روش‌های تعدیل بی‌پاسخی و مدل‌بندی وضعیت اشتغال افراد با استفاده از معادلات برآوردگر وزنی

مجتبی گنجعلی[i]؛ حسن رنجی[ii]؛ زهرا رضایی قهرودی[iii]

چکیده

با تأکید بر استفاده از روش‌های کاهش و تعدیل بی‌پاسخی؛ که در اکثر آمارگیری‌ها روی می‌دهند، و با استفاده از معادلات برآوردگر وزنی، که راهکاری مؤثر برای تعدیل اریبی بی‌پاسخی است، به بررسی عوامل مؤثر بر وضعیت اشتغال افراد در ایران پرداخته شده است. به منظور بهبود برآوردها، از تعدیل وزن در چند مرحله؛ که از اطلاعات وزن پایه، وزن بی‌پاسخی و پیش‌بینی‌های جمعیتی بهره می‌برد، استفاده شده است. برآوردهای حاصل از معادلات برآوردگر غیروزنی؛ که از تعدیل‌های بی‌پاسخی و پیش‌بینی‌های جمعیتی چشم‌پوشی می‌کند، با برآوردهای حاصل از معادلات برآوردگر وزنی نیز مقایسه شده‌اند.

کلمات کلیدی

کاهش، تعدیل، معادلات برآوردگر وزنی، بیکاری، اریبی بی‌پاسخی.

Nonresponse Adjustment Methods and Modeling Employment Status of People Using Weighted Estimating Equations

M. Ganjali, H. Ranji and Z. Rezaee Ghahroodi

ABSTRACT

With underlining the use of reduction and adjustment methods for nonresponse, which occurs in most surveys, and with the use of weighted estimating equations, which are an effective way of adjusting nonresponse bias, influential factors on employment status of people in Iran are discussed. To improve the estimates, weighting adjustment is used in some steps, which include the use of base weight information, nonresponse weight and population predictions. The estimate from a method of non-weighted estimating equations, which ignores nonresponse adjustment and population predictions is also compared with the estimate of weighted estimating equations.

KEYWORDS

Reduction; Adjustment; Weighted estimating equations; Unemployment; Nonresponse bias.

۱- مقدمه

مسأله اساسی در آمارگیری‌ها به دست آوردن اطلاعات کامل از پاسخگویان است. امکان دستیابی به برخی از پاسخگویان ممکن است به‌علت نقص اطلاع نشانی، مسافرت یا جابه‌جایی دشوار باشد و یا ممکن است برخی پاسخگویان از مشارکت در آمارگیری سر باز زنند. علاوه بر این، برخی از

[i] دانشیار گروه آمار، دانشکده علوم ریاضی، دانشگاه شهید بهشتی: m-ganjali@sbu.ac.ir

[ii] پژوهشگر پژوهشکدهٔ آمار

[iii] دانشجوی دکتری، گروه آمار دانشکده علوم ریاضی، دانشگاه شهید بهشتی

پاسخگویانی که حاضر به مشارکت در آمارگیری شده‌اند، ممکن است یا از پاسخ دادن به برخی پرسش‌ها امتناع کنند و یا به اشتباه برخی از پرسش‌ها را بدون پاسخ بگذارند.

دو راهبرد برای مقابله با بی‌پاسخی، کاهش و تعدیل است [۶]. کاهش بی‌پاسخی به معنی به‌کارگیری راهکارهایی برای افزایش احتمال مشارکت واحدهای نمونه در آمارگیری است؛ به گونه‌ای که بی‌پاسخی واحد و بی‌پاسخی قلم را تا حد ممکن بکاهیم. تعدیل بی‌پاسخی به معنی استفاده از روش‌هایی در مرحله تحلیل، برای کمینه کردن اثرهای بی‌پاسخی بر برآوردهای آماری است. در این مقاله با ذکر روش‌های مختلف تعدیل بی‌پاسخی، روشی مناسب برای مدل‌بندی وضعیت اشتغال افراد در ایران پیشنهاد و برای طرح آمارگیری نیروی کار مرکز آمار ایران استفاده شده است.

در بخش بعد، تعاریف و مفاهیم مرتبط با بی‌پاسخی بیان شده؛ سپس اثرات و اریبی ناشی از بی‌پاسخی تشریح می‌شود. در بخش ۳، روش‌های مختلف تعدیل بی‌پاسخی به اجمال بررسی می‌شود. در بخش ۴، روش پیشنهادی برای مدل‌بندی معرفی می‌شود. در بخش ۵، به‌عنوان یک مثال عملی، داده‌های طرح نیروی کار مرکز آمار ایران؛ که در آن پاسخ مورد علاقه وضعیت اشتغال (بیکار بودن یا نبودن) فرد است، استفاده شده است. تعدیل وزن در این طرح در چند مرحله با استفاده از اطلاعات وزن پایه، وزن بی‌پاسخی و پیش‌بینی جمعیتی انجام می‌شود. برآوردهای تعدیل‌شده مربوط به مدل ارائه شده؛ که یک مدل نیمه‌پارامتری است، با استفاده از معادلات برآوردگر وزنی به دست می‌آیند. در انتها نتیجه‌گیری و چند پیشنهاد ارائه شده است.

۲- تعاریف و مفاهیم بی‌پاسخی

پژوهشگران از عبارت «بی‌پاسخی» برای عدم موفقیت در به دست آوردن اطلاع از واحد نمونه‌ای استفاده می‌کنند. گاهی اوقات این عدم موفقیت (بی‌پاسخی) به‌صورت کامل است. برای مثال، فرد نمونه پاسخ می‌دهد که «من هرگز در آمارگیری‌ها شرکت نمی‌کنم، لطفاً با من تماس نگیرید.» گاهی عدم موفقیت به‌صورت جزئی است و تنها در خصوص اندازه‌گیری یک یا چند قلم آمارگیری می‌باشد؛ برای مثال، پاسخگو در پاسخ به پرسش مصاحبه‌گر در خصوص درآمد کل خانوار، اظهار بی‌اطلاعی می‌کند. عدم موفقیت کامل «بی‌پاسخی واحد»[۱] و عدم موفقیت جزئی «بی‌پاسخی قلم»[۲] نامیده می‌شود.

بی‌پاسخی به عنوان نوعی خطای غیرنمونه‌گیری می‌تواند کیفیت آماره‌های مورد نظر در آمارگیری را تحت تأثیر قرار دهد. اگر مقادیر متغیر هدف برای پاسخگویان و غیرپاسخگویان برای محاسبهٔ آمارهٔ مورد نظر متفاوت باشند، مقدار آماره بر اساس پاسخگویان متفاوت از مقدار آن بر اساس کل نمونه خواهد بود. در این صورت، نتایج دچار اریبی بی‌پاسخی خواهند بود. در برخی آماره‌های ساده (مثل میانگین نمونه)، اریبی بی‌پاسخی تابعی از نرخ بی‌پاسخی و اختلاف میانگین بین پاسخ‌داده‌ها و پاسخ‌نداده‌هاست [۷].

گیریم $Y=(y_{ij})$ یک ماتریس داده‌های $(n\times k)$، بدون دادهٔ گم‌شده باشد که سطر iام آن را $\underline{y}_i=(y_{i1},\ldots,y_{ik})$ تشکیل می‌دهد و y_{ij} مقدار متغیر jام برای واحد iام است. برای حالتی که با **داده‌های** گم‌شده مواجه هستیم، یک ماتریس نشان‌گر گم‌شدن داده‌ها، $R=(r_{ij})$، تعریف می‌کنیم؛ به طوری که اگر y_{ij} گم‌شده باشد، $r_{ij}=0$ و اگر y_{ij} مشاهده شده باشد، $r_{ij}=1$ باشد؛ آنگاه ماتریس R الگوی گم‌شدن داده‌ها را تعریف می‌کند. این پرسش مهم است که آیا در حقیقت، گم‌شدگی متغیرها به مقدار واقعی این متغیرها در مجموعه داده‌ها وابسته است یا خیر. سازوکارهای گم‌شدن داده‌ها یک مسأله حائز اهمیت است، چرا که خصوصیات روش‌های برخورد با گم‌شدن داده‌ها به‌شدت به ماهیت این سازوکارها مربوط می‌شود. این سازوکارها تا قبل از آن‌که رابین [۱۴] سازمان‌دهی کند، تقریباً مورد اغماض قرار می‌گرفت. در روش رابین، این سازمان‌دهی از طریق در نظر گرفتن یک متغیر نشانگر برای وضعیت گم‌شدگی و در نظر گرفتن توزیعی برای آن صورت پذیرفت.

سازوکار گم‌شدن داده‌ها به‌وسیله R به شرط y؛ یعنی $f(R|y,\phi)$ مشخص می‌شود که ϕ پارامترهای نامعلوم سازوکار گم‌شدن است. اگر گم‌شدگی به مقادیر y، اعم از مشاهده شده یا گم‌شده، مربوط نباشد؛ یعنی اگر برای تمامی yها و ϕها:

$$f(R|y,\phi)=f(R|\phi) \qquad (۱)$$

گم‌شدگی داده‌ها را کاملاً تصادفی (MCAR)[۳] گویند. در صورت MCAR، استفاده از واحدهایی که برای هیچ متغیری داده گم‌شده ندارند (CC)[۴]، برآوردهایی نااریب برای پارامترها به دست می‌دهد؛ اما این برآوردها به علت کاهش حجم نمونه موجب کاهش کارایی می‌شوند.

گیریم y_{obs} نشان‌دهنده اجزا یا درایه‌هایی از y باشد که مشاهده شده‌اند و y_{mis} به‌صورت مشابه برای نمایش درایه‌های با مقادیر گم‌شده به کار رود. فرض دیگری که نسبت به MCAR از محدودیت کمتری برخوردار است، آن است که

گم‌شدگی تنها به اجزای مشاهده شده از y (y_{obs}) وابسته است و به شرط y_{obs} به مقادیر واقعی y_{mis} ارتباط ندارد؛ یعنی برای تمامی y_{mis} و ϕها:

$$f(R|y,\phi)=f(R|y_{obs},\phi) \qquad (۲)$$

این سازوکار را گم‌شدگی تصادفی (MAR)[۶] نامند. در MCAR و MAR سازوکار گم‌شدن قابل چشم‌پوشی است؛ اما در صورتی که توزیع R به مقادیر گم‌شدگی در ماتریس داده‌های y وابسته باشد، سازوکار گم‌شدگی غیر تصادفی (MNAR)[۷] و غیر قابل چشم‌پوشی نامیده می‌شود. در این مقاله، فرض می‌کنیم که گم‌شدن داده‌ها قابل چشم‌پوشی است و سعی خواهیم کرد تا از تمام داده‌های موجود (AC)[۷] یعنی داده‌هایی که در آنها واحدهای با داده گم‌شده حذف نمی‌شوند؛ بلکه تنها از داده‌های مشاهده شده در تحلیل‌های یک یا دو متغیره استفاده می‌شود، و روش‌های تعدیل برای بهبود برآوردها استفاده کنیم.

۳- روش‌های تعدیل

به دلیل بروز مشکلات ناشی از بی‌پاسخی در برآوردهای میانگین، مجموع و یا هر پارامتر جامعه، بی‌پاسخی از اهمیت قابل توجهی برخوردار است. این مشکلات عبارتند از:

۱) اریبی در برآوردگرهای نقطه‌ای

۲) افزایش واریانس برآوردگرهای نقطه‌ای

۳) اریبی در برآوردگرهای دقت

در حالت ۱) این مشکلات به نوع پارامتر نیز وابسته است، برای پارامترهای میانگین و مجموع کمتر و برای پارامترهای ارتباط همچون ضریب همبستگی و ضرایب رگرسیونی بیشترند. سازمان‌های آماری، اریبی در برآوردگرهای نقطه‌ای را بسیار مورد توجه قرار داده‌اند و از روش‌های کاهش و تعدیل بی‌پاسخی استفاده می‌کنند. به‌هر حال؛ بررسی اریبی در برآوردگرهای دقت و افزایش واریانس برآوردگرهای نقطه‌ای نیز حائز اهمیت است. برآوردهای به دست آمده از روش‌های ساده؛ مانند جانهی ساده میانگین به علت وجود داده‌های گم‌شده از دقت کمی برخوردار است و موجب کم‌برآوردی واریانس برآوردگر نقطه‌ای می‌شود. برای بررسی اریبی در برآوردگرهای نقطه‌ای، روش‌های تعدیل مختلف در ذیل آورده می‌شود.

۳-۱- تعدیل به روش‌های وزنی

جامعه متناهی U را که از N واحد با مقادیر Y_i برای $i=1,2,...,N$، تشکیل شده است در نظر گیرید. میانگین جامعه $\mu=N^{-1}\sum_{i=1}^{N}Y_i$ است. فرض کنید S مجموعهٔ اندیس‌های nتایی از واحدهای نمونه باشد که با یک طرح نمونه‌گیری به‌دست آمده است. طرح فوق ممکن است ترکیبی از طبقه‌بندی، خوشه‌ای یا انتخاب با احتمال متغیر را در بر داشته باشد. برای واحد iام جامعه، فرض کنید π_i احتمال انتخاب واحد i در نمونه S باشد. وزن احتمالی $w_i=\pi_i^{-1}$ را؛ که نشان‌دهنده آن است که واحد iام نمایندهٔ π_i^{-1} واحد جامعه است برای واحد iام در نظر گیرید. برای مثال در نمونه‌گیری تصادفی طبقه‌بندی شده یک واحد انتخاب شده در طبقه jام، N_j/n_j واحد جامعه را معرفی می‌کند.برآوردگر نقطه‌ای معمول میانگین جامعه، μ عبارت است از:

$$\hat{\mu}=\left(\sum_{i\in S}w_i\right)^{-1}\sum_{i\in S}w_iY_i \qquad (۳)$$

در طرح‌های معمول این برآوردگر، برای میانگین جامعه نااریب است و برآوردگر هورویتز-تامپسون [۹] نامیده می‌شود. در حضور بی‌پاسخی یک یا چند عضو $i\in S$، برآوردگر نمی‌تواند بر اساس (۳) محاسبه شود. یک روش ساده استفاده از برآوردگر غیر تعدیل‌شدهٔ[۸] زیر است:

$$\hat{\mu}_{UA}=\left(\sum_{i\in R}w_i\right)^{-1}\sum_{i\in R}w_iY_i \qquad (۴)$$

که در آن S به دو زیر مجموعهٔ R: واحدهای پاسخ دهنده و M: واحدهای گم‌شده تقسیم شده است.

این برآوردگر را می‌توان به فرم:

$$\hat{\mu}_{UA}=\left(\sum_{i\in S}w_ir_i\right)^{-1}\sum_{i\in S}w_ir_iY_i \qquad (۵)$$

نیز نمایش داد که در آن r_i نشانگر پاسخ واحد iام است و مقدار آن ۱ است، اگر واحد i پاسخ دهد؛ و در غیر این‌صورت برابر با صفر تعریف می‌شود. اگر بی‌پاسخی‌ها به‌طور تصادفی رخ نداده باشند، این برآوردگر اریب است.

تعدیل وزنی بی‌پاسخی، روشی برای کاهش اریبی است. فرض کنید بتوان احتمال پاسخ iامین واحد $p_i=p(r_i=1)$ را با استفاده از اطلاعات موجود آمارگیری یا یک منبع خارجی برآورد کرد. آنگاه برآوردگر:

$$\hat{\mu}_A=\left(\sum_{i\in R}w_{pi}\right)^{-1}\sum_{i\in R}w_{pi}Y_i \qquad (۶)$$

که در آن $w_{pi}=w_i/\hat{p}_i$ وزنی است که با عکس احتمال انتخاب برای واحد iام (تعدیل شده) تعیین می‌شود. در این

صورت $\hat{\mu}_A$ دارای واریانسی کمتر از $\hat{\mu}_{UA}$ است [۶].

۳-۲- تعدیل به روش‌های مدل-مبنا

در زیر، وزن‌دهی بر حسب تمایل به پاسخ و استفاده از اطلاعات کمکی را به اجمال ذکر کرده‌ایم.

۳-۲-۱ وزن‌دهی بر حسب تمایل به پاسخ

فرض کنید؛ x، مجموعه متغیرهایی را نشان دهد که برای هر دو گروه پاسخگو و غیرپاسخگو (به متغیر پاسخ) مشاهده شده‌اند. برآوردگرهای وزنی؛ حتی اگر مجموعه متغیرهای x محدود باشند، نیز در کاربردها می‌توانند به کار روند. به هر حال، در برخی حالت‌ها، مانند بررسی‌های پانلی؛ که اطلاعات از بررسی‌های قبلی برای بی‌پاسخ‌ها موجود است، رده‌بندی بر اساس متغیرهای ثبت شده غیر عملی است. این به دلیل آن است که رده‌های وزن‌دهی خیلی زیاد می‌شوند و خانه‌هایی را در بر خواهند داشت که بی‌پاسخ زیاد دارند؛ ولی پاسخ‌دهنده ندارند (بنابراین در این خانه‌ها وزن بی‌پاسخی نامتناهی است). نظریه نمرات تمایل[۹] [۱۵]، [۱۶] که در [۱۱] برای آمارگیری بحث شده است، نسخه‌ای برای کاهش بعد x به متغیر ردهٔ وزن‌دهی C ارائه می‌دهد. فرض کنید داده‌ها MAR هستند؛ یعنی:

$$pr(R|x,y)=pr(R|x) \qquad (۷)$$

بنا براین، C را می‌توان x در نظر گرفت. تمایل پاسخ فرد iام را با:

$$p(x_i)=pr(R_i=1|x_i) \qquad (۸)$$

تعریف و فرض کنید که اکیداً مثبت است $(\forall x_i)$. بنا بر این، پاسخگویان یک زیر نمونهٔ تصادفی داخل طبقاتی هستند که به وسیلهٔ نمرهٔ تمایل $[p(x)]$ تعریف می‌شوند. در عمل، متغیر تبدیل‌کنندهٔ $p(x)$ نامعلوم و نیاز به برآورد آن با استفاده از داده‌های نمونه است. یک شیوه کاربردی عبارت است از:

۱- برآورد $p(x)$، $[\hat{p}(x)]$، با استفاده از رگرسیون لوژستیک یا پروبیت R روی X بر اساس داده‌های موجود برای پاسخگویان و غیرپاسخگویان.

۲- تشکیل متغیر رده‌بندی شده با تجزیهٔ $\hat{p}(x)$ به پنج یا شش مقدار.

۳- C را متغیر رده‌بندی فرض کنید که داخل ردهٔ تعدیل‌کنندهٔ j، تمام پاسخگویان و غیرپاسخگویان مقدار نمرهٔ تمایل یکسانی دارند.

در این شیوه، پاسخگوی iام با عکس نمرهٔ تمایل برآورد شده $[\hat{p}(x)]^{-1}$ وزن‌دهی می‌شود [۲]. توجه کنید که برآورد ردهٔ وزنی حالت خاصی از این روش است که X در آن تنها یک متغیر رده‌بندی شده است و مدل لوژستیک R روی X اشباع شده است.

تحت فرض‌های مدل‌بندی $p(R|X)$ این روش، اریبی بی‌پاسخی را حذف می‌کند؛ اما ممکن است برآوردگرهایی با واریانس بسیار بالا را نتیجه دهد؛ زیرا برای پاسخگویان با تمایل پاسخ برآوردشدهٔ کم، وزن‌های بی‌پاسخی بزرگی نتیجه می‌دهند و این ممکن است روی برآوردهای میانگین و مجموع، تأثیر نامناسب بگذارد. همچنین، وزن‌دهی مستقیم به‌وسیلهٔ $[\hat{p}(x)]^{-1}$ اتکای زیادی روی مدل مشخص شده رگرسیون R روی X دارد؛ اما طبقه‌بندی بر اساس تمایل پاسخ $\hat{p}(x)$ را تنها برای تشکیل رده‌های تعدیل استفاده می‌کند.

۳-۲-۲ وزن‌دهی بر حسب تمایل به پاسخ

در این روش، تعدیل‌های وزنی بر مبنای استفاده از اطلاعات کمکی به‌دست می‌آیند. اطلاعات کمکی به‌عنوان مجموعه‌ای از متغیرها؛ که نه تنها در آمارگیری اندازه‌گیری شده‌اند؛ بلکه برای آن‌ها اطلاعات نمونه یا توزیع جامعه نیز موجود است، تعریف می‌شوند. اغلب در بی‌پاسخی واحد هیچ متغیر کمکی که توزیع نمونه برای آن موجود باشد، وجود ندارد؛ یعنی نمی‌توان از وزن‌دهی نمونه-مبنا برای آن استفاده کرد. استثنا حالتی است که چارچوب نمونه‌گیری حاوی متغیرهای کمکی همچون جنس و سن است که با استفاده از اطلاعات آمارهای ثبتی به دست می‌آیند و در این صورت می‌توان از وزن‌دهی نمونه-مبنا برای آن استفاده کرد. اگر توزیع نمونه‌ای موجود نباشد، وزن‌ها می‌توانند بر مبنای توزیع جامعه‌ای متغیرهای کمکی به‌دست آیند (جامعه-مبنا). اغلب، سازمان‌های آماری این نوع اطلاعات را برای تعداد کمی از متغیرهای کمکی در اختیار دارند.

روش‌های دیگر تعدیل عبارتند از: پس‌طبقه‌بندی [۸]، [۱۱]، [۱۲]، [۱۷]، وزن‌دهی خطی [۱]، چنگک‌زنی [۴]، وزن‌دهی ضربی [۵]، و برآورد کالیبدینی [۳]، [۴]. پس‌طبقه‌بندی به روش تحلیل داده‌ای اشاره می‌کند که شامل ساخت واحدها به گروه‌های همگن پس از مشاهدهٔ نمونه است. این اصطلاح معمولاً به مواردی محدود می‌شود که اطلاعات اضافی در نمونه برای زیرگروه‌های مورد نظر، در دسترس باشد. در حالت ایده‌آل این اطلاعات اضافی مقادیر واقعی نسبت‌ها مانند وزن‌ها در گروه‌ها می‌باشند. معمولاً ما به برآورد وزن‌ها بر اساس رکوردهای اداری و اجرایی، نتایج سرشماری‌های قبلی و یا یک آمارگیری نسبتاً بزرگ؛ که قبلاً انجام شده است، دسترسی پیدا می‌کنیم. وزن‌دهی بر اساس برآوردگرهای تولید شده از مدل‌های خطی

می‌تواند روش پس‌طبقه‌بندی را بهبود بخشند؛ که دو مشکل اساسی وجود طبقه‌های خالی و غیر کارا بودن در صورت نداشتن اطلاعات کافی جامعه را داراست.

چنگک‌زنی یک روش پس‌طبقه‌بندی است و زمانی مورد استفاده قرار می‌گیرد که پس‌طبقه‌ها با استفاده از بیش از یک متغیر تشکیل شده باشند، ولی فقط مجموع‌های حاشیه‌ای جامعه معلوم باشند.

کالیبدن، یک سیستم وزن‌دهی است که از یک تابع فاصله و مجموعه‌ای از معادله‌های کالیبدن کمک گرفته و با استفاده از متغیرهای کمکی و مقادیر معلوم این متغیرها، برآوردگرهایی تحت عنوان برآوردگرهای کالیبدنی تولید می‌کند. برآوردگرهای کالیبدنی، وزن‌های کالیبدنی را به‌کار می‌گیرند که بر اساس یک تابع فاصله، تا حد امکان به وزن‌های اصلی طرح نمونه‌گیری نزدیک باشند. این در حالی است که وزن‌های کالیبدن باید در یک مجموعه معادله‌های کالیبدن نیز صدق کنند. برای تشریح کامل این روش‌ها به [۶] و [۷] رجوع کنید.

٤- معادلات برآوردگر و برآوردگر وزنی

وقتی فرض‌های توزیعی درست هستند برآوردهای درستنمایی بیشینه (ML)[10] می‌توانند برای داده‌های گم‌شده مورد استفاده قرار گیرند. اگر چنین فرض‌هایی برقرار نباشند این برآوردگرها از مطلوبیت کمتری برخوردارند. به این دلیل، در این مقاله از معادلات برآوردگر استفاده می‌کنیم [۱۸]، [۱۳]. برآوردهای حاصل از این معادلات نسبت به برآوردهای درستنمایی بیشینه (در صورت برقراری فرض‌ها) کارایی کمتری دارند؛ ولی استوار بوده و به برآوردهای سازگار منتهی می‌شوند. این معادلات برآوردگر (EE)[11] عبارتند از:

$$\sum_{i=1}^{n} r_i d_i v_i^{-1}[y_i - \mu_i] = 0 \qquad (۹)$$

که در آن، r_i نشانگر پاسخ فرد iام است ($r_i = 1$ اگر فرد پاسخ دهد، $r_i = 0$ اگر فرد پاسخ ندهد) μ_i میانگین پاسخ فرد iام است که به‌عنوان تابعی از متغیرهای کمکی مدل‌بندی می‌شود. برای مثال، برای پاسخ دودویی احتمال موفقیت، میانگین پاسخ است و تابع ربط مورد نظر را می‌توان تابع لوجیت در نظر گرفت.

$$v_i = v_i(\boldsymbol{\beta}) = \mathrm{var}(y_i \mid \mathbf{x}_i) \qquad (۱۰)$$

برای پاسخ دودویی وضعیت اشتغال داریم:

$$v_i = \mu_i(1 - \mu_i) \qquad (۱۱)$$

و $d_i = \partial\mu_i / \partial\boldsymbol{\beta}$ که در آن‌ها $\boldsymbol{\beta}$ بردار پارامترهای نامعلوم و $\mathbf{x}_i$ بردار متغیرهای کمکی است. اگر داده‌ها به‌طور کاملاً تصادفی گم‌شده باشند، این معادلات برآورد سازگاری از $\boldsymbol{\beta}$ فراهم می‌کنند. اگر داده‌ها به‌صورت تصادفی گم شوند معادلات برآوردگر وزنی (WEE)[12]، معادلهٔ بالا را با:

$$\sum_{i=1}^{n} r_i w_i d_i v_i^{-1}[y_i - \mu_i] = 0 \qquad (۱۲)$$

تعدیل می‌کند که در آن، w_i وزن داده شده به واحد iام است.

توجه کنید که در این معادلات تنها تابع ربط و رابطهٔ بین واریانس و میانگین قید شده است و از هیچ‌گونه فرضی بر توزیع y_iها استفاده نمی‌شود. در معادلات برآوردگر واریانس مجانبی برآوردگرها بر اساس روش زیر محاسبه می‌شود. ابتدا:

$$V = \left(\sum_{i=1}^{n} r_i d_i v_i^{-1}[y_i - \mu_i]\right)^{-1} \qquad (۱۳)$$

در نظر گرفته و سپس از برآورد $\boldsymbol{\beta}$ برای محاسبهٔ برآورد آن استفاده می‌شود. این برآورد واریانس را برآورد واریانس مدل‌مبنا می‌نامند. برآورد استوار واریانس [۱۰] با استفاده از فرمول زیر تعریف می‌شود:

$$\hat{V}_1 = \hat{V}\left(\sum_{i=1}^{n} r_i d_i v_i^{-1}[y_i - \mu_i]^2 v_i^{-1} d_i\right)\hat{V} \qquad (۱۴)$$

در حالت استفاده از WEE واریانس‌ها می‌توانند از واریانس مدل‌مبنای:

$$V_w = \left(\sum_{i=1}^{n} r_i w_i d_i v_i^{-1}[y_i - \mu_i]\right)^{-1} \qquad (۱۵)$$

و یا برآورد استوار واریانس زیر محاسبه شود:

$$\hat{V}_{w1} = \hat{V}_w\left(\sum_{i=1}^{n} r_i w_i d_i v_i^{-1}[y_i - \mu_i]^2 v_i^{-1} d_i\right)\hat{V}_w. \qquad (۱۶)$$

در بخش بعد، از برآوردهای استوار واریانس استفاده خواهد شد.

۵- مدل‌بندی وضعیت اشتغال در طرح نیروی کار

در این بخش، طرح نیروی کار در مرکز آمار ایران را در نظر گرفته و روش وزن‌دهی در این طرح را به‌عنوان روشی مناسب مورد استفاده قرار می‌دهیم و وضعیت اشتغال افراد را به‌عنوان پاسخ در نظر گرفته و تأثیر برخی متغیرهای مؤثر را بر احتمال بیکار بودن با استفاده از معادلات برآوردگری که از وزن‌های مربوط استفاده می‌کنند، به دست می‌آوریم و با برآوردهای ساده که از وزن‌ها چشم‌پوشی می‌کنند، مقایسه

می‌کنیم. سیستم وزن‌دهی در این طرح؛ که در ذیل توضیح داده خواهد شد، به وزن‌هایی منجر می‌شود که برای تعدیل از احتمال انتخاب، احتمال پاسخ و پیش‌بینی‌های جمعیتی استفاده می‌کند.

۵-۱- معرفی طرح نیروی کار

طرح نیروی کار با هدف شناخت ساختار و وضعیت جاری نیروی کار و تغییرات آن اجرا می‌شود. جامعهٔ هدف این طرح، شامل همهٔ افرادی است که طبق تعریف، عضو خانوار معمولی ساکن در نقاط شهری یا روستایی کشور محسوب می‌شوند.

روش نمونه‌گیری دراین طرح، خوشه‌ای دو مرحله‌ای با طبقه‌بندی است که واحد نمونه‌گیری مرحلهٔ اول، خوشه و واحد نمونه‌گیری مرحلهٔ دوم، گروهی شامل سه خانوار معمولاً مجاور است. واحد اطلاع‌گیری، خانوار ساکن در محل سکونت خانوار نمونه است. در این طرح، علاوه بر برآورد تأثیر متغیرهای مهم، برآورد تغییرات نیز مورد نظر است؛ بنا بر این از روش نمونه‌گیری چرخشی استفاده می‌شود.

۵-۲- فرایند وزن‌دهی در طرح نیروی کار

فرایند وزن‌دهی طرح نیروی کار در چهار مرحلهٔ جداگانه انجام می‌شود. نخست، وزن پایه اعمال می‌شود که عبارت است از عکس احتمال انتخاب هر خانوار در نمونهٔ متعلق به استان، دومین مرحلهٔ وزن‌دهی برای جبران اثر خوشه‌های بی‌پاسخ انجام می‌شود. در مرحلهٔ سوم وزن‌دهی تعدیل بی‌پاسخی برای خانوار در خوشه انجام می‌شود و در آخرین مرحلهٔ وزن‌دهی، وزن‌های حاصل از مرحلهٔ سوم، برای پیش‌بینی‌های جمعیتی تعدیل می‌شوند. عامل تعدیل در مرحلهٔ دوم عبارت است از عکس نرخ خوشه‌های آمارگیری شده در استان؛ که روی وزن پایه اعمال می‌شود، و عامل تعدیل مرحلهٔ سوم عبارت است از عکس نرخ پاسخ خانوار در خوشه؛ که روی وزن حاصل از مرحلهٔ دوم اعمال می‌شود. در مرحلهٔ آخر وزن‌دهی، یک حالت خاص از برآوردگرهای کالیبدنی (برآوردگر پس‌طبقه‌بندی کامل) مورد استفاده قرار گرفته است. برای این کار با استفاده از پیش‌بینی‌های جمعیتی، جمعیت خانوارهای معمولی ساکن در ۱۲ زیرگروه حاصل از شهری/ روستایی، مرد/ زن و سه گروه سنّی «کمتر از ۱۰ سال»، «۱۰ تا ٦٤ سال» و «٦٥ سال و بیشتر» در فصل مورد نظر محاسبه می‌شود. همچنین، مجموع وزن‌های حاصل از مرحلهٔ قبل در این ۱۲ زیرگروه نیز محاسبه می‌شود. خارج قسمت جمعیت پیش‌بینی شده و مجموع وزن‌های حاصل از مرحلهٔ قبل در هر زیرگروه، عامل تعدیل وزن برای افراد متعلق به زیرگروه مورد نظر است که روی وزن حاصل از مرحلهٔ قبل اعمال می‌شود. برای جزئیات بیشتر، به طرح نمونه‌گیری نیروی کار؛ که به‌وسیلهٔ دفتر تهیهٔ طرح‌های نمونه‌گیری و چارچوب‌های آماری مرکز آمار ایران تهیه شده است، مراجعه کنید.

۵-۳- داده‌های طرح نیروی کار

از داده‌های طرح نیروی کار در زمستان ۱۳۸٤ استفاده کرده و وضعیت اشتغال افراد را به‌عنوان متغیر پاسخ مورد علاقه در نظر گرفته و از روش ناپارامتری معادلات برآوردگر برای یافتن تأثیر برخی متغیرهای مهم بر بیکاری افراد استفاده می‌کنیم. این روش می‌تواند به آسانی برای فصول مختلف سال استفاده شود. در ادامه، توضیحات کامل به‌گونه‌ای ارائه شده است که محقق علاقه‌مند بتواند به شیوه‌ای مشابه بررسی یکسانی را برای فصول دیگر انجام دهد.

پس از استخراج افراد فعال (شاغل یا بیکار) ۱۰ ساله و بیشتر در نمونه متغیرهای زیر انتخاب شدند.

۱) متغیر وضعیت اشتغال (بیکار: ۱، شاغل: ۰) به‌عنوان متغیر پاسخ. با در نظر گرفتن تنها این متغیر، نرخ بیکاری برای افراد ۱۰ ساله و بیشتر در فصل زمستان سال ۱۳۸٤، ۱۲/۱ درصد به‌دست آمد.

۲) جنسیت (۱: مرد، ۲: زن؛ که به‌ترتیب ۸۰/۰۲ و ۱۹/۹۸ درصد نمونه را شامل می‌شوند).

۳) سن (متوسط سن: ۳۵/۷).

٤) میزان تحصیلات (۱: بی‌سواد، ۲: زیر دیپلم، ۳: دیپلم یا فوق دیپلم، ٤: لیسانس، ٥: فوق لیسانس یا دکترا، ٦: سایر؛ که به ترتیب ۱٤/۹۰، ٤۸/۳۸، ۲۳/۷٤، ۷/۳۷، ۱/۰۹ و ٤/٥۱ درصد نمونه را شامل می‌شوند).

٥) تعداد اعضای خانوار (۱، ۲، ۳، ٤ یا بیش‌از ٤ نفر؛ که به ترتیب ۰/٥۹، ۷/۱٤، ۱٦/۲٤، ۲۲/۳۹ و ٥۳/٦۲ درصد نمونه را شامل می‌شوند).

٦) وضعیت تأهل (۱: دارای همسر، ۲: بی همسر بر اثر فوت، ۳: بی همسر بر اثر طلاق، ٤: هرگز ازدواج نکرده؛ که به ترتیب ٦۸/٥۷، ۱/۲٥، ۰/٤۹ و ۲۹/٦۹ درصد نمونه را شامل می‌شوند).

۷) محل اقامت (۱: شهری، ۲: روستایی؛ که به ترتیب ٦۰/۳۲ و ۳۹/٦۸ درصد نمونه را شامل می‌شوند).

متغیر سن به‌عنوان یک متغیر تبیینی پیوسته و سایر متغیرها به‌عنوان متغیرهای تبیینی رده‌بندی‌شده در مدل‌بندی وضعیت اشتغال استفاده خواهند شد.

۵-٤- نتایج استفاده از معادلات برآوردگر و معادلات برآوردگر وزنی

برآورد پارامترها و خطای استاندارد مجانبی به دو روش EE و WEE با استفاده از برنامهٔ نوشته شده در نرم‌افزار S-Plus در جدول (۱) آورده شده‌اند. در جدول (۱) (در صورت کنترل دیگر متغیرها) با استفاده از روشEE به نتایج زیر دست می‌یابیم:

جدول (۱): نتایج برآورد پارامترها و خطای استاندارد مجانبی آن‌ها با استفاده از دو روش معادلات برآوردگر (EE) و معادلات برآوردگر وزنی (WEE) (× معنی‌دار در سطح ۵ درصد)

متغیر		روش معادلات برآوردگر (EE)		روش معادلات برآوردگر وزنی (WEE)	
		برآورد	خطای استاندارد مجانبی	برآورد	خطای استاندارد مجانبی
ثابت		-۲،۳۷۹×	۰،۲۵۳	-۲،۲۸۳×	۰،۲۱۷
سن		-۰،۰۳۶×	۰،۰۰۲	-۰،۰۳۰×	۰،۰۰۲
جنسیت (مرد)		-۰،۳۰۱×	۰،۰۳۴	-۰،۳۹۸×	۰،۰۳۴
بعد خانوار	یک نفره (مبنای‌مقایسه)	-	-	-	-
	دو نفر	۰،۶۴۸×	۰،۲۳۴	۰،۴۰۲×	۰،۱۹۴
	سه نفره	۰،۵۹۱×	۰،۲۳۰	۰،۳۶۲×	۰،۱۸۹
	چهار نفره	۰،۵۶۳×	۰،۲۲۹	۰،۳۵۶×	۰،۱۸۸
	بیشتر از چهار نفر	۰،۶۷۰×	۰،۲۲۷	۰،۴۵۶×	۰،۱۸۷
وضعیت تاهل	دارای همسر (مبنای مقایسه)	-	-	-	-
	بی همسر بر اثر فوت	۰،۰۳۳	۰،۲۰۴	۰،۲۰۲	۰،۱۷۸
	بی همسر بر اثر طلاق	۰،۸۷۷×	۰،۱۷۶	۰،۶۵۵×	۰،۱۶۲
	هرگز ازدواج نکرده	۱،۰۱۱×	۰،۰۴۰	۰،۹۶۰×	۰،۰۴۰
محل سکونت (شهر)		۰،۳۸۴×	۰،۰۳۲	۰،۴۲۷×	۰،۰۳۴
میزان تحصیلات	بی‌سواد (مبنای‌مقایسه)	-	-	-	-
	زیر دیپلم	۰،۲۵۹×	۰،۰۶۹	۰،۳۳۰×	۰،۰۷۲
	دیپلم یا فوق دیپلم	۰،۷۴۲×	۰،۰۷۱	۰،۷۱۰×	۰،۰۷۴
	لیسانس	۰،۶۹۲×	۰،۰۸۰	۰،۶۰۸×	۰،۰۸۲
	فوق لیسانس یا دکترا	-۰،۳۲۳	۰،۲۰۴	-۰،۲۱۰	۰،۱۶۷
	سایر	-۰،۱۶۳	۰،۱۲۳	-۰،۳۱۱×	۰،۱۳۷

۱- هرچه سن در میان افراد فعال افزایش می‌یابد احتمال بیکاری کاهش می‌یابد

۲- احتمال بیکاری برای مردان کمتر از زنان است.

۳- احتمال بیکاری برای افراد در خانوارهای ۲، ۳، ۴ یا بیشتر از ۴ نفر بیشتر از افراد در خانوارهای تک‌نفری است.

۴- احتمال بیکاری برای متأهلین یا متأهلین قبل از فوت همسر، کمتر از بی‌همسران براثر طلاق یا مجردهاست.

۵- احتمال بیکاری برای افراد ساکن در نقاط شهری بیشتر از نقاط روستایی است.

۶- احتمال بیکاری برای افراد دارای مدرک لیسانس یا کمتر، بیشتر از افراد بی‌سواد و برای افراد دارای مدرک فوق‌لیسانس یا دکتری کمتر از افراد بیسواد است.

در جدول (۱) همین تعابیر را می‌توان با استفاده از روش WEE به‌دست آورد. به هر صورت، برآورد واریانس‌ها در روش WEE اغلب کمتر از روش EE می‌باشد. برآوردهای سطوح مختلف بعد خانوار وضعیت تأهل در WEE کاهش یافته‌اند و برآورد تأثیر جنسیت نیز در WEE کاهش بیشتری را نشان می‌دهد.

از آنجا که تعابیر نتایج وضعیت کاری افراد در ایران با وجود اثرات دوتایی کار پیچیده‌ای است، تنها به آوردن نتایج برای اثرات اصلی اکتفا کرده‌ایم. به‌هرحال، برخی از نتایج؛ که با در نظر گرفتن اثرات دوتایی می‌توانند حاصل شوند، عبارتند از:

مردان شهری احتمال بیکاری را کاهش می‌دهند؛ ولی مردان مجرد احتمال بیکاری را افزایش می‌دهند. افراد مجرد شهری نرخ بیکاری را افزایش می‌دهند. مردان باسواد دیپلم به بالا نرخ بیکاری را کاهش می‌دهند. شهری‌های دیپلم تا لیسانس نرخ بیکاری را کاهش می‌دهند. بی‌همسران بر اثر طلاق لیسانسه، نرخ بیکاری را کاهش می‌دهند. مردان با سن بالا و شهری‌های با سن بالا، نرخ بیکاری را افزایش می‌دهند، در حالی که لیسانسه‌های با سن بالا نرخ بیکاری را کاهش می‌دهد.

۶- نتیجه‌گیری و پیشنهادها

استفاده از معادلات برآوردگر وزنی با وزن‌هایی که در محاسبهٔ آن‌ها احتمالات انتخاب و پاسخ دادن دخیل هستند می‌توانند در تعدیل اریبی ناشی از بی‌پاسخی مفید باشند. این معادلات وقتی از توزیع پاسخ مطمئن نیستیم برآوردهایی سازگار از پارامترها در اختیار محقق قرار می‌دهند. در استفاده از این معادلات در وضعیت اشتغال افراد ایرانی تنها دو مشخصهٔ (۱) تابع ربط‌دهندهٔ میانگین پاسخ و بردار متغیرهای کمکی و (۲) رابطهٔ تابعی بین میانگین پاسخ و واریانس آن را مشخص کردیم.

در داده‌های طرح نیروی کار استفاده از روش WEE در مقایسه با روش EE باعث شد تا برآورد واریانس‌ها در اکثر موارد کاهش یابد. همچنین برآوردهای سطوح مختلف بعد خانوار، وضعیت تأهل و جنسیت کاهش یافتند. در استفاده از داده‌های طرح نیروی کار فرض کردیم که افراد از هم مستقل هستند. اگر داده‌ها بتوانند نوعی خوشه‌بندی (همچون افراد متعلق به خانوار یکسان) را متمایز سازند استفاده از معادلات برآوردگر تعمیم‌یافتهٔ وزنی (WGEE)[۱۳] با محاسبهٔ وزن‌هایی که همچون وزن‌های مورد استفاده در این مقاله از اطلاعات وزن پایه، وزن بی‌پاسخی و پیش‌بینی‌های جمعیتی بهره گیرد، پیشنهاد می‌شود. در این صورت، محقق می‌تواند تأثیر همبستگی افراد متعلق به خانوار یکسان را نیز مد نظر قرار دهد. استفادهٔ مدل‌های یکسانی برای داده‌های پانلی نیز پیشنهاد می‌شود.

۷- تقدیر و تشکر

از پژوهشکدهٔ آمار برای حمایت مالی این مقاله و از مرکز آمار ایران برای در اختیار قرار دادن داده‌های خام طرح نیروی کار تشکر می‌کنیم.

۸- مراجع

[۱] Bethlehem, J.G., Reduction of the nonresponse bias through regression estimation. *Journal of Official Statistics*, **4**, 251-260, 1988.

[۲] Cassel, C.M., Särndel, C.E., and Wretman, J.H., Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys, vol. III: Symposium on Incomplete Data*, Proceedings, W.G. Madow and I. Olkin, eds. Academic Press, New York, 1983.

[۳] Deville, J.C., and Särndal, C.E., Calibration estimators in survey sampling. *J. Amer. Stat. Assoc.* **87**, 376-382, 1992.

[۴] Deville, J.C., Sarndal C.E. and Sautory. O., Generalized raking procedures in survey sampling. *J. Amer. Stat. Assoc.* **88**(423), 1013-1020, 1993.

[۵] Deming, W.E. and Stephan, F.F., On a least squares adjustment of a sample frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **11** 427-444, 1940.

[۶] Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A., *Survey Nonresponse*. Wiley, New York, 2002.

[۷] Groves, R.M., Presser, S. and Dipko, S., The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, **68** (1), 2-31 2004.

[۸] Holt, D. and Smith, T.M.F., Post stratification. *Journal of the Royal Statistical Society*, Ser. A. **142**, 33-36, 1979.

[۹] Horvitz, D. G., and Thompson, D. J., A generalization of sampling without replacement from a finite population. *J. Amer. Stat. Assoc.* **47** 663-685, 1952.

[۱۰] Liang, K.Y., and Zegar, S.L. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika, 73, 13-22 1986.

[۱۱] Little, R.J.A., Survey nonresponse adjustments. *International Statistical Review*, **54**, 139-157, 1986.

[۱۲] Little, R.J.A., Post stratification: a modeler's perspective. *J. Amer. Stat. Assoc.* **88**, 1001-012 , 1993.

[۱۳] McCullagh, P., Quasi-likelihood functions. *Annals of Statistics*, 1 59-67, 1983.

[۱۴] Rubin, D.B., Inference and missing data. *Biometrika*, **63** (3), 581-92 1976.

[۱۵] Rosenbaum, P.R. and Rubin, D.B., The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70** 41-55 , 1983.

[۱۶] Rosenbaum, P.R., and Rubin, D.B., Reducing Bias in Observational Studies Using Subclassification on the Propensity Scores. *J. Amer. Stat. Assoc.* **79** 516-524, 1995.

[۱۷] Thomsen, I., A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift*, **11**, 278-283, 1973.

[۱۸] Wedderburn, R.W.M., Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 61 439-47 1974.

۹- زیرنویس‌ها

۱ Unit Nonresponse
۲ Item Nonresponse
۳ Missing Completely At Random
۴ Complete Cases
۵ Missing At Random
۶ Missing Not At Random
۷ Available Cases
۸ Unadjusted
۹ Propensity Scores
۱۰ Maximum Likelihood
۱۱ Estimating Equations
۱۲ Weighted Estimating Equations
۱۳ Weighted Generalized Estimating Equations