مجله علوم آماری، پاییز و زمستان ۱۳۹۲

جلد ۷، شماره ۲، ص ۲۳۳-۲۴۸

تحلیل دو سطحی با اثرات تصادفی چوله نرمال و مدل‌بندی داده‌های طولی

محمد غلامی فشارکی [۱]، انوشیروان کاظم‌نژاد [۱]، فرید زایری [۲]

[۱] گروه آمار زیستی، دانشگاه تربیت مدرس

[۲] گروه آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی

تاریخ دریافت: ۱۳۹۲/۲/۱۷ تاریخ آخرین بازنگری: ۱۳۹۲/۱۰/۲۶

چکیده : مدل‌سازی داده‌های دوسطحی با فرض نرمال بودن مولفه تصادفی و خطا انجام می‌شود. عدم برقراری این فرض باعث استنباط غلط در مورد پارامترهای مدل می‌گردد. در این مقاله، استفاده از خانواده توزیع چوله نرمال که خانواده‌ای انعطاف‌پذیرتر از توزیع نرمال است مطرح می‌شود. سپس در یک مطالعه شبیه‌سازی نشان داده می‌شود عدم در نظر گرفتن چولگی مثبت (منفی) در مدل باعث بیش برآوردی (کم برآوردی) عرض از مبدا و کم برآوردی (بیش برآوردی) شیب مدل می‌گردد سپس با مدل به‌دست آمده رابطه نوبت کاری و کلسسترول خون تعیین می‌شود.

واژه‌های کلیدی : تحلیل دوسطحی، توزیع چوله نرمال، رهیافت بیزی.

---

* آدرس الکترونیک مسئول مقاله: انوشیروان کاظم‌نژاد، kazem_an@modares.ac.ir

* کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲F۱۵، ۶۰E۰۵

۱ مقدمه

بسیاری از داده‌های حوزه علوم بهداشتی دارای ساختار خوشه ای یا سلسله مراتبی هستند (گلدستین، ۲۰۰۸). از آنجا که استفاده از روش‌های متداول مدل‌بندی رگرسیونی در تحلیل داده‌های خوشه‌ای به دلیل عدم در نظر گرفتن همبستگی بین مشاهدات منجر به کم برآوردی خطای برآورد ضرایب رگرسیونی و معنی‌داری نادرست ضرایب و به تناسب آن افزایش خطای نوع اول می‌گردد (پینهایرو و بتس، ۲۰۰۰) از این رو محققان برای تحلیل داده‌های سلسله مراتبی از روش تحلیل چندسطحی [۱] استفاده می‌نمایند (گلدستین، ۲۰۰۸؛ لوک، ۲۰۰۴).

تحلیل چندسطحی بسط مدل‌های خطی تعمیم‌یافته است که در آن علاوه بر مدل‌بندی متغیر پاسخ ضرایب رگرسیونی نیز مدل‌بندی می‌گردد. این روش کارا در مدل‌سازی داده‌هایی با ساختار آشیانه‌ای بوده و هدف آن مدل‌بندی متغیر وابسته براساس تابعی از متغیرهای پیشگو در بیشتر از یک سطح است (هاکس، ۲۰۱۰). معمولاً برای سادگی فرض می‌شود مولفه‌های تصادفی و خطای مدل دارای توزیع نرمال هستند. از آنجا که عدم برقراری این فرض منجر به برآورد نادرست پارامترهای مدل و استنباط نامعتبر در مورد پارامترهای مدل می‌شود، محققان به روش‌هایی چون تبدیل داده‌ها (به‌ویژه استفاده از تبدیل باکس-کاکس [۲])، استفاده از روش‌های ناپارامتری و توزیع‌های غیر نرمال رو می‌آورند. استفاده از روش‌های ناپارامتری علی رغم کاهش اطلاعات از لحاظ عملی دارای محدودیت‌های زیادی برای محققان بوده و از طرف دیگر استفاده از روش تبدیل متغیرها نیز باعث مشکلاتی چون تفسیر نامناسب پارامترهای تحت بررسی، عدم احراز فرض نرمال توام متغیرها در حالات چند متغیره (تی سانگ و لی، ۲۰۰۳؛ لی و همکاران، ۲۰۰۵؛ جارا و همکاران، ۲۰۰۸)، پنهان شدن فرآیند تولید داده‌ها و کاهش اطلاعات (جارا و همکاران، ۲۰۰۸)، یکتا نبودن تبدیل در مطالعات مختلف می‌گردد (بندیوپدهایی و همکاران، ۲۰۱۰). همچنین استفاده از توزیع‌های غیر نرمال، محقق را از دسترسی به خصوصیات مناسب توزیع‌های بیضوی مانند نرمال دور می‌سازد. راهکار دیگر در

---

۱ Multilevel Modeling

۲ Box-Cox

حل این معضل استفاده از توزیع‌هایی است که به لحاظ نظری قادر به تبیین تغییرات مشاهده بدون استفاده از فرض نرمال و با داشتن خصوصیات توزیع‌های بیضوی می‌باشد که از آن جمله می‌توان به مطالعات پینهایرو و همکاران (۲۰۰۱)، ژو و هی (۲۰۰۸)، روسا و همکاران (۲۰۰۳)، لین و لی (۲۰۰۶ ، ۲۰۰۷ و ۲۰۰۸)، لانگ و سینشیمر (۱۹۹۳)، ما و همکاران (۲۰۰۴)، لاچوس و همکاران (۲۰۰۹)، جارا و همکاران (۲۰۰۸) و بندیوپدهایی و همکاران (۲۰۱۰) در برازش تابع درستنمایی و رهیافت بیزی و همچنین مقایسه روش‌های انتخاب مدل اشاره نمود. اما علی رغم مطالعات متعدد، وجود مطالعه‌ای که در آن به مشاهده تاثیر مولفه تصادفی و خطای غیرنرمال بر پارامترهای یک مدل دوسطحی می پردازد احساس می‌گردد. از این رو این موضوع با انجام یک مطالعه شبیه‌سازی بررسی و نهایتا این مدل آماری در بررسی طولی رابطه نوبت کاری با کلسترول خون به‌کار گرفته شده است.

۲ مدل آماری و تحلیل بیزی

۱.۲ تابع چگالی توام مدل دو سطحی

فرض کنید محققی می‌خواهد تاثیر کار در نوبت کاری را بر میزان کلسترول خون، وقتی که مشاهدات کارگران بصورت طولی جمع‌آوری شده است را اندازه‌گیری نماید. از آنجا که کلسترول خون افراد در زمان‌های مختلفی اندازه گیری شده‌اند، داده‌ها مطالعه طولی و به یکدیگر وابسته هستند. در اینجا مجموعه اندازه‌های گرفته شده در زمان‌های مختلف برای هر یک از اعضای نمونه، به‌عنوان واحدهای سطح اول و هر یک از افراد انتخاب شده به‌عنوان سطح دوم در نظر گرفته می‌شود. این ساختار در مدل های چندسطحی به‌صورت

$$\begin{aligned} y_{i(j)} &= \beta_{0(j)} + \beta_1 x_{i(j)} + \epsilon_{i(j)} \\ \beta_{0(j)} &= \beta_0 + u_{(j)} \end{aligned} \qquad (1)$$

به ترتیب در سطح اول و دوم قابل ارائه است، که با ادغام آن‌ها به‌صورت

$$y_{i(j)} = \beta_0 + \beta_1 x_{i(j)} + u_{(j)} + \epsilon_{i(j)} \qquad i = 1, \ldots, M_j, \;\; j = 1, \ldots, M \qquad (2)$$

تبدیل می‌شود. فرض‌های متداول در برازش مدل‌های چندسطحی، استقلال و نرمال بودن مولفه های تصادفی و خطا است، یعنی

$$Cov(\epsilon_{i(j)}, u_{(j)}) = ۰$$
$$u_{(j)} \sim N(۰, \sigma_u^۲), \quad \epsilon_{i(j)} \sim N(۰, \sigma_\epsilon^۲) \qquad (۳)$$

چون فرض نرمال بودن مولفه تصادفی و خطا در بسیاری از مواقع به دلیل مشاهدات دور افتاده و یا فرم نامتقارن توزیعی برقرار نیستند می‌توان به‌جای آن از توزیع چوله نرمال استفاده نمود.

$$u_{(j)} \sim SN(۰, \lambda_u, \sigma_u^۲), \quad \epsilon_{i(j)} \sim SN(۰, \lambda_\epsilon, \sigma_\epsilon^۲) \qquad (۴)$$

که در آن، ($Z \sim SN(۰, \sigma^۲, \lambda)$) نماد توزیع چوله نرمال با میانگین صفر، واریانس $\sigma^۲$ و پارامتر چولگی λ با تابع چگالی (آزالینی، ۱۹۸۵)

$$f(z) = \frac{۲}{\sigma}\varphi(\frac{z-\mu}{\sigma})\Phi(\lambda\frac{z-\mu}{\sigma}) \qquad (۵)$$

است به‌طوری‌که $\varphi(۰)$ و $\Phi(۰)$ به ترتیب توابع چگالی و توزیع احتمال نرمال استاندارد هستند. آزالینی (۱۹۸۵) نشان داد اگر $X_۱$ و $X_۲$ هر دو دارای توزیع نرمال استاندارد باشند، آنگاه توزیع متغیر تصادفی ، Z

$$Z = \lambda|X_۱| + X_۲ \qquad (۶)$$

چوله نرمال استاندارد با پارامتر λ است. بنابراین دو بردار تصادفی u و ϵ را می‌توان به‌صورت

$$\begin{aligned} u_{(j)} &= \lambda_u|t_{(j)}^{u_۱}| + t_{(j)}^{u_۲}, & t_{(j)}^{u_۱}, t_{(j)}^{u_۲} &\sim N(۰, \sigma_u^۲) \\ \epsilon_{i(j)} &= \lambda_\epsilon|t_{i(j)}^{\epsilon_۱}| + t_{i(j)}^{\epsilon_۲}, & t_{i(j)}^{\epsilon_۲}, t_{i(j)}^{\epsilon_۲} &\sim N(۰, \sigma_\epsilon^۲) \end{aligned} \qquad (۷)$$

نوشت. با در نظر گرفتن

$$X = \begin{pmatrix} ۱ & x_{۱۱} \\ . & . \\ . & . \\ . & . \\ ۱ & x_{MM} \end{pmatrix}, \quad \beta' = \begin{pmatrix} \beta_۰ \\ \beta_۱ \end{pmatrix}, \quad Y = \begin{pmatrix} y_{۱۱} \\ . \\ . \\ . \\ y_{M۱} \end{pmatrix}, \quad T^{\epsilon_۱} = \begin{pmatrix} t_{۱۱}^{\epsilon_۱} \\ . \\ . \\ . \\ t_{M۱}^{\epsilon_۱} \end{pmatrix},$$

$$U = \begin{pmatrix} u_1 \\ . \\ . \\ . \\ u_{M1} \end{pmatrix}, \quad T^{u_1} = \begin{pmatrix} t_1^{u_1} \\ . \\ . \\ . \\ t_M^{u_1} \end{pmatrix},$$

و با توجه به ساختار آشیانی داده‌ها، تابع چگالی توام داده‌ها براساس مولفه‌های تصادفی را می‌توان به‌صورت

$$f(Y, T^{\epsilon_1}, U, T^{u_1}) = f(Y|T^{\epsilon_1}, U, T^{u_1}) f(T^{\epsilon_1}|U, T^{u_1}) f(U|T^{u_1}) f(T^{u_1}) \quad (۸)$$

نوشت. با در نظر گرفتن $n = \sum_{j=1}^{M} M_k$ داریم

$$Y|T^{\epsilon_1}, U, T^{u_1} \sim N_n(X\beta + U + \lambda_\epsilon T^{\epsilon_1{}^t}, \sigma_\epsilon^2 I_n), \quad T^{u_1} \sim N_M(0, \sigma_u^2 I_M),$$
$$U|T^{u_1} \sim N_M(\lambda_u T^{u_1{}^t}, \sigma_u^2 I_M), \qquad T^{\epsilon_1}|U, T^{u_1} \sim N_n(0, \sigma_\epsilon^2 I_n) \quad (۹)$$

بنابراین تابع چگالی توام به‌صورت زیر حاصل می‌شود.

$$\begin{aligned} f(Y, T^{\epsilon_1}, U, T^{u_1}) = & \\ & \prod_{j=1}^{M} \prod_{i=1}^{M_j} (2\pi\sigma_\epsilon^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma_\epsilon^2}(y_{i(j)} - \beta_0 - \beta_1 x_{i(j)} - u_{(j)} - \lambda_\epsilon |t_{i(j)}^{\epsilon_1}|\}^2 \\ & \times \prod_{j=1}^{M} \prod_{i=1}^{M_j} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\{-\frac{t_{i(j)}^{\epsilon_1}{}^2}{2\sigma_\epsilon^2}\} \\ & \times \prod_{j=1}^{M} \sqrt{\frac{1}{2\pi\sigma_u^2}} \exp\{-\frac{(u_j - \lambda_u |t_{(j)}^{u_1}|)^2}{2\sigma_u^2}\} \\ & \times \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\{-\frac{t_{(j)}^{u_1}{}^2}{2\sigma_u^2}\} \end{aligned} \quad (۱۰)$$

۲.۲ توزیع پیشین و توزیع پسین

در این مطالعه توزیع‌های توابع پیشین با استفاده از کارهای هوبرت و کسلا (۱۹۹۶) و ژاو و هممکاران (۲۰۰۶) به‌صورت جدول ۱ انتخاب گردید. با در نظر گرفتن

جدول ۱: توزیع‌های پیشین برای پارامترهای مدل

ضرایب بتا	پارامتر چولگی	پارامتر پراکندگی
$\beta_0 \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0})$	$\lambda_u \sim N(\mu_{\lambda_u}, \sigma^2_{\lambda_u})$	$\sigma^2_u \sim IG(\alpha_u, \beta_u)$
$\beta_1 \sim N(\mu_{\beta_1}, \sigma^2_{\beta_1})$	$\lambda_u \sim N(\mu_{\lambda_\epsilon}, \sigma^2_{\lambda_\epsilon})$	$\sigma^2_\epsilon \sim IG(\alpha_\epsilon, \beta_\epsilon)$

$\theta = (\beta_0, \beta_1, \sigma^2_u, \sigma^2_e, \lambda_u, \lambda_\epsilon)$ و مستقل فرض کردن پارامترهای مدل، توزیع توام پیشین θ به‌صورت

$$\pi(\theta) = \pi(\beta_0)\pi(\beta_1)\pi(\sigma^2_u)\pi(\sigma^2_e)\pi(\lambda_u)\pi(\lambda_\epsilon) \qquad (11)$$

است. با ضرب توزیع پیشین θ در تابع چگالی توام، توزیع‌های شرطی کامل پارامترها با انجام محاسبات تحلیلی به‌صورت زیر حاصل می‌شوند.

$$\beta_0|others \sim N\Big(\frac{\frac{\mu_{\beta_0}}{\sigma^2_{\beta_0}} + \frac{1}{\sigma^2_\epsilon}\sum_{j=1}^{M}\sum_{i=1}^{M_k}(y_{i(j)} - \beta_1 x_{i(j)} - u_{j(k)} - \lambda_\epsilon |t^{\epsilon}_{i(j)}|)}{\frac{1}{\sigma^2_{\beta_0}} + \frac{n}{\sigma^2_\epsilon}}, \Big(\frac{1}{\sigma^2_{\beta_0}} + \frac{n}{\sigma^2_\epsilon}\Big)^{-1}\Big)$$

$$\beta_1|others \sim N\Big(\frac{\frac{\mu_{\beta_1}}{\sigma^2_{\beta_1}} + \frac{1}{\sigma^2_\epsilon}\sum_{j=1}^{M}\sum_{i=1}^{M_k}(y_{i(j)} - \beta_0 - u_j - \lambda_\epsilon |t^{\epsilon}_{i(j)}|)x_{i(jk)}}{\frac{1}{\sigma^2_{\beta_1}} + \frac{\sum_{j=1}^{M}\sum_{i=1}^{M_k} x^2_{i(jk)}}{\sigma^2_\epsilon}}, \Big(\frac{1}{\sigma^2_{\beta_1}} + \frac{\sum_{j=1}^{M}\sum_{i=1}^{M_k} x^2_{i(jk)}}{\sigma^2_\epsilon}\Big)^{-1}\Big)$$

$$\lambda_\epsilon|others \sim N\Big(\frac{\frac{\mu_{\lambda_\epsilon}}{\sigma^2_{\lambda_\epsilon}} + \frac{1}{\sigma^2_\epsilon}\sum_{j=1}^{M}\sum_{i=1}^{M_k}(y_{i(j)} - \beta_0 - \beta_1 xi(j) - u_{(j)})|t^{\epsilon}_{i(j)}|}{\frac{1}{\sigma^2_{\lambda_\epsilon}} + \frac{1}{\sigma^2_\epsilon}\sum_{j=1}^{M}\sum_{i=1}^{M_k} t^{\epsilon\,2}_{i(j)}}, \Big(\frac{1}{\sigma^2_{\lambda_\epsilon}} + \frac{1}{\sigma^2_\epsilon}\sum_{j=1}^{M}\sum_{i=1}^{M_k} t^{\epsilon\,2}_{i(j)}\Big)^{-1}\Big)$$

$$\lambda_u|others \sim N\Big(\frac{\frac{\mu_{\lambda_u}}{\sigma^2_{\lambda_u}} + \frac{1}{\sigma^2_u}\sum_{j=1}^{M}\sum_{i=1}^{M_k}|t^{u}_{i(j)}|}{\frac{1}{\sigma^2_{\lambda_u}} + \frac{1}{\sigma^2_u}\sum_{j=1}^{M}\sum_{i=1}^{M_k} t^{u\,2}_{j(i)}}, \Big(\frac{1}{\sigma^2_{\lambda_u}} + \frac{1}{\sigma^2_u}\sum_{j=1}^{M}\sum_{i=1}^{M_k} t^{u\,2}_{j(i)}\Big)^{-1}\Big)$$

$$\begin{aligned}
\sigma_\epsilon^{-2}|others &\sim IG(n+\alpha_\epsilon, 0/5\Sigma_{j=1}^{M}\Sigma_{i=1}^{M_k}(t_{i(j)}^{\epsilon 1})^2 \\
&\quad +(y_{i(j)}-\beta_0-\beta_1 x_{i(j)}-u(j)-\lambda|t_{i(j)}^{\epsilon 1}|)^2)+\beta_\epsilon) \\
\sigma_u^{-2}|others &\sim IG(M+\alpha_u, 0/5\Sigma_{j=1}^{M}(t_{(j)}^{u 1})^2+(u_{(j)}-\lambda_u|t_{(j)}^{u 1}|)^2)+\beta_u)
\end{aligned}$$

۳ مطالعه شبیه‌سازی

برای شبیه‌سازی، ابتدا مدل دوسطحی (۱۲) را با پارامترهای $\beta_0 = 2$، $\beta_1 = -3$، $x_{i(j)} = (x_1, x_2, \ldots, x_{50})$، $x_j = (1, 2, 3, 4, 5)$ در نظر گرفته شد، سپس با چهار استراتژی، ۱- مولفه خطا نرمال، مولفه تصادفی نرمال، ۲- مولفه خطا چوله نرمال، مولفه تصادفی نرمال، ۳- مولفه خطا نرمال، مولفه تصادفی چوله نرمال و ۴- مولفه خطا چوله نرمال، مولفه تصادفی چوله نرمال و با استفاده از مقادیر جدول ۲ و به کمک دستور sn در بسته نرم افزاری R شبیه‌سازی شده است. سپس در هر شبیه سازی هر چهار مدل با بسته نرم‌افزاری R2WinBUGS و نرم افزار WinBUGS به داده ها برازش داده شده است. این عمل ۱۰۰ بار تکرار و خلاصه نتایج در جدول ۳ و ۴ ارائه شده است.

در برازش بیزی بعد از بررسی نمودارهای لازم، زمان داغیدن ۱۰۰۰ تعیین شده سپس ۱۰۰۰۰ نمونه از توزیع پسین استخراج و از هر ۱۰۰ نمونه یک نمونه برای تحلیل نهایی در نظر گرفته شده است. همچنین در این مطالعه برای β_0، β_1 و λ توزیع پیشین $N(0, 100)$ و برای σ_u^2 و σ_e^2 توزیع پیشین $IG(0/001, 0/001)$ در نظر گرفته شده است.

$$y_{i(j)} = \beta_0 + \beta_1 x_{i(j)} + u_{(j)} + \epsilon_{i(j)}, \quad i = 1, \ldots, 5, \quad j = 1, \ldots, 50 \qquad (12)$$

برای مقایسه مدل‌های برازش داده شده از ملاک متوسط قدرمطلق خطا به‌صورت

$$MAE(\theta_k) = \frac{\Sigma_{i=1}^{n}|\hat{\theta}_{ik} - \theta_k|}{n} \qquad (13)$$

استفاده شده است، که در آن θ_k پارامتر مورد بررسی و $\hat{\theta}_k$ برآورد بیزی آن است.

جدول ۲: چهار استراتژی برای مولفه تصادفی $u_{(j)}$ و $\epsilon_{i(j)}$

استراتژی	نماد	توزیع مولفه تصادفی	
		$u_{(j)}$	$\epsilon_{i(j)}$
۱	$N-N$	$N(۰,۴)$	$N(۰,۹)$
۲	$SN-N$	$N(۰,۴)$	$SN(۰,۹,۳)$
۳	$N-SN$	$SN(۰,۴,۲)$	$N(۰,۹)$
۴	$SN-SN$	$SN(۰,۴,۲)$	$SN(۰,۹,۳)$

نتایج شبیه‌سازی برآورد پارامتر، میزان خطای برآورد براساس ملاک MAE و درصد بیش برآوردی در جداول ۳ و ۴ خلاصه شده ند. میزان خطای واقعی برابر با میزان خطای محاسبه شده براساس شاخص MAE است. اما میزان خطای مربوط به مدل میزان انحراف برآورده‌ها در ۱۰۰ بار شبیه‌سازی است. همان‌طور که ملاحظه می‌شود در مدل دوسطحی وقتی مولفه تصادفی و مولفه خطا هر دو نرمال هستند، برازش مدل‌های چوله در مولفه تصادفی و خطای مدل باعث کم برآوردی واریانس مولفه چوله می‌شود. این در حالی است که هنگامی که واقعا مولفه خطا چوله است، و ما از مدل دوسطحی معمولی (توزیع خطا نرمال و توزیع مولفه تصادفی نرمال استفاده نماییم) در صورتی که پارامتر چولگی مثبت باشد مقدار عرض از مبدا با بیش برآوردی و مقدار ضریب متغیر کمکی با کم برآوردی مواجه و در صورت چولگی منفی این رفتار با کم برآوردی عرض از مبدا و بیش برآوردی ضریب متغیر تبیینی مواجه خواهد داد. از طرف دیگر نتایج این شبیه‌سازی نشان داد که در سطح دوم حتی وقتی داده‌ها واقعا چوله هستند، پارامتر چولگی معنی دار نمی‌شود.

۴ مثال کاربردی

داده های این پژوهش از نوع مطالعات طولی گذشته‌نگر بوده و جامعه مورد مطالعه آن را کلیه کارکنان شاغل در کارخانه فولاد مبارکه اصفهان طی سالهای ۱۳۷۰ تا ۱۳۹۰ تشکیل می‌داد. تعداد هم‌گروه مورد نظر در این مطالعه ۵۷۴ نفر بوده که از میان ۶۷۱۳ نفر پرسنل این کارخانه با روش نمونه‌گیری طبقه‌ای خوشه‌ای از نواحی کاری و سپس مشاهدات یک فرد با مراجعه به پرونده‌های پزشکی کارکنان انجام

پذیرفت. هدف از این مطالعه بررسی رابطه کار در نوبت کاری و شاخص چاقی (BMI) با کلسترول خون بود. در شکل ۱ بافت‌نگار میزان کلسترول و متوسط کلسترول هر فرد نمایش داده شده است. میزان چولگی برای کلیه مشاهدات ۴/۱۶ و برای میانگین هر نفر در زمانهای متفاوت ۳/۳۷ محاسبه گردید. براساس مقادیر چولگی و بافت‌نگار به خوبی می‌توان به اختلاف اساسی فرض نرمال بودن این دو مولفه در مدلسازی دوسطحی اطمینان نمود.

جدول ۳: برآورد خطای واقعی برآورد پارامترها

مدل در نظر گرفته شده				برآورد خطای واقعی پارامترهای مدل					
مدل واقعی	e	u		β_0	β_1	σ_e^2	σ_u^2	λ_e	λ_u
$N-N$	N	N	E	۰/۳۸	۰/۰۹	۰/۷۲	۰/۸۵	---	---
			P	۰/۴۸	۰/۵۸	۰/۵۷	۰/۴۲	---	---
		SN	E	۱/۰۹	۰/۱	۲/۲	۰/۸۶	۲/۰۴	---
			P	۰/۴۶	۰/۵۵	۰/۰۲	۰/۴۱	۰/۰۱	---
$SN-N$	SN	N	E	۰/۹۴	۰/۰۹	۰/۷۲	۲/۰۸	---	۳/۳
			P	۰/۵۱	۰/۵۸	۰/۵۳	صفر	---	۰/۰۳
		SN	E	۱/۱۶	۰/۰۹	۲/۲۳	۲/۰۶	۲/۰۶	۳/۲۴
			P	۰/۵۴	۰/۵۷	۰/۰۱	صفر	۲/۰۶	۰/۰۳
$N-SN$	N	N	E	۶/۸۶	۰/۵	۱۵/۲۲	۰/۷۴	---	---
			P	۱	صفر	۱	صفر	---	---
		SN	E	۱/۵۵	۰/۴۹	۱/۲	۰/۲	۰/۵۷	---
			P	۱	صفر	صفر	صفر	۱	---
	SN	N	E	۷/۳۹	۰/۵	۱۴/۸۹	۲/۴۸	---	۴/۲۷
			P	۱	صفر	۱	صفر	---	صفر
		SN	E	۲/۵۳	۰/۴۹	۰/۴۶	۲/۵۷	۰/۳۶	۴/۴۷
			P	۱	صفر	صفر	صفر	۰/۳۶	صفر
$N-SN$	N	N	E	۴/۷۱	۰/۱	۰/۷۸	۱۳/۲۷	---	---
			P	۱	۰/۶	۰/۴۷	۱	---	---
		SN	E	۴/۲۶	۰/۱	۲/۲۷	۱۳/۳۳	۱/۷۴	---
			P	۱	۰/۵۸	۰/۰۳	۱	صفر	---
$SN-SN$	SN	N	E	۱/۶۴	۰/۱	۰/۷۸	۲/۹	---	۱/۱۲
			P	۰/۷۶	۰/۵۹	۰/۴۵	۰/۸۵	---	۰/۴۰
		SN	E	۱/۷۵	۰/۱	۲/۳۵	۲/۸۲	۱/۷۸	۱/۱۲
			P	۰/۶۲	۰/۵۴	۰/۰۲	۰/۸۱	۱/۷۸	۰/۴۸
$SN-SN$	N	N	E	۹/۴۸	۰/۱۵	۱۳/۳۳	۱۴/۶۷	---	---
			P	۱	۰/۵۷	۱	۱	---	---
		SN	E	۵/۲۷	۰/۱۶	۳/۰۸	۱۴/۶۴	۰/۷۹	---
			P	۱	۰/۵۵	۰/۵۴	۱	۰/۵۷	---
	SN	N	E	۶/۰۵	۰/۱۵	۱۳/۱۸	۳/۳۱	---	۱/۰۸
			P	۱	۰/۵۴	۱	۰/۸۴	---	۰/۴۱
		SN	E	۲/۵۳	۰/۱۶	۳/۲۳	۳/۳۸	۰/۸۳	/۲۵
			P	۰/۸	۰/۵۲	۰/۵۶	۰/۸۹	۰/۸۳	۰/۴۱

E: متوسط میزان خطا براساس شاخص MSE، P: درصد بیش برآوردی

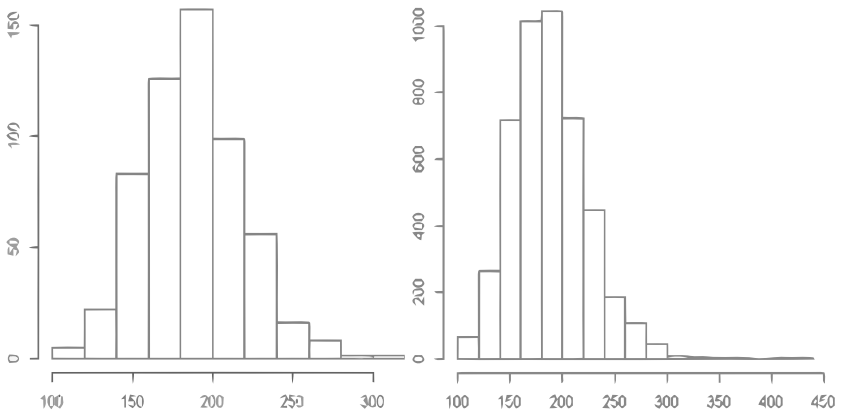
جدول ۴: برآورد میانگین و انحراف معیار برآورد پارامترها

برآورد پارامترهای مدل						مدل در نظر گرفته شده			
λ_u	λ_e	σ_u^2	σ_e^2	β_1	β_0		u	e	مدل واقعی
---	۳/۹۲	---	۹/۲۱	-۲/۹۸	۱/۹۰	C	N	N	
---	۱/۲۱	---	/۹۳	۰/۱۴	۰/۵۳	D			
---	۳/۹۵	-۰/۰۴	۶/۸۱	-۲/۹۹	۱/۹۷	C	N	SN	
---	۱/۲۵	۰/۸۸	۱/۸۱	۰/۱۳	۱/۶۱	D			$N-N$
-۰/۲۶	۱/۹۲	---	۹/۱۸	-۲/۹۹	۲/۰۵	C	SN	N	
۲/۳۷	۱/۳۶	---	۰/۹۴	۰/۱۳	۱/۶۱	D			
۱/۹۴	۱/۹۴	-۰/۰۴	۶/ ۷۸	-۲/۹۹	۲/۱۲	C	SN	SN	
۱/۳۹	۱/۳۹	۰/۹۲	۱/۸۶	۰/۱۳	۲/۱۷	D			
---	۳/۲۶	---	۲۴/۲۲	-۳/۵	۲/۸۶	C	N	N	
---	۱/۹۵	---	۲/۵۵	۰/۲۲	۰/۷۷	D			
---	۳/۸	۲/۵۷	۷/۸	-۳/۴۹	۳/۵۵	C	N	SN	
---	۱/۶۸	۰/۷۷	۳/۲۴	۰/۲۱	۰/۹۷	D			$SN-N$
-۱/۲۷	۱/۵۲	---	۲۳/۸۹	-۳/۵	۹/۳۹	C	SN	N	
۳/۱۹	۱/۵۶	---	۲/۴۳	۰/۲۲	۱/۸۷	D			
۱/۴۳	۱/۱۴۳	۲/۳۶	۸/۵۴	-۳/۴۹	۴/۵۳	C	SN	SN	
۱/۳۷	۱/۳۷	۰/۶۳	۲/۸۸	۰/۲۱	۱/۸۵	D			
---	۱۷/۲۷	---	۹/۱	-۲/۹۹	۶/۷۱	C	N	N	
---	۳/۹۸	---	۰/۹	۰/۱۳	۰/۷۴	D			
---	۱۷/۳۳	۰/۲۶	۶/۷۴	-۲/۹۹	۶/۲۶	C	N	SN	
---	۴/۰۴	۰/۹۴	۱/۸۸	۰/۱۳	۱/۷۵	D			$N-SN$
۲/۵۲	۶/۶۴	---	۹/۰۵	-۲/۹۹	۳/۳۶	C	SN	N	
۱/۸۲	۴/۴	---	۰/۹۲	۰/۱۳	۲	D			
۶/۵۲	۶/۵۲	۰/۲۳	۶/۶۷	-۳	۲/۹	C	SN	SN	
۴/۳۸	۴/۳۸	۰/۹۴	۱/۸۵	۰/۱۳	۲/۵	D			
---	۱۸/۶۷	---	۲۲/۳۳	-۲/۹۷	۱۱/۴۸	C	N	N	
---	۴/۸۴	---	۲/۲۲	۰/۲۱	۰/۹۳	D			
---	۱۸/۶۵	۲/۱۶	۹/۷۱	-۲/۹۸	۷/۲۷	C	N	SN	
---	۴/۸۸	۰/۹۶	۳/۸۷	۰/۲	۱/۶۱	D			$SN-SN$
۲/۵۸	۷/۰۱	---	۲۲/۱۸	-۲/۹۸	۸/۰۵	C	SN	N	
۱/۹۹	۵/۱۳	---	۲/۲۶	۰/۲۱	۲/۲۹	D			
۷/۱۸	۷/۱۸	۲/۱۲	۹/۸۴	-۲/۹۹	۴/۰۴	C	SN	SN	
۵/۰۴	۵/۰۴	۰/۹۵	۳/۸۷	۰/۲	۲/۶۲	D			

در مدل برازش داده شده (۱۴) متغیر پاسخ کلسترول خون ($Cholesterol$)، متغیر نشانگر کار در نوبت کار چرخشی ($Shift\mathbf{1}$) ،متغیر نشانگر کار در نوبت کار هفتگی ($Shift\mathbf{1}$)، شاخص توده بدنی (BMI)،مولفه تصادفی تکرار ($u_{(j)}$) و مولفه تصادفی خطا ($\epsilon_{i(j)}$) می‌باشد.

$$\begin{aligned} Cholesterol_{i(j)} &= \beta_0 + \beta_1 Shift\mathbf{1}_{i(j)} + \beta_2 Shift\mathbf{2}_{i(j)} \\ +\beta_3 BMI_{i(j)} &+ u_{(j)} + \epsilon_{i(j)}, \quad i = \mathbf{1},\ldots,n_j \quad j = \mathbf{1},\ldots,\mathbf{574} \end{aligned} \quad (۱۴)$$

پس از بررسی توصیفی داده‌های مطالعه، جهت انجام تحلیل بیزی بعد از بررسی



(ب) (الف)

شکل ۱: (الف) بافت‌نگار میزان کلسترول کلیه مشاهدات، (ب) بافت‌نگار میانگین میزان کلسترول هر فرد .

نمودارهای لازم ، زمان داغیدن ۱۰۰۰ تعیین گردید. سپس ۱۰۰۰۰ نمونه از توزیع پسین استخراج گردید. از هر ۱۰۰ نمونه به دست آمده یکی به عنوان نمونه نهایی در نظر گرفته شد. به عبارتی دیگر در پایان از ۱۰۰ نمونه برای استنباط‌ها استفاده گردید. در این تحلیل برای مقادیر β و λ از توزیع پیشین $N(۰, ۱۰۰)$ و برای پارامترهای واریانس مدل از توزیع پیشین $IG(۰/۰۰۱, ۰.۰۰۱)$ استفاده شده است.

برای مقایسه مدل‌های برازش داده شده از ملاک اطلاع انحراف $(DIC = \overline{D(\theta)} + P_D)$ استفاده شد، که در آن $\overline{D(\theta)}$ میانگین پسین انحراف بوده و میزان برازش را نشان می‌دهد، P_D برابر با تعداد پارامترهای موثر است و میزان پیچیدگی مدل را نشان می‌دهد. بر اساس این معیار مدلی که دارای کمترین مقدار DIC باشد به عنوان بهترین مدل انتخاب می‌شود. این معیار برای هر اندازه نمونه قابل استفاده بوده و به آسانی توسط روش‌های مونت کار لوی زنجیره مارکوفی قابل محاسبه است (اسپیگرهالتر و همکاران، ۲۰۰۲)

در جدول ۵ خلاصه اطلاعات مربوط به مقادیر DIC محاسبه شده برای ۴ مدل

ارائه شده‌اند. با توجه به مقادیر DIC مدل شماره ۲ (توزیع نرمال چوله برای مولفه خطا و توزیع نرمال برای مولفه های تصادفی فرد) به عنوان بهترین مدل برازش داده شده انتخاب گردید. نتایج حاصل از برآورد و فواصل اطمینان ناپارامتری بیزی ضرایب در جدول ۶ نشان‌دهنده آن است که متغیر BMI بر کلسترول خون تاثیر مثبت معنی‌دار دارد و متغیر نوبت کاری بر میزان کلسترول تاثیر معنی‌داری ندارد. همچنین پارامتر چولگی، واریانس سطح اول و خطا معنی‌دار است.

جدول ۵: خلاصه اطلاعات DIC برای مدل های برازش داده شده

شماره مدل	مولفه تصادفی		pD	$\overline{D(\theta)}$	DIC	زمان برازش
	$u_{(j)}$	$\epsilon_{i(j)}$				
۱	نرمال	نرمال	۵۳۵	۴۱۲۵۵	۴۱۷۶۰	۱۵
۲	نرمال	چوله نرمال	۳۲۱۰	۳۷۱۴۰	۴۰۳۵۰	۱۸۲
۳	چوله نرمال	نرمال	۵۳۱	۴۰۱۵۹	۴۰۶۹۰	۲۹
۴	چوله نرمال	چوله نرمال	۲۵۰۹	۳۷۹۳۱	۴۰۴۴۰	۱۴۷

جدول ۶: نتایج حاصل از برآورد و بازه اطمینان ناپارامتری بیزی ضرایب

	β_0	β_1	β_2	β_3	σ_e^2	σ_u^2	λ_e
برآورد	۷۲/۹۲	−۳/۹۰	−۲/۶۸	۳/۷۷	۲۱۰/۴	۷۴۱/۶	۱/۸۵
انحراف معیار	۵/۵۵	۲/۷۳	۴/۷۱	۰/۲۰	۲۹/۷۱	۵۷/۷۹	۰/۴۱
حد پایین	۶۱/۲۷	−۸/۷۱	−۱۱/۸۴	۲/۳۰	۸۳/۳۵	۶۵۰/۸	۱/۵۵
حد بالا	۸۵/۲۶	۳/۱۴	۶/۶۲	۴/۲۱	۲۴۵/۱	۸۴۷/۵	۳/۸۱

۵ بحث و نتیجه‌گیری

در این مقاله با توجه به اینکه کلسترول خون متغیری ذاتاً چوله است (لاچوس و همکاران، ۲۰۰۹) بنابراین انتظار ما پیش از برازش مدل این بود که انتخاب مدل دوسطحی معمول برازش ضعیفی را نسبت به مدل‌های چوله ارائه نماید. این مطلب با برازش ۴ مدل پیشنهادی مورد بررسی قرار گرفته و همانگونه که مقادیر DIC ارائه شده در جدول ۵ نمایش می‌دهد از بین تمام مدل‌های برازش داده شده مدل شماره ۲ (توزیع نرمال چوله برای مولفه خطا و توزیع نرمال برای مولفه‌های تصادفی فرد)

دارای برازش بهتری نسبت به مدل دوسطحی معمول با توزیع نرمال بوده و توزیع های چوله بخوبی توانسته بود تا اثر تخطی داده‌ها از توزیع نرمال را کنترل نماید.

همچنین نتایج حاصل از مطالعه شبیه‌سازی نشان داد که در حضور چولگی مثبت مولفه خطای مدل، مقدار عرض از مبدا بیشتر از مقدار واقعی برآورد و در صورتی که پارامتر چولگی منفی باشد این میزان کمتر از مقدار واقعی برآورد می‌شود. بیش برآوردی عرض از مبدا باعث کم برآوردی ضریب متغیر تبیینی و همچنین کم برآوردی عرض از مبدا باعث بیش برآوردی ضرایب متغیرهای تبیینی می‌گردد (نتر و همکاران، ۱۹۸۷) نتایج بیشتر نشان دهنده عدم معنی داری پارامتر چولگی متغیر تصادفی سطح دوم در داده‌هایی که سطح دوم آن از توزیع چوله شبیه‌سازی شده بود می‌باشد. از آنجایی که برآورد تقریبی مولفه تصادفی خوشه iام برابر با $\overline{y_i} - \overline{y}$ است، که در آن $\overline{y_i}$ متوسط متغیر پاسخ برای همه مشاهدات خوشه iام می‌باشد و بخاطر متوسط گیری و طبق قضیه حد مرکزی که میانگین نمونه به توزیع نرمال میل می‌نماید می‌توان گفت که چولگی مولفه تصادفی سطح دو با متوسط گیری به نوعی از داده‌ها حذف می‌شود.

بنابراین استفاده از خانواده توزیع‌های نامتقارن به جای در نظر گرفتن توزیع نرمال برای مولفه‌های تصادفی و خطای مدل روشی کارا در تحلیل داده‌های دوسطحی با متغیر چوله نرمال بوده و به عنوان جایگزینی مناسب برای روش هایی مانند روش تبدیل متغیر و استفاده از روش‌های ناپارامتری پیشنهاد می‌گردد.

تقدیر و تشکر

نویسندگان از پیشنهادات ارزنده داوران گرامی مجله که باعث ارائه بهتر و بهبود مقاله شده است، کمال تشکر را دارند.

مراجع

Arellano-Valle, R. B., Bolfarine, H. and Lachos V. H. (2005), Skew-Normal Linear Mixed Models, *Journal of Data Science*, **3**, 415-438.

Azzalini A. (1985), A Class of Distributions which Includes the Normal Ones, *Scandinavian Journal of Statistics Theory and Applications*, **12**, 171-178.

Bandyopadhyay, D., Lachos, V. H., Abanto-Valle, C. A. and Ghosh, P. (2010), Linear Mixed Models for Skew-Normal/Independent Bivariate Responses with an Application to Periodontal Disease, *Statistics in Medicine*, **29**, 2643-2655.

Goldstein, H. (2005), *Handbook of Multilevel Analysis*, Springer, New York.

Hobert, J. and Casellas, G. (1996), The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models, *Journal of the American Statistical Association*, **91**, 1461-1473.

Hox, J. J. (2010), *Multilevel Analysis: Techniques and Applications (Quantitative Methodology Series) Great Britain*, Routledge.

Jara, A., Quintana, F. and San Martin, E. (2008), Linear Mixed Models with Skew-Elliptical Distributions: a Bayesian Approach, *Computational Statistics and Data Analysis*, **52**, 5033-5045.

Lachos, V. H., Bolfarine, H. and Arellano-Valle, B. (2009), Bayesian Inference for Skew-Normal Linear Mixed Models, *Journal of Applied Statistics*, **34**, 663-682.

Lange, K. and Sinsheimer J. S. (1993), Normal/Independent Distributions and Their Applications in Robust Regression, *Journal of Computational and Graphical Statistics*, **2**, 175-198.

Lee, J. C., Lin, T. I., Lee, K. J. and Hsu, Y. L. (2005), Bayesian Analysis of Box-Cox Transformed Linear Mixed Models wtth (p,q)ARMA Dependence, *Journal of Statistical Planning and Inference*, **133**, 435-451.

Lin, T. I. and Lee, J. C. (2006), A Robust Approach to t Linear Mixed Models Applied to Multiple Sclerosis Data, *Statistics in Medicine*, **25**, 1397-1412.

Lin, T. I. and Lee, J. C. (2007), Bayesian Analysis of hierarchical Linear Mixed Modeling Using Multivariate t Distributions, *Journal of Statistical Planning and Inference*, **137**, 484-495.

Lin, T. I. and Lee, J. C. (2008), Estimation and Prediction in Linear Mixed Models with Skew-Normal Random Effects for Longitudinal Data, *Statistics in Medicine*, **27**, 1490-1507.

Luke, A. D. (2004), *Multilevel Modeling (Quantitative Applications in the Social Sciences*, A Sage University Paper Series.

Ma, Y. M., Genton, G. and Davidian, M. (2004), *Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality, Linear Mixed Models with Flexible Generalized Skew-Elliptical Random Effect*, Chapman & Hall, CRC Press, Boca Raton, FL.

Neter, J., Wasserman, W. and Whitmore, G. A. (1987), *Appliede Statistics*, Boston Allyn and Bacon Publisher.

Pinheiro, J. C. and D. M. Bates (2000), *Mixed-Effects Models in S and S-PLUS*, New York, Springer-Verlag.

Pinheiro, J. C., Liu, C. H. and Wu, Y. N. (2001), Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using a Multivariate t-Distribution, *Journal of Computational and Graphical Statistics*, **10**, 249-276.

Rosa, G. J. M., Padovani, C. R. and Gianola, D. (2003), Robust Linear Mixed Models with Normal/Independent Distributions and Bayesian MCMC Implementation, *Biometrical Journal*, **45**, 573-590.

Spiegelhalter, D. J., Best, N .G., Carlin, B. P. and Van Der Linde, A. (2002), Bayesian Measures of Model Complexity and Fit (with discussion), *Journal of Royal Statistical Society*, B, **64**, 583-640.

Tsung, I. L. and Lee, J. C. (2003), On Modelling Data from Degradation Sample Paths over Time, *Australian and New Zealand Journal of statistics*, **45**, 257-270.

Zhao, Y., Staudenmayer, J., Coull, B. and Wand, M. (2006), General Design Bayesian Generalized Linear Mixed Models, *Statistical Science*, **21**, 35-51.

Zhou, T. and He, X. (2008), Three-Step Estimation in Linear Mixed Models with Skew-t Distributions, *Journal of Statistical Planning and Inference*, **138**, 1542-1555.