T \hat{X})^{-1}] (\hat{\Gamma}^{-1})^T$$

به دست می آید که در آن  $D = [\mathbf{1} \quad \hat{W} \quad Z]$

## ۴.۲ رویکرد ماکسیمم درستنمایی

تا اینجا روش هایی که ارائه شد همگی مبتنی بر پیشگویی مقادیر مشاهده نشده متغیر توضیحی در موقعیت های متناظر با متغیر پاسخ و استفاده از مقادیر حاصل در مدل رگرسیونی بودند. اکنون روش ماکسیمم درستنمایی ارائه می شود که در آن برای برآورد ضرایب رگرسیونی، مستقیماً از بردار  $W$  استفاده می شود.

مدل (۱) با توابع میانگین و کوواریانس

$$E \begin{bmatrix} Y \\ X \\ W \end{bmatrix} = \begin{bmatrix} [\beta_0 + \beta_1 \mu_X + Z^T(s_i) \beta_Z]_{i=1}^{n_y} \\ \mu_X \mathbf{1}_{(n_y+n_w) \times 1}, \end{bmatrix} \quad (9)$$

و

$$cov \begin{bmatrix} Y \\ X \\ W \end{bmatrix} = \begin{bmatrix} \beta_0^T \Sigma_X + \Sigma_\epsilon & \beta_1 \Sigma_X & \beta_Z \Sigma_{XW} \\ \beta_1 \Sigma_X & \Sigma_X & \Sigma_{XW} \\ \beta_Z \Sigma_{XW}^T & \Sigma_{XW}^T & \Sigma_W \end{bmatrix} \quad (10)$$

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

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

که در آن

$$\Sigma = \begin{bmatrix} \beta_0^T \Sigma_X + \Sigma_\epsilon & \beta_1 \Sigma_{XW} \\ \beta_1 \Sigma_{XW}^T & \Sigma_W \end{bmatrix}.$$

شبیه سازی که در بخش های قبل ارائه شدند، می توان از رویکرد رگرسیون کالبدنی که توسط گری پاریس و همکاران (۲۰۰۹) ارائه شده است، استفاده نمود.

در این روش، ابتدا بردار مشاهدات متغیر توضیحی یعنی  $W = (X(t_1), X(t_2), \dots, X(t_{n_w}))^T$  که در مکان های متفاوت با مشاهدات متغیر پاسخ مشاهده شده اند، به دو بردار  $W_1$  با طول  $k$  و  $W_2$  با طول  $n_w - k$  افزایش می شود. سپس  $W_1$  با استفاده از بردار  $W_2$  پیشگویی شده و با  $\hat{W}_1$  نشان داده می شود. همچنین ماتریس  $Z$  را به طور متناظر با  $W$ ، به دو ماتریس  $Z_1$  و  $Z_2$  افزایش می کنیم. اکنون مبتنی بر  $W_1$  رگرسیون  $W_1$  روی  $\hat{W}_1$  برازش داده می شود. برای این منظور، مدل

$$\begin{aligned} W_1(t_i) &= \gamma_0 + \gamma_1 \hat{W}_1(t_i) + \gamma_Z^T Z_1(t_i) \\ &+ \delta(t_i), \quad i = 1, \dots, k, \end{aligned} \quad (6)$$

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 توان های دوم، برآوردهای  $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_Z^T)^T$  به دست می آید که از آن ها برای کالیبره کردن مقادیر پیشگویی  $\hat{X}$  استفاده می شود. با توجه به رابطه (۶) داریم:

$$E(X(s_i) | \hat{X}(s_i)) = \gamma_0 + \gamma_1 \hat{X}(s_i) + \gamma_Z^T Z(s_i), \quad i = 1, \dots, n_y,$$

با جایگذاری  $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_Z^T)$  به جای  $\gamma = (\gamma_0, \gamma_1, \gamma_Z^T)$  در عبارت فوق، مقادیر کالیبره شده  $\hat{X}$  به صورت

$$\hat{X}_{cal}(s_i) = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{X}(s_i) + \hat{\gamma}_Z^T Z(s_i), \quad i = 1, \dots, n_y,$$

به دست می آید و لذا می توان نوشت

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

که در آن  $\hat{X} = [\mathbf{1}_{n_y \times 1} \quad \hat{X} \quad Z]$  و  $\mathbf{1}$  برداری از یک هاست. ماتریس تبدیل  $\Gamma$  را به صورت زیر تعریف می کنیم:

$$\Gamma = \begin{bmatrix} \mathbf{1} & \gamma_0 & \mathbf{0}_{1 \times q} \\ \mathbf{0} & \gamma_1 & \mathbf{0}_{1 \times q} \\ \mathbf{0}_{q \times 1} & \gamma_Z & I_{q \times q} \end{bmatrix}.$$

لازم به ذکر است که برآورد ماتریس  $\Gamma$ ، ماتریس  $[\mathbf{1}_{n_y \times 1} | \hat{X} | Z]$  را به ماتریس  $[\mathbf{1}_{n_y \times 1} | \hat{X}_{cal} | Z]$  تبدیل می کند. حال با جایگذاری

بنابراین چگالی توأم  $Y$  و  $W$  عبارت از

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

است که در آن

$$V = \begin{bmatrix} V_Y \\ V_X \end{bmatrix} = \begin{bmatrix} Y - [\beta_0 + \beta_1 \mu_X + Z^T(s_i)\beta_Z]_{i=1}^{n_y} \\ W - \mu_X \mathbf{1}_{m \times 1} \end{bmatrix}.$$

فرض کنید  $\theta_X$  و  $\theta_\epsilon$  به ترتیب بردارهای پارامترهای نیم‌تغییرنگار فرایندهای پیشگو و خطا باشند. برآورد ماکسیمم درستمایی  $\beta$  با مینیم کردن مقدار

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

نسبت به پارامترهای  $\phi = [\beta^T \mu_X \theta_X^T \theta_\epsilon^T]^T$  به دست می‌آید.

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

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

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

و برآورد ML پارامتر شیب  $\beta_1$ ، عنصر دوم  $\hat{\phi}$  خواهد بود که با  $\hat{\beta}_{1,ML} = \hat{\phi}_{2,ML}$  نشان داده می‌شود.

### برآورد واریانس $\hat{\beta}_{1,ML}$

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

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

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

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

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

### ۳ مطالعه شبیه‌سازی

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

$$X = g + \delta$$

تولید می‌شود که در آن

$$g \sim N(\mu, R(\phi, \nu)),$$

و برای  $R$  تابع همبستگی مترن به صورت

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

در نظر گرفته می‌شود که در آن  $t$  فاصله  $\phi$  دامنه فضایی،  $\nu < 0$  پارامتر همواری و  $K_\nu$  تابع بسل اصلاح شده نوع دوم از مرتبه  $\nu$  است. همچنین  $\delta$  تغییرات مقیاس کوچک است و فرض می‌شود

$$\delta \sim N(0, \sigma_\delta^2 I_{n_w}),$$

برای بردار متغیر پاسخ فرض می‌کنیم

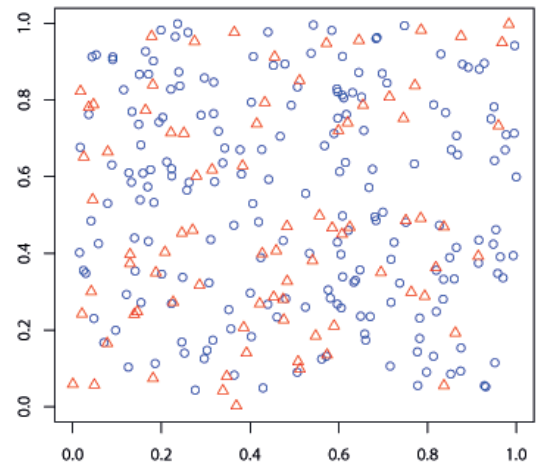
$$Y \sim N(\beta_0 \mathbf{1} + \beta_1 X, \sigma_\epsilon^2 I),$$

و در کلیه سناریوها  $\beta_0 = 1$  و  $\beta_1 = 0$  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  $X$  و  $Y$  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  $Y$  و ۲۰۰ مقدار

آخر بردار  $X$  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 نقطه 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  $\Delta$  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

مطالعه شبیه‌سازی

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

- سناریوی A: یک فرایند فضایی با همواری بالا و دامنه  $1/6$

$$g \sim N(\circ, R(1/6, 1)), \quad \delta \sim N(\circ, \sigma_\delta^2 I_{\Lambda^\circ}),$$

$$\sigma_\delta^2 = \circ/1^2, \quad \sigma_\epsilon^2 = \circ/8^2$$

- سناریوی B: یک فرایند فضایی با همواری بالا و دامنه ۱

$$g \sim N(\circ, R(1, 1)), \quad \delta \sim N(\circ, \sigma_\delta^2 I_{\Lambda^\circ}),$$

$$\sigma_\delta^2 = \circ/2^2, \quad \sigma_\epsilon^2 = \circ/8^2$$

- سناریوی C: یک فرایند فضایی با همواری کم و دامنه ۱

$$g \sim N(\circ, R(1, \circ/5)), \quad \delta \sim N(\circ, \sigma_\delta^2 I_{\Lambda^\circ}),$$

$$\sigma_\delta^2 = \circ/2^2, \quad \sigma_\epsilon^2 = \circ/8^2$$

- سناریوی D: یک فرایند فضایی با همواری کم و دامنه ۱ و اثر قطعه‌ای بالا

$$g \sim N(\circ, R(1, \circ/5)), \quad \delta \sim N(\circ, \sigma_\delta^2 I_{\Lambda^\circ}),$$

$$\sigma_\delta^2 = \circ/4^2, \quad \sigma_\epsilon^2 = \circ/8^2$$

- سناریوی E: یک فرایند فضایی با همواری کم و دامنه  $\circ/3$

$$g \sim N(\circ, R(\circ/3, \circ/5)), \quad \delta \sim N(\circ, \sigma_\delta^2 I_{\Lambda^\circ}),$$

$$\sigma_\delta^2 = \circ/2^2, \quad \sigma_\epsilon^2 = \circ/8^2$$

بدیهی است هرچه فرایند ناهموارتر باشد، برآورد پارامترها چالش برانگیزتر خواهد بود. شکل ۳ تحقیق‌هایی 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 بسته geoR در R محاسبه و سپس پارامترهای مدل رگرسیون خطی  $Y$  روی مقادیر کریگی حاصل برآورد می‌شوند.

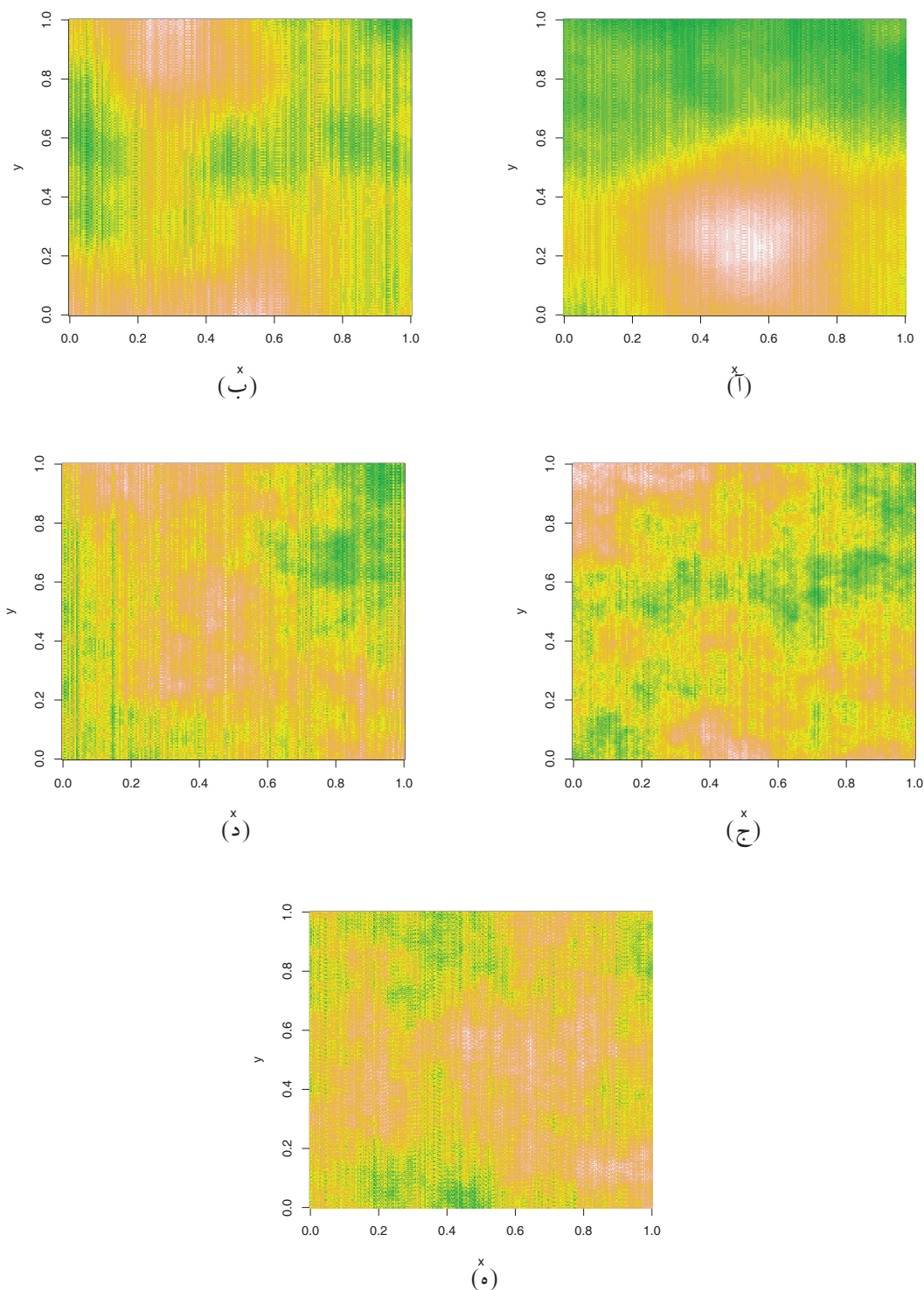
در رویکرد شبیه‌سازی، ابتدا ماتریس هم‌تغییرنگار مانا را با استفاده از تابع cov.spatial در بسته geoR و تجزیه چولسکی این ماتریس را با استفاده از تابع chol محاسبه می‌کنیم. سپس برای هر مجموعه داده به تعداد  $M = 100$  تکرار عمل شبیه‌سازی متغیر توضیحی در موقعیت‌های متناظر با متغیر پاسخ را، همانگونه که در بخش ۲.۲ شرح داده شد، انجام می‌دهیم و در نتیجه ۱۰۰ مقدار برای پارامتر شیب خط رگرسیون حاصل می‌شود. میانگین این ۱۰۰ مقدار، برآورد  $\beta_1$  برای آن مجموعه داده است.

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

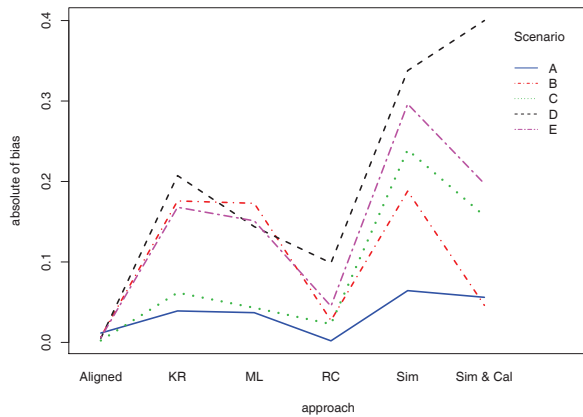
در رویکرد رگرسیون کالبدنی برای هر مجموعه داده تمامی عناصر بردار  $W$  را به نوبت با استفاده از  $n_w - 1$  عنصر دیگر این بردار پیشگویی و سپس بردار  $W$  را روی بردار مقادیر کریگی آن مدل و بردار پارامترهای  $\gamma$  برآورد می‌کنیم. سپس با استفاده از بردار  $\hat{\gamma}$ ، مقادیر کریگیدن در رویکرد

سپس، برآوردهای ماکسیمم درستنمایی پارامترهای مدل رگرسیونی محاسبه می‌شوند.

باجایگذاری را، همانگونه که در بخش ۳.۲ شرح داده شد، کالبره و در نهایت رگرسیون خطی  $Y$  روی مقادیر کالبره شده برازش داده می‌شوند.

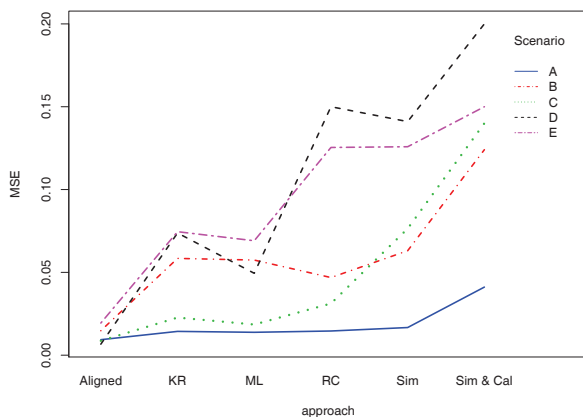


شکل ۳. تحقیقاتی از فرایند پیشگو (آ) سناریوی  $A$ ، (ب) سناریوی  $B$ ، (ج) سناریوی  $C$ ، (د) سناریوی  $D$ ، (ه) سناریوی  $E$ .



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.



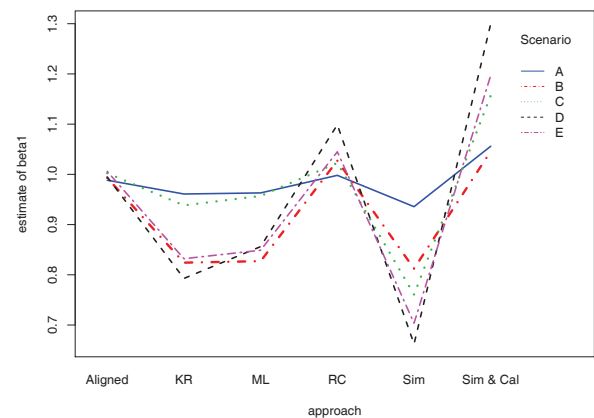
شکل ۶:  $MSE(\hat{\beta}_1)$  در رویکردهای مختلف

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 مطالعه شبیه‌سازی نیز مشاهده شد، روش *KR* برآوردهایی اریب ایجاد می‌کند و از این رو روش *RC* نسبت به این روش برتری دارد.

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 کالبدنی را با *RC*، رویکرد ماکسیمم درست‌نمایی را با *ML*، رویکرد شبیه‌سازی متغیر توضیحی را با *Sim* و رویکرد شبیه‌سازی و کالبدن را با *Sim&Cal* نشان داده‌ایم.

شکل ۴ برآوردهای  $\beta_1$  را تحت رویکردهای مختلف نشان می‌دهد. ملاحظه می‌شود که روش‌های رگرسیون کالبدنی و شبیه‌سازی و کالبدن، برای  $\beta_1$  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.

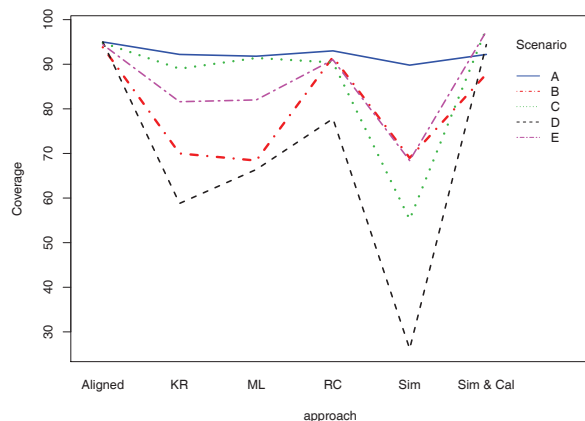


شکل ۴: برآورد شیب خط رگرسیون تحت رویکردهای مختلف

برای بررسی بهتر این موضوع، شکل ۵ را ببینید. در این شکل، قدر مطلق اریبی موجود در برآورد  $\beta_1$  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.

جدول ۲: نتایج مطالعه شبیه‌سازی برای  $\beta_1$ 

| 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 کالبدنی    | 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 کالبدن | 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 کالبدنی    | 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 کالبدن | 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 کالبدنی    | 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 کالبدن | 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 کالبدنی    | 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 کالبدن | 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 کالبدنی    | 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 کالبدن | ۱/۱۷۱           | ۰/۱۷۱   | ۲/۲۲۱        | ۴/۲۰۳  | ۹۶/۸     |



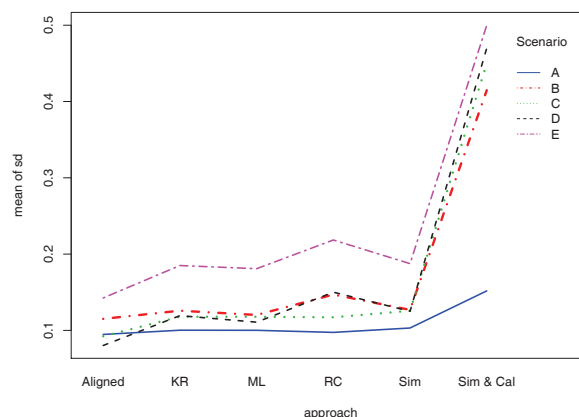
شکل ۸: درصد پوشش بازه‌های اطمینان ۹۵٪ در رویکردهای مختلف

## بحث و نتیجه‌گیری

مسئله رگرسیون داده‌های بد تراز فضایی یکی از موضوعات مورد توجه محققان و تحلیل‌گران در بسیاری از زمینه‌های علوم است. در این مقاله رویکردهای مختلفی برای رگرسیون داده‌های بد تراز فضایی شامل رویکرد باجایگذاری، شبیه‌سازی، رگرسیون کالبدنی و ماکسیمم درستمایی ارائه شدند. در دو رویکرد باجایگذاری و شبیه‌سازی وجود خطای اندازه‌گیری منجر به ایجاد اریبی در برآورد پارامترهای مدل رگرسیونی شدند. لذا رویکرد رگرسیون کالبدنی برای تعدیل این اریبی پیشنهاد شد. نتایج مطالعه 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.

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  $\hat{\beta}_1 \pm 1/96sd(\hat{\beta}_1)$  محاسبه شد و در نهایت ۵۰۰ بازه اطمینان به دست آمد. سپس بررسی شد که چند درصد از این بازه‌ها مقدار واقعی  $\beta_1$  یعنی ۱ را در برمی‌گیرند. شکل ۸ نشان می‌دهد هنگامی که فرایند ناهموار می‌شود، درصد پوشش بازه‌های اطمینان ۹۵٪ در رویکردهای *ML*، *KR* 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  $\beta_1$  فراهم کرده‌اند و از این نظر نسبت به سایر روش‌های برآورد برتری دارد.

## 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.