یک مقایسه بین برآوردگرها در داده‌های طولی با پیگیری نامنظم و وابسته به برآمد

امیدعلی آقابابائی - مجتبی گنجعلی

گروه آمار، دانشگاه شهید بهشتی

چکیده: آزمودنی‌ها در مطالعات طولی معمولاً در یک مجموعهٔ مشترک از زمان‌های ملاقات از پیش تعیین شده مورد اندازه‌گیری قرار می‌گیرند. با این وجود در عمل، آزمودنی‌ها معمولاً زمان‌های ملاقات شده را لغو و یا در بین زمان‌های برنامه‌ریزی شده مورد بررسی قرار می‌گیرند که در نتیجه ساختار داده‌ها شدیداً نامتعادل‌می‌شود. به‌علاوه زمان‌های مشاهده ممکن است مستقیماً به اندازه برآمد یا بعضی متغیرهای کمکی که وابسته به برآمد هستند، وابسته باشد. همچنین در بسیاری از موارد، نمی‌توان در مدل میانگین پاسخ از فرم پارامتری برای تابع عرض از مبدأ استفاده کرد، بنابراین از مدل رگرسیونی نیم‌پارامتری بایستی استفاده کرد. در این مقاله با تعمیم مطالعهٔ بوزکوا و لاملی (۲۰۰۸) علاوه بر مقایسهٔ برآوردگر وزنی و برآوردگر تعدیل نشده تحت مدل رگرسیونی نیم‌پارامتری برای داده‌های طولی با پیگیری نامنظم و وابسته به برآمد، به مقایسهٔ این دو برآوردگر نیم‌پارامتری با برآوردگر پارامتری متداول حاصل از روش معادلات برآوردگر تعمیم یافته خواهیم‌پرداخت. همچنین عملکرد برآوردگرهای مذکور را برای داده‌هایی از آزمایش بالینی ایدز با نهایت عدم رعایت در زمان‌های ملاقات برنامه‌ریزی شده ارزیابی خواهیم کرد.

واژه‌های کلیدی: معادلات برآوردگر، داده‌های طولی، فرایند شمارشی، پیگیری نا-منظم، پیگیری وابسته به برآمد، رگرسیون نیم‌پارامتری، آزمایش بالینی ایدز

۱ مقدمه

مطالعات طولی، اندازه‌گیری تکراری واحدها در طول زمان به منظور شناسایی تغییرات متغیر پاسخ در طول زمان و عوامل مؤثر بر این تغییرات است. یکی از حالات متداول در هنگام جمع‌آوری داده‌های طولی، پیگیری نامنظم آزمودنی‌ها است؛ به طوری که در بسیاری از موارد، زمان‌های مشاهدهٔ آزمودنی‌ها با زمان‌های برنامه‌ریزی شده مطابقت ندارند. به عبارت دیگر آزمودنی‌ها ممکن است به منظور اندازه‌گیری متغیر پاسخ، در

بین زمان‌های برنامه‌ریزی شده حضور یابند و یا حتی ملاقات مربوطه را لغو نمایند. این مسئله باعث نامتعادل شدن ساختار داده‌ها می‌گردد. مسئلهٔ دیگر امکان وجود ارتباط بین زمان و فراوانی ملاقات آزمودنی‌ها و مقدار متغیر پاسخ گذشته یا متغیرهای وابسته به آن است. هنگامی که زمان‌های مشاهده گسسته و از پیش تعیین شده هستند، پیگیری وابسته به برآمد معادل آگاهی‌بخش بودن مکانیسم گمشدن خواهد بود. بنابراین همانطور که ملاحظه خواهد شد، با توجه به شرط کاملاً تصادفی بودن مکانیسم گمشدن داده‌ها در معادلات برآوردگر تعمیم یافته (GEE)، استفاده از این روش برای برآورد پارامترها باعث ایجاد اریبی در برآوردگرها می‌شود.

لین و پِنگ (۲۰۰۱) از مدل شرطی برای ادغام دو مدل که به صورت حاشیه‌ای مدل‌بندی شده‌اند، استفاده کردند؛ به طوری که مدل‌های میانگین پاسخ و زمان مشاهده از طریق متغیرهای کمکی مدل میانگین پاسخ ارتباط داده شدند. با این وجود در مدل آن‌ها، متغیر پاسخ و زمان‌های مشاهده به شرط متغیرهای کمکی مدل میانگین پاسخ، مستقل در نظر گرفته شدند. به عبارت دیگر زمان‌های مشاهده ناآگاهی‌بخش فرض شدند. لیپ‌شیتس و همکاران (۲۰۰۲) به منظور مدل‌بندی توأم مدل پاسخ و زمان‌های مشاهده تابع درستنمایی را به دو مؤلفهٔ فرایند پاسخ و فرایند زمان‌های مشاهده تقسیم کردند و درستنمایی زمان سپری شده بین دو مشاهده را محاسبه و به منظور چشم‌پوشی از مؤلفهٔ دوم، از فرض نسبتاً قوی در مورد وابستگی مدل زمان مشاهده و سابقهٔ اندازه های مکرر پاسخ‌های مشاهده شده استفاده کردند. لین و همکاران (۲۰۰۴) با استفاده از برآوردگر حاصل از مدل شدّت ملاقات، وزن‌هایی را به دست آوردند که با استفاده از معکوس آن به عنوان وزن در معادلات برآوردگر تعمیم‌یافته، اریبی برآوردگرهای مدل میانگین پاسخ تصحیح می‌شود. سان و همکاران (۲۰۰۷) یک مدل رگرسیون خطی با عرض از مبدأ نامعلوم را تحت زمان‌های مشاهدهٔ آگاهی‌بخش مورد مطالعه قرار دادند. آنها مدل برآمد را با استفاده از متغیر پنهان، به مدل زمان‌های مشاهده پیوند دادند. در نهایت بوزکوا و لاملی (۲۰۰۸) با تعمیم روش لین و پِنگ (۲۰۰۱) به تحلیل داده‌های طولی با زمان‌های مشاهدهٔ نامنظم و وابسته به برآمد پرداختند.

در این مقاله ضمن بررسی عملکرد برآوردگرهای پارامتری و نیم‌پارامتری در شرایط مختلف، لزوم استفاده از برآوردگر تعدیل یافته در داده‌های طولی با پیگیری وابسته به برآمد نشان داده خواهد شد. در بخش بعد، مدل‌های رگرسیونی نیم‌پارامتری میانگین پاسخ و زمان‌های مشاهده و ساختار معادلات برآوردگر وزنی معرفی خواهد شد. در بخش آخر، ضمن مقایسهٔ برآوردگر وزنی حاصل با برآوردگر غیروزنی و برآورد گر پارامتری حاصل از روش GEE، به تحلیل داده‌های مربوط به مقایسهٔ روش‌های درمان بیماری ایدز خواهیم پرداخت.

۲ ساختار مدل‌های رگرسیونی نیم‌پارامتری میانگین پاسخ و زمان مشاهده

استفاده از روش‌های پارامتری در مطالعات طولی، بسیار مفید و متداول هستند. با این وجود این روش‌ها نیازمند تعیین پارامتری تابع میانگین پایه (عرض از مبدأ) برای متغیر پاسخ پارامتری است که در عمل کار ساده‌ای نیست. به علاوه ضعف اصلی در استفاده از این روش‌ها، محدودیت و عدم انعطاف‌پذیری آن‌ها در بسیاری از مطالعات است؛ به طوری که در بسیاری از موارد، مدل‌های کاملاً پارامتری قادر به تعیین صحیح نحوهٔ ارتباط متغیر پاسخ و متغیرهای کمکی نیستند. از این رو استفاده از روش‌های نیم-پارامتری، با وجود محاسبات و روش‌های تحلیل پیچیده‌تر، به طور روزافزون در حال گسترش است.

روش‌های رگرسیون نیم‌پارامتری و ناپارامتری برای داده‌های طولی بر خلاف داده‌های مقطعی، به طور قابل قبولی گسترش نیافته‌اند. دلیل اصلی این مسئله، وجود همبستگی در پاسخ‌های هر آزمودنی در داده‌های طولی است؛ به طوری که وجود این همبستگی، چالش اصلی در پیشرفت روش‌های اسپلاین و هسته‌ای در داده‌های طولی است. هم-چنین روش‌های درستنمایی موضعی بر حسب روش هموارسازی هسته‌ای، همبستگی داخل آزمودنی‌ها را به طور مؤثر محاسبه نمی‌کنند. به علاوه تعمیم روش‌های هموار-سازی اسپلاین برای داده‌های‌طولی، مستلزم محاسبهٔ ضمنی همبستگی‌داخلی آزمودنی-ها در تابع درستنمایی تاوانیده است (لین و کارُل، ۲۰۰۰).

برای مدل میانگین پاسخ در این مقاله، از مدل میانگین کاملاً حاشیه‌ای برای مدل‌بندی متغیر پاسخ $Y(t)$ و متغیرهای کمکی $X(t)$ در بازهٔ زمانی $[\circ, \tau]$ استفاده می‌شود. مقدار ثابت و از قبل تعیین شدهٔ τ، به عنوان پایان مطالعه یا زمانی که آخرین آزمودنی از ادامهٔ مطالعه انصراف می‌دهد، در نظر گرفته می‌شود. بنابراین مدل میانگین پاسخ به صورت

$$E\left(Y_i(t)\,|X_i(t)\right) = \alpha_\circ(t) + \boldsymbol{\beta}_\circ^T X_i(t), \quad \forall t \in [\circ, \tau], \qquad (۱)$$

است که در آن تابع عرض از مبدأ $\alpha_\circ(t)$، پارامتر مزاحم با بعد نامتناهی است. در واقع این تابع برابر میانگین متغیر پاسخ با فرض صفر بودن متغیرهای کمکی است.

در مدل میانگین پاسخ (۱)، که به صورت یک مدل فرایندْ مانند است، در هر زمان از بازهٔ $[\circ, \tau]$، میانگین متغیر پاسخ به شرط متغیرهای کمکی در آن زمان مدل‌بندی می‌شود. بنابراین با توجه به مدل‌بندی کاملاً حاشیه‌ای متغیر پاسخ، فرضی دربارهٔ توزیع فرایند پاسخ $\{Y(t) : t \in [\circ, \tau]\}$ در نظر گرفته نمی‌شود. به علاوه، در این روش نیازی به تعیین ساختار همبستگی پاسخ‌های هر آزمودنی نیست.

زمان‌های اندازه‌گیری متغیر پاسخ مورد نـــظر یا همان زمان‌های مشاهدهٔ آزمودنی به صورت یک مجموعه از پیشامدها، با تعداد تصادفی عضو در نظر گرفته می‌شود. این مجموعه برای آزمودنی iاُم، $i \in \{۱,...,n\}$، به صورت $\{T_{i۱}, T_{i۲}, ..., T_{iK_i}\}$ است که در آن متغیر تصادفی K_i، تعداد دفعات مشاهدهٔ آزمودنی iاُم است.
فرایند شمارشی تعداد مشاهدات تا زمان t برای آزمودنی iاُم را به صورت

$$N_i(t) = \sum_{k=۱}^{K_i} I(T_{ik} \leq t),$$

تعریف می‌کنیم که در آن $I(A)$، تابع نشانگر پیشامد A است.
فرایند شمارشی $\{N_i(t) : t \in [۰, \tau]\}$، یک فرایند از راست پیوسته، غیرنزولی و پله‌ای با پرش‌های برابر با یک است. متغیر پاســـخ $Y_i(t)$ تنها در نقاط پرش فرایند $N_i(t)$ مشاهده می‌شود. بنابراین با توجه به امکان وجود داده‌های ناکامل در پاسخ‌های هر آزمودنی، بایستی مسئلهٔ سانسورشدن از سمت راست در محاسبات مدل اِعمال شود. بر این منظور، از متغیر زمان انصراف یا پایان پیگیری C_i در مدل‌بندی زمان‌های مشاهدهٔ آزمودنی iاُم استفاده می‌شود. حال به همراه فرایند شمارشی $N_i(t)$، که نشان دهندهٔ تعداد زمان های مشاهدهٔ آزمودنی i اُم تا زمان t است، از فرایند شمارشی مقدماتی $\{N_i^*(t), t \in [۰, \tau]\}$ که نشان‌دهندهٔ عدم وجود سانسور در داده‌ها است، استفاده‌می شود.
برای مدل‌بندی زمان‌های مشاهده از تابع نرخ حاشیه‌ای استفاده می‌شود. بر این اساس، مدل زمان‌های مشاهدهٔ مقدماتی برای هر آزمودنی i، $i \in \{۱,...,n\}$، در هر زمان t، $t \in [۰, \tau]$، به صورت

$$E[dN_i^*(t) | Z_i(t)] = \exp\{\gamma_۰^T Z_i(t)\} d\Lambda_۰(t), \quad \forall t \in [۰, \tau] \qquad (۲)$$

است که در آن $\gamma_۰$ بردار پارامترها و $Z_i(t)$ بردار متغیرهای کمکی برای مدل زمان‌های مشاهده است. همچنین تابع $\Lambda_۰(t)$، میانگین تعدادتجمعی مشاهدات تا زمان t، با فرض صفر بودن متغیرهای کمکی و نبود سانسور در داده‌ها است.
برای برآورد بردار پارامترهای مدل زمان‌های مشاهـــده، از فرض استقلال‌نمونه‌گیری استفاده می‌شود. یعنی فرض‌می‌شود که زمان‌مشاهدهٔ آزمودنی به شرط متغیرهای کمکی مدل زمان‌های مشاهده، از متغیر پاسخ و متغیرهای کمکی آن و متغیر سانسورشدن مستقل است. بنابراین داریم:

$$E[dN_i^*(t) | Z_i(t), X_i(t), Y_i(t), C_i \geq t] = E[dN_i^*(t) | Z_i(t)]. \qquad (۳)$$

با فرض اینکه متغیرهای کمکی در مدل زمان‌های مشاهده بخشی از متغیرهای کمکی در مدل میانگین پاسخ است، این فرض معادل ناآگاهی‌بخش بودن زمان‌های مشاهده است. با این وجود در این مقاله، متغیرهای کمکی در دو مدل دلخواه در نظر گرفته می‌شود. بنابراین با توجه به آگاهی‌بخش بودن زمان‌های مشاهده، بایستی از اطلاعات زمان‌های مشاهده در معادلات برآوردگر میانگین پاسخ استفاده کرد. برای این منظور از برآوردگر وزنی برای بردار پارامتر استفاده می‌شود.

حال به معرفی ساختـــار معادلات برآوردگر وزنی برای برآورد بردار پارامتر در مدل میانگین پاسخ در زمان‌های مشاهدهٔ نامنظم و وابسته به برآمد می‌پردازیم. برای این منظور، ابتدا وزن هر یک از مشاهدات در معادلات برآوردگر معرفی می‌شود. وارون وزن‌ها که با $\pi_i(.)$ نشان داده می‌شود، متناسب با احتمال این است که آزمودنی $i \in \{۱,...,n\}$ در زمان $t \in [۰,\tau]$، نسبت به سایر آزمودنی‌ها تحت مدل زمان‌های مشاهدهٔ (۲)، مشاهده شود. به عبارت دیگر، وارون وزن آزمودنی iاُم در زمان t به صورت

$$\pi_i(t,\gamma,h) = \frac{\exp\{\gamma^T Z_i(t)\}}{h(X_i(t))}, \qquad (۴)$$

تعریف می‌شود که در آن تابع $h(.)$، هر تابع قطعی از متغیرهای کمکی مدل میانگین پاسخ است. این تابع باعث افزایش دقّت برآوردگر حاصل از معادلات برآوردگـــر می‌شود. بر این اساس با استفاده از بررسی هِرنان و همکاران (۲۰۰۲)، تابع $h(.)$ به صورت زیر تعریف می‌شود.

$$h_۰(X_i(t)) = \exp\left\{\boldsymbol{\delta}_۰^T X_i(t)\right\},$$

به منظور به دست آوردن برآوردگر سازگار و به طور مجانبی نرمال، فرض می‌شود که برای آزمودنی iاُم، میانگین‌متغیر پاسخ‌مورد نظر به شرط متغیرهای کمکی مدل میانگین پاسخ، از زمان سانسور شدن مستقل است. به عبارت دیگر

$$E[Y_i(t)\,|X_i(t), C_i \geq t] = E[Y_i(t)\,|X_i(t)]. \qquad (۵)$$

حال برای یافتن معادلات برآوردگـــر، فرایند تصادفی $\{\mathcal{M}^W(t), t \in [۰,\tau]\}$، برای آزمودنی iاُم در زمان t به صورت

$$\mathcal{M}_i^W(t;\mathcal{A}(.),\boldsymbol{\beta},\gamma,h(.)) = \int_۰^t \frac{۱}{\pi_i(s;\gamma,h)}\{\left[Y_i(s) - \boldsymbol{\beta}^T X_i(s)\right] dN_i(s) - \xi_i(s)\exp\{\gamma^T Z_i(s)\}\, d\mathcal{A}(s)\}, \qquad (۶)$$

تعریف می‌شود که در آن تابع تجمعی میانگین پاسخ پایه به صورت زیر تعریف می‌شود:

$$\mathcal{A}\left(t;\alpha,\Lambda\left(.\right)\right)=\int_{0}^{t}\alpha\left(s\right)d\Lambda\left(s\right).$$

قضیه ۶ برای هر آزمودنی i، $i\in\{1,...,n\}$ و هر تابع قطعی $h\left(.\right)$، میانگین فرایند $d\mathcal{M}_{i}^{W}\left(t;\mathcal{A}_{0}\left(.\right),\beta_{0},\gamma_{0},h\left(.\right)\right)$ در زمان t، $t\in[0,\tau]$، به شرط $X_{i}\left(t\right)$ برابر صفر است (بوزکوا و لاملی، ۲۰۰۸). به عبارت دیگر می‌توان نوشت:

$$E\left[d\mathcal{M}_{i}^{W}\left(t\right)|X_{i}\left(t\right)\right]=0,\quad\forall t\in[0,\tau].$$

حال با استفاده از فرایند $\{\mathcal{M}^{W}\left(t\right):t\in[0,\tau]\}$ و قضیهٔ (۱)، مجموعهٔ معادلات برآوردگرِ همزمان برای برآورد بردار پارامتر در مدل میانگین پاسخ به صورت

$$\sum_{i=1}^{n}\mathcal{M}_{i}^{W}\left(t;A\left(.\right),\beta,\gamma,h\left(.\right)\right)=0\quad\forall t\in[0,\tau]\qquad(۷)$$

$$\sum_{i=1}^{n}\int_{0}^{\tau}W\left(t\right)X_{i}\left(t\right)d\mathcal{M}_{i}^{W}\left(t;\mathcal{A}\left(.\right),\beta,\gamma,h\left(.\right)\right)=0,\qquad(۸)$$

است که در آن فرایند $\{W\left(t\right),t\in[0,\tau]\}$، یک فرایند وزنی دلخواه و وابسته به داده‌ها است. حال با برآورد بردار پارامتر γ در مدل زمان‌های مشاهده با استفاده از روش درستنمایی جزئی و جایگذاری آن در معادلات (۷)، تابع $\mathcal{A}\left(t\right)$ برآورد می‌شود. سپس با جایگذاری دو برآوردگر در معادلات (۸)، برآوردگر بردار پارامتر β_{0} در مدل میانگین پاسخ به دست می‌آید.

۳ کاربرد

در این بخش با استفاده از یک مطالعهٔ شبیه‌سازی، به مقایسهٔ برآوردگر وزنی حاصل از مدل رگرسیونی نیم‌پارامتری با برآوردگر غیروزنی نیم‌پارامتری و برآوردگر پارامتری روش GEE برای داده‌های طولی تولید شده با پیگیری نامنظم و وابسته به برآمد می‌پردازیم. برای این منظور، میزان اریبی به همراه کارایی این برآوردگرها تحت توابع عرض از مبدأ و مدت زمان پیگیری آزمودنی‌ها مورد ارزیابی قرار می‌گیرد.

جدول (۱) میزان صحت و کارایی نسبی برآوردگر وزنی و نیم‌پارامتری را نسبت به برآوردگر غیروزنی نیم‌پارامتری و برآوردگر پارامتری حاصل از روش GEE نشان می‌دهد. همانطور که ملاحظه‌می‌شود، در حالت زمان‌های مشاهدهٔ آگاهی‌بخش، برآوردگر

تابع عرض از مبدأ	مدت زمان	میزان اریبی برآوردگر			کارایی برآوردگرها	
		برآوردگر وزنی	برآوردگر غیروزنی	برآوردگر GEE	RE^* (وزنی|GEE)	RE (غیروزنی|وزنی)
$\propto_i(t) = \sqrt{t}$	$\tau = ٤$	۰/۰٦۹	−۲/۷۳٦	−۲/۵٤۲	۱/۷۹۲	۳/٤۵۷
	$\tau = ۸$	۰/۰۵۵	−۲/٦۵۳	−۲/٤٦۵	۲/۳۲۲	٦/۵٤۲
$\propto_i(t) = \sin(t)$	$\tau = ٤$	۰/۰۹۵	−۲/٦۱۹	−۲/۵۱۵	۱/۸۳٦	۳/۷۹۵
	$\tau = ۸$	۰/۰۷٤	−۲/۹۲۰	−۲/۸۲۳	۲/۱٦۷	٦/۷۸٦
$\propto_i(t) = exp(۲\|\sin(t)\|)$	$\tau = ٤$	۰/۰۸۸	−۲/۷۳۸	−۲/۹٤٤	۱/۷۳۵	۳/۵۵۷
	$\tau = ۸$	۰/۰۹۲	−۲/۸۷۵	−۳/۰٦٦	۲/۱٦٦	٦/٤۵۲

* Relative Efficiency

جدول (۱): صحت و کارایی برآوردگرهای وزنی و غیروزنی نیم‌پارامتری و برآوردگر پارامتری GEE برای داده‌های طولی با پیگیری نامنظم و وابسته به برآمد

غیروزنی حاصل از روش نیم‌پارامتری و برآوردگر پارامتری حاصل از روش GEE، اریب هستند و پارامتر مورد نظر را کم برآورد می‌کنند. همچنین فرم تابع عرض از مبدأ، تأثیری بر سازگاری برآورد پارامتر $\beta_{۰۱}$ ندارد.

با توجه به جدول (۱)، کارایی نسبی برآوردگر وزنی نسبت به برآوردگر GEE از کارایی آن نسبت به برآوردگر غیروزنی کمتر است. زیرا در حالت نیم‌پارامتری، تابع عرض از مبدأ نامعلوم در نظر گرفته می‌شود. بنابراین واریانس برآوردگر نیم‌پارامتری از واریانس برآوردگر پارامتری GEE بزرگتر است. به طور کلی هنگامی که تابع عرض از مبدأ معلوم است، کارایی روش پارامتری از روش نیم‌پارامتری بیشتر است، با این وجود بایستی در نظر داشت که کاهش دقت برآوردگرهای پارامتری در صورت عدم تشخیص صحیح تابع عرض از مبدأ، قابل ملاحظه است (لین و ینگ، ۲۰۰۱).

نکته قابل توجه دیگر در این جدول، افزایش کارایی نسبی برآوردگر وزنی با افزایش مدت زمان آزمایش (τ) است. به‌طوری‌که به عنوان مثال، کارایی برآوردگر وزنی نسبت به برآوردگر GEE با فرض $\tau = ۴$ برابر ۱/۷۹۲ و در حالت $\tau = ۸$ برابر ۲/۳۲۲ است. علت این مسئله، افزایش زمان‌های مشاهدهٔ آگاهی‌بخش برای هـــر آزمودنی به دلیل افزایش طول مدت مطالعه است.

مثال ۱ مطالعهٔ ACTG[۱] به مقایسهٔ ۴ روش درمان متفاوت در بیماران مبتلا به HIV+ پرداخته است. این ۴ روش درمان عبارتند از:

۱) مصرف ۶۰۰ میلی گرم zidovudine به طوری که هر ماه به تناوب جایگزین ۴۰۰ میلی گرم didanosine می‌شود.

۲) مصرف ۶۰۰ میلی گرم zidovudine به همراه ۲/۲۵ میلی گرم zalcitabine

[۱] AIDS Clinical Trial Group

۳) مصرف ۶۰۰ میلی گرم zidovudine به همراه ۴۰۰ میلی گرم didanosine

۴) مصرف ۶۰۰ میلی گرم zidovudine به همراه ۴۰۰ میلی گرم didanosine و ۴۰۰ میلی گرم nevirapine

متغیر پاسخ مورد نظر در این بررسی، لگاریتم تعداد سلول‌های CD4 به اضافهٔ یک بیمار است. به عبارت دیگر، $Y(t) = \log\{CD4count(t) + 1\}$ است. هدف از این مطالعه، برآورد تغییر منحنی میانگین متغیر پاسخ در طول زمان در بین ۴ روش مختلف درمان بیماری ایدز است.

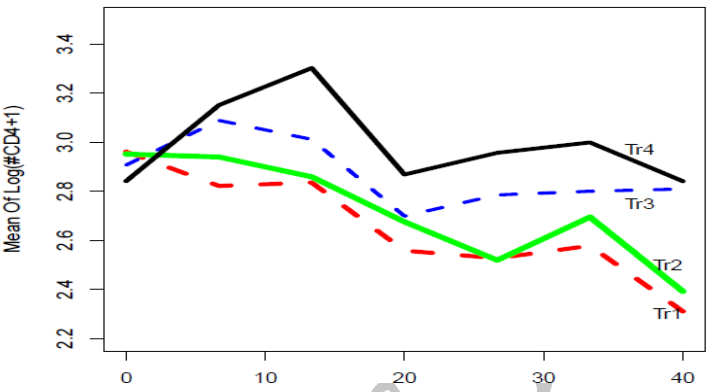
برای این منظور، ۱۳۰۹ بیمار به تصادف در این ۴ نوع درمان تخصیص یافته، و مقرر گردید تعداد CD4 هر بیمار در فواصل ۸ هفته‌ای به مدت ۴۰ هفته جمع‌آوری شوند. تعداد بیماران تخصیص یافته در این گروه‌ها به ترتیب ۳۲۵، ۳۲۴، ۳۳۰ و ۳۳۰ نفر می‌باشد. همچنین از ۵۰۳۶ ملاقات صورت گرفته به منظور شمارش تعداد سلول‌های CD4 بیماران، تعداد ملاقات در هر گروه به ترتیب ۱۲۳۹، ۱۲۵۱، ۱۲۵۴ و ۱۲۹۲ می‌باشد.

با وجود تلاش‌های زیاد برای حضور منظم بیماران برای آزمایش (فواصل ۸ هفته‌ای)، آزمودنی‌ها اغلب این ملاقات‌ها را لغو یا در بین زمان‌های مقرّر در مطالعه حضور می‌یافتند، که در نتیجه فراوانی و زمان‌های مشاهده در بین آزمودنی‌ها کاملاً متفاوت و ساختار داده‌ها نامتعادل گردید. به نحوی که در طول ۴۰ هفته، تعداد دفعات حضور بیماران بین ۱ تا ۹ با میانهٔ ۴ بود.

نگرانی دیگر در تحلیل داده‌ها، وجود ارتباط بین زمان و فراوانی مشاهدات با مقدار متغیر پاسخ در آخرین ملاقات، تعداد دفعات حضور قبلی آزمودنی‌ها و متغیرهای دیگر بود. به طوری که با بررسی انجام شده بر روی داده‌ها، به طور متوسط، بیمارانی که تعداد سلول CD4 بیشتری داشتند، برای حضور در آزمایش تمایل بیشتری داشتند. همچنین بیشترین و کمترین تعداد حضور، به ترتیب مربوط به بیماران درمان‌های نوع چهارم و اول است که به طور متوسط دارای بیشترین و کمترین تعداد CD4 هستند [شکل (۱)]. به علاوه بیماران مرد به طور متوسط بیشتر از بیماران زن برای اندازه‌گیری متغیر پاسخ تمایل داشته‌اند.

این موارد باعث می‌شود که با مدل‌بندی متغیر برآمد (لگاریتم تعداد CD4 به اضافهٔ یک) و متغیر کمکی (نوع درمان) به روش معمولی، برآوردگر حاصل برای کل جامعه نااریب نباشد. این اریبی ممکن است با در نظر گرفتن برآمدهای گذشته و تعداد ملاقات‌های قبلی یا متغیرهای دیگر به عنوان متغیر کمکی در مدل میانگین پاسخ، تعدیل شود. ولی این کار، ممکن است، اثر نوع درمان را بر متغیر پاسخ مورد نظر تعدیل نماید. بنابراین

در این وضعیت استفاده از تحلیل رگرسیون استاندارد، پارامترهای مورد نظر را به شکل قابل اطمینان برآورد نخواهد کرد.



شکل ۱.: میانگین تعداد سلول‌های CD4 بیماران در هر روش درمان در طول زمان (هفته)

شکل (۱) وضعیت توصیفی میانگین تعدادسلول‌های CD4 بیماران را در روش‌های مختلف درمان در طول مدت زمان مطالعه نشان می‌دهد. همان‌طور که مشاهده می‌شود، استفاده از ۳ نوع دارو (درمان نوع چهارم) و استفاده از داروی جایگزین در طول درمان (درمان نوع اول) به ترتیب باعث بیشترین و کمترین میانگین تعداد سلول CD4 در طول دورهٔ درمان می‌شوند. با این حال به منظور برآورد اثر هر یک از روش‌های درمان و بررسی‌وجود تفاوت معنی‌دار بین روش‌های درمان، از جدول تحلیل‌واریانس استفاده خواهیم کرد.

برای برازش مدل و یافتن فرم بستهٔ برآوردگر، کافی است متغیرهای کمکی و متغیر پاسخ توسط منحنی میانگین، مرکزی شوند. علاوه بر این، برای یافتن برآوردگر وزنی، ابتدا بایستی با استفاده از برآوردگرهای مدل زمان مشاهده، وزن‌ها را پیدا کرد. برای مدل زمان مشاهده، از مدل رگرسیونی نیم‌پارامتری نرخ حاشیه‌ای زیر استفاده می‌شود:

$$E\left[dN_i^*(t)\,|N_i(t^-), Y_i(t^-), g_i\right] = \exp\left\{\gamma_{01}N_i(t^-) + \gamma_{02}Y_i(t^-) + \gamma_{03}g_i\right\} d\Lambda_0(t)$$

که در آن متغیرهای زمان وابستهٔ $N_i(t^-)$ و $Y_i(t^-)$، به‌ترتیب نشان‌دهندهٔ تعداد ملاقات های قبلی و لگاریتم تعداد CD4 به اضافهٔ یک در آخرین ملاقات، برای آزمودنی iاُم است. به علاوه متغیر زمان ثابت g_i نشان دهندهٔ جنسیت آزمودنی iاُم است.

جدول (۲) برآورد پارامترهای مدل نرخ حاشیه‌ای را نشان می‌دهد. تعداد دفعات حضور قبلی و لگاریتم تعداد CD4 به اضافهٔ یک بیمار در آخرین ملاقات، در سطح ۰/۰۵ به

طور معنی‌دار بر نرخ حضور فرد مؤثر هستند. همچنین تفاوت معنی‌داری در حضور بیماران بر اساس جنسیت آن‌ها وجود دارد. به طوری که با توجه به نسبت نرخ حضور بیماران مرد به زن، (۱/۲۵۳)، می‌توان نتیجه گرفت که با فرض ثابت بودن دو متغیر دیگر، نرخ حضور بیماران مرد ۲۵/۳ درصد بیشتر از بیماران زن است.

متغیر کمکی	برآورد پارامتر	$\exp(\hat{\beta})$	برآورد استوار انحراف معیار برآوردگر	P- مقدار
تعداد ملاقات قبلی	-۱/۰۰۸	۰/۳۶۵	۰/۰۳۸	۰/۰۰۰
لگاریتم تعداد CD4 به اضافه یک آخرین ملاقات	-۰/۰۵۰	۰/۹۵۱	۰/۰۲۵	۰/۰۴۵
جنسیت	۰/۲۲۵	۱/۲۵۳	۰/۱۰۳	۰/۰۲۹

جدول (۲): برآورد پارامترهای مدل زمان مشاهده

حال با استفاده از برآوردگرهای حاصل از روش ماکسیمم درستنمایی جزئی در جدول (۲) برای مدل زمان مشاهده، ضمن به دست آوردن برآوردگر وزنی در مدل رگرسیونی نیم‌پارامتری (۱)، به مقایسهٔ این برآوردگر با برآوردگر غیروزنی و برآوردگر پارامتری روش GEE می‌پردازیم.

مدل رگرسیونی نیم‌پارامتری به منظور ارزیابی روش‌های مختلف درمان بیماری ایدز به صورت زیر است:

$$E\left[Y_i(t)\,|I_i\right] = \alpha_{\circ}(t) + \beta_{\circ 1} I(I_i = Tr_1) + \beta_{\circ 2} I(I_i = Tr_2) + \beta_{\circ 3} I(I_i = Tr_3) \quad (۹)$$

که در آن $I(.)$ تابع نشانگر و Tr_j $(j = ۱,۲,۳)$ نشان دهندهٔ درمان نوع j است. همچنین درمان نوع چهارم به عنوان متغیر کمکی پایه در نظر گرفته می‌شود.

جدول (۳) برآورد و برآورد انحراف معیار اثر هر یک از روش‌های درمان را نشان می‌دهد. همان‌طور که ملاحظه می‌شود، مانند شکل (۱)، تأثیر درمان نوع چهارم بر میان‌گین پاسخ (لگاریتم تعداد سلول‌های CD4 به اضافهٔ یک) از سایر روش‌های درمان بیشتر است. با این وجود در سطح ۰/۰۵، تنها تفاوت معنی‌داری بین روش‌های درمان اول و چهارم وجود دارد. به عبارت دیگر، اثربخشی معنی‌داری در استفاده از سه دارو (درمان نوع چهارم) به جای دو دارو (درمان دوم یا سوم) وجود ندارد. ولی استفاده از سه داروی مذکور به طور معنی‌دار اثربخشی بیشتری نسبت به استفاده از داروی جایگزین در درمان (درمان نوع اول) خواهد داشت. به طوری که در هر زمان میانگین لگاریتم تعداد سلول‌های CD4 به اضافهٔ یک حاصل از درمان چهارم، نزدیک به ۰/۲۵

متغیر کمکی	برآورد پارامتر			برآورد انحراف معیار برآوردگر			P-مقدار	
	برآوردگر وزنی	برآوردگر غیروزنی	برآوردگر GEE	برآوردگر وزنی	برآوردگر غیروزنی	برآوردگر GEE	برآوردگر وزنی	برآوردگر غیروزنی
درمان نوع اول	-۰/۲۴۱۹	-۰/۳۹۶۶	-۰/۳۶۸۸	۰/۰۸۴۲	۰/۰۶۶۰	۰/۰۴۲۷	۰/۰۰۴۱	۰/۰۰۰۷
درمان نوع دوم	-۰/۱۱۰۶	-۰/۲۸۴۵	-۰/۲۴۹۷	۰/۰۸۴۶	۰/۰۶۶۰	۰/۰۴۲۶	۰/۱۹۱۶	۰/۲۰۰۶
درمان نوع سوم	-۰/۰۷۴۹	-۰/۲۲۸۷	-۰/۲۰۹۰	۰/۰۸۳۹	۰/۰۶۵۱	۰/۰۴۲۵	۰/۳۷۲۲	۰/۶۵۹۳

جدول (۳): مقایسهٔ برآوردگرهای وزنی و غیروزنی نیم‌پارامتری و برآوردگر پارامتری GEE برای داده‌های روش‌های درمان بیماری ایدز

بیشتر از درمان نوع اول است. با استفاده از جدول (۳) می‌توان برآوردهای حاصل از روش وزنی را با برآوردگر غیروزنی نیم‌پارامتری و برآوردگر پارامتری مقایسه نمود. نکته قابل توجه تفاوت در برآورد پارامترها و انحراف معیار برآوردگرها در هر ۳ روش است. در روش پارامتری، برآورد انحراف معیار از روش نیم‌پارامتری کمتر است. به این دلیل که در روش‌های نیم‌پارامتری، تابع عرض از مبدأ نامعلوم در نظر گرفته شده‌اند. با این وجود همان‌طور که در مطالعهٔ شبیه‌سازی نیز اشاره شد، برآوردهای حاصل از روش وزنی برای داده‌های طولی با زمان‌های مشاهدهٔ آگاهی‌بخش، سازگار هستند.

۴ نتیجه‌گیری

در این مقاله به مقایسهٔ برآوردگرهای مختلف برای داده‌های طولی با پیگیری نامنظم و وابسته به برآمد پرداختیم. برای به دست آوردن معادلات برآوردگر وزنی از اطلاعات زمان‌های مشاهده استفاده شد. به علاوه با استفاده از مرکزی کردن متغیرهای کمکی و متغیر پاسخ، پارامتر مزاحم $\alpha_0(t)$ از معادلات برآوردگر حذف گردید. در واقع روش برآورد از این جهت مانند روش کمترین توان‌های دوم است. همانطور که ملاحظه شد، فرم تابع عرض از مبدأ، نقشی در سازگاری برآوردگر وزنی نداشت. استفاده از روش‌های هموارسازی برای برآورد تابع عرض از مبدأ، محاسبات پیچیده‌ای دارند. به علاوه این روش‌ها ممکن است بر سازگاری برآوردگر مدل تأثیرگذار باشند.
سازگاری برآوردگرهای به دست آمده، مستلزم کاملاً مشاهده شدن متغیر کمکی $X(t)$

است. زیرا کامل بودن متغیرهای کمکی در تحلیل داده‌های بقا، یک فرض اساسی است. با این وجود این فرض برای متغیر پاسخ الزامی نیست و استفاده از تقریب برای آن تنها بر دقت برآوردگر اثر می گذارد.

همان‌طور که ملاحظه گردید، هنگامی که زمان‌های مشاهده ناآگاهی بخش و تابع عرض از مبدأ معلوم است، با توجه به معیار دقت، استفاده از برآوردگر پارامتری بر نیم‌پارامتری اولویت دارد. در حالی که اگر تابع عرض از مبدأ نامعلوم باشد، زیان ناشی از عدم تشخیص صحیح تابع عرض از مبدأ در روش پارامتری قابل ملاحظه خواهد بود. در نهایت اینکه، همان‌طور که ملاحظه شد، هنگامی کـــه زمان‌های مشاهده آگاهی‌بخش هستند، استفاده از برآوردگر وزنی بر خلاف برآوردگرهای دیگر برای پارامتر مورد نظر سازگار خواهد بود.

مراجع

1. Buzkova, P. and Lumley, T. (2008). Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Stat Med*, 28(6): 987-1003.
2. Hernan, M. A., Brumback, B. A., Robins, J.M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for
3. Lin, D. Y. , Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association*, 96, 103-113.
4. Lin, H, Scharfstein, D.O. and Rosenheck, R. A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society*, B 66, 791-813.
5. Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without-with error. *Journal of the American Statistical Association*, 95, 520-534.
6. Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R. and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, 58, 621-630.
7. Sun, J., Sun, L., Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *Journal of the American Statistical Association*, 102, 1397-1406.