

مجله علوم آماری، پاییز و زمستان ۱۳۹۴

جلد ۹، شماره ۲، ص ۱۴۹-۱۶۷

## رگرسیون چندکی بیزی با تاوان لاسو و لاسوی تطبیق پذیر برای داده‌های طولی دودویی

علی آقامحمدی، سکینه محمدی

گروه آمار، دانشگاه زنجان

تاریخ دریافت: ۱۳۹۳/۳/۱۱ تاریخ آخرین بازنگری: ۱۳۹۳/۱۱/۴

**چکیده:** در بسیاری از مطالعات علوم پزشکی برای بیان سیر بیماری و تأثیر درمان از مطالعات طولی استفاده می‌شود، که در آن پاسخ‌ها به طور مکرر در طول زمان اندازه‌گیری می‌شوند. اما گاهی این پاسخ‌ها دو حالتی و گسسته هستند. در این مقاله مدل رگرسیون چندکی با تاوان لاسو و لاسوی تطبیق پذیر برای داده‌های طولی با پاسخ‌های دو حالتی ارائه شده و هر دو روش از دیدگاه آمار بیزی مورد تحلیل قرار می‌گیرد. با توجه به این که در هر دو روش توزیع‌های پسینی پارامترها به شکل بسته قابل حصول نیستند، توزیع‌های پسینی شرطی کامل پارامترها محاسبه شده و از الگوریتم نمونه‌گیری گیبز برای استنباط استفاده می‌شود. برای مقایسه کارایی روش‌های ارائه شده با روش‌های متداول، مطالعه شبیه‌سازی انجام شده و در پایان نیز نحوه کاربرد مدل‌ها در قالب مثال کاربردی شرح داده خواهد شد.

**واژه‌های کلیدی:** رگرسیون چندکی دودویی، تاوان لاسو، تاوان لاسوی تطبیق پذیر، داده‌های طولی، نمونه‌گیری گیبز، استنباط بیزی.

---

آدرس الکترونیک مسئول مقاله: علی آقامحمدی، [Aghamohammadi.ali@znu.ac.ir](mailto:Aghamohammadi.ali@znu.ac.ir)  
کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲J۹۹، ۶۲F۱۵

در بسیاری از مطالعه‌های مربوط به علوم پزشکی و اجتماعی، متغیر پاسخ برای هر فرد در چندین نوبت متوالی مشاهده می‌شود. برای مثال یک پزشک، سلامت بیماران را به طور هفتگی ارزیابی می‌کند تا دریابد، آیا داروی جدید موفقیتی داشته است یا خیر. به مطالعه‌ای که اندازه‌گیری مربوط به یک صفت در طول زمان مورد بررسی قرار می‌گیرد، مطالعه طولی<sup>۱</sup> گویند. ممکن است فاصله‌های زمانی که هر واحد آزمایشی مورد مشاهده قرار می‌گیرد، یکسان نباشد. مهم‌ترین مزیت این نوع مطالعات تفکیک اثر زمان از تفاوت‌های فردی است که در مطالعه‌های مقطعی (پاسخ برای هر واحد آزمایشی، تنها یک بار مشاهده می‌شود) امکان‌پذیر نیست. لذا مطالعه‌های طولی نسبت به مطالعه‌های مقطعی دارای دقت بیشتری هستند (غلامی فشارکی و همکاران، ۱۳۹۲). گاهی پاسخ‌های ثبت شده به صورت کیفی، به‌ویژه دو حالتی هستند. به‌عنوان مثال تأثیر یا عدم تأثیر یک داروی خاص روی نوعی بیماری. بدیهی است تحلیل این نوع داده‌ها با توجه به نوع متغیر پاسخ که دودویی است، روش‌های خاص خود را می‌طلبد. یکی از روش‌های تحلیل این متغیرها، رگرسیون چندکی<sup>۲</sup> دودویی است که مانند رگرسیون چندکی در مقایسه با روش رگرسیون میانگین دارای دو مزیت مهم است. نخست این که رگرسیون چندکی دودویی نسبت به ناپایداری واریانس و داده‌های دورافتاده حساس نیست و دوم این که اطلاعات جزئی‌تری نسبت به تأثیر متغیرهای توصیفی در چندک‌های مختلف توزیع متغیر پاسخ ارائه می‌دهد. به همین دلیل مطالعاتی نیز در زمینه تحلیل داده‌های طولی با پاسخ‌های دودویی صورت گرفته است. فیزماریس و لارد (۱۹۹۳) روشی بر پایه درست‌نمایی برای تحلیل داده‌های طولی با پاسخ‌های دو حالتی ارائه کردند. سپس گانکلوز (۲۰۰۲) این روش را برای تحلیل داده‌های طولی گسسته تعمیم داد. گانکلوز و آزلاینی (۲۰۰۸) استفاده از زنجیرهای مارکوفی را در روش درست‌نمایی مورد مطالعه قرار دادند. در ادامه گانکلوز و همکاران (۲۰۱۲) روش رگرسیون لوزستیک را برای تحلیل داده‌های طولی با پاسخ‌های دودویی ارائه کردند. در مطالعه

<sup>۱</sup> Longitudinal study

<sup>۲</sup> Quantile regression

داده‌های طولی علاوه بر اثرات ثابت که تغییرات درون گروه‌ها (تغییرات در طول زمان) را کنترل می‌کند، اثرات تصادفی بین متغیرهای پاسخ نیز در مدل لحاظ می‌شود که در واقع تغییرات بین گروهی را کنترل می‌کند. در بسیاری از مسائل، تعداد مشاهدات از متغیر پاسخ نسبت به مشاهدات انجام گرفته در طول زمان از متغیر مربوطه، زیاد است. بنابراین در این مدل‌ها تعداد اثرات تصادفی (تعداد پارامترها) نیز خیلی زیاد شده و لذا تعبیر و تفسیر مدل مشکل می‌شود. کوئنکر (۲۰۰۴) رگرسیون چندکی را برای تحلیل این داده‌ها با افزودن تاوان لاسو روی اثرات تصادفی را مورد بررسی قرار داد که با ایجاد تاوان، اثرهای تصادفی را به سمت صفر منقبض<sup>۳</sup> کرده و اثرهای کم اهمیت را از مدل حذف و مدلی تنک<sup>۴</sup> ایجاد می‌کند.

اولین بار تیشیرانی (۱۹۹۶) رگرسیون لاسو که با ایجاد تاوان لاسو روی پارامترها انجام می‌شود را ارائه کرد. سپس زو (۲۰۰۶) رگرسیون لاسوی تطبیق‌پذیر<sup>۵</sup> را به منظور گسترش رگرسیون لاسو پیشنهاد کرد. در این روش برخلاف تاوان لاسو که برای کلیه ضرایب مدل اندازه تاوان را یکسان در نظر می‌گیرد، اندازه تاوان‌های مختلفی برای ضرایب رگرسیونی متفاوت لحاظ می‌شود. بررسی‌های زو (۲۰۰۶) مشخص کرد که در رگرسیون با تاوان لاسوی تطبیق‌پذیر اریبی در برآورد پارامترها در مقایسه با روش لاسو کمتر است. الحمزوی و همکاران (۲۰۱۲) نیز با تعریف توزیع پیشینی به صورت لاپلاس متقارن روی ضرایب رگرسیونی، مدل رگرسیون چندکی با تاوان لاسوی تطبیق‌پذیر را برای داده‌های مقطعی مورد مطالعه قرار دادند. در این مقاله هدف بررسی رگرسیون چندکی با تاوان لاسو و لاسوی تطبیق‌پذیر روی اثرات تصادفی در داده‌های طولی با پاسخ‌های دو حالتی از دیدگاه آمار بیزی است. برای این منظور تاوان لاسو و لاسوی تطبیق‌پذیر روی اثرات تصادفی در داده‌های طولی دودویی از طریق تعریف توزیع‌های پیشینی مناسب در نظر گرفته می‌شود و هر دو روش از دیدگاه آمار بیزی مورد تحلیل و ارزیابی قرار می‌گیرند. در بخش ۲ مطالب کلی در مورد مدل رگرسیون چندکی برای

<sup>۳</sup> Shrink

<sup>۴</sup> Sparse

<sup>۵</sup> Adaptive Lasso

داده‌های طولی بیان می‌شود. در بخش‌های ۳ و ۴ مدل رگرسیون چندکی دودویی با توان لاسو و لاسوی تطبیق‌پذیر از دیدگاه آمار بیزی ارائه می‌شود. در بخش ۵ در مطالعه‌ای شبیه‌سازی، کارایی مدل‌های ارائه شده با مدل‌های دیگر مورد مقایسه قرار می‌گیرد. بخش ۶ نیز شامل تحلیل داده‌های واقعی و ارزیابی کارایی مدل‌ها است.

## ۲ تعریف مدل

مدل رگرسیون خطی

$$y_i = x_i' \beta + \epsilon_i, \quad i = 1, \dots, n$$

را در نظر بگیرید، که در آن بردار مربوط به متغیرهای توصیفی،  $\beta$  یک بردار  $1 \times k$  بعدی از پارامترها و  $\epsilon_i$  مولفه خطا است. در مدل رگرسیونی میانگین، هدف استنباط در مورد متوسط مقدار متغیر پاسخ به ازای مقادیر متفاوت متغیرهای توصیفی است، لذا  $E(y_i|x_i) = x_i' \beta$  مورد توجه است. اما در رگرسیون چندکی، تابع چندک شرطی

$$Q_\tau(y_i|x_i) = x_i' \beta_\tau \quad i = 1, \dots, n$$

مورد توجه است، که در آن  $Q_\tau(\cdot)$  معکوس تابع توزیع تجمعی متغیر پاسخ  $y_i$  به شرط معلوم بودن بردار  $x_i$  است. به عبارت دیگر  $x_i' \beta_\tau$  چندک  $\tau$ ام شرطی متغیر  $y_i$  را نشان می‌دهد. در رگرسیون معمولی میانگین توزیع خطاها صفر در نظر گرفته می‌شود، اما در رگرسیون چندکی، چندک  $\tau$ ام  $\epsilon_i$ ها برابر صفر هستند، یعنی  $\int_{-\infty}^{\tau} f_\tau(\epsilon_i) d\epsilon_i = \tau$  برای  $i = 1, \dots, n$ . در رگرسیون چندکی، ضرایب رگرسیونی یعنی پارامتر  $\beta_\tau$  از مینیمم کردن رابطه

$$\sum_{i=1}^n \rho_\tau(y_i - x_i' \beta_\tau), \quad (1)$$

نسبت به  $\beta_\tau$  برآورد می‌شود، که در آن  $\rho_\tau(u) = \frac{|u| + (\tau-1)u}{2}$  تابع زیان است (کوئنکر و ماکادو، ۱۹۹۹). چون رابطه (۱) نسبت به  $\beta_\tau$  در مبدأ مشتق پذیر نیست، راه‌حل تحلیلی برای برآورد ضرایب رگرسیونی وجود ندارد. برای رفع این مشکل

کوئنکر و ماکادو (۱۹۹۹) استفاده از توزیع لاپلاس نامتقارن را پیشنهاد کردند. متغیر تصادفی  $Y$  را دارای توزیع لاپلاس نامتقارن<sup>۶</sup> (ALD) با پارامترهای  $\tau$ ،  $\sigma$  و  $\mu$  گویند و با نماد  $Y \sim ALD(\mu, \sigma, \tau)$  نشان می‌دهند، هر گاه تابع چگالی آن به صورت

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_{\tau}\left(\frac{y-\mu}{\sigma}\right)\right\},$$

باشد، که در آن  $0 < \tau < 1$  پارامتر چولگی،  $\sigma$  پارامتر مقیاس و  $-\infty < \mu < +\infty$  پارامتر مکانی است. با فرض این که  $Y_i \sim ALD(x'_i\beta_{\tau}, \sigma, \tau)$ ، آن گاه تابع درستنمایی به صورت

$$L(\beta_{\tau}|y, x) \propto \left(\frac{1}{\sigma}\right)^n \exp\left\{-\sum_{i=1}^n \rho_{\tau}\left(\frac{y_i - x'_i\beta_{\tau}}{\sigma}\right)\right\}$$

به دست می‌آید، که ماکسیمم کردن آن در حضور پارامتر مزاحم  $\sigma$  با مینیمم کردن رابطه (۱) نسبت به  $\beta_{\tau}$  معادل است. به همین دلیل در بسیاری از موارد، تحلیل رگرسیون چندکی با استفاده از این توزیع مورد بررسی قرار می‌گیرد. مدل رگرسیون چندکی برای داده‌های طولی با اثرات تصادفی به صورت

$$y_{ij} = x'_{ij}\beta_{\tau} + \alpha_i + \epsilon_{ij}; \quad i = 1, \dots, n, \quad j = 1, \dots, n_i, \quad \sum_i n_i = N$$

تعریف می‌شود، که در آن  $y_{ij}$ ،  $j$ امین پاسخ اندازه‌گیری شده روی  $i$ امین واحد آزمایشی و  $x_{ij}$  بردار مربوط به متغیرهای توصیفی و  $\beta_{\tau}$  بردار  $1 \times k$  بعدی از پارامترها و  $\alpha_i$  اثر تصادفی مربوط به پاسخ  $y_{ij}$  را نشان می‌دهد. تابع چندک شرطی در این مدل به صورت  $Q_{\tau}(y_{ij}|x_{ij}) = x'_{ij}\beta_{\tau} + \alpha_i$  بیان می‌شود. با فرض نتوان لاسو روی اثرات تصادفی به صورت  $p(\alpha) = \sum_{i=1}^n |\alpha_i|$  برآورد پارامترهای  $\beta_{\tau}$  و  $\alpha = (\alpha_1, \dots, \alpha_n)$  از دیدگاه آمار بسامدی از مینیمم کردن تابع زیان

$$\sum_{i=1}^n \sum_{j=1}^{n_i} \rho_{\tau}(y_{ij} - x'_{ij}\beta_{\tau} - \alpha_i) + \lambda \sum_{i=1}^n |\alpha_i|$$

نسبت به  $\beta_{\tau}$  و  $\alpha$  به دست می‌آید (کوئنکر، ۲۰۰۴)، که در آن  $\lambda$  را پارامتر تاوانیدن گویند و میزان انقباض به سمت صفر پارامترهای  $\alpha_i$  را کنترل می‌کند. هر چه مقدار

<sup>۶</sup> Asymmetric Laplace Distribution

آن بزرگتر باشد، انقباض به صفر نیز بیشتر و اثرهای تصادفی زیادی از مدل حذف خواهند شد. در این رهیافت برآورد  $\beta_\tau$  تحت تأثیر اندازه  $\lambda$  است که معمولاً از روش اعتبارسنجی متقابل<sup>۷</sup> محاسبه می‌شود.

### ۳ رگرسیون چندکی دودویی با تاوان لاسو

یکی از روش‌های معمول برای تعریف مدل‌های رگرسیونی با پاسخ‌های دودویی استفاده از متغیرهای پنهان<sup>۸</sup> است (بنویت و همکاران، ۲۰۱۳). مدل رگرسیون چندکی دودویی در داده‌های طولی به صورت

$$y_{ij}^* = x'_{ij}\beta_\tau + \alpha_i + \epsilon_{ij}, \quad y_{ij} = g(y_{ij}^*), \quad i = 1, \dots, n, \quad j = 1, \dots, n_i, \quad (2)$$

تعریف می‌شود، که در آن  $g(\cdot)$  تابع پیوند<sup>۹</sup> و به صورت  $g(y_{ij}^*) = I(y_{ij}^* > 0)$  تعریف شده  $I(\cdot)$  تابع مشخصه است) و  $y_{ij}$ ، زامین پاسخ مشاهده شده برای  $i$ امین واحد آزمایشی و  $y_{ij}^*$  متغیر پنهان و غیر قابل مشاهده است. چون چندک  $\tau$ ام  $\epsilon_{ij}$ ها برابر صفر فرض می‌شوند،  $x'_{ij}\beta_\tau + \alpha_i$  چندک  $\tau$ ام  $y_{ij}^*$  خواهد بود. از آنجایی که تابع پیوند یک تابع یکنوا است،  $g(x'_{ij}\beta_\tau + \alpha_i)$  چندک  $\tau$ ام متغیر  $y_{ij} = g(y_{ij}^*)$  است. بنابراین در این مدل تابع زیان با تاوان لاسو روی اثرات تصادفی به صورت

$$\sum_{i=1}^n \sum_{j=1}^{n_i} \rho_\tau(y_{ij}^* - x'_{ij}\beta_\tau - \alpha_i) + \lambda \sum_{i=1}^n |\alpha_i| \quad (3)$$

به دست می‌آید. با فرض این که  $\epsilon_{ij} \sim ALD(0, \sigma, \tau)$  و توزیع پیش‌بینی برای  $\alpha_i$  به صورت  $\nu = \frac{\lambda}{\sqrt{\sigma}} \exp\{-\frac{\lambda|\alpha_i|}{\sqrt{\sigma}}\}$  در نظر گرفته شود، به ازای  $\nu = \frac{\lambda}{\sqrt{\sigma}}$  توزیع پسینی  $\alpha$  به صورت

$$\pi(\alpha|y, x, \sigma, \lambda, \beta_\tau) = \sigma^{-N} \exp\{-\sigma^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} \rho_\tau(y_{ij}^* - x'_{ij}\beta_\tau - \alpha_i)\} \prod_{i=1}^n \frac{\nu}{\sqrt{\sigma}} \exp\{-\nu|\alpha_i|\} \quad (4)$$

<sup>۷</sup> Cross validation

<sup>۸</sup> Hidden variable

<sup>۹</sup> Link function

به دست می آید. مینیمم کردن تابع هدف (۳) نسبت به  $\alpha$  و  $\beta_\tau$  معادل ماکسیمم کردن تابع درست‌نمایی (۴) در حضور پارامتر مزاحم  $\sigma$  است. بنابراین می توان توزیع لاپلاس نامتقارن را برای این مدل نیز مورد استفاده قرار داد. چون استفاده از تابع درست‌نمایی توزیع لاپلاس نامتقارن به دلیل وجود قدرمطلق آسان نیست، در تحلیل های بیزی معمولاً از شکل آمیخته این توزیع استفاده می شود. اگر  $u \sim ALD(0, \sigma, \tau)$  و متغیرهای تصادفی  $z$  و  $e$  به ترتیب دارای توزیع نرمال استاندارد و نمایی با میانگین  $\frac{\sigma}{\tau(1-\tau)}$  و مستقل از هم باشند، آنگاه

$$u = k_1 e + \sqrt{2\sigma} z$$

که در آن  $k_1 = (1 - 2\tau)$  (کازومی و کابایاشی، ۲۰۱۱). با استفاده از این خاصیت توزیع لاپلاس نامتقارن و با توجه به این که  $\epsilon_{ij} \sim ALD(0, \sigma, \tau)$ ، مدل رابطه (۲) به صورت

$$y_{ij}^* = x'_{ij}\beta_\tau + \alpha_i + k_1 e_{ij} + \sqrt{2\sigma} z_{ij} \quad (5)$$

خواهد بود، که در آن  $z_{ij}$  و  $e_{ij}$  به ترتیب دارای توزیع نرمال استاندارد و نمایی با میانگین  $\frac{\sigma}{\tau(1-\tau)}$  و مستقل از هم هستند. چون توزیع پیشینی  $\alpha_i$  ها لاپلاس متقارن است، می توان آن را به صورت توزیع آمیخته از دو توزیع

$$\alpha_i | s_i \sim N(0, s_i), \quad s_i | \nu^2 \sim \text{Exp}\left(\frac{\nu^2}{\tau}\right) \quad (6)$$

نوشت، که در آن  $\text{Exp}\left(\frac{\nu^2}{\tau}\right)$  توزیع نمایی با میانگین  $\frac{\tau}{\nu^2}$  همان توزیع آمیختگی است (پارک و کسلا، ۲۰۰۸). چون هدف تحلیل بیزی است، برای ابرپارامتر  $\nu^2$  (به جای  $\nu$ ) توزیع پیشینی مزدوج نمایی به صورت  $\pi(\nu^2 | \phi) = \phi \exp\{-\phi \nu^2\}$  و برای  $\sigma$  و  $\beta_\tau$  به ترتیب وارون گاما و توزیع نرمال چند متغیره در نظر گرفته می شود. حال با توجه به مدل رابطه (۵)، (۶) و توزیع های پیشینی تعریف شده، مدل سلسله مراتبی را می توان به صورت

$$y_{ij} = g(y_{ij}^*), \quad y_{ij}^* | e_{ij}, \beta_\tau, \sigma, \alpha_i \sim N(x'_{ij}\beta_\tau + \alpha_i + k_1 e_{ij}, 2\sigma e_{ij}), \\ e_{ij} | \sigma \sim \text{Exp}\left(\frac{\tau(1-\tau)}{\sigma}\right), \quad \beta_\tau | b_0, B_0 \sim N_k(b_0, B_0), \quad \alpha_i | s_i \sim N(0, s_i),$$

$$s_i | \nu^2 \sim \text{Exp}\left(\frac{\nu^2}{\sigma}\right), \nu^2 | \phi \sim \text{Exp}(\phi), \pi(\phi) \propto \frac{1}{\phi}, \sigma | c_0, d_0 \sim IG(c_0, d_0),$$

در نظر گرفت، که در آن  $IG(a, b)$  توزیع وارون گاما با پارامترهای  $a$  و  $b$  است. با توجه به شکل سلسله مراتبی فوق، توزیع پیسینی کلیه پارامترها و متغیرهای پنهان به صورت

$$\begin{aligned} \pi(y^*, \beta_\tau, \sigma, s, \nu^2, \alpha, \phi | y, x) &\propto \prod_{i=1}^n \prod_{j=1}^{n_i} \frac{1}{\sqrt{\nu^2 \sigma \pi e_{ij}}} \\ &\times \exp\left\{-\frac{1}{\nu^2 \sigma e_{ij}} (y_{ij}^* - x'_{ij} \beta_\tau - \alpha_i - k_1 e_{ij})^2\right\} \\ &\times \{I(y_{ij} = 0)I(y_{ij}^* \leq 0) + I(y_{ij} = 1)I(y_{ij}^* > 0)\} \\ &\times \left(\frac{1}{\sigma}\right)^N \exp\left\{-\sum_{i=1}^n \sum_{j=1}^{n_i} \frac{e_{ij}}{\sigma}\right\} \times \prod_{i=1}^n \frac{1}{\sqrt{\nu^2 \pi s_i}} \exp\left\{-\frac{\alpha_i^2}{\nu^2 s_i}\right\} \frac{\nu^2}{\nu^2} \exp\left\{-\frac{\nu^2 s_i}{\nu^2}\right\} \\ &\times \exp\left\{-\frac{1}{\nu^2} (\beta_\tau - b_0) B_0^{-1} (\beta_\tau - b_0)\right\} \exp\{-\phi \nu^2\} \times \left(\frac{1}{\sigma}\right)^{c_0+1} \exp\left\{-\frac{d_0}{\sigma}\right\} \end{aligned}$$

به دست می آیند، که در آن  $s = (s_1, \dots, s_n)$ . ملاحظه می شود که توزیع پیسینی کامل پارامترها و متغیرهای پنهان توزیع شناخته شده ای نیست. اما با اندکی محاسبات جبری توزیع های پیسینی شرطی کامل متغیرهای پنهان و پارامترها به صورت

$$\pi(y_{ij}^* | y_{ij}, e_{ij}, \alpha, \beta_\tau) = \begin{cases} N(x'_{ij} \beta_\tau + \alpha_i + k_1 e_{ij}, \nu^2 \sigma e_{ij}) I(y_{ij}^* > 0), & y_{ij} = 1 \\ N(x'_{ij} \beta_\tau + \alpha_i + k_1 e_{ij}, \nu^2 \sigma e_{ij}) I(y_{ij}^* \leq 0), & y_{ij} = 0 \end{cases}$$

$$\pi(e_{ij} | \cdot) \sim GIG\left(\frac{1}{\nu^2}, \hat{\gamma}_{ij}, \hat{\delta}_{ij}\right)$$

به دست می آیند، که در آن  $\hat{\gamma}_{ij} = \frac{(y_{ij}^* - x'_{ij} \beta_\tau - \alpha_i)^2}{\nu^2 \sigma}$  و  $\hat{\delta}_{ij} = \frac{k_1^2}{\nu^2 \sigma} + \frac{\nu^2 (1-\tau)}{\sigma}$ .  $GIG(r, m, n)$  نشان دهنده توزیع تعمیم یافته وارون نرمال<sup>۱۰</sup> با پارامترهای  $r, m$  و  $n$  است. تابع چگالی این توزیع به صورت

$$f(x | r, m, n) \propto x^{r-1} \exp\left\{-\frac{1}{\nu^2} (m^2 x^{-1} + n^2 x)\right\}, x > 0, -\infty < r < \infty, m, n > 0$$

است، که با استفاده از تابع  $rgig()$  در پکیج  $ghyp$  در نرم افزار  $R$  می توان از این توزیع نمونه تصادفی تولید کرد (لیوزی و بریمن، ۲۰۱۴).

<sup>۱۰</sup> Generalized Invers Gaussian



توزیع شرطی کامل  $\sigma$  نیز به صورت

$$\pi(\sigma|\cdot) \sim IG(\nu_1, w_1)$$

است، که در آن  $w_1 = \sum_{i=1}^n \sum_{j=1}^{n_i} \left( \frac{(y_{ij}^* - x'_{ij}\beta_\tau - \alpha_i - k_1 e_{ij})^2}{\nu e_{ij}} + \tau(1-\tau)e_{ij} \right) + d_0$ .

$\nu_1 = \frac{\nu N}{\nu} + c_0$ . همچنین  $\pi(\alpha_i|\cdot) \sim N(\bar{\mu}_i, \bar{\sigma}_i)$  که در آن

$$\bar{\mu}_i = \bar{\sigma}_i \left( \frac{1}{\nu\sigma} \sum_{j=1}^{n_i} \frac{(y_{ij}^* - x'_{ij}\beta_\tau - k_1 e_{ij})}{e_{ij}} \right), \quad \bar{\sigma}_i = \left( \frac{1}{s_i} + \frac{1}{\nu\sigma} \sum_{j=1}^{n_i} \frac{1}{e_{ij}} \right)^{-1}$$

$$\pi(\phi|\cdot) \sim Exp(\nu^2), \quad \pi(s_i|\cdot) \sim GIG\left(\frac{1}{\nu}, \sqrt{\alpha_i^2}, \sqrt{\nu^2}\right),$$

به علاوه  $\pi(\nu^2|\cdot) \sim G(n+1, \frac{\sum_{i=1}^n s_i}{\nu} + \phi)$  که در آن  $G(a, b)$  نشان دهنده توزیع

گاما با پارامترهای  $a$  و  $b$  است. بالاخره توزیع پسینی شرطی کامل  $\beta_\tau$  به صورت

$$\pi(\beta_\tau|\cdot) \sim N_k(Bb, B)$$

به دست می آید، که در آن  $b = \frac{1}{\nu\sigma} \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{x_{ij}(y_{ij}^* - \alpha_i - k_1 e_{ij})}{e_{ij}} + B_0^{-1} b_0$ .

حالت با مشخص شدن توزیع پسینی  $B^{-1} = \frac{1}{\nu\sigma} \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{x_{ij}x'_{ij}}{e_{ij}} + B_0^{-1}$ .

پارامترها و متغیرهای پنهان می توان با روش نمونه گیری گیبز نمونه های کافی تولید کرده و استنباطها را انجام داد.

#### ۴ رگرسیون چندکی دودویی با تاوان لاسوی تطبیق پذیر

در این روش به منظور بسط روش قبل اندازه تاوان های متفاوتی روی هر یک از اثرات تصادفی مدل در نظر گرفته می شود. بنابراین تابع زیان در این مدل به صورت

$$\sum_{i=1}^n \sum_{j=1}^{n_i} \rho_\tau(y_{ij} - g(x'_{ij}\beta_\tau + \alpha_i)) + \sum_{i=1}^n \lambda_i |\alpha_i| \quad (7)$$

تعریف می شود، که در آن عبارت سمت راست را تاوان لاسوی تطبیق پذیر گویند (زو، ۲۰۰۶). اگر در این مدل توزیع پیشینی برای  $\alpha_i$  ها به صورت

$\pi(\alpha_i|\lambda_i, \sigma) = \frac{\lambda_i}{\nu\sqrt{\sigma}} \exp\left\{-\frac{\lambda_i|\alpha_i|}{\nu\sqrt{\sigma}}\right\}$  تعریف شود و قرار دهیم،  $\nu_i = \frac{\lambda_i}{\nu\sqrt{\sigma}}$ ، آنگاه توزیع

پسینی بردار  $\alpha$  به صورت

$$\pi(\alpha|y, x, \sigma, \nu, \beta_\tau) =$$

$$(\sigma)^{-N} \exp\{-\sigma^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} \rho_{\tau}(y_{ij}^* - x'_{ij} \beta_{\tau} - \alpha_i)\} \prod_{i=1}^n \frac{\nu_i}{\gamma} \exp\{-\nu_i |\alpha_i|\}$$

به دست می آید. مینیمم کردن رابطه (۷) با ماکسیمم کردن توزیع پسینی  $\alpha$  در حضور پارامتر مزاحم  $\sigma$  معادل است. لذا در این مدل نیز می توان از توزیع لاپلاس نامتقارن استفاده کرد. اگر توزیع پیشینی  $\alpha_i$  به شکل آمیخته از دو توزیع

$$\alpha_i | s_i \sim N(0, s_i), \quad s_i | \nu_i \sim \text{Exp}\left(\frac{\nu_i}{\gamma}\right)$$

در نظر گرفته شود، با استفاده از شکل آمیخته توزیع لاپلاس نامتقارن، مدل را می توان مانند بخش ۳ به صورت سلسله مراتبی زیر نوشت:

$$\begin{aligned} y_{ij} &= g(y_{ij}^*), \quad y_{ij}^* | \beta_{\tau}, \sigma, e, \alpha_i \sim N(x'_{ij} \beta_{\tau} + \alpha_i + k_1 e_{ij}, \gamma \sigma e_{ij}) \\ e_{ij} | \sigma &\sim \text{Exp}\left(\frac{\tau(1-\tau)}{\sigma}\right), \quad \beta_{\tau} | b_0, B_0 \sim N_k(b_0, B_0), \quad s_i | \nu_i \sim \text{Exp}\left(\frac{\nu_i}{\gamma}\right) \\ \alpha_i | s_i &\sim N(0, s_i), \quad \nu_i | \phi \sim \text{Exp}(\phi), \quad \pi(\phi) \propto \frac{1}{\phi}, \quad \sigma | c_0, d_0 \sim IG(c_0, d_0). \end{aligned}$$

توزیع پسینی کامل پارامترها و متغیرهای پنهان در این مدل نیز به شکل بسته قابل حصول نیستند. اما توزیع پسینی شرطی کامل پارامترها و متغیرهای پنهان قابل محاسبه اند. توزیع های پسینی  $y^*$ ,  $\beta_{\tau}$ ,  $\alpha$  و  $\sigma$  مانند بخش ۳ می باشد. اما توزیع های پسینی برای سایر پارامترها در این مدل به صورت

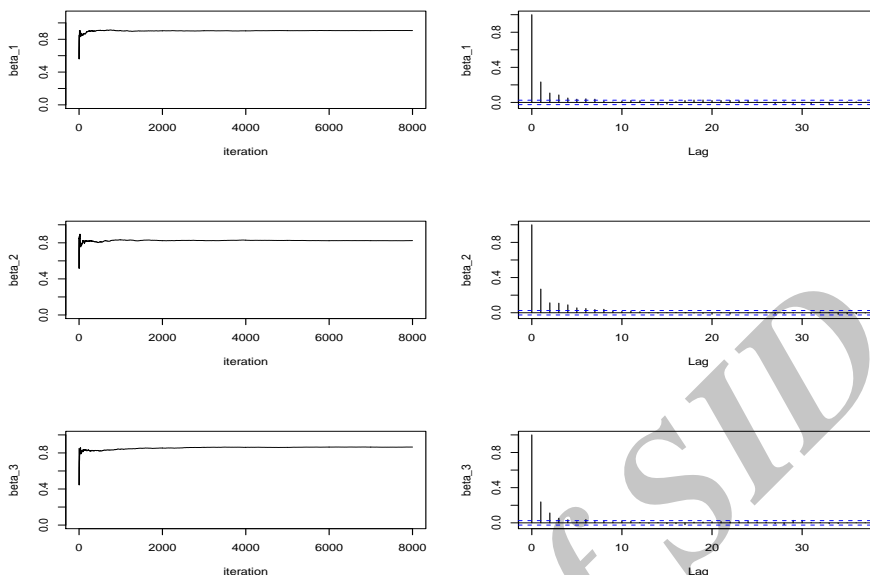
$$\begin{aligned} \pi(\nu_i | \cdot) &\sim G\left(\gamma, \frac{s_i}{\gamma} + \phi\right), \quad \pi(\phi | \cdot) \sim G(n, \sum_{i=1}^n \nu_i), \\ \pi(s_i | \cdot) &\sim GIG\left(\frac{1}{\gamma}, \sqrt{\alpha_i}, \sqrt{\nu_i}\right), \end{aligned}$$

به دست می آیند. در این روش نیز با مشخص شدن توزیع پسینی هر یک از پارامترها و متغیرهای پنهان می توان با روش نمونه گیری گیبز، نمونه های کافی تولید کرده و استنباطها را انجام داد.

## ۵ مطالعه شبیه سازی

در این بخش برای مقایسه کارایی روش های ارائه شده با روش های متداول دیگر مطالعه ای شبیه سازی انجام می پذیرد. مدل به صورت

$$y_{ij}^* = x'_{ij} \beta_1 + x'_{ij} \beta_2 + x'_{ij} \beta_3 + x'_{ij} \beta_4 + x'_{ij} \beta_5 + x'_{ij} \beta_6 + \alpha_i + \epsilon_{ij}$$



شکل ۱: نمودار اثر و تابع خودهمبستگی در روش رگرسیون چندکی با تاوان لاسو ( $\tau = 0/5$ )

در نظر گرفته شده است، که در آن  $i = 1, \dots, 50$  و  $j = 1, \dots, 5$ . توجه شود که تعداد متغیرهای پاسخ در مقایسه با متغیرهای توصیفی زیاد است. متغیرهای توصیفی یعنی  $x_{ij}^k$  ها به صورت  $x_{ij}^k \sim N(0, 1)$  برای  $k = 1, \dots, 6$  تولید کرده و توزیع اثرهای تصادفی به صورت  $\alpha_i \sim N(0, 4)$  فرض شده است. برای  $\beta_\tau$  نیز بردارهای

$$\beta_\tau = (5, 0, 0, 0, 0, 1), \quad \beta_\tau = (3, 1/5, 0, 0, 0, 1),$$

$$\beta_\tau = (0/85, 0/85, 0/85, 0/85, 0/85, 0/85)$$

در نظر گرفته شده است، که در آن بردار اول منظر با یک مدل کاملاً تنک است. برای مولفه خطا نیز توزیع‌های نرمال استاندارد، تی با سه درجه آزادی و توزیع لاپلاس متقارن در نظر گرفته شده است. به منظور کاهش حساسیت مدل به توزیع‌های پیشینی برای پارامتر  $\beta_\tau$  توزیع پیشینی را  $N_6(0, 100I)$  و برای ابرپارامتر  $\sigma$  توزیع پیشینی تخت<sup>۱۱</sup> به صورت  $IG(0/01, 0/01)$  فرض می‌شود. روش

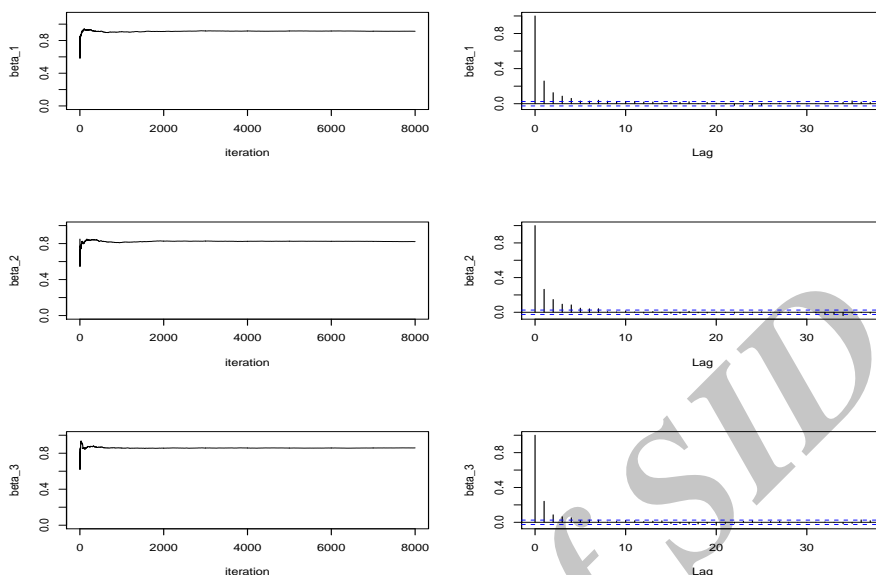
<sup>۱۱</sup> Flat

رگرسیون چندکی دودویی با تاوان لاسو (BL) و روش رگرسیون چندکی دودویی با تاوان لاسوی تطبیق‌پذیر (BAL) برای داده‌های طولی، با دو روش بیزی و بسامدی مورد مقایسه قرار می‌گیرد. روش اول (بنویت و همکاران، ۲۰۱۳) مدل رگرسیون چندکی دودویی با تاوان لاسو از دیدگاه آمار بیزی برای داده‌های مقطعی است که تاوان لاسو روی اثرات ثابت اعمال شده و اثرات تصادفی در مدل لحاظ نمی‌شود، این مدل به اختصار مدل BA نامیده می‌شود. روش دوم (گانکلوز و همکاران، ۲۰۱۲) مدل رگرسیون لوژستیک برای داده‌های طولی با پاسخ‌های دودویی (به اختصار RLG) است که یکجیح محاسبه آن (bild) در نرم‌افزار R موجود است.

برای برآورد پارامترها با روش نمونه‌گیری گیبز، ۸۰۰۰ نمونه از توزیع‌های پسینی شرطی کامل پارامترها تولید کرده و ۲۰۰۰ نمونه اول به عنوان دوره همگرایی مطلوب کنار گذاشته شده است. برای ارزیابی همگرایی زنجیر و تعیین تقریبی تعداد نمونه‌ها به عنوان همگرایی مطلوب از آماره  $\hat{R}$  که توسط گلنن و همکاران (۱۹۹۵) ارائه شده استفاده می‌شود. برای کلیه روش‌ها تقریباً بعد از ۲۰۰۰ نمونه اولیه، همگرایی مطلوب رخ می‌دهد که شکل ۱ نمودارهای اثر و تابع خود همبستگی نمونه‌ای زنجیر را برای روش رگرسیون چندکی با تاوان لاسو و شکل ۲ این نمودارها را برای روش رگرسیون چندکی با تاوان لاسوی تطبیق‌پذیر برای  $(\beta_{\tau} = (0/85, 0/85, 0/85, 0/85, 0/85, 0/85))$  هر دو نمودار به خوبی همگرایی زنجیرهای مارکف تولید شده از توزیع‌های پسینی ضرایب را نشان می‌دهند. با توجه به نمودارهای اثر نیز در هر دو روش می‌توان گفت که تقریباً بعد از ۲۰۰۰ تکرار، همگرایی مطلوب رخ داده است. جدول ۱ برآورد پارامتر  $\beta_{\tau}$  برای چهار روش را در مقایسه با مقدار واقعی  $\beta_{\tau}$  نشان می‌دهد. همان‌طور که مشاهده می‌شود روش‌های ارائه شده عملکرد بهتری در برآورد پارامتر  $\beta_{\tau}$  در مقایسه با مقدار واقعی دارند. برای مقایسه کارایی چهار روش عنوان شده و به منظور لحاظ کردن تغییرپذیری نتایج حاصل از شبیه‌سازی، ۲۰۰ مرتبه روند شبیه‌سازی را تکرار کرده و سه معیار ارزیابی<sup>۱۲</sup> پارامترها، متوسط قدرمطلق خطا<sup>۱۳</sup> و مجذور

<sup>۱۲</sup> Bias

<sup>۱۳</sup> Mean absolute error



شکل ۲: نمودار اثر و تابع خودهمبستگی در روش رگرسیون چندکی با تاوان لاسوی تطبیق پذیر ( $\tau = 0/5$ )

متوسط توان دوم خطا<sup>۱۴</sup> به صورت

$$Bias(\hat{\beta}_k) = \frac{1}{n} \sum_{i=1}^n (\hat{\beta}_{ik} - \beta_k),$$

$$MAE(\hat{\beta}_k) = \frac{1}{n} \sum_{i=1}^n |\hat{\beta}_{ik} - \beta_k|,$$

$$RMSE(\hat{\beta}_k) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\beta}_{ik} - \beta_k)^2}$$

محاسبه می شود. واضح است هر چه اندازه این معیارها کوچک باشد، حاکی از کارایی بالای مدل می باشد. چون نوشتن تمام مقادیر با توجه به وجود چهار مدل، سه روش برآورد برای سه توزیع خطا و همجده پارامتر در قالب جدول امکان پذیر نیست، بنابراین نتایج در قالب نمودارهای جعبه ای خلاصه شده که نتایج آن در شکل های شماره ۳، ۴ و ۵ ارائه شده است. هر چه نمودارهای جعبه ای کوچکتر و

<sup>۱۴</sup> Root mean square error

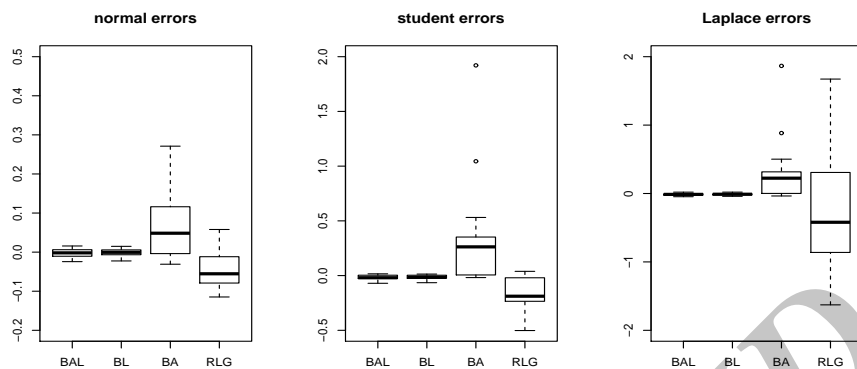
جدول ۱: میانگین پسینی پارامترها با توزیع خطاهای نرمال و  $\tau = 0/5$

$\beta_6$	$\beta_5$	$\beta_4$	$\beta_3$	$\beta_2$	$\beta_1$	مدل
۱	۰	۰	۰	۰	۵	$\beta_{True}$
۰/۸۶۳	۰/۰۶۸	۰/۰۱۴	۰/۰۴۸	-۰/۰۰۲	۵/۰۲۵	BAL
۰/۸۶۱	۰/۰۸۶	-۰/۰۰۲	۰/۰۵۵	-۰/۰۰۷	۵/۰۱۲	BL
۰/۶۸۲	-۰/۳۸۶	۰/۳۳۰	۰/۱۹۸	-۰/۲۲۱	۴/۵۶۴	BA
۱/۰۳۱	۰/۰۰۱	-۰/۲۵۴	۰/۰۳۳	۰/۰۹۱۳	۴/۸۳۰	RGL
۱	۰	۰	۰	۱/۵	۳	$\beta_{True}$
۱/۰۵۷	-۰/۰۴۲	۰/۰۴۶	-۰/۲۰۲	۱/۵۲۳	۳/۰۸۵	BAL
۱/۲۲	-۰/۱۴۶	۰/۰۸۶	-۰/۰۷۴	۱/۵۷	۳/۱۷	BL
۰/۹۹۶	-۰/۱۸۳	-۰/۳۵۰	-۰/۰۵۸	۱/۲۲	۲/۱۱۵	BA
۱/۰۹۱	۰/۰۰۱	-۰/۲۵۴	۰/۰۳۳	۰/۰۹۱۳	۳/۰۷۱	RGL
۰/۸۵	۰/۸۵	۰/۸۵	۰/۸۵	۰/۸۵	۰/۸۵	$\beta_{True}$
۰/۸۲۴	۰/۸۰۴	۰/۹۴۴	۰/۸۷۳	۰/۸۸۱	۰/۸۹۰	BAL
۰/۸۱۷	۰/۸۰۳	۰/۹۴۳	۰/۸۶۷	۰/۸۷۴	۰/۸۸۹	BL
۰/۶۷۲	۰/۵۱۳	۰/۸۲۳	۰/۶۶۸	۰/۸۷۳	۰/۹۲۷	BA
۰/۴۵۰	۱/۱۲۰	۰/۹۷۶	۰/۶۳۹	۱/۰۰۱	۰/۷۶۵	RGL

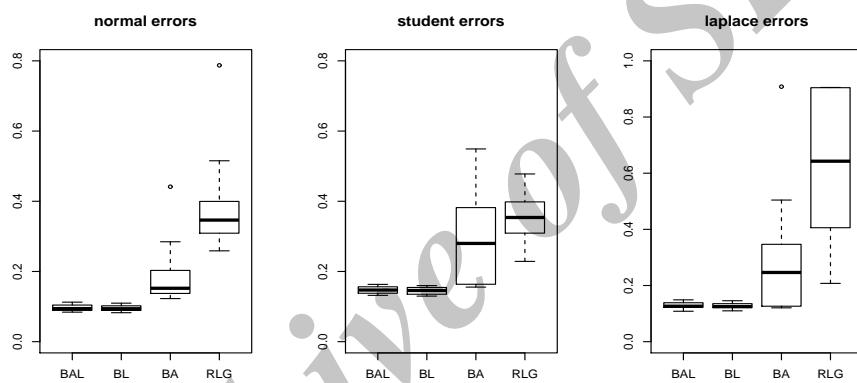
نزدیک صفر باشد، نشان دهنده آریبی کمتر و برآورد بهتر پارامترها است. شکل ۳ عملکرد روش‌ها را از نقطه نظر آریبی مورد مقایسه می‌دهد. با توجه به این شکل دو روش BAL و BL در تمامی توزیع‌های فرض شده برای خطاها عملکرد بهتری نسبت به سایر روش‌ها دارد. اما این دو روش تقریباً عملکرد یکسانی در برآورد پارامترها دارند. شکل‌های ۴ و ۵ نیز عملکرد روش‌ها را به لحاظ معیارهای RMSE و MAE مورد بررسی قرار می‌دهد. با توجه به این دو شکل نیز مشاهده می‌شود که همانند نمودارهای آریبی، در این نمودارها نیز روش‌های ارائه شده برآورد بهتری از پارامترها در تمامی توزیع‌های فرض شده برای خطا دارند.

## ۶ تحلیل داده‌های واقعی

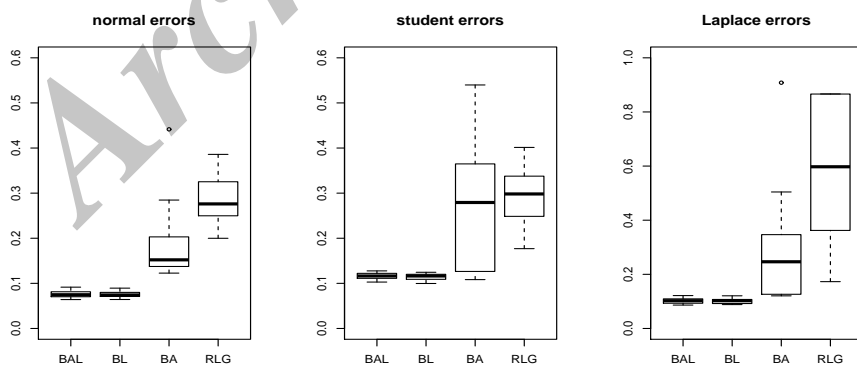
داده‌های واقعی شامل رشد ۵۰ کودک در شهرستان زنجان، از بدو تولد تا دوسالگی است که از درمانگاه‌های شهرستان به تصادف انتخاب شده‌اند. هدف بررسی تأثیر



شکل ۳: عملکرد روش‌ها به لحاظ معیار اریبی در چندک ۵٪



شکل ۴: عملکرد روش‌ها به لحاظ معیار RMSE در چندک ۵٪



شکل ۵: عملکرد روش‌ها به لحاظ معیار MAE در چندک ۵٪

برخی عوامل بر رشد کودک در طول زمان می‌باشد. این عوامل عبارت‌اند از قد، وزن و دور سر کودک که در فواصل زمانی بدو تولد، ۲، ۴، ۶، ۹، ۱۲، ۱۸ و ۲۴ ماهگی کودکان اندازه‌گیری شده است. متغیر پاسخ نیز به صورت صفر و یک در همین فواصل زمانی است که عدد یک نشان دهنده رشد مناسب و عدد صفر نشان‌دهنده رشد نامناسب کودک است. برای برآورد پارامترها با استفاده از روش نمونه‌گیری گیبز، ۸۰۰۰ نمونه از توزیع‌های پسینی شرطی کامل پارامترها تولید کرده و ۲۰۰۰ نمونه اول را به عنوان دوره همگرایی مطلوب کنار می‌گذاریم. در جدول ۲ برآورد ضرایب رگرسیونی برای سه چندک  $\tau = (0/25, 0/5, 0/75)$  برای سه روش بیزی محاسبه شده است. همچنین با مدل برازش شده، متغیرهای پاسخ  $y_{ij}$ ها برای کلیه داده‌ها که تعداد آن‌ها برابر ۴۰۰ مشاهده است، برآورد شده و تعداد پاسخ‌های برآورد شده که با پاسخ‌های واقعی مطابقت دارد در سطرهای آخر جدول ۲ (سطر مربوط به  $\hat{y}_{True}$ ) ارائه شده است. در جدول ۳ نیز برآورد پارامترها و تعداد پاسخ‌های برآورد شده که با پاسخ‌های واقعی مطابقت دارد با روش RLG آمده است. با توجه به این دو جدول ملاحظه می‌شود که هر دو روش ارائه شده در چندک‌های پایین بخصوص برای چندک ۰/۵ از دو مدل دیگر عملکرد بهتری داشته‌اند. ولی برای چندک بالا ( $\tau = 0/75$ ) مدل غیر بیزی RLG عملکرد بهتری نسبت به مدل‌های بیزی دارد.

### بحث و نتیجه‌گیری

در مطالعات طولی علاوه بر اثرات ثابت که تأثیر هر یک از متغیرهای توصیفی را بر متغیر پاسخ بیان می‌کند، اثرات تصادفی نیز در مدل لحاظ می‌شود. این اثرات تغییرات بین متغیرهای پاسخ (تغییرات بین گروهی) را کنترل می‌کند. در مطالعاتی که تعداد مشاهدات از متغیر پاسخ زیاد است، لذا اثرات تصادفی نیز زیاد شده و پارامترهای زیادی وارد مدل شده و بنابراین دقت مدل کم و تفسیر آن مشکل می‌شود. در این مقاله با ایجاد تاوان لاسو و لاسوی تطبیق‌پذیر روی اثرات تصادفی آن‌ها را به سمت صفر منقبض کرده و اثرات کم اهمیت از مدل حذف شدند. هر دو



جدول ۲: برآورد ضرایب رگرسیونی در چندک‌های متفاوت

مدل	۰/۲۵	۰/۵	۰/۷۵
BAL			
$\beta_1$	۰/۰۱۶۷	۰/۰۰۹۶	۰/۰۲۱۴
$\beta_2$	۰/۱۲۸۹	۰/۰۸۲۰	۰/۱۵۳۴
$\beta_3$	۰/۰۰۲۳۱	۰/۰۱۶۴	۰/۰۰۵۷
$\hat{y}_{True}$	۳۰۰	۳۵۰	۲۵۰
BL			
$\beta_1$	۰/۰۱۵۸	۰/۰۰۹۰	۰/۰۰۰۳۵
$\beta_2$	۰/۱۲۲۲	۰/۰۸۲۶	۰/۰۵۰۵
$\beta_3$	۰/۰۰۳۶۱	۰/۰۱۶۷	۰/۰۲۷۷
$\hat{y}_{True}$	۳۰۰	۳۵۰	۲۵۰
BA			
$\beta_1$	۰/۱۵۰۲	۰/۰۱۲۸	۰/۰۱۴۹
$\beta_2$	۰/۲۱۰۰	۰/۰۲۴۱	۰/۹۰۶۲
$\beta_3$	۰/۲۵۴۱۱	۰/۰۳۱۰۷	۰/۰۲۲۶
$\hat{y}_{True}$	۲۴۹	۲۸۵	۲۷۵

جدول ۳: برآورد ضرایب رگرسیونی با استفاده از مدل رگرسیون لوژستیک

مدل	$\beta_1$	$\beta_2$	$\beta_3$	$\hat{y}_{True}$
RGL	۰/۰۲۳۶	۰/۰۴۹۹	۰/۰۴۹۵	۲۸۵

نوع تاوان با تعریف توزیع‌های پیشین لاپلاس متقارن که تابع چگالی آن در نقطه صفر نوک‌دار است، در مدل لحاظ شده و هر دو مدل از دیدگاه آمار بیزی مورد بررسی قرار گرفتند. نتایج حاصل از شبیه‌سازی و تحلیل داده‌های واقعی نشان داد که روش‌های ارائه شده در چندک ۰/۵ نسبت به روش بیزی (بنویت و همکاران، ۲۰۱۳) که اثرات تصادفی و تاوان روی آن‌ها را در مدل لحاظ نمی‌کند و همچنین روش بسامدی که اثرات تصادفی را در مدل لحاظ می‌کند، ولی هیچ‌گونه تاوانی روی آن‌ها قرار نمی‌دهد، در برآورد ضرایب رگرسیونی عملکرد بهتری دارد. نتایج حاصل از تحلیل داده‌های واقعی نشان داد که روش‌های ارائه شده در چندک‌های ۰/۲۵ و ۰/۵ دقت بالایی نسبت به دو روش دیگر دارند، اما در چندک بالا ( $\tau = ۰/۷۵$ ) رهیافت بسامدی تا حدودی از روش‌های بیزی عملکرد بهتری

دارد. اما از آنجایی که مدل‌های ارائه شده بر اساس روش رگرسیون چندکی است، لذا در مقایسه با مدل بسامدی که بر اساس روش رگرسیون میانگین است، نسبت به داده‌های پرت و توزیع خطاها استوار است. بنابراین استفاده از مدل‌های ارائه شده در تحلیل داده‌های واقعی توصیه می‌شود.

### تقدیر و تشکر

نویسندگان مقاله از داوران و سردبیر محترم مجله به خاطر ارایه پیشنهادات ارزنده که سبب بهبود مقاله گردید، کمال تشکر و قدردانی را دارند.

### مراجع

غلامی فشارکی، م.، کاظم نژاد، ا. و زایری، ف. (۱۳۹۲)، تحلیل دو سطحی با اثرات تصادفی چوله نرمال و مدل‌بندی داده‌های طولی، مجله علوم آماری، ۷، ۲۳۳-۲۴۸.

Alhamzawi, R., Yu, K. and Benoit, D. F. (2012), Bayesian Adaptive Lasso Quantile Regression, *Statistical Modeling*, **12**, 279-297.

Benoit, D. F., Alhamzawi, R. and Keming, Y. U. (2013), Bayesian Lasso Binary Quantile Regression, *Computational Statistics*, **28**, 2861-2873.

Fitzmaurice, G. M. and Laird, N. M. (1993), A Likelihood-Based Method for Analyzing Longitudinal Binary Responses, *Biometrika*, **80**, 141-151.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995), *Bayesian Data Analysis*, Chapman and Hall, London.

Goncalves, M. H. (2002), *Likelihood Methods for Discrete Longitudinal Data*, Ph.D. Thesis, University of Lisbon.

- Goncalves, M. H. and Azzalini, A. (2008), Using Markov Chains for Marginal Modeling of Binary Longitudinal Data in Exact Likelihood Approach, *Metron*, **LXVI**, 157-181.
- Goncalves, M. H., Cabral, M. S. and Azzalini, A. (2012), The R Package Bild for the Analysis of Binary Longitudinal Data, *Journal of Statistical Software*, **46**, 1-17.
- Koenker, R. (2004), Quantile Regression for Longitudinal Data, *Journal of Multivariate Analysis*, **91**, 74-89.
- Koenker, R. and Machado, J. (1999), Goodness of Fit and Related Inference Processes for Quantile Regression, *Journal of the American Statistical Association*, **94**, 1296-1310.
- Kozumi, H. and Kabayashi, G. (2011), Gibbs Sampling Methods for Bayesian Quantile Regression, *Journal of Statistical Computation and Simulation*, **81**, 1565-1578.
- Luethi, D. and Breyman, W. (2014), A Package on the Generalized Hyperbolic Distribution and Its Special Cases, R Package Version 1.5.6, URL <http://www.r-project.org>
- Park, T. and Casella, T. (2008), The Bayesian Lasso, *Journal of the American Statistical Association*, **103**, 681-686.
- Tibshirani, R. (1996), Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society, B*, **58**, 267-288.
- Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, **101**, 1418-1429.