مجله علوم آماری، بهار و تابستان ۱۳۹۹
جلد ۱۴، شماره ۱، ص ۹۵ – ۱۱۲
DOI: 10.29252/jss.14.1.95

تحلیل نیم‌پارامتری مدل‌های رگرسیونی برای پاسخ‌های سری توانی آماسیدهٔ صفر با متغیرهای تبیینی گم‌شده

نفیسه خجسته و احسان بهرامی

گروه آمار، دانشکده علوم ریاضی، دانشگاه شهید بهشتی
تاریخ دریافت: ۱۳۹۷/۰۱/۲۰ تاریخ پذیرش: ۱۳۹۸/۱۷/۱۳

چکیده: در این مقاله پاسخ‌های شمارشی با تعداد صفر زیاد، که داده‌های آماسیدهٔ صفر نامیده می‌شوند، مورد تحلیل قرار گرفته‌اند. فرض می‌شود پاسخ‌ها از سری توانی آماسیدهٔ صفر پیروی می‌کنند. همچنین به دلیل وجود گم‌شدگی از نوع تصادفی در متغیرهای تبیینی برخی از داده‌های کاربردی، انواع روش‌های برآوردیابی پارامترهای مدل براساس تابع امتیاز با و بدون درنظر گرفتن گم‌شدگی برای مدل رگرسیونی ارایه شده است. در این میان، معلوم یا نامعلوم بودن احتمال انتخاب متغیر تبیینی گم‌شده منجر به ارایه روش نیم‌پارامتری برای برآورد پارامترها در مدل رگرسیونی سری توانی آماسیدهٔ صفر می‌شود. به منظور تشریح روش پیشنهادی، مطالعه‌ای شبیه‌سازی برای مدل رگرسیونی دوجمله‌ای منفی آماسیدهٔ صفر با متغیرهای تبیینی گم‌شده به عنوان یک مدل رگرسیونی سری توانی انجام می‌شود و سپس مثالی از داده‌های واقعی ارایه می‌شود. در انتها، عملکرد روش نیم‌پارامتری در مقایسه با روش ماکسیمم درستنمایی، مورد–کامل، احتمال وارون وزنی مورد بررسی و ارزیابی قرار گرفته است.

واژه‌های کلیدی: آماسیدهٔ صفر، داده‌های گم‌شده، تابع امتیاز، گم‌شدگی تصادفی، تحلیل نیم‌پارامتری، احتمال انتخاب.

آدرس الکترونیکی نویسنده مسئول مقاله: احسان بهرامی سامانی، ehsan_bahrami_samani@yahoo.com
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J05، 62J12

۱ مقدمه

در بسیاری از رشته‌ها مانند زیست‌شناسی، پزشکی، جرم‌شناسی، اقتصاد، محیط زیست، علوم‌سیاسی، جامعه‌شناسی و غیره داده‌های شمارشی رخ می‌دهند، به طوری که اغلب این داده‌ها از توزیعی از خانوادهٔ سری‌های توانی مانند توزیع پواسون، دوجمله‌ای منفی و مانند آن تبعیت می‌کنند. بیشتر محققان برای تحلیل این داده‌ها از مدل‌های خطی تعمیم‌یافته مانند مدل رگرسیونی پواسون یا مدل رگرسیونی دوجمله‌ای منفی استفاده می‌نمایند. همچنین در تحلیل این داده‌ها با مسأله تعداد صفر زیاد، سروکار خواهیم داشت. به عنوان مثال می‌توان تحلیل داده‌های بیمهٔ شخص ثالث، تعداد مفصل‌های آسیب دیدهٔ یک فرد، تعداد دفعاتی که یک دانشجو غیبت غیر موجه دارد و تعداد مقالات منتشر شده در هر سال توسط یک محقق اشاره کرد. به دلیل وجود تعداد صفر زیاد در این داده‌ها، برای مدل‌سازی نمی‌توان از مدل رگرسیونی پواسون یا دوجمله‌ای منفی استفاده نمود. بنابراین برای رفع این مشکل از مدل رگرسیونی پواسون آماسیدهٔ صفر[1] یا مدل رگرسیونی دوجمله‌ای منفی آماسیدهٔ صفر استفاده می‌شود. مدل رگرسیونی پواسون آماسیدهٔ صفر برای مدل‌بندی داده‌های شمارشی با تعداد صفر زیاد در حالت تک متغیره توسط لمبرت (۱۹۹۲) مورد مطالعه قرار گرفته است. سپس مدل رگرسیونی دوجمله‌ای منفی آماسیدهٔ صفر توسط گرین و همکاران (۱۹۹۴) بیان شد. علاوه بر این، آزمون امتیاز برای مدل رگرسیونی پواسون آماسیدهٔ صفر در مقابل مدل رگرسیونی پواسون توسط جانساکل و هیند (۲۰۰۲) مطرح شد.

از طرفی موضوع گم‌شدگی داده‌ها، امری است که نمی‌توان از آن چشم پوشی کرد. ساختار گم‌شدگی نخستین بار توسط روبین (۱۹۷۶) مطرح شد. وی معتقد بود داده‌های گم‌شده دارای سه نوع ساختار گم‌شدگی هستند. ساختار اول: گم‌شدگی کاملاً تصادفی[2] $(MCAR)$، ساختار دوم: گم‌شدگی تصادفی[3] (MAR) و ساختار سوم: گم‌شدگی غیرتصادفی[4] $(NMAR)$، در همین راستا لیتل و روبین (۲۰۱۴) و ابراهیم و همکاران (۱۹۷۶) استنباط‌هایی را برای مدل‌های خطی تعمیم‌یافته با حضورگم‌شدگی انجام دادند. همچنین روش انتخاب مدل، زمانی که متغیرهای تبیینی به طور تصادفی گم‌شده هستند، توسط چن و فو ارائه شد. از طرفی برآورد پارامترهای مدل رگرسیونی سری توانی، زمانی که متغیرهای تبیینی دارای گم‌شدگی هستند، مسأله مهمی است. در همین راستا برآورد پارامترها در این نوع از مدل‌های رگرسیونی به روش بیزی توسط ماسون و همکاران (۲۰۱۰) مورد پژوهش قرار گرفت. همچنین تحلیل نیم‌پارامتری به عنوان یک روش

[1] Zero Inflated
[2] Missing Completely At Random
[3] Missing At Random
[4] Not Missing At Random

بسیار مهم و کارساز در برآورد پارامترهای مدل مورد استفاده قرار می‌گیرد، از سوی دیگر تحلیل نیم پارامتری مربوط به پارامترهای مدل رگرسیونی اولین‌بار توسط وانگ و همکاران (۱۹۹۷) مورد بررسی و تحقیق قرار گرفت. از سویی دیگر، تحلیل نیم‌پارامتری مربوط به پارامترهای مدل‌های خطی تعمیم‌یافته توسط لینتون و نیلسن (۱۹۹۵) معرفی شدند. تحلیل نیم پارامتری روی مدل‌های رگرسیونی سری توانی آماسیدهٔ صفر تاکنون مورد بررسی قرار نگرفته است و تنها می‌توان به تحلیل نیم‌پارامتری مدل‌های رگرسیونی پواسون آماسیدهٔ صفر با امکان گم‌شدگی متغیرهای تبیینی گم‌شده که توسط لوکسا و همکاران (۲۰۱۶) اشاره نمود. نکته قابل توجه آن است لوکسا و همکاران (۲۰۱۶) داده‌هایی که مورد بررسی قرار داده‌اند از توزیع پواسون می‌باشند که با مسأله بیش‌پراکنش مواجه نیستند. زمانی مسأله بیش پراکنش رخ می‌دهد که واریانس متغیر شمارشی از مقدار مورد انتظار آن متغیر بیشتر شود. بنابراین نمی‌توان از مدل رگرسیونی پواسون آماسیدهٔ صفر استفاده نمود. از این رو کار انجام شده توسط لوکسا و همکاران (۲۰۱۶)، ناکارآمد می‌شود. بنابراین با ارایه مدل‌های رگرسیونی دوجمله ای منفی آماسیدهٔ صفر که نوعی از مدل‌های رگرسیونی سری توانی آماسیدهٔ صفر هستند و دارای بیش‌پراکنش ذاتی‌اند، می‌توان مدل‌های مناسبی را ارایه نمود. در این مقاله با تعمیم کار لوکسا و همکاران (۲۰۱۶)، تحلیل نیم‌پارامتری مدل‌های رگرسیونی سری توانی آماسیده صفر با امکان گم‌شدگی تصادفی در متغیرهای تبیینی انجام می‌شود.

در بخش ۲ مدل رگرسیونی سری توانی آماسیدهٔ صفر معرفی می‌شود. سپس در بخش ۳ انواع شیوه‌های برآورد درحالت با و بدون گم‌شدگی در متغیرهای تبیینی مطرح می‌شود. سپس با معرفی روش برآورد نیم‌پارامتری وزنی برای مدل‌های رگرسیونی سری توانی آماسیدهٔ صفر، جزئیات این روش معرفی می‌شود. در بخش ۴ به مطالعهٔ شبیه‌سازی روش‌های پیشنهادی بیان شده در بخش ۳ بر روی مدل رگرسیونی دوجمله‌ای منفی آماسیدهٔ صفر با متغیرهای تبیینی گم‌شده پرداخته می‌شود. در نهایت در بخش ۵، روش‌های پیشنهادی در مورد داده‌های واقعی پیاده‌سازی می‌شوند.

۲ مدل رگرسیونی سری توانی آماسیدهٔ صفر

متغیر تصادفی Y دارای توزیع سری‌توانی[5] است، هرگاه دارای تابع جرم احتمال به صورت

$$f(y|\theta) = \frac{a(y)\theta^y}{g(\theta)}, \qquad y = ۰, ۱, ۲, ...$$

[5]Power Series

باشد، که در آن $a(y)$ تابعی نامنفی $(a(y) > ۰)$ و $g(\theta) = \sum_{y=۰}^{\infty} a(y)\theta^y$ تابعی مثبت، متناهی و مشتق‌پذیر از θ است و به عنوان ثابت نرمال کننده در این توزیع نقش ایفا می‌کند. واضح است با در نظر گرفتن $a(y) = \binom{m+y-۱}{m-۱}$ و $g(\theta) = (\frac{۱}{p})^m = (\frac{۱}{۱-\theta})^m$ به طوری‌که $\theta = ۱ - p$ ،توزیع سری توانی به توزیع دوجمله‌ای منفی با پارامترهای (m, p) تبدیل می‌شود. هم‌چنین با درنظرگرفتن $a(y) = \frac{۱}{y!}$ و $g(\theta) = e^\theta$، این توزیع به توزیع پواسون با پارامتر θ تبدیل می‌شود.

عموماً توزیع داده‌های آماسیدهٔ صفر به صورت آمیخته از توزیع تباهیده با جرم ۱ در صفر و توزیع سری‌های توانی (PS) مانند پواسون و دوجمله‌ای منفی تعریف می‌شوند و با نماد $ZIPS(\theta, \omega)$ قابل نمایش هستند. تابع جرم احتمال توزیع سری‌های توانی آماسیدهٔ صفر به صورت

$$P(Y = y|\omega, \theta) = \begin{cases} \omega + (۱ - \omega)f(۰|\theta) & y = ۰ \\ (۱ - \omega)f(y|\theta) & y = ۱, ۲, \dots \end{cases}$$

در نظرگرفته می‌شود، که در آن $f(۰|\theta)$ تابع جرم احتمال توزیع سری‌های توانی و ω پارامتر آمیختگی است که در بازهٔ $\frac{-f(۰|\theta)}{۱ - f(۰|\theta)} < \omega < ۱$ تغییر می‌کند. به عبارت دیگر این پارامتر، احتمال صفر بودن y را که نمی‌تواند به طور کامل توسط فرض مدل برآورد شود، معرفی می‌نماید و برای مدل‌هایی با تعداد صفرهای زیاد به کار می‌رود. احتمال تعداد صفر برابر با $\omega + (۱ - \omega)f(۰|\theta)$ است در حالی‌که احتمال برای yهایی که مقدار مثبت را اختیار می‌کنند، برابر با $(۱ - \omega)f(y|\theta)$ خواهد بود.

فرض کنید Y متغیر پاسخ از توزیع سری توانی آماسیدهٔ صفر با پارامترهای ω و θ باشد، همچنین با درنظر گرفتن X و Z به عنوان ماتریس طرح مربوط به متغیرهای تبیینی، مدل رگرسیونی سری توانی آماسیدهٔ صفر به صورت

$$Y \sim ZIPS(\omega, \theta), \quad \log(\theta) = \boldsymbol{\beta}^\top \chi, \quad \text{logit}(\omega) = \boldsymbol{\gamma}^\top \chi,$$

است، که در آن $\chi = (۱^\top, X^\top, Z^\top)^\top$ ماتریس طرح مربوط به متغیرهای تبیینی و $\boldsymbol{\eta} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$ بردار پارامترهای مدل هستند. فرض کنید

$$\omega = H(\boldsymbol{\gamma}^\top \chi), \quad \theta = \exp(\boldsymbol{\beta}^\top \chi),$$

که در آن $H(u) = (۱ + \exp(-u))^{-۱}$ تابع پیوند لوژستیک و $\text{logit}(u) = H^{-۱}(u)$ است. در

این صورت مدل رگرسیونی سری توانی آماسیدهٔ صفر معرفی می‌شود. تابع جرم احتمال Y در این مدل به صورت

$$P(Y=y|X,Z)=H\left(\boldsymbol{\gamma}^{\top}\chi\right)I(y=۰)+[۱-H\left(\boldsymbol{\gamma}^{\top}\chi\right)]\frac{a(y)[\exp(\boldsymbol{\beta}^{\top}\chi)]^{y}}{g(\exp(\boldsymbol{\beta}^{\top}\chi))},$$

به دست می‌آید. از سوی دیگر

$$\begin{aligned}P(Y=۰|X,Z)&=H\left(\boldsymbol{\gamma}^{\top}\chi\right)+[۱-H\left(\boldsymbol{\gamma}^{\top}\chi\right)]\frac{a(y)}{g(\exp(\boldsymbol{\beta}^{\top}\chi))}\\&=H\left(\boldsymbol{\gamma}^{\top}\chi\right)[۱-\frac{a(y)}{g(\exp(\boldsymbol{\beta}^{\top}\chi))}]+\frac{a(y)}{g(\exp(\boldsymbol{\beta}^{\top}\chi))},\end{aligned}$$

و تابع لگاریتم درستنمایی برای نمونه تصادفی $\{(Y_i,\chi_i):i=۱,\ldots,n\}$ به صورت

$$\begin{aligned}\ell(\boldsymbol{\eta})=\textstyle\sum_{i=۱}^{n}I(Y_i=۰)\{\log[H\left(\boldsymbol{\gamma}^{\top}\chi_i\right)(۱-\frac{a(۰)}{g(\exp(\boldsymbol{\beta}^{\top}\chi))})+\frac{a(۰)}{g(\exp(\boldsymbol{\beta}^{\top}\chi))}]\}\\+I(Y_i>\\۰)\left\{\log\left[۱-H\left(\gamma^{\top}\chi_i\right)\right]+\log(a(Y_i))+Y_i\beta^{\top}\chi-\log(g(\exp(\beta^{\top}\chi)))\right\},\end{aligned}$$

محاسبه می‌شود، که در آن $I(Y_i=۰)$ و $I(Y_i>۰)$ توابع نشانگر هستند. در ادامه شیوه‌های مختلف برآورد پارامترهای این مدل در حالت با و بدون گم‌شدگی در متغیرهای تبیینی مورد بررسی قرار می‌گیرند. نکته قابل تأمل آن است که با توجه به نوع مدل رگرسیونی و نوع روش برآورد پارامترها، برآوردهای پارامتری، ناپارامتری یا نیم‌پارامتری حاصل می‌شوند. یادآوری می‌شود که مدل رگرسیونی از نوع پارامتری است، یعنی رابطه بین متغیرهای پاسخ و متغیرهای تبیینی معلوم است اما یک جزء مدل به روش ناپارامتری برآورد می‌شود، این موضوع به تحلیل و برآورد نیم‌پارامتری می‌انجامد، درحالی که مدل نیم‌پارامتری فرض نشده است.

۳ برآوردیابی پارامترهای مدل با و بدون گم‌شدگی در متغیرهای تبیینی

وقتی متغیرهای تبیینی بدون گم‌شدگی هستند، یک روش برآورد بردار پارامترهای مدل رگرسیونی سری توانی آماسیدهٔ صفر، بر اساس تابع امتیاز

$$U_{F,n}(\boldsymbol{\eta})=\frac{۱}{\sqrt{n}}\frac{\partial\ell(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}=\frac{۱}{\sqrt{n}}\begin{pmatrix}\frac{\partial\ell(\boldsymbol{\eta})}{\partial\gamma}\\\frac{\partial\ell(\boldsymbol{\eta})}{\partial\beta}\end{pmatrix}=\frac{۱}{\sqrt{n}}\textstyle\sum_{i=۱}^{n}S_i(\boldsymbol{\eta}),$$

است، که درآن

$$S_i(\boldsymbol{\eta}) = \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \left(S_{i1}^\top(\boldsymbol{\eta}), S_{i2}^\top(\boldsymbol{\eta})\right)^\top, \quad i = 1, \ldots, n,$$

$$S_{i1}(\boldsymbol{\eta}) = \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \gamma}, \tag{۱}$$

$$S_{i2}(\boldsymbol{\eta}) = \frac{\partial \ell_i(\boldsymbol{\eta})}{\partial \beta}. \tag{۲}$$

از طرفی $E(U_{F,n}(\boldsymbol{\eta})) = 0$، نشان می‌دهد $U_{F,n}(\boldsymbol{\eta})$ برآوردگر نااریب است. همچنین با قرار دادن $U_{F,n}(\boldsymbol{\eta}) = 0$، برآوردگر ماکسیمم درستنمایی $\boldsymbol{\eta}$ که با $\widehat{\boldsymbol{\eta}}_F$ نشان داده می‌شود، به دست می‌آید.

فرض کنید X_i متغیر تبینی برای فرد i ام با گم‌شدگی از نوع MAR و

$$\delta_i = \begin{cases} 1 & X_i \text{ مشاهده شود} \\ 0 & X_i \text{ مشاهده نشود} \end{cases}$$

متغیری نشانگر باشد. همچنین W_i را یک متغیر جانشین برای X_i در نظر بگیرید، به طوری‌که W_i و Z_i همواره مشاهده می‌شوند. تحت ساختار MAR احتمال انتخاب به صورت

$$P(\delta_i = 1 | Y_i, X_i, V_i) = \pi(Y_i, V_i),$$

است (روبین، ۱۹۷۶)، که در آن $V = (Z^T, W^\top)^\top$، در این حالت روش‌های برآورد پارامترهای مدل رگرسیونی سری توانی آماسیدهٔ صفر، وقتی متغیرهای تبیینی با گم‌شدگی هستند، به شرح زیر خواهند بود.

الف) روش برآورد مورد کامل:

برآورد مورد کامل [۶] یکی از روش‌های برخورد با داده‌های گم‌شدگی است و در بسیاری از نرم افزارهای آماری به صورت پیش فرض وجود دارد. در این روش داده‌های گم‌شده حذف می‌شوند. با توجه به آنکه ساختار گم‌شدگی داده‌ها MAR است، تابع برآورد CC به صورت

$$U_{CC,n}(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta_i S_i(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta_i \begin{pmatrix} S_{i1}(\boldsymbol{\eta}) \\ S_{i2}(\boldsymbol{\eta}) \end{pmatrix},$$

[6] Complete Case

به دست می‌آید، که در آن $S_{i1}(\boldsymbol{\eta})$ و $S_{i2}(\boldsymbol{\eta})$ به ترتیب در روابط (۱) و (۲) تعریف شده‌اند.

با حل معادله $U_{CC,n}(\boldsymbol{\eta}) = 0$، برآوردگر CC به دست می‌آید، که با $\widehat{\boldsymbol{\eta}}_{CC}$ نشان داده می‌شود. نکته قابل توجه آن است که $E(U_{CC,n}(\boldsymbol{\eta})) \neq 0$ (وانگ و همکاران، ۱۹۹۷). بنابراین برآوردگر $U_{CC,n}(\boldsymbol{\eta})$، اریب است. اگرچه تحلیل CC ساده است اما دو معضل به همراه دارد: ۱-از دست دادن میزان کارایی به علت حذف نمونه‌های ناقص، ۲- عامل بالقوه برای برآوردهای ناسازگار وقتی مجموعه داده‌های باقیمانده حاصل از حذف داده‌های گم‌شده از مجموعهٔ داده‌های کامل، زیرنمونه‌ای تصادفی از داده‌های اصلی نیست.

ب) روش برآورد احتمال معکوس وزنی:

ایده اصلی این روش از برآوردگر هورویتز-تامپسون (۱۹۵۲) است که توسط ژائو و همکاران (۱۹۹۲) توسعه یافت و منجر به معرفی برآوردگر شد. روش احتمال معکوس وزنی [۷] وقتی متغیرهای تبینی دارای گم‌شدگی هستند برای داده‌های مشاهده شده وزن‌های وارون در نظر می‌گیرد و باعث می‌شود که اریبی حاصل از نمونه‌های ناقص حذف شده، کاهش یابد. از سوی دیگر اگر X_i مشاهده شود $\delta_i = 1$ و اگر X_i مشاهده نشود، $\delta_i = 0$. همچنین تحت ساختار MAR، احتمال انتخاب یا مشاهدهٔ X_i به صورت

$$P(\delta_i = 1|Y_i, X_i, V_i) = \pi(Y_i, V_i),$$

در نظر گرفته می‌شود. بنابراین، تابع برآورد وزنی به صورت

$$U_{W,n}(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\pi(Y_i, V_i)} S_i(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\pi(Y_i, V_i)} \binom{S_{i1}(\boldsymbol{\eta})}{S_{i2}(\boldsymbol{\eta})},$$

به دست می‌آید، که درآن $S_{i1}(\boldsymbol{\eta})$ و $S_{i2}(\boldsymbol{\eta})$ به ترتیب در روابط (۱) و (۲) تعریف شده‌اند. با توجه به شرایط $\pi(Y_i, V_i)$، تابع برآورد متفاوت و در نتیجه برآوردگرهای متفاوتی به شرح زیر حاصل خواهند شد.

الف) $\pi(Y_i, V_i)$ معلوم باشد:

در بسیاری از مطالعات $\pi(Y_i, V_i)$ در مرحله طراحی مشخص می‌شود (برسلو و همکاران، ۱۹۸۸). در این‌حالت اگر $\pi(Y_i, V_i)$ از ابتدا به طور درست مشخص شده باشد، تابع برآورد وزنی بدون تغییر خواهد شد. در این صورت با حل معادلهٔ $U_{W,n}(\boldsymbol{\eta}) = 0$، برآوردگر وزنی برای $\boldsymbol{\eta}$، که با $\widehat{\boldsymbol{\eta}}_W$ نشان داده می‌شود، به دست می‌آید. از طرفی دیگر $E(U_{W,n}(\boldsymbol{\eta})) = 0$، (وانگ و همکاران، ۱۹۹۷)، بنابراین $U_{W,n}(\boldsymbol{\eta})$ برآوردگر نااریب خواهد شد.

[7] Inverse Probability Weighted

ب) $\pi(Y_i, V_i)$ **نامعلوم باشد:**

برآورد $\pi(Y_i, V_i)$ به روش ناپارامتری یا روش پارامتری امکان‌پذیر است. فرض کنید $v_1, \ldots, v_m$ مقادیری متمایز از V_i را وقتی $V_i = (Z_i^T, W_i^T)$ گسسته در نظر گرفته شود را نشان دهند، آنگاه برآورد ناپارامتری $\pi(y, v)$ به صورت

$$\widehat{\pi}(y_i, v_i) = \frac{\sum_{k=1}^{n} \delta_k I(Y_k = y, V_k = v)}{\sum_{i=1}^{n} I(Y_i = y, V_i = v)},$$

به دست می‌آید، که در آن مقادیر Y به صورت $y = 0, 1, \ldots$ و مقادیر V نیز به صورت $\{v_1, \ldots, v_m\}$ در نظر گرفته می‌شوند. در این مقاله فرض می‌شود که V_i گسسته است. (در حالتی که V_iها پیوسته باشند برای برآورد $\pi(Y_i, V_i)$، به وانگ و همکاران (۱۹۹۷) مراجعه شود.) بنابراین با جایگذاری $\widehat{\pi}(Y_i, V_i)$ در تابع برآورد وزنی، تابع برآورد نیم‌پارامتری وزنی به صورت

$$U_{Ws,n}(\boldsymbol{\eta}, \widehat{\pi}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\widehat{\pi}(Y_i, V_i)} S_i(\Theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\widehat{\pi}(Y_i, V_i)} \left(\begin{smallmatrix} S_{i1}(\boldsymbol{\eta}) \\ S_{i2}(\boldsymbol{\eta}) \end{smallmatrix}\right),$$

به دست می‌آید، که در آن $S_{i1}(\boldsymbol{\eta})$ و $S_{i2}(\boldsymbol{\eta})$ به ترتیب در روابط (۱) و (۲) تعریف شده‌اند. با حل معادلهٔ $U_{Ws,n}(\boldsymbol{\eta}) = 0$، برآوردگر نیم‌پارامتری وزنی برای $\boldsymbol{\eta}$، که با $\widehat{\boldsymbol{\eta}}_{Ws}$ نشان داده می‌شود، به دست می‌آید. از سوی دیگر رابطه

$$U_{Ws,n}(\boldsymbol{\eta}, \widehat{\pi}) - U_{Ws,n}(\boldsymbol{\eta}, \pi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\delta_i - \pi(Y_i, V_i)}{\pi(Y_i, V_i)}\right] S_i^*(\boldsymbol{\eta}) + o_p(1),$$

برقرار است، که در آن $S_i^*(\boldsymbol{\eta}) = E(S_i(\boldsymbol{\eta}))|Y_i, V_i)$، $i = 1, \ldots, n$، نماد o (کوچک) برای تقریب فرمول‌ها به کار می‌رود به این مفهوم که تقریب‌های نامتناهی و تقریب‌های چند جمله‌ای نامتناهی را وقتی متغیر به سمت صفر میل می‌کند، مشخص می‌کند. ($o_p(1)$ مجموعه‌ای از عبارات هستند که بزرگترین توان آنها، توان یک است. وقتی متغیر به سمت صفر میل می‌کند، این عبارت برابر با صفر است.) در این صورت مشخص می‌شود که $U_{Ws,n}(\boldsymbol{\eta}, \widehat{\pi})$، تابع برآورد تقریبا نااریب است.

۴ مطالعه شبیه‌سازی

تولید داده‌ها در یک فرایند دو مرحله‌ای شبیه‌سازی می‌شوند. ابتدا داده‌های کامل تولید می‌شود. برای این منظور ابتدا متغیرهای تبیینی و متغیر پاسخ بر اساس مدل رگرسیون دوجمله‌ای منفی آماسیدهٔ صفر تولید

می‌شود. سپس با ایجاد گم‌شدگی از نوع تصادفی در متغیر تبیینی،تولید داده‌ها صورت می‌پذیرد. آنگاه نحوهٔ برآورد پارامترهای مدل بیان می‌شود. ابتدا متغیرهای تبیینی X و Z به‌ترتیب از توزیع یکنواخت روی بازه $(-۱, ۲)$و توزیع برنولی با احتمال موفقیت ۰/۴ به تصادف تولید می‌شوند. در مرحلهٔ دوم در متغیرهای تبیینی X، گم‌شدگی از نوع تصادفی ایجاد شده است. پس از تولید متغیرهای تبیینی، ابتدا متغیر پاسخ Y_i^{NB} از روی مدل رگرسیونی دوجمله‌ای منفی به صورت

$$\begin{aligned} Y_i^{NB} &\sim NB(\Theta_i, m), \quad m = ۱۰, \\ \log \theta_i &= \beta_۰ + \beta_۲ X_i + \beta_۳ Z_i, \end{aligned}$$

تولید می‌شود. حال برای تولید Y_i از مدل رگرسیونی دوجمله‌ای منفی آماسیدهٔ صفر

$$\begin{aligned} Y_i &\sim ZINB(\theta, m, \omega), \quad m = ۱۰, \\ \text{logit}(\omega_i) &= \gamma_۰ + \gamma_۱ X_i + \gamma_۲ Z_i, \\ \log \theta_i &= \beta_۰ + \beta_۱ X_i + \beta_۲ Z_i, \end{aligned}$$

استفاده می‌شود، که درآن پارامتر آمیختگی از روی مدل به صورت

$$\omega_i = \frac{\exp(\gamma_۰ + \gamma_۱ X_i + \gamma_۲ Z_i)}{۱ + \exp(\gamma_۰ + \gamma_۱ X_i + \gamma_۲ Z_i)}$$

به‌دست می‌آید. با در نظر گرفتن متغیر تصادفی $U_i^{(۱)}$ که از توزیع یکنواخت روی بازه $(۰, ۱)$ تولید می‌شود. می‌توان $U_i^{(۱)}$ را با پارامتر محاسبه شدهٔ ω_i مورد مقایسه قرار داد و نتایج زیر را به دست آورد.
اگر $U_i^{(۱)} \geq \omega_i$، آنگاه مقدار $Y_i = Y_i^{NB}$ در نظر گرفته می‌شود. به عبارت دیگر مقدار Y_i همان نمونهٔ تولید شده از توزیع دوجمله‌ای منفی است.
اگر $U_i^{(۱)} < \omega_i$، آنگاه مقدار Y_i صفر درنظر گرفته می‌شود. شایان یادآوری است مقادیر واقعی بردار پارامترهای مدل رگرسیونی دو جمله‌ای منفی آماسیدهٔ صفر به صورت

$$\begin{aligned} \boldsymbol{\eta} &= (\boldsymbol{\gamma}, \boldsymbol{\beta})^\top = (\gamma_۰, \gamma_۱, \gamma_۲, \beta_۰, \beta_۱, \beta_۲)^\top, \\ \boldsymbol{\eta}_۱ &= (-۱, -۱, ۰/۵, ۱, ۰/۷, ۱)^\top, \\ \boldsymbol{\eta}_۲ &= (-۰/۱, -۱, ۰/۵, ۱, ۰/۷, ۱)^\top, \end{aligned}$$

است. بنابراین به این شیوه یک نمونهٔ تصادفی از مدل رگرسیون دوجمله‌ای منفی آماسیدهٔ صفر تولید شد. در این بخش با استفاده از مکانیسم گم‌شدگی تصادفی، احتمال انتخاب متغیر تبیینی X_i به صورت

$$\pi(Y_i, V_i) = P(\delta_i = ۱|Y_i, X_i, V_i)$$
$$= H(\mu_۰ + \mu_۱ Y_i I(Y_i < ۶) + \mu_۲ I(Y_i \geq ۶) + \mu_۳ Z_i + \mu_۴ W_i),$$

در نظر گرفته می‌شود، که در آن $V_i = (Z_i, W_i)$ و W_i متغیر تصادفی دودویی به صورت

$$W_i = \begin{cases} ۱ & X_i \leq ۰ \\ ۰ & X_i > ۰ \end{cases}$$

و H تابع پیوند لوژستیک است. همانند بخش ۱، یک متغیر تصادفی مانند $U_i^{(۲)}$ در نظر گرفته می‌شود که از توزیع یکنواخت روی بازه $(۰, ۱)$ تولید می‌شود. سپس با محاسبه $\pi(Y_i, V_i)$ و مقایسه آن با $U_i^{(۲)}$ نتایج زیر در مورد تولید داده‌های گم‌شده به دست می‌آید:

- اگر $U_i^{(۲)} < \pi(y_i, v_i)$، آن‌گاه مقدار X_i گم‌شده در نظر گرفته می‌شود.
- اگر $U_i^{(۲)} \geqslant \pi(y_i, v_i)$، آن‌گاه مقدار X_i تغییر نمی‌کند.

حال با در نظر گرفتن مقادیر واقعی برای بردار پارامترهای احتمال انتخاب $\boldsymbol{\mu} = (\mu_۰, \ldots, \mu_۴)^\top$، احتمال انتخاب متغیر X_i محاسبه می‌شود. برای این منظور، می‌توان دو حالت در نظر گرفت:

الف-.نسبت گم‌شدگی متغیر تبیینی X_i کم‌تر از ۵۰ درصد باشد.

ب- نسبت گم‌شدگی متغیر تبیینی X_i بیش‌تر از ۵۰ درصد باشد.

برای اینکه حالت (الف) اتفاق بیافتد لازم است مقادیر واقعی پارامترهای مربوط به احتمال انتخاب X_i به صورت

$$\boldsymbol{\mu}_۱ = (\mu_۰, \ldots, \mu_۴)^\top = (-۱, ۰/۳, ۱۰, ۰/۵, ۱)^\top,$$

در نظر گرفته شوند. در این حالت درصد داده‌های گم‌شده در متغیر تبیینی X کمتر از ۵۰ درصد خواهد شد. همچنین برای این‌که حالت (ب) رخ دهد مقادیر واقعی بردار پارامترهای مربوط به احتمال انتخاب X_i به صورت

$$\boldsymbol{\mu}_۲ = (\mu_۰, \ldots, \mu_۴)^\top = (-۱/۸, ۰/۳, ۱۰, ۰/۵, ۱)^\top,$$

در نظر گرفته می‌شود. حال احتمال انتخاب متغیر تبیینی X_i برای دو حالت (الف) و (ب) به طور جداگانه محاسبه و با متغیر $U_i^{(۲)}$ مقایسه و دو سری نمونه تصادفی برای متغیر X_i (حالت الف و ب) تولید می‌شود.

به طوری‌که متغیر X_i درکنار متغیر تبیینی Z_i روی متغیر پاسخ Y_i که در بخش اول تولید شده‌اند برازش داده می‌شوند.

حال براساس مراحل ب و مدل مورد نظر، چهار روش برآوردیابی برای برآورد بردار پارامترهای Θ مورد استفاده قرار می‌گیرد و چهار برآوردگر به صورت زیر حاصل می‌شوند.

- $\widehat{\boldsymbol{\eta}}_F$: برآورگر ماکسیم درستنمایی براساس تابع امتیاز $U_{F,n}(\boldsymbol{\eta})$ برای داده‌های کامل.
- $\widehat{\boldsymbol{\eta}}_{CC}$: برآوردگر CC براساس تابع برآورد $U_{CC,n}(\boldsymbol{\eta})$.
- $\widehat{\boldsymbol{\eta}}_W$: برآوردگر IPW وزنی براساس تابع برآورد $U_{W,n}(\boldsymbol{\eta},\pi)$.
- $\widehat{\boldsymbol{\eta}}_{Ws}$: برآوردگر IPW نیم‌پارامتری براساس تابع برآورد نیم پارامتری وزنی $U_{Ws}(\boldsymbol{\eta},\widehat{\pi})$.

در جدول ۱ نتایج شبیه‌سازی برای حالت (الف) ارائه شده است، که در آن میزان اریبی $(Bias)$، خطای استاندارد (SE)، احتمال پوشش (CP) هریک از پارامترهای مدل برای حالت با و بدون گم‌شدگی در متغیر تبیینی X برای حجم نمونه‌های $n = ۵۰۰$ و $n = ۱۰۰۰$ محاسبه شده است. دراین حالت درصد گم‌شدگی کمتر از ۵۰ درصد X است. در جدول ۲ نتایج شبیه‌سازی برای حالت (ب) بیان شده است. در این جدول نیز میزان اریبی، خطای استاندارد و احتمال پوشش هر یک از پارامترهای مدل برای حالت با و بدون گم‌شدگی در متغیر تبیینی برای حجم نمونه‌های $n = ۱۰۰۰$ و $n = ۲۰۰۰$ محاسبه شده است. همچنین درصد گم‌شدگی بالاتر از ۵۰ درصد X_i، فرض شده است. شایان یادآوری است که برآوردگر ماکسیمم درستنمایی براساس داده‌های بدون گم‌شدگی($\widehat{\boldsymbol{\eta}}_F$)به عنوان یک معیار مقایسه در مقابل برآوردگرهای دارای گم‌شدگی درمطالعات شبیه‌سازی به‌کار برده می‌شود. در جدول ۱ میزان گم‌شدگی در متغیر تبیینی X_i، برای حجم‌های $n = ۵۰۰$ و $n = ۱۰۰۰$ به ترتیب ۰/۲۶ و ۰/۳۳ است. از سوی دیگر درصد صفرهای تولید شده $(Y = ۰)$ از مجموعه‌داده‌های کامل (بدون گم‌شدگی) برای حجم‌های $n = ۵۰۰$ و $n = ۱۰۰۰$ به ترتیب ۰/۴۲ و ۰/۴۷ است. همچنین درصد صفرهای تولید شده $(Y = ۰)$ از مجموعه داده‌های ناکامل (با گم‌شدگی) برای حجم‌های $n = ۵۰۰$ و $n = ۱۰۰۰$ به ترتیب ۰/۱۷ و ۰/۲۳ است. جدول ۱ نشان می‌دهد که میزان اریبی برای هر کدام از برآوردگرهای $\widehat{\boldsymbol{\eta}}_F$ ، $\widehat{\boldsymbol{\eta}}_{CC}$ ، $\widehat{\boldsymbol{\eta}}_W$ و $\widehat{\boldsymbol{\eta}}_{Ws}$ وجود دارد. به طورخاص، برآوردگر CC برای پارامترهای رگرسیون لوژستیک $(\lambda_۰, \lambda_۱, \lambda_۲)^T$ دارای میزان اریبی بالایی است. همچنین مقادیر SE با افزایش تعداد نمونه، کاهش یافته‌اند. همچنین مقادیر SE تحت برآوردگر $\widehat{\boldsymbol{\eta}}_{Ws}$ از مقادیر SE تحت برآوردگر $\widehat{\boldsymbol{\eta}}_{CC}$ کمتر هستند. به عبارت دیگر می‌توان نشان داد که $SE(\widehat{\boldsymbol{\eta}}_F) \leq SE(\widehat{\boldsymbol{\eta}}_{Ws}) \leq SE(\widehat{\boldsymbol{\eta}}_W)$. احتمال پوشش براساس برآوردگر CC بدترین وضعیت ممکن را دارا است. از آنجا که احتمال پوشش ۰/۹۵ فرض می‌شود، مقادیر احتمال پوشش CC اختلاف

بسیاری با احتمال پوشش ۰/۹۵ دارند. احتمال پوشش با روش ماکسیمم درستنمایی از روش‌های دیگر بهتر است، اما این روش به مجموعه داده‌های کامل نیاز دارد. از طرفی احتمال پوشش $\widehat{\boldsymbol{\eta}}_W$ از احتمال پوشش $\widehat{\boldsymbol{\eta}}_{Ws}$ بهتر بوده است اما این روش برآورد IPW وزنی، ارزش عملی محدودی دارد به این خاطر که استفاده از آن به یک دانش خاصی برای تعیین مدل احتمال انتخاب نیازمند است. روش نیم‌پارامتری در مقابل روش IPW دارای مزیت بوده و در این روش شکل پارامتری برای احتمال انتخاب در نظر گرفته نمی‌شود.

در جدول ۲ میزان گم‌شدگی در متغیر تبیینی X، برای حجم‌های $n = ۱۰۰۰$ و $n = ۲۰۰۰$ به ترتیب ۰/۶۰ و ۰/۶۷ است. از سوی دیگر درصد صفرهای تولید شده $(Y = ۰)$ از مجموعه داده‌های کامل برای حجم‌های $n = ۱۰۰۰$ و $n = ۲۰۰۰$ به ترتیب ۰/۵۳ و ۰/۵۷ است. همچنین درصد صفرهای تولید شده از مجموعه داده‌های ناکامل (با گم‌شدگی) برای حجم‌های $n = ۱۰۰۰$ و $n = ۲۰۰۰$ به ترتیب ۰/۳۲ و ۰/۳۶ است. با این وجود با در نظر گرفتن فرض‌ها و نتایج عددی متفاوت برای دو جدول، نتایج تحلیلی یکسان هستند. برای مثال، میزان بالای گم‌شدگی در متغیر تبینی X_i در جدول ۲ باعث می‌شود احتمال پوشش تحت برآوردگر $\widehat{\boldsymbol{\eta}}_{CC}$ بسیار کم باشد. درحالی که در جدول ۱ نیز احتمال پوشش کم است اما در جدول ۲ این مقدار بیشتر است. هدف از در نظر گرفتن درصدهای متفاوت گم‌شدگی برای متغیر تبیینی بررسی عملکرد برآوردگرها درحالت‌های متفاوت بوده است. نتایج نشان از عملکرد یکسان در حالت‌های متفاوت دارند.

۵ مثال کاربردی

داده‌ها مربوط به تعداد ماهی‌هایی است $(count)$ که در یک پارک آبی توسط بازدیدکنندگان گرفته شده‌اند. در این مطالعه از ۲۵۰ گروه از بازدیدکنندگان تشکیل شده است. عوامل موثر بر تعداد ماهی‌های گرفته شده، تعداد کودکان در گروه $(C.child)$، تعداد بزرگسالان در گروه $(C.persons)$ و وجود یا عدم وجود راهنما $(leder)$ هستند. تعداد ماهی‌ها به عنوان متغیر پاسخ شمارشی و تعداد کودکان، تعداد بزرگسالان و وجود یا عدم وجود راهنما به عنوان متغیر تبیینی گسسته در نظر گرفته شده‌اند. برخی شاخص‌های مربوط به متغیرهای مورد علاقه به طور خلاصه در جدول ۳ گردآوری شده است.

همان‌طور که ملاحظه می‌شود، تعداد صفر زیادی در متغیر پاسخ تعداد ماهی‌ها ملاحضه می‌شود. همچنین در متغیر تبیینی تعداد بزرگسالان درصد گم‌شدگی ۰/۲۹ درصد وجود داشته است. با وجود تعداد زیاد صفر در داده‌ها برای اطمینان از این که کدام مدل برای برازش روی داده‌های مربوطه مناسب است

جدول ۱: نتایج شبیه‌سازی برای حالت (الف) و $m = ۱۰$

حجم نمونه	$n = ۵۰۰$				$n = ۱۰۰۰$			
پارامتر	$\widehat{\eta}_F$	$\widehat{\eta}_{CC}$	$\widehat{\eta}_W$	$\widehat{\eta}_{Ws}$	$\widehat{\eta}_F$	$\widehat{\eta}_{CC}$	$\widehat{\eta}_W$	$\widehat{\eta}_{Ws}$
$\gamma_۰$								
$Bias$	۰/۰۳۱	−۰/۷۱۰	−۰/۰۴۲	−۰/۰۳۳	−۰/۰۳۰	−۰/۶۳	−۰/۰۴۰	−۰/۰۱۲
$S.E.$	۰/۳۱۲	۰/۴۶۸	۰/۴۷۳	۰/۳۴۴	۰/۲۹۱	۰/۴۰۷	۰/۴۱۳	۰/۳۰۶
CP	۰/۹۴۸	۰/۴۰۸	۰/۹۳۷	۰/۹۳۴	۰/۹۴۴	۰/۵۰۲	۰/۹۴۱	۰/۹۴۰
$\gamma_۱$								
$Bias$	−۰/۰۲۴	−۰/۸۱۲	−۰/۰۴۵	−۰/۰۳۹	−۰/۰۲۰	−۰/۶۴۴	−۰/۰۳۵	−۰/۰۳۲
$S.E.$	۰/۳۰۵	۰/۴۲۷	۰/۴۳۳	۰/۳۲۲	۰/۲۷۶	۰/۴۰۶	۰/۴۱۲	۰/۲۹۳
CP	۰/۹۵۰	۰/۳۰۵	۰/۹۴۸	۰/۹۳۷	۰/۹۵۱	۰/۴۰۶	۰/۹۵۰	۰/۹۴۵
$\gamma_۲$								
$Bias$	۰/۰۴۱	−۰/۷۲۰	−۰/۰۴۰	۰/۰۳۲	۰/۰۳۱	−۰/۵۶۴	−۰/۰۳۸	۰/۰۲۹
$S.E.$	۰/۲۹۸	۰/۴۴۱	۰/۴۶۷	۰/۳۱۱	۰/۲۵۴	۰/۴۳۴	۰/۴۴۲	۰/۲۶۳
CP	۰/۹۵۰	۰/۴۰۷	۰/۹۴۷	۰/۹۳۹	۰/۹۴۷	۰/۵۰۹	۰/۹۵۰	۰/۹۴۶
$\beta_۰$								
$Bias$	−۰/۰۲۰	۰/۰۶۰	−۰/۰۱۱	۰/۰۱۰	۰/۰۱۲	۰/۰۵۹	−۰/۰۱۰	۰/۰۰۹
$S.E.$	۰/۰۵۴	۰/۰۶۱	۰/۰۶۶	۰/۰۵۹	۰/۰۵۲	۰/۰۵۸	۰/۰۵۹	۰/۰۵۳۱
CP	۰/۹۴۹	۰/۳۲۴	۰/۹۴۱	۰/۹۲۵	۰/۹۵۴	۰/۳۵۴	۰/۹۴۶	۰/۹۳۶
$\beta_۱$								
$Bias$	۰/۰۱۱	−۰/۰۵۹	۰/۰۱۲	−۰/۰۱۶	۰/۰۱۰	−۰/۰۴۹	۰/۰۱۰	−۰/۰۱۵
$S.E.$	۰/۰۴۴	۰/۰۵۴	۰/۰۵۳	۰/۰۴۸	۰/۰۴۳	۰/۰۵۱	۰/۰۵۲	۰/۰۴۵
CP	۰/۹۵۲	۰/۴۵۱	۰/۹۳۴	۰/۹۳۵	۰/۹۴۶	۰/۶۵۰	۰/۹۴۵	۰/۹۳۹
$\beta_۲$								
$Bias$	۰/۰۳۰	−۰/۰۵۲	−۰/۰۱۲	۰/۰۱۵	۰/۰۲۹	−۰/۰۵۰	−۰/۰۱۲	۰/۰۱۴
$S.E.$	۰/۰۴۳	۰/۰۵۲	۰/۰۵۲	۰/۰۴۵	۰/۰۳۹	۰/۰۵۲	۰/۰۴۹	۰/۰۴۱
CP	۰/۹۴۹	۰/۵۰۵	۰/۹۴۳	۰/۹۳۹	۰/۹۵۴	۰/۶۴۸	۰/۹۳۹	۰/۹۳۶

چهار مدل را بر روی داده‌ها برازش داده می‌شود. با برازش چهار مدل رگرسیونی پواسون، دوجمله‌ای منفی، پواسون آماسیدهٔ صفر و دوجمله‌ای منفی آماسیدهٔ صفر با گم‌شدگی در متغیرهای تبیینی برروی داده‌های مثال واقعی که در متغیر پاسخ آن تعداد زیادی صفر مشاهده شده است به کمک معیار AIC استدلال می‌شود.

همان‌طور که در جدول ۴ ملاحظه می‌شود، مدل رگرسیونی دوجمله‌ای منفی آماسیدهٔ صفر که نسبت به مدل‌های دیگر دارای کمترین مقدار AIC است، مدل برتر محسوب می‌شود. این مدل به صورت

$$
\begin{aligned}
&Y_i \sim ZINB(\theta_i, m, \omega_i),\\
&\text{logit}(\omega_i) = \gamma_۰ + \gamma_۱ C.child_i + \gamma_۲ C.persons_i + \gamma_۳ lider_i,\\
&\log \theta_i = \beta_۰ + \beta_۱ C.child_i + \beta_۲ C.persons_i + \beta_۳ lider_i,
\end{aligned}
$$

در نظر گرفته می‌شود، که در آن متغیر نشانگر δ_i و متغیر W_i به عنوان متغیر جانشین برای $C.persons_i$

جدول ۲: نتایج شبیه‌سازی برای حالت (ب) و $m = ۱۰$

حجم نمونه	$n = ۱۰۰۰$				$n = ۲۰۰۰$			
پارامتر	$\widehat{\eta}_F$	$\widehat{\eta}_{CC}$	$\widehat{\eta}_W$	$\widehat{\eta}_{Ws}$	$\widehat{\eta}_F$	$\widehat{\eta}_{CC}$	$\widehat{\eta}_W$	$\widehat{\eta}_{Ws}$
$\gamma_۰$								
$Bias$	۰/۰۱۱	−۰/۹۴۴	۰/۰۴۵	۰/۰۴۵	−۰/۰۱۱	−۰/۸۵۴	۰/۰۴۲	۰/۰۳۰
$S.E.$	۰/۲۰۱	۰/۳۱۹	۰/۳۳۲	۰/۳۰۲	۰/۲۰۰	۰/۳۱۴	۰/۳۱۵	۰/۲۰۰
CP	۰/۹۵۵	۰/۵۵۰	۰/۹۴۵	۰/۹۳۸	۰/۹۴۹	۰/۰۰۲	۰/۹۴۲۲	۰/۹۴۰
$\gamma_۱$								
$Bias$	−۰/۰۴۰	−۰/۸۰۳	−۰/۰۷۲	۰/۰۷۵	−۰/۰۴۰	−۰/۸۰۰	۰/۰۷۰	۰/۰۷۲
$S.E.$	۰/۳۱۴۳	۰/۳۳۲	۰/۳۳۳	۰/۳۱۵	۰/۳۱۶	۰/۳۳۲	۰/۳۳۲	۰/۳۱۵
CP	۰/۹۳۹	۰/۰۰۸	۰/۹۳۰	۰/۸۹۵	۰/۹۲۶	۰/۰۰۲	۰/۹۲۱	۰/۹۱۶
$\gamma_۲$								
$Bias$	−۰/۰۰۹	−۱/۰۰۳	−۰/۰۲۱	−۰/۰۳۷	۰/۰۰۳	۰/۹۹۱	−۰/۰۱۱	۰/۰۲۲
$S.E.$	۰/۳۵۹	۰/۳۹۵	۰/۳۹۶	۰/۳۵۸	۰/۳۴۴	۰/۳۸۷	۰/۳۸۹	۰/۳۵۵
CP	۰/۹۵۳	۰/۳۸۱	۰/۹۴۰	۰/۹۳۹	۰/۹۳۷	۰/۰۰۲	۰/۹۴۵	۰/۹۳۹
$\beta_۰$								
$Bias$	−۰/۰۱۳	۰/۴۰۹	−۰/۰۲۰	۰/۰۳۰	−۰/۰۱۱	۰/۴۰۸	۰/۰۲۰	۰/۰۲۸
$S.E.$	۰/۰۴۸	۰/۰۵۸	۰/۰۵۹	۰/۰۴۵	۰/۰۴۶	۰/۰۵۶	۰/۰۵۱	۰/۰۴۰
CP	۰/۹۵۰	۰/۰۳۰	۰/۹۳۷	۰/۸۷۶	۰/۹۳۴	۰/۰۰۱	۰/۹۴۰	۰/۹۰۱
$\beta_۱$								
$Bias$	۰/۰۱۱	−۰/۲۰۱	−۰/۰۲۱	۰/۰۲۸	۰/۰۱۰	۰/۰۱۹۸	۰/۰۲۰	۰/۰۲۵
$S.E.$	۰/۰۳۴	۰/۰۴۸	۰/۰۴۶	۰/۰۳۷	۰/۰۳۳	۰/۰۴۰	۰/۰۴۱	۰/۰۳۴
CP	۰/۹۴۲	۰/۵۰۰	۰/۹۴۴	۰/۸۶۶	۰/۹۳۳	۰/۰۱۵	۰/۹۴۱	۰/۹۱۳
$\beta_۲$								
$Bias$	۰/۰۱۰	−۰/۲۱۵	۰/۰۱۰	۰/۰۱۶	۰/۰۱۱	−۰/۲۰۴	۰/۰۱۰	۰/۰۱۲
$S.E.$	۰/۰۴۳	۰/۰۵۵	۰/۰۵۶	۰/۰۴۳	۰/۰۴۲	۰/۰۵۴	۰/۰۵۵	۰/۰۵۰
CP	۰/۹۴۰	۰/۰۵۰	۰/۹۳۵	۰/۹۲۲	۰/۹۴۴	۰/۰۰۱	۰/۹۳۰	۰/۹۲۵

جدول ۳: شاخص‌های مربوط به متغیر پاسخ

متغیر	مقادیر	فراوانی	درصدفراوانی
تعداد ماهی‌ها	۰	۱۱۸	۰/۴۷
	۱	۴۶	۰/۱۸
	$\geqslant ۲$	۸۶	۰/۳۵

به صورت

$$\delta_i = \begin{cases} ۱ & C.persons_i \text{ مشاهده شود} \\ ۰ & \text{در غیر این‌صورت} \end{cases} \qquad W_i = \begin{cases} ۱ & C.persons_i \geqslant ۲/۵ \\ ۰ & C.persons_i < ۲/۵ \end{cases}$$

جدول ۴: شاخص‌های مربوط به متغیرهای تبیینی

متغیر	درصدگم‌شدگی	بیشترین مقدار	کمترین مقدار	میانگین	انحراف معیار
تعداد بزرگسالان	۰/۲۹	۴	۱	۲/۵۱	۰/۱۴۷
تعداد کودکان	۰	۳	۰	۰/۶۸	۰/۰۲۴
تعداد راهنما	۰	۱	۰	۰/۲۴	۰/۰۰۳

جدول ۵: معیار AIC برای مدل‌های مختلف با گم‌شدگی در متغیرهای تبیینی منفی آماسیدهٔ صفر

مدل رگرسیونی	پواسون	دوجمله‌ای منفی	پواسون آماسیدهٔ صفر	دوجمله‌ای منفی آماسیدهٔ صفر
AIC	۶۵۴/۲	۷۱۱/۸۱	۵۴۲/۶	۴۳۲/۱

تعریف می‌شوند. در نهایت احتمال انتخاب تحت ساختار MAR به صورت

$$\begin{aligned}\text{logit}(\pi(Y_i, V_i)) &= \text{logit}P(\delta_i = ۱|Y_i, V_i)\\ &= \eta_۰ + \eta_۱ Count_i + \eta_۲ W_i + \eta_۳ C.child_i + \eta_۴ lider_i.\end{aligned}$$

فرض می‌شود. در داده‌های واقعی نسبت تعداد صفرهای مربوط به متغیر پاسخ در $n = ۲۵۰$ از داده‌های کامل برابر با ۰/۵۶ درصد و نسبت تعداد صفرهای مربوط به متغیر پاسخ در $n = ۲۵۰$ از مجموعه داده‌های ناکامل ۰/۱۵ درصد است. همچنین میزان گم‌شدگی متغیر تعداد بزرگسالان ۰/۲۹ درصد است. نتایج جدول ۵ نشان می‌دهد که میزان اریبی برای هر کدام از برآوردکننده $\widehat{\theta}_F$، $\widehat{\theta}_{CC}$، $\widehat{\theta}_W$ و $\widehat{\theta}_{Ws}$ وجود دارد.

به طور کلی تحلیل‌های داده‌های واقعی با داده‌های شبیه‌سازی نتایج یکسان را نشان می‌دهند و این موضوع را مشخص می‌کنند که روش برآورد نیم‌پارامتری برای مدل رگرسیونی دوجمله‌ای منفی آماسیدهٔ صفر با متغیرهای تبیینی گم‌شده یک روش مطلوب است و می‌تواند به عنوان یک شیوهٔ بامزیت به کار برده شود.

۶ بحث و نتیجه‌گیری

در این مقاله به بررسی و تحلیل نیم‌پارامتری مدل‌های رگرسیونی سری توانی آماسیدهٔ صفر مانند مدل رگرسیونی دوجمله‌ای منفی آماسیدهٔ صفر با و بدون داده‌های گم‌شده در متغیرهای تبیینی پرداخته شد. این

جدول ۶: برآورد پارامترهای مدل رگرسیون دوجمله‌ای منفی آماسیدهٔ صفر با امکان گم‌شدگی تصادفی در متغیرهای تبیینی

روش‌های برآوردیابی								
$\widehat{\Theta}_{Ws}$	$\widehat{\Theta}_{W}$	$\widehat{\Theta}_{CC}$	$\widehat{\Theta}_{F}$	$\widehat{\Theta}_{Ws}$	$\widehat{\Theta}_{W}$	$\widehat{\Theta}_{CC}$	$\widehat{\Theta}_{F}$	
$\beta_0(constant)$				$\gamma_0(constant)$				
۰/۰۹۳	۰/۱۹۶	۰/۵۶۱	۰/۰۴۹	۰/۱۹۰	۰/۲۳۰	−۰/۹۶۱	۰/۰۷۲	$Bias$
$\beta_1(C.child)$				$\gamma_1(C.child)$				
۰/۱۳۲	۰/۱۸۹	۰/۷۵۲	۰/۰۴۰	۰/۱۶۵	۰/۲۲۰	−۰/۸۶۰	۰/۰۵۱	$Bias$
$\beta_2(C.persian)$				$\gamma_2(C.persian)$				
۰/۱۶۱	۰/۱۵۱	۰/۲۱۹	۰/۰۳۶	۰/۱۹۰	۰/۲۳۵	−۰/۸۳۱	۰/۰۴۷	$Bias$
$\beta_3(lider)$				$\gamma_3(lider)$				
۰/۱۵۳	۰/۲۱۱	۰/۶۰۱	۰/۰۳۲	۰/۱۷۴	۰/۲۳۰	−۰/۸۳۰	۰/۰۴۲	$Bias$

تحلیل نیم‌پارامتری را می‌توان برای مدل‌های رگرسیونی آماسیدهٔ در یک نقطه غیر از صفر نیز مورد بررسی قرار داد. با تعمیم این روش‌ها روی مدل‌های رگرسیونی سری توانی آماسیدهٔ در بیش از یک نقطه، تحلیل‌های جدیدی مورد بررسی و به عنوان پژوهش‌های آتی مورد توجه قرار خواهد گرفت. هم‌چنین با در نظر گرفتن نوع مکانیسم گم‌شدگی غیر تصادفی، تحلیل‌های نیم‌پارامتری جدیدی را می‌توان برای مدل‌های بیان شده ارایه داد.

۷ تقدیر و تشکر

از داوران و ویراستاران محترم مجله که با توصیه‌های بسیار مفید سبب ارتقای این مقاله و ارایه بهتر آن شده‌اند کمال تشکر را داریم.

مراجع

اسماعیل‌زاده، م. بهرامی سامانی، ا.(۱۳۹۷)، مدل‌های اثر تصادفی دومتغیره آماسیده برای پاسخ‌های آمیخته سری توانی نرمال، مجله علوم آماری، مقاله آماده انتشار.

Lambert, D. (1992), Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, *Technometrics*, **34**, 1-14.

Greene, W. H. (1994), Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models, *Leonard N, Department of Economics, New York.*

Jansakul, N., and Hinde, J. P. (2002), Score Tests for Zero-Inflated Poisson Models, *Computational Statistics and Data Analysis*, **40** , 75-96.

Rubin, D. B. (1976), Inference and Missing Data, *Biometrika*, **63**, 581-592.

Little, R. J., and Rubin, D. B. (2014), *Statistical Analysis with Missing Data*, John Wiley and Sons, New York.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., and Herring, A. H. (2005), Missing-Data Methods for Generalized Linear Models: A Comparative Review, *Journal of the American Statistical Association*, **100**, 332-346.

Chen, X. D., and Fu, Y. Z. (2011), Model Selection for Zero-Inflated Regression with Missing Covariates, *Computational Statistics and Data Analysis*, **55**, 765-773.

Mason, A., Best, N., Richardson, S., and PLEWIS, I. (2010), Strategy for Modelling Non-Random Missing Data Mechanisms in Observational Studies using Bayesian Methods, *Journal of Official Statistics*, **28** , 279-302.

Wang, C. Y., Wang, S., Zhao, L. P., and Ou, S. T. (1997), Weighted Semiparametric Estimation in Regression Analysis With Missing Covariate Data, *Journal of the American Statistical Association*, **92**, 512-525.

Linton, O., and Nielsen, J. P. (1995), A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration, *Biometrika*, **82**, 93-100.

Lukusa, T. M., Lee, S. M., and Li, C. S. (2016), Semiparametric Estimation of a Zero-Inflated Poisson Regression Model with Missing Covariates, *Metrika*, **79**, 457-483.

Horvitz, D. G., and Thompson, D. J. (1952), A Generalization of Sampling without Replacement from a Finite Universe, *Journal of the American Statistical Association*, **47**, 663-685.

Zhao, L. P.,and Lipsitz, S. (1992), Designs and Analysis of Two Stage Studies, *Statistics in Medicine*, **11**, 769-782.

Breslow, N. E., and Cain, K. C. (1988), Logistic Regression for Two-Stage Case-Control Data, *Biometrika*, **75** , 11-20.

Wang, C. Y., Wang, S., Zhao, L. P., and Ou, S. T. (1997),Weighted Semiparametric Estimation in Regression Analysis with Missing Covariate Data, *Journal of the American Statistical Association*, **75**, 11-20.