مجله علوم آماری، پاییز و زمستان ۱۳۹۲
جلد ۷، شماره ۲، ص ۲۴۹-۲۶۸

برآورد بازگشتی پارامترهای مدل‌های چندسطحی در حضور خطاهای اندازه‌گیری

هاشم محمودنژاد، موسی گل‌علی‌زاده
گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۹۱/۹/۱۶ تاریخ آخرین بازنگری: ۱۳۹۲/۱۲/۱۸

چکیده : اگرچه خطای اندازه‌گیری در اکثر آزمایشات علمی وجود دارد اما معمولا برای ساده‌سازی مدل‌بندی وجود آن در مطالعات آماری نادیده گرفته می‌شود. در این مقاله، روش‌های مختلف برآورد پارامترهای مدل‌های چندسطحی در حضور خطای اندازه‌گیری مورد مطالعه قرار می‌گیرد. علاوه بر این، روشی جدید برای برآورد پارامترها در این حالت پیشنهاد می‌شود که در مقایسه با روش‌های مرسوم قبلی از دقت بالا و سرعت همگرایی قابل قبول‌تری برخوردار است. همچنین، به‌کمک مطالعه شبیه‌سازی و تحلیل داده‌های مربوط به هزینه و درآمد تعدادی از خانوارهای شهر تهران در سال ۱۳۸۷ عمل‌کرد روش پیشنهادی با روش‌های مرسوم مورد ارزیابی و مقایسه قرار می‌گیرد.

واژه‌های کلیدی : مدل‌های چند سطحی، خطای اندازه‌گیری، الگوریتم بازگشتی، برآورد برون‌یابی، داده‌های هزینه-درآمد.

° آدرس الکترونیک مسئول مقاله: موسی گل‌علی‌زاده، golalizadeh@modares.ac.ir
° کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲J۹۹

۱ مقدمه

معمولا مشاهدات آزمایشگاهی حاصل از هر آزمایش نمی‌تواند مقادیری دقیق باشند و همواره آلوده به خطا هستند. این موضوع به‌خصوص در زمینه‌هایی مانند علوم آموزشی، زیستی و کشاورزی کاملا مشهود است. به عنوان مثال در تعیین سطح نمرات یک کلاس یا تعیین ضریب هوشی افراد، بسته به موقعیت زمانی، نوع سوالات و وضعیت روحی افراد چنین خطایی وجود دارد. از نقطه نظر آماری چنین خطایی باعث بروز مشکلاتی در تحلیل داده‌ها شده و می‌تواند اعتبار نتایج را زیر سوال ببرد و در عین حال نادیده گرفتن آن معمولا منجر به نتیجه‌های گمراه کننده‌ای خواهد شد. در بین محققین علوم آماری چنین خطایی در مشاهدات به خطای اندازه‌گیری[۱] معروف است (فولر، ۱۹۸۷).

مفهوم خطای اندازه‌گیری در مدل‌های چند سطحی را اولین بار لانکفورد (۱۹۹۳) مطرح کرد. وی خطای اندازه‌گیری را در مدل دو سطحی ساده (مدل عرض از مبدا تصادفی) به‌کار برد و مشکلات ناشی از آلوده بودن متغیر تبیینی به خطا، در برآورد پارامترهای مدل را مورد کنکاش قرار داد. سپس، وودهوس و همکاران (۱۹۹۶) مدلی را مورد تحلیل قرار دادند که در آن زیر مجموعه‌ای از متغیرهای تبیینی مدل رگرسیونی چند سطحی آلوده به خطای اندازه‌گیری است. آن‌ها که مدل رگرسیونی چند سطحی را بر روی داده‌های آموزشی ایالات متحده به‌کار بردند، نشان دادند در صورت لحاظ کردن خطای اندازه‌گیری در مدل رگرسیونی، متغیر وضعیت اجتماعی- اقتصادی خانواده‌ها بر روی معدل دانش‌آموزان تاثیری نخواهد داشت. روش مرسوم برآورد پارامترها در این مطالعات روش ماکسیمم درستنمایی[۲] (ML) است (گلداستاین، ۲۰۱۰). از آنجاکه این روش برای مدل‌های چندسطحی از مشکل کم برآوردی رنج می‌برد، روش ماکسیمم درستنمایی مقید[۳] (REML) پیشنهاد شده است (پینهریو و بیتس، ۲۰۰۰).

---

۱ Measurement error

۲ Maximum Likelihood

۳ Restricted Maximum Likelihood

یکی از روش‌های مناسب برای تصحیح اریبی برآوردگر روش بوت‌استرپ است (افرون، ۱۹۸۲). با این دیدگاه فعالیت‌هایی موثری در استفاده از ترکیب این دو روش در مدل‌های چندسطحی در حضور خطای اندازه‌گیری صورت گرفته است. به‌عنوان مثال هاتچیسون و همکاران (۲۰۰۳) برای برآورد پارامترهای مدل رگرسیونی چند سطحی عرض از مبدا تصادفی وقتی‌که متغیرهای تبیینی مدل آلوده به خطای اندازه‌گیری است، از این ایده استفاده کردند. فرائو و گلداستاین (۲۰۰۹) نیز تاثیر وجود خطای اندازه‌گیری را در یک مدل رگرسیونی چند سطحی مربوط به داده‌های آموزش و پرورش کشور پرتغال با دو متغیر تبیینی که یکی از آن‌ها آلوده به خطای اندازه‌گیری است مورد مطالعه قرار دادند. اخیرا، باتاوز و همکاران (۲۰۱۱) با در نظر گرفتن خطای اندازه‌گیری در متغیر تبیینی در مدل چند سطحی با شیب تصادفی روش‌هایی برای بهبود برآورد پارامترهای مدل پیشنهاد دادند. در این مقاله با بررسی موضوع خطای اندازه‌گیری در مدل‌های‌چندسطحی یک راهکار جدیدی معرفی می‌شود که نسبت به روش‌های موجود دارای مزیت‌های برجسته‌ای است. ایده‌های پیشنهادی در مطالعه شبیه‌سازی و هم‌چنین مثال واقعی بکار گرفته می‌شود.

ادامه مقاله حاضر به‌صورت زیر تدوین شده است. در بخش ۲ روش بسیار کارای شبیه‌سازی برون‌یابی [۴] (SIMEX) و در بخش ۳ روش بوت‌استرپ در برآورد پارامترهای مدل‌های چندسطحی در حضور خطای اندازه‌گیری توصیف می‌شود. سپس در بخش ۴ روش پیشنهادی این مقاله ارائه می‌گردد. مطالعه شبیه‌سازی و تحلیل داده‌های واقعی به ترتیب بخش‌های ۵ و ۶ مقاله را تشکیل می‌دهند.

۲ روش شبیه‌سازی برون‌یابی

در این بخش نحوه برآورد پارامترهای مدل‌های چند سطحی در حضور خطای اندازه‌گیری به روش شبیه‌سازی برون‌یابی (کوک و استفانسکی، ۱۹۹۳) و با روش REML تشریح می‌شود.

[۴] Simulation Extrapolation

خطای اندازه‌گیری به‌صورت اختلاف مقادیر مشاهده شده و مقادیر واقعی تعریف می‌شود. به زبانی دقیق‌تر، اگر X_i، x_i و m_iها به‌ترتیب مقادیر مشاهدات، مقادیر واقعی (متغیرهای پنهان[۵]) و خطای اندازه‌گیری باشند آنگاه:

$$X_i = x_i + m_i. \qquad (۱)$$

معمولا خطای اندازه‌گیری دارای توزیع نرمال با میانگین صفر و واریانس $\sigma_m^۲$ فرض می‌شود. به‌علاوه، در مدل‌هایی که با مبحث خطای اندازه‌گیری سر و کار دارند، این خطا مستقل از بقیه خطاهای مرسوم در ادبیات آماری از جمله خطای برآورد، خطای انتخاب، خطای مشاهده و خطای غیر مشاهده‌ای و همچنین مستقل از متغیرهای پنهان در نظر گرفته می‌شود (فولر، ۱۹۸۷). در ادامه مدلی که بر اساس مقادیر واقعی (x_iها) است، مدل واقعی[۶] و مدلی که بر اساس مقادیر مشاهده شده (X_iها) است، مدل ناپخته[۷] نامیده می‌شود طوری‌که θ_{TRUE} و θ_{NAIVE} به ترتیب پارمترهای مدل واقعی و مدل ناپخته هستند. همچنین، عبارت

$$K_x = \frac{\sigma_x^۲}{\sigma_X^۲} = \frac{\sigma_x^۲}{\sigma_x^۲ + \sigma_m^۲} \qquad (۲)$$

نرخ قابلیت اعتماد[۸] متغیر تبیینی در نظر گرفته می‌شود.

فرض کنید یک مدل چند سطحی با عرض از مبدا تصادفی به‌صورت

$$y_{ij} = \alpha + \beta x_{ij} + u_{۰j} + \epsilon_{ij}, \quad i = ۱, \ldots, n_j, \qquad j = ۱, \ldots, J, \qquad (۳)$$

نوشته شود، که در آن $u_{۰j}$ خطای عرض از مبدا تصادفی و متغیر تبیینی x_{ij} آلوده به خطای اندازه‌گیری و y_{ij} متغیر پاسخ است. همچنین ϵ_{ij} خطای iام در سطح j و α و β به ترتیب عرض از مبدا و ضریب رگرسیونی هستند. در صورت نگارش رابطه (۳) به‌صورت ماتریسی بردار مشاهدات با y، ماتریس طرح با X و ماتریس کواریانس با V نمایش داده می‌شود. به‌منظور از بین بردن تاثیر خطای اندازه‌گیری با

۵ Latent variables

۶ True

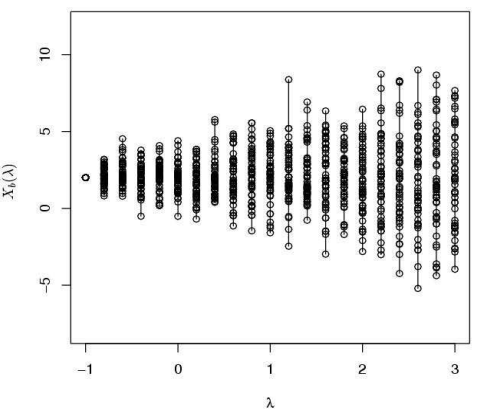
۷ Naive

۸ Reliability rate

کمک روش شبیه‌سازی برون‌یابی ، خطانمای[۹] $\sqrt{\lambda}\sigma_m z_{b,ij}$ به‌ازای $b = ۱, \ldots, B$ به‌صورت تصادفی تولید می‌شود، که در آن $z_{b,ij}$ دارای توزیع نرمال استاندارد است. با اضافه نمودن این خطانما به مقادیر مشاهده شده X_{ij}، داده‌نماهای[۱۰] $X_{b,ij}(\lambda)$، به‌صورت $X_{b,ij}(\lambda) = X_{ij} + \sqrt{\lambda}\sigma_m z_{b,ij}$ به‌دست می‌آیند. آنگاه، با جایگذاری $X_{ij} = x_{ij} + m_{ij}$ در رابطه $X_{b,ij}(\lambda) = X_{ij} + \sqrt{\lambda}\sigma_m z_{b,ij}$، به‌دست می‌آید:

$$X_{b,ij}(\lambda) \Big| x_{ij} \sim N(x_{ij}, (۱ + \lambda)\sigma_m^۲). \qquad (۴)$$

همان‌طور که از رابطه (۴) مشخص است، وقتی λ به‌سمت -۱ میل کند داده‌نماهای تولید شده ($X_{b,ij}$ها) رفتاری شبیه داده‌هایی خواهند داشت که از توزیع نرمال با میانگین x_{ij} و واریانس صفر تولید می‌شوند. به عبارتی دیگر مقادیر تولید شده با از بین بردن اثر خطای اندازه‌گیری در مقادیر مشاهده شده X_{ij}ها، رفتاری مانند مقادیر واقعی (x_{ij}ها) خواهند داشت. همان‌طور که در شکل ۱ ملاحظه می‌شود X_bهای تولید شده در مقابل λ شکلی مانند قیف دارد و هرچقدر λ به -۱ نزدیک می‌شود، پراکندگی X_b نیز کاهش می‌یابد. در نهایت به‌ازای $\lambda = -۱$ تاثیر خطای اندازه‌گیری در مقادیر مشاهده شده از بین رفته و آن‌ها را به مقدار واقعی (x_{ij}) می‌رساند.

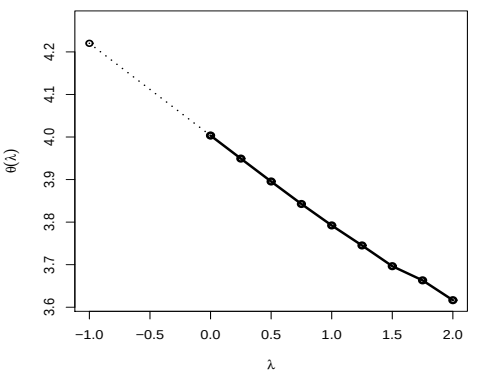


شکل ۱: تغییرات $X_b(\lambda)$ در مقابل λ

[۹] Pseudo Error

[۱۰] Pseudo Data

چون $X_{b,ij}(\lambda)$ در $\lambda = -1$ تاثیر خطای اندازه‌گیری m_{ij} را از بین می‌برد لذا طبیعی است به‌جای برآورد پارامترها با مقادیر مشاهده شده X_{ij}، که همراه با خطا است و در نتیجه باعث برآوردهای غیر معقول می‌شود، پارامترهای مدل در نقاط مختلف $\lambda \geq 0$ با استفاده از $X_{b,ij}(\lambda)$ها $(b = 1, \ldots, B)$ برآورد و پارامترهای برآورد شده در $\lambda = -1$ برون‌یابی شوند. به‌طور دقیق‌تر، فرض کنید $X_{1,ij}(\lambda), X_{2,ij}(\lambda), \ldots, X_{B,ij}(\lambda)$ داده‌نماهای تولید شده به‌ازای $b = 1, \ldots, B$ باشند. به کمک این داده‌نماهای تولید شده برآوردهای $\hat{\theta}_1(\lambda), \hat{\theta}_2(\lambda), \ldots, \hat{\theta}_B(\lambda)$ با استفاده از روش‌های مرسوم مانند REML به‌دست آورده و قرار داده می‌شود $\hat{\theta}(\lambda) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b(\lambda)$. سپس، برآورد $\hat{\theta}(\lambda)$ را به‌ازای نقاط مختلف $\lambda \geq 0$ به‌دست آورده و نمودار $\hat{\theta}(\lambda)$ها در مقابل λها رسم می‌شود (شکل ۲). با امتداد نمودار به‌دست آمده به‌ازای $\lambda = -1$ برآورد شبیه‌سازی برون‌یابی حاصل می‌شود.



شکل ۲: برون‌یابی $\theta(\lambda)$ در نقطه $\lambda = -1$

۳ تصحیح اریبی پارامترها با استفاده از بوت‌استرپ

در این بخش به نحوه تصحیح اریبی ناشی از وجود خطای اندازه‌گیری در برآورد پارامترها با روش بوت‌استرپ در یک مدل رگرسیونی چند سطحی با عرض از مبدا تصادفی پرداخته می‌شود. مدل رگرسیونی چند سطحی با عرض از مبدا تصادفی (۳)

را در نظر بگیرید که در آن متغیر تبیینی x_{ij} آلوده به خطای اندازه‌گیری است. بنابراین مقادیر مشاهده شده به‌صورت $X_{ij} = x_{ij} + m_{ij}$ خواهد بود، که در آن m_{ij} خطای اندازه‌گیری با میانگین صفر و واریانس σ_m^2 است. مدل رگرسیونی براساس مقادیر مشاهده شده که به مدل ناپخته معروف است را می‌توان به‌صورت

$$y_{ij} = \alpha_{NAIVE} + \beta_{NAIVE} X_{ij} + u_{0j} + e_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J$$

نوشت، که در آن پارامترهای α_{NAIVE} و β_{NAIVE} با روش REML برآورد می‌شوند.

به‌منظور کاهش اریبی برآوردهای به‌دست آمده حاصل از مقادیر مشاهده شده، مراحل تصحیح اریبی برآوردها به روش بوت‌سترپ به‌صورت الگوریتم زیر دنبال می‌شود.

الگوریتم ۱:

گام ۱: بردار پارامتر $\theta_{NAIVE} = (\alpha_{NAIVE}, \beta_{NAIVE})$ برآورد شود.

گام ۲: مقادیر مشاهده شده (X_{ij}) برآورد شده و با $\hat{x}_{ij}$ نشان داده شوند.

گام ۳: $\hat{y}_{ij}$ها با استفاده از $\hat{\theta}_{NAIVE}$ و $\hat{x}_{ij}$ها تولید شوند.

گام ۴: به $\hat{x}_{ij}$ها خطای اندازه‌گیری اضافه کرده و با $\hat{X}_{ij}$ نشان داده شوند.

گام ۵: با استفاده از $\hat{X}_{ij}$های و $\hat{y}_{ij}$ها تولید شده در گام‌های ۳ و θ_t برآورد شود.

با تکرار گام‌های ۳ تا ۵ به‌ازای $t = 1, \ldots, T$، اریبی پارامترهای مدل به‌صورت $bias = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}_t - \hat{\theta}_{NAIVE}$ به‌دست می‌آید. در این صورت پارامتر مورد نظر برابر $\hat{\theta}_b = \hat{\theta}_{NAIVE} - bias$ خواهد بود. برآورد $\hat{\theta}_b$ به‌ازای $b = 1, \ldots$ تا زمانی ادامه می‌یابد که میانگین آنها به همگرایی برسد.

۴ برآورد بازگشتی پارامترهای مدل‌های چند سطحی

روش‌های مختلفی برای برآورد و کاهش اریبی پارامترهای مدل‌های چندسطحی در حضور خطاهای اندازه‌گیری ارائه شده، که هریک از این روش‌ها از معایب و مزایای خاصی برخوردار هستند. الگوریتم‌های موجود در برآورد پارامترهای مدل‌های چند سطحی در حضور خطای اندازه‌گیری نیز از این اوامر مستثنا نیستند. لذا در این بخش روش جدیدی برای برآورد پارامترهای مدل‌های مورد مطالعه پیشنهاد می‌شود که از دو مزیت سرعت بالای همگرایی و میزان اریبی کمتر نسبت به روش‌های شبیه‌سازی برون‌یابی و بوت‌استرپ برخوردار است.

مدل دو سطحی با عرض از مبدا تصادفی (۳) را در نظر بگیرید، که در آن متغیر تبیینی آلوده به خطای اندازه‌گیری است. با ترکیب معادله (۳) و رابطه (۱) می‌توان نوشت:

$$\begin{pmatrix} y_{ij}-\alpha \\ X_{ij} \end{pmatrix} = \begin{pmatrix} \beta \\ ۱ \end{pmatrix} x_{ij} + \begin{pmatrix} u_{۰j}+\epsilon_{ij} \\ m_{ij} \end{pmatrix},\ i=۱,..n_j\ ,\quad j=۱,\ldots,J. \qquad (۵)$$

در این صورت برآورد کمترین توان‌های دوم تعمیم یافته متغیر پنهان x_{ij}ها با استفاده از برآوردهای مدل ناپخته به‌صورت

$$\hat{x}_{ij} = \left((\hat{\beta}_{NAIVE},۱)\Sigma^{-۱}(\hat{\beta}_{NAIVE},۱)^T\right)^{-۱}(\hat{\beta}_{NAIVE},۱)\Sigma^{-۱}d_{ij} \qquad (۶)$$

به‌دست می‌آید، که در آن $d_{ij}=(y_{ij}-\hat{\alpha}_{NAIVE},X_{ij})^T$ و Σ ماتریس کواریانس بردار $(u_{۰j}+\epsilon_{ij}\ ,\ m_{ij})^T$ به‌صورت $\Sigma=diag\{\hat{\sigma}_u^۲,\hat{\sigma}_m^۲\}$ است. همچنین $\hat{\alpha}_{NAIVE}$ و $\hat{\beta}_{NAIVE}$ به ترتیب برآورد پارامترهای عرض از مبدا و شیب با روش REML در مدل ناپخته هستند. با استفاده از مدل دوسطحی (۳) و رابطه (۶) می‌توان پارامترهای مدل رگرسیونی و متغیر پنهان را به‌صورت بازگشتی توسط الگوریتم زیر برآورد کرد.

الگوریتم ۲:

گام ۱: قرار داده شود $t=۰$.

گام ۲: بردار پارامتر $\theta^{(t)}=(\alpha_{NAIVE},\beta_{NAIVE})$ برآورد شود.

گام ۳: به کمک برآوردهای حاصل در گام ۲ متغیرهای پنهان $x_{ij}^{(t)}$ها برآورد شوند.

گام ۴: قرار داده شود $t = t + ۱$.

گام ۵: با توجه به $\hat{x}_{ij}^{(t-۱)}$ها، پارامتر $\theta^{(t)}$ برآورد شود.

گام ۶: با توجه به $\theta^{(t)}$های برآورد شده در گام ۵ x_{ij}ها برآورد شوند.

گام ۷: اگر پراکندگی $\hat{x}^{(t)} - X$ بیشتر از $\hat{x}^{(t-۱)} - X$ باشد، به گام ۴ رجوع شود.

با استفاده از الگوریتم ۲ پارامترهای مدل و متغیرهای پنهان به‌طور هم‌زمان برآورد می‌شوند. همان‌گونه که انتظار می‌رود اگر چه در گام ۱ برآورد پارامترهای مدل و متغیرهای پنهان دارای اریبی هستند، اما در طی گام‌های بعد این اریبی کم شده تا زمانی که پراکندگی خطای اندازه‌گیری مینیمم مقدار خود را اختیار می‌کند. از جمله ویژگی‌های بارز الگوریتم ۲ علاوه بر دقت بالا، سرعت همگرایی خوب آن است.

۵ مطالعه شبیه‌سازی

الگوریتم ۲ برای یک مدل دو سطحی شیب تصادفی بکار گرفته شده است. در این شبیه‌سازی اندازه‌های نمونه به‌صورت $i = ۱,\ldots,۱۰$ ، $j = ۱,\ldots,۱۰$ و $n_j = ۱۰$ و همچنین مقادیر ثابت $\beta_۰$ ، $\beta_۱$ و σ_e به‌ترتیب برابر ۴، ۵ و ۲۰ در نظر گرفته شده‌اند. جدول ۱ نشان دهنده برآورد پارامترهای مدل واقعی و مدل ناپخته با روش REML است. همان‌طور که ملاحظه می‌شود برآورد پارامترهای مدل واقعی به مقادیر ثابت $\beta_۰ = ۴$، $\beta_۱ = ۵$ و $\sigma_e = ۲۰$ خیلی نزدیک است. اما چون پارامترهای مدل ناپخته براساس مقادیر مشاهده شده X_{ij} که همراه با خطای اندازه گیری است، برآورد می‌شوند با مقادیر ثابت اختلاف قابل توجهی دارند.

جدول ۱: برآورد پارامترهای مدل واقعی و مدل ناپخته با روش REML

مدل	$\hat{\sigma}_e$	$\hat{\beta}$	$\hat{\alpha}$
واقعی	۲۰/۳۳	۵/۰۷	۴/۶
ناپخته	۴۳/۸۶	۴/۵۶	۲۰/۰

برآورد پارامترهای عرض از مبدا، ضریب ثابت رگرسیونی، انحراف معیار خطای سطح اول و میزان زمان مصرفی سیستم (به ثانیه) در رسیدن به جواب به روش‌های شبیه‌سازی برون‌یابی ، بوت‌استرپ و بازگشتی محاسبه و نتایج حاصل به‌صورت جدول ۲ خلاصه شده‌اند. همان‌طور که ملاحظه می‌شود برآورد پارامترها به روش بازگشتی علاوه بر دقت، از سرعت بالایی در مقایسه با روش‌های شبیه‌سازی برون‌یابی و بوت‌استرپ برخوردار است. در بین برآوردگرها، برآوردگر بازگشتی در مقایسه با دیگر روش‌ها به برآوردهای مدل واقعی نزدیک‌تر بوده و با مقایسه $\hat{\sigma}_e$ها در جدول ۲ ملاحظه می‌شود که روش بازگشتی بهتر از بقیه روش‌ها عمل کرده است. انحراف معیار هر کدام از برآوردگرها در کل تکرار شبیه‌سازی در جدول ۳

جدول ۲: مقایسه سرعت و دقت روش‌های برآورد در یک مدل دو سطحی شیب تصادفی

روش	$\hat{\sigma}_e$	$\hat{\beta}$	$\hat{\alpha}$	زمان(ثانیه)
شبیه‌سازی برون‌یابی	۳۸/۱۶	۵/۵۰	۱۶/۷	۹۳۷/۱۲
بوت‌استرپ	۲۲/۳۶	۴/۹۵	۱۸/۷	۵۶۱/۶۳
بازگشتی	۲۱/۴۶	۵/۰۴	۱۷/۹	۶/۸۶۰۰

خلاصه شده‌اند. واضح است که مقادیر کوچک انحراف معیار برآوردگرها نشان از دقت بالای هر کدام از برآوردگرها است. اگر چه به خاطر مقادیر بسیار کوچک انحراف معیار نمی‌توان تمییزی بین روش‌ها قائل شد، اما می‌توان ادعا نمود بطور کلی روش بازگشتی بهتر از بقیه موارد (بجز برای یک حالت از روش بوت استرپ) عمل کرده است. حال که ارجحیت روش برآورد بازگشتی مشخص شد بهتر است

جدول ۳: انحراف معیار برآوردگرها در مدل دو سطحی شیب تصادفی

روش	$\hat{\sigma}_e$	$\hat{\beta}$	$\hat{\alpha}$
شبیه‌سازی برون‌یابی	۰/۰۰۲	۰/۰۰۲۲	۰/۰۳۹
بوت‌استرپ	۰/۰۰۱	۰/۰۰۱۲	۰/۰۱۹
بازگشتی	۰/۰۰۱	۰/۰۰۱۰	۰/۰۲۶

عمل‌کرد آن در برآورد متغیرهای پنهان مدل نیز مطالعه شود. جدول ۴ نشان دهنده برآورد متغیرهای پنهان در یک مدل چند سطحی شیب تصادفی است. ستون‌های

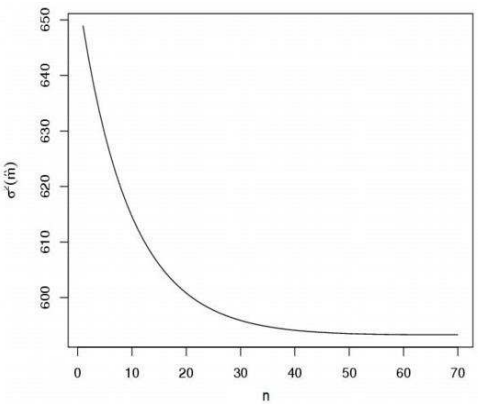
دوم تا چهارم این جدول به‌ترتیب از راست به چپ نشان دهنده برآورد متغیرهای پنهان، مقادیر واقعی متغیرهای پنهان و مقادیر مشاهدات است. با مقایسه این سه ستون می‌توان ملاحظه نمود که برآورد متغیرهای پنهان به روش بازگشتی از دقت بالایی برخوردار است. به‌خصوص ردیف‌های سوم، هفتم، هشتم و یازدهم این جدول دقت برآورد متغیرهای پنهان را به‌خوبی نشان می‌دهد. در این ردیف‌ها مقادیر مشاهده شده(X_{ij}) با مقادیر واقعی(x_{ij}) اختلاف بسیار زیادی دارند، (به علامت منفی آن‌ها دقت شود) ولی با استفاده از روش بازگشتی اختلاف ناشی از خطای اندازه‌گیری برطرف شده است و الگوریتم موفق شد متغیر پنهان مربوطه را به خوبی برآورد کند.

جدول ۴: برآورد بازگشتی تعدادی از متغیرهای پنهان در مدل چند سطحی شیب تصادفی با حضور خطای اندازه‌گیری

ردیف	$\hat{x}_{ij}$	x_{ij}	X_{ij}
۱	۱۰۷/۲	۱۰۹/۷۴	۱۰۵/۳۱
۲	۵۲/۵۳	۵۵/۰۲	۲۵/۴۷
۳	−۱۲/۵۵	−۱۴/۹۸	۲۱/۹۶
۴	−۱۱۲/۵۴	−۱۱۰/۱۰	−۱۲۱/۴۳
۵	−۵۹/۳۴	−۵۵/۴۸	−۶۴/۱۳
۶	۱۴۰/۳۸	۱۴۱/۸۶	۱۵۰/۴۵
۷	۶/۵۹	۸/۰۹	−۵/۸۴
۸	−۳/۰۲	−۲/۱۰	۱۴/۸۲
۹	۸۴/۹۶	۸۴/۷۹	۸۸/۱۷
۱۰	۷۲/۲۹	۷۵/۱۸	۳۷/۸۲
۱۱	−۲۰/۳۴	−۲۰/۰۲	۲۴/۸۵
۱۲	−۳۳/۵۶	−۳۲/۱۹	−۴۱/۳۵
۱۳	−۵۷/۴۸	−۵۵/۴۸	−۷۵/۰۸
۱۴	۵۱/۸۷	۵۳/۰۵	۶۱/۰۳

شکل ۳ نشان دهنده نمودار واریانس برآوردهای خطای اندازه‌گیری به روش بازگشتی در طی مراحل برآورد است. همان‌طور که ملاحظه می‌شود واریانس خطای اندازه‌گیری در برآورد پارامترهای مدل کاهش پیدا کرده تا اینکه در نهایت به

مینیمم مقدار خود رسیده است. قابل ذکر است که برآوردهای مدل با حضور خطای اندازه‌گیری با کمترین اریبی در این حالت اتفاق می‌افتد. ملاحظه می‌شود که روند کاهشی در ابتدای امر شدید ولی در ادامه به حالت تقریبا ثابتی رسیده است.

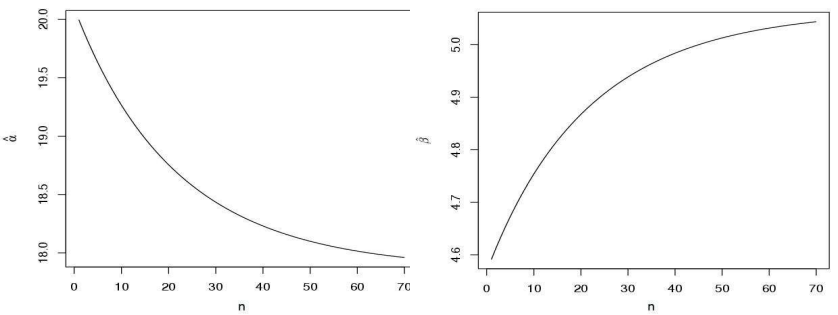


شکل ۳: واریانس خطای اندازه‌گیری در طی مراحل روش برآورد بازگشتی

شکل ۴ نشان دهنده نمودارهای برآورد پارامترهای عرض از مبدا و ضریب رگرسیونی مدل دو سطحی شیب تصادفی در مقابل مراحل تکرار الگوریتم است. همان‌گونه که ملاحظه می‌شود برآورد β در طی مراحل برآورد به مقدار واقعی نزدیک شده است (جدول ۱). به‌طور دقیق‌تر نمودار ۴.الف نشان دهنده برآورد پارامتر β به روش بازگشتی با شروع از ۴/۵۶ و با کاهش اثر خطای اندازه‌گیری در برآورد پارامترها به سمت مقدار واقعی ۵/۰۷ میل کرده است. همین روند در نمودار ۴.ب نیز در برآورد پارامتر عرض از مبدا تصادفی اتفاق افتاده است. این نمودار نیز نشان می‌دهد که برآورد پارامتر عرض از مبدا به روش برآورد بازگشتی از برآورد ناپخته شروع شده و با کاهش اریبی به سمت برآورد واقعی میل کرده است.

۶ مثال کاربردی

داده‌های مورد استفاده در این بخش اطلاعات هزینه –درآمد ۱۹۶۰۰ خانوار شهر تهران است که در سال ۱۳۸۷ توسط مرکز آمار ایران جمع‌آوری شده است. به دلیل



(ب) (الف)

شکل ۴: نمودار تغییرات برآورد پارامتر الف: ثابت ضریب رگرسیونی و ب: عرض از مبدا در یک مدل چند سطحی با شیب تصادفی در مقابل مراحل تکرار

عدم گزارش دقیق درآمدها توسط سرپرست خانوار، مقادیر گزارش شده غیر واقعی و آلوده به خطای اندازه‌گیری هستند. چون خانوارها در مناطق ۲۲ گانه خوشه‌بندی شده‌اند می‌توان از یک مدل‌دوسطحی برای تحلیل این داده‌ها استفاده نمود.

واضح است که افراد و خانوارها سعی می‌کنند مخارج خود را بر اساس میزان درآمدشان تنظیم کنند. به‌عبارتی دیگر هزینه یک فرد یا یک خانوار تابعی از درآمد آن‌ها است. به‌همین دلیل هزینه خانوار به‌عنوان متغیر پاسخ و درآمد خانوار متغیر مستقل در نظر شده است. تحلیل اکتشافی بیانگر آن است که داده‌های هزینه–درآمد دارای چولگی به راست هستند. با استفاده از تبدیل باکس –کاکس و انجام آزمون کلموگرف–اسمیرنف، p–مقدار آزمون بیشتر از ۰/۰۵ به‌دست آمده که تاییدی بر نرمال بودن داده‌های تبدیل یافته هزینه–درآمد است. نکته قابل اشاره این است که برای جلوگیری از تکرار کلمات از این‌جا به بعد منظور از داده‌های هزینه –درآمد، مقادیر تبدیل یافته آن‌ها است.

با فرض غیر همبسته بودن درون گروهی داده‌ها، ابتدا رابطه رگرسیونی ساده بین دو متغیر برازش داده شد. رابطه رگرسیونی خطی ساده بین دو متغیر درآمد و هزینه خانوار به‌صورت

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = ۱, \ldots, ۱۹۶۰۰ \qquad (۷)$$

است، که در آن y_i ،x_i و $\epsilon_i \sim N(۰, \sigma_\epsilon^۲)$ به‌ترتیب متغیر درآمد، متغیر هزینه و خطای برازش مدل است. با توجه به P–مقدار و آماره F در جدول ۵ رابطه رگرسیونی خطی ساده بین متغیر درآمد و هزینه هر خانوار معنی‌دار است.

جدول ۵: جدول تحلیل واریانس مدل رگرسیونی (۷) برای داده‌های هزینه–درآمد

	df	SS	MS	آماره F	p–مقدار
هزینه	۱	۵۲۴۵۹۸	۵۲۴۵۹۸	۱۸۹۴۰	۰/۰۰۰۰۱
باقیمانده	۱۹۵۹۸	۵۴۲۸۲۸	۲۸		

برآورد پارامترهای مدل (۷) در جدول ۶ آمده است. بنا به این جدول P–مقدار برآورد پارامترهای عرض از مبدا و شیب مدل (۷) در سطح ۰/۰۵ معنی‌دار هستند.

جدول ۶: نتایج برازش داده‌های درامد مدل رگرسیونی (۹)

پارامتر	برآورد	خطای استاندارد	p–مقدار
α	۳۰/۳۱۲	۰/۲۷۰	$< ۲/۲e-۱۶$
β	۰/۳۰۲	۰/۰۰۲	$< ۲/۲e-۱۶$

ضریب تعیین مدل رگرسیونی (۷) برابر ۰/۴۳ به‌دست آمد که نشان دهنده ناکارآمدی مدل رگرسیونی خطی ساده برای داده‌های هزینه–درآمد شهر تهران است. علاوه بر این، مقدار آماره دوربین–واتسون برابر ۰/۰۱۷ به‌دست آمد که نشان دهنده همبستگی درون‌گروهی بین مشاهدات پاسخ است. این کمیت هشدار می‌دهد که از مدلی استفاده شود که قادر به لحاظ نمودن همبستگی درون‌گروهی داده‌هاست. یکی از مدل‌های کاندید مدل چند سطحی است. با سلسله مراتبی در نظر گرفتن داده‌های هزینه–درآمد، می‌توان مدل دوسطحی برای این داده برازش داد، که خانوارها (سطح اول) در مناطق (سطح دوم) مختلف آشیانه کرده‌اند. برای داده‌های هزینه درآمد، چهار مدل تصادفی دو سطحی مؤلفه واریانس، مدل عرض از مبدا تصادفی، مدل شیب تصادفی و مدل عرض از مبدا و شیب تصادفی برازش داده شد. با توجه به جدول ۷ مدل عرض از مبدا و شیب تصادفی دارای کمترین مقدار

جدول ۷: نتایج برازش مدل‌های مختلف به داده‌های هزینه-درآمد

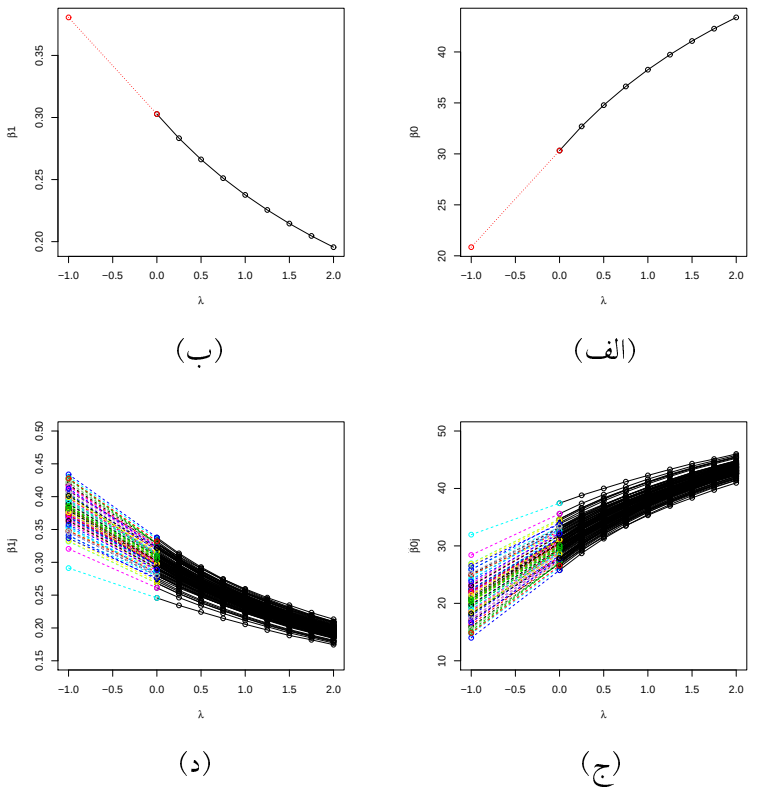
مدل	پارامتر	برآورد	خطای استاندارد	AIC	BIC	$logLik$
مؤلفه واریانس	μ	۶۷/۲۰۵	۰/۲۹۸	۱۳۳۹۷۹	۱۳۴۰۰۳	−۶۶۹۸۶
	σ^2_u	۰/۰۲۱				
	σ^2_e	۵۴/۳۷۵				
عرض از مبدا تصادفی	α	۳۰/۳۱۳	۰/۲۷۱	۱۲۰۷۳۷	۱۲۰۷۶۸	−۶۰۳۶۴
	β	۰/۳۰۲	۰/۰۰۲			
	$\sigma^2_{u_0}$	۰/۰۱۳				
	σ^2_e	۲۷/۶۴۶				
شیب تصادفی	α	۳۰/۳۱۲	۰/۲۷۰	۱۲۰۷۳۹	۱۲۰۷۷۰	−۶۰۳۶۵
	β	۰/۳۰۲	۰/۰۰۲			
	$\sigma^2_{u_0}$	۰/۰۰۰۰۶				
	σ^2_e	۲۷/۶۵۷				
عرض از مبد و شیب تصادفی	α	۳۰/۳۲۶	۰/۳۹۹	۱۲۰۷۰۳	۱۲۰۷۵۰	−۶۰۳۴۵
	β	۰/۳۰۲	۰/۰۰۳			
	$\sigma^2_{u_0}$	۸/۶۵				
	$\sigma^2_{u_1}$	۰/۰۰۰۵۴				
	$\sigma_{u_{01}}$	−۰/۹۹۸				
	σ^2_e	۲۷/۴۹۲				

معیارهای $logLik$، BIC و AIC نسبت به سه مدل دیگر است. بنابراین از بین چهار مدل فوق، مدل عرض از مبدا و شیب تصادفی به مدل‌های دیگر ترجیح داده می‌شود. به عبارتی دیگر مدل انتخابی به‌صورت

$$y_{ij} = \alpha + \beta x_{ij} + u_{1j}x_{ij} + u_{0j} + \epsilon_{ij}, \quad i = 1, \ldots, n_i, \quad j = 1, \ldots, 22 \qquad (8)$$

است که در آن i در سطح خانوار(سطح اول) و J در سطح مناطق (سطح دوم) تغییر می‌کند. همچنین، y_{ij}، x_{ij}، (u_{0j}, u_{1j}) و ϵ_{ij} به ترتیب متغیر پاسخ (هزینه)، متغیر تبیینی (درآمد)، بردار خطای سطح دوم و خطای ناشی از برازش مدل است.

در ادامه برآورد پارامترهای مدل (۸) در حضور خطای اندازه‌گیری مورد مطالعه قرار می‌گیرد. از آن‌جایی که در این داده‌ها $\sigma^2_X = 291/78$ و با توجه به رابطه $X_{ij} = x_{ij} + m_{ij}$، واضح است که $\sigma^2_X > \sigma^2_m$. بنابراین با فرض $\sigma^2_m = 80$ به برآورد پارامترهای مدل برازش داده شده با در نظر گرفتن خطای اندازه‌گیری با روش‌های مختلف شامل روش پیشنهادی پرداخته می‌شود.

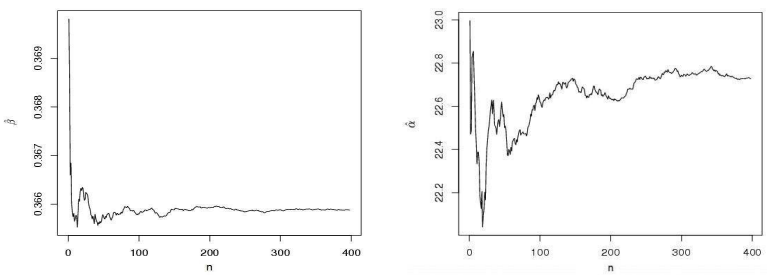


شکل ۵: نمودار برآورد پارامترهای عرض از مبدا و شیب ثابت و تصادفی در حضور داده‌های خطا به روش شبیه‌سازی برون‌یابی

شکل‌های ۵.الف و ۵.ب به‌ترتیب نشان دهنده نمودارهای برآورد عرض از مبدا و ضریب رگرسیونی به روش شبیه‌سازی برون‌یابی است. با استفاده از این روش پارامترهای α و β به ازای $\lambda = ۰, ۰/۲۵, ۰/۵, ۰/۷۵, ۱, ۱/۲۵, ۱/۵, ۱/۷۵, ۲$ برآورد شده و با استفاده از آن‌ها پارامترهای مورد نظر به‌ازای $\lambda = -۱$ برون‌یابی شده‌اند. همان‌طور که ملاحظه می‌شود نمودارهای شبیه‌سازی برون‌یابی از نقاط $(۰, \hat{\alpha}_{NAIVE})$ و $(۰, \hat{\beta}_{NAIVE})$ عبور کرده و به‌ازای $\lambda = -۱$ به برآورد شبیه‌سازی برون‌یابی رسیده است. شکل‌های ۵.ج و ۵.د نیز نشان دهنده برآورد پارامترهای تصادفی عرض از مبدا و ضریب رگرسیونی مدل مورد نظر با استفاده از

روش شبیه‌سازی برون‌یابی و در حضور خطای اندازه‌گیری است.

شکل ۶.الف و ۶.ب به‌ترتیب نشان دهنده برآورد پارامترهای عرض از مبدا و شیب ثابت مدل (۸) به روش بوت‌استرپ است. همان‌طور که ملاحظه می‌شود روش بوت‌استرپ بعد از ۴۰۰ مرحله تکرار در برآورد عرض از مبدا و شیب ثابت پارامتر مدل مورد نظر به همگرایی رسیده است.

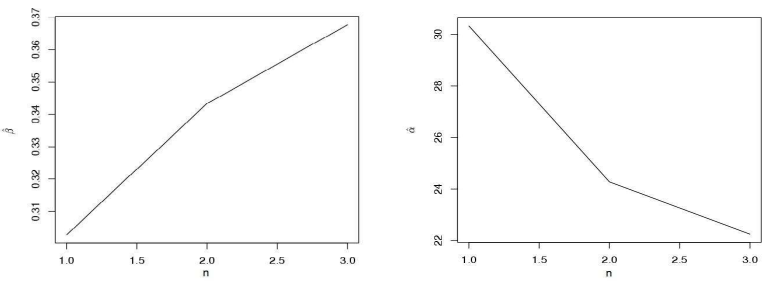


(ب) (الف)

شکل ۶: نمودار برآورد پارامترهای عرض از مبدا و شیب ثابت مدل به روش بوت‌استرپ

شکل ۷ نشان دهنده نمودارهای برآورد پارامترهای عرض از مبدا ثابت و ضریب رگرسیونی ثابت مدل (۸) به روش بازگشتی است. همان طور که ملاحظه می‌شود روش بازگشتی تنها در سه مرحله موفق شده بعد از کاهش تاثیر خطای اندازه‌گیری به پارامترهای مورد نظر دست پیدا کند. قابل ذکر است که تعداد مراحل و سرعت روش بازگشتی مانند روش‌های بوت‌استرپ و شبیه‌سازی برون‌یابی به واریانس خطای اندازه‌گیری متغیرها و به نوع مدل مورد مطالعه بستگی دارد.

در این مقاله به‌همراه روش‌های شبیه‌سازی برون‌یابی و بوت‌استرپ، روش بازگشتی را به‌منظور بهبود پارامترهای مدل‌های چند سطحی در حضور خطای اندازه‌گیری شد که از لحاظ دقت و سرعت همگرایی در مقایسه با روش‌های شبیه‌سازی برون‌یابی و بوت‌استرپ عمل‌کرد بهتری دارد. برآورد پارامترهای مدل (۸) به‌وسیله روش بوت‌استرپ و بازگشتی به‌منظور مقایسه بهتر همراه برآوردهای شبیه‌سازی برون‌یابی در جدول ۸ ارائه شده‌اند. همان‌طور که ملاحظه می‌شود



(ب) (الف)

شکل ۷: نمودار الف: برآورد پارامتر عرض از مبدا و ب: برآورد پارامتر ضریب رگرسیونی به روش بازگشتی

برآوردهای حاصل از دو روش بوت‌استرپ و بازگشتی تقریبا نزدیک به هم هستند. اما با توجه به برآورد واریانس خطای سطح اول، روش بازگشتی عمل‌کرد نسبتا بهتری در دست‌یابی به برآورد پارامترهای مدل ناپخته داشته است. نکته قابل توجه در روش بازگشتی نسبت به روش‌های شبیه‌سازی برون‌یابی و بوت‌استرپ سرعت همگرایی این روش است. روش بازگشتی در ۱۶ ثانیه ولی روش‌های شبیه‌سازی برون‌یابی و بوت‌استرپ به‌ترتیب در ۵۹۷۸۴ و ۲۱۵۹۶ ثانیه به برآورد پارامترهای مدل رسیده است. باید توجه شود که زمان عمل‌کرد روش‌های شبیه‌سازی برون‌یابی بوت‌استرپ و بازگشتی بسته به نوع مدل متفاوت است.

بحث و نتیجه‌گیری

گرچه خطای اندازه‌گیری نقش بسزایی در مدل‌بندی آماری ایفا می‌کند اما در بسیاری از تحقیقات این خطا نادیده گرفته می‌شود. در اکثر موارد نادیده گرفتن این خطا نتایج گمراه کننده‌ای در اختیار محقق قرار می‌دهد. در این مقاله با برجسته نمودن خطای اندازه‌گیری در مدل‌بندی آماری نقش آن را در بهبود مدل‌های مختلف مورد بحث قرار داده شد. روش‌های مختلفی برای بهبود مدل‌هایی که متغیرهای آن آلوده به خطای اندازه‌گیری است، مورد مطالعه قرار گرفت و روش بازگشتی برای بهبود مدل پیشنهاد شد. این روش از لحاظ سرعت و دقت در بهبود مدل با

جدول ۸: برآورد پارامترهای مدل به روش‌های بوت‌استرپ، شبیه‌سازی برون‌یابی و بازگشتی

روش	پارامتر	برآورد	زمان(ثانیه)
بوت‌استرپ	α	۲۲/۷	۲۱۵۹۶
	β	۰/۳۶	
	$\sigma^2_{u_0}$	۴/۰۱	
	$\sigma^2_{u_1}$	۰/۰۰۰۲	
	$\sigma_{u_{01}}$	$-$۰/۰۶	
	σ^2_e	۱۵/۸	
شبیه‌سازی برون‌یابی	α	۲۰/۸	۵۹۷۸۴
	β	۰/۳۸	
	$\sigma^2_{u_0}$	۹/۳۱	
	$\sigma^2_{u_1}$	۰/۰۰۰۶	
	$\sigma_{u_{01}}$	$-$۰/۰۸	
	σ^2_e	۱۶/۵	
بازگشتی	α	۲۲/۰	۱۶
	β	۰/۳۶	
	$\sigma^2_{u_0}$	۱/۶۶	
	$\sigma^2_{u_1}$	۰/۰۰۰۱	
	$\sigma_{u_{01}}$	$-$۰/۰۳	
	σ^2_e	۱۵/۳	

روش‌های بوت‌استرپ و شبیه‌سازی برون‌یابی مورد مقایسه قرار گرفت و مشخص شد که آن برتری قابل توجهی در سرعت و دقت بهبود مدل نسبت به روش‌های مورد اشاره دارد.

مراجع

Battauz, M., Bellio, R. and Gori, E. (2011), Covariate Measurement Error Adjustment for Multilevel Models with Application to Educational

Data, *Journal of Educational and Behavioral Statistics*, **36**, 283-306.

Cook, J. R. and Stefanski, L. A. (1993), Simulation-Extrapolation in Measurement Error Models, *Journal of American Statistical Association*, **89**, 1314-1328.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, SIAM, Philadelohia.

Ferrao, M. E. and Goldstein, H. (2009), Adjusting for Measurement Error in the Value Added Model: Evidence from Portugal, *Quality and Quantity*, **43**, 951-963.

Fuller, W. A. (1978), *Measurement Error Models*, John Wiley, New York.

Goldstein, H. (2010), *Multilevel Statistical Models*, 4th Ed., Arnold Publishing, London.

Hutchison, D., Morrison, J. and Felgate, R. (2003), Bootstraping the Effects of Measurement Errors, *Multilevel Modeling Newspaper*, **15**, 2-10.

Longford, N. T. (1993), Regression Analysis of Multilevel Data with Measurement Error, *British Journal of Mathematical and Statistical Psychology*, **47**, 301-312.

Pinherio, J. C. and Bates, D. M. (2000), *Mixed-Effects Models in R and S-plus*, Springer-Verlag, New York.

Woodhouse, G., Yang, M., Goldstein, H. and Rasbash, J. (1996), Adjusting for Measurement Error in Multilevel Analysis, *Journal of the Royal Statistical Society*, A, **159**, 201-212.