# مقایسه چند روش برای برآورد ناحیه کوچک و کاربرد آن در برآورد متوسط هزینه خانوار در استان‌های کشور

پروین جلیلی، محمدرضا فقیهی
بانک مرکزی جمهوری اسلامی ایران
گروه آمار، دانشگاه شهید بهشتی

اهمیت و گستردگی برآورد برخی پارامترها برای ناحیه‌های کوچک، در طرح‌هایی مانند طرح بررسی هزینه و درآمد خانوار و طرح اشتغال و بیکاری که نتایج حاصل از آنها مبنای بسیاری از سیاست‌های مالی و اقتصادی دولت قرار می‌گیرند و اغلب برای ناحیه‌های کوچک دربردارنده تعداد نمونه کافی نیستند، سبب شده است روش‌های برآورد ناحیه کوچک به سرعت گسترش یابند. در مطالعات پیش از این در بسیاری از موارد برتری روش بیز سلسله مراتبی بر برخی روش‌ها، به اثبات رسیده است. در این مقاله با استفاده از مدل‌های سطح ناحیه، سه روش، ناپارامتری دو مرحله‌ای، بهترین پیشگوی نااریب خطی تجربی و بیز سلسله مراتبی، تحت دو ساختار خطی و ناخطی را با یکدیگر مقایسه کرده و با توجه به ملاک مقایسه‌ای مناسب بهترین روش انتخاب شده است. نتایج حاصل از مقایسه برتری روش بیز سلسله مراتبی در مقایسه با دو روش دیگر نشان می‌دهند. همچنین از روش بیز سلسله مراتبی برای برآورد متوسط هزینه خانوار در استان‌های کشور در سال ۸۵، بر اساس اطلاعات جمع‌آوری شده توسط بانک مرکزی جمهوری اسلامی ایران استفاده شده است.

**واژه‌های کلیدی:** برآورد ناحیه کوچک، مدل سطح ناحیه، رگرسیون ناپارامتری، برآوردگر ناداریا- واتسون، اعتبارسنجی متقابل، بهترین پیشگوی نااریب خطی تجربی، بیز سلسله مراتبی، نمونه‌گیری گیبس.

## ۱ مقدمه

در طرح‌های نمونه‌ای پیمایشی، اندازه نمونه برای ناحیه بزرگ (حوزه بزرگ) مانند کشور، استان و غیره تعیین می‌گردد. بنابراین اطلاعات جمع‌آوری شده برای استنباط در مورد حوزه بزرگ مناسب بوده و از دقت لازم برخوردار است. حال آنکه در اکثر موارد حوزه بزرگ شامل



Archive of SID
www.SID.ir

زیرجامعه‌هایی است که کسب اطلاع در مورد ویژگی‌های این زیرجامعه‌ها برای ما مطلوب بوده ولی متاسفانه در اکثر موارد تعداد نمونه اخذ شده از زیرجامعه‌های مورد نظر کمتر از حد مورد نیاز برای انجام یک برآورد با دقت مناسب است. به این زیرجامعه‌ها در اصطلاح ناحیه‌های کوچک گفته می‌شود. از آنجا که بهینه کردن اندازه نمونه درسطح ناحیه‌های کوچک باعث افزایش قابل توجه هزینه آمارگیری می‌شود و همچنین استفاده از روش‌های مستقیم برآورد در مورد ناحیه‌های کوچک از دقت لازم برخوردار نخواهند بود، آمارشناسان سعی کرده‌اند با استفاده از اطلاعات کمکی برآوردگرهای مناسبی برای ناحیه‌های کوچک ارائه دهند.

در این راستا دو رویکرد مدل سطح ناحیه و مدل سطح واحد وجود دارد (رائو ۲۰۰۳). درمدل سطح ناحیه علاوه بر اطلاعات کمکی اغلب از اثرهای ناحیه نیز برای برآورد پارامتر مورد نظر استفاده می‌شود.

در این مقاله از مدل سطح ناحیه استفاده می‌شود. مدل فِی-هریوت (فِی-هریوت، ۱۹۷۹) یکی از معروف‌ترین مدل‌های سطح ناحیه است. هدف، استنباط درباره‌ی تابع مناسبی از میانگین‌های ناحیه کوچک $i$ام ( $i = ۱, ۲, \cdots, m$، که $m$ تعداد نواحی کوچک است.)، یعنی $\bar{Y}_i$، به صورت $\theta_i = g(\bar{Y}_i)$ است که $\theta_i$ با بردار متغیرهای کمکی $\boldsymbol{x}_i$ به صورت زیر در ارتباط است.

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + b_i \nu_i \qquad i = ۱, ۲, \cdots, m. \tag{۱}$$

در این رابطه $\boldsymbol{x}_i = (x_{i۱}, \cdots, x_{ip})^T$ یک بردار $p$-متغیره از متغیرهای کمکی ناحیه $i$ام، $\boldsymbol{\beta} = (\beta_۱, \cdots, \beta_p)^T$ بردار ضرایب رگرسیونی که نامعلوم است، $\nu_i$، متغیر تصادفی اثر ناحیه $i$ام، که فرض می‌شود، دارای میانگین صفر و واریانس، $\sigma_\nu^۲$ است (معمولاً فرض می‌شود $\nu_i \sim N(۰, \sigma_\nu^۲)$.)، $b_i$ها نیز مقادیر ثابت، مثبت و معلومی هستند که معمولاً ۱ در نظر گرفته می‌شوند. و $m$ نیز تعداد ناحیه‌های کوچک مورد بررسی هستند. برای استنباط در مورد میانگین ناحیه کوچک $i$ام، $(\bar{Y}_i)$ بر اساس مدل (۱)، اگر $\hat{\bar{Y}}_i$ برآوردگر مستقیم میانگین ناحیه $i$ام باشد، فرض می‌کنیم:

$$\hat{\theta}_i = g(\hat{\bar{Y}}_i) = \theta_i + e_i, \qquad i = ۱, ۲, \cdots, m. \tag{۲}$$

در این رابطه $e_i$ها خطاهای نمونه‌گیری هستند. که فرض می‌شود مستقل از هم و دارای میانگین و واریانس به صورت زیر باشند.

$$E(e_i|\theta_i) = ۰, \qquad Var(e_i|\theta_i) = \psi_i$$



Archive of SID

www.SID.ir

معمولاً فرض می‌شود واریانس‌های نمونه‌ای ($\psi_i$ها) معلوم هستند.
از ترکیب دو رابطه‌ی (۱) و (۲)، داریم:

$$\hat{\theta}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + b_i\nu_i + e_i\,, \qquad i = ۱, ۲, \cdots, m. \qquad (۳)$$

در اکثر موارد فرض می‌شود $\nu_i$ها و $e_i$ها از هم مستقل هستند.
در این مقاله برای برآورد میانگین‌های ناحیه‌های کوچک از سه روش، ناپارامتری دو مرحله‌ای، بهترین پیشگوی نااریب خطی تجربی ($EBLUP$) و بیز سلسله مراتبی ($HB$) استفاده می‌شود.
در بخش ۲ سه روش $EBLUP$، $HB$ و ناپارامتری دومرحله‌ای معرفی می‌شوند. که در مورد روش ناپارامتری دومرحله‌ای در حالتی که یک متغیر کمکی در مدل حضور داشته باشد و حالتی که بیش از یک متغیر کمکی در مدل باشد، بحث می‌شود.
در بخش ۳ سه روش مذکور بااستفاده از شبیه‌سازی در دو حالت وجود ارتباط خطی بین $\hat{\theta}_i$ و $\boldsymbol{x}_i$ و حالت وجود ارتباط ناخطی بحث می‌شود. و با استفاده از یک معیار مقایسه مناسب، سه روش مقایسه شده و بهترین روش انتخاب می‌شود.
در بخش ۴ با استناد به نتایج بخش ۳ بهترین روش از بین سه روش معرفی شده جهت برآورد متوسط هزینه خانوار در استان‌های کشور براساس اطلاعات جمع‌آوری شده توسط بانک مرکزی جمهوری اسلامی ایران در سال ۸۵، مورد استفاده قرار می‌گیرد.

## ۲ معرفی سه روش $EBLUP$، $HB$ و ناپارامتری دو مرحله‌ای

در این بخش سه روش $EBLUP$، $HB$ و ناپارامتری دومرحله‌ای معرفی می‌شوند. که روش ناپارامتری دومرحله‌ای در حالتی که یک متغیر کمکی در مدل حضور داشته باشد و حالتی که بیش از یک متغیر کمکی در مدل باشد، مورد بررسی قرار می‌گیرد.

### ۲-۱ روش بهترین پیشگوی نااریب خطی تجربی

با توجه به آنکه مدل (۳) یک مدل آمیخته خطی است. هندرسون (۱۹۷۵)، با فرض معلوم بودن واریانس اثرهای ناحیه و واریانس خطای نمونه‌گیری، برآوردگر زیر را برای $\theta_i$ پیشنهاد



Archive of SID

www.SID.ir

کرد.

$$\begin{aligned}\tilde{\theta}_i &= \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}} + \gamma_i(\hat{\theta}_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) \\ &= \gamma_i\hat{\theta}_i + (1-\gamma_i)\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}\end{aligned}$$

که در آن

$$\tilde{\boldsymbol{\beta}} = \left(\frac{\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T}{\sigma_\nu^2 b_i^2 + \psi_i}\right)^{-1}\left(\frac{\sum_{i=1}^m \boldsymbol{x}_i\hat{\theta}_i}{\sigma_\nu^2 b_i^2 + \psi_i}\right).$$

در روابط فوق $\sigma_\nu^2$ مجهول است، گوش و رائو (۱۹۹۴) با استفاده از گشتاورها و با فرض نرمال بودن خطا و اثرناحیه، برآوردگر زیر را ارائه کرده‌اند.

$$\hat{\sigma}_\nu^2 = \frac{1}{m-p}\left(\sum_{i=1}^m(\hat{\theta}_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*)^2 - \sum_{i=1}^m\psi_i\big(1 - \boldsymbol{x}_i^T(\sum_{i=1}^m \boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1}\boldsymbol{x}_i\big)\right)$$

که در آن $p$ بعد بردارضرایب رگرسیونی است، $\beta^*$ برآوردگر کمترین توان‌های دوم خطا و $m$ تعداد ناحیه‌های کوچک است.

## ۲-۲ روش بیز سلسله مراتبی

در این قسمت به منظور برآورد $\theta_i$ها، از روش بیز سلسله مراتبی با استفاده از روش نمونه‌گیری گیبس استفاده می‌کنیم (رائو، ۲۰۰۳). همانطور که ذکر شد معمولاً $b_i = 1$ در نظر گرفته می‌شود.

در این حالت فرضیات مدل (۳) را با در نظر گرفتن یک توزیع مسطح برای $\boldsymbol{\beta}$ و با فرض $\nu_i \sim N(0, \sigma_\nu^2)$ و $e_i \sim N(0, \psi_i)$ به صورت زیر در نظر می‌گیریم.

$(i)$ $\hat{\theta}_i|\theta_i, \boldsymbol{\beta}, \sigma_\nu^2 \sim N(\theta_i, \psi_i)$

$(ii)$ $\theta_i|\boldsymbol{\beta}, \sigma_\nu^2 \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_\nu^2)$

$(iii)$ $\pi(\boldsymbol{\beta}) \propto 1$

$(iv)$ $\pi(\boldsymbol{\beta}, \sigma_\nu^2) = \pi(\boldsymbol{\beta})\pi(\sigma_\nu^2) \propto \pi(\sigma_\nu^2)$

$(v)$ $\pi(\sigma_\nu^2) \propto IG(a, b)$



Archive of SID

www.SID.ir

در روابط بالا $\pi(\sigma_\nu)^{۲}$ توزیع پیشین $\sigma_\nu^{۲}$ 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 مثبتی هستند که در اغلب موارد $a = b = ۰/۰۰۱$ در نظر گرفته می‌شود. مانند قبل فرض می‌شود $\psi_i$ها معلوم هستند، حالتی که در آن $\psi_i$ها نامعلوم هستند توسط زارعی و همکاران (۱۳۸۶) بحث شده است.

برآوردگر نقطه‌ای $\theta_i$ در روش $HB$، $E(\theta_i|\hat{\boldsymbol{\theta}})$ است. برای محاسبه‌ی امید باید چگالی‌های پسین را به دست آورد. این کار اغلب به انتگرال‌های دارای بعد زیاد می‌رسد که حل آن‌ها مشکل است. بنابراین برای به دست آوردن جواب‌ها از روش شبیه‌سازی مونت کارلوی زنجیر مارکوفی مانند نمونه گیری گیبس استفاده می‌کنند (گلفند و اسمیت، ۱۹۹۰).

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

$$
\begin{aligned}
(\theta_i|\boldsymbol{\beta},\sigma_\nu^{۲},\hat{\boldsymbol{\theta}}) &\sim N(\gamma_i\hat{\theta}_i+(۱-\gamma_i)\boldsymbol{x}_i^T\boldsymbol{\beta},\gamma_i\psi_i),\\
(\boldsymbol{\beta}|\boldsymbol{\theta},\sigma_\nu^{۲},\hat{\boldsymbol{\theta}}) &\sim N_p\left(\Big(\sum_{i=۱}^{m}x_i\boldsymbol{x}_i^T\Big)^{-۱}\Big(\sum_{i=۱}^{m}x_i\theta_i\Big),\sigma_\nu^{۲}\Big(\sum_{i=۱}^{m}x_i\boldsymbol{x}_i^T\Big)^{-۱}\right),\\
(\sigma_\nu^{۲}|\boldsymbol{\beta},\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) &\sim IG\left(\frac{m}{۲}+a,b+\frac{۱}{۲}\sum_{i=۱}^{m}(\theta_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^{۲}\right).
\end{aligned}
$$

با در نظر گرفتن $\boldsymbol{\eta}=(\boldsymbol{\beta},\boldsymbol{\theta},\sigma_\nu)$، مراحل نمونه‌گیری گیبس را می‌توان به صورت زیر تشریح کرد.

- گام اول: انتخاب مقادیر اولیه $\boldsymbol{\eta}^{(۰)}$.
- گام دوم: تولید $\boldsymbol{\eta}^{(k+۱)}=(\boldsymbol{\beta}^{(k+۱)},\boldsymbol{\theta}^{(k+۱)},\sigma_\nu^{(k+۱)})$ به صورتی که برای تولید $\boldsymbol{\beta}^{(k+۱)}$ از توزیع شرطی

$$
(\beta|\boldsymbol{\theta}^{(k)},\sigma_\nu^{۲,(k)},\hat{\theta}) \sim N_p\left(\Big(\sum_{i=۱}^{m}x_i\boldsymbol{x}_i^T\Big)^{-۱}\Big(\sum_{i=۱}^{m}x_i\theta_i^{(k)}\Big),\sigma_\nu^{۲,(k)}\Big(\sum_{i=۱}^{m}x_ix_i^T\Big)^{-۱}\right),
$$

برای تولید $\boldsymbol{\theta}^{(k+۱)}$ از توزیع شرطی

$$
(\theta_i|\boldsymbol{\beta}^{(k+۱)},\sigma_\nu^{۲,(k)},\hat{\boldsymbol{\theta}}) \sim N(\gamma_i^{(k)}\hat{\theta}_i+(۱-\gamma_i^{(k)})\boldsymbol{x}_i^T\boldsymbol{\beta}^{(k+۱)},\gamma_i^{(k)}\psi_i),
$$

و برای تولید $\sigma_\nu^{۲,(k+۱)}$ از توزیع شرطی

$$
(\sigma_\nu^{۲}|\boldsymbol{\beta}^{(k+۱)},\boldsymbol{\theta}^{(k+۱)},\hat{\boldsymbol{\theta}}) \sim IG\left(\frac{m}{۲}+a,b+\frac{۱}{۲}\sum_{i=۱}^{m}(\theta_i^{(k+۱)}-\boldsymbol{x}_i^T\boldsymbol{\beta}^{(k+۱)})^{۲}\right).
$$



Archive of SID

www.SID.ir

استفاده می‌شود.

- گام سوم: تکرار گام‌های ۱ و ۲، تا زمانی که $|\boldsymbol{\eta}^{(k+1)} - \boldsymbol{\eta}^{(k)}|$ از مقدار از پیش تعیین شده $\delta$ کمتر شود.

در نهایت با فرض اینکه $\{(\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma_\nu^{2,(K)}), k = d+1, \cdots, d+D\}$ نمونه‌های تولید شده با استفاده از روش گیبس بعد از مرحله $d$ و $D$ تعداد تکرارهای نمونه‌گیری باشند، برآوردگر $HB$ به صورت زیر بیان می‌شود.

$$\hat{\theta}_i^{HB} = \frac{1}{D} \sum_{k=d+1}^{d+D} \left( \gamma_i^{(k)} \hat{\theta}_i + (1 - \gamma_i^{(k)}) \boldsymbol{x}_i^T \boldsymbol{\beta}^{(k)} \right)$$

## ۳-۲ روش ناپارامتری دو مرحله‌ای

همانطور که ملاحظه شد در دو روش بیان شده با فرض خطی بودن رابطه‌ی $\hat{\theta}_i$ و $\boldsymbol{x}_i$، برآورد ناحیه کوچک حاصل گردید. در این زیر بخش سعی برآن است تا با استفاده از یک روش ناپارامتری که برآورد حاصل از آن کمتر تحت تاثیر نوع ارتباط $\hat{\theta}_i$ و $\boldsymbol{x}_i$ قرار می‌گیرد، برای برآورد ناحیه‌های کوچک ارائه کنیم. برای این منظور ابتدا حالتی را که یک متغیر در مدل حضور دارد را بررسی می‌کنیم.
موخپادهیای و مایتی (۲۰۰۴)، یک روش ناپارامتری دو مرحله‌ای را با استفاده از هموارساز هسته‌ی ناداریا- واتسون ارائه کرده‌اند که به صورت زیر تشریح می‌شود.
در این روش رابطه‌ی (۱) به صورت

$$\theta_i = f(x_i) + \nu_i \qquad (۴)$$

که $f$ تابعی نامعلوم و هموار (دارای مشتق پیوسته از مرتبه دوم) است.
به منظور برآورد تابع $f(x_i)$، موخپادهیای و مایتی (۲۰۰۴)، برآوردگر هسته ناداریا-واتسون را پیشنهاد کرده‌اند که به صورت زیر محاسبه می‌شود.

$$\hat{f}_h(x) = \frac{\sum_{i=1}^m K_h(x - x_i)\hat{\theta}_i}{\sum_{i=1}^m K_h(x - x_i)} \qquad (۵)$$

که $K_h(.)$ یک تابع هسته با پهنای نوار $h$ است و

$$K_h(u) = \frac{1}{h} K(\frac{u}{h}).$$



Archive of SID

www.SID.ir

که $K(.)$ دارای ویژگی‌های زیر است

۱. $K(.)$ متقارن است.

۲. روی دامنه $x$ کراندار ومتقارن است.

۳. $\int_x K(u)du = ۱$

برآوردگر (۵) را می‌توان به صورت $\hat{f}_h(x) = \frac{۱}{m}\sum_{i=۱}^{m} W_{hi}(x)\hat{\theta}_i$ بازنویسی کرد که

$$W_{hi} = \frac{K_h(x-x_i)}{۱/m\sum_{i=۱}^{m} K_h(x-x_i)}.$$

بر اساس بهترین برآورد ناحیه کوچک برای میانگین $\theta_i$، با فرض معلوم بودن $\sigma_\nu^۲$، می‌توان به صورت زیر نوشت.

$$E(\theta_i|\hat{\theta}_i) = \tilde{\theta}_i = \gamma_i\hat{\theta}_i + (۱-\gamma_i)\hat{f}_h(x_i)\,.$$

که مانند قبل $\gamma_i = \frac{\sigma_\nu^۲}{\sigma_\nu^۲+\psi_i}$. حال در مرحله دوم با برآورد $\sigma_\nu^۲$ داریم:

$$\hat{\theta}_i^{Non} = \hat{\gamma}_i\hat{\theta}_i + (۱-\hat{\gamma}_i)\hat{f}_h(x_i)\,. \qquad (۶)$$

که $\hat{\gamma}_i = \frac{\hat{\sigma}_\nu^۲}{\hat{\sigma}_\nu^۲+\psi_i}$. و $\hat{\sigma}_\nu^۲$ یک برآوردگر سازگار برای $\sigma_\nu^۲$ است.
موخپادهیای و مایتی (۲۰۰۴)، تحت مدل (۴) نشان داده‌اند $\hat{f}_h(x)$ در هر نقطه‌ی پیوستگی تابع $f$، در احتمال به $f(x)$ میل می‌کند و برای $\hat{\sigma}_\nu^۲$، برآوردگر زیر را پیشنهاد کرده‌اند.

$$\hat{\sigma}_\nu^۲ = max\Big\{۰, \frac{۱}{m-۱}\sum_{i=۱}^{m} W_{hi}\Big(\hat{\theta}_i - \hat{f}_h(x_i)\Big)^۲ - \psi\Big\}. \qquad (۷)$$

که در این برآوردگر واریانس خطاهای ناحیه‌ها ثابت و برابر $\psi$ فرض شده است. برآوردگر پیشنهادی ممکن است منفی شود ولی ثابت می‌شود زمانی که $m\to\infty$، $P(\sigma_\nu^۲ < ۰)\to ۰$.
در برآوردگر(۵)، $h$، تحت عنوان پهنای نوار، پارامتر این هموارساز بوده و باید برآورد شود. در مورد انتخاب $h$، باید به این نکته توجه داشت که پهنای نوار کوچک می‌تواند به تابع رگرسیون برآوردشدهٔ ناهمواری منجر گردد. در حالی که پهنای نوار بزرگ می‌تواند سبب هموار شدن بیش از حد و حذف بعضی از تغییرات موضعی گردد. در حالت اول اریبی کاهش ولی واریانس افزایش می‌یابد و در حالت دوم اریبی افزایش ولی واریانس کاهش می‌یابد. بنابراین در انتخاب $h$ باید تعادلی بین این دو حالت برقرار گردد. برای این منظور معمولاً از معیارهایی



Archive of SID

www.SID.ir

استفاده می‌شود که سعی در برقراری این تعادل دارند، مانند معیار اعتبار سنجی متقابل $CV$[1]. در محاسبه $CV$، فرض کنید $\hat{f}^{[-i]}(x)$ و $i = ۱, ۲, ..., m$ ، مدل برازش‌شده پس از کنار گذاشتن مشاهده $i$ام باشد، معیار اعتبارسنجی متقابل به صورت زیر تعریف می‌شود.

$$CV = \frac{۱}{m}\sum_{i=۱}^{m}\left(\hat{f}^{[-i]}(x_i) - \hat{\theta}_i\right)^۲ \qquad (۸)$$

در واقع $\hat{f}^{[-i]}(x_i) - \hat{\theta}_i$ بیانگر خطای پیش‌بینی مشاهده $i$ام به عنوان مشاهده جدید حاصل از مدل برازش‌شده به کلیه مشاهدات بعد از کنار گذاشتن مشاهده $i$ام است و $CV$ میانگین توان دوم این خطا را در مجموعهٔ مشاهدات اندازه می‌گیرد.
با کمی محاسبات جبری فرمول معادل زیر برای محاسبهٔ $CV$ حاصل می‌شود

$$CV(h) = \frac{۱}{m}\sum_{i=۱}^{m}\Big(\frac{\hat{\theta}_i - \hat{f}(x_i)}{۱ - s_{ii}}\Big)^۲$$

که $s_{ii}$ها عناصر قطری ماتریس هموارساز $\boldsymbol{S}_h$ هستند. ماتریس هموارساز $\boldsymbol{S}_h$ همان ماتریس تبدیل بردار $\hat{\boldsymbol{f}} = (\hat{f}(x_۱), ..., \hat{f}(x_m))^T$ به $\hat{\boldsymbol{\theta}} = (\hat{\theta}_۱, ..., \hat{\theta}_m)'$ است به طوری که داریم، $\hat{\boldsymbol{f}} = \boldsymbol{S}_h\hat{\boldsymbol{\theta}}$. $tr(\boldsymbol{S}_h)$ تحت عنوان درجهٔ آزادی هموارساز شناخته می‌شود که معیاری برای اندازه‌گیری هموارای هموارساز است.
در حالتی که بیش از یک متغیر کمکی در مدل حضور دارد، مدل (۴) به صورت زیر بیان می‌شود.

$$\theta_i = f(\boldsymbol{x}_i) + \nu_i$$

که $\boldsymbol{x}_i = (x_{i۱}, \cdots x_{ip})^T$، و $p$ تعداد متغیرهای کمکی موجود در مدل است. هردل و مولر (۲۰۰۰)، برآوردگر زیر را برای $f(.)$ پیشنهاد کرده‌اند.

$$\hat{f}_H(\boldsymbol{x}) = \frac{\sum_{i=۱}^{m}\mathcal{K}_H(\boldsymbol{x}_i - \boldsymbol{x})\hat{\theta}_i}{\sum_{i=۱}^{m}\mathcal{K}_H(\boldsymbol{x}_i - \boldsymbol{x})} \qquad (۹)$$

و در آن $H$ به عنوان پهنای نوار و $H = diag(h_۱, \cdots, h_p)$، که $h_i$ پهنای نوار مربوط به متغیر کمکی $i$ام است، و در صورتی که $\boldsymbol{u} = (u_۱, \cdots, u_q)^T$، $\mathcal{K}_H(\boldsymbol{u})$، به صورت زیر تعریف می‌شود.

$$\mathcal{K}_H(\boldsymbol{u}) = K_{h_۱}(u_۱).\cdots.K_{h_q}(u_q)$$

[1] Cross Validation



Archive of SID

www.SID.ir

که $K_{h_i}(.)$، تابع هسته با پهنای نوار $h_i$ است.
برای انتخاب پهنای نوار $(H)$، نیز مانند قبل از معیار $CV$ استفاده شده است که برای این برآوردگرهای به صورت زیر تعریف می‌شود.

$$\begin{aligned} CV(H) &= \frac{1}{m}\sum_{i=1}^{m}\Big(\hat{\theta}_i - \hat{f}_H^{[-i]}(\boldsymbol{x}_i)\Big)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m}\Big(\hat{\theta}_i - \hat{f}_H(\boldsymbol{x}_i)\Big)^2 \left(1 - \frac{\mathcal{K}_H(0)}{\sum_{i=1}^{m}\mathcal{K}_H(\boldsymbol{x}_i - \boldsymbol{x}_j)}\right). \end{aligned}$$

برای محاسبه $\hat{\theta}_i^{Non}$ در این حالت کافی است در روابط (۶) و (۷) به جای $\hat{f}_h(.)$، $\hat{f}_H(.)$ و به جای $W_{hi}$، $W_{Hi}$ را قرار دهیم که

$$W_{Hi} = \frac{\mathcal{K}_H(\boldsymbol{x}_i - \boldsymbol{x})}{1/m\sum_{i=1}^{m}\mathcal{K}_H(\boldsymbol{x}_i - \boldsymbol{x})}.$$

در بخش بعد سه روش معرفی شده را با استفاده از شبیه‌سازی مقایسه می‌کنیم.

## ۳ شبیه‌سازی

به منظور مقایسه روش‌های معرفی شده، سه تابع میانگین به صورت زیر در نظر گرفته شده است.

$(i)$ $f_1(x_1) = 5 + 2x_1$

$(ii)$ $f_2(x_1) = 0/01 + 0/2x_1 - 0/005x_1^3$

$(iii)$ $f_3(\boldsymbol{x}) = x_2 exp(-0/5x_1)$.

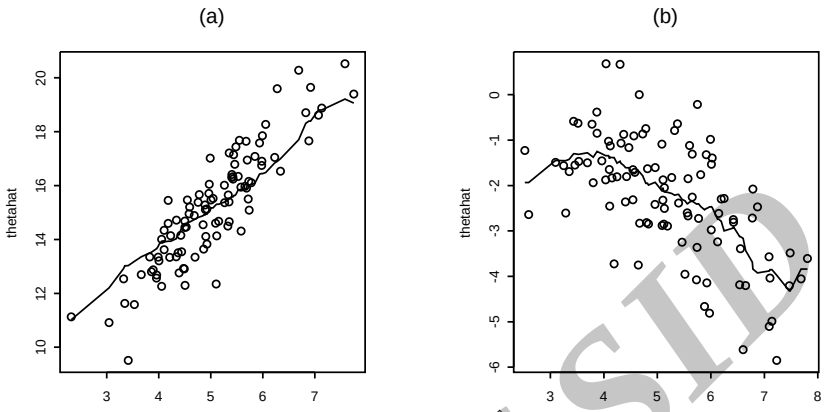
که $x_{1i}$ها و $x_{2i}$ها دارای توزیع یکنواخت $(0, 10)$، برای ۱۰۰ ناحیه کوچک $i = 1, 2, \cdots, 100$ هستند، $\nu_i \sim N(0, 0/25)$ و $\epsilon_i \sim N(0, 0/1)$. از این جوامع به تعداد $R$ بار شبیه سازی شده است. که در این شبیه‌سازی برای توابع $f_1$ و $f_2$، $R = 500$ و برای تابع $f_3$، $R = 200$ در نظر گرفته شده است.

در نمودار $(a)$ شکل (۱)، نمودار پراکنش $x$ در مقابل $\hat{\theta}$ برای یک نمونه شبیه‌سازی شده به همراه منحنی برآوردگر ناداریا-واتسون برای $h = 2$ برای تابع $f_1$ رسم شده است و در نمودار $(b)$ شکل (۱)، نیز نمودار پراکنش $x$ در مقابل $\hat{\theta}$ برای یک نمونه شبیه‌سازی شده به



Archive of SID

www.SID.ir




شکلِ ۱: $(a)$ نمودار پراکنش $x$ در مقابل $\hat{\theta}$ برای یک نمونه شبیه‌سازیِ شده به همراه منحنی برآوردگر ناداریا-واتسون برای $h = ۲$ برای تابع $m_۱$ و $(b)$ نمودار پراکنش $x$ در مقابل $\hat{\theta}$ برای یک نمونه شبیه‌سازی شده به همراه منحنی برآوردگر ناداریا-واتسون برای $h = ۰/۹$ برای تابع $m_۲$

همراه منحنی برآوردگر ناداریا-واتسون برای $h = ۰/۹$ برای تابع $f_۲$ رسم شده است، و در شکل (۲) نیز، نمودار سطح شبیه‌سازی شده برای تابع $f_۳$ و برای یک نمونه نوعی رسم شده است. از معیار

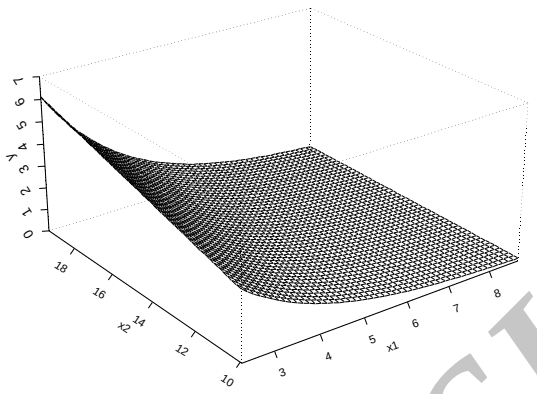
$$ASE(\hat{\boldsymbol{a}}) = \frac{۱}{m}\frac{۱}{R}\sum_{i=۱}^{m}\sum_{r=۱}^{R}(\hat{a}_{ir} - a_i)^۲$$

برای مقایسه سه روش استفاده شده است که در رابطه‌ی فوق $\hat{\boldsymbol{a}} = (\hat{a}_۱, \cdots, \hat{a}_m)^T$ بردار برآوردهای حاصل برای هریک از روش‌های معرفی شده برای $m$ ناحیه کوچک است.
در جدول (۱)، مقادیر $ASE$ حاصل از برآوردهای چهار روش مستقیم، ناپارامتری، $EBLUP$ و $HB$ برای مشاهدات شبیه‌سازی شده براساس سه تابع $f_۱$، $f_۲$ و $f_۳$ دیده می‌شود.

| روش | مستقیم | ناپارامتری | $EBLUP$ | $HB$ |
|---|---|---|---|---|
| $ASE_{f_۱}$ | ۰/۸۲۲ | ۰/۷۱۱ | ۰/۵۰۲ | ۰/۲۵۳ |
| $ASE_{f_۲}$ | ۷/۶۳۰ | ۷/۱۴ | ۷/۱۷۳ | ۶/۹۵۹ |
| $ASE_{f_۳}$ | ۴/۱۷۲ | ۳/۰۴۹ | ۳/۱۳۰ | ۲/۱۵۸ |



Archive of SID
www.SID.ir




شکلِ ۲: نمودار سطح شبیه‌سازی شده برای یک نمونه

جدول ۱ :مقادیر $ASE$، حاصل از برآوردهای چهار روش مستقیم، ناپارامتری، $EBLUP$ و $HB$ برای مشاهدات شبیه‌سازی شده براساس تابع‌های $f_1$، $f_2$ و $f_3$.

در جدول (۱) مقادیر $ASE$، حاصل از برآوردهای چهار روش مستقیم، ناپارامتری، $EBLUP$ و $HB$ برای مشاهدات شبیه‌سازی شده براساس تابع‌های $f_1$، $f_2$ و $f_3$ دیده می‌شود. همانطور که ملاحظه می‌شود با توجه به خطی بودن تابع $f_1$، برآوردگر حاصل از روش ناپارامتری دو مرحله‌ای بعد از برآوردگر حاصل از روش مستقیم دارای بزرگترین مقدار از نظر معیار $ASE$ است و برآوردگر حاصل از روش $HB$، دارای کمترین مقدار از نظر این معیار است.

با خارج شدن فرم تابع از حالت خطی ($f_2$)، برآوردگر حاصل از روش ناپارامتری بهتر شده و مقادیر $ASE$ آن از مقدار معیار $ASE$ روش $EBLUP$ کمتر شده است. ولی همچنان برآوردگر حاصل از روش $HB$ دارای کمترین مقدار $ASE$ است. در مورد تابع دومتغیره‌ی $f_3$ نیز وضع به همین ترتیب است یعنی روش ناپارامتری بهتر از روش $EBLUP$ است ولی برآوردگر حاصل از روش $HB$ دارای کمترین مقدار $ASE$ است. به طور کلی با حرکت از حالت خطی به حالت ناخطی برآوردگر حاصل از روش ناپارامتری بر برآوردگر حاصل از روش $EBLUP$ از نظر معیار $ASE$ پیشی می‌گیرد ولی در تمام حالات بررسی شده برآوردگر حاصل از روش $HB$ بهترین برآوردگر در مقایسه با سه روش مستقیم، ناپارامتری و $EBLUP$ بود.



Archive of SID

www.SID.ir

# ۴ کاربرد

همانطور که در بخش قبل ملاحظه شد، با استناد به نتایج حاصل از شبیه‌سازی روش بیز سلسله مراتبی به عنوان بهترین روش در مقایسه با سه روش مستقیم، ناپارامتری دو مرحله‌ای و $EBLUP$ شناخته شد. در این قسمت سعی بر آن است تا با استفاده از روش بیز سلسله مراتبی متوسط هزینه خانوار را در استان‌های کشور بر اساس اطلاعات جمع‌آوری شده توسط بانک مرکزی جمهوری اسلامی ایران در سال ۸۵، برآورد کنیم.

اصولاً یکی از اساسی‌ترین مطالعات آماری که به منظور نیل به اهداف مختلف اقتصادی و اجتماعی، در اغلب کشورهای جهان صورت می‌گیرد، بررسی بودجه خانوار است. از طریق این بررسی می‌توان به چگونگی هزینه‌ها و درآمدهای خانوارها و روند تغییرات آنها و نیز آمار و اطلاعات گوناگون دیگری پی برد.

بررسی هزینه و درآمد خانوار مناطق شهری ایران نیز یکی از اساسی‌ترین بررسی‌های آماری در ایران است و از آنجا که نتایج آن پایه بسیاری از محاسبات و تحقیقات اقتصادی و اجتماعی قرار می‌گیرد، حائز کمال اهمیت است.

این طرح یک طرح نمونه‌گیری است و چنانچه در برخی از استان‌ها، نمونه استفاده شده از نظر تعداد کمتر از سایر استان‌ها باشد روش بیز سلسله مراتبی بهتر از روش برآورد مستقیم عمل می‌کند. از این رو در این بخش سعی بر آن است تا با استفاده از اطلاعات کمکی موجود و روش بیز سلسله مراتبی، برآورد متوسط هزینه خانوار برای هر استان را با دقت قابل قبولی ارائه کنیم.

متغیر پاسخ مورد استفاده متوسط هزینه ناخالص خانوار است که شامل کلیه هزینه‌های خانوار به استثنای هزینه‌های شغلی و سرمایه‌گذاری است.

بعد از بررسی‌های مربوط به همبستگی و با استفاده از روش‌های رگرسیونی متغیرهای زیر به عنوان متغیرهای کمکی انتخاب شدند.

- درآمد ناخالص خانوار (شامل کلیه وجوه و ارزش پولی کالاهایی که در مقابل کار انجام شده، سرمایه به کار افتاده و یا از محل سایر منابع مانند حقوق بازنشستگی، واگذاری املاک (اجاره) و مانند آن‌ها در دوره مورد بررسی به خانوار تعلق می‌گیرد.)

- تعداد اعضای شاغل خانوار

- تعداد اعضای باسواد خانوار



Archive of SID

www.SID.ir

با بررسی روی باقیمانده‌های مدل رگرسیونی، تبدیل لگاریتمی برای متوسط هزینه و متوسط درآمد هر استان مناسب تشخیص داده شد.

روش بیز معرفی شده با فرض معلوم بودن واریانس خطای نمونه‌گیری بیان شد به این منظور از برآورد واریانس نمونه‌گیری تصادفی ساده که به صورت

$$\hat{V}ar(\bar{Y}_i) = \frac{S_i^2}{n_i}$$

استفاده شد. که $S_i^2$ واریانس نمونه‌ای ناحیه کوچک $i$ام و $n_i$ تعداد نمونه ناحیه کوچک $i$ام است. همانطور که در بخش ۲.۲ بیان شد فرض می‌شود $\sigma_\nu \sim IG(a,b)$، برای ابرپارامترهای $a$ و $b$ نیز مقدار ۰/۰۰۱ در نظر گرفته شده است و تعداد تکرار نمونه‌گیری گیبس برابر ۵۰۰۰ انتخاب شد.

با انجام محاسبات نتایج زیر برای برآورد متوسط خانوار هر استان برای سال ۸۵ با استفاده از روش مستقیم و بیز سلسله مراتبی حاصل شد.

| نام استان | روش مستقیم | روش $HB$ | نام استان | روش مستقیم | روش $HB$ |
|---|---|---|---|---|---|
| آذربایجان شرقی | ۵۷۵۲۲۹۱۴ | ۶۰۶۹۱۲۷۱ | فارس | ۸۵۳۵۷۳۷ | ۸۳۳۲۱۲۸۶ |
| آذربایجان غربی | ۵۹۶۱۷۴۱۸ | ۶۳۵۲۳۲۳۶ | قزوین | ۸۰۸۵۷۱۸۱ | ۸۲۱۴۰۲۰۰ |
| اردبیل | ۷۹۱۹۰۱۳۱ | ۷۱۷۵۹۲۶۶ | قم | ۵۳۰۸۵۶۹۶ | ۵۳۶۱۷۳۶۳ |
| اصفهان | ۷۸۳۲۳۷۷۸ | ۷۶۳۲۷۶۵۴ | کردستان | ۶۵۰۳۰۷۸۲ | ۶۲۵۸۱۴۳۴ |
| ایلام | ۸۵۱۹۵۹۰۴ | ۸۳۳۰۶۹۰۴ | کرمان | ۷۵۰۹۸۵۶۹ | ۷۳۰۹۳۹۹۶ |
| بوشهر | ۸۸۸۸۷۳۷۱ | ۸۵۲۷۹۲۲۸ | کرمانشاه | ۶۷۱۴۳۵۱۵ | ۶۶۸۴۹۳۲۶ |
| تهران | ۱۰۴۳۱۰۷۵۹ | ۱۰۳۶۹۶۷۰۱ | کهکیلویه و بویراحمد | ۹۶۲۱۵۳۵۷ | ۹۷۰۸۶۶۸۶ |
| چهارمحال وبختیاری | ۸۷۴۶۷۷۱۴ | ۸۵۹۲۶۲۶۴ | گلستان | ۶۸۰۱۷۷۴۷ | ۷۴۶۵۱۹۷۶ |
| خراسان رضوی | ۵۴۷۶۳۷۴۷ | ۵۶۱۸۳۵۱۸ | گیلان | ۷۶۱۶۶۵۶۸ | ۷۶۷۲۶۴۶۱ |
| خراسان شمالی | ۵۵۸۰۸۹۸۱ | ۵۵۶۵۲۵۷۱ | لرستان | ۶۳۷۳۶۸۵۷ | ۶۴۵۵۶۲۹۶ |
| خراسان جنوبی | ۶۵۹۸۱۷۷۸ | ۶۴۴۲۳۸۵۷ | مازندران | ۸۵۴۷۰۶۶۰ | ۸۳۷۱۰۶۷۹ |
| خوزستان | ۹۰۶۱۸۱۵۸ | ۹۲۵۷۹۶۸۰ | مرکزی | ۷۲۴۹۲۲۳۴ | ۷۰۳۶۱۷۱۵ |
| زنجان | ۶۸۶۷۲۵۶ | ۶۹۹۶۱۴۴۳ | هرمزگان | ۵۶۳۵۷۵۵ | ۶۱۱۶۴۵۸۰ |
| سمنان | ۶۹۵۹۵۹۲۲ | ۶۶۲۶۸۷۷۱ | همدان | ۷۰۳۰۷۳۴۲ | ۷۱۹۱۰۸۸۶ |
| سیستان وبلوچستان | ۵۲۵۴۳۲۳۳ | ۵۴۷۳۱۵۳۶ | یزد | ۷۸۳۳۸۱۳۱ | ۷۶۵۸۳۵۰۱ |

جدول ۲ :مقادیربرآورد متوسط هزینه خانوار استان‌ها با استفاده از روش مستقیم و روش بیز سلسله مراتبی.

در جدول (۲)، مقادیر برآورد متوسط هزینه خانوار استان‌ها با استفاده از روش مستقیم و روش بیز سلسله مراتبی دیده می‌شود، همانطور که ملاحظه می‌شود تفاوت وضعیت استان‌ها از نظر متوسط هزینه خانوار در روش مستقیم با روش بیز سلسله مراتبی کاملاً مشخص است. اگر استان‌های کشور را براساس بزرگی مقادیر برآورد روش مستقیم نسبت به مقادیر برآورد روش بیز سلسله مراتبی به سه دسته تقسیم کنیم، دسته‌ای از استان‌ها که مقدار برآورد روش مستقیم آنها بزرگتر از مقدار برآورد حاصل از روش $HB$ است، به ترتیب بزرگی اختلاف



Archive of SID

www.SID.ir

شامل استان‌های اردبیل، بوشهر، سمنان، کردستان، مرکزی، فارس، کرمان، اصفهان، ایلام، مازندران، یزد، خراسان جنوبی و چهارمحال بختیاری می‌شود. استان‌های تهران، کرمانشاه، خراسان شمالی، قم، گیلان، لرستان، کهکیلویه و بویراحمد نیز دسته‌ای از استان‌ها را شامل می‌شوند که مقدار برآورد روش مستقیم آنها نزدیک به مقدار برآورد روش $HB$ است. و در آخر دسته‌ای از استان‌ها که مقدار برآورد روش مستقیم آنها کمتر از مقدار برآورد روش $HB$ است عبارتند از:
گلستان، هرمزگان، آذربایجان غربی، آذربایجان شرقی، سیستان و بلوچستان، خوزستان، همدان، خراسان رضوی، زنجان و قزوین.

## ۵ نتیجه‌گیری

در این مقاله با استفاده از مدل‌های سطح ناحیه، سه روش، ناپارامتری دو مرحله‌ای، بهترین پیشگوی نااریب خطی تجربی و بیز سلسله مراتبی، تحت دو ساختار خطی و ناخطی را با یکدیگر مقایسه شده و با توجه به ملاک مقایسه‌ای $ASE$، بهترین روش انتخاب شد. به طور کلی با حرکت از حالت خطی به حالت ناخطی برآوردگر حاصل از روش ناپارامتری بر برآوردگر حاصل از روش $EBLUP$ از نظر معیار $ASE$ پیشی می‌گیرد ولی در تمام حالات بررسی شده برآوردگر حاصل از روش $HB$ بهترین برآوردگر در مقایسه با سه روش مستقیم، ناپارامتری و $EBLUP$ بود.

## مراجع

[1] Fay, R. E and Herriot, R. A. (1979). Estimation of income from small places: an application of James-Stein procedures to cunsus Dsata.*J. Amer. Stat. Assoc.* **74**, 269-277.



Archive of SID
www.SID.ir

[2] Gelfand, A. E. and Smith, A. F. M. (1990). Smapling based approaches to calculating marginal densities. *J. Amer. Stat. Assoc.* **8**5, 398-409.

[3] Ghosh, M. and Rao, J. N. K. (1994). Small ares estimation: an appraisal ( with discussion). *Stat. Sci.* **9**, 55-93.

[4] Härdle, W. and Müller, M. (2000). *Multivariate and Semiparametric kernel regression.* Wiley, NewYork.

[5] Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics.* **3**1, 423-447.

[6] Mukhopadhyay, P. and Maiti, T. (2004). Two stage non-parametric approach for small area estimation. *ASA Section on Survey Research Methods.*

[7] Rao, J. N. K. (2003). *Small Area Estmation.* Wiley, New York.

[۸] زارعی، ش.، گرامی، ع. و جعفری خالدی، م. (۱۳۸۶). مقایسه‌ی برآورد ناحیه‌ی کوچک متوسط درآمد خانوار در برخی از استان‌های کشور با روش بیز سلسله مراتبی. پژوهش‌های آماری ایران، ۴. ۷۱ـ۹۰.

Archive of SID



www.SID.ir