کاربردی از کاهش اریبی بی‌پاسخی با روش‌های امتیاز تمایل

مریم شکری‌ساز،[۱] حمیدرضا نواب‌پور[۲]

تاریخ دریافت: ۱۳۹۴/۹/۲۰
تاریخ پذیرش: ۱۳۹۶/۲/۲

چکیده:

در بسیاری از آمارگیری‌ها، برخی از واحدهای نمونه به تعدادی از پرسش‌ها یا همهٔ آنها پاسخ نمی‌دهند. این امر موجب بروز بی‌پاسخی می‌شود. اریبی و تورم واریانس، دو اثر مهم بی‌پاسخی بر آماره‌های آمارگیری هستند. اگرچه افزایش اندازهٔ نمونه از تورم واریانس براوردها جلوگیری می‌کند، لزوماً اثری بر کاهش اریبی آماره‌ها ندارد. از این رو روش‌های مختلفی برای تعدیل اریبی بی‌پاسخی مورد استفاده قرار می‌گیرد. هنگامی که ساختار گم‌شدگی تصادفی باشد، تعدیل موزون برای جبران اثر بی‌پاسخی واحد آماری مناسب است. یکی از روش‌های وزن‌دهی، روش امتیاز تمایل است. تخصیص وزن در روش امتیاز تمایل، بر مبنای براورد احتمال پاسخ واحدهای نمونه‌ای انجام می‌شود. این براوردها با برازش مدل‌های پارامتری مناسب به‌دست می‌آیند. در این مقاله روش امتیاز تمایل و براوردگرهای تعدیل‌شدهٔ حاصل از آن معرفی می‌شوند. سپس به مقایسهٔ عمل‌کرد سه براوردگر تعدیل‌شدهٔ امتیاز تمایل پرداخته می‌شود. در آخر با استفاده از مجموعه‌داده‌های آمارگیری هزینه و درامد خانوارهای شهری مرکز آمار ایران در سال ۱۳۹۰، براوردگرهای تعدیل‌شدهٔ امتیاز تمایل از نظر معیارهای ریشهٔ دوم میانگین توان دوم خطای نسبی و کارایی نسبی با هم مقایسه می‌شوند.

واژه‌های کلیدی: رگرسیون خطی فاصله‌ای، فاصله‌های فازی، شمول کل، تشخیص مدل، برنامه‌ریزی خطی.

۱ مقدمه

به‌طور کلی بی‌پاسخی نتیجهٔ تلاشی ناموفق در به‌دست آوردن پاسخ برای چند پرسش یا همهٔ پرسش‌های آمارگیری از یک واحد واجد شرایط است. خطای بی‌پاسخی بر دقت آماره‌های آمارگیری تأثیر دارد و باعث اریبی براوردها و تورم واریانس[۳] آن‌ها می‌شود. بی‌پاسخی را نمی‌توان به‌طور کامل برطرف کرد، اما می‌توان با بکارگیری روش‌های مختلف برای جبران اثرهای آن در آمارگیری‌ها اقدام کرد. روش‌های مختلفی برای تعدیل اثر بی‌پاسخی وجود دارد. این روش‌ها را می‌توان در سه مرحلهٔ طراحی آمارگیری، گردآوری داده‌ها و پس از گردآوری داده‌ها به کار گرفت. روش‌های وزن‌دهی از جمله شیوه‌هایی هستند که پس از گردآوری داده‌ها برای تعدیل اثر بی‌پاسخی مورد استفاده قرار می‌گیرند. این روش‌ها برای جبران اثر بی‌پاسخی واحد آماری به کار می‌روند. روش امتیاز تمایل[۴] یک روش وزن‌دهی است که در آن از متغیرهای کمکی برای محاسبهٔ احتمال پاسخ واحد نمونه استفاده می‌شود.

جامعهٔ متناهی Nعضوی را که با U نشان داده می‌شود، در نظر بگیرید. برای هر واحد i در جامعهٔ هدف، y_i مقدار متغیر مورد بررسی، Y، و x_i بردار $k \times 1$ از متغیرهای کمکی است. هدف، براورد میانگین، $\theta = N^{-1}\sum_{i=1}^{N} y_i$، متغیر مورد بررسی است. جامعهٔ متناهی

$$F_N = \{(y_1, x'_1), (y_2, x'_2), \ldots, (y_N, x'_N)\}$$

یک نمونهٔ تصادفی به اندازهٔ N از توزیع ابرجامعه‌ای $F(y,x)$ فرض می‌شود.

فرض می‌شود نمونهٔ احتمالاتی A به اندازهٔ n از جامعهٔ هدف U مطابق با یک طرح نمونه‌گیری احتمالاتی گزینش شده است.

[۱] دانش‌آموختهٔ کارشناس ارشد آمار اقتصادی و اجتماعی، دانشگاه علامه طباطبایی تهران
[۲] دانشیار گروه آمار دانشگاه علامه طباطبایی تهران

[۳] variance inflation

[۴] propensity score

در صورت پاسخ‌گویی همهٔ اعضای نمونه، از براوردگر هورویتز-تامپسون [۵]، $\hat{\theta}_{HT} = N^{-1}\sum_{i\in A} w_i y_i$، برای براورد میانگین جامعه‌ای استفاده می‌شود. در این براوردگر، w_i وزن طرحی است که با توجه به طرح نمونه‌گیری احتمالاتی به‌دست می‌آید. در صورتی که در نمونهٔ انتخابی واحدهای بی‌پاسخ وجود داشته باشند، براوردگر هورویتز-تامپسون برای براورد میانگین متغیر مورد بررسی مناسب نیست و باعث اریبی براورد می‌شود. در این صورت برای تعدیل اثر بی‌پاسخی می‌توان از براوردگر کامل مورد [۶] استفاده کرد. این براوردگر به‌صورت

$$\hat{\theta}_{cc} = \frac{\sum_{i\in A} w_i r_i y_i}{\sum_{i\in A} w_i r_i}$$

تعریف می‌شود، که در آن r_i متغیر نشانگر پاسخ است. در صورت پاسخ‌گویی واحد iام مقدار r_i برابر با ۱ و در غیر این صورت مقدار آن برابر با صفر است. متغیر نشانگر پاسخ بازای هر واحد جامعه مقدار می‌پذیرد. با توجه به متغیر نشانگر پاسخ r_i، احتمال این‌که یک واحد گزینش‌شده (واحد iام) پاسخ دهد، به‌صورت $p_i = Pr(r_i = 1|x_i, y_i)$ تعریف می‌شود. در واقع این احتمال، تمایل به پاسخ‌گویی واحد iام را نشان می‌دهد. براوردگر کامل مورد در احتمال به $E(Y|r = 1)$ همگرا است. در صورتی که ساختار بی‌پاسخی، کاملاً تصادفی باشد، براوردگر کامل مورد نااریب است. در غیر این صورت، این براوردگر اریب می‌باشد [۱۱]. برای تصحیح اریبی براوردگر کامل مورد از براوردگر کامل مورد موزون استفاده می‌شود. این براوردگر به‌صورت زیر و نااریب است.

$$\hat{\theta}_{wcc} = N^{-1}\sum_{i\in A}\frac{w_i r_i y_i}{p(x_i, y_i)},$$

زیرا

$$E\{N^{-1}\sum_{i\in A} w_i \frac{r_i y_i}{p(x_i, y_i)}|F_N\} = E\{N^{-1}\sum_{i=1}^{N}\frac{r_i y_i}{p(x_i, y_i)}|F_N\} = N^{-1}\sum_{i=1}^{N} y_i.$$

براوردگر کامل مورد موزون هنگامی به کار برده می‌شود که احتمال پاسخ $p(x, y)$ معلوم باشد. در صورتی که احتمال پاسخ نامعلوم باشد، باید آن را براورد کرد. این براوردها با برازش مدل‌های پارامتری مناسب از قبیل مدل رگرسیون لوژستیکی و مدل رگرسیون خطی به‌دست می‌آیند. براورد احتمال پاسخ به‌دلیل شباهت به نظریهٔ امتیاز تمایل [۱۶] برای مطالعه‌های مشاهده‌ای [۷]، امتیاز تمایل (تمایل به پاسخ‌گویی) نامیده می‌شود [۴]. با استفاده از این مفهوم براوردگرهای تعدیل‌شدهٔ امتیاز تمایل معرفی می‌شوند.

از براوردگرهای امتیاز تمایل برای کاهش اریبی بی‌پاسخی به‌طور گسترده در آمارگیری‌ها استفاده می‌شود. مثال‌هایی در خصوص کاربرد براوردگرهای امتیاز تمایل در آمارگیری‌ها را می‌توان در مطالعه‌های فولر و همکاران [۸] و ریزو و همکاران [۱۴] ملاحظه کرد. نظریهٔ امتیاز تمایل برای اولین بار توسط روزن‌بوم و روبین [۱۶] مطرح شد. آنها از نظریهٔ امتیاز تمایل در مطالعه‌های مشاهده‌ای استفاده کردند. روزن‌بوم [۱۵] استفاده از روش امتیاز تمایل را برای براورد اثرهای تیماری در مطالعه‌های مشاهده‌ای پیشنهاد کرد. لیتل [۱۳] به‌صورت گسترده‌تر به مطالعه در مورد روش امتیاز تمایل برای کنترل واحدهای بی‌پاسخ در آمارگیری‌های نمونه‌ای پرداخت. استفاده از یک مدل رگرسیونی برای براورد احتمال پاسخ را می‌توان در مطالعه‌های آلهو [۱] و ای‌خولم و لاکسونن [۶] ملاحظه کرد. فالسام [۷] و ایاناکچیون و همکاران [۹] یک مدل رگرسیون لوژستیکی را برای براورد احتمال پاسخ به کار بردند. دوآگوستینو [۳] ضرورت استفاده از امتیاز تمایل را برای کاهش اریبی در مطالعه‌های مشاهده‌ای مطرح کرد. وی امتیاز تمایل را در روش‌های مختلف تعدیل اریبی نظیر جورسازی، طبقه‌بندی و تعدیل رگرسیونی به کار برد. کاربرد این نظریه تنها به مطالعه‌های مشاهده‌ای محدود نمی‌شود. دانکن و استسنی [۵] از رهیافت امتیاز تمایل برای کنترل اریبی همپوشانی در آمارگیری‌های تلفنی استفاده کردند. لی [۱۲] روش امتیاز تمایل را برای آمارگیری وبی پانلی داوطلبانه [۸] به کار برد.

داسیلوا و اوپسومر [۴] از روش‌های ناپارامتری برای به‌دست آوردن براوردگرهای امتیاز تمایل کمک گرفتند. کیم و کیم [۱۰] ویژگی‌های مجانبی براوردگرها را در این روش بررسی کردند. آنها نشان دادند که براوردگر امتیاز تمایل با استفاده از براورد احتمال پاسخ کاراتر از براوردگری با احتمال پاسخ معلوم

[۵] Horvitz-Thompson estimator

[۶] complete-case estimator

[۷] observational studies

[۸] volunteer panel web survey

است. کیم و ریدلز [۱۱] نیز ویژگی‌های مجانبی براوردگرهای امتیاز تمایل را مورد بحث قرار دادند و بر اساس یک مدل رگرسیونی، براوردگر بهین را معرفی کردند. آنها همچنین براوردگرهای متفاوت امتیاز تمایل را معرفی کردند و به مقایسهٔ آنها و انتخاب براوردگر بهین پرداختند.

در این مقاله، در ابتدا به معرفی سه براوردگر تعدیل‌شدهٔ امتیاز تمایل می‌پردازیم. اغلب براوردگرها در روش امتیاز تمایل به‌طور مجانبی نااریب هستند. براوردگر تعدیل‌شدهٔ امتیاز تمایل بهین با استفاده از یک مدل امتیاز تمایل افزوده به‌دست می‌آید. سپس با توجه به ویژگی‌های مجانبی براوردگرهای امتیاز تمایل، کران پایین واریانس مجانبی آنها معرفی می‌شود. در نهایت، با انجام دادن یک مطالعهٔ شبیه‌سازی و با استفاده از مجموعهٔ داده‌های آمارگیری هزینه و درامد خانوارهای شهری مرکز آمار ایران در سال ۱۳۹۰، براوردگرهای مورد نظر با یک براوردگر بهین مقایسه می‌شوند. سپس اثر نرخ بی‌پاسخی و اندازهٔ نمونه بر کارایی این براوردگرها بررسی و یافته‌های مطالعه تشریح می‌شوند.

۲ روش امتیاز تمایل

اساس فرایند وزن‌دهی تمایل بر مبنای افزایش وزن‌های نمونه‌ای پاسخگویان در نمونه است. وزن‌های نمونه‌ای با استفاده از براورد احتمال تمایل پاسخگویی واحدهای نمونه به آمارگیری محاسبه می‌شوند. این احتمال، امتیاز تمایل نامیده می‌شود [۱۶]. همانطور که بیان شد، براورد احتمال پاسخ با برازش مدل‌های پارامتری از قبیل مدل رگرسیونی لوژستیکی و مدل رگرسیونی خطی و با استفاده از متغیرهای کمکی مشاهده‌شده در نمونه به‌دست می‌آید. در صورتی که مدل پارامتری تعیین‌شده برای براورد احتمال پاسخ بدرستی انتخاب نشود، براوردگرهای تعدیل‌یافتهٔ حاصل از روش امتیاز تمایل اریب خواهند شد. در ادامه شکل کلی براوردگرهای تعدیل‌شدهٔ امتیاز تمایل معرفی می‌شوند.

در عمل، احتمال پاسخ $p(x,y)$ نامعلوم است و نیاز به براورد آن با استفاده از داده‌های نمونه‌ای است. در این صورت، احتمال پاسخ را به‌صورت یک مدل پارامتری $p(x,y,\phi)$ در نظر می‌گیریم، که در آن $\phi \in \Omega$. برای براورد احتمال پاسخ، براوردگر $\sqrt{n}$ سازگار $\hat{\phi}$ برای پارامتر ϕ را در نظر بگیرید. این براوردگر در

$$\sqrt{n}(\hat{\phi}-\phi)=O_p(۱)$$

صدق می‌کند، که در آن $g_n = O_p(۱)$ نشان می‌دهد که g_n در احتمال کران‌دار است. با استفاده از این براوردگر می‌توان احتمال پاسخ برای واحد iام را به‌شکل $\hat{p}_i = p(x_i, y_i, \hat{\phi})$ براورد کرد. در این صورت براوردگر تعدیل‌شدهٔ امتیاز تمایل (PSA)[۹] عبارت است از [۱۱]:

$$\hat{\theta}_{PSA} = N^{-۱}\sum_{i\in A} w_i \frac{r_i}{\hat{p}_i} y_i. \qquad (۱)$$

تعریف نمادهای r_i و w_i همان است که در قبل بیان شد.

برای معرفی واریانس مجانبی براوردگرهای PSA، فرض می‌کنیم که ساختار پاسخ به متغیر y وابسته نیست (گم‌شدگی تصادفی). بنا بر این، برابری زیر برای هر $\phi. \in \Omega$ برقرار است:

$$p(r=۱|x,y)=p(r=۱|x)=p(x;\phi.). \qquad (۲)$$

با توجه به رابطهٔ ۲، مدل احتمال پاسخ با پارامتر $\phi.$ شناخته می‌شود.

برای براورد احتمال پاسخ باید براوردگرهای $\sqrt{n}$ سازگار برای $\phi.$ را در نظر گرفت. براوردگر $\sqrt{n}$ سازگار برای $\phi.$ در رابطهٔ ۲، با حل معادلهٔ براورد زیر بر حسب ϕ به‌دست می‌آید.

$$\hat{U}_h(\phi) \equiv \sum_{i\in A} w_i\{r_i - p_i(\phi)\}h_i(\phi) = ۰. \qquad (۳)$$

در این رابطه $p_i(\phi)=p(x_i;\phi)$ است و تابع $h_i(\phi)=h(x_i;\phi)$ یک تابع هموار از متغیر کمکی x_i و پارامتر ϕ است. با حل معادلهٔ ۳ بر حسب ϕ ، براوردگر $\hat{\phi}$ بهدست می‌آید. بنا بر این جواب رابطهٔ ۳ به انتخاب صحیح تابع $h_i(\phi)$ وابسته است.

در عبارت ۲، هر براوردگر $\sqrt{n}$ سازگار برای پارامتر $\phi.$ با حل معادلهٔ ۳ بر حسب ϕ به‌دست می‌آید. بنا بر این انتخاب مناسب تابع هموار $h_i(\phi)$ در عبارت ۳، میزان کارایی براوردگرهای PSA را تعیین می‌کند [۲].

قضیهٔ ۱.۲ برخی از ویژگی‌های مجانبی براوردگر PSA را معرفی می‌کند. با استفاده از این قضیه یک معادل مجانبی برای $\hat{\theta}_{PSA}$ به‌دست می‌آید. قبل از بیان قضیه، دو فرض زیر را در نظر می‌گیریم [۱۱]:

الف) در ساختار گم‌شدگی تصادفی، احتمال پاسخ $P(x;\phi)$ در ϕ پیوسته و دارای مشتق مرتبهٔ اول و مرتبهٔ دوم پیوسته

[۹] propensity score adjusted estimator

در یک ناحیهٔ باز شامل ϕ. است. همچنین پاسخ‌ها از هم مستقل هستند. به عبارت دیگر، $Cov(r_i, r_j|x) = ۰$ برای هر $i \neq j$.

ب) تابع براورد $\hat{U}_h(\phi)$ در احتمال به

$$U_h(\phi) = \sum_{i=۱}^{N} \{r_i - p_i(\phi)\} h_i(\phi)$$

همگرا است. همچنین مشتق جزئی $\partial\{\hat{U}_h(\phi)\}/\partial\phi$ در احتمال به $\partial\{U_h(\phi)\}/\partial\phi$ همگرا است. در حالت کلی با حل معادلهٔ $U_h(\phi) = ۰$ بر حسب ϕ، ϕ_N به‌دست می‌آید. تحت ساختار گم‌شدگی تصادفی، ϕ_N در عبارت $\sqrt{N}(\phi_N - \phi.) = O_p(۱)$ صدق می‌کند.

قضیهٔ ۱.۲. با فرض برقراری فرض‌های الف و ب و تحت توزیع توأم ساختار نمونه‌گیری و ساختار پاسخ، براوردگر تعدیل‌شدهٔ امتیاز تمایل در عبارت زیر صدق می‌کند:

$$\sqrt{n}(\hat{\theta}_{PSA} - \tilde{\theta}_{PSA,h}) = o_p(۱),$$

که در آن

$$\tilde{\theta}_{PSA,h} = \frac{۱}{N} \sum_{i \in A} w_i \{p_i h_i' \gamma_h^* + \frac{r_i}{p_i}(y_i - p_i h_i' \gamma_h^*)\}$$

و

$$\gamma_h^* = (\sum_{i=۱}^{N} r_i z_i p_i h_i')^{-۱} (\sum_{i=۱}^{N} r_i z_i y_i). \qquad (۴)$$

در رابطهٔ ۴، $p_i = p(x_i; \phi.)$، $z_i = \partial\{p^{-۱}(x_i; \phi.)\}/\partial\phi.$ و $h_i = h(x_i; \phi.)$. علاوه بر این، اگر جامعهٔ متناهی یک نمونهٔ تصادفی از مدل ابرجامعه‌ای [۱۰] باشد، واریانس مجانبی براوردگر *PSA* به‌شکل زیر معرفی می‌شود:

$$Var(\tilde{\theta}_{PSA,h}) \geq V_l \equiv V(\hat{\theta}_{HT}) + \frac{۱}{N^۲} E\Big\{ \sum_{i \in A} w_i^۲ \Big(\frac{۱}{p_i} - ۱\Big) V(Y|x_i) \Big\} \qquad (۵)$$

در صورتی برابری برقرار می‌شود که $\hat{\phi}$ در شرط زیر صدق کند:

$$\sum_{i \in A} w_i \{\frac{r_i}{p(x_i; \hat{\phi})}\} E(Y|x_i) = ۰. \qquad (۶)$$

واریانس $\tilde{\theta}_{PSA,h}$ به‌شکل زیر به‌دست می‌آید.

$$\begin{aligned} V(\tilde{\theta}_{PSA,h}) &= V(\hat{\theta}_{HT}) + \frac{۱}{N^۲} E\{\sum_{i \in A} w_i^۲ \Big(\frac{۱}{p_i} - ۱\Big)(y_i - p_i h_i' \gamma_h^*)\} \\ &= V(\hat{\theta}_{HT}) + \frac{۱}{N^۲} E[\sum_{i \in A} w_i^۲ \Big(\frac{۱}{p_i} - ۱\Big)(y_i - E(Y|x_i) \\ &+ E(Y|x_i) - p_i h_i' \gamma_h^*)\}^۲] \\ &= V(\hat{\theta}_{HT}) + \frac{۱}{N^۲} E\{\sum_{i \in A} w_i^۲ \Big(\frac{۱}{p_i} - ۱\Big) V(Y|x_i)\} \\ &+ \frac{۱}{N^۲} E[\sum_{i \in A} w_i^۲ \Big(\frac{۱}{p_i} - ۱\Big)\{E(Y|x_i) - p_i h_i' \gamma_h^*\}^۲]. \end{aligned} \qquad (۷)$$

متغیر y_i با شرطی کردن بر x_i، مستقل از $E(Y|x_i) - p_i h_i' \gamma_h^*$ است. این نکته باعث می‌شود که به برابری آخر در رابطهٔ ۷ دست یابیم. از آن‌جا که عبارت آخر در ۷ نامنفی است، نابرابری ۵ را خواهیم داشت. علاوه بر این، فرض کنید $E(Y|x_i) = p_i h_i' \alpha$، که در آن $\alpha = E(\gamma_h^*|x_i)$. بنا بر این، با توجه به

$$\begin{aligned} E(Y|x_i) - p_i h_i' \gamma_h^* &= p_i h_i' \alpha - p_i h_i' \gamma_h^* = \\ &- p_i h_i' \{\gamma_h^* - E(\gamma_h^*|x_i)\} = o_p(۱) \end{aligned}$$

عبارت آخر در ۷ قابل چشم‌پوشی است و داریم:

$$\begin{aligned} V(\tilde{\theta}_{PSA,h}) &= V(\hat{\theta}_{HT}) \\ &+ \frac{۱}{N^۲} E\Big\{ \sum_{i \in A} w_i^۲ \Big(\frac{۱}{p_i} - ۱\Big)\{E(Y|x_i) - p_i h_i' \gamma_h^*\}^۲ \Big\} \\ &= V(\hat{\theta}_{HT}) + \frac{۱}{N^۲} E\Big\{ \sum_{i \in A} w_i^۲ \Big(\frac{۱}{p_i} - ۱\Big) V(Y|x_i) \Big\}. \end{aligned}$$

همان‌گونه که قبلاً بیان شد، با برقراری برابری بالا شرط ۶ نیز برقرار می‌شود (برای اثبات به [۱۱] مراجعه شود).

قضیهٔ ۱.۲. نشان می‌دهد که $\hat{\theta}_{PSA}$ به‌طور مجانبی با $\tilde{\theta}_{PSA,h}$ معادل است و $E(\tilde{\theta}_{PSA,h}|F_N) = \theta$.

در عبارت ۵، V_l کران پایین واریانس مجانبی براوردگرهای PSA به‌شکل رابطهٔ ۱ است. هر براوردگر امتیاز تمایل که دارای واریانس مجانبی V_l باشد، براوردگر بهین است. به عبارت دیگر، براوردگر تعدیل‌شدهٔ امتیاز تمایل بهین دارای کمترین واریانس در ردهٔ براوردگرهای PSA است.

با استفاده از روش امتیاز تمایل، براوردگرهای متفاوتی برای پارامتر جامعه معرفی می‌شوند. همهٔ این براوردگرها در ردهٔ براوردگرهای PSA قرار دارند. این براوردگرها ممکن است از لحاظ مقدار اریبی و کارایی با یکدیگر متفاوت باشند. بنا

[۱۰] superpopulation model

بر این هرکدام از آنها که با یک براوردگر بهین (که لزوماً براوردگر PSA نیست) معادل باشد، کارایی بیش‌تری نسبت به سایر براوردگرهای امتیاز تمایل دارد. استفاده از براوردگر PSA بهین، اریبی بی‌پاسخی را تا حد امکان کاهش می‌دهد. در ادامه به معرفی مختصر برخی از براوردگرهای تعدیل‌شدهٔ امتیاز تمایل می‌پردازیم، که می‌توان از آنها برای کاهش اریبی بی‌پاسخی در آمارگیری‌ها استفاده کرد.

براوردگر تعدیل‌شدهٔ امتیاز تمایل ماکسیمم درست‌نمایی

شکل کلی براوردگرهای امتیاز تمایل به‌صورت رابطهٔ ۱ است. تنها تفاوت آن‌ها در روش براورد احتمال پاسخ است. یکی از این براوردگرها، براوردگر تعدیل‌شدهٔ امتیاز تمایل ماکسیمم درست‌نمایی[11] (PSA.MLE) است. این براوردگر PSA به‌شکل رابطهٔ ۱ است که در آن $\hat{p}_i = p_i(\hat{\phi})$. براوردگر $\hat{\phi}$ در مدل احتمال پاسخ، براوردگر ماکسیمم درست‌نمایی برای ϕ است، که از حل معادلهٔ براورد ۳ به‌دست می‌آید.

براوردگر تعدیل‌شدهٔ امتیاز تمایل کالیبدنی

عبارت

$$\sum_{i\in A} w_i\{\frac{r_i}{p(x_i;\hat{\phi}_h)}\}E(Y|x_i) = 0$$

راهی برای معرفی یکی دیگر از براوردگرهای امتیاز تمایل ارائه می‌دهد. به این صورت که در ابتدا مدل رگرسیونی $E(Y|x)$، به‌صورت $E(Y|x) = \beta_0 + \beta_1' x$ فرض می‌شود، که در آن β_0 عرض از مبدأ و β بردار ضریب‌های رگرسیونی و x بردار متغیرهای کمکی است. بنا بر این امتیاز تمایل با حل معادلهٔ ۸ بر حسب ϕ به‌دست می‌آید.

$$\sum_{i\in A} w_i \frac{r_i}{p(\phi)}(1, x_i) = \sum_{i\in A} w_i(1, x_i) \qquad (۸)$$

با براورد ϕ و جایگذاری این مقدار در مدل احتمال پاسخ می‌توان امتیاز تمایل را محاسبه کرد. رابطهٔ ۸ در آمارگیری‌های نمونه‌ای شرط کالیبدنی نامیده می‌شود. در واقع $\hat{p}_i$ در این نوع براوردگر PSA، در شرط کالیبدنی صدق می‌کند. بنا بر این براوردگر تعدیل‌شدهٔ امتیاز تمایل کالیبدنی[12] (PSA.CAL) برای براورد پارامتر مورد نظر به‌شکل رابطهٔ ۱ است، که در آن براورد احتمال پاسخ با استفاده از شرط کالیبدنی به‌دست می‌آید.

براوردگر تعدیل‌شدهٔ امتیاز تمایل افزوده

فرض کنید $E(Y|x) = m(x;\beta)$، که در آن $m(x;.)$ تابعی معلوم از بردار x است که با بردار پارامترهای β مشخص می‌شود. مدل فرض‌شده، مدلی ابرجامعه‌ای است. با استفاده از $m(x_i;\beta)$ در مدل پاسخ برای واحد iاُم، می‌توان یک براوردگر امتیاز تمایل معرفی کرد. فرض کنید $\hat{m}_i = m(x_i;\hat{\beta})$، $\hat{p}_i = p(\hat{\phi})$ و $\hat{h}_i = h_i(\hat{\phi})$، که در آن $\hat{\phi}$ از معادلهٔ براورد ۳ به‌دست می‌آید. در این صورت مدل احتمال پاسخ به‌شکل

$$p_i^*(\hat{\phi},\lambda) \equiv \frac{\hat{p}_i}{\hat{p}_i + (1-\hat{p}_i)exp(\lambda_0 + \lambda_1 \hat{m}_i)} \qquad (۹)$$

گسترش می‌یابد، که در آن $\lambda = (\lambda_0, \lambda_1)'$، بردار ضریب‌های نامعلوم لاگرانژ است. اگر $\lambda = (\lambda_0, \lambda_1)' = 0$ باشد، احتمال پاسخ افزوده به $\hat{p}_i$ تبدیل خواهد شد.

با استفاده از مدل پاسخ ۹ براوردگر تعدیل‌شدهٔ امتیاز تمایل افزوده[13] (PSA.AUG) به‌صورت زیر به‌دست می‌آید:

$$\hat{\theta}_{PSA} = N^{-1}\sum_{i\in A} w_i \frac{r_i}{p_i^*(\hat{\phi},\hat{\lambda})} y_i. \qquad (۱۰)$$

در براوردگر امتیاز تمایل افزوده، $\hat{\phi}$ با استفاده از روش‌های ماکسیمم درست‌نمایی محاسبه می‌شود. همچنین $\hat{\lambda}$ در رابطهٔ زیر صدق می‌کند.

$$\sum_{i\in A} w_i(\frac{r_i}{p_i^*(\hat{\phi},\hat{\lambda})})(1, \hat{m}_i) = \sum_{i\in A} w_i(1, \hat{m}_i).$$

تحت مدل پاسخ ۲، $(\hat{\phi},\hat{\lambda})$ در رابطهٔ ۱۰ در احتمال به $(\phi_0, 0)$ همگرا است. بنا بر این امتیاز تمایل از مدل پاسخ افزوده $(\hat{p}_i^*)$ به احتمال پاسخ (p_i) همگرا می‌شود. به‌دلیل این‌که $\hat{\lambda}$ در احتمال به صفر می‌گراید، انتخاب $\hat{\beta}$ در $\hat{m}_i = m(x_i;\hat{\beta})$ نقشی در نااریبی مجانبی براوردگر PSA ندارد؛ اما واریانس مجانبی با انتخاب $\hat{\beta}$های مختلف تغییر می‌کند [۱۱].

[11] maximum likelihood propensity score adjusted estimator

[12] calibration propensity score adjusted estimator

[13] augmented propensity score adjusted estimator

براوردگر بهین

برای شناسایی بهترین براوردگر امتیاز تمایل، می‌توان براوردگرهای امتیاز تمایل مختلف را بر اساس معیارهای کارایی براوردگرها مقایسه کرد. همچنین برای شناسایی این براوردگر، یک براوردگر بهین[۱۴] (OPT) معرفی می‌شود. براوردگر بهین دارای واریانس مجانبی V_l است، اما لزوماً به‌شکل براوردگرهای امتیاز تمایل نیست. با در نظر گرفتن ویژگی‌های براوردگر بهین می‌توان در بارهٔ کارایی براوردگرهای امتیاز تمایل مختلف بحث کرد. در واقع هرچه براوردگر امتیاز تمایل منطبق بر براوردگر بهین مورد نظر باشد، کارایی بیش‌تری دارد.

با در نظر گرفتن یک مدل رگرسیونی برای $E(Y|x)$، براوردگری با واریانس مجانبی V_l ساخته می‌شود. بنا بر این با استفاده از مدل $E(Y|x)=m(x;\beta)$، براوردگر بهین عبارت است از:

$$\hat{\theta}_{opt}=\frac{1}{N}\sum_{i\in A}w_i\Big\{m(x_i;\hat{\beta})+\frac{r_i}{p_i(\hat{\phi})}\{y_i-m(x_i;\hat{\beta})\}\Big\}, \quad (11)$$

که در آن $\hat{\beta}$ براوردگر $\sqrt{n}$ سازگار برای پارامتر β تحت مدل $E(Y|x)=m(x;\beta)$ است. همچنین $\hat{\phi}$ براوردگر $\sqrt{n}$ سازگار برای پارامتر $\phi.$ است، که با استفاده از معادلهٔ ۳ محاسبه می‌شود.

قضیهٔ زیر نشان می‌دهد که واریانس مجانبی براوردگر بهین در رابطهٔ ۱۱ برابر با V_l است.

قضیهٔ ۲.۲. شرط‌های قضیهٔ ۱.۲ را در نظر بگیرید. همچنین فرض کنید که $\hat{\beta}$ در عبارت $\hat{\beta}=\beta+O_p(n^{\frac{-1}{2}})$ صدق می‌کند و مدل ابرجامعه‌ای، $m(x;\beta)$، مشتق‌های جزئی پیوسته در ناحیه‌ای باز شامل β دارد. براوردگر $\hat{\theta}_{opt}$ در رابطهٔ زیر صدق می‌کند:

$$\sqrt{n}(\hat{\theta}_{opt}-\tilde{\theta}^*_{opt})=o_p(1),$$

که در آن

$$\tilde{\theta}^*_{opt}=\frac{1}{N}\sum_{i\in A}w_i\Big\{m(x_i;\beta)+\frac{r_i}{p_i}\{y_i-m(x_i;\beta)\}\Big\}.$$

واریانس براوردگر $\tilde{\theta}^*_{opt}$ برابر با V_l، حد پایین واریانس مجانبی، است (برای اثبات به [۱۱] مراجعه شود).

۳ مطالعهٔ شبیه‌سازی

در این بخش با اجرای یک مطالعهٔ شبیه‌سازی، سه براوردگر معرفی‌شده در بخش ۲ با هم مقایسه می‌شوند. معیارهای مقایسهٔ براوردگرهای تعدیل‌شدهٔ امتیاز تمایل، «ریشهٔ دوم میانگین توان دوم خطای نسبی» و «کارایی نسبی» هستند. شایان ذکر است که مجموعه‌داده‌های استفاده‌شده در این مطالعه، داده‌های حاصل از طرح آمارگیری هزینه و درامد خانوارهای شهری در سال ۱۳۹۰ می‌باشد که توسط مرکز آمار ایران اجرا شده است. این داده‌ها مربوط به ۱۸۷۰۷ خانوار شهری در کل کشور است، که ما آن را در این مطالعه به‌عنوان جامعهٔ هدف فرض می‌کنیم.

در این مطالعهٔ شبیه‌سازی، از جامعهٔ هدف $N=18707$ عضوی $K=10000$ بار نمونه‌های مستقل nتایی به‌روش نمونه‌گیری تصادفی ساده بدون جایگذاری گزینش می‌شوند. هدف این مطالعه بررسی اثر بی‌پاسخی بر عملکرد براوردگرها است. برای این منظور سه نرخ بی‌پاسخی ۵، ۱۰ و ۱۵ درصد در نظر گرفته می‌شوند. همچنین برای بررسی اثر اندازهٔ نمونه بر روی عملکرد براوردگرها، سه اندازهٔ نمونهٔ ۱۰۰، ۳۰۰ و ۵۰۰ در نظر گرفته می‌شوند. هر یک از نرخ‌های بی‌پاسخی در سه اندازهٔ نمونه اعمال می‌شود. ساختار گم‌شدگی نیز تصادفی در نظر گرفته می‌شود. بنا بر این اگر y_i هزینهٔ خالص برای خانوار iام، متغیر کمکی x_{1i} هزینهٔ غیر خوراکی برای خانوار iام، و x_{2i} درامد خانوار iام (برای هر $i=1,\dots,n$) باشند، بی‌پاسخی فقط برای متغیر y رخ می‌دهد که وابسته به متغیرهای کمکی است. برای محاسبهٔ امتیاز تمایل، مدل رگرسیون لوژستیکی زیر در نظر گرفته می‌شود:

$$p(x;\phi)=\frac{\exp(\phi_0+\phi_1x_2)}{1+\exp(\phi_0+\phi_1x_2)} \quad (12)$$

و برای دست‌یابی به براوردگرهای PSAی بهین مدل رگرسیون خطی به‌شکل

$$m(x;\beta)=\beta_0+\beta_1x_1 \quad (13)$$

در نظر گرفته می‌شود.

در براوردگر PSA.AUG، براوردگر $\hat{\phi}$ با روش ماکسیمم درست‌نمایی محاسبه می‌شود. تحت مدل ۱۳، براوردگر ماکسیمم درست‌نمایی برای $\phi=(\phi_0,\phi_1)'$ با حل معادلهٔ براورد ۳ که در آن $h_i(\phi)=(1,x_{2i})'$، محاسبه می‌شود. پارامترهای (β_0,β_1)

[۱۴] optimal estimator

[۱۵] ordinary least squares

در مدل رگرسیونی ۱۳ با استفاده از روش کمترین توان‌های دوم عادی [۱۵] و برازش مدل رگرسیونی y بر x_1 برآورد می‌شوند.

۴ معیارهای مقایسه

برآوردگرهای معرفی‌شده در این مقاله با توجه به معیارهای ریشهٔ دوم میانگین توان دوم خطای نسبی[۱۶] و کارایی نسبی [۱۷] در سه اندازهٔ نمونه و سه نرخ بی‌پاسخی اشاره‌شده مقایسه می‌شوند. این معیارها در زیر تعریف می‌شوند.

۱.۴ ریشهٔ دوم میانگین توان دوم خطای نسبی

معیار ریشهٔ دوم میانگین توان دوم خطای نسبی (RRMSE) برای برآوردگر $\hat{\theta}$ به‌صورت زیر تعریف می‌شود.

$$RRMSE = \frac{\sqrt{MSE(\hat{\theta})}}{\theta}$$

۲.۴ کارایی نسبی

معیار کارایی نسبی (eff) برای مقایسهٔ دو برآوردگر $\hat{\theta}_1$ و $\hat{\theta}_2$ به‌صورت زیر تعریف می‌شود.

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)} = \frac{MSE(\hat{\theta}_{OPT})}{MSE(\hat{\theta}_{PSA})}$$

هرچه میزان کارایی برآوردگر تعدیل‌شدهٔ امتیاز تمایل PSA نسبت به برآوردگر بهین OPT)) به مقدار ۱ نزدیک باشد، برآوردگر امتیاز تمایل، کارایی بیش‌تری دارد.

۵ یافته‌های مطالعه

با توجه به سه اندازهٔ نمونه و سه نرخ بی‌پاسخی فرض‌شده، جدول ۱، کارایی نسبی (eff) و جدول ۲، میزان ریشهٔ دوم میانگین توان دوم خطای نسبی (RRMSE) برای چهار برآوردگر مورد نظر را نشان می‌دهند. شایان ذکر است که در این جدول‌ها مقدارهای RRMSE تا چهار رقم اعشار و مقدارهای eff تا دو رقم اعشار گرد شده‌اند. در جدول‌های ۱ و ۲ نماد nr نشان‌دهندهٔ نرخ بی‌پاسخی است.

با توجه به مقدارهای کارایی نسبی در جدول ۱، در هر سه اندازهٔ نمونه و سه میزان بی‌پاسخی، برآوردگر PSA.AUG کاراتر از دو برآوردگر PSA.MLE و PSA.CAL است؛ زیرا کارایی نسبی برای این برآوردگر در همهٔ حالت‌ها تقریباً برابر با ۱ است. این بدان معنی است که برآوردگر PSA.AUG، دارای میانگین توان دوم خطاهای تقریباً برابر با MSE برآوردگر بهین می‌باشد. همچنین در جدول ۲ برآوردگر PSA.AUG، مقدار معیار RRMSE کمتری نسبت به برآوردگرهای PSA.MLE و PSA.CAL دارد. بیش‌ترین مقدار این معیار برای برآوردگر PSA.MLE است. با مقایسهٔ مقدارهای RRMSE در جدول ۲، تفاوت چندانی بین برآورد بهین و برآورد PSA.AUG مشاهده نمی‌شود. بنا بر این، برآوردگر PSA.AUG به‌طور مجانبی با برآوردگر بهین معادل است. از این رو، با وجود بی‌پاسخی در نمونه، برآوردگر امتیاز تمایل افزوده برای برآورد میانگین هزینهٔ خالص خانوارهای شهری بهتر عمل کرده است.

کمترین مقدار معیار RRMSE برای هر چهار برآوردگر مورد نظر در هر سه نرخ بی‌پاسخی برای اندازهٔ نمونهٔ ۵۰۰ به‌دست آمده است. با کاهش اندازهٔ نمونه این مقدار برای همهٔ برآوردگرها افزایش می‌یابد. هم‌چنین با افزایش نرخ بی‌پاسخی در هر سه اندازهٔ نمونه‌ای، شاهد افزایش معیار RRMSE و کاهش کارایی نسبی برآوردگرها هستیم. بیش‌ترین مقدار RRMSE برای برآوردگر PSA.MLE برای نرخ بی‌پاسخی ۱۵ درصد و اندازهٔ نمونهٔ ۱۰۰ به‌دست آمده است. کمترین مقدار این معیار نیز برای برآوردگر PSA.AUG و برآوردگر OPT در اندازهٔ نمونه‌ای ۵۰۰ با نرخ بی‌پاسخی ۵ درصد حاصل شده است. این نتیجه‌ها در شکل‌های ۱ تا ۶ قابل مشاهده‌اند.

خلاصهٔ نتایج

نتیجه‌های حاصل از مطالعهٔ شبیه‌سازی را به‌طور خلاصه می‌توان به شرح زیر برشمرد:

۱. برآوردگر PSA.CAL اندکی کاراتر از برآوردگر PSA.MLE است؛ زیرا با توجه به واریانس مجانبی برآوردگر PSA که به‌شکل

$$V(\tilde{\theta}_{PSA,h}) = V(\hat{\theta}_{HT}) + \frac{1}{N^2}E\{\sum_{i\in A} w_i^2\Big(\frac{1}{p_i} - 1\Big)V(Y|x_i)\}$$
$$+ \frac{1}{N^2}E[\sum_{i\in A} w_i^2\Big(\frac{1}{p_i} - 1\Big)\{E(Y|x_i) - p_i h_i'\gamma_h^*\}^2]$$

[۱۶] root relative mean square error

[۱۷] relative efficiency

است، عبارت آخر در واریانس مجانبی برای روش کالیبدنی کوچک‌تر است؛ زیرا پیش‌بینی برای $E(Y|x_i) = \beta_0 + \beta_1 x_{1i}$ توسط یک تابع خطی از $(1, x_{2i})$ (شرط کالیبدنی) بهتر از یک تابع خطی از $(\hat{p}_i, \hat{p}_i x_{2i})$ (معادلهٔ براورد) تقریب زده می‌شود.

۲. براوردگر PSA.AUG کاراتر از براوردگر PSA به‌شکل رابطهٔ ۱ است. این براوردگر با استفاده از مدل رگرسیونی ۱۳ که به‌درستی انتخاب شده است، ساخته می‌شود. بنا بر این براوردگر PSA.AUG به‌طور مجانبی با براوردگر بهین در عبارت ۱۱ معادل است.

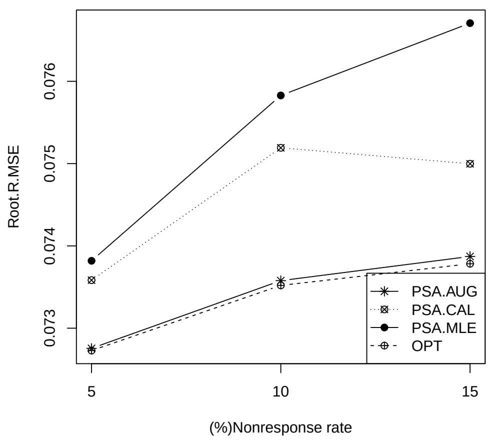
۳. با کاهش اندازهٔ نمونه و افزایش میزان بی‌پاسخی، از کارایی براوردگرها کاسته می‌شود.

جدول ۱. کارایی نسبی سه براوردگر تعدیل‌شدهٔ امتیاز تمایل نسبت به براوردگر بهین

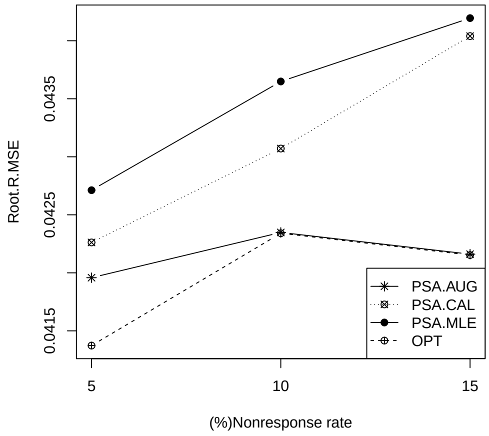
		$n = ۱۰۰$			$n = ۳۰۰$			$n = ۵۰۰$	
براوردگر	$nr = ۰/۱۵$	$nr = ۰/۱$	$nr = ۰/۰۵$	$nr = ۰/۱۵$	$nr = ۰/۱$	$nr = ۰/۰۵$	$nr = ۰/۱۵$	$nr = ۰/۱$	$nr = ۰/۰۵$
PSA.MLE	۰/۹۲	۰/۹۴	۰/۹۷	۰/۹۱	۰/۹۴	۰/۹۴	۰/۹۱	۰/۹۳	۰/۹۷
PSA.CAL	۰/۹۶	۰/۹۶	۰/۹۸	۰/۹۲	۰/۹۶	۰/۹۶	۰/۹۱	۰/۹۴	۰/۹۸
PSA.AUG	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۰/۹۷	۱/۰۰	۱/۰۰	۱/۰۰

جدول ۲. ریشهٔ دوم میانگین توان دوم خطای نسبی چهار براوردگر

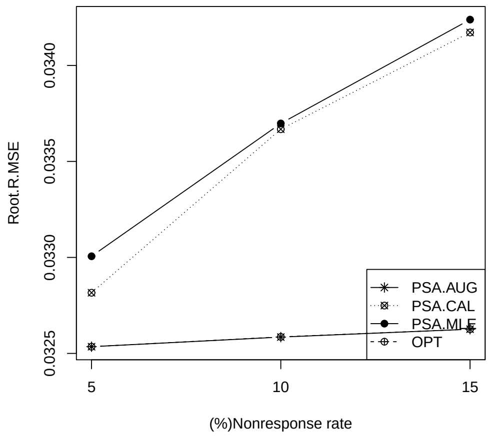
		$n = ۱۰۰$			$n = ۳۰۰$			$n = ۵۰۰$	
براوردگر	$nr = ۰/۱۵$	$nr = ۰/۱$	$nr = ۰/۰۵$	$nr = ۰/۱۵$	$nr = ۰/۱$	$nr = ۰/۰۵$	$nr = ۰/۱۵$	$nr = ۰/۱$	$nr = ۰/۰۵$
PSA.MLE	۰/۰۷۶۷	۰/۰۷۵۸	۰/۰۷۳۸	۰/۰۴۴۲	۰/۰۴۳۶	۰/۰۴۲۷	۰/۰۳۴۲	۰/۰۳۳۷	۰/۰۳۳۰
PSA.CAL	۰/۰۷۵۰	۰/۰۷۵۲	۰/۰۷۳۶	۰/۰۴۴۰	۰/۰۴۳۱	۰/۰۴۲۳	۰/۰۳۴۲	۰/۰۳۳۷	۰/۰۳۲۸
PSA.AUG	۰/۰۷۳۹	۰/۰۷۳۶	۰/۰۷۲۷	۰/۰۴۲۲	۰/۰۴۲۳	۰/۰۴۱۹	۰/۰۳۲۶	۰/۰۳۲۶	۰/۰۳۲۵
OPT	۰/۰۷۳۸	۰/۰۷۳۵	۰/۰۷۲۷	۰/۰۴۲۱	۰/۰۴۲۳	۰/۰۴۱۴	۰/۰۳۲۶	۰/۰۳۲۶	۰/۰۳۲۵



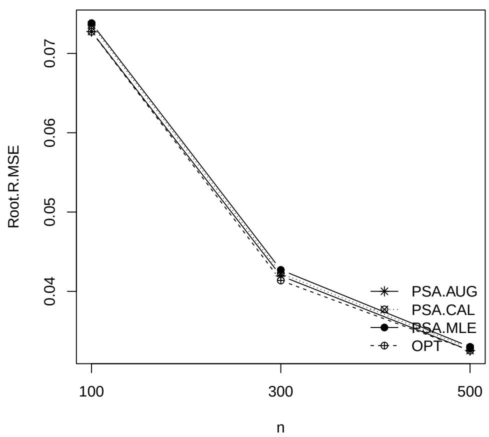
شکل ۱. اثر نرخ بی‌پاسخی بر عملکرد براوردگرها با اندازهٔ نمونهٔ ۱۰۰



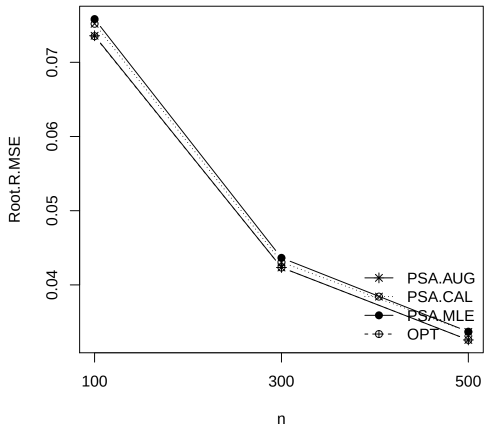
شکل۲. اثر نرخ بی‌پاسخی بر عملکرد براوردگرها با اندازهٔ نمونهٔ ۳۰۰



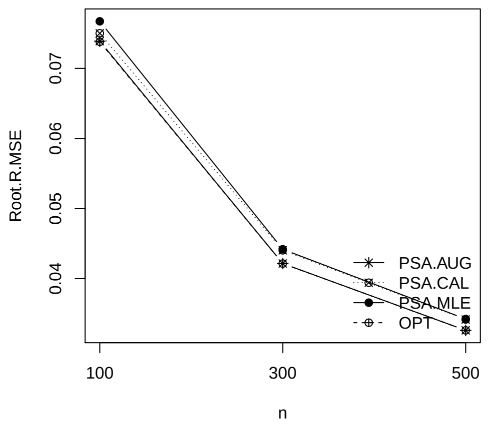
شکل۳. اثر نرخ بی‌پاسخی بر عملکرد براوردگرها با اندازهٔ نمونهٔ ۵۰۰



شکل۴. اثر اندازهٔ نمونه بر عملکرد براوردگرها با نرخ بی‌پاسخی ۵ درصد



شکل۵. اثر اندازهٔ نمونه بر عملکرد براوردگرها با نرخ بی‌پاسخی ۱۰ درصد



شکل۶. اثر اندازهٔ نمونه بر عملکرد براوردگرها با نرخ بی‌پاسخی ۱۵ درصد

مراجع

[1] Alho, J. M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika*, 617-624.

[2] Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, **34**, 305-334.

[3] D’Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparsion of a treatment to a non-randomize control group. *Statistics in Medicine*, **14**, 2265-2281.

[4] Da Silva, D. N. and Opsomer, J. D. (2009). Nonparametric propensity weighting for survey nonrespone through local polynomial regression. *Survey Methodology*, **35**, 165-176.

[5] Duncan, K. B. and Stasny, E. A. (2001). Using propensity scores to control coverage bias in telephone surveys. *Survey Methodology*, **27**, 121-130.

[6] Ekholm, A. and Laaksonen, S (1991). Weighting via response Modeling in the Finnish household budget survey. *Journal of Official Statistics*, **7**, 325-337.

[7] Folsom, R. E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. Proceedings of the Social Statistics Section, *American Statistical Association*, **7**, 197-202.

[8] Fuller, W. A., Loughin, M. M. and Baker, H. D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide food consumption survey. *Survey Methodology*, **20**, 75-85.

[9] Iannacchione, V. G., Milne, J. G. and Folsom, R. E. (1991). Response probability weight adjustments using logistic regression. Proceedings of the Survey Research Methods Section, *American Statistical Association*, **20**, 637-642.

[10] Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, **35**, 501-514.

[11] Kim, J. K. and Riddles, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, **38**, 157-165.

[12] Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, **22**, 329-349.

[13] Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, **6**, 287-296.

[14] Rizzo, L., Kalton, G. and Brick, J. M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, **22**, 43-53.

[15] Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, **82**, 387-394.

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