سیزدهمین کنفرانس آمار ایران
۴-۲ شهریور ۱۳۹۵

کاهش اریبی برآوردگرهای مدل نیمه‌پارامتری بقا به کمک خودگردان‌سازی

احسان اسحقی[۱] ، حسین باغیشنی[۲]

[۱]دانشگاه علوم پزشکی بقیه‌الله

[۲]دانشگاه صنعتی شاهرود

چکیده: برای تحلیل داده‌های بازگشتی بقا، مدل‌های مختلفی پیشنهاد شده‌اند. با توجه به ماهیت بازگشتی بودن این نوع داده‌ها در طول زمان، رده مدل‌های نیمه‌پارامتری که شامل ضرایب وابسته به زمان هستند، بسیار مفید و منعطف می‌باشند. برآوردگرهای پارامترها در این مدل‌ها معمولا صورت بسته ندارند و لازم است برای محاسبه آن‌ها از تکرار معادله‌های برآوردیاب، به ازای هر نقطه زمانی، استفاده کرد. یک نتیجه معمول این نوع برآوردیابی، ناهمواری برآوردگرهاست و برای حل این مشکل می‌توان از روش هموارسازی هسته استفاده کرد. به دلیل اریبی القاشده به برآوردگرهای حاصل به واسطه اریبی ذاتی روش هسته در کران‌ها، به‌ویژه در نمونه‌های کوچک، در این مقاله از روش خودگردان‌سازی پارامتری برای کاهش آن استفاده می‌کنیم. در یک مطالعه شبیه‌سازی عملکرد روش پیشنهادی را تشریح خواهیم کرد.

واژه‌های کلیدی: خودگردان‌سازی، داده‌های بازگشتی، مدل نیمه‌پارامتری بقا.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62F40، 62N86.

۱ مقدمه

در بسیاری از علوم، گاهی اوقات یک پیشامد در طول زمان ممکن است بیش از یک بار رخ دهد. این قبیل پیشامدها را پیشامدهای بازگشتی [۱] می‌گویند. برای مثال، برخی بیماری‌ها مانند ایدز که عفونت‌هایی تکرارشونده دارند یا خرابی‌های پی در پی یک سیستم در خط تولید یک دستگاه، نمونه‌هایی از این نوع داده‌ها هستند. یکی از ویژگی‌های این نوع داده‌ها وابستگی پیشامدهای متفاوت برای یک فرد است. برای مدل‌بندی پیشامدهای بازگشتی، استفاده از مدل‌های رگرسیونی نیمه‌پارامتری یک انتخاب مناسب است. مدل‌های نیمه‌پارامتری با داشتن ضرایب وابسته و مستقل از زمان در خود، در مدل‌بندی پیشامدهای بازگشتی موثر و منعطف هستند؛ زیرا در واقعیت، به دلیل ماهیت بازگشتی بودن پیشامدها، ممکن است هر دو عوامل وابسته و مستقل از زمان بر آن‌ها موثر باشند.

[۱]احسان اسحقی : ehsan_eshaghi@yahoo.com

[۱] Recurrent events

مدل‌های نیمه‌پارامتری مختلفی برای تحلیل داده‌های بازگشتی به‌عنوان زیرده‌هایی از مدل‌های شرطی و حاشیه‌ای پیشنهاد شده‌اند. در تحلیل بقا مدل‌های رگرسیونی تعریف‌شده بر حسب تابع مخاطره شرطی، به‌طور گسترده برای توصیف وابستگی زمان‌های بقا به متغیرهایی تبیینی مورد استفاده قرار می‌گیرند. مدل مخاطره نسبی کاکس[۲]، یکی از معروف‌ترین این مدل‌هاست. در عمل، مدل‌بندی میانگین تعداد پیشامدها در مقابل مدل مخاطره نسبی کاکس، برای این داده‌های بازگشتی قابل فهم‌تر است (سان و همکاران (۲۰۱۱)). برخی از محققان از مدل‌های رگرسیونی مبتنی بر توابع میانگین و نرخ استفاده کرده‌اند. لین و همکاران (۲۰۰۰) مدلی را برای تابع میانگین و نرخ حاشیه‌ای بر اساس یک تابع پیوند از نوع کاکس معرفی کردند و استنباط‌ها را با فرض پیوستگی زمان تعمیم دادند. همچنین مارتینسون و همکاران (۲۰۰۲) مطالعاتی را برای برآورد ضرایب وابسته و مستقل از زمان در مدل‌های کاکس انجام دادند. اخیرا نیز سان و همکاران (۲۰۱۱)، یک مدل رگرسیونی حاشیه‌ای را با ترکیبی از ضرایب وابسته و مستقل از زمان برای تحلیل داده‌های بازگشتی پیشنهاد کرده‌اند.

تاکنون کاربردهای روش‌های بازنمونه‌گیری[۳]، مانند خودگردان‌سازی[۴]، در تحلیل داده‌های بازگشتی چندان مورد توجه و مطالعه قرار نگرفته است. گنزالز و پنا (۲۰۰۳) روش خودگردان‌سازی را برای برآورد توزیع نمونه‌ای برآوردگرهای میانه توزیع زمان، بین وقوع پیشامدها در داده‌های بازگشتی مورد استفاده قرار دادند. آناکلتو و لوزادا (۲۰۱۲) روشی بر مبنای خودگردان‌سازی برای برآورد فاصله اطمینان در داده‌های بازگشتی ارایه دادند.

برآوردگرهای پارامترها در مدل‌های نیمه‌پارامتری بقا معمولا صورت بسته ندارند و لازم است برای محاسبه آن‌ها از تکرار معادله‌های برآوردیاب تعمیم‌یافته[۵]، به ازای هر نقطه زمانی، استفاده کرد. نتیجه معمول این نوع برآوردیابی، ناهمواری برآوردگرهاست. برای حل این مشکل معمولا از روش‌های هموارسازی استفاده می‌شود. در این مقاله، ما از روش هسته[۶] برای اجرای این مرحله استفاده می‌کنیم. به دلیل اریبی القاشده به برآوردگرهای حاصل به واسطه اریبی ذاتی روش هسته در کران‌ها، به‌ویژه برای نمونه‌های با حجم کم، در این مقاله از روش خودگردان‌سازی پارامتری برای کاهش آن استفاده می‌کنیم. در بخش بعدی مدل نیمه‌پارامتری بقا (سان و همکاران (۲۰۱۱)) و روش برازش مدل را تشریح می‌کنیم. در بخش ۳ روش خودگردان‌سازی پارامتری را به اختصار معرفی می‌کنیم. کاهش اریبی حاصل از به‌کارگیری روش خودگردان‌سازی را نیز با یک مطالعه شبیه‌سازی در بخش ۴ مورد ارزیابی قرار می‌دهیم. در پایان بحث و نتیجه‌گیری خواهیم کرد.

۲ معرفی مدل

برای ساخت مدل، n نفر را در طول زمان مورد بررسی قرار می‌دهیم. فرض کنید $N_i^*(t)$ تعداد پیشامدهایی باشد که در طول بازه $[۰, t]$ برای فرد iام رخ می‌دهد. همچنین فرض کنید $X_i(\cdot)$ و $Z_i(\cdot)$ بردارهای فرآیندهایی به ترتیب p و q بعدی برای فرد iام باشند. در اکثر کاربردها، بازه زمانی مورد بررسی محدود است و ممکن است برخی از پیشامدها سانسور شوند. از این رو، $N_i^*(t)$ به‌طور کامل مشاهد نمی‌شود. اگر C_i را به عنوان زمان سانسور تعریف کنیم، آن‌گاه فرآیند قابل مشاهد به صورت $N_i(t) = N_i^*(t \wedge C_i)$ تعریف می‌شود، که در آن $a \wedge b = \min(a, b)$. فرض می‌کنیم با شرط داشتن $X_i(\cdot)$ و $Z_i(\cdot)$، C_i و $N_i^*(t)$ مستقل از هم هستند. تعریف می‌کنیم $Y_i(t) = I(C_i \geq t)$، به طوری که $I(\cdot)$ تابع نشان‌گر است. بنابراین، مجموعه داده مشاهده‌شده به صورت $\{N_i(t), Y_i(t), X_i(\cdot), Z_i(\cdot)\}$, $i = ۱, \cdots, n$ است.

[۲] Cox proportional hazard model
[۳] Resampling
[۴] Bootstrap
[۵] Generalized estimating equations
[۶] Kernel method

مدل رگرسیونی نرخ حاشیه‌ای[۷] پیشنهادی توسط سان و همکاران (۲۰۱۱) با ترکیبی از ضرایب وابسته و مستقل از زمان به صورت زیر تعریف می‌شود:

$$E\{dN_i^*(t)|\mathbf{X}_i(t),\mathbf{Z}_i(t)\} = \exp\{\boldsymbol{\beta}_\cdot(t)^T\mathbf{X}_i(t) + \boldsymbol{\gamma}_\cdot^T\mathbf{Z}_i(t)\}d\mu_\cdot(t), \qquad (۱.۲)$$

به طوری که $\boldsymbol{\beta}_\cdot(\cdot)$ بردار p بعدی ضرایب رگرسیونی وابسته به زمان، $\boldsymbol{\gamma}_\cdot$ بردار q بعدی ضرایب رگرسیونی مستقل از زمان و $\mu_\cdot(\cdot)$ تابع میانگین پایه‌ای است. فرض کنید

$$\mathbf{N}(t) = (N_1(t),\cdots,N_n(t))^T \quad X(t) = (\mathbf{X}_1(t),\cdots,\mathbf{X}_n(t))^T \quad Z(t) = (\mathbf{Z}_1(t),\cdots,\mathbf{Z}_n(t))^T,$$

و $N_\cdot(t) = n^{-1}\sum_{i=1}^n N_i(t)$. تعریف می‌کنیم

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\exp\{\boldsymbol{\beta}_\cdot(s)^T\mathbf{X}_i(s) + \boldsymbol{\gamma}_\cdot^T\mathbf{Z}_i(s)\}d\mu_\cdot(s), \quad i = 1,\ldots,n.$$

تحت مدل (۱.۲)، $M_i(t)$ها فرآیندهایی با میانگین صفر هستند. صفر بودن میانگین $M_i(t)$ ها با توجه به تعریف $Y_i(t)$ و مدل (۱.۲)، به سادگی قابل درک است. بنابراین به ازای مقادیر $\beta(t)$ و γ، برآوردگری منطقی برای $\mu_\cdot(t)$ از حل عبارت

$$\sum_{i=1}^n[dN_i(t) - Y_i(t)\exp\{\boldsymbol{\beta}_\cdot(t)^T\mathbf{X}_i(t) + \boldsymbol{\gamma}_\cdot^T\mathbf{Z}_i(t)\}d\mu_\cdot(t)] = 0, \quad 0 \le t \le \tau,$$

به‌دست می‌آید، که در آن τ مقدار تعیین‌شده‌ای است که $P(C_i \ge \tau) > 0$. در نتیجه برآوردگر $\mu_\cdot(t)$ به صورت

$$\hat{\mu}_\cdot(t;\boldsymbol{\beta}(t),\boldsymbol{\gamma}) = \int_0^t S_\cdot(t;\boldsymbol{\beta}(t),\boldsymbol{\gamma})^{-1}dN_\cdot(t),$$

محاسبه می‌شود.

۱.۲ برازش مدل

برای برآورد $\boldsymbol{\beta}_\cdot(t)$ و $\boldsymbol{\gamma}_\cdot$، با استفاده از معادله‌های برآوردیاب تعمیم‌یافته (لیانگ و زگر (۱۹۸۶)) و با استفاده از $\hat{\mu}_\cdot(t;\boldsymbol{\beta}_\cdot(t),\boldsymbol{\gamma}_\cdot)$، معادله‌های برآوردیاب زیر را داریم:

$$X(t)^T[d\mathbf{N}(t) - \boldsymbol{\phi}(t;\boldsymbol{\beta}(t),\boldsymbol{\gamma})S_\cdot(t;\boldsymbol{\beta}(t),\boldsymbol{\gamma})^{-1}dN_\cdot(t)] = 0, \quad 0 \le t \le \tau,$$
$$\int_0^\tau Z(t)^T[d\mathbf{N}(t) - \boldsymbol{\phi}(t;\boldsymbol{\beta}(t),\boldsymbol{\gamma})S_\cdot(t;\boldsymbol{\beta}(t),\boldsymbol{\gamma})^{-1}dN_\cdot(t)] = 0. \qquad (۲.۲)$$

معادله‌های (۲.۲) هر کدام معادله‌ای با دو مجهول هستند و بنابراین نمی‌توان به صورت بسته‌ای از برآوردگرها دست یافت. پس با به کار بردن روش‌های عددی نظیر نیوتون-رافسون به وسیله بسط تیلور تابع $\boldsymbol{\phi}(t;\boldsymbol{\beta}(t),\boldsymbol{\gamma})S_\cdot(t;\boldsymbol{\beta}(t),\boldsymbol{\gamma})^{-1}$ حول برآورد جاری $(\boldsymbol{\beta}^l(t),\boldsymbol{\gamma}^l)$، می‌توان برآوردهایی برای $\boldsymbol{\gamma}_\cdot$ و $\boldsymbol{\beta}_\cdot(t)$ به‌دست آورد. با استفاده از الگوریتم نیوتن-رافسون، معادله‌های به‌هنگام کننده به‌صورت زیر تعریف می‌شوند (سان و همکاران (۲۰۱۱)):

$$\{\boldsymbol{\beta}^{l+1}(t) - \boldsymbol{\beta}^l(t)\}S_\cdot^l(t)^{-1}dN_\cdot(t) = n^{-1}E_{xx}^l(t)^{-1}\{X(t) - \bar{X}^l(t)\}^T[d\mathbf{N}(t)$$

[۷] Marginal rate regression

$$-\Phi^l(t)\{Z(t)-\bar{Z}^l(t)\}(\boldsymbol{\gamma}^{l+1}-\boldsymbol{\gamma}^l)S^l_{\cdot}(t)^{-1}dN_{\cdot}(t)], \tag{۳.۲}$$

$$n^{-1}\int_{\cdot}^{\tau}\{Z(t)-\bar{Z}^l(t)\}^T d\mathbf{N}(t)-\int_{\cdot}^{\tau}E^l_{zx}(t)\{\boldsymbol{\beta}^{l+1}(t)-\boldsymbol{\beta}^l(t)\}S^l_{\cdot}(t)^{-1}dN_{\cdot}(t)$$
$$=\int_{\cdot}^{\tau}E^l_{zz}(t)(\boldsymbol{\gamma}^{l+1}-\boldsymbol{\gamma}^l)S^l_{\cdot}(t)^{-1}dN_{\cdot}(t), \tag{۴.۲}$$

با قرار دادن رابطه (۳.۲) در (۴.۲) و حل معادله بر اساس γ^{l+1}، به رابطه تکراری $\gamma^{l+1}=\Psi_r(\gamma^l)$ دست می‌یابیم به‌طوری که

$$\Psi_r(\boldsymbol{\gamma}^l)=\boldsymbol{\gamma}^l+\frac{A^l(\tau)^{-1}}{n}\int_{\cdot}^{\tau}[\{Z(t)-\bar{Z}^l(t)\}^T-E^l_{zx}(t)E^l_{xx}(t)^{-1}\{X(t)-\bar{X}^l(t)\}^T d\mathbf{N}(t)], \tag{۵.۲}$$

که در آن $A^l(\tau)=A(\tau;\boldsymbol{\beta}^l(t),\boldsymbol{\gamma}^l)$ و

$$A(\tau;\boldsymbol{\beta}(t),\boldsymbol{\gamma})=\int_{\cdot}^{\tau}[E^l_{zz}(t)-E^l_{zx}(t)E^l_{xx}(t)^{-1}E^l_{zx}(t)^T]S^l_{\cdot}(t)^{-1}dN_{\cdot}(t).$$

برای برآورد $\beta_{\cdot}(t)$ با استفاده از معادله (۳.۲)، تکرار معادله به ازای هر t باعث ناهمواری برآوردگر منحنی $\beta(t)$ می‌شود و حتی ممکن است موجب عدم سازگاری برآوردگر شود. بنابراین از روش هسته برای هموارسازی آن استفاده می‌کنیم. مارتینسون و همکاران (۲۰۰۲) و شیکه و مارتینسون (۲۰۰۴)، پیشنهاد کردند به جای استفاده از $\beta_{\cdot}(t)$ از ضرایب رگرسیونی تجمعی $B_{\cdot}(t)=\int_{\cdot}^{t}\beta_{\cdot}(s)ds$ استفاده کنیم که منجر به نتایج پایدارتری می‌شود.

فرض کنید $\mu^l_{\cdot}(t)=\hat{\mu}_{\cdot}(t;\beta^l,\gamma^l)$. همچنین $\beta^l(t)$ و $\lambda^l(t)$ به ترتیب برآوردگرهای هسته $\beta_{\cdot}(t)$ و $\lambda_{\cdot}(t)=\frac{d}{dt}\mu_{\cdot}(t)$ بر پایه $B^l(t)$ و $\mu^l_{\cdot}(t)$ با پهنای نوار[۸] h باشند. یعنی

$$\boldsymbol{\beta}^l(t)=\int h^{-1}K\left(\frac{u-t}{h}\right)\boldsymbol{\beta}^l(u)du,$$

و

$$\lambda^l_{\cdot}(t)=\int h^{-1}K\left(\frac{u-t}{h}\right)\lambda^l_{\cdot}(u)du,$$

که در آن $K(\cdot)$ یک تابع هسته متقارن است. با قرار دادن $\lambda^l_{\cdot}(t)dt$ در رابطه (۳.۲) به جای $\frac{dN_{\cdot}(t)}{S^l_{\cdot}(t)}$، داریم

$$\boldsymbol{\beta}^{l+1}(t)\lambda^l_{\cdot}(t)dt-\boldsymbol{\beta}^l(t)\lambda^l_{\cdot}(t)dt$$
$$=n^{-1}E^l_{xx}(t)^{-1}\{X(t)-\bar{X}^l(t)\}^T[d\mathbf{N}(t)-\Phi^l(t)\{Z(t)-\bar{Z}^l(t)\}(\boldsymbol{\gamma}^{l+1}-\boldsymbol{\gamma}^l)S^l_{\cdot}(t)^{-1}dN_{\cdot}(t)].$$

بنا به پایداری ضرایب رگرسیونی تجمعی، بنابراین با تقسیم طرفین بر $\lambda^l_{\cdot}(t)$ و انتگرال‌گیری، به رابطه تکراری $\mathbf{B}^{l+1}(t)=\Psi_b(\mathbf{B}^l)(t)$ دست می‌یابیم، به طوری که

$$\Psi_b(\mathbf{B}^l)(t)=\int_{\cdot}^{t}\boldsymbol{\beta}^l(u)du+n^{-1}\int_{\cdot}^{t}\lambda^l_{\cdot}(u)^{-1}E^l_{xx}(u)^{-1}\{X(u)-\bar{X}^l(u)\}^T$$
$$\times[d\mathbf{N}(u)-\Phi^l(u)\{Z(u)-\bar{Z}^l(u)\}(\boldsymbol{\gamma}^{l+1}-\boldsymbol{\gamma}^l)S^l_{\cdot}(u)^{-1}dN_{\cdot}(u)]. \tag{۶.۲}$$

با استفاده متوالی از معادله‌های تکراری (۵.۲) و (۶.۲)، برآوردگرهای $\mathbf{B}_{\cdot}(t)$ و $\boldsymbol{\gamma}_{\cdot}$ به‌هنگام می‌شوند. مراحل تا جایی ادامه می‌یابند که اختلاف بین برآوردها در دو مرحله متوالی، کمتر از یک آستانه مشخص کوچک، مثلا 10^{-6}، شود.

[۸] Bandwidth

۳ خودگردان‌سازی پارامتری

به دلیل استفاده از روش هسته برای هموارسازی برآوردگرهای پارامترهای مدل (۱.۲)، انتظار داریم یک اریبی به برآوردگرها تحمیل شود. دلیل آن هم وجود اریبی ذاتی القاشده توسط روش هسته به‌ویژه در کران‌ها است. از طرفی اگر حجم نمونه‌ها کم باشد، این اریبی در میانگین توان دوم خطا (MSE) واریانس را مغلوب می‌کند. بنابراین تصحیح این اریبی از اهمیت ویژه‌ای در به‌دست آوردن استنباط‌های دقیق برخوردار است. این تصحیح نیاز به شناخت توزیع برآوردگرها دارد. اما به دلیل پیچیدگی صورت برآوردگرهای حاصل، محاسبه توزیع نمونه‌ای آن‌ها ممکن نیست و باید از نسخه‌های تقریبی استفاده کرد. در آمار، برای تقریب توزیع‌های نمونه‌ای دو راهکار کلی وجود دارد: ۱) نظریه مجانبی توزیع‌ها و ۲) روش‌های بازنمونه‌گیری مانند خودگردان‌سازی. در راهکار اول، همگرایی کند به توزیع‌های حدی برای مدل‌های پیچیده استفاده از تقریب‌های حاصل را برای نمونه‌های کوچک و متوسط با تردید مواجه می‌کند. بنابراین، در این موارد، راهکار معمول استفاده از روش‌های بازنمونه‌گیری است.

در این مقاله، رهیافت منتخب ما استفاده از روش خودگردان‌سازی پارامتری است. در خودگردان‌سازی پارامتری، ابتدا مدل (۱.۲) را به داده‌ها برازش می‌دهیم. سپس نمونه‌های خودگردان را از مدل برازش‌شده شبیه‌سازی می‌کنیم. با تکرار این مکانیسم، مجموعه داده‌های خودگردان تولید می‌شوند که دارای توزیع مشابه داده‌های واقعی هستند. در مرحله بعدی برای هر نمونه خودگردان (مشابه داده‌های واقعی) مدل را برازش می‌دهیم و برآورد پارامترها را به‌دست می‌آوریم. با این عمل، تحققی از توزیع نمونه‌ای برآوردگرها (با حجمی برابر تعداد تکرارهای خودگردان) به‌دست می‌آید. در نهایت، توزیع نمونه‌ای برآوردگرها توسط توزیع تجربی تحقق نتیجه‌شده از آن‌ها تقریب زده می‌شود.

۴ مطالعه شبیه‌سازی

در مطالعه شبیه‌سازی، مدلی با یک ضریب وابسته به زمان و یک ضریب مستقل از زمان در نظر گرفتیم. در مدل مفروض، به ازای $i = ۱, ۰۰۰, n$، زمان‌های رخداد پیشامدها را از فرآیند پواسن با مدل حاشیه‌ای زیر تولید کردیم:

$$E\{dN_i^*(t)|X_i(t), Z_i(t)\} = \exp\{-۰/۵ + ۰/۵\sin(۲t)X_i(t) + ۰/۵Z_i(t)\}dt, \qquad (۱.۴)$$

به طوری که X_i از توزیع نرمال استاندارد، Z_i از توزیع برنولی با پارامتر ۰/۵ و زمان‌های سانسور C_i از توزیع یکنواخت در بازه $(۲, ۵)$ تولید شدند. با این تنظیمات، برای هر فرد به‌طور متوسط ۳ پیشامد بازگشتی شبیه‌سازی شد.

با تولید یک مجموعه داده از مدل (۱.۴)، به عنوان داده‌های اصلی، ابتدا ضرایب $\beta(t)$ و γ را برآورد کردیم. سپس با استفاده از روش خودگردان‌سازی پارامتری، ۱۰۰۰ مجموعه داده خودگردان را از مدل برازش‌شده شبیه‌سازی کردیم. برای این ۱۰۰۰ مجموعه، برآوردهای $\beta^{(\ell)}(t)$، $\gamma^{(\ell)}$ و در نتیجه $B^{(\ell)}(t)$، $\ell = ۱, \ldots, ۱۰۰۰$، را به دست آوردیم. در نهایت، میانگین ضرایب مستقل از زمان $\gamma^{(\ell)}$ و ضرایب تجمعی $B(t)^{(\ell)}$ را محاسبه کردیم. نتایج برای حجم‌های نمونه ۵۰، ۱۰۰، ۲۰۰ و ۴۰۰ برای ضرایب وابسته و مستقل از زمان به ترتیب در جدول‌های ۱ و ۲ گزارش شده‌اند.

جدول ۱، مقادیر اریبی جمع‌بسته توان دوم[۹] (ISB) و میانگین جمع‌بسته توان دوم خطاها[۱۰] (MISE) را برای برآورد ضریب رگرسیونی تجمعی واقعی $B.(\cdot)$ در مدل (۱.۴) با استفاده از تابع هسته اپانچنیکوف[۱۱] نشان می‌دهد. این دو کمیت به صورت‌های

$$ISB = \int_.^t (E(\hat{B}(s)) - B.(s))^۲ ds, \quad MISE = E(\int_.^t (\hat{B}(s) - B.(s))^۲) ds,$$

[۹] Integrated Squared Bias
[۱۰] Mean Integrated Squared Error
[۱۱] Epanechnikov

جدول ۱: نتایج شبیه‌سازی برای برآورد ضریب رگرسیونی وابسته به زمان در مدل (۱.۴)

n	h	ISB	MISE
۵۰	۰/۴	۰/۰۶۴۵	۰/۱۳۰۹
۱۰۰	۰/۴	۰/۰۲۹۳	۰/۰۷۰۹
۲۰۰	۰/۴	۰/۰۲۳۰	۰/۰۴۶۶
۴۰۰	۰/۴	۰/۰۰۶۴	۰/۰۱۷۳

جدول ۲: نتایج شبیه‌سازی برای برآورد ضریب رگرسیونی مستقل از زمان در مدل (۱.۴)

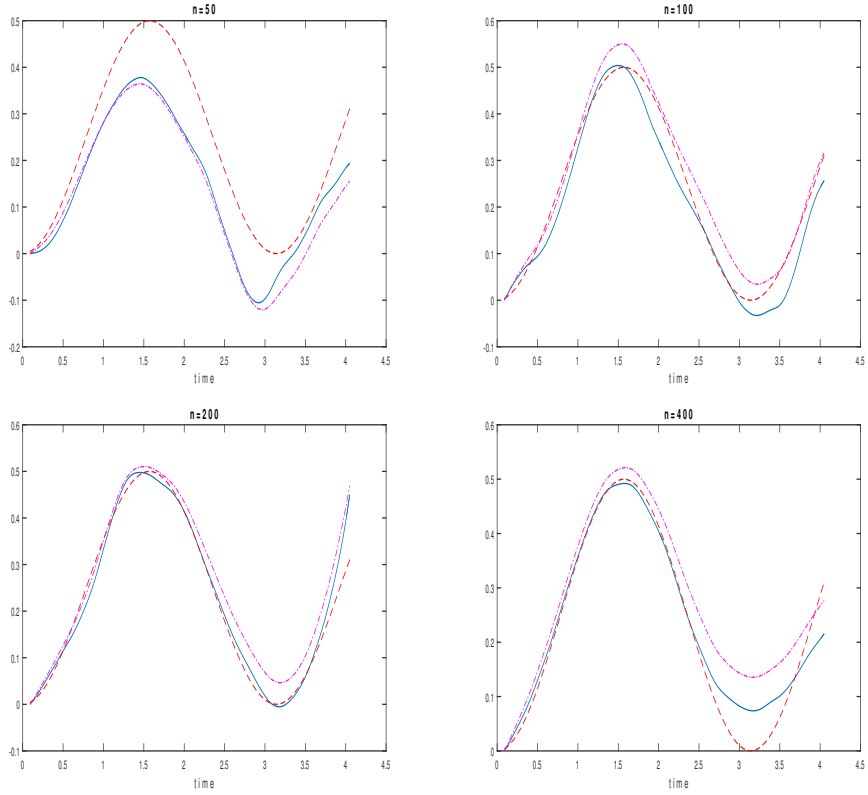
n	h	Bias	SSE	SEE
۵۰	۰/۴	۰/۰۷۳۴	۰/۰۲۴۰	۰/۰۴۴۸
۱۰۰	۰/۴	۰/۱۷۳۹	۰/۰۱۱۲	۰/۰۳۶۶
۲۰۰	۰/۴	۰/۱۶۰۷	۰/۰۰۴۴	۰/۰۲۲۲
۴۰۰	۰/۴	۰/۰۷۹۱	۰/۰۰۴۰	۰/۰۱۴۵

محاسبه می‌شوند. همان‌طور که در جدول ۱ مشهود است، با افزایش حجم نمونه عملکرد برآوردگر بهتر شده و مقادیر اریبی و خطا کاهش می‌یابند به‌طوری که برای حجم نمونه ۴۰۰، مقادیر ISB و MISE ناچیز بوده و تقریبا برآوردهای نااریب حاصل می‌شوند. اما برای حجم نمونه کوچک ۵۰ و متوسط ۱۰۰، میزان اریبی و MISE قابل صرفنظر کردن نیستند. این نتیجه در شکل ۱ نیز تایید می‌شود. جدول ۲ شامل اریبی (Bias) (میانگین نمونه‌ای برآوردهای $\hat{\gamma}$ منهای مقدار واقعی)، میانگین نمونه‌ای خطاهای استاندارد برآوردشده $\hat{\gamma}$ (SEE) و خطای استاندارد نمونه‌ای $\hat{\gamma}$ (SSE) می‌باشد که بر مبنای ۱۰۰۰ مجموعه داده شبیه‌سازی‌شده از مدل برازش‌شده، محاسبه شده‌اند. مانند نتایج جدول ۱، با افزایش حجم نمونه مقدار اریبی و خطای برآوردگر کاهش می‌یابد.

برای بررسی عملکرد نقطه‌ای برآوردگر $\hat{B}(t)$ ،$B.(t)$، برآوردهای ضرایب رگرسیونی تجمعی را در ۱۰۰ نقطه

$$t_k = ۰/۰۵ + ۰/۰۴k, \quad k = ۱, \ldots, ۱۰۰,$$

محاسبه کردیم. نمودار برآوردها برای چهار حجم نمونه ۵۰، ۱۰۰، ۲۰۰ و ۴۰۰ در شکل ۱ نمایش داده شده‌اند. در این نمودارها منحنی خط‌چین نشان دهنده تابع رگرسیونی تجمعی واقعی $B.(t) = \int_۰^t ۰/۵\sin(۲s)ds$، منحنی نقطه-خط‌چین برآورد ضریب تجمعی وابسته به زمان با استفاده از مجموعه داده اصلی و منحنی توپر، برآورد تصحیح اریبی‌شده $\hat{B}(t)$ به ازای ۱۰۰۰ نمونه خودگردان است. همان‌طور که در نمودارها مشخص است، برآورد تصحیح اریبی‌شده مبتنی بر روش خودگردان‌سازی در اکثر نقاط زمانی و در حجم نمونه‌های مختلف، اریبی کمتری نسبت به برآورد ضریب تجمعی وابسته به زمان با استفاده از مجموعه داده اصلی ارایه می‌دهد. همچنین، عملکرد استفاده از این روش با افزایش حجم نمونه، بهتر و منحنی برآورد به منحنی اصلی نزدیک‌تر می‌شود.



شکل ۱: برآورد ضرایب رگرسیونی تجمعی وابسته به زمان

بحث و نتیجه‌گیری

مدل‌های نیمه‌پارامتری بقا به دلیل در نظر گرفتن عوامل وابسته و مستقل از زمان در توجیه رخداد پیشامدها دارای انعطاف و نتایج دقیقی می‌باشند. هر چند این برآوردگرها از دقت قابل قبولی برخوردار هستند و از طرفی دقت آن‌ها با افزایش حجم نمونه افزایش می‌یابد (اسحقی و همکاران (۱۳۹۲))، اما در نمونه‌های کوچک دارای اریبی قابل ملاحظه‌ای هستند که عمده این اریبی را از اریبی ذاتی روش هسته به ارث می‌برند. در این مقاله نشان دادیم که با استفاده از روش خودگردان‌سازی می‌توان اریبی برآوردگرهای حاصل را کاهش داد و استنباط‌های دقیق‌تری گزارش کرد.

مراجع

اسحقی، ا.، باغیشنی، ح. و شاهسونی، د. (۱۳۹۲). برازش مدل‌های نیمه‌پارامتری بقا با اثرات وابسته به زمان برای داده‌های بازگردنده با روش هسته، *مجله علوم آماری* ۷، ۲۴-۱.

Anacleto O. and Louzada F. (2012), Bootstrap confidence intervals for industrial recurrent event data,

Pesquisa Operacional 32, 103 - 120.

González J.R. and Pena E.A. (2003), Bootstraping median survival with recurrent event data, *IX Conferencia Española de Biometría*, 28 - 30.

Liang K.Y. and Zeger S.L. (1986), Longitudinal data analysis using generalized linear models, *Biometrika* 73, 13 - 22.

Lin D.Y., Wei L.J., Yang I. and Ying Z. (2000), Semiparametric regression for the mean and rate function of recurrent events, *Journal of the Royal Statistical Society, Series B* 62, 711 - 730.

Martinussen T., Scheike T.H. and Skovgaard I.M. (2002), Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models, *Scandinavian Journal of Statistics* 29, 57 - 74.

Scheike T.H. and Martinussen T. (2004), On estimation and tests of time-varying effects in the proportional hazard models, *Scandinavian Journal of Statistics* 31, 51 - 62.

Sun L., Zhou X. and Guo S. (2011), Marginal regression models with time-varying coefficients for recurrent event data, *Statistics in Medicine* 30, 2265 - 2277.