کاربرد نمونه‌گیری مونت‌کارلویی در تحلیل پاسخ‌های دو حالتی طولی ناقص به روش بیزی

دکتر حبیب‌الله اسماعیلی[1]، دکتر محمد رضا مشکانی، دکتر ناصر رضا ارقامی، دکتر انوشیروان کاظم نژاد

چکیده مقاله

قبل از دهه ۱۹۷۰ بکارگیری روش‌های بیزی به دلیل محاسبات طولانی، کمتر مورد استفاده آمارشناسان قرار می‌گرفت . اما در حال حاضر سرعت فوق العاده ریز پردازشگرها این امکان را فراهم نموده است که روش‌های بیزی به سرعت مورد توجه تحلیل گران قرار گیرد.

در بسیاری از مطالعه‌ها ، متغیر پاسخ برای هر فرد در چندین نوبت متوالی مشاهده میشود و ممکن است به صورت دو حالته باشد.این نوع مطالعه‌ها را مطالعه‌های طولی (Longitudinal Study) می‌گویند. مطالعه‌های طولی علی رغم دقت بیشتر نسبت به مطالعه‌های مقطعی دارای این اشکال هست که به دلایل مختلف ، مشاهدات برخی از واحدها در همه زمان‌ها ثبت نمی‌شوند.بنابر این با داده‌های گمشده مواجه هستیم.

کاظم نژاد و مشکانی با در نظر گرفتن اثرتصادفی برای تیمار یا عامل‌ها از یک مدل بندی خاصی که بر گرفته از مدل (Grizzel, ۱۹۶۹ Starmer, Koch,)، با به کارگیری تبدیل لوجیت برای پاسخ‌های دو حالتیِ طولیِ با داده‌های گمشده یک مدل خطیِ آمیخته را پیشنهاد کردند و با استفاده از روش سوم هندرسون مؤلفه‌های واریانس را برآورد کردند. اما هنگام به کارگیری روش سوم هندرسن امکان تولید واریانس صفر یا حتی منفی وجود دارد ، که از نظر کاربردی نتیجه‌ای بی‌فایده است. برای اجتناب از این وضعیت نامطلوب ، در این مقاله ضمن به کارگیری روش کاظم نژاد و همکاران ، ۱۳۷۵ مؤلفه‌های واریانس به روش بیزی برآورد شد.در این روش از نمونه‌گیری ردی که دارای انعطاف پذیری زیادی می‌باشد ،استفاده شده است. این نوع نمونه‌گیری یکی از انواع نمونه‌گیری مونت کارلویی می‌باشد که روشی کلی است و برای پیشین‌های مزدوج و نامزدوج کاربرد دارد.در این مطالعه توزیع پیشین برای مولفه واریانس پیشین جفریز در نظر گرفته شد. برای توضیح چگونگی استفاده از روش پیشنهادی ، مجموعه‌ای از داده‌ها که مربوط به تاثیر چهار نوع رژیم غذایی بر میزان کلسترول خون می‌باشد، مورد تجزیه و تحلیل قرار گرفت.نتایج نشان داد برآوردهای بیزی به دست آمده خطای معیار کمتری نسبت به روش‌های کلاسیک داشته و به علت در نظر گرفتن پیشین جفریز برای مولفه‌های واریانس امکان تولید واریانس صفر یا منفی وجود ندارد.

● *واژه‌های کلیدی: مؤلفه واریانس، طولی، دوحالتی، گمشده، لوجیت، آمیخته، بیز، نمونه‌گیری ردی.*

مقدمه

در بسیاری از مطالعه‌ها ، متغیر پاسخ برای هر فرد در چندین نوبت متوالی مشاهده می‌شود. به مطالعه‌ای که اندازه‌گیری مربوط به یک صفت در طول زمان مورد بررسی قرار می‌گیرد، مطالعه طولی می‌گویند(۱). پاسخ ثبت شده، ممکن است کمی یا به صورت رسته‌ای[2] به ویژه دو حالته باشد. مطالعه‌های طولی نسبت به مطالعه‌های مقطعی (پاسخ هر فرد، تنها یک بار مشاهده می‌شود) دارای دقت بیشتری است، اما دارای این اشکال می‌باشد که به دلایل مختلف، مشاهدات مربوط به برخی از واحدها در همه زمان‌ها ثبت نمی‌شوند که در این صورت با داده‌های گمشده مواجه هستیم. حذف واحدهایی که دارای داده گمشده هستند، علاوه بر کاهش حجم نمونه و از بین رفتن هزینه‌های مصرف شده، ممکن است باعث اریبی در تجزیه و تحلیل نهایی گردد(۲). جهت تحلیل این گونه پاسخ‌ها راه حل‌های گوناگونی ارایه گردیده است.گریزل و همکاران در سال ۱۹۶۹ روشی را به کار بردند که بعدا به روش GSK[3] معروف شد، سپس افرادی همچون کوچ، ایمری و رینفورت[4] در سال ۱۹۷۲ و ولسون و کلارک[5] در سال ۱۹۸۴ روش فوق را برای مجموعه داده‌های دو حالته طولی با پاسخ‌های گمشده تعمیم دادند، (۳). مدل فوق مدل اثرات ثابت می‌باشد. اما در بسیاری از مطالعه ممکن است تیمارها یا عامل‌ها از جامعه تیمارها یا عامل‌ها به تصادف انتخاب شوند، بنابراین نمی‌توان از مدل‌های اثرات ثابت استفاده نمود.

در این مقاله با فرض اینکه تیمار یا متغیر تبیینی دارای اثر تصادفی هستند، پاسخ های دو حالتی طولی برای داده هایی که دارای مقادیر گمشده نیز بودند، مورد بررسی قرار گرفت و پارامترها به روش بیزی برآورد شدند و با روش ماکسیمم درستنمایی مقید مقایسه گردیدند.

مدل آماری برای مطالعه‌های طولی با پاسخ دو حالتی

مدل بندی به کار گرفته شده ، روشی است که با تعمیم روش GSKتوسط کاظم نژاد و همکاران ، در سال۱۳۷۵ به کار گرفته شده است.در این مدل داده‌های گمشده طبق تعریف لیتل ورابین (۲) کاملا تصادفی (Missing Completely At Random ,MCAR) می‌باشد. یک جدول توافقی از رسته‌های پاسخ و زیر گروه طبق جدول ۱ تشکیل می‌شود. بردارهای به وجود آمده به صورت یک بردار ستونی نوشته می‌شود که به آن بردار مشاهدات می‌گویند. با این مدل‌بندی حتی اگر در یک مورد هم داده‌های

۱- استادیار دانشگاه علوم پزشکی مشهد

2. Categorical
3. Grizzel, Starmer, Koch
4. Koch, Imrey & Reinfurt
5. Woolson & Clarke

گمشده مشاهده شود ، مد نظر قرار می‌گیرد.با در دست داشتن بردار محاسبه مشاهدات و عملیات ماتریسی مقادیر آن در Logit(P)گردید. (جزییات قسمت کاربرد و منابع ۲، ۴ و ۱۲ توضیح داده شده است) بردار برداری از Logit(P) F(pبا F(p) را می‌نامند ، p نسبت‌های موفقیت می‌باشد که توجه به زیر جامعه‌ها و زمان‌های تکرار ساخته می‌شود. بنابراین (تابعی از تابعی از بطور p می‌باشد و p متغیرهای تبیینی یا تیماری است. با فرض اینکه عامل تیمار تصادفی است و (در نظرF(p مجانبی دارای توزیع نرمال است، عبارت نوشته F(p)به صورت y گرفته شده و مدل آمیخته ذیل می‌شود (۴).

$y = x\beta + z\gamma + e$ (مدل ۱)

که در آن β و γ ضرایب مدل هستند و به ترتیب مربوط به عوامل ثابت و تصادفی می‌باشند و e جمله خطاست. ماتریس‌های X و Z به ترتیب ماتریس‌های طرح اثرات ثابت و تصادفی در مدل ۱ می‌باشند. برای برآورد پارامترهای مدل ۱ روشهای گوناگونی در آمار کلاسیک وجود دارد (۵). اما به کارگیری روش‌های بیزی به دلیل استفاده از اطلاعات پیشین، برآوردهای دقیق تری از پارامترها ارایه می‌دهد. بادردست داشتن مقادیر لوجیت و مدل ۱ با استفاده از روشهای نمونه‌گیری مونت کارلو[1] پارامترهای مدل، به شیوه نمونه‌گیری ردی[2] برآورد می‌شوند. (۶،۷)

توزیع‌های پیشین

پارامتر اثر تصادفی به طور معمول دارای میانگین صفر و ماتریس واریانس D و بردار e دارای توزیع نرمال با میانگین صفر و ماتریس واریانس Σ فرض می‌شود. در اینجا D و Σ مولفه‌های واریانس هستند، که باید برآورد شوند. روش به کار رفته در این قسمت طوری است که علاوه بر وجود دقت ، اجرای آن راحت بوده و برای به دست آوردن برآوردها امکان استفاده از نرم افزارهای متداول وجود دارد.

ساختار مدل آمیخته ۱ به گونه‌ای است که منجر به دو مرحله می‌شود. چنانچه پارامترها به صورت β و γ و θ در نظر گرفته شود، (پارامتر θ شامل مؤلفه‌های واریانس D و Σ بوده و D و Σ قطری فرض می‌شوند.)مدل فوق از نوع مدل مؤلفه‌های واریانس معمولی است. حال توزیع پسین توأم آن‌ها به صورت ذیل نوشته می‌شود.

$$P(\beta,\gamma,\theta \mid y) = P(\beta,\gamma \mid \theta,y)\, P(\theta \mid y) \qquad (۲)$$

دو عامل سمت راست به طور جداگانه مورد بررسی قرار می‌گیرد. در مرحله اول تحت تابع زیان درجهٔ دوم خطا ، نخست با استفاده از بر آورد θ به صورت میانگین توزیع پسین به دست آورده می‌شود و سپس با جایگذاری این برآورد در یک توزیع نرمال چند متغیرهٔ $P(\beta,\gamma,\theta \mid y)$ پارامترهای β و γ به صورت میانگین این توزیع به دست آورده می‌شود. توزیع توأم β و γ به شرط داده‌ها به صورت ذیل می‌باشد،(۸) .

$$\beta,\gamma \mid y \sim MVN \begin{matrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & Z'\Sigma^{-1}Z+D^{-1} \end{matrix}^{-1} \begin{matrix} X'\Sigma^{-1}Y \\ Z'\Sigma^{-1}Y \end{matrix}, \begin{matrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & Z'\Sigma^{-1}Z+D^{-1} \end{matrix}^{-1} \qquad (۳)$$

در مرحله دوم با داشتن توزیع فوق وبا استخراج نمونه‌ای از آن می‌توان در مورد β و γ استنباطهایی انجام داد .

پسین کناری

استفاده ازرابطه (۳) به در دست داشتن Σ و D یا بطور کلی θ نیاز دارد که این‌ها همان مؤلفه‌های واریانس هستند.برای این منظور توزیع پسین θ را به صورت ذیل می‌نویسیم.

$$P(\theta \mid y) \propto L(\theta)\, \pi(\theta) \qquad (۴)$$

در اینجا $L(\theta)$ می‌تواند تابع درستنمایی یا درستنمایی مقید[3] باشد و $\pi(\theta)$ توزیع پیشین θ است. در این بحث ما همه جا $L(\theta)$ را تابع درستنمایی مقید در نظر می‌گیریم.

$$L(\theta) = |kvk'|^{-1/2} \cdot (2\pi)^{-1/2(n-p)} \exp[-\frac{1}{2} y'k' (kvk')ky]$$

که در آن Kیک ماتریس بصورت (n-p)×n بصورت $k= I - x(x'x)^{-1}x$ می‌باشد. برای $\pi(\theta)$ با توجه به نظر خبرگان می‌توان توزیع مناسبی انتخاب کرد ولی چنانچه اطلاعی راجع به آن نباشد می‌توان از توزیع‌های مرجع استفاده نمود. در این خصوص پیشین‌های مرجع زیادی وجود دارند که می‌توان به منبع(۹) مراجعه کرد. مهمترین آنها پیشین جفریز می‌باشد. این پیشین متناسب با ریشه دوم دترمینان ماتریس اطلاع فیشر است. که توسط Wolfinger R.D. و دیگران محاسبه گردیده است .لذا توزیع پسین به صورت ذیل خواهد شد.

$$p(\theta \mid y) \propto |I_R(\theta)\, L(\theta,y)| \qquad (۵)$$

که در آن $I_R(\theta)$ همان ماتریس اطلاع فیشر است. با داشتن فرم توزیع $p(\beta,\gamma) \mid \theta,y)$ و $p(\theta \mid y)$ ، استفاده از شبیه‌سازی در دو مرحله به ترتیب زیر صورت می‌پذیرد.

۱) یک عدد شبه تصادفی θ^* را از توزیع $p(\theta \mid y)$ استخراج می‌کنیم.

۲) به ازای θ^* حاصل از مرحله (۱) وبا استفاده از رابطه (۳) مقادیری را برای β و γ از توزیع توأم $p(\beta,\gamma \mid y)$ استخراج می‌نماییم.

همان طور که گفتیم مرحله دوم به آسانی قابل انجام است ، برای اجرای مرحله نخست و استخراج نمونه از $p(\theta \mid y)$ از روش نمونه‌گیری ردی[4] استفاده می‌شود. اما این روش به یک توزیع پیشنهادی[5] به نام $g(\theta)$ نیاز داردکه طریقه به دست آوردن و متعاقب آن نحوهٔ استخراج نمونه از $p(\theta \mid y)$ را ارائه خواهیم کرد.

توزیع پیشنهادی

در برآورد کردن پارامترها با روش نمونه‌گیری ردی ،از دیدگاه بیزی نیاز به

1. Mont Carlo
2. Regection Sampling
3. Restricted Maximum Likelihood
4. Regection Sampling
5. Proposal distribution

تعیین یک توزیع پیشنهادی به نام $g(\theta)$ داریم تا با θ_j تولید از آن بتوان توزیع تقریبی پسین ومتعاقب آن پارامترها را برآورد نمود. لازم به ذکر است $g(\theta)$ تا آنجا که ممکن است، باید به توزیع پسین نزدیک باشد.

باکس و تیائو(۱۰) اظهار می‌کنند که در مؤلفه‌های واریانس ، توزیع پسین $p(\theta|y)$ دقیقا برابر حاصلضرب چگالی گاماهای وارون یا تقریب خوبی از این حاصلضرب می‌باشد. همچنین سیرل و دیگران(۸) ، گامای وارون برای توزیع σ^2ها را نه فقط برای متغیر تصادفی مثبت، واقع گرایانه ، بلکه نتایج توزیع پسین آن را به مراتب ساده‌تر می‌دانند. چگالی گامای وارون به صورت ذیل نوشته می‌شود.

$$IG(x|a,b) = \frac{b^a}{\Gamma(a)x^{a+1}} e^{-b/a} \qquad x>0 \text{ و } a,b>0 \qquad (۶)$$

چگالی‌هایی که شامل این حاصلضرب‌ها هستند ، نوعاً بر حسب مؤلفه‌های واریانس موجود در θ نیستند، بلکه بیشتر برحسب امید ریاضی توان دوم اثرها در یک جدول تحلیل واریانس می‌باشند. به طوری که امید ریاضی میانگین توان دوم اثرها ترکیبی خطی از مؤلفه‌های واریانس هستندکه ضرایب آنها متاثر از حجم نمونه است .برای مثال در یک آزمایش بلوک تصادفی متعادل با k . مشاهده در هر بلوک ، حاصلضرب چگالی‌ها براساس σ^2_e و $\sigma^2_e + k\sigma^2_b$ می‌باشد. σ^2_c و σ^2_b به ترتیب مؤلفه‌های واریانس باقیمانده و بلوک می‌باشند.

برای تعریف یک چگالی پیشنهادی $g(\theta|y)$ از حاصلضرب چگالی گاماهای وارون استفاده می‌شود.برای تعیین آنها ، ابتدا یک تبدیل خطی از $\theta \to \tau$ تعریف می‌کنیم. این ترکیب خطی را می‌توان با استفاده از روش‌های کلاسیک در برآورد مؤلفه‌های واریانس مانند روش سوم هندرسن تعیین نمود.

اگر فرض کنیم هدف برآورد کردن i امین مؤلفهٔ واریانس باشد، τ_i را تعیین کنیم. در این ترکیب خطی از مولفه‌های واریانس ، برآورد ماکسیمم درستنمایی مقید به جای مؤلفه‌های واریانس قرار داده می‌شود. لذا برای هر یک از τ_iها یک عدد به دست آورده می‌شود. فرض کنید τ_i دارای چگالی گامای وارون است،لگاریتم آن به صورت ذیل در نظر گرفته می‌شود.

$$\log(\tau_i|y) = c_i-(a_i+1)\log\tau_i - \frac{b_i}{\tau_i} \qquad (۸)$$

رابطهٔ فوق یک رگرسیون خطی با عرض از مبدأ c_i می‌باشد. با برازش یک رگرسیون خطی می‌توان a_i و b_i را برآورد کرد. برای این کار باید برای هر یک از τ_i ها تعدادی مقدار داشته باشیم. مقدار τ_i را که قبلاً از روی ترکیب خطی با قرار دادن برآوردهای ماکسیمم درستنمایی مقید به دست آورده بودیم ، یک بار بر دو تقسیم و یک بار در دو ضرب می‌کنیم. دو مقدار به دست می‌آید. بین این دو عدد را به ۲۰ قسمت مساوی تقسیم می‌کنیم. حال به ازای این ۲۰ مقدار لگاریتم چگالی پسین طبق رابطه (۵) به دست آورده می‌شود. با در دست داشتن این مقادیر از رابطهٔ (۹) استفاده گردیده ، برآوردهای a_i و b_i را به دست می‌آوریم. این شیوه توسط ول فینگر و کَس[1] بکار برده شده و آنها اظهار می‌کنند که هر چند این کار قدری مبتدیانه است ولی در عمل خوب کار می‌کند.با یافتن برآوردهای a_i و b_i توزیع τ_i به دست می‌آید. حال اگر ما به تعداد q مؤلفهٔ واریانس داشته باشیم، توزیع g را از حاصلضرب‌ها τ_iها به ازای q، ... ،i=۱،۲ به دست آورده می‌شود.

$$g(\tau|y) \propto \prod_{i=1}^{q} IG(\tau_i|a_i, b_i) \qquad (۹)$$

می‌دانیم τ_i ها تبدیل‌های خطی از مؤلفه‌های واریانس بودند، حال با کار بردن تبدیل عکس $g(\theta|y)$ به دست آورده می‌شود ، که به آن توزیع پیشنهادی می‌گویند.

با در دست داشتن توزیع پیشنهادی یعنی $g(\theta|y)$ از طریق گرفتن نمونه، توزیع پسین تقریبی با استفاده از روش نمونه‌گیری ردی به دست آورده می‌شود(۶).

نمونه‌گیری ردی

این روش یک روش کلی و مشابه روش هایی است که اعداد تصادفی تولید می‌کنند. در این روش توزیع پسین استاندارد شده تقریب زده می‌شودلم ذیل موید این مطلب است.

لم (حالت گسسته)

فرض کنید g(x) و h(x) توابع جرم احتمال دو توزیع گسسته با تکیه گاه مشترک باشند، الگوریتم ذیل را در نظر بگیرید.

۱- مشاهدهٔ x را از توزیع g(x) تولید می‌کنیم.

۲- مشاهدهٔ u را از توزیع یکنواخت در بازهٔ (۱و۰) تولید می‌کنیم.

اگر به ازای عددی مانند M>1 داشته باشیم $u<\frac{h(x)}{Mg(x)}$ ، آنگاه y =x و توقف می‌کنیم. در غیر این صورت به مرحله (۱) بر می‌گردیم و نمونه‌گیری را ادامه میدهیم تا شرط $u<\frac{h(x)}{Mg(x)}$ حاصل شود. آنگاه y~h(y) .

برآورد پارامترها

به روش نمونه‌گیری ردی، به تعداد دلخواه از توزیع پسین، نمونه استخراج می‌کنیم و با توجه به تابع زیان، مؤلفه‌های واریانس برآورد می‌شود. به این صورت که اگر تابع زیان را توان دوم خطا در نظر بگیریم، میانگین نمونه‌ها، و چنانچه تابع زیان را قدر مطلق خطا در نظر بگیریم، میانه نمونه‌ها برآورد بیزی از پارامتر خواهد شد. در هر صورت مولفه‌های واریانس برآورد خواهد شد. با در دست داشتن مؤلفه‌های واریانس، توزیع توام γ و β به شرط داده‌ها مطابق با رابطه (۳) معلوم می‌شود. با داشتن توزیع توام، نمونه‌های مورد نیاز استخراج خواهدشد و از روی نمونه‌ها پارامترهای γ و β برآورد می‌شود.

کاربرد

داده‌های مورد استفاده مربوط به چهار نوع رژیم غذایی می‌باشد که کوچ، ایمری و رینفورت (۱۲) آن را به کار برده‌اند. این چهار نوع رژیم به چهار

1. Wolfinger & Kass

گروه افراد تجویز و نمونه خون آنها در پایان هر دوره گرفته شده است و سپس نرمال یا غیر نرمال بودن کلسترول خون برای هر فرد ثبت گردیده است. در این مطالعه سه دوره زمانی (هفته اول، دوم و چهارم) در نظر گرفته شده است. از آنجا که در رژیم غذایی گروه چهارم بعضی از افراد در بعضی از زمان‌ها مراجعه نکرده‌اند با داده‌های گمشده نیز روبرو هستیم. با توجه به اینکه پاسخ ها، (نرمال یا غیر نرمال بودن کلسترول) دوحالتی، طولی و دارای مقادیر گمشده می‌باشند، برای مثال کاربردی مناسب تشخیص داده شد. در اینجا این سؤال را مطرح می‌کنیم که چه سهمی از واریانس پاسخ مربوط به رژیم غذایی و چه سهمی مربوط به سایر عوامل محیطی و ژنتیکی است. جمع بندی پاسخ‌های ثبت شده در جدول ۱ آمده است. در جدول ۱ نرمال بودن را (normal) N را و غیر نرمال بودن را با A (abnormal) نشان داده‌ایم.

ستون‌های مربوط به رژیم‌های ۱ و ۲ و۳ و ۴ را زیر هم نوشته و به صورت یک بردار آورده شد،که آن را P_G نامیدیم. سپس آن را در ماتریس ذیل ضرب کردیم.

$$A=0 \quad \begin{vmatrix} A_1 & 0 & 0 & 0 \\ 0 & A_1 & 0 & 0 \\ 0 & 0 & A_1 & \\ 0 & 0 & 0 & A_2 \end{vmatrix}$$

که درآن A_1 و A_2 به صورت ذیل تعریف می‌شوند.

$$A_1= \begin{vmatrix} 11110000 \\ 00001111 \\ 11001100 \\ 00110011 \\ 10101010 \\ 01010101 \end{vmatrix}$$

$$A_2= \begin{vmatrix} 1100 & & & & & & & & & \\ 0011 & & & & \text{صفر} & & & & & \\ 1010 & & & & & & & & & \\ 0101 & & & & & & & & & \\ & & 1100 & & & & & & & \\ & & 0011 & & & & & & & \\ & & 1010 & & & & & & & \\ & & 0101 & & & & & & & \\ & & & & & 1100 & & & & \\ & & & & & 0011 & & & & \\ & & & & \text{صفر} & 1010 & & & & \\ & & & & & 0101 & & & & \\ & & & & & & & & 10 & \\ & & & & & & & & 01 & \end{vmatrix}$$

سپس $F=AP_G$ را به دست آوردیم و با توجه به آن مقادیر لوجیت با استفاده از رابطه $Y= k \log F$ به دست آورده شد . در عبارت فوق kیک ماتریس قطری است که روی قطر اصلی مقادیر [1- 1] قرار دارند. با عملیات ماتریسی فوق مقادیر Yبه دست آورده شد، کاظم نژاد و مشکانی ثابت کرده‌اند که Yدارای توزیع مجانبی نرمال می‌باشد، (۴).

با تبدیل فوق مقادیر Yدر مدل ۱ مشخص شد و با داشتن ماتریس‌های طرح با روشی که در قسمت نظری به آن اشاره گردید، پارامترهای مدل را برآورد کردیم. پس از برازش مدل نتایج ذیل به دست آورده شد.

نتایج

متغیر تغذیه و زمان را در مدل ۱ وارد کردیم . تغذیه دارای اثر تصادفی و زمان دارای اثر ثابت فرض شد. با در دست داشتن توزیع پسین پارامترها با استفاده از نمونه‌گیری ردی ، پارامترهای مدل با حجم نمونه ۱۰۰۰۰ برآورد شدند که نتایج در جداول ۲ و ۳ آمده است.

تغییرات ناشی از تغذیه (عامل تصادفی) دربرآورد بیزی σ^2_1= ۰/۹۸۴۶۲ و واریانس خطا σ^2_0= ۰/۱۱۱۰۰۴ باشد. به عبارتی ۸۹/۸ درصد کل خطا مربوط به تغذیه است، $\frac{\sigma^2_1}{\sigma^2_0+\sigma^2_1} = \frac{۰/۹۸۴۶۲}{۱/۰۹۵۶۶} = ۰/۸۹۸$ و بقیه مربوط به سایر عوامل محیطی و ژنتیکی است.این نسبت در بر آورد ماکسیمم درستنمایی مقید برابر ۷۸/۳ درصد می‌باشد، (جدول ۲). بر اساس برآورد مؤلفه‌های واریانس فوق ،ضرایب مدل ۱ با استخراج نمونه از توزیع نرمال دو متغیره طبق رابطهٔ (۳) برآورد شده ، که نتایج در جدول (۳) آمده است.

با توجه به جدول (۳) می‌توان گفت: در روش بیزی بازهٔ باورمندی ۹۸ درصدی مربوط به ضریب رگرسیونی زمان عبارت است از (۰/۷۱۰۹۳ ، ۰/۲۵۵۶) این بازه حاکی از آن است که عامل زمان در نرمال شدن کلسترول موثر است و چون ضریب آن مثبت است ، نشان دهندهٔ آن است که با گذشت زمان احتمال نرمال شدن کلسترول افزایش می‌یابد و همچنین تغذیه نوع چهارم اثر معنی داری بر نرمال شدن کلسترول خون دارد.

بحث

تحلیل پاسخ‌های دوحالته طولی به سال ۱۹۶۹ برمی‌گردد که گریزل و همکاران آن را مطرح نمودند.مدل گریزل تنها داده‌های کامل را مورد تجزیه و تحلیل قرار می‌داد. کوچ، ایمری و رینفورت با تعمیم مدل گریزل پاسخ‌های دو حالته طولی ناقص را تجزیه و تحلیل نمودند. ولسون و کلارک در سال ۱۹۸۴ با تغییراتی که در ماتریس تبدیل کوچ، ایمری و رینفورت انجام دادند، توانستند با به کارگیری نرم‌افزارهای در دسترس داده‌های فوق را تحلیل کنند. اما مدل به کار رفته توسط ولسون و کلارک مدل اثرات ثابت را مورد تجزیه و تحلیل قرار می‌داد. در سال ۱۳۷۵، برای اولین بار کاظم‌نژاد و همکاران این مدل را برای اثرات ثابت و تصادفی (مدل آمیخته) تعمیم داده و مؤلفه‌های واریانس را با استفاده از روش سوم هندرسن که مبتنی بر حداقل مربعات وزنی است، برآورد کردند.

جدول ۱. توزیع فراوانی افراد تحت مطالعه برحسب رسته‌های پاسخ و زیرجامعه

زیرجامعه / رسته‌های پاسخ	رژیم غذایی ۴	۳	۲	۱
NNN	2	7	16	31
NNA*	2	2	13	0
NAN	8	5	9	6
ANNN	9	2	3	0
ANN	9	31	14	22
ANA	15	5	4	2
AAB	27	32	15	9
AAA	28	6	6	0
NN-**	1			
NA-	4			
AN-	6			
AA-	12			
N-N	1			
N-A	3			
A-N	7			
A-A	14			
-NN	3			
-NA	4			
-AN	8			
-AA	10			
N- -	6			
A - -	19			

NNA : به این معنی می‌باشد که کلسترول خون در هفته اول نرمال، در هفته دوم نیز نرمال و در هفته چهارم غیرنرمال است و ستون‌ها تعداد افراد را در هر وضعیت نشان می‌دهد.

_NN: به این معنی می‌باشد که در هفته اول و دوم نرمال بوده و در هفته چهارم حضور نداشته است.

روش سوم هندرسن روشی گشتاوری است هر چند روشی ساده و برآوردهای ناریب به دست می‌آورد ولی امکان تولید واریانس صفر و حتی منفی وجود دارد که از نظر کاربردی نتیجه‌ای بی حاصل است. روش به کار رفته در این مقاله با استفاده از نظرهای باکس و تیائو (۱۹۷۳)، سیرل (۱۹۹۳) و ول فینگر و همکاران (۲۰۰۰) توزیعی برای مؤلفه‌های واریانس در نظر گرفته شد و به روش بیزی مؤلفه‌های واریانس برآورد گردید.

یکی از مشکلاتی که روش بیزی نسبت به روش کلاسیک دارد انتخاب پیشین مناسب برای پارامتر است. ما یکی از پیشین‌های مرجع را که پیشین جفریز از مهمترین آنها است را برای مؤلفه‌های واریانس برگزیدیم که مقادیر مثبت را اختیار می‌کند، لذا برآوردهایی که به دست می‌آید مقادیر مثبت دارند، بنابراین امکان تولید واریانس صفر یا منفی وجود ندارد.

از طرفی در روش سوم هندرسن اگر ترتیب به کارگیری متغیرها برای محاسبه پارامتر تغییر کند، نتایج یکسانی به دست نمی‌آید که در روش بیزی چنین مشکلی وجود ندارد. البته برآوردهای بیزی اغلب ناریب نیستند (برخلاف روش سوم هندرسن) اما ناریبی چقدر می‌تواند معیاری برای انتخاب برآوردگرها باشد خود جای بحث دارد.

روش‌های مونت کارلویی این حسن را دارند که می‌توان به هر مقدار دلخواه نمونه انتخاب کرد و خطای معیار را کم کرد. همانطور که در جدول ۱ و ۲ دیده می‌شود روش بیزی نسبت به روش کلاسیک برآوردهایی با خطای معیار کمتری ارائه می‌کند.

از فواید دیگر روش به کار رفته در این مقاله آن است که به سادگی می‌توان از نرم‌افزار SAS با به کارگیری Proc Mixed Model و Proc IML جهت برنامه‌نویسی استفاده نمود.

تقدیر و تشکر

نویسندگان از همکاری کارکنان کتابخانه پروفسور محمد تقی فاطمیِ دانشکده علوم ریاضی دانشگاه فردوسی مشهد، آقایان اتحاد، داوودی نژاد، خالقی و سرکار خانم حسینی تشکر می‌کنند.

جدول ۲. برآورد مؤلفه‌های واریانس به روش بیزی و ماکسیمم درست‌نمایی مقید

مؤلفه‌های واریانس	روش بیزی					روش کلاسیک	
	برآورد بیزی	خطای معیار	میانه	صدک ۱	صدک ۹۹	برآورد ماکسیمم درست‌نمایی مقید	خطای معیار
تغذیه (σ^2_1)	۰/۹۸۴۶۲	۰/۰۳۰۱۷۷	۰/۴۴۷۹	۰/۰۷۲۵	۸/۷۲۱	۰/۳۴۶۷	۰/۳۰۰۹
باقیمانده (σ^2_0)	۰/۱۱۱۰۰۴	۰/۰۰۰۴۸۲	۰/۱۰۰۵۸	۰/۰۴۵۹	۰/۲۸۰۶	۰/۰۹۶۰۵	۰/۰۳۶۲۱

جدول ۳. برآورد ضریب مدل به روش بیزی و ماکسیمم درست‌نمایی مقید

مؤلفه‌های واریانس	روش بیزی					روش کلاسیک	
	برآورد بیزی	خطای معیار	میانه	صدک ۱	صدک ۹۹	برآورد ماکسیمم درست‌نمایی مقید	خطای معیار
عرض از مبدأ	۴/۱۸۷۳	۰/۰۰۵۳	۴/۱۸۵۴	۲/۸۳۰۹	۵/۶۲۹	۴/۱۸۸۳	۰/۳۴۹۶
زمان	۰/۴۸۰۴	۰/۰۰۰۸۳	۰/۴۸۰۰۹	۰/۲۵۵۶۰	۰/۷۱۰۹۳	۰/۴۸۱۰	۰/۰۸۶۲۸
تغذیه ۱	-۰/۶۶۰۱۰	۰/۰۰۵۰۵	-۰/۶۵۰۰۲	-۲/۱۱۵۳	۰/۶۲۵۳	-۰/۶۶۴۴	۰/۳۱۲۵
تغذیه ۲	-۰/۱۲۲۴۹	۰/۰۰۵۱۵	-۰/۱۱۴۷۳	-۱/۵۴۸۵	۱/۱۹۹۸	-۰/۱۲۲۷	۰/۳۲۷۸
تغذیه ۳	۰/۰۷۴۱۸	۰/۰۰۵۱۴	۰/۰۷۸۲۲	-۱/۳۶۴۸	۱/۴۶۷۴	۰/۰۷۶۹۳	۰/۳۲۷۸
تغذیه ۴	۰/۷۰۳۳۴	۰/۰۰۵۱۶۹	۰/۶۹۶۲۳	۰/۶۷۷۴	۲/۱۲۱۳	۰/۷۱۰۲	۰/۳۲۷۸

مراجع

1- *Diggle, J.D. Liang, K.Y.Zeger, S.L. Analysis of Longitudinal data, New York, Oxford University press, (1996).*

2- *Little, R. J. A. Rubin, D.B. Statistical Analysis with Missing Data. New York, John wiley & Sons, (1987).*

3- *Woolson, R. F., Clark, W.C., Analysis of Categorical Incomplete Longitudinal Data, Jornal of the Royal Statistical Society, 147-87 - 99, (1984).*

۴ ـ کاظم نژاد، انوشیروان، مشکانی، محمد رضا. بررسی اثرات تصادفی در مطالعات طولی همراه با داده‌های گمشده. دانشور، سال چهارم، شماره ۱۳ و ۱۴. پاییز و زمستان (۱۳۷۵).

5- *Cristensen, R. Plane answer to Complex Questions the Theory of Linear Models. New York, Springer verlag, (2000).*

6- *Carlin B. P. Louis T. U. Bayes and Empirtical Bayes Methoes for data analysis. (Second edition), Chapman and Hall / CRC, (2000).*

7- *SAS Institute, SAS.STAT User's Guid, version 8, Cary, North Carolina: SAS Institute, (2000).*

8- *Searl, S. R., Casella, G. M. and MCCulloch, C. E., Variance Components, New York, John Wiley & Sons, (1993).*

9- *Kass, R. S., Wasserman L., The selection of prior Distribution by formal valuse. 3.JASA, vol. 91, No. 435, (1996).*

10- *Box, G. E. P., and Tiao, G. C., Bayesian Inference in statistical analysis, Reading, MA: Addision - Wesley,(1973).*

11- *Wolfinger, R. D., Kass, R. E., Nonconjugate bayesian analysis of variance component models. Biometrica, 56:768-774, (2000).*

12- *Koch, C., Imrey, P., and Reinfort, D., Linear model analysis of categorical data with incomplete respons evectors. Biometrics, 28, 633-692, (1972).*