

کاربرد از کاهش اریبی بی پاسخی با روش های امتیاز تمایل

مریم شکری ساز،^۱ حمیدرضا نواب پور^۲

تاریخ دریافت: ۱۳۹۴/۹/۲۰

تاریخ پذیرش: ۱۳۹۶/۲/۲

چکیده:

در بسیاری از آمارگیری ها، برخی از واحدهای نمونه به تعدادی از پرسش ها یا همه آنها پاسخ نمی دهند. این امر موجب بروز بی پاسخی می شود. اریبی و تورم واریانس، دو اثر مهم بی پاسخی بر آمارهای آمارگیری هستند. اگرچه افزایش اندازه نمونه از تورم واریانس برآوردها جلوگیری می کند، لزوماً اثری بر کاهش اریبی آمارها ندارد. از این رو روش های مختلفی برای تعدیل اریبی بی پاسخی مورد استفاده قرار می گیرد. هنگامی که ساختار گم شدگی تصادفی باشد، تعدیل موزون برای جبران اثر بی پاسخی واحد آماری مناسب است. یکی از روش های وزن دهی، روش امتیاز تمایل است. تخصیص وزن در روش امتیاز تمایل، بر مبنای برآورد احتمال پاسخ واحدهای نمونه ای انجام می شود. این برآوردها با برازش مدل های پارامتری مناسب به دست می آیند. در این مقاله روش امتیاز تمایل و برآوردهای تعدیل شده حاصل از آن معرفی می شوند. سپس به مقایسه عمل کرد سه برآوردگر تعدیل شده امتیاز تمایل پرداخته می شود. در آخر با استفاده از مجموعه داده های آمارگیری هزینه و درآمد خانوارهای شهری مرکز آمار ایران در سال ۱۳۹۰، برآوردهای تعدیل شده امتیاز تمایل از نظر معیارهای ریشه دوم میانگین توان دوم خطای نسبی و کارایی نسبی با هم مقایسه می شوند.

واژه های کلیدی: رگرسیون خطی فاصله ای، فاصله های فازی، شمول کل، تشخیص مدل، برنامه ریزی خطی.

۱ مقدمه

آماري به کار می روند. روش امتیاز تمایل^۴ یک روش وزن دهی است که در آن از متغیرهای کمکی برای محاسبه احتمال پاسخ واحد نمونه استفاده می شود.

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

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

یک نمونه تصادفی به اندازه N از توزیع ابرجامعه ای $F(y, x)$ فرض می شود.

فرض می شود نمونه احتمالاتی A به اندازه n از جامعه هدف U مطابق با یک طرح نمونه گیری احتمالاتی گزینش شده است.

به طور کلی بی پاسخی نتیجه تلاشی ناموفق در به دست آوردن پاسخ برای چند پرسش یا همه پرسش های آمارگیری از یک واحد واجد شرایط است. خطای بی پاسخی بر دقت آمارهای آمارگیری تأثیر دارد و باعث اریبی برآوردها و تورم واریانس^۳ آنها می شود. بی پاسخی را نمی توان به طور کامل برطرف کرد، اما می توان با بکارگیری روش های مختلف برای جبران اثرهای آن در آمارگیری ها اقدام کرد. روش های مختلفی برای تعدیل اثر بی پاسخی وجود دارد. این روش ها را می توان در سه مرحله طراحی آمارگیری، گردآوری داده ها و پس از گردآوری داده ها به کار گرفت. روش های وزن دهی از جمله شیوه هایی هستند که پس از گردآوری داده ها برای تعدیل اثر بی پاسخی مورد استفاده قرار می گیرند. این روش ها برای جبران اثر بی پاسخی واحد

^۱ دانش آموخته کارشناس ارشد آمار اقتصادی و اجتماعی، دانشگاه علامه طباطبائی تهران

^۲ دانشیار گروه آمار دانشگاه علامه طباطبائی تهران

^۳ variance inflation

^۴ propensity score

مدل‌های پارامتری مناسب از قبیل مدل رگرسیون لوژیستیک و مدل رگرسیون خطی به‌دست می‌آیند. برآورد احتمال پاسخ به‌دلیل شباهت به نظریه امتیاز تمایل [۱۶] برای مطالعه‌های مشاهده‌ای^۷، امتیاز تمایل (تمایل به پاسخ‌گویی) نامیده می‌شود [۴]. با استفاده از این مفهوم برآوردهای تعدیل‌شده امتیاز تمایل معرفی می‌شوند.

از برآوردهای امتیاز تمایل برای کاهش اریبی بی‌پاسخی به‌طور گسترده در آمارگیری‌ها استفاده می‌شود. مثال‌هایی در خصوص کاربرد برآوردهای امتیاز تمایل در آمارگیری‌ها را می‌توان در مطالعه‌های فولر و همکاران [۸] و ریزو و همکاران [۱۴] ملاحظه کرد. نظریه امتیاز تمایل برای اولین بار توسط روزن‌بوم و روبین [۱۶] مطرح شد. آنها از نظریه امتیاز تمایل در مطالعه‌های مشاهده‌ای استفاده کردند. روزن‌بوم [۱۵] استفاده از روش امتیاز تمایل را برای برآورد 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 برآورد احتمال پاسخ را می‌توان در مطالعه‌های آلهو [۱] و ای‌خولم و لاکسونن [۶] ملاحظه کرد. فالسام [۷] و ایاناکیچون و همکاران [۹] یک مدل رگرسیون لوژیستیک را برای برآورد احتمال پاسخ به کار بردند. دو آگوستینو [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 وی‌پانی داوطلبانه^۸ به کار برد.

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

در صورت پاسخ‌گویی همه اعضای نمونه، از برآوردهای هورویتز-تامپسون^۵، $\hat{\theta}_{HT} = N^{-1} \sum_{i \in A} w_i y_i$ ، برای برآورد میانگین جامعه‌ای استفاده می‌شود. در این برآوردها، w_i وزن طرحی است که با توجه به طرح نمونه‌گیری احتمالاتی به‌دست می‌آید. در صورتی که در نمونه انتخابی واحدهای بی‌پاسخ وجود داشته باشند، برآوردهای هورویتز-تامپسون برای برآورد میانگین متغیر مورد بررسی مناسب نیست و باعث اریبی برآورد می‌شود. در این صورت برای تعدیل اثر بی‌پاسخی می‌توان از برآوردهای کامل مورد استفاده کرد. این برآوردها 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 i 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 این که یک واحد گزینش شده (واحد i) پاسخ دهد، به‌صورت $p_i = Pr(r_i = 1 | x_i, y_i)$ تعریف می‌شود. در واقع این احتمال، تمایل به پاسخ‌گویی واحد i را نشان می‌دهد. برآوردهای کامل مورد در احتمال به $E(Y | r = 1)$ همگرا است. در صورتی که ساختار بی‌پاسخی، کاملاً تصادفی باشد، برآوردهای کامل مورد ناریب است. در غیر این صورت، این برآوردها اریب می‌باشد [۱۱]. برای تصحیح اریبی برآوردهای کامل مورد از برآوردهای کامل مورد موزون استفاده می‌شود. این برآوردها به‌صورت زیر و ناریب است.

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

زیرا

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

برآوردهای 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 برآورد کرد. این برآوردها با برازش

^۵ Horvitz-Thompson estimator

^۶ complete-case estimator

^۷ observational studies

^۸ volunteer panel web survey

برآوردگر در

$$\sqrt{n}(\hat{\phi} - \phi) = O_p(1)$$

صدق می‌کند، که در آن $g_n = O_p(1)$ نشان می‌دهد که g_n در احتمال کران‌دار است. با استفاده از این برآوردگر می‌توان احتمال پاسخ برای واحد نام را به شکل $\hat{p}_i = p(x_i, y_i, \hat{\phi})$ برآورد کرد. در این صورت برآوردگر تعدیل‌شده امتیاز تمایل^۹ (PSA) عبارت است از [۱۱]:

$$\hat{\theta}_{PSA} = N^{-1} \sum_{i \in A} w_i \frac{r_i}{\hat{p}_i} y_i. \quad (1)$$

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

برای معرفی واریانس مجانبی برآوردگرهای 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:

$$p(r = 1 | x, y) = p(r = 1 | x) = p(x; \phi). \quad (2)$$

با توجه به رابطه ۲، مدل احتمال پاسخ با پارامتر ϕ شناخته می‌شود. برای برآورد احتمال پاسخ باید برآوردگرهای $\sqrt{n}$ سازگار برای ϕ را در نظر گرفت. برآوردگر $\sqrt{n}$ سازگار برای ϕ در رابطه ۲، با حل معادله برآورد زیر بر حسب ϕ به دست می‌آید.

$$\hat{U}_h(\phi) \equiv \sum_{i \in A} w_i \{r_i - p_i(\phi)\} h_i(\phi) = 0. \quad (3)$$

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 معادله ۳ بر حسب ϕ ، برآوردگر $\hat{\phi}$ به دست می‌آید. بنا بر این جواب رابطه ۳ به انتخاب صحیح تابع $h_i(\phi)$ وابسته است.

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

قضیه ۱۰۲ برخی از ویژگی‌های مجانبی برآوردگر PSA را معرفی می‌کند. با استفاده از این قضیه یک معادل مجانبی برای $\hat{\theta}_{PSA}$ به دست 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 مرتبه اول و مرتبه دوم پیوسته

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

در این مقاله، در ابتدا به معرفی سه برآوردگر تعدیل‌شده امتیاز تمایل می‌پردازیم. اغلب برآوردگرها در روش امتیاز تمایل به طور مجانبی ناریب هستند. برآوردگر تعدیل‌شده امتیاز تمایل بهین با استفاده از یک مدل امتیاز تمایل افزوده به دست می‌آید. سپس با توجه به ویژگی‌های مجانبی برآوردگرهای امتیاز تمایل، کران پایین واریانس مجانبی آنها معرفی می‌شود. در نهایت، با انجام دادن یک مطالعه شبیه‌سازی و با استفاده از مجموعه داده‌های آمارگیری هزینه و درآمد خانوارهای شهری مرکز آمار ایران در سال ۱۳۹۰، برآوردگرهای مورد نظر با یک برآوردگر بهین مقایسه می‌شوند. سپس اثر نرخ بی‌پاسخی و اندازه نمونه بر کارایی این برآوردگرها 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 برآورد احتمال تمایل پاسخگویی واحدهای نمونه به آمارگیری محاسبه می‌شوند. این احتمال، امتیاز تمایل نامیده می‌شود [۱۶]. همانطور که بیان شد، برآورد احتمال پاسخ با برازش مدل‌های پارامتری از قبیل مدل رگرسیونی لوژیستیک و مدل رگرسیونی خطی و با استفاده از متغیرهای کمکی مشاهده‌شده در نمونه به دست می‌آید. در صورتی که مدل پارامتری تعیین شده برای برآورد احتمال پاسخ بدرستی انتخاب نشود، برآوردگرهای تعدیل‌یافته حاصل از روش امتیاز تمایل اریب خواهند شد. در ادامه شکل کلی برآوردگرهای تعدیل‌شده امتیاز تمایل معرفی می‌شوند.

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

^۹ propensity score adjusted estimator

واریانس $\tilde{\theta}_{PSA,h}$ به شکل زیر به دست می‌آید.

$$\begin{aligned} V(\tilde{\theta}_{PSA,h}) &= V(\hat{\theta}_{HT}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) (y_i - p_i h'_i \gamma_h^*) \right\} \\ &= V(\hat{\theta}_{HT}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) (y_i - E(Y|x_i)) \right. \\ &\quad \left. + E(Y|x_i) - p_i h'_i \gamma_h^* \right\}^2 \\ &= V(\hat{\theta}_{HT}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) V(Y|x_i) \right\} \\ &\quad + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) \{E(Y|x_i) - p_i h'_i \gamma_h^*\}^2 \right\}. \end{aligned} \quad (۷)$$

متغیر y_i با شرطی کردن بر x_i مستقل از $p_i h'_i \gamma_h^*$ است. این نکته باعث می‌شود که به برابری آخر در رابطه 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

$$\begin{aligned} E(Y|x_i) - p_i h'_i \gamma_h^* &= p_i h'_i \alpha - p_i h'_i \gamma_h^* = \\ &= -p_i h'_i \{\gamma_h^* - E(\gamma_h^* | x_i)\} = o_p(1) \end{aligned}$$

عبارت آخر در ۷ قابل چشم‌پوشی است و داریم:

$$\begin{aligned} V(\tilde{\theta}_{PSA,h}) &= V(\hat{\theta}_{HT}) \\ &\quad + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) \{E(Y|x_i) - p_i h'_i \gamma_h^*\}^2 \right\} \\ &= V(\hat{\theta}_{HT}) + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) V(Y|x_i) \right\}. \end{aligned}$$

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

قضیه 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 برآوردهای PSA به شکل رابطه ۱ است. هر برآوردهای امتیاز تمایل که دارای واریانس مجانبی V_l باشد، برآوردهای بهین است. به عبارت دیگر، برآوردهای تعدیل‌شده امتیاز تمایل بهین دارای کمترین واریانس در رده برآوردهای PSA است.

با استفاده از روش امتیاز تمایل، برآوردهای متفاوتی برای پارامتر جامعه معرفی می‌شوند. همه این برآوردها در رده برآوردهای PSA قرار دارند. این برآوردها ممکن است از لحاظ مقدار اریبی و کارایی با یکدیگر متفاوت باشند. بنا

در یک ناحیه باز شامل ϕ است. همچنین پاسخ‌ها از هم مستقل هستند. به عبارت دیگر، $Cov(r_i, r_j | x) = 0$ برای هر $j \neq i$.

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

$$U_h(\phi) = \sum_{i=1}^N \{r_i - p_i(\phi)\} h_i(\phi)$$

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 معادله $U_h(\phi) = 0$ بر حسب ϕ ، ϕ_N به دست می‌آید. تحت ساختار گم‌شدگی تصادفی، ϕ_N در عبارت $\sqrt{N}(\phi_N - \phi_0) = O_p(1)$ صدق می‌کند.

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

$$\sqrt{n}(\hat{\theta}_{PSA} - \tilde{\theta}_{PSA,h}) = o_p(1),$$

که در آن

$$\tilde{\theta}_{PSA,h} = \frac{1}{N} \sum_{i \in A} w_i \{p_i h'_i \gamma_h^* + \frac{r_i}{p_i} (y_i - p_i h'_i \gamma_h^*)\}$$

و

$$\gamma_h^* = \left(\sum_{i=1}^N r_i z_i p_i h'_i \right)^{-1} \left(\sum_{i=1}^N r_i z_i y_i \right). \quad (۴)$$

در رابطه ۴، $p_i = p(x_i; \phi)$ ، $z_i = \partial\{p^{-1}(x_i; \phi)\}/\partial\phi$ ، اگر جامعه متناهی یک نمونه تصادفی از مدل ابرجامعه‌ای^{۱۰} باشد، واریانس مجانبی برآوردهای PSA به شکل زیر معرفی می‌شود:

$$\begin{aligned} Var(\tilde{\theta}_{PSA,h}) &\geq V_l \equiv V(\hat{\theta}_{HT}) \\ &\quad + \frac{1}{N^2} E \left\{ \sum_{i \in A} w_i^2 \left(\frac{1}{p_i} - 1 \right) V(Y|x_i) \right\} \end{aligned} \quad (۵)$$

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

$$\sum_{i \in A} w_i \left\{ \frac{r_i}{p(x_i; \hat{\phi})} \right\} E(Y|x_i) = 0. \quad (۶)$$

^{۱۰} superpopulation model

برآوردگر تعدیل شده امتیاز تمایل کالبدنی^{۱۲} (PSA.CAL) برای برآورد پارامتر مورد نظر به شکل رابطه ۱ است، که در آن برآورد احتمال پاسخ با استفاده از شرط کالبدنی به دست می آید.

برآوردگر تعدیل شده امتیاز تمایل افزوده

فرض کنید $E(Y|x) = m(x; \beta)$ ، که در آن $m(x; \cdot)$ تابعی معلوم از بردار x است که با بردار پارامترهای β مشخص می شود. مدل فرض شده، مدلی ابرجامعه ای است. با استفاده از $m(x_i; \beta)$ در مدل پاسخ برای واحد نام، می توان یک برآوردگر امتیاز تمایل معرفی کرد. فرض کنید $\hat{m}_i = m(x_i; \hat{\beta})$ و $\hat{p}_i = p(\hat{\phi})$ و $\hat{h}_i = h_i(\hat{\phi})$ که در آن $\hat{\phi}$ از معادله برآورد ۳ به دست می آید. در این صورت مدل احتمال پاسخ به شکل

$$p_i^*(\hat{\phi}, \lambda) \equiv \frac{\hat{p}_i}{\hat{p}_i + (1 - \hat{p}_i) \exp(\lambda_0 + \lambda_1 \hat{m}_i)} \quad (9)$$

گسترش می یابد، که در آن $\lambda = (\lambda_0, \lambda_1)'$ بردار ضریب های 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 برآوردگر تعدیل شده امتیاز تمایل افزوده^{۱۳} (PSA.AUG) به صورت زیر به دست می آید:

$$\hat{\theta}_{PSA} = N^{-1} \sum_{i \in A} w_i \frac{r_i}{p_i^*(\hat{\phi}, \hat{\lambda})} y_i. \quad (10)$$

در برآوردگر امتیاز تمایل افزوده، $\hat{\phi}$ با استفاده از روش های ماکسیمم درست نمایی محاسبه می شود. همچنین $\hat{\lambda}$ در رابطه زیر صدق می کند.

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

تحت مدل پاسخ ۲، $(\hat{\phi}, \hat{\lambda})$ در رابطه ۱۰ در احتمال به $(\phi, 0)$ همگرا است. بنا بر این امتیاز تمایل از مدل پاسخ افزوده $(\hat{p}_i^*)$ به احتمال پاسخ (p_i) همگرا می شود. به دلیل این که $\hat{\lambda}$ در احتمال به صفر می گراید، انتخاب $\hat{\beta}$ در $\hat{m}_i = m(x_i; \hat{\beta})$ نقشی در ناریبی مجانبی برآوردگر PSA ندارد؛ اما واریانس مجانبی با انتخاب $\hat{\beta}$ های مختلف تغییر می کند [۱۱].

بر این هر کدام از آنها که با یک برآوردگر بهین (که لزوماً برآوردگر PSA نیست) معادل باشد، کارایی بیش تری نسبت به سایر برآوردگرهای امتیاز تمایل دارد. استفاده از برآوردگر PSA بهین، اریبی بی پاسخی را تا حد امکان کاهش می دهد. در ادامه به معرفی مختصر برخی از برآوردگرهای تعدیل شده امتیاز تمایل می پردازیم، که می توان از آنها برای کاهش اریبی بی پاسخی در آمارگیری ها استفاده کرد.

برآوردگر تعدیل شده امتیاز تمایل ماکسیمم درست نمایی

شکل کلی برآوردگرهای امتیاز تمایل به صورت رابطه ۱ است. تنها تفاوت آنها در روش برآورد احتمال پاسخ است. یکی از این برآوردگرها، برآوردگر تعدیل شده امتیاز تمایل ماکسیمم درست نمایی^{۱۱} (PSA.MLE) است. این برآوردگر PSA به شکل رابطه ۱ است که در آن $\hat{p}_i = p_i(\hat{\phi})$ برآوردگر $\hat{\phi}$ در مدل احتمال پاسخ، برآوردگر ماکسیمم درست نمایی برای ϕ است، که از حل معادله برآورد ۳ به دست می آید.

برآوردگر تعدیل شده امتیاز تمایل کالبدنی

عبارت

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

راهی برای معرفی یکی دیگر از برآوردگرهای امتیاز تمایل ارائه می دهد. به این صورت که در ابتدا مدل رگرسیونی $E(Y|x)$ به صورت $E(Y|x) = \beta_0 + \beta_1 x$ فرض می شود، که در آن β_0 عرض از مبدأ و β_1 بردار ضریب های رگرسیونی و x بردار متغیرهای کمکی است. بنا بر این امتیاز تمایل با حل معادله ۸ بر حسب ϕ به دست می آید.

$$\sum_{i \in A} w_i \frac{r_i}{p(\phi)} (1, x_i) = \sum_{i \in A} w_i (1, x_i) \quad (8)$$

با برآورد ϕ و جایگذاری این مقدار در مدل احتمال پاسخ می توان امتیاز تمایل را محاسبه کرد. رابطه ۸ در آمارگیری های نمونه ای شرط کالبدنی نامیده می شود. در واقع $\hat{p}_i$ در این نوع برآوردگر PSA، در شرط کالبدنی صدق می کند. بنا بر این

^{۱۱} maximum likelihood propensity score adjusted estimator

^{۱۲} calibration propensity score adjusted estimator

^{۱۳} augmented propensity score adjusted estimator

برآورد گر بهین

۳ مطالعه شبیه سازی

در این بخش با اجرای یک مطالعه شبیه سازی، سه برآورد گر معرفی شده در بخش ۲ با هم مقایسه می شوند. معیارهای مقایسه برآورد گرهای تعدیل شده امتیاز تمایل، «ریشه دوم میانگین توان دوم خطای نسبی» و «کارایی نسبی» هستند. شایان ذکر است که مجموعه داده های استفاده شده در این مطالعه، داده های حاصل از طرح آمارگیری هزینه و درآمد خانوارهای شهری در سال ۱۳۹۰ می باشد که توسط مرکز آمار ایران اجرا شده است. این داده ها مربوط به ۱۸۷۰۷ خانوار شهری در کل کشور است، که ما آن را در این مطالعه به عنوان جامعه هدف فرض می کنیم.

در این مطالعه شبیه سازی، از جامعه هدف $N = 18707$ عضو $K = 10000$ بار نمونه های مستقل n تایی به روش نمونه گیری تصادفی ساده بدون جایگذاری گزینش می شوند. هدف این مطالعه بررسی اثر بی پاسخی بر عمل کرد برآورد گرها است. برای این منظور سه نرخ بی پاسخی ۵، ۱۰ و ۱۵ درصد در نظر گرفته می شوند. همچنین برای بررسی اثر اندازه نمونه بر روی عملکرد برآورد گرها، سه اندازه نمونه ۱۰۰، ۳۰۰ و ۵۰۰ در نظر گرفته می شوند. هر یک از نرخ های بی پاسخی در سه اندازه نمونه اعمال می شود. ساختار گم شدگی نیز تصادفی در نظر گرفته می شود. بنا بر این اگر y_i هزینه خالص برای خانوار i ام، متغیر کمکی x_{1i} هزینه غیر خوراکی برای خانوار i ام، و x_{2i} درآمد خانوار i ام (برای هر $i = 1, \dots, n$) باشند، بی پاسخی فقط برای متغیر y رخ می دهد که وابسته به متغیرهای کمکی است. برای محاسبه امتیاز تمایل، مدل رگرسیون لوژیستیک زیر در نظر گرفته می شود:

$$p(x; \phi) = \frac{\exp(\phi_0 + \phi_1 x_1)}{1 + \exp(\phi_0 + \phi_1 x_1)} \quad (12)$$

و برای دستیابی به برآورد گرهای PSA ی بهین مدل رگرسیون خطی به شکل

$$m(x; \beta) = \beta_0 + \beta_1 x_1 \quad (13)$$

در نظر گرفته می شود.

در برآورد گر PSA.AUG، برآورد گر $\hat{\phi}$ با روش ماکسیم درست نمایی محاسبه می شود. تحت مدل ۱۳، برآورد گر ماکسیم درست نمایی برای $\phi = (\phi_0, \phi_1)$ با حل معادله برآورد ۳ که در آن $h_i(\phi) = (1, x_{2i})'$ ، محاسبه می شود. پارامترهای (β_0, β_1)

برای شناسایی بهترین برآورد گر امتیاز تمایل، می توان برآورد گرهای امتیاز تمایل مختلف را بر اساس معیارهای کارایی برآورد گرها مقایسه کرد. همچنین برای شناسایی این برآورد گر، یک برآورد گر بهین^{۱۴} (OPT) معرفی می شود. برآورد گر بهین دارای واریانس مجانبی V_I است، اما لزوماً به شکل برآورد گرهای امتیاز تمایل نیست. با در نظر گرفتن ویژگی های برآورد گر بهین می توان در باره کارایی برآورد گرهای امتیاز تمایل مختلف بحث کرد. در واقع هرچه برآورد گر امتیاز تمایل منطبق بر برآورد گر بهین مورد نظر باشد، کارایی بیش تری دارد.

با در نظر گرفتن یک مدل رگرسیونی برای $E(Y|x)$ ، برآورد گری با واریانس مجانبی V_I ساخته می شود. بنا بر این با استفاده از مدل $E(Y|x) = m(x; \beta)$ ، برآورد گر بهین عبارت است از:

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

که در آن $\hat{\beta}$ برآورد گر $\sqrt{n}$ سازگار برای پارامتر β تحت مدل $E(Y|x) = m(x; \beta)$ است. همچنین $\hat{\phi}$ برآورد گر $\sqrt{n}$ سازگار برای پارامتر ϕ است، که با استفاده از معادله ۳ محاسبه می شود. قضیه زیر نشان می دهد که واریانس مجانبی برآورد گر بهین در رابطه ۱۱ برابر با V_I است.

قضیه ۲.۲. شرط های قضیه ۱.۲ را در نظر بگیرید. همچنین فرض کنید که $\hat{\beta}$ در عبارت $\hat{\beta} = \beta + O_p(n^{-\frac{1}{2}})$ صدق می کند و مدل ابرجامعه ای، $m(x; \beta)$ ، مشتق های جزئی پیوسته در ناحیه ای باز شامل β دارد. برآورد گر $\hat{\theta}_{opt}$ در رابطه زیر صدق می کند:

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

که در آن

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

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

^{۱۴} optimal estimator

^{۱۵} ordinary least squares

کارا تر از دو برآوردگر PSA.CAL و PSA.MLE است؛ زیرا کارایی نسبی برای این برآوردگر در همه حالات تقریباً برابر با ۱ است. این بدان معنی است که برآوردگر PSA.AUG، دارای میانگین توان دوم خطاهای تقریباً برابر با MSE برآوردگر بهین می باشد. 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 مقایسه مقادارهای RRMSE در جدول ۲، تفاوت چندانی بین برآورد بهین و برآورد PSA.AUG مشاهده نمی شود. بنا بر این، برآوردگر PSA.AUG به طور مجانبی با برآوردگر بهین معادل است. از این رو، با وجود بی پاسخی در نمونه، برآوردگر امتیاز تمایل افزوده برای برآورد میانگین هزینه خالص خانوارهای شهری بهتر عمل کرده است.

کمترین مقدار معیار RRMSE برای هر چهار برآوردگر مورد نظر در هر سه نرخ بی پاسخی برای اندازه نمونه ۵۰۰ به دست آمده است. با کاهش اندازه نمونه این مقدار برای همه برآوردگرها افزایش می یابد. هم چنین با افزایش نرخ بی پاسخی در هر سه اندازه نمونه ای، شاهد افزایش معیار RRMSE و کاهش کارایی نسبی برآوردگرها هستیم. بیشترین مقدار RRMSE برای برآوردگر PSA.MLE برای نرخ بی پاسخی ۱۵ درصد و اندازه نمونه ۱۰۰ به دست آمده است. کمترین مقدار این معیار نیز برای برآوردگر PSA.AUG و برآوردگر OPT در اندازه نمونه ای ۵۰۰ با نرخ بی پاسخی ۵ درصد حاصل شده است. این نتیجه ها در شکل های ۱ تا ۶ قابل مشاهده اند.

خلاصه نتایج

نتیجه های حاصل از مطالعه شبیه سازی را به طور خلاصه می توان به شرح زیر برشمرد:

۱. برآوردگر PSA.CAL اندکی کارا تر از برآوردگر PSA.MLE است؛ زیرا با توجه به واریانس مجانبی برآوردگر PSA که به شکل

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

^{۱۶} root relative mean square error

^{۱۷} relative efficiency

در مدل رگرسیونی ۱۳ با استفاده از روش کمترین توان های دوم عادی ۱۵ و برازش مدل رگرسیونی y بر x_1 برآورد می شوند.

۴ معیارهای مقایسه

برآوردگرهای معرفی شده در این مقاله با توجه به معیارهای ریشه دوم میانگین توان دوم خطای نسبی^{۱۶} و کارایی نسبی^{۱۷} در سه اندازه نمونه و سه نرخ بی پاسخی اشاره شده مقایسه می شوند. این معیارها در زیر تعریف می شوند.

۱.۴ ریشه دوم میانگین توان دوم خطای نسبی

معیار ریشه دوم میانگین توان دوم خطای نسبی (RRMSE) برای برآوردگر $\hat{\theta}$ به صورت زیر تعریف می شود.

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

۲.۴ کارایی نسبی

معیار کارایی نسبی (eff) برای مقایسه دو برآوردگر $\hat{\theta}_1$ و $\hat{\theta}_2$ به صورت زیر تعریف می شود.

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

هرچه میزان کارایی برآوردگر تعدیل شده امتیاز تمایل PSA نسبت به برآوردگر بهین (OPT) به مقدار ۱ نزدیک باشد، برآوردگر امتیاز تمایل، کارایی بیش تری دارد.

۵ یافته های مطالعه

با توجه به سه اندازه نمونه و سه نرخ بی پاسخی فرض شده، جدول ۱، کارایی نسبی (eff) و جدول ۲، میزان ریشه دوم میانگین توان دوم خطای نسبی (RRMSE) برای چهار برآوردگر مورد نظر را نشان می دهند. شایان ذکر است که در این جدول ها مقادارهای RRMSE تا چهار رقم اعشار و مقادارهای eff تا دو رقم اعشار گرد شده اند. در جدول های ۱ و ۲ نماد nr نشان دهنده نرخ بی پاسخی است.

با توجه به مقادارهای کارایی نسبی در جدول ۱، در هر سه اندازه نمونه و سه میزان بی پاسخی، برآوردگر PSA.AUG

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

۳. با کاهش اندازه نمونه و افزایش میزان بی‌پاسخی، از کارایی برآوردگرها کاسته می‌شود.

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

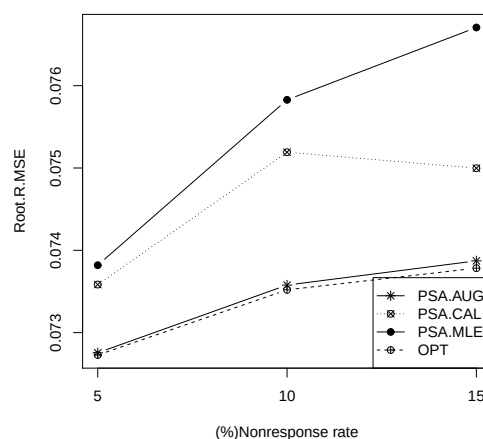
۲. برآوردگر PSA.AUG کارا تر از برآوردگر PSA به‌شکل

جدول ۱. کارایی نسبی سه برآوردگر تعدیل‌شده امتیاز تمایل نسبت به برآوردگر بهین

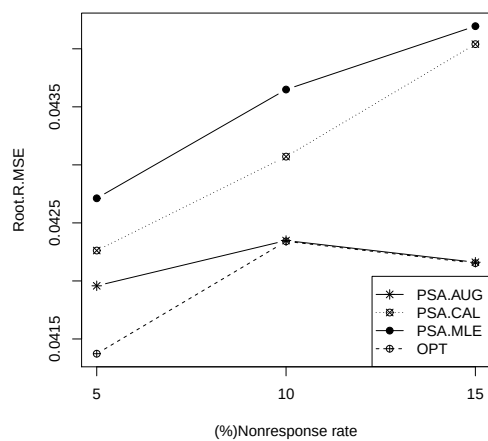
		$n = 500$		$n = 300$		$n = 100$			
برآوردگر	$nr = 0.15$	$nr = 0.1$	$nr = 0.05$	$nr = 0.15$	$nr = 0.1$	$nr = 0.05$	$nr = 0.15$	$nr = 0.1$	$nr = 0.05$
PSA.MLE	۰/۹۲	۰/۹۴	۰/۹۷	۰/۹۱	۰/۹۴	۰/۹۴	۰/۹۳	۰/۹۷	۰/۹۷
PSA.CAL	۰/۹۶	۰/۹۶	۰/۹۸	۰/۹۲	۰/۹۶	۰/۹۶	۰/۹۴	۰/۹۸	۰/۹۸
PSA.AUG	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰

جدول ۲. ریشه دوم میانگین توان دوم خطای نسبی چهار برآوردگر

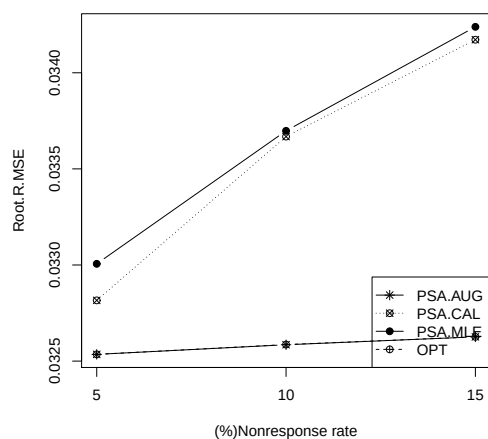
		$n = 500$		$n = 300$		$n = 100$			
برآوردگر	$nr = 0.15$	$nr = 0.1$	$nr = 0.05$	$nr = 0.15$	$nr = 0.1$	$nr = 0.05$	$nr = 0.15$	$nr = 0.1$	$nr = 0.05$
PSA.MLE	۰/۰۷۶۷	۰/۰۷۵۸	۰/۰۷۳۸	۰/۰۴۴۲	۰/۰۴۳۶	۰/۰۴۲۷	۰/۰۳۴۲	۰/۰۳۳۷	۰/۰۳۳۰
PSA.CAL	۰/۰۷۵۰	۰/۰۷۵۲	۰/۰۷۳۶	۰/۰۴۴۰	۰/۰۴۳۱	۰/۰۴۲۳	۰/۰۳۴۲	۰/۰۳۳۷	۰/۰۳۲۸
PSA.AUG	۰/۰۷۳۹	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



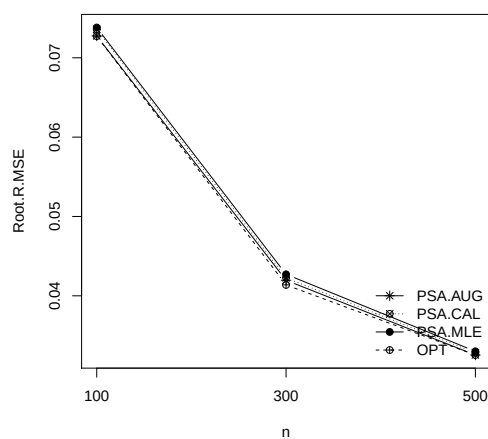
شکل ۱. اثر نرخ بی‌پاسخی بر عملکرد برآوردگرها با اندازه نمونه ۱۰۰



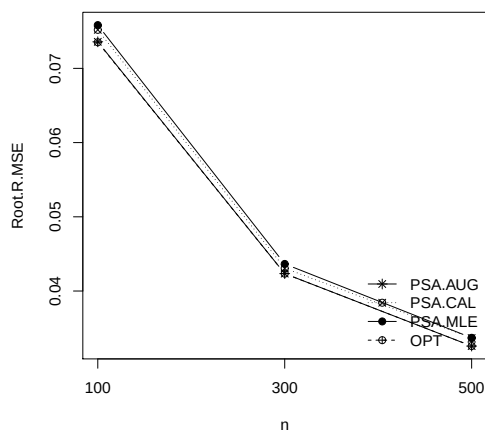
شکل ۲. اثر نرخ بی پاسخی بر عملکرد برآورد گرها با اندازه نمونه ۳۰۰



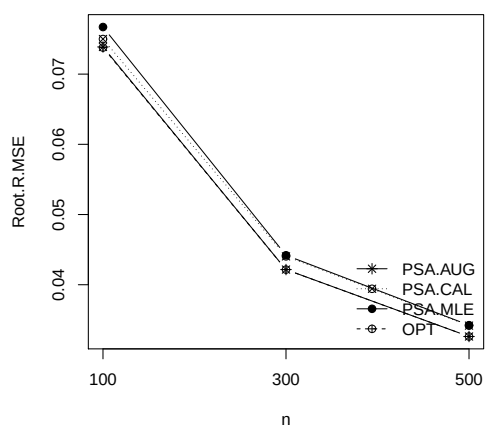
شکل ۳. اثر نرخ بی پاسخی بر عملکرد برآورد گرها با اندازه نمونه ۵۰۰



شکل ۴. اثر اندازه نمونه بر عملکرد برآورد گرها با نرخ بی پاسخی ۵ درصد



شکل ۵. اثر اندازه نمونه بر عملکرد برآورد گرها با نرخ بی‌پاسخی ۱۰ درصد



شکل ۶. اثر اندازه نمونه بر عملکرد برآورد گرها 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 comparison 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 nonresponse 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 effects. *Biometrika*, **70**, 41-55.