مجلّه‌ی پژوهش‌های آماری ایران
دوره‌ی ۴، شماره‌ی ۱، بهار و تابستان ۱۳۸۶، صص ۷۱-۹۰
J. Statist. Res. Iran 4 (2007): 71-90

مقایسه‌ی برآورد ناحیه‌ی کوچک متوسط درامد خانوار در برخی استان‌های کشور با روش بیز سلسله‌مراتبی

شاهو زارعی،[†] عباس گرامی،[‡,*] و مجید جعفری خالدی[†]

[†] دانشگاه تربیت مدرس

[‡] دانشگاه تهران

چکیده. در سال‌های اخیر نیاز به تعیین آمارهایی با دقت لازم برای برآورد پارامترهای توصیفی ناحیه‌ی کوچک به‌شدت افزایش یافته است. در ناحیه‌ی کوچک، برآوردگرهای مستقیم به‌علت اندازه‌ی کم نمونه ممکن است از دقت لازم برخوردار نشوند. از این رو با به‌کارگیری مدل‌ها که قابلیت لحاظ کردن اثرهای ناحیه و اطلاعات کمکی را در تحلیل‌ها دارند، می‌توان دقت برآوردگر مستقیم را افزایش داد. در این میان، مدل سطح ناحیه، به‌علت سادگی و دسترسی آسان به اطلاعات کمکی سطح ناحیه و امکان بررسی فرضیات مورد استفاده به‌وسیله‌ی داده‌های نمونه‌ای، از اهمیت بیش‌تری برخوردار شده است. در این مقاله با به‌کارگیری چنین مدلی، روش بیز سلسله‌مراتبی برای برآورد میانگین درامد چند استان کشور در حالت‌های مختلف مورد استفاده و بررسی قرار گرفته و با برآوردهای به دست آمده از رهیافت بهترین پیشگوی نااریب خطی تجربی و بیز تجربی مقایسه شده است. نتایج این تحقیق به‌طور کلّی برتری رهیافت بیز سلسله‌مراتبی بر دو رهیافت مذکور و برآوردگر مستقیم را نشان می‌دهد.

واژگان کلیدی. برآورد ناحیه‌ی کوچک؛ مدل سطح ناحیه؛ بیز سلسله‌مراتبی؛ نمونه‌گیری گیبس؛ قدرت قرضی.

* نویسنده‌ی عهده‌دار مکاتبات

۱ مقدمه

امروزه تقاضا برای تعیین آمارهایی با دقت لازم برای برآورد پارامترهای توصیفی ناحیه‌ی کوچک به‌شدت افزایش یافته است. در ناحیه‌ی کوچک، برآوردگرهای مستقیم به‌علت اندازه‌ی کم نمونه از دقت لازم برخوردار نیستند؛ لذا لازم است با استفاده از منابع مختلف اطلاعات کمکی به برآوردگرها قدرت قرضی داده، دقت آن‌ها را افزایش دهیم.

یکی از راه‌های دادن قدرت قرضی به برآوردگرها استفاده از مدل‌ها است. در مدل ناحیه‌ی کوچک علاوه بر اطلاعات کمکی از اثرهای ناحیه نیز برای برآورد پارامتر مورد نظر استفاده می‌شود. برتری روش‌های برآورد ناحیه‌ی کوچک مبتنی بر مدل بر سایر روش‌های برآورد ناحیه‌ی کوچک این است که با توجه به داده‌ها می‌توان مناسب‌ترین مدل را انتخاب و فرض‌های در نظر گرفته شده را ارزیابی کرد.

مدل مورد بررسی در این مقاله، مدل سطح ناحیه است. این مدل را اغلب به‌عنوان مدل فِی-هریوت (فِی و هریوت، ۱۹۷۹) می‌شناسند. در این مدل، هدف، استنباط درباره‌ی تابع مناسبی از میانگین‌های ناحیه‌ی کوچک iام، یعنی $\bar{Y}_i$، به‌صورت $\theta_i = g(\bar{Y}_i)$ است که با بردار متغیرهای کمکی $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^{\mathrm{T}}$، به‌صورت خطی

$$\theta_i = \mathbf{x_i}^{\mathrm{T}}\beta + b_i v_i, \qquad i = 1, \ldots, m \tag{۱}$$

ارتباط دارد. در این رابطه b_iها ضرایب ثابت معلومی هستند که معمولاً برابر با ۱ فرض می‌شوند و $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ یک بردار p-بعدی از ضرایب رگرسیونی است و باید برآورد شود. علاوه بر این v_i اثر تصادفی خاص ناحیه‌ی کوچک iام است، که فرض می‌شود میانگین و واریانس آن به‌ترتیب ۰ و σ_v^2 هستند. اغلب فرض می‌شود که v_iها متغیرهای تصادفی از توزیع نرمال و مستقل از یکدیگر هستند.

برای انجام استنباط‌های آماری در باره‌ی میانگین‌های ناحیه‌ی کوچک $\bar{Y}_i$ بر اساس مدل (۱)، فرض می‌شود برآوردگرهای مستقیم میانگین‌ها یعنی $\hat{\bar{Y}}_i$، در دسترس است. رابطه‌ی بین برآوردگرهای مستقیم و میانگین‌های ناحیه‌ی کوچک به‌صورت

$$y_i = g(\hat{\bar{Y}}_i) = \theta_i + e_i, \qquad i = 1, \ldots, m \tag{۲}$$

می‌باشد، که در آن e_iها خطاهای نمونه‌گیری‌اند و با توجه به طرح نمونه‌گیری p، فرض می‌شود مستقل و دارای میانگین و واریانس

$$E_p(e_i|\theta_i) = 0, \qquad V_p(e_i|\theta_i) = \psi_i$$

باشند. معمولاً فرض می‌شود واریانس‌های نمونه‌ای ψ_i معلوم‌اند. از آن‌جا که این یک فرض محدودکننده است، در این مقاله با توجه به نتایج به دست آمده توسط یو و چیمن (۲۰۰۶) حالت مجهول آن نیز بررسی

می‌شود. از ترکیب دو فرمول بالا، یعنی (۱) و (۲)، مدل

$$y_i = \mathbf{x_i}^{\mathrm{T}}\beta + b_i v_i + e_i, \qquad i = ۱, \ldots, m \tag{۳}$$

به دست می‌آید. این مدل شامل اثر تصادفی ناحیه‌ی iام و خطاهای نمونه‌گیری است. اغلب (از جمله در این مقاله) فرض می‌شود e_iها و v_iها مستقل هستند.

برای برآورد کردن پارامترهای ناحیه‌ی کوچک معمولاً از سه رهیافت بهترین پیشگوی نااریب خطی تجربی (EBLUP)، بیز تجربی (EB) و بیز سلسله‌مراتبی (HB) استفاده می‌شود که در بخش‌های بعد این مقاله شرح داده می‌شوند.

در این مقاله، در بخش ۲، رهیافت بهترین پیشگوی نااریب خطی تجربی را برای برآورد کردن میانگین ناحیه‌ی کوچک به کار می‌بریم و در بخش ۳ رهیافت بیز تجربی را برای برآورد کردن میانگین مورد استفاده قرار می‌دهیم. سپس در بخش ۴ برای دو حالت معلوم و نامعلوم بودن واریانس، خطای نمونه‌گیری میانگین ناحیه‌ی کوچک را به دست می‌آوریم. در بخش ۵ این رهیافت‌ها را برای برآورد کردن میانگین درامد در چند استان کشور به کار می‌بریم و مقایسه می‌کنیم. در انتها، نتیجه‌گیری در بخش ۶ بیان می‌شود.

۲ بهترین پیشگوی نااریب خطی تجربی در مدل سطح ناحیه

اگر مدل (۳) را برای m ناحیه به‌صورت برداری بنویسیم به یک مدل آمیخته‌ی خطی (سیرل، ۱۹۷۱) تبدیل می‌شود. با توجه به نتایج کلّی مدل‌های آمیخته‌ی خطی می‌توان برآوردگر پارامتر ناحیه‌ی کوچک θ_i را به دست آورد.

هندرسون (۱۹۷۵) با انجام این کار و با فرض معلوم بودن واریانس اثرهای ناحیه و واریانس خطای نمونه‌گیری، برآوردگر BLUP برای پارامتر θ_i را به‌صورت

$$\begin{aligned} \tilde{\theta}_i &= \mathbf{x}_i^{\mathrm{T}}\tilde{\beta} + \gamma_i(y_i - \mathbf{x}_i^{\mathrm{T}}\tilde{\beta}) \\ &= \gamma_i y_i + (۱ - \gamma_i)\mathbf{x_i}^{\mathrm{T}}\tilde{\beta} \end{aligned} \tag{۴}$$

به دست آورد، که در آن برآوردگر β به‌صورت

$$\tilde{\beta} = \left[\frac{\sum_{i=۱}^{m} \mathbf{x_i}\mathbf{x_i}^{\mathrm{T}}}{\sigma_v^۲ b_i^۲ + \psi_i}\right]^{-۱} \left[\frac{\sum_{i=۱}^{m} \mathbf{x_i} y_i}{\sigma_v^۲ b_i^۲ + \psi_i}\right]$$

و

$$\gamma_i = \frac{\sigma_v^2 b_i^2}{\sigma_v^2 b_i^2 + \psi_i}$$

می‌باشد. برآوردگر (۴) یک میانگین وزنی از برآوردگر مستقیم θ_i و برآوردگر رگرسیونی $\mathbf{x}_i^{\mathrm{T}}\tilde{\beta}$ با وزن‌های γ_i و $1-\gamma_i$ است. وزن γ_i میزان تغییرپذیری درون هر ناحیه نسبت به کل تغییرپذیری آن ناحیه را نشان می‌دهد.

در روابط فوق σ_v^2 مجهول است، که در روش EBLUP با یکی از روش‌های فراوانی‌گرا برآورد می‌شود. گوش و رائو (۱۹۹۴) با استفاده از روش گشتاورها و با فرض توزیع نرمال برای اجزای تصادفی، برآوردگر σ_v^2 را به‌صورت

$$(۵) \quad \hat{\sigma}_v^2 = \frac{1}{m-p}\left[\sum_{i=1}^{m}(y_i - \mathbf{x}_i^{\mathrm{T}}\beta^*)^2 - \sum_{i=1}^{m}\psi_i\left(1 - \mathbf{x_i}^{\mathrm{T}}\left(\sum_{\mathbf{i}}\mathbf{x_i}\mathbf{x_i}^{\mathrm{T}}\right)^{-1}\mathbf{x_i}\right)\right]$$

ارائه دادند، که در آن p بعد بردار ضرایب رگرسیونی، β^* برآوردگر کم‌ترین توان‌های دوم خطا و m تعداد ناحیه‌های کوچک است.

۳ برآورد بیز تجربی پارامترهای مدل سطح ناحیه

در رهیافت بیزی با فرض آن‌که پارامترهای توزیع پیشین (ابرپارامترها) معلوم هستند، برآوردگر بیزی را می‌توان به دست آورد. اما چون در عمل، اغلب این ابرپارامترها مجهول هستند، در روش بیز تجربی با یکی از روش‌های فراوانی‌گرا از قبیل روش ماکسیمم درست‌نمایی یا روش گشتاوری یا هر روش دیگری آن‌ها را برآورد می‌کنند. سپس با قرار دادن برآورد حاصل به‌جای ابرپارامترها، برآورد بیز تجربی پارامتر مورد نظر را به دست می‌آورند (کارلین و لوئیس، ۲۰۰۰).

موریس (۱۹۸۳) با فرض $v_i \overset{iid}{\sim} N(0,\sigma_v^2)$ و $e_i \overset{ind}{\sim} N(0,\psi_i)$ برآورد بیز تجربی پارامتر θ_i را با در نظرگرفتن تابع زیان درجه‌ی دوم به‌صورت

$$\hat{\theta}_i^{\mathrm{EB}} = \tilde{\gamma}_i y_i + (1-\tilde{\gamma}_i)\mathbf{x}_i^{\mathrm{T}}\tilde{\beta}$$

به دست آورد، که در آن

$$\tilde{\gamma}_i = \frac{\hat{\sigma}_v^2 b_i^2}{\hat{\sigma}_v^2 b_i^2 + \psi_i}$$

است. برآورد ماکسیمم درست‌نمایی پارامترهای σ_v^2 و β با حل تکراری دستگاه معادلات

$$(۶) \qquad \begin{cases} \sigma_v^2 = \frac{\sum_{i=1}^{m} w_i \{\frac{m}{m-p}((y_i - \mathbf{x_i}^{\mathrm{T}}\beta)^2 - \psi_i)\}}{\sum_{i=1}^{m} w_i} \\ \beta = (X^{\mathrm{T}} V^{-1} X)^{-1} X^{\mathrm{T}} V^{-1} Y \end{cases}$$

به دست می‌آید. در این روابط، $Y = (y_1, \ldots, y_m)^{\mathrm{T}}$، $w_i = \frac{1}{\sigma_v^2 + \psi_i}$، X ماتریس اطلاعات کمکی و $V = \mathrm{Diag}(b_1^2\sigma_v^2 + \psi_1, \ldots, b_m^2\sigma_v^2 + \psi_m)$ است.

چون در رهیافت بیز تجربی ابرپارامترها باید برآورد شوند، نوعی عدم حتمیت در برآورد لحاظ می‌شود. برای رفع این مشکل از رهیافت بیز سلسله‌مراتبی استفاده می‌کنند.

۴ رهیافت بیز سلسله‌مراتبی در مدل سطح ناحیه

در مدل‌های بیز سلسله‌مراتبی اطلاعات موجود در سطح اول برای تعیین توزیع پسین کافی نیست و به این دلیل این اطلاعات را در چند سطح تقسیم می‌کنند. سپس با ترکیب این اطلاعات با توجه به قضیه‌ی بیز، توزیع پسین به دست می‌آید. میانگین توزیع پسین به‌عنوان برآوردگر بیز سلسله‌مراتبی و واریانس پسین به‌عنوان معیار ارزیابی برآوردگرها محاسبه می‌شوند.

حال با استفاده از رهیافت HB برآورد پارامتر θ_i مدل (۳) را مورد بررسی قرار می‌دهیم. بدون این‌که از کلّیت مسئله کاسته شود فرض می‌کنیم $b_i = 1$ باشد. این رهیافت را در دو حالت معلوم بودن و نامعلوم بودن واریانس خطای نمونه‌گیری مورد بررسی قرار می‌دهیم.

۴٫۱ برآورد میانگین با واریانس خطای نمونه‌گیری معلوم

در این حالت، فرضیات مدل (۳) را با در نظرگرفتن یک توزیع مسطح برای β و با فرض $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ و $e_i \overset{ind}{\sim} N(0, \psi_i)$ به‌صورت

$$\begin{aligned} &(i) \quad y_i|\theta_i, \beta, \sigma_v^2 \overset{ind}{\sim} N(\theta_i, \psi_i) \\ &(ii) \quad \theta_i|\beta, \sigma_v^2 \overset{ind}{\sim} N(\mathbf{x}_i^{\mathrm{T}}\beta, \sigma_v^2) \\ &(iii) \quad \pi(\beta) \propto 1 \\ &(iv) \quad \pi(\beta, \sigma_v^2) = \pi(\beta)\pi(\sigma_v^2) \propto \pi(\sigma_v^2) \end{aligned}$$

$$(v) \quad \pi(\sigma_v^2) \propto \mathrm{IG}(a,b)$$

در نظر می‌گیریم (رائو، ۲۰۰۳). در روابط بالا $\pi(\sigma_v^2)$ توزیع پیشین σ_v^2 است، که برای اجتناب از ناسره شدن، توزیع پیشین آن را گامای معکوس (IG) در نظر می‌گیریم، که ابرپارامترهای a و b اعدادی معلوم و مثبت هستند.

برآوردگر نقطه‌ای θ_i در رهیافت HB، $E(\theta_i|Y)$ است. برای محاسبه‌ی این امید باید چگالی‌های پسین را به دست آورد. این کار اغلب به انتگرال‌های دارای بعد زیاد می‌رسد که حل آن‌ها مشکل است. بنا بر این برای به دست آوردن جواب‌ها از روش‌های شبیه‌سازی مونت‌کارلوی زنجیر مارکوفی مانند نمونه‌گیری گیبس (گلفند و اسمیت، ۱۹۹۰) استفاده می‌کنند. نمونه‌گیری گیبس یک ابزار شبیه‌سازی است که در آن نمونه‌ای تصادفی از یک تابع چگالی حاشیه‌ای، بدون این‌که نیاز به محاسبه‌ی خود تابع چگالی باشد، تولید می‌کند. از نمونه‌ی به دست آمده برای برآورد پارامترهای مورد نظر استفاده می‌شود.

برای این منظور با استفاده از فرضیات (i) تا (v) توزیع‌های شرطی کامل محاسبه می‌شوند. پس از انجام محاسبات لازم (رائو، ۲۰۰۳) خواهیم داشت:

$$(\theta_i|\beta,\sigma_v^2,Y) \sim N(\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i^{\mathrm{T}}\beta, \gamma_i\psi_i),$$

$$(\beta|\theta,\sigma_v^2,Y) \sim N_p\left(\left(\sum_i \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\left(\sum_i \mathbf{x}_i\theta_i\right),\ \sigma_v^2\left(\sum_i \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\right),$$

$$(\sigma_v^2|\beta,\theta,Y) \sim \mathrm{IG}\left(\frac{m}{2}+a,\ b+\frac{1}{2}\sum_{i=1}^{m}(\theta_i-\mathbf{x}_i^{\mathrm{T}}\beta)^2\right).$$

اگر G نشان‌دهنده‌ی تعداد تکرارهای نمونه‌گیری گیبس بعد از مرحله‌ی داغیدن، و L تعداد دفعات انجام شبیه‌سازی باشد و $\hat{\beta}_{lg}$ و $\hat{\sigma}_v^{2(lg)}$ نشان‌دهنده‌ی مقادیر به دست آمده برای σ_v^2 و β در تکرار gام و بار lام باشند، برآورد HB برای θ_i (رائو، ۲۰۰۳) به‌صورت زیر خواهد شد:

$$\frac{1}{LG}\sum_{l=1}^{L}\sum_{g=1}^{G}\left\{\frac{\hat{\sigma}_v^{2(lg)}}{\hat{\sigma}_v^{2(lg)}+\psi_i}y_i + \frac{\psi_i}{\hat{\sigma}_v^{2(lg)}+\psi_i}\mathbf{x}_i^{\mathrm{T}}\hat{\beta}_{lg}\right\}.$$

۴٫۲ برآورد میانگین با واریانس خطای نمونه‌گیری نامعلوم

در این حالت فرض می‌کنیم که مشاهدات نمونه از جامعه‌ی نرمال اخذ شده‌اند و واریانس نمونه‌ای S_i^2، برآوردگر ناااریب ψ_i باشد. بنا بر این اگر n_i اندازه‌ی نمونه‌ی اخذ شده از ناحیه‌ی کوچک iام باشد و

$d_i = n_i - ۱$، آنگاه $d_i S_i^۲ \sim \psi_i \chi_{d_i}^۲$ و $S_i^۲$ از برآوردگر مستقیم میانگین ناحیه‌ی کوچک y_i مستقل است. با استفاده از این فرضیات و با به‌کارگیری مدل بیز سلسله‌مراتبی در دو حالت (یو و چپمن، ۲۰۰۶) مجهول بودن و معلوم بودن واریانس خطای نمونه‌گیری (به‌کارگیری $S_i^۲$ به‌جای ψ_i) برآوردگر پارامترهای ناحیه‌ی کوچک را به دست می‌آوریم.

در حالت اول، فرضیات مدل فِی-هریوت را به‌صورت زیر می‌نویسیم.

مدل ۱:

$$
\begin{aligned}
&(i) \quad y_i|\theta_i, \psi_i \overset{ind}{\sim} N(\theta_i, \psi_i), \qquad i = ۱, \ldots, m \\
&(ii) \quad \theta_i|\beta, \sigma_v^۲ \overset{ind}{\sim} N(\mathbf{x}_i^{\mathrm{T}}\beta, \sigma_v^۲), \qquad i = ۱, \ldots, m \\
&(iii) \quad d_i S_i^۲|\psi_i \sim \psi_i \chi_{d_i}^۲, \qquad i = ۱, \ldots, m, \quad d_i = n_i - ۱ \\
&(iv) \quad \pi(\beta) \propto ۱, \qquad i = ۱, \ldots, m \\
&(v) \quad \pi(\psi_i) \sim \mathrm{IG}(a_i, b_i), \qquad i = ۱, \ldots, m \\
&(vi) \quad \pi(\sigma_v^۲) \sim \mathrm{IG}(a, b), \qquad i = ۱, \ldots, m
\end{aligned}
$$

در این روابط، پارامترهای a، b، a_i و b_i برای $i = ۱, \ldots, m$ اعدادی ثابت و معلوم می‌باشند.

در مدل ۱ برای برآورد واریانس نمونه‌گیری ψ_i، آن را به‌عنوان یک متغیر تصادفی با توزیع مشخص وارد مدل می‌کنند. روش دیگری که در عمل ممکن است مورد استفاده قرارگیرد استفاده از $S_i^۲$ به‌عنوان مقدار واقعی ψ_i است.

مدل ۲:

$$
\begin{aligned}
&(i) \quad y_i|\theta_i \overset{ind}{\sim} N(\theta_i, \psi_i), \qquad i = ۱, \ldots, m \\
&(ii) \quad \theta_i|\beta, \sigma_v^۲ \overset{ind}{\sim} N(\mathbf{x}_i^{\mathrm{T}}\beta, \sigma_v^۲), \qquad i = ۱, \ldots, m \\
&(iii) \quad \pi(\beta) \propto ۱, \\
&(iv) \quad \pi(\sigma_v^۲) \sim \mathrm{IG}(a, b), \qquad a, b > ۰
\end{aligned}
$$

برای استفاده از نمونه‌گیری گیبس، چگالی‌های شرطی کامل را به‌صورت

$$
(\psi_i|Y, \theta, \beta, \sigma_v^۲) \sim \mathrm{IG}\left(a_i + \frac{d_i + ۱}{۲},\ b_i + \frac{(y_i - \theta_i)^۲ + d_i S_i^۲}{۲}\right), \quad i = ۱, \ldots, m, \quad d_i = n_i - ۱
$$

$$
(\theta_i|Y, \beta, \psi_i, \sigma_v^۲) \sim N(\gamma_i y_i + (۱ - \gamma_i)\mathbf{x}_i^{\mathrm{T}}\beta, \gamma_i \psi_i), \qquad i = ۱, \ldots, m
$$

$$(\beta|Y,\theta_i,\psi_i,\sigma_v^2) \sim N_p\left(\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{m}\mathbf{x}_i\theta_i\right),\ \sigma_v^2\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\right),$$

$$(\sigma_v^2|Y,\theta,\psi_i,\beta) \sim \mathrm{IG}\left(a+\frac{m}{2},\ b+\frac{1}{2}\sum_{i=1}^{m}(\theta_i-\mathbf{x}_i^{\mathrm{T}}\beta)^2\right)$$

به دست می‌آورند (یو و چپمن، ۲۰۰۶)، که در این روابط $\gamma_i = \frac{\sigma_v^2}{\sigma_v^2+\psi_i}$ است.

تولید یک نمونه‌ی تصادفی از توزیع‌های شرطی بالا ساده است و بنا بر این با استفاده از نمونه‌گیری گیبس می‌توان برآورد پارامترهای ناحیه‌ی کوچک را در این حالت به دست آورد. یو و چپمن (۲۰۰۶) با استفاده از روش گلفند و اسمیت (۱۹۹۰) برآوردگر بیز میانگین پسین حاصل از نمونه‌گیری گیبس را به‌صورت

$$\hat{\theta}_i^{\mathrm{HB}} = \frac{1}{\mathrm{LG}}\sum_{i=1}^{L}\sum_{g=1}^{G}(\gamma_i^{(lg)}y_i + (1-\gamma_i^{(lg)})\mathbf{x}_i^{\mathrm{T}}\beta^{(lg)}) \qquad (۷)$$

ارائه کردند، که در آن $\gamma_i^{(lg)} = \frac{\sigma_v^{2\,(lg)}}{\sigma_v^{2\,(lg)}+\psi_i^{(lg)}}$ است و L و G به‌ترتیب تعداد دفعات انجام شبیه‌سازی و تعداد تکرارها بعد از مرحله‌ی داغیدن می‌باشند.

به طریق مشابه برای مدل ۲ نیز توزیع‌های شرطی کامل به‌صورت

$$(\theta_i|Y,\beta,\sigma_v^2) \sim N(\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i^{\mathrm{T}}\beta, \gamma_i\psi_i), \qquad \gamma_i = \frac{\sigma_v^2}{\sigma_v^2+\psi_i},$$
$$i = 1,\ldots,m$$

$$(\beta|Y,\theta_i,\sigma_v^2) \sim N_p\left(\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{m}\mathbf{x}_i\theta_i\right),\ \sigma_v^2\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\right),$$

$$(\sigma_v^2|Y,\theta,\beta) \sim \mathrm{IG}\left(a+\frac{m}{2},\ b+\frac{1}{2}\sum_{i=1}^{m}(\theta_i-\mathbf{x}_i^{\mathrm{T}}\beta)^2\right)$$

خواهند شد (یو و چپمن، ۲۰۰۶). برای مدل ۲ نیز برآورد میانگین HB به همان صورت (۷) است، با این تفاوت که ψ_i جایگزین $\psi_i^{(lg)}$ می‌شود.

در ادامه مدل‌ها و روش‌های ارائه‌شده در این بخش و بخش‌های قبلی مقاله را برای به دست آوردن میانگین درامد در چند استان نمونه‌ی کشور مورد ارزیابی و مقایسه قرار می‌دهیم.

۵ محاسبه‌ی میانگین درامد در برخی استان‌های کشور

در این بخش با تحلیل یک مجموعه‌داده‌ی واقعی و به کارگیری شبیه‌سازی، مدل‌ها و روش‌های ارائه‌شده در بخش‌های قبل را مورد مقایسه و ارزیابی قرار می‌دهیم. از آن‌جا که هزینه و درامد از مسائل مهم خانواده و کشور است، داده‌های طرح آمارگیری از هزینه و درامد خانوارهای شهری و روستایی در سال ۱۳۸۵ برای ارزیابی نتایج مورد استفاده قرار گرفت.

طرح آمارگیری از هزینه و درامد خانوارهای شهری و روستایی، هر سال با هدف بررسی درامدها و هزینه‌های مصرفی خانوارهای شهری و روستایی، مقدار مصرف مواد خوراکی، تعیین سهم هزینه‌های مختلف در هزینه‌ی کل، تشخیص انواع درامدهای خانوار، و مطالعه‌ی تغییرات آن‌ها در طول زمان با توجه به یک طرح نمونه‌گیری دومرحله‌ای صورت می‌گیرد. واحد نمونه‌گیری مرحله‌ی اول، بلوک (آبادی) و واحد نمونه‌گیری مرحله‌ی دوم، خانوار است که با بهینه کردن اندازه‌ی نمونه‌ی مورد نیاز در سطح استان‌های کشور، واحدهای نمونه از هر استان به دست می‌آیند.

برای بررسی درستی مدل‌ها و روش‌های ارائه‌شده با در اختیار داشتن واحدهای نمونه‌ی به دست آمده از هر استان، فرض می‌شود بخشی از این واحدها را (با همان طرح نمونه‌گیری) در اختیار داریم. هدف این است که بدون طراحی و اجرای یک نمونه‌گیری جدید و با همین اندازه‌ی کم نمونه، برآوردهایی برای متغیر مورد نظر به دست آوریم که از نظر دقت، تفاوت چندانی با برآورد به دست آمده از کل نمونه نداشته باشند.

متغیر پاسخ مورد استفاده، متوسط درامد خالص سرانه‌ی اعضای شاغل یک خانوار شهری از مشاغل مزد و حقوق بگیری در استان‌های مذکور است. با توجه به مفهوم همبستگی، پس از بررسی اثرهای هم‌خطی احتمالی و به‌روش رگرسیونی گام‌به‌گام، متغیرهای کمکی مناسب از فایل موجود انتخاب می‌شوند. پس از بررسی نتایج تحلیل رگرسیونی، به‌ترتیب، متغیرهای متوسط تعداد اعضای شاغل خانوار، متوسط تعداد روزهای کار در هفته‌ی اعضای شاغل خانوار، و متوسط تعداد اعضای باسواد خانوار در استان‌های مذکور انتخاب شده‌اند.

پس از انتخاب متغیرهای پاسخ و کمکی به ترتیبی که بیان شد، از آن‌جا که اطلاعات موجود در مورد هر خانوار در قسمت‌های مختلف فایل مورد استفاده قرار داشت، لازم بود تا این اطلاعات مرتب شود. با انجام این کار و با توجه به شرایطی که در مورد اثرهای ناحیه‌ها (مانند استقلال اثرهای ناحیه‌ها) وجود داشت، مجموعه‌داده‌ی مورد استفاده به ترتیبی که در ادامه شرح داده می‌شود به دست آمد.

مجموعه‌داده‌ی مورد استفاده پس از مرتب شدن، از ۱۷۰۰ خانوار شهری، که در چهار ناحیه‌ی کوچک شامل استان‌های خراسان رضوی، لرستان، همدان و تهران زندگی می‌کنند، تشکیل شده است. این استان‌ها با توجه به متوسط درامد هر خانوار و اندازه‌ی نمونه‌ی اخذ شده از آن‌ها در طرح آمارگیری از هزینه و درامد

خانوارهای شهری و روستایی در سال ۱۳۸۵، برای ساختن جامعه‌ای مشابه با ساختار کشورمان، از میان استان‌های مختلف کشور انتخاب شده‌اند. شایان ذکر است که در طرح مذکور از استان تهران بیش‌ترین، از استان لرستان کم‌ترین، و از دو استان دیگر اندازه‌ی متوسطی نمونه انتخاب شده است. یک طرح نمونه‌گیری دومرحله‌ای را (با همان روش مرکز آمار ایران) در جامعه‌ای متشکل از این چهار استان اجرا می‌کنیم. با بهینه کردن اندازه‌ی نمونه (با دقت ۰/۹۵) در سطح استان‌ها، اندازه‌ی کل نمونه‌ی مورد نیاز تقریباً باید برابر ۴۰۰ باشد، اما امکان دسترسی به فقط ۲۱۲ واحد از این نمونه وجود دارد. از این رو مسئله‌ی مورد تحقیق به یک مسئله‌ی برآورد ناحیه‌ی کوچک تبدیل می‌شود و می‌توان با توجه به در اختیار بودن نمونه‌ی بهینه‌شده، به مقایسه‌ی برآوردگرها پرداخت؛ بدین ترتیب که از ۲۱۲ واحد موجود و اطلاعات کمکی آن‌ها، برآوردهای مستقیم و برآوردهای رهیافت‌های مختلف در استان‌های مذکور را به دست آورده، با برآوردهای مستقیم حاصل از ۱۷۰۰ خانوار مقایسه می‌کنیم.

با توجه به اختلاف زیاد اندازه‌ی نمونه‌ی در اختیار با اندازه‌ی نمونه‌ی بهینه، نمونه‌ی در اختیار نمی‌تواند معرف کل جامعه باشد. بنا بر این امکان انجام استنباط‌هایی در مورد پارامترهای ناحیه‌ی کوچک با همان دقت وجود ندارد. هدف این است که با استفاده از همین اندازه‌ی نمونه و اطلاعات کمکی قابل دسترس، دقت برآوردگر مستقیم ناحیه، یعنی میانگین نمونه را افزایش دهیم.

معیارهای ارزیابی مورد استفاده، متوسط توان دوم خطا (ASE)، متوسط قدر مطلق خطای نسبی (AARE)، متوسط قدر مطلق اریبی (AAB) و درصد بهبود (IP) است (گوش و رائو، ۱۹۹۴)، که به‌ترتیب به‌صورت

$$\begin{aligned}
\mathrm{ASE} &= \frac{1}{m}\sum_{i=1}^{m}(p_i-\hat{p}_i)^2,\\
\mathrm{AARE} &= \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{p}_i-p_i}{p_i}\right|,\\
\mathrm{AAB} &= \frac{1}{m}\sum_{i=1}^{m}|\hat{p}_i-p_i|,\\
\mathrm{IP} &= 100\times\frac{\mathrm{ASEDE}-\mathrm{ASE}}{\mathrm{ASEDE}}
\end{aligned}$$

تعریف می‌شوند. در این روابط، m تعداد ناحیه‌ها، p_i برآورد پارامتر مورد نظر از نمونه‌ی بهینه‌شده، $\hat{p}_i$ برآورد پارامتر مورد نظر با اندازه‌ی نمونه‌ی کم‌تر از بهینه، و ASEDE متوسط توان دوم خطای برآوردگر مستقیم (میانگین نمونه) است.

قبل از انجام شبیه‌سازی، لازم است تا درستی فرضیات مدل‌های مورد استفاده بررسی شود. برای

این منظور، مانده‌ها را مورد تحلیل قرار می‌دهیم. فرض‌های اساسی در تعیین برآوردگرها، فرض همسانی واریانس اثرهای ناحیه‌ها، نرمال بودن داده‌ها و استقلال اثرهای ناحیه‌ها است.

در تحلیل داده‌های درامد، اقتصاددانان معمولاً تبدیل لگاریتمی را به کار می‌برند. علت استفاده از این تبدیل، محاسبه‌ی ساده‌تر ضرایب کشش و نتایج تجربی بهتر آن است. در این مقاله ضمن بررسی تبدیل مذکور، از خانواده‌ی تبدیلات باکس-کاکس (باکس و کاکس، ۱۹۶۴)، به‌منظور نرمال‌سازی و تثبیت واریانس استفاده می‌شود و سپس نتایج حاصل مورد مقایسه قرار می‌گیرد. به‌علاوه، لزوم این تبدیلات نیز بیان خواهد شد. ابتدا لزوم به‌کارگیری تبدیلات لگاریتمی و باکس-کاکس را نشان داده، سپس درستی فرضیات مدل‌ها را پس از انجام این تبدیلات بررسی می‌کنیم.

برای بررسی همسانی واریانس ناحیه‌ها از آماره‌ی لون و برای بررسی نرمال بودن داده‌ها از آماره‌ی شاپیرو-ویلک استفاده می‌کنیم. سطح معنی‌داری آماره‌ی لون برای داده‌ها ۰/۰۵ و سطح معنی‌داری آماره‌ی شاپیرو-ویلک در استان‌های خراسان رضوی، همدان، لرستان و تهران به‌ترتیب ۰/۰۱، ۰/۱، ۰/۰۳ و ۰/۰۱ است. بنا بر این، درستی فرضیات برای داده‌های اصلی در سطح معنی‌داری ۰/۰۵ رد می‌شود یا به‌سختی قابل قبول است. لذا لازم است در صورت وجود، با تبدیلات مناسب، داده‌هایی به دست آورد که در آن‌ها این فرضیات با اطمینان بیش‌تری برقرار باشد. از این رو از دو تبدیل لگاریتمی و باکس-کاکس با مقدار ضریب برآورد شده‌ی ۰/۳۲۸۲۸۹ استفاده می‌کنیم.

برای بررسی نرمال بودن داده‌ها پس از به‌کارگیری تبدیلات، مقدار آماره‌ی شاپیرو-ویلک را برای هر دو تبدیل لگاریتمی و باکس-کاکس محاسبه می‌کنیم. سطح معنی‌داری آماره‌ی شاپیرو-ویلک برای آزمون فرض نرمال بودن تحت تبدیل لگاریتمی برای استان‌های مذکور به‌ترتیب ۰/۱، ۰/۰۵۱، ۰/۷۸ و ۰/۰۴۸ و برای تبدیل باکس-کاکس، به همان ترتیب، ۰/۳۶، ۰/۱۵، ۰/۷۴ و ۰/۰۸۴ است. سطح معنی‌داری آماره‌ی لون برای تبدیل لگاریتمی ۰/۳۲ و برای تبدیل باکس-کاکس ۰/۴۳ است. با مقایسه‌ی سطح‌های معنی‌داری هر دو تبدیل مشخص می‌شود که تبدیل باکس-کاکس، درستی فرضیات را بیش‌تر از تبدیل لگاریتمی توجیه می‌کند. همچنین باید توجه داشت که فرض نرمال بودن داده‌ها در تبدیل لگاریتمی برای استان تهران به‌سختی قابل قبول است. مقایسه‌ی سطح معنی‌داری آماره‌ها قبل و بعد از اعمال تبدیل لگاریتم و باکس-کاکس، لزوم و مناسب بودن به‌کارگیری این تبدیلات را نشان می‌دهد.

چون واریانس ناحیه‌ها نامعلوم است، در رهیافت‌های EB و EBLUP با توجه به روابط (۵) و (۶)، این واریانس برآورد می‌شود. به‌علاوه با توجه به حضور عرض از مبدأ در ضرایب رگرسیونی، در این رهیافت‌ها فقط از دو متغیر کمکی اول استفاده می‌شود و برای این‌که امکان مقایسه‌ی برآوردگرهای HB با این رهیافت‌ها موجود باشد، رهیافت HB علاوه بر سه متغیر کمکی با همان دو متغیر نیز در حالت‌های مختلف به دست خواهد آمد.

از آن‌جا که واحدهای نمونه با توجه به طرح نمونه‌گیری تصادفی ساده به دست آمده‌اند، در مدل‌هایی که فرض معلوم بودن واریانس خطای نمونه‌گیری را دارند، از برآورد واریانس نمونه‌گیری تصادفی ساده که به‌صورت

$$\hat{V}(\bar{Y}) = \left(\frac{۱}{n} - \frac{۱}{N}\right) S^{۲}$$

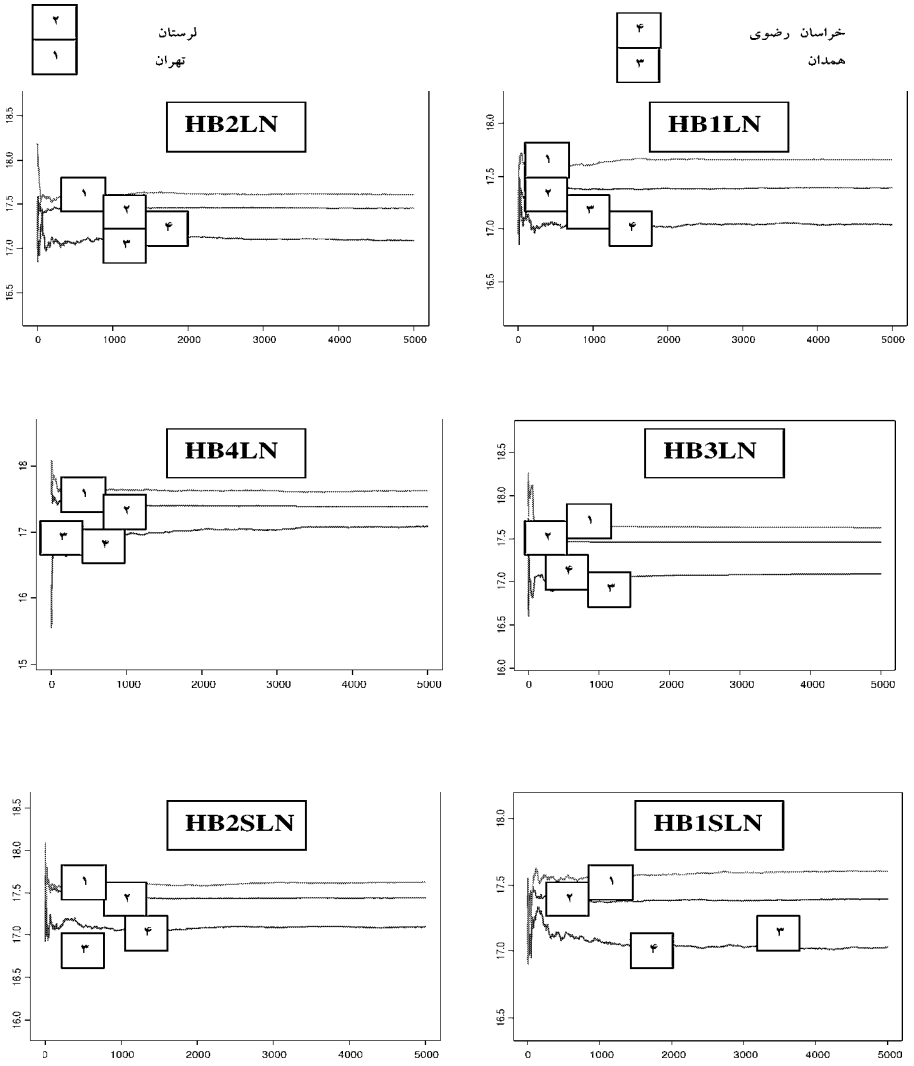
می‌باشد، استفاده می‌شود. در این رابطه n اندازه‌ی نمونه، N اندازه‌ی جامعه و $S^{۲}$ واریانس نمونه‌ای است. علاوه بر این در رهیافت HB این واریانس، با توجه به مدل‌های ۱ و ۲ (بخش ۳) نیز برآورد می‌شود و نتایج مورد مقایسه قرار می‌گیرد. حالت‌های مختلفی که برآوردگرها برای آن محاسبه شده‌اند عبارت‌اند از: ($\mathrm{EBLUP}^{\mathrm{ln}}$) بهترین پیشگوی نااریب خطی تجربی میانگین‌های جامعه تحت تبدیل لگاریتمی، ($\mathrm{EB}^{\mathrm{ln}}$) برآورد بیز تجربی میانگین‌ها تحت تبدیل لگاریتمی، ($\mathrm{HB}۱^{\mathrm{ln}}$) و ($\mathrm{HB}۲^{\mathrm{ln}}$) برآورد بیز سلسله‌مراتبی میانگین‌های ناحیه‌ی کوچک تحت تبدیل لگاریتمی، با دو متغیر کمکی (متوسط تعداد شاغلان و تعداد روزهای کار در هفته) و به‌ترتیب با به‌کارگیری واریانس نمونه‌ای و واریانس نمونه‌گیری تصادفی ساده به‌جای برآورد واریانس خطای نمونه‌گیری. ($\mathrm{HB}۳^{\mathrm{ln}}$) و ($\mathrm{HB}۴^{\mathrm{ln}}$) برآورد بیز سلسله‌مراتبی میانگین‌های ناحیه‌ی کوچک تحت تبدیل لگاریتمی، با هر سه متغیر کمکی و به‌ترتیب با به‌کارگیری واریانس نمونه‌ای و واریانس نمونه‌گیری تصادفی ساده به‌جای برآورد واریانس خطای نمونه‌گیری هستند. ($\mathrm{HB}۱S^{\mathrm{ln}}$) و ($\mathrm{HB}۲S^{\mathrm{ln}}$) برآورد بیز سلسله‌مراتبی میانگین‌های ناحیه‌ی کوچک تحت تبدیل لگاریتمی، با در نظر گرفتن توزیع پیشین گامای معکوس برای واریانس خطای نمونه‌گیری، به‌ترتیب با همان دو متغیر کمکی و هر سه متغیر کمکی هستند.

تعریف برآوردها برای تبدیل باکس-کاکس نیز به همین صورت است، با این تفاوت که توان ln، که نماد استفاده از تبدیل لگاریتمی است، با نماد bc، برای نشان دادن استفاده از تبدیل باکس-کاکس جایگزین می‌شود.

در بخش‌های قبل فرض کردیم واریانس‌های خطای نمونه‌گیری، ψ_iها، تصادفی و مستقل با توزیع $\mathrm{IG}(a_i, b_i)$ هستند. همچنین برای واریانس ناحیه‌ها، $\sigma_v^{۲}$، نیز (هنگام تصادفی بودن) توزیع $\mathrm{IG}(a, b)$ در نظر گرفته شد. ثابت‌های معلوم a_i و b_i، $(i \leqslant m \leqslant ۱)$، و a و b در این توزیع‌ها، اغلب خیلی کوچک انتخاب می‌شوند تا از کم‌ترین اطلاعات در باره‌ی $\sigma_v^{۲}$ و ψ_i استفاده شود یا اصطلاحاً توزیع پیشین مبهم باشد. به همین دلیل در شبیه‌سازی‌های انجام گرفته تمام مقادیر این ثابت‌ها برابر با ۰٫۰۰۱ انتخاب شده است. شایان ذکر است که حساس نبودن تحلیل‌ها به این مقادیر نیز مورد بررسی قرار گرفته است.

برای انجام شبیه‌سازی‌ها به‌روش نمونه‌گیری گیبس از نرم‌افزار S-PLUS استفاده شده است. برای این کار با استفاده از ۲۱۲ واحد نمونه، برآوردهای مستقیم میانگین ناحیه‌های کوچک و مقادیر متغیرهای

کمکی محاسبه می‌شود. بعد از آن با استفاده از این مقادیر و با انتخاب مقادیر اولیه‌ی مناسب، برآورد پارامترهای یکی از چگالی‌های شرطی کامل به دست می‌آید. سپس با استفاده از این برآوردها، یک نمونه



شکل ۱. بررسی همگرایی تبدیل لگاریتمی

از چگالی مربوط تولید می‌شود. از مقدار تولید شده برای تولید یک نمونه از چگالی شرطی کامل بعدی استفاده می‌شود و این روند تا تولید یک نمونه از تمام چگالی‌های شرطی کامل ادامه می‌یابد. این کار را تا زمانی که داده‌های تولید شده در هر مرحله با مرحله‌ی قبل تفاوت زیادی نداشته باشد، انجام می‌دهیم. چون اندازه‌ی نمونه‌ی تولید شده توسط نرم‌افزار در هر بار انجام شبیه‌سازی متفاوت با دفعات دیگر است، برای بررسی این امر، شبیه‌سازی‌ها را چند بار تکرار کرده‌ایم و نتایج را مورد ارزیابی قرار داده‌ایم.

شایان ذکر است، در رهیافت HB، چون برآوردها با توجه به الگوریتم نمونه‌گیری گیبس به دست می‌آیند، نخست باید همگرایی الگوریتم‌ها را مشخص کرد. برای این کار، نمودار مجموع تجمعی داده‌های تولید شده را برای برآوردگرهای مختلف در هر بار تکرار شبیه‌سازی رسم می‌کنیم. نتایج انجام این کار برای یک بار تکرار شبیه‌سازی در شکل‌های ۱ و ۲ برای حالت‌های مختلف برآورد HB با توجه به هر تبدیل به‌طور جداگانه رسم شده است. در همه‌ی این نمودارها خطوط ۱ و ۲ و ۳ و ۴ به‌ترتیب نمودار داده‌های شبیه‌سازی‌شده از استان‌های تهران، لرستان، همدان و خراسان رضوی را نشان می‌دهند. در این نمودارها محور عمودی، مقدار شبیه‌سازی‌شده برای پارامترها، و محور افقی، تعداد دفعات انجام شبیه‌سازی است.

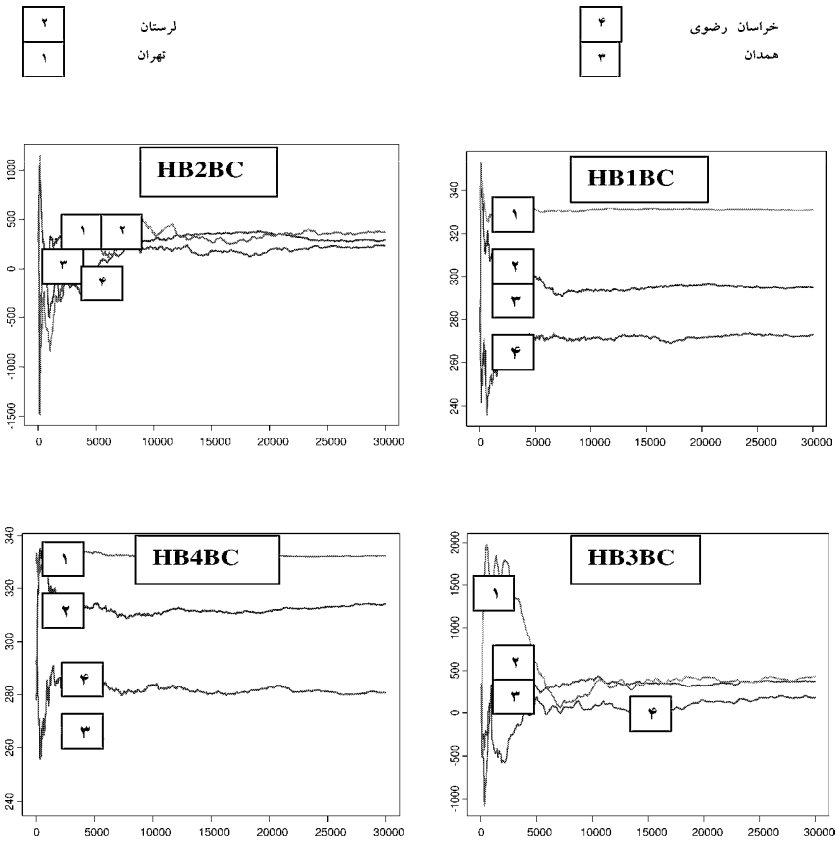
همان‌طور که از این شکل‌ها مشخص است، داده‌های شبیه‌سازی‌شده در ابتدا روند کاملاً تصادفی دارند، ولی در همه‌ی این نمودارها بعد از چند مرحله تولید داده، روندی ثابت پدیدار می‌شود به‌طوری که داده‌های تولید شده در هر مرحله تفاوت زیادی با مرحله‌ی قبل ندارد، یا به تعبیری نمودار همگرا می‌شود. تعداد دفعات رسیدن به این همگرایی یا مرتبه‌ی داغیدن برای مدل‌های مختلف و نوع تبدیل یکسان نیست، ولی همان‌طور که از این شکل‌ها پیدا است در تبدیل لگاریتمی ۵۰۰۰ تکرار و برای تبدیل باکس-کاکس حدود ۳۰۰۰۰ تکرار برای رسیدن به همگرایی لازم است.

همچنین چون الگوریتم گیبس برای شروع نیاز به مقادیر اولیه دارد و درست انتخاب کردن این مقادیر سبب کاهش مرتبه‌ی داغیدن و زمان همگرایی می‌شود، تحلیل تجربی داده‌ها نشان داد که بهتر است برای شروع، از برآوردهای روش‌های فراوانی‌گرا مانند روش درست‌نمایی ماکسیمم و روش گشتاوری استفاده شود.

۶ نتیجه‌گیری

نتایج انجام محاسبات برآوردهای مختلف و معیارهای دقت متناظرشان برای تبدیل لگاریتمی در جدول‌های ۱ و ۲ و برای تبدیل باکس-کاکس در جدول‌های ۳ و ۴ آمده است. همان‌طور که از این جدول‌ها مشخص است، برآوردگر مستقیم با توجه به هر سه معیار، برآورد نامناسب‌تری برای میانگین ناحیه‌ی کوچک ارائه می‌دهد. همچنین تحلیل جدول‌های به دست آمده نتایج تقریباً متفاوتی در استفاده از تبدیل باکس-کاکس و تبدیل لگاریتمی را نشان می‌دهد. در تبدیل لگاریتمی، بیش‌ترین

متوسط درصد بهبود معیار ASE، برابر با ۳۴/۳۷ بوده و مربوط به رهیافت HB، هنگام برآورد واریانس خطای نمونه‌گیری با توزیع پیشین IG و با دو متغیر کمکی متوسط تعداد افراد شاغل خانوار و متوسط تعداد روزهای کار در هفته، به دست آمده است. به‌علاوه در این تبدیل، $HB1^{ln}$ نیز ASE برآورد مستقیم را به‌طور متوسط ۲۸/۸۴ درصد بهبود بخشیده است. همچنین در استفاده از این تبدیل، دو رهیافت EBLUP و EB نیز ASE برآوردگر مستقیم را به‌ترتیب ۱۵/۸ و ۸/۷۸ درصد بهبود بخشیده‌اند، که نشان می‌دهد درصد بهبود رهیافت EBLUP تقریباً دو برابر رهیافت EB است.



شکل ۲. بررسی همگرایی تبدیل باکس-کاکس

جدول ۱. برآورد متوسط درامد خانوار (ریال) برای استان‌های نمونه در تبدیل لگاریتمی

	استان			
برآورد	خراسان رضوی	همدان	لرستان	تهران
اندازه‌ی نمونه	۴۲	۲۸	۲۰	۱۲۲
میانگین جامعه	۲۶۳۰۶۵۹۹	۲۸۰۷۸۹۸۳	۳۲۸۵۳۳۴۳	۳۹۰۸۱۴۹۶
میانگین نمونه	۲۳۰۳۷۰۶۵	۲۶۶۲۹۵۷۵	۳۸۰۶۴۰۱۹	۴۵۱۴۱۲۹۴
$\mathrm{EBLUP}^{\ln}$	۲۸۱۳۹۷۹۸	۲۲۰۴۳۹۵۱	۳۳۴۷۰۵۶۵	۴۵۰۸۴۳۸۸
$\mathrm{EB}^{\ln}$	۲۷۵۳۶۲۷۸	۲۲۵۰۰۰۲۴	۳۳۹۴۱۴۹۵	۴۵۰۹۰۵۶۷
$\mathrm{HB}۱^{\ln}$	۲۶۳۵۱۱۲۹	۲۵۰۳۶۳۹۱	۳۵۸۰۷۶۹۰	۴۵۱۰۹۹۰۰
$\mathrm{HB}۲^{\ln}$	۲۳۵۵۹۴۱۲	۲۶۲۵۶۰۲۸	۳۷۵۲۲۵۶۶	۴۵۱۱۳۴۴۱
$\mathrm{HB}۳^{\ln}$	۲۳۱۶۱۵۰۸	۲۶۳۱۰۶۹۹	۳۷۷۷۴۹۳۳	۴۵۱۶۶۲۰۸
$\mathrm{HB}۴^{\ln}$	۲۳۰۳۲۰۲۷	۲۶۶۳۳۲۰۰	۳۸۰۷۵۱۵۷	۴۵۱۳۲۶۴۸
$\mathrm{HB}۱S^{\ln}$	۲۶۶۵۳۴۱۲	۲۵۶۰۵۷۹۳	۳۵۸۱۷۱۳۴	۴۴۸۷۸۸۰۱
$\mathrm{HB}۲S^{\ln}$	۲۳۱۷۳۵۵۰	۲۶۶۰۵۸۸۵	۳۷۹۷۰۱۶۲	۴۴۴۹۷۹۳۰

جدول ۲. ارزیابی برآوردهای به دست آمده در تبدیل لگاریتمی

	معیارها		
برآورد	AARE	AAB	درصد بهبود ASE
میانگین نمونه	۰٫۱۲۲۴	۳۹۹۷۳۵۴	—
$\mathrm{EBLUP}^{\ln}$	۰٫۱۱۳۳	۳۷۱۹۴۹۴	۱۵٫۸
$\mathrm{EB}^{\ln}$	۰٫۱۰۸	۳۴۷۶۴۶۵	۸٫۷۸
$\mathrm{HB}۱^{\ln}$	۰٫۰۹۱	۳۰۴۲۰۸۱	۲۸٫۸۴
$\mathrm{HB}۲^{\ln}$	۰٫۱۱۶۴	۳۸۲۵۱۴۷	۹٫۲۷
$\mathrm{HB}۳^{\ln}$	۰٫۱۲۳۲	۳۹۷۹۹۱۸	۱٫۷
$\mathrm{HB}۴^{\ln}$	۰٫۱۲۲۵	۴۰۰۰۰۵۴	۰٫۰۷
$\mathrm{HB}۱S^{\ln}$	۰٫۰۸۸۸	۲۹۷۶۴۷۵	۳۴٫۳۷
$\mathrm{HB}۱S^{\ln}$	۰٫۱۱۸۹	۳۸۹۸۸۹۱	۶٫۷

در تبدیل باکس-کاکس، نتایج تا حدودی متفاوت است و به‌طور کلّی مجدداً برتری رهیافت HB در برآورد میانگین جامعه را نسبت به برآورد مستقیم و دو رهیافت دیگر نشان می‌دهد. بیش‌ترین درصد بهبود (۴۸٫۵۴) هنگام استفاده از سه متغیر کمکی و واریانس نمونه‌گیری تصادفی ساده حاصل شده است.

جدول ۳. برآورد متوسط درامد خانوار (ریال) استان‌های نمونه در تبدیل باکس-کاکس

	استان			
برآورد	خراسان رضوی	همدان	لرستان	تهران
اندازه‌ی نمونه	۴۲	۲۸	۲۰	۱۲۲
میانگین جامعه	۲۶۳۰۶۵۹۹	۲۸۰۷۸۹۸۳	۳۲۸۵۳۳۴۳	۳۹۰۸۱۴۹۶
میانگین نمونه	۲۳۰۳۷۰۶۵	۲۶۶۲۹۵۷۵	۳۸۰۶۴۰۱۹	۴۵۱۴۱۲۹۴
EBLUP^{bc}	۲۳۲۶۳۰۷۴	۲۶۴۴۷۲۷۰	۳۷۷۹۴۴۶۳	۴۵۱۳۸۰۵۹
EB^{bc}	۲۴۴۴۱۶۷۶	۲۵۵۲۸۴۱۹	۳۶۲۵۷۹۳۱	۴۵۱۲۱۵۴۴
$\mathrm{HB1}^{bc}$	۲۷۵۹۳۴۱۲	۲۵۳۶۵۴۴۴	۲۸۵۶۰۲۷۳	۴۴۵۶۱۲۰۰
$\mathrm{HB2}^{bc}$	۲۷۶۰۰۶۱۷	۲۵۶۷۶۹۱۰	۲۸۳۰ ۷۸۶۹	۴۴۳۶۲۸۱۹
$\mathrm{HB3}^{bc}$	۲۴۰۱۰۰۹۵	۲۷۱۶۴۳۰۶	۳۰۶۱۶۸۸۹	۴۴۴۱۴۲۱۳
$\mathrm{HB4}^{bc}$	۲۴۱۵۴۷۹۶	۲۷۰۵۵۹۸۳	۳۰۶۲۰۸۹۱	۴۱۱۶۳۶۷۶
$\mathrm{HB1}S^{bc}$	۲۳۱۷۱۴۰۵	۲۶۵۵۶۹۷۵	۳۷۴۰۳۰۵۲	۴۵۱۳۹۴۷۱
$\mathrm{HB2}S^{bc}$	۲۳۱۷۱۴۰۵	۲۶۵۵۶۹۷۵	۳۷۴۰۳۰۵۲	۴۵۱۳۹۴۷۱

جدول ۴. ارزیابی برآوردهای به دست آمده از تبدیل باکس-کاکس

	معیار		
برآورد	AARE	AAB	درصد بهبود ASE
میانگین نمونه	۰٫۱۲۲۴	۳۹۹۷۳۵۴	—
EBLUP^{bc}	۰٫۱۱۹۵	۳۹۱۰۷۳۰	۵٫۱۳
EB^{bc}	۰٫۱۰۵	۳۴۶۵۰۳۱	۲۴٫۲۷
$\mathrm{HB1}^{bc}$	۰٫۱۰۰۸	۳۳۷۸۲۲۹	۲۹٫۳۶
$\mathrm{HB2}^{bc}$	۰٫۰۹۳۶	۳۲۸۷۷۶۱	۳۴٫۹۹
$\mathrm{HB3}^{bc}$	۰٫۰۸۱۳	۲۷۴۸۴۰۳	۴۶٫۹۸
$\mathrm{HB4}^{bc}$	۰٫۰۸۰۹	۲۵۷۷۰۹۱	۴۸٫۷۱

با مقایسه‌ی برآوردهای به دست آمده برای استان‌های مختلف با مقدار واقعی متناظرشان بر اساس هر دو تبدیل (جدول‌های ۱ و ۳) کم‌ترین بهبود برآوردگر مستقیم در استان تهران دیده می‌شود، که با توجه به اندازه‌ی نمونه‌ی زیاد این استان در مقایسه با استان‌های دیگر، این نتیجه توجیه‌پذیر است. همچنین در رهیافت‌های مختلف بیزی استفاده‌شده در تبدیل لگاریتمی، استفاده از سه متغیرکمکی باعث بهبود

جدول ۵. بررسی درصد بهبودها در تکرارهای مختلف شبیه‌سازی گیبس

برآورد	تعداد شبیه‌سازی گیبس	کم‌ترین IP	بیش‌ترین IP	انحراف معیار
$HB1^{ln}$	۶۰	۲٫۸۲	۴۸٫۶۲	۱۵٫۰۶
$HB2^{ln}$	۶۰	۸٫۹۵	۹٫۸	۰٫۳۷
$HB3^{ln}$	۶۰	−۳۰	۱۹٫۱	۱۵٫۲۵
$HB4^{ln}$	۶۰	−۳٫۳۷	۵٫۱۱	۱٫۸۵
$HB1S^{ln}$	۶۰	۸٫۶۴	۵۷٫۹۷	۱۷٫۶۹
$HB2S^{ln}$	۶۰	۰٫۶۱	۱۳٫۶	۵٫۹۹
$HB1^{bc}$	۴۰	۲۵٫۰۲	۳۲٫۲۷	۳٫۸۳
$HB2^{bc}$	۴۰	۱۸٫۴	۴۷٫۴۶	۱۴٫۹۶
$HB3^{bc}$	۴۰	۳۶٫۶۵	۵۶٫۳۶	۷٫۱۴
$HB4^{bc}$	۴۰	۳۴٫۵۸	۵۷٫۹۷	۷٫۱۶

قابل ملاحظه‌ای در دقت برآوردگر مستقیم نشده است و این نکته را متذکر می‌شود که در انتخاب متغیر کمکی باید متغیرهای با بیش‌ترین تأثیر و تا حد ممکن با تعداد کم‌تر انتخاب شوند.

با وجود کوچک‌تر بودن واریانس نمونه‌گیری تصادفی ساده نسبت به واریانس نمونه‌ای، استفاده از این واریانس به‌جای واریانس خطای نمونه‌گیری در تبدیل لگاریتمی موجب بهبود نتیجه نشده است و استفاده از واریانس نمونه‌ای مناسب‌تر است.

همچنین مقایسه‌ی دو برآورد رهیافت HB بر اساس مدل‌های ۱ و ۲ حاکی از برتری نسبی رهیافت HB در حالت نامعلوم بودن واریانس (مدل ۱) را دارد، اما این روش مستلزم محاسبات بیش‌تر بوده، ممکن است در برخی موارد نمونه‌گیری گیبس مناسب نباشد (یو و چپمن، ۲۰۰۶). علاوه بر این، مقدار واریانس نمونه‌ای نیز در همگرایی الگوریتم‌ها مؤثر است، چنان‌که مقدار بزرگ واریانس نمونه‌ای در تبدیل باکس-کاکس سبب همگرا نشدن الگوریتم‌ها در هنگام برآورد واریانس خطای نمونه‌گیری با استفاده از رهیافت بیز سلسله‌مراتبی شده است. مقدار این واریانس برای استان‌های خراسان رضوی، همدان، لرستان و تهران با توجه به تبدیل لگاریتمی به‌ترتیب ۰٫۸۵، ۱٫۴۵، ۰٫۴۲ و ۰٫۸۷ و با توجه به تبدیل باکس-کاکس ۴۸۷۸٫۵، ۶۵۵۶٫۶، ۳۵۸۱٫۵ و ۷۱۸۳ می‌باشد.

بررسی معیارهای AARE و ABB نیز نتایج مشابهی با معیار ASE برای برآوردها در استفاده از هر دو تبدیل را نشان می‌دهد.

از نظر زمان همگرایی که مربوط به دو رهیافت EB و HB است، در تبدیل لگاریتمی، زمان همگرایی و مرتبه‌ی داغیدن نسبت به تبدیل باکس-کاکس کم‌تر است، که درست انتخاب کردن مقادیر اولیه نیز موجب

کاهش زمان همگرایی و مرتبه‌ی داغیدن در هر دو تبدیل می‌شود. همچنین تعداد تکرارها، به تعداد پارامترهای برآوردشده و تعداد متغیرهای مدل بستگی دارد. هرچه این تعدادها بیش‌تر باشد مرتبه‌ی داغیدن و زمان همگرایی بالاتر است. از این رو رهیافت HB برای داده‌های با حجم بالا با مشکلات محاسبه‌ای مواجه است.

همچنین اگر انحراف معیار درصدهای بهبود را در شبیه‌سازی‌های مختلف مورد بررسی قرار دهیم (جدول ۵)، می‌توان گفت با توجه به انحراف معیار کم تبدیل باکس-کاکس و درصدهای بهبود آن، استفاده از این تبدیل در هنگام همگرا شدن الگوریتم‌ها، نتایج بهتری را نسبت به تبدیل لگاریتمی ارائه می‌دهد.

نتایج این تحقیق به‌وضوح برتری رهیافت‌های EBLUP، EB و HB را بر برآوردگر مستقیم نشان می‌دهد، اما باید توجه کرد که دو رهیافت EBLUP و EB نمی‌توانند عدم حتمیتی که ناشی از برآورد کردن پارامترها است در مدل لحاظ کنند. از این رو این دو رهیافت ممکن است باعث کم‌برآوردی MSE شوند. همچنین با توجه به رابطه‌ی برآورد واریانس ناحیه‌ها، تعداد ناحیه‌های کوچک نیز به‌عنوان یک عامل در برآوردها تأثیرگذار است. علاوه بر این، رهیافت EBLUP را فقط در مدل سطح ناحیه و سطح واحد می‌توان به کار برد و در مدل‌های غیر خطی ناحیه‌ی کوچک غیر قابل استفاده است (رائو، ۲۰۰۳).

با توجه به این نتایج و محدودیت‌های ذکر شده برای رهیافت‌های EBLUP و EB، به‌طورکلّی می‌توان گفت اگر توزیع‌های پیشین و متغیرهای کمکی مناسب در اختیار باشند رهیافت HB باعث بهبود قابل ملاحظه‌ای نسبت به دو رهیافت دیگر و برآوردگر مستقیم می‌شود.

سپاس‌گزاری

نویسندگان از تمامی کارشناسان محترم مرکز آمار ایران و پژوهشکده‌ی آمار که در ارائه‌ی نظرهای کارشناسانه، در اختیار قرار دادن فایل داده‌ها و سایر امکانات همکاری داشته‌اند، صمیمانه تشکر و قدردانی می‌کنند.

مرجع‌ها

Box, G.E.P.; Cox, D.R. (1964). An analysis of transformations, *J. Roy. Stat. Soc.* **26**, 211-224.

Carlin, B.P.; Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, London.

Fay, R.E.; Herriot, R.A. (1979). Estimation of income from small places: an application of James-Stein procedures to census Ddata. *J. Amer. Stat. Assoc.* **74**, 269-277.

Gelfand, A.E.; Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Stat. Assoc.* **85**, 398-409.

Ghosh, M.; Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Stat. Sci.* **9**, 55-93.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423-447.

Morris, C.A. (1983). Parametric empirical Bayes inference: theory and application. *J. Amer. Stat. Assoc.* **78**, 47-54.

Rao, J.N.K. (2003). *Small Area Estimation.* Wiley, New York.

Searil, S.R. (1971). *Linear Models.* Wiley, New York.

You, Y.; Chapman, B. (2006). Small area estimation using area level model and estimated sampling variances. *Sur. Meth.* **32**, 97-103.

شاهو زارعی
گروه آمار، دانشکده‌ی علوم پایه،
دانشگاه تربیت مدرس،
تهران، ایران.
پیام‌نگار: *sh.zarei@uok.ac.ir*

عباس گرامی
گروه آمار، دانشکده‌ی ریاضی، آمار و علوم کامپیوتر،
دانشگاه تهران،
تهران، ایران.
پیام‌نگار: *agerami@ut.ac.ir*

مجید جعفری خالدی
گروه آمار، دانشکده‌ی علوم پایه،
دانشگاه تربیت مدرس،
تهران، ایران.
پیام‌نگار: *jafari-m@modares.ac.ir*