رگرسیون چندکی بیزی با تاوان الاستیک‌نت سازوار برای داده‌های طولی

علی آقامحمدی*، سکینه محمدی؛ دانشگاه زنجان، گروه آمار

دریافت ۹۳/۱۰/۱۳ پذیرش ۹٤/۱۰/٦

چکیده

بررسی داده‌های طولی قسمت مهمی از پژوهش‌*های* اپیدمیولوژی، بررسی‌های بالینی و تحقیقات اجتماعی را شامل می‌شود. در بررسی‌های طولی، اندازه‌گیری پاسخ‌ها به‌طور مکرر در طول زمان انجام می‌شود. اغلب هدف اصلی تشخیص تغییر در متغیر پاسخ در طول زمان و عواملی است که روی این تغییر اثر می‌گذارند. اخیراً به رگرسیون چندکی برای تجزیه و تحلیل این نوع داده‌ها توجه شده است. در این مقاله مدل رگرسیون چندکی با ایجاد تاوان الاستیک‌نت سازوار روی اثرهای تصادفی برای داده‌های طولی ارائه شده و از دیدگاه آمار بیزی تجزیه و تحلیل می‌شود. چون در این روش توزیع پسینی پارامترها به شکل بسته قابل حصول نیستند، از این‌رو، توزیع‌های پسینی شرطی کامل پارامترها محاسبه شده و از الگوریتم نمونه‌گیری گیبس برای استنباط استفاده می‌شود. برای مقایسهٔ کارایی روش ارائه شده با روش‌های متداول، بررسی شبیه‌سازی انجام شده و در پایان نحوهٔ کاربست مدل‌ها در قالب مثال کاربردی شرح داده می‌شود.

واژه‌های کلیدی: رگرسیون چندکی، داده‌های طولی، تاوان الاستیک‌نت سازوار، اثرهای تصادفی، استنباط بیزی.

مقدمه

در بسیاری از تحقیقات مربوط به علوم‌پزشکی و علوم‌اجتماعی و اقتصادی، برای بررسی تغییرات ویژگی‌های واحدهای آزمایشی در طول زمان و تحلیل آن‌ها از بررسی‌های طولی استفاده می‌شود. در این نوع پژوهش‌ها، متغیر پاسخ در چندین نوبت متوالی مشاهده می‌شود، به‌عبارت دیگر یک واحد آزمایشی تحت اندازه‌گیری مکرر در طول زمان قرار می‌گیرد. زیرا اندازهٔ صفت بررسی شده، در طول زمان در واحدهای آزمایشی تغییر می‌کند. در واقع به پژوهشی که اندازه‌گیری مربوط به صفتی در طول زمان بررسی می‌شود، مطالعهٔ طولی[1] و به داده‌های جمع‌آوری شده در این نوع پژوهش‌ها، داده‌های طولی[2] گویند. هدف اولیّه این تحقیقات تشخیص عواملی است که متغیّرهای پاسخ را در طول زمان تحت تأثیر قرار می‌دهد. واضح است این نوع پژوهش‌ها نسبت به بررسی‌های مقطعی که متغیر پاسخ برای هر واحد آزمایشی تنها یک بار اندازه‌گیری می‌شود، نتایج کارآمدتری ارائه می‌دهد. یکی از روش‌های تجزیه و تحلیل این نوع داده‌ها رگرسیون چندکی[3] است. این مدل رگرسیونی را اولین بار کوئنکر و باست [۱] ارائه کرد. این روش نسبت به‌روش‌های دیگر دارای دو مزیت مهم است. نخست این‌که رگرسیون چندکی نسبت به ناپایداری واریانس و داده‌های دورافتاده حسّاس نیست و دوم این‌که اطلاعات جزئی‌تری نسبت به تأثیر متغیرهای تبیینی در چندک‌های مختلف توزیع متغیر پاسخ ارائه می‌دهد.

***نویسندهٔ مسئول** Aghamohammadi.ali@znu.ac.ir

1. Longitudinal study
2. Longitudinal data
3. Quantile regression

در بررسی داده‌های طولی علاوه بر اثرهای ثابت که تغییرات درون گروه‌ها (تغییرات در طول زمان) را کنترل می‌کند، اثرات تصادفی[٤] بین متغیرهای پاسخ نیز در مدل لحاظ می‌شود که تغییرات بین گروهی را کنترل می‌کنند. اما در بسیاری از مسائل، تعداد مشاهدات از متغیر پاسخ نسبت به مشاهدات انجام گرفته در طول زمان از متغیر مربوط بسیار زیاد است. بنابراین در این مدل تعداد اثرات تصادفی (تعداد پارامترها) نیز خیلی زیاد شده از این‌رو، تعبیر و تفسیر مدل مشکل شده و دقت آن نیز کاهش می‌یابد. برای رفع این مشکل، کوئنکر [۲] رگرسیون چندکی را برای تحلیل این داده‌ها با افزودن تاوان لاسو[٥] (تاوان با نرم L_1) روی اثرات تصادفی بررسی کرد. در واقع این روش با ایجاد تاوان، اثرات تصادفی کم‌اهمیت را به‌سمت صفر ترنجش[٦] کرده و مدلی تنک[٧] ایجاد می‌کند.

اولین بار تیبشیرانی [۳] رگرسیون لاسو که با ایجاد تاوان لاسو روی پارامترها انجام می‌شود را ارائه کرد. سپس زو [٤] رگرسیون لاسوی سازوار[٨] را به‌منظور گسترش رگرسیون لاسو پیشنهاد کرد. در این روش بر خلاف تاوان لاسو که برای همهٔ ضرایب اندازه تاوان را یکسان در نظر می‌گیرد، اندازهٔ تاوان‌های مختلفی برای ضرایب رگرسیونی متفاوت لحاظ می‌شود. بررسی‌های زو [٤] مشخص کرد که در رگرسیون چندکی با تاوان لاسوی سازوار اریبی در برآورد پارامترها در مقایسه با روش لاسو کم‌تر است. اما روش‌های رگرسیونی لاسو عمل‌کرد بهتری ندارند مگر این‌که اندازه کران تاوان نرم L_1 از مقداری مشخص کوچک‌تر باشد. درغیر این صورت عمل‌کرد این روش مانند روش کم‌ترین توان‌های دوم عادی[٩] (OLS) خواهد بود. در واقع دقت روش‌های رگرسیونی لاسو به انتخاب کران بسیار حساس است [٥]. از طرف دیگر در حالت $n > k$ (k تعداد پارامترهای موجود در مدل و n حجم نمونه است)، اگر هم‌بستگی زیادی بین متغیرهای تبیینی در مدل وجود داشته باشد، آن‌گاه به‌صورت تجربی مشاهده می‌شود که روش رگرسیونی ستیغی[١٠] به لحاظ پیش‌بینی نسبت به‌روش‌های لاسو عمل‌کرد بهتری دارد [۳]. از این‌رو، زو و هستی [٥]، روش رگرسیونی الاستیک‌نت[١١] که تابع تاوان آن ترکیبی اکیداً محدب از توابع تاوان لاسو و ستیغی است، را معرفی کردند. چنان‌که دیدیم رگرسیون ستیغی را اولین بار هورل و کنارد [٦]، به‌منظور غلبه بر هم‌خطی یعنی زمانی که هم‌بستگی زیادی بین متغیرهای تبیینی در مدل‌های خطی موجود است، ارائه کرده است. با توجه به این ویژگی رگرسیون ستیغی، مدل رگرسیونی با تاوان الاستیک‌نت علاوه بر ترنجش ضرایب کم‌اهمیت مدل به‌سمت صفر و در نهایت حذف برخی از آن‌ها از مدل، می‌تواند متغیرهای تبیینی که اثر یکسان بر متغیر پاسخ دارند یا دارای هم‌بستگی زیاد هستند را به‌طور یکسان و برابر برآورد کند [٥]. از این‌روش در مدل‌های رگرسیونی که متغیرهای تبیینی گروه‌بندی شده دارند، به‌خصوص اگر تعداد آن‌ها زیاد باشد می‌توان استفاده کرد. اولین بار لی و لین [۷] مدل رگرسیونی با تاوان الاستیک‌نت را در داده‌های مقطعی از دیدگاه آمار بیزی بررسی کردند. در این روش پارامترهای تنظیم[١٢]

4. Random effects
5. Lasso penalty
6. Shrink
7. Sparse
8. Adaptive Lasso
9. Ordinary least square
10. Ridge
11. Elastic net
12. Tuning parameters

مربوط به هر دو تاوان لاسو و ستیغی از روش اعتبارسنجی متقابل[۱۳] محاسبه می‌شوند. سپس چن و همکاران [۸] با تعمیم روش لی ولین [۷] مدل رگرسیونی با تاوان الاستیک‌نت سازوار را در داده‌های مقطعی از دیدگاه آمار بیزی ارائه کردند. این روش در مقایسه با روش لی ولین [۷] دارای دو مزیت است. اول این‌که به‌جای تاوان الاستیک‌نت از تاوان الاستیک‌نت سازوار استفاده شده است و دوم این‌که در این روش برای پارامترهای تنظیم در تاوان لاسو توزیع‌های پیشینی مزدوج موجود بوده از این‌رو، برای برآورد آن‌ها برخلاف روش لی و لین [۷] نیاز به اعتبارسنجی متقابل نیست.

در این مقاله هدف تحلیل داده‌های طولی با استفاده از روش رگرسیون چندکی با در نظر گرفتن تاوان الاستیک‌نت سازوار روی اثرهای تصادفی از دیدگاه آمار بیزی است. برخلاف روش‌های لی و لین [۷] و چن و همکاران [۸] در مدل ارائه شده برای پارامترهای تنظیم در هر دو تاوان لاسو و ستیغی توزیع‌های پیشینی مزدوج موجود است از این‌رو، تمام پارامترهای مدل در فرآیند نمونه‌گیری گیبس قابل برآورد هستند. بنابراین بخش دوم، شامل مطالب کلی در مورد مدل رگرسیون چندکی و رگرسیون چندکی در داده‌های طولی است. بخش سوم مدل رگرسیون چندکی با تاوان الاستیک‌نت سازوار از دیدگاه آمار بیزی ارائه و در بخش چهارم نیز با روش شبیه‌سازی، کارایی مدل ارائه شده با مدل‌های دیگر مقایسه می‌شود. بخش ششم شامل تحلیل داده‌های واقعی و ارزیابی کارایی مدل‌ها است.

رگرسیون چندکی

مدل رگرسیونی خطی $y_i = x_i'\beta + \varepsilon_i, i = ۱,\ldots,n$ را در نظر بگیرید که در آن x_i بردار مربوط به متغیرهای تبیینی، β یک بردار $k \times ۱$ بعدی از پارامترها و ε_i مؤلفهٔ خطا است. در مدل رگرسیونی میانگین، هدف استنباط دربارهٔ متوسط مقدار پاسخ به‌ازای مقادیر متفاوت متغیرهای تبیینی است، از این‌رو، به $E(y_i \mid x_i) = x_i'\beta$ توجه شده است. اما در رگرسیون چندکی، به این تابع چندک شرطی توجه شده است:

$$Q_\tau \ y_i|x_i \ = x_i'\beta, \qquad i = ۱,\ldots,n,$$

که در آن $Q_\tau(.)$ معکوس تابع توزیع تجمعی متغیر پاسخ y_i به شرط معلوم بودن بردار x_i است. به‌عبارت دیگر $x_i'\beta$ چندک τ-ام شرطی متغیر y_i به‌شرط معلوم بودن بردار x_i را نشان می‌دهد. در رگرسیون معمولی میانگین توزیع خطاها صفر در نظر گرفته می‌شود، امّا در رگرسیون چندکی، چندک τـام توزیع ε_iها برابر صفر هستند، یعنی $\int_{-\infty}^{\cdot} f_{\varepsilon_i} \ t \ dt = \tau$. در رگرسیون چندکی، ضرایب رگرسیونی یعنی β ازمینیمم کردن این عبارت نسبت به β برآورد می‌شود:

$$\sum_{i=۱}^{n} \rho_\tau \ y_i - x_i'\beta \ , \qquad (۱)$$

که در آن $\rho_\tau(.)$ نشان‌دهنده تابع زیان و به‌صورت $\rho_\tau \ u \ = \frac{|u| + \ ۲\tau - ۱ \ u}{۲}$ تعریف می‌شود [۱].

13. *Cross validation*

چون رابطهٔ (۱) نسبت به β در مبدأ مشتق‌پذیر نیست، از این‌رو،راه حل تحلیلی برای برآورد ضرایب رگرسیونی وجود ندارد. برای رفع این مشکل یو و مویید [۹] رگرسیون چندکی را از دیدگاه آمار بیزی با استفاده از توزیع لاپلاس نامتقارن پیشنهاد کردند. متغیر تصادفی y را دارای توزیع لاپلاس نامتقارن گویند و با نماد $y \sim ALD(\mu,\sigma,\tau)$ نشان می‌دهند، هر گاه تابع چگالی آن بدین‌صورت باشد:

$$f\ y\ = \frac{\tau\ \ ۱-\tau}{\sigma}\exp\left\{-\rho_{\tau}\left(\frac{y-\mu}{\sigma}\right)\right\}, \qquad -\infty < y < +\infty$$

که در آن $۰ < \tau < ۱$ پارامتر چولگی، σ پارامتر مقیاس و $-\infty < \mu < \infty$ پارامتر مکان است. از این‌رو، با فرض $y_i \sim ALD(x_i'\beta,\sigma,\tau)$، تابع درست‌نمایی بردار $y = (y_۱,\ldots,y_n)$ بدین‌صورت به‌دست می‌آید:

$$L\ \ \beta,\sigma \mid y,x,\tau\ \ = \left(\frac{\tau\ \ ۱-\tau}{\sigma}\right)^n \exp\left\{-\sum_{i=۱}^{n}\rho_{\tau}\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right\}, \qquad (۲)$$

بنابراین ماکسیمم کردن رابطهٔ (۲) در حضور پارامتر مزاحم σ با مینیمم کردن رابطهٔ (۱) نسبت به β معادل است. به‌همین دلیل در بسیاری از موارد، تحلیل رگرسیون چندکی با استفاده از این توزیع بررسی می‌شود. یو و مویید [۹] با فرض $\sigma = ۱$ و تعریف توزیع پیشینی ناآگاهی‌بخش برای β $(\pi(\beta) \propto ۱)$ رگرسیون چندکی بیزی را در داده‌های مقطعی معرفی و بررسی کردند.

مدل خطی برای داده‌های طولی یک مدل آمیخته خطی بدین‌صورت است:

$$y_{ij} = x_{ij}'\beta + \alpha_i + \varepsilon_{ij}, \qquad i = ۱,\ldots,n, \qquad j = ۱,\ldots,n_i, \qquad (۳)$$

است که در آن y_{ij}، j-امین پاسخ اندازه‌گیری شده روی i-امین واحد آزمایشی و x_{ij} بردار $k \times ۱$ بعدی مربوط به متغیرهای تبیینی، β یک بردار $k \times ۱$ بعدی از پارامترها و α_i اثرتصادفی مربوط به پاسخ y_{ij} را نشان می‌دهد. توجه شود که در این مدل α_iها اثرات بین گروهی و β اثرات درون گروهی را کنترل می‌کنند. تابع چندک شرطی در این مدل بدین‌صورت بیان می‌شود:

$$Q_{\tau}\ \ y_{ij}|x_{ij},\alpha_i\ \ = x_{ij}'\beta + \alpha_i,$$

کوئنکر [۲] با فرض تاوان لاسو روی بردار $\alpha = (\alpha_۱,\ldots,\alpha_n)$ $(p(\alpha) = \lambda\sum_{i=۱}^{n}|\alpha_i|)$ پارامترهای مدل یعنی α و β را از مینیمم کردن تابع زیان

$$\sum_{i=۱}^{n}\sum_{j=۱}^{n_i} w_{\tau}\rho_{\tau}\ \ y_{ij} - x_{ij}'\beta - \alpha_i\ \ + \lambda\sum_{i=۱}^{n}|\alpha_i|,$$

نسبت به α و β برآورد کردند که در آن وزن w_{τ} تأثیر نسبی چندک τ-ام را در برآورد اثرات تصادفی کنترل می‌کند. کوئنکر [۲] استدلال کرد که اگر i در مقایسه با j خیلی بزرگ باشد، آنگاه تابع زیان تاوانیده فوق نسبت به‌روش‌های غیرتاوانیده عمل‌کرد بهتری دارند. گراسی و باتای [۱۰] نیز با تعریف توزیع خاص برای α_i و با در نظر گرفتن توزیع لاپلاس نامتقارن برای y_{ij}ها مدل رگرسیونی چندکی در داده‌های طولی را

از دیدگاه آمار بیزی بررسی شدند. در این روش، با فرض این‌که توزیع α_iها لاپلاس متقارن هستند، مدل کوئنکر [۲] به‌دست می‌آید. با توجه به مشکلات تاوان لاسو که در بخش اول به آن‌ها اشاره شد، هدف این مقاله بررسی مدل رگرسیون چندکی درداده‌های طولی با تاوان الاستیک‌نت سازوار روی اثرات تصادفی از دیدگاه آمار بیزی است. با در نظر گرفتن این تاوان روی اثرات تصادفی، تابع زیان تاوانیده به‌صورت

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_\tau\left(y_{ij}-x_{ij}'\beta-\alpha_i\right)+\sum_{i=1}^{n}\nu_i\left|\alpha_i\right|+\lambda\sum_{i=1}^{n}\alpha_i^2 \qquad (۴)$$

به‌دست می‌آید که در آن مؤلفه $\sum_{i=1}^{n}\nu_i\left|\alpha_i\right|+\lambda\sum_{i=1}^{n}\alpha_i^2$ را تاوان الاستیک‌نت سازوار گویند. در این تاوان به عبارت $\sum_{i=1}^{n}\nu_i\left|\alpha_i\right|$ تاوان لاسوی سازوار و به ν_i پارامتر تنظیم گویند. هرچه مقدار این پارامتر بزرگ‌تر باشد، ترنجش به‌سمت صفر نیز بیش‌تر و اثرهای تصادفی کم‌اهمیت از مدل حذف خواهند شد. عبارت $\lambda\sum_{i=1}^{n}\alpha_i^2$ همان تاوان ستیغی و λ نیز پارامتر تنظیم آن است. برآورد بردار پارامترهای α و β از دیدگاه آمار بسامدی از مینیمم کردن عبارت رابطهٔ (۴) نسبت به α و β به‌دست می‌آید.

رگرسیون چندکی با تاوان الاستیک‌نت سازوار از دیدگاه آمار بیزی

چنان‌که عنوان شد در مدل رگرسیون چندکی با تاوان الاستیک‌نت سازوار در داده‌های طولی، تابع زیان تاوانیده با تاوان الاستیک‌نت سازوار روی اثرهای تصادفی، به‌صورت رابطهٔ (٤) است. اگر فرض کنیم در رابطهٔ (۳)، $\varepsilon_{ij}\overset{iid}{\sim}ALD(۰,\sigma,\tau)$ و توزیع پیشینی برای α_i را مانند چن و همکاران[۸] بدین‌صورت در نظر بگیریم:

$$\pi(\alpha_i\mid\lambda,\sigma,\nu_i)\propto\frac{\nu_i}{۲\sigma}\exp\left\{-\frac{\nu_i\mid\alpha_i\mid}{\sigma}-\lambda\alpha_i^2\right\}$$

آن‌گاه توزیع پسینی بردار α بدین‌صورت به‌دست می‌آید:

$$\pi(\alpha\mid y,x,\sigma,\lambda,\nu,\beta)\propto$$
$$\sigma^N\exp\left\{-\sigma^{-۱}\left(\sum_{i=1}^{n}\sum_{j=1}^{n_i}\rho_\tau(y_{ij}-x_{ij}'\beta-\alpha_i)-\sum_{i=1}^{n}\nu_i\left|\alpha_i\right|\right)-\lambda\sum_{i=1}^{n}\alpha_i^2\right\}, \qquad (۵)$$

که در آن $N=\sum_{i=1}^{n}n_i$، تعداد کل مشاهدات را نشان می‌دهد. از این‌رو، مینیمم کردن تابع هدف (٤) معادل ماکسیمم کردن توزیع پسینی در رابطهٔ (۵) در حضور پارامتر مزاحم σ است. بنابراین می‌توان توزیع لاپلاس نامتقارن را برای این مدل استفاده کرد. با توجه به رابطهٔ (۵) توزیع پسینی بردار α توزیع شناخته شده‌ای نیست و نمی‌توان برای استنباط از آن استفاده کرد. از این رو، برای سادگی مدل و استفاده از روش نمونه‌گیری گیبس، ابتدا توزیع لاپلاس نامتقارن و عبارت $(۲\sigma)^{-۱}\nu_i\exp\ -(\sigma)^{-۱}\nu_i\mid\alpha_i\mid$ را به‌صورت توزیع‌های آمیخته در

نظر می‌گیریم. اگر $u \sim ALD(\cdot,\sigma,\tau)$ و متغیرهای z و e به‌ترتیب دارای توزیع نرمال استاندارد و نمایی با میانگین $(\tau(1-\tau))^{-1}\sigma$ و مستقل از هم باشند، آن‌گاه:

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

که در آن $k_1 = (1-2\tau)$ است [۱۱]. با استفاده از این خاصیت توزیع لاپلاس نامتقارن و با توجه به این‌که $\varepsilon_{ij} \sim ALD(\cdot,\sigma,\tau)$ است، آن‌گاه مدل رابطهٔ (۳) را می‌توان بدین‌صورت بازنویسی کرد:

$$y_{ij} = x'_{ij}\beta + \alpha_i + k_1 e_{ij} + \sqrt{2\sigma e_{ij}}\, z_{ij}, \qquad (۴)$$

که در آن z_{ij} و e_{ij} به‌ترتیب دارای توزیع نرمال استاندارد و نمایی با میانگین $\sigma(\tau(1-\tau))^{-1}$ و مستقل از هم هستند. تابع چگالی $(2\sigma)^{-1}\nu_i \exp\ -(\sigma)^{-1}\nu_i \mid \alpha_i \mid$ که همان توزیع لاپلاس متقارن است را می‌توان به‌صورت توزیعی آمیخته از هسته‌های فیجر- بارتلت با توزیع آمیختگی گاما بدین‌صورت نوشت [۱۲]:

$$\frac{\nu_i}{2\sigma}\exp\left\{-\frac{\nu_i \mid \alpha_i \mid}{\sigma}\right\} = \int_0^\infty \frac{\nu_i}{\sigma}\left\{1-\frac{\nu_i \mid \alpha_i \mid}{w_i\sigma}\right\}_+ f(w_i)dw_i$$

که در آن $f(w_i) = w_i \exp\ -w_i$ و $a_+ = \max\{a,0\}$. چون هدف تحلیل بیزی است، از این‌رو، برای پارامترهای تنظیم ν_i و λ توزیع‌های پیشینی مزدوج را نمایی به‌صورت $\pi(\lambda \mid \phi_2) = \phi_2 \exp\{-\phi_1\lambda\}$ و $\pi(\nu_i \mid \phi_1) = \phi_1 \exp\{-\phi_1\nu_i\}$ قرار داده و برای پارامترهای σ و β به‌ترتیب توزیع معکوس گاما و توزیع نرمال چندمتغیره در نظر می‌گیریم. واضح است که هر چه مقدار ϕ_1 کوچک باشد، آن‌گاه ترنجش به‌سمت صفر α_iها بیش‌تر و تاوانیدن آن‌ها نیز بزرگ است. چون تعداد α_iها زیاد است، از این‌رو، برای پارامتر ϕ_1 توزیع پیشینی را به‌صورت پخ[14] یعنی $\pi(\phi_1) \propto \phi_1^{-1}$ در نظر می‌گیریم. برای ϕ_2 نیز توزیع پیشینی را به‌صورت $\pi(\phi_2) \propto \phi_2^{-1}$ قرار می‌دهیم. حال با توجه به مدل روابط (۵)، (٦) و توزیع‌های پیشینی تعریف شده، مدل سلسله مراتبی را می‌توان بدین‌صورت خلاصه کرد:

$$y_{ij} \mid x_{ij},\beta,\alpha_i,\sigma,e_{ij} \sim N(x'_{ij}\beta + \alpha_i + k_1 e_{ij}, 2\sigma e_{ij})$$

$$e_{ij} \mid \sigma \sim \mathrm{Exp}\left(\frac{\tau(1-\tau)}{\sigma}\right), \quad \beta \mid b_\cdot, B_\cdot \sim N_k(b_\cdot, B_\cdot),$$

$$\pi(\alpha_i \mid \nu_i,\lambda,\sigma,w_i) \propto \exp\{-\lambda\alpha_i^2\} \times \frac{\nu_i}{\sigma}\left\{1-\frac{\nu_i \mid \alpha_i \mid}{w_i\sigma}\right\}_+, \ w_i \sim Gamma(2,1)$$

$$\nu_i \mid \phi_1 \sim \mathrm{Exp}(\phi_1), \quad \lambda \mid \phi_2 \sim \mathrm{Exp}(\phi_2), \qquad \pi(\phi_1) \propto \frac{1}{\phi_1},$$

$$\pi(\phi_2) \propto \frac{1}{\phi_2}, \qquad \sigma \mid c_\cdot, d_\cdot \sim IGamma(c_\cdot, d_\cdot),$$

که در آن $Gamma(a,b)$ و $IGamma(a,b)$ به‌ترتیب نشان‌دهندهٔ توزیع گاما و معکوس گاما و $\mathrm{Exp}(\theta)$ نشان‌دهندهٔ توزیع نمایی با میانگین θ^{-1} است. با توجه به شکل سلسله مراتبی مذکور، توزیع پسینی همهٔ

14. Flat

پارامترها و ابرپارامترها به‌صورت بسته قابل حصول نیست. اما می‌توان توزیع‌های پسینی شرطی کامل همهٔ پارامترها و ابرپارامترها را بدین‌صورت محاسبه کرد:

هستهٔ توزیع پسینی α_i بدین‌صورت به‌دست می‌آید:

$$\pi(\alpha_i \mid .) \propto \exp\left\{-\frac{۱}{۴\sigma}\sum_{j=۱}^{n_i}\frac{(y_{ij}-x'_{ij}\beta-\alpha_i-k_۱ e_{ij})^۲}{e_{ij}}\right\}\times\exp\{-\lambda\alpha_i^۲\}\times\left\{۱-\frac{\nu_i \mid \alpha_i \mid}{w_i\sigma}\right\}_+ .$$

ملاحظه می‌شود که توزیع مذکور، توزیع شناخته شده‌ای نیست. برای تولید نمونه از این توزیع به پیروی از پولسون و همکاران [۱۲]، ابتدا با تعریف متغیرهای تصادفی برشی[15] $u_۱,\ldots,u_n$ که دارای توزیع یک‌نواخت در بازهٔ $(۰,۱)$ هستند، توزیع پسینی α_i را به‌صورت توزیع توأم

$$\pi(\alpha_i,u_i \mid .) \propto \exp\left\{-\frac{۱}{۴\sigma}\sum_{j=۱}^{n_i}\frac{(y_{ij}-x'_{ij}\beta-\alpha_i-k_۱e_{ij})^۲}{e_{ij}}\right\}$$
$$\times\exp\{-\lambda\alpha_i^۲\}\times I(۰<u_i<۱-\frac{\nu_i\left|\alpha_i\right|}{w_i\sigma}),$$

در نظر می‌گیریم که در آن $I(.)$ همان تابع مشخصه است. حال با توجه به رابطهٔ مذکور داریم:

$$(u_i \mid \alpha_i,.) \sim Unif(۰,۱-\frac{\nu_i \mid \alpha_i \mid}{w_i\sigma}), i=۱,\ldots,n,$$

که در آن $Unif(a,b)$ توزیع یک‌نواخت در بازهٔ (a,b) است. همچنین

$$(\alpha_i \mid u_i,.) \sim N(\bar{\mu}_i,\tilde{\sigma}_i)\times I(\mid \alpha_i \mid \leq \frac{(۱-u_i)\sigma w_i}{\nu_i}), i=۱,\ldots,n$$

که در آن

$$\bar{\mu}_i=\tilde{\sigma}_i\left[\frac{۱}{۲\sigma}\sum_{j=۱}^{n_i}\frac{(y_{ij}-x'_{ij}\beta-k_۱e_{ij})}{e_{ij}}\right] \text{ و } \tilde{\sigma}_i=(۲\lambda+\frac{۱}{۲\sigma}\sum_{j=۱}^{n_i}\frac{۱}{e_{ij}})$$

است. بنابراین توزیع پسینی شرطی کامل α_i نرمال بریده شده در نقاط $-\frac{(۱-u_i)\sigma w_i}{\nu_i}$ و $\frac{(۱-u_i)\sigma w_i}{\nu_i}$ با میانگین $\bar{\mu}_i$ و واریانس $\tilde{\sigma}_i$ است.

$$\pi(w_i \mid .) \propto w_i\exp\{-w_i\}\times I(w_i\geq\frac{\nu_i\left|\alpha_i\right|}{(۱-u_i)\sigma}), i=۱,\ldots,n.$$

اگر $a_i=\nu_i\left|\alpha_i\right|\left[(۱-u_i)\sigma\right]^{-۱}$، آنگاه توزیع پسینی w_i توزیع گامای بریده شده در نقطهٔ a_i است. برای سهولت در نمونه‌گیری از این توزیع، فرض می‌کنیم $q_i=w_i-a_i$، آنگاه توزیع q_i بدین‌صورت به‌دست می‌آید:

$$\pi(q_i \mid .)=c(q_i+a_i)\exp\ -(q_i+a_i)\ \times I(q_i\geq۰)$$
$$=cq_i\exp\{-q_i\}\exp\{-a_i\}+\mathrm{c}\,a_i\exp\ -q_i\ \exp\{-a_i\}, \quad (۷)$$

15. Slice

که در آن c عبارتست از $\exp(a_i)(1+a_i)^{-1}$، حال با جای‌گذاری c در رابطهٔ (۷) توزیع پسینی q_i بدین‌صورت به‌دست می‌آید:

$$\pi(q_i \mid .) = \frac{1}{1+a_i} Gamma(2,1) + \frac{a_i}{1+a_i} Gamma(1,1),$$

از این‌رو، توزیع q_i توزیع آمیخته‌ای از دو توزیع گاما است. حال با تولید q_i می‌توان از رابطهٔ $w_i = q_i + a_i$، w_i را نیز تولید کرد.

$$\pi(e_{ij} \mid .) \propto (e_{ij})^{-1/2} \exp\left\{-\frac{1}{4\sigma e_{ij}}(y_{ij} - x'_{ij}\beta - \alpha_i - k_1 e_{ij})^2 - \frac{\tau(1-\tau)e_{ij}}{\sigma}\right\}$$
$$\propto (e_{ij})^{-1/2} \exp\left\{-\frac{1}{2}(\hat{\delta}_{ij}^2 e_{ij}^{-1} + \hat{\gamma}_{ij}^2 e_{ij})\right\},$$

از این‌رو، توزیع پسینی شرطی کامل e_{ij} بدین‌صورت است:

$$(e_{ij} \mid .) \sim GIG(\frac{1}{2}, \hat{\delta}_{ij}, \hat{\gamma}_{ij})$$

که در آن

$$\hat{\delta}_{ij}^2 = \sigma^{-1}\ \ 2^{-1}k_1^2 + 2\tau(1-\tau)\text{، }\hat{\gamma}_{ij}^2 = (2\sigma)^{-1}(y_{ij} - x'_{ij}\beta - \alpha_i)^2 \text{ و } GIG(r,m,n)$$

نشان‌دهندهٔ توزیع گاوسی وارون تعمیم یافته[16] با تابع چگالی

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

است. برای تولید نمونهٔ تصادفی از این توزیع، از تابع ()rgig که در بسته ghyp در نرم‌افزار R موجود است، می‌توان استفاده کرد [۱۳].

$$\pi(\sigma \mid .) \propto \sigma^{-N/2}\exp\left\{-\frac{1}{4\sigma}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{(y_{ij} - x'_{ij}\beta - \alpha_i - k_1 e_{ij})^2}{e_{ij}}\right\}$$
$$\times\sigma^{-N}\exp\left\{-\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\tau(1-\tau)e_{ij}}{\sigma}\right\} \times \sigma^{-n}\exp\left\{-\sum_{i=1}^{n}\frac{\nu_i|\alpha_i|}{\sigma}\right\} \times \left(\frac{1}{\sigma}\right)^{c.+1}\exp\left\{-\frac{d.}{\sigma}\right\}$$

بنابراین توزیع پسینی شرطی کامل σ بدین‌صورت به‌دست می‌آید:

$$(\sigma \mid .) \sim IGamma(a,b)$$

که در آن $a = 2^{-1}(3N) + n + c.$ و

$$b = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{\frac{(y_{ij} - x'_{ij}\beta - \alpha_i - k_1 e_{ij})^2}{4e_{ij}} + \frac{\tau(1-\tau)e_{ij}}{\sigma}\right\} + \sum_{i=1}^{n}\frac{\nu_i|\alpha_i|}{\sigma} + d.$$

است. توزیع پسینی شرطی کامل پارامترهای دیگر بدین‌صورت به‌دست می‌آیند:

16. Generalized inverse Gaussian

$$(\nu_i \mid .) \sim Gamma(۲, \frac{|\alpha_i|}{\sigma} + \phi_۱), i = ۱,\ldots,n, \quad (\phi_۱ \mid .) \sim Gamma(n, \sum_{i=۱}^{n} \nu_i),$$

$$(\lambda \mid .) \sim \mathrm{E}\,\mathrm{xp}(\phi_۲ + \sum_{i=۱}^{n} \alpha_i^۲), \qquad (\phi_۲ \mid .) \sim \mathrm{E}\,\mathrm{xp}(\lambda).$$

توجه کنیم که توزیع پسینی شرطی کامل σ و ν_i با انتگرال‌گیری از توزیع پسینی همهٔ پارامترها نسبت به u_i ها و α_i ها محاسبه شده‌اند. بالاخره توزیع پسینی شرطی کامل β بدین‌صورت به‌دست می‌آید:

$$(\beta \mid .) \sim N_k(Bb, B),$$

که در آن

$$b = \frac{۱}{۲\sigma} \sum_{i=۱}^{n} \sum_{j=۱}^{n_i} \frac{x_{ij}(y_{ij} - \alpha_i - k_۱ e_{ij})}{e_{ij}} + B_۰^{-۱} b_۰ \quad \text{و} \quad B^{-۱} = \sum_{i=۱}^{n} \sum_{j=۱}^{n_i} \frac{x_{ij} x_{ij}'}{e_{ij}} + B_۰^{-۱}.$$

حال با مشخص شدن توزیع پسینی شرطی کامل پارامترها و ابرپارامترها می‌توان با استفاده از روش نمونه‌گیری گیبس نمونه‌های کافی تولید کرده و استنباط‌ها را انجام داد. از این‌رو، چنان‌که عنوان شد در این مدل نیازی به برآورد پارامترها به‌روش اعتبارسنجی متقابل نیست.

بررسی شبیه‌سازی

در این بخش برای مقایسه کارایی روش ارائه شده با روش‌های متداول دیگر به بررسی شبیه‌سازی می‌پردازیم. در این شبیه‌سازی مدل را بدین‌صورت در نظر می‌گیریم:

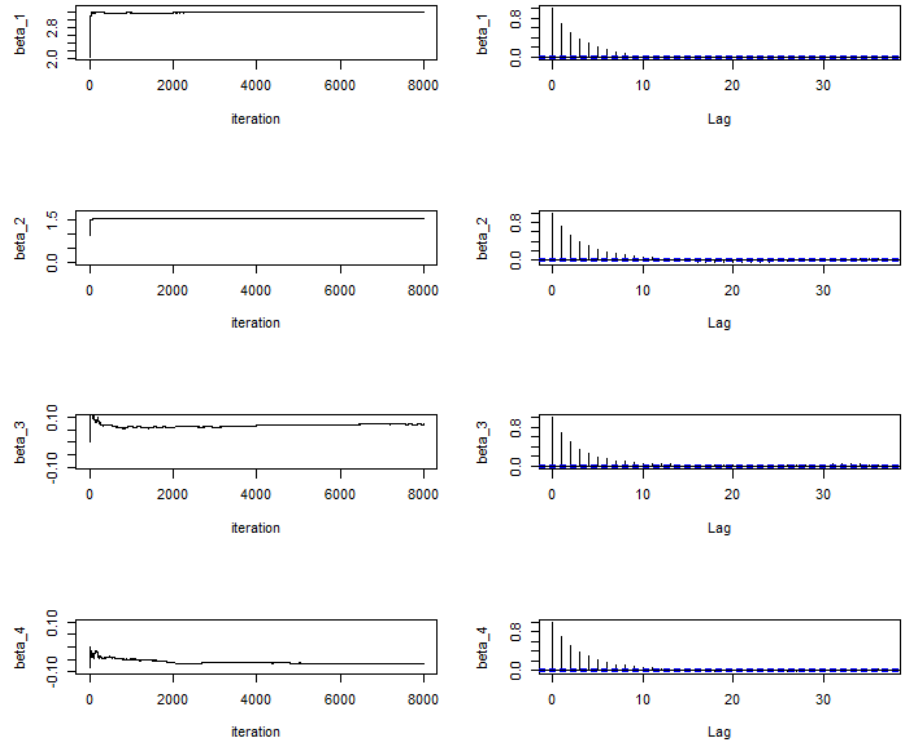
$$y_{ij} = x_{ij}^{۱}\beta_۱ + x_{ij}^{۲}\beta_۲ + \cdots + x_{ij}^{۷}\beta_۷ + x_{ij}^{۸}\beta_۸ + \alpha_i + \varepsilon_{ij},$$

که در آن $i = ۱,\ldots,۱۰۰$ و $j = ۱,\ldots,۵$ است. توجّه کنیم که تعداد متغیرهای پاسخ در مقایسه با متغیرهای تبیینی زیاد است. از این‌رو، استفاده از تاوان الاستیک‌نت سازوار منجر به بهبود برآورد پارامترها و دقت مدل خواهد شد. همانند شبیه‌سازی لیو و همکاران [۱۴] متغیرهای تبیینی یعنی x_{ij}^k را به‌صورت $x_{ij}^k \sim N(۰,۱)$ برای $k = ۱,\ldots,۸$ تولید کرده و توزیع اثرهای تصادفی را نیز برابر $\alpha_i \sim N(۰,۴)$ فرض می‌کنیم. برای β نیز بردارهای

$$\beta = \ ۵,۰,۰,۰,۰,۰,۰,۰ \ , \quad \beta = \ ۳,۱ \,/\, ۵,۰,۰,۲,۰,۰,۰ \ ,$$
$$\beta = \ ۰ \,/\, ۸۵,۰ \,/\, ۸۵,۰ \,/\, ۸۵,۰ \,/\, ۸۵,۰ \,/\, ۸۵,۰ \,/\, ۸۵,۰ \,/\, ۸۵$$

را در نظر می‌گیریم که در آن بردار اول متناظر با یک مدل کاملاً تنک و بردار سوم مربوط به مدلی کاملاً چگال است. برای مؤلفهٔ خطا نیز توزیع‌های نرمال استاندارد، تی با سه درجهٔ آزادی و توزیع لاپلاس متقارن در نظر می‌گیریم. به‌منظور کاهش حساسیت مدل به توزیع‌های پیشینی، برای پارامتر β توزیع پیشینی را $\beta \sim N_۸(۰,۱۰۰I)$ و برای ابرپارامتر σ نیز توزیع پیشینی پخ به‌صورت $IGamma(۰/۰۱, ۰/۰۱)$ فرض می‌شود. مدل ارائه شده یعنی رگرسیون چندکی با تاوان الاستیک‌نت سازوار از دیدگاه آمار بیزی (BENQR) را با روش رگرسیون چندکی معمولی بدون در نظر گرفتن اثرهای تصادفی (QR) و روش رگرسیون چندکی در

داده‌های طولی با در نظر گرفتن تاوان لاسو روی اثرهای تصادفی از دیدگاه آمار بسامدی (PQR) که کوئنکر [۲] ارائه کرده است، مقایسه می‌کنیم. برای برآورد پارامترها در مدل بیزی با استفاده از روش نمونه‌گیری گیبس، ۸۰۰۰ نمونه از توزیع‌های پسینی شرطی کامل پارامترها تولید کرده و ۲۰۰۰ نمونه اول را به‌عنوان دورهٔ همگرایی مطلوب کنار می‌گذاریم. برای ارزیابی همگرایی زنجیر و تعیین تقریبی تعداد نمونه‌ها به‌عنوان همگرایی مطلوب از آمارهٔ $\hat{R}$ که گلمن و همکاران [۱۵] ارائه کرده‌اند، استفاده می‌کنیم. برای همهٔ پارامترها تقریباً بعد از ۲۰۰۰ نمونه اولیه، همگرایی مطلوب رخ می‌دهد که شکل ۱ نمودارهای اثر و تابع خودهمبستگی را با فرض توزیع نرمال برای مؤلفه‌های خطا و $\beta = (3{,}1, 5{,}0{,}0{,}2{,}0{,}0{,}0)$ نشان می‌دهند. برای اجتناب از طولانی شدن نمودارها، فقط برای چهار مؤلفه اول بردار β ، نمودارها نمایش داده شده‌اند.



شکل۱. نموداراثر و تابع خود همبستگی در روش رگرسیون چندکی با تاوان الاستیک‌نت در چندک $\tau = ۰/۵$

هر دو نمودار به‌خوبی همگرایی زنجیره‌های مارکف تولید شده از توزیع‌های پسینی ضرایب را بعد از تقریباً ۲۰۰۰ نمونه نشان می‌دهند. جدول ۱ برآورد و انحراف معیار پارامترهای β برای سه‌روش عنوان شده را در مقایسه با مقدار واقعی β نشان می‌دهد. چنان‌که مشاهده می‌شود روش ارائه شده عملکرد بهتری در برآورد پارامتر β در مقایسه با مقدار واقعی آن‌ها دارد (توجه کنیم در تحلیل بیزی میانگین توزیع پسینی به‌عنوان برآورد در نظر گرفته شده است). برای مقایسهٔ کارایی روش‌ها و به‌منظور لحاظ کردن تغییرپذیری نتایج حاصل از شبیه‌سازی، روند شبیه‌سازی را ۱۰۰ مرتبه تکرار کرده و دو معیار اریبی[17] پارامترها و جذر میانگین توان دوم خطا[18] را طبق این روابط محاسبه می‌کنیم:

17. Bias
18. Root mean square error

$$Bias(\hat{\beta}_k)=\left|\frac{۱}{۱۰۰}\sum_{i=۱}^{۱۰۰}(\hat{\beta}_{ik}-\beta_{ik})\right|, k=۱,\ldots,۸ \quad RMSE(\hat{\beta}_k)=\sqrt{\frac{۱}{۱۰۰}\sum_{i=۱}^{۱۰۰}(\hat{\beta}_{ik}-\beta_{ik})^۲}$$

جدول ۱. برآورد پارامترها و انحراف معیارآن‌ها زمانی که توزیع خطاها نرمال استاندارد و $\tau = ۰/۵$ است

مدل		$\beta_۱$	$\beta_۲$	$\beta_۳$	$\beta_۴$	$\beta_۵$	$\beta_۶$	$\beta_۷$	$\beta_۸$
β_{True}		۵	۰	۰	۰	۰	۰	۰	۰
BENQR	برآورد	۵/۰۹۷	-۰/۰۷۲	۰/۰۰۲	-۰/۰۲۹	-۰/۰۴۰	۰/۰۴۱	۰/۰۰۹	۰/۰۵۶
	انحراف معیار	۰/۱۱۴	۰/۱۲۴	۰/۱۲۱	۰/۱۱۰	۰/۱۲۶	۰/۱۱۸	۰/۱۳۵	۰/۱۱۹
PQR	برآورد	۵/۱۷۳	-۰/۰۸۲	۰/۲۰۹	۰/۰۲۱	۰/۰۵۴	۰/۰۶۶	-۰/۰۴۵	۰/۰۲۰
	انحراف معیار	۰/۱۷۸	۰/۱۵۸	۰/۱۵۷	۰/۱۳۳	۰/۱۴۸	۰/۱۵۳	۰/۱۴۲	۰/۱۸۸
QR	برآورد	۵/۱۲۰	-۰/۲۵۰	-۰/۱۰۳	-۰/۰۷۰	-۰/۱۰۰	۰/۰۳۳	-۰/۰۴۹	۰/۱۷۱
	انحراف معیار	۰/۱۲۲	۰/۱۲۱	۰/۱۲۴	۰/۱۱۲	۰/۱۲۰	۰/۱۲۱	۰/۱۲۶	۰/۱۲۶
β_{True}		۳	۱/۵	۰	۰	۲	۰	۰	۰
BENQR	برآورد	۳/۰۲۶	۱/۴۱۰	۰/۰۲۶	۰/۱۱۶	۲/۰۴۱	۰/۰۴۷	-۰/۱۱۸	-۰/۰۶۱
	انحراف معیار	۰/۱۲۸	۰/۱۱۶	۰/۱۴۱	۰/۱۳۲	۰/۱۲۰	۰/۱۲۰	۰/۱۲۹	۰/۱۲۴
PQR	برآورد	۳/۰۶۴	۱/۴۶۳	۰/۰۴۰	۰/۲۴۴	۲/۱۷۲	۰/۱۳۹	-۰/۱۷۳	-۰/۱۵۴
	انحراف معیار	۰/۱۵۶	۰/۱۵۹	۰/۱۳۶	۰/۱۴۱	۰/۱۶۸	۰/۱۲۷	۰/۱۳۶	۰/۱۴۷
QR	برآورد	۳/۰۴۳	۱/۴۰۰	۰/۲۰۵	-۰/۲۶۳	۲/۰۴۶	۰/۰۶۶	-۰/۱۰۵	۰/۰۸۲
	انحراف معیار	۰/۱۳۹	۰/۱۳۲	۰/۱۴۲	۰/۱۴۹	۰/۱۴۳	۰/۱۳۶	۰/۱۴۶	۰/۱۳۵
β_{True}		۰/۸۵	۰/۸۵	۰/۸۵	۰/۸۵	۰/۸۵	۰/۸۵	۰/۸۵	۰/۸۵
BENQR	برآورد	۰/۷۳۴	۰/۹۰۹	۰/۸۶۳	۰/۹۲۱	۰/۸۶۵	۰/۸۶۸	۰/۷۳۹	۰/۸۸۴
	انحراف معیار	۰/۱۲۹	۰/۱۵۷	۰/۱۱۸	۰/۱۴۵	۰/۱۳۵	۰/۱۳۶	۰/۱۳۶	۰/۱۳۱
PQR	برآورد	۰/۸۶۱	۱/۰۲۸	۰/۹۰۲	۱/۰۲۳	۱/۰۶۵	۰/۹۸۷	۰/۵۷۲	۰/۹۸۵
	انحراف معیار	۰/۱۵۲	۰/۱۲۵	۰/۱۵۸	۰/۱۵۵	۰/۱۶۹	۰/۱۷۳	۰/۱۵۵	۰/۱۳۹
QR	برآورد	۰/۷۶۲	۰/۹۶۵	۰/۸۲۳	۰/۹۴۹	۰/۸۸۲	۰/۹۴۳	۰/۶۴۳	۰/۹۳۹
	انحراف معیار	۰/۱۳۳	۰/۱۳۰	۰/۱۳۰	۰/۱۳۶	۰/۱۳۴	۰/۱۳۳	۰/۱۳۳	۰/۱۳۱

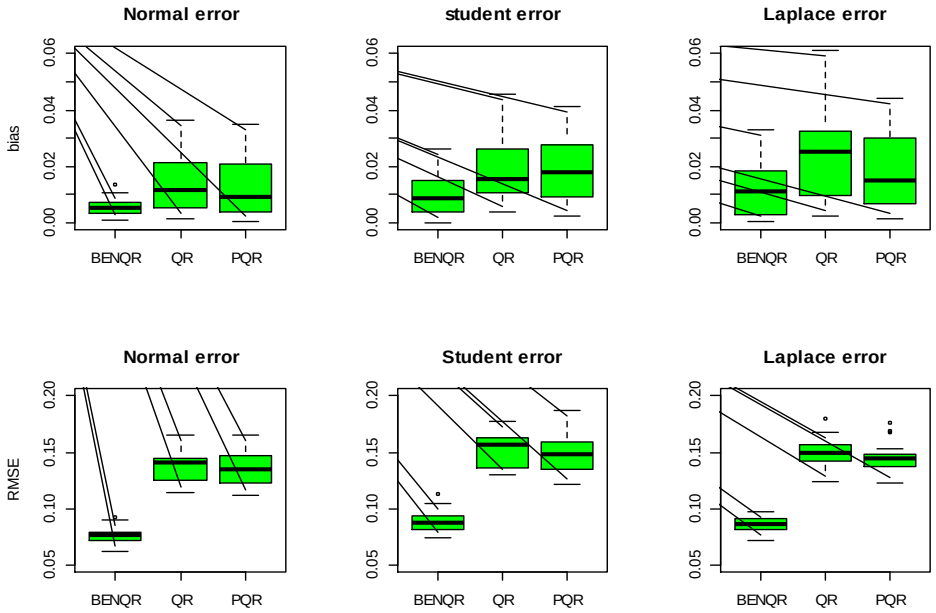
هرچه اندازهٔ این معیارها کوچک باشد، حاکی از کارایی زیاد مدل است. چون نوشتن تمام مقادیر با توجه به وجود سه مدل، دو روش برآورد برای سه توزیع خطا و بیست و چهار پارامتر در قالب جدول امکان‌پذیر نیست، بنابراین نتایج در قالب نمودارهای جعبه‌ای برای چندک‌های $\tau = (۰/۲۵, ۰/۵, ۰/۷۵)$، خلاصه شده‌اند که نتایج آن در شکل‌های ۲، ۳ و ٤ ارائه شده است. هر چه نمودارهای جعبه‌ای کوچک‌تر و نزدیک به صفر باشند، نشان‌دهندهٔ اریبی کم‌تر و برآورد بهتر پارامترها است. در شکل ۲ عملکرد روش‌ها از نقطه نظر اریبی و RMSE در چندک $\tau = ۰/۲۵$ مقایسه می‌شود. با توجه به این شکل روش BENQR در تمامی توزیع‌های فرض شده برای خطاها عملکرد بهتری نسبت به‌سایر روش‌ها دارد. شکل‌های ۳ و٤ نیز عملکرد روش‌ها را از نظر معیارهای ذکر شده در چندک‌های $\tau = ۰/۵$ و $\tau = ۰/۷۵$ مقایسه می‌کند. با توجه به این دو شکل مشاهده می‌شود که مانند شکل ۲، در این نمودارها نیز روش ارائه شده برآورد بهتری از پارامترها در تمامی توزیع‌های فرض شده برای خطا دارند.

تحلیل داده‌های واقعی

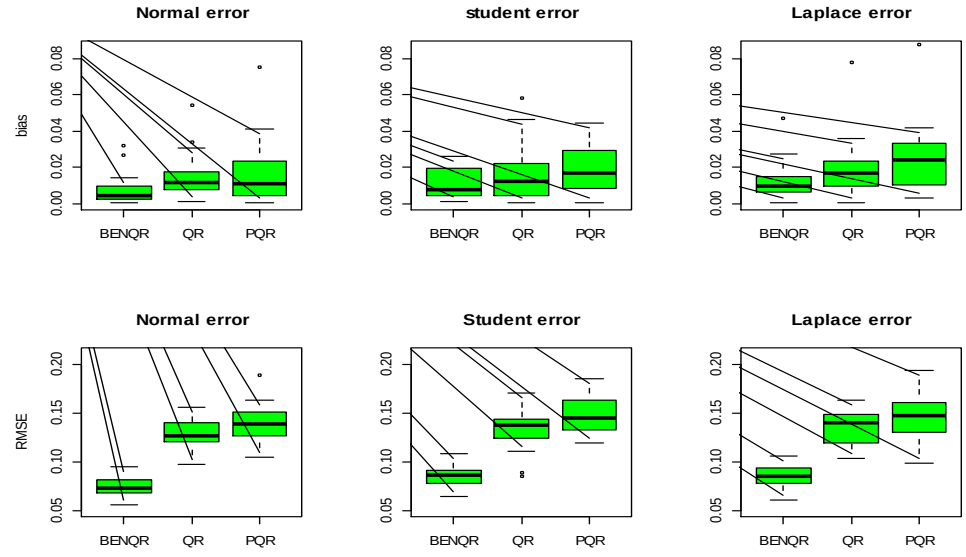
در این بخش به بررسی و تحلیل داده‌های واقعی می‌پردازیم. این داده‌ها شامل رشد اندازهٔ قد ۵۰۰ کودک از بدو تولد تا دوسالگی است که از سال ۱۳۹۱ تا سال ۱۳۹۲ (به‌مدت دو سال) از درمانگاه‌های شهرستان زنجان به تصادف انتخاب و جمع‌آوری شده است. هدف بررسی تأثیر برخی عوامل بر میزان رشد قد کودکان است. این

عوامل یعنی متغیرهای تبیینی عبارت‌اند از: جنسیت کودک $x^{(۱)}$، فاصلهٔ زایمان با کودک قبل $x^{(۲)}$، شاغل بودن یا نبودن مادر $x^{(۳)}$، سن مادر $x^{(۴)}$، چندمین نوزاد خانواده بودن $x^{(۵)}$، تحت مراقبت ویژه بودن مادر $x^{(۶)}$، وزن مادر $x^{(۷)}$، قد مادر $x^{(۸)}$، سن تولد نوزاد $x^{(۹)}$، چندقلو بودن نوزاد $x^{(۱۰)}$، مدت شیردهی انحصاری مادر $x^{(۱۱)}$، مدت شیردهی مادر در کنار غذای کمکی $x^{(۱۲)}$ و متغیر پاسخ نیز اندازهٔقد کودک از بدو تولد تا دو سالگی y است که در فاصله‌های زمانی بدو تولد،۱،۲، ٤، ٦، ۹، ۱۲، ۱۸ و ۲٤ ماهگی اندازه‌گیری شده است. بنابراین مدل با اثرات تصادفی برای این داده‌ها را بدین‌صورت در نظر می‌گیریم:

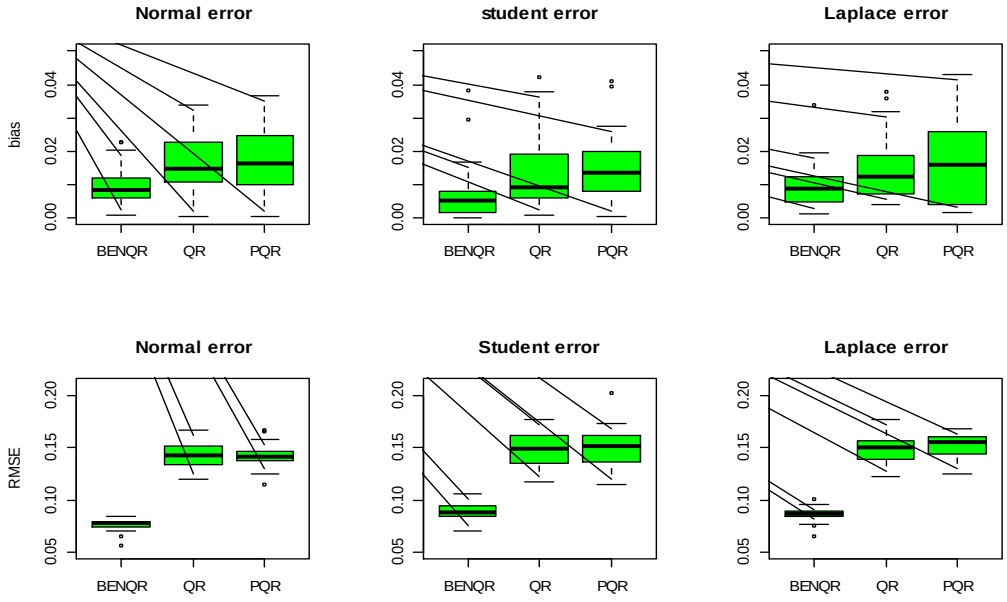
$$y_{ij} = x_{ij}^{(۱)}\beta_۱ + x_{ij}^{(۲)}\beta_۲ + \cdots + x_{ij}^{(۱۲)}\beta_{۱۲} + \alpha_i + \varepsilon_{ij},$$



شکل ۲. عملکرد روش‌ها به لحاظ معیار اریبی و RMSE در چندک $\tau = ۰/۲۵$



شکل ۳. عملکرد روش‌ها به لحاظ معیار اریبی و RMSE در چندک $\tau = ۰/۵$



شکل ٤ . عمل‌کرد روش‌ها به لحاظ معیار اریبی و RMSE در چندک $\tau = ۰/۷۵$

که در آن $i = ۱,\ldots,۵۰۰, j = ۰,۱,۲,۴,۶,۹,۱۲,۱۸,۲۴$ است. در برازش مدل به‌روش BENQR توزیع مؤلفه‌های خطا ε_{ij} را لاپلاس نامتقارن، $ALD(۰,\sigma,\tau)$ در نظر گرفته و با استفاده از روش نمونه‌گیری گیبس، ۲۰۰۰۰ نمونه از توزیع‌های پسینی شرطی کامل پارامترها تولید کرده و ۱۰۰۰۰ نمونه اول را به‌عنوان دورهٔ هم‌گرایی مطلوب کنار می‌گذاریم. برای پارامترهای β و σ، توزیع‌های پیشینی را ناآگاهی‌بخش به‌صورت $\beta \sim N_{۸}(۰,۱۰۰I)$ و $\sigma \sim IGamma(۰/۰۱, ۰/۰۱)$ در نظر می‌گیریم. جدول ۲ برآورد پارامترهای مدل و انحراف معیار آن‌ها را برای سه روش ذکر شده در چندک $\tau = ۰/۵$ نشان می‌دهد. ملاحظه می‌شود که هر سه روش ضرایب مدل را تقریباً یک‌سان برآورد کرده‌اند. اندازهٔ انحراف معیار برآوردها در هر سه روش برای متغیرهای تبیینی $x^{(۴)}$ و $x^{(۱۱)}$ در مقایسه با اندازهٔ قدر مطلق برآوردها کوچک هستند که حاکی از تأثیر معنی‌دار آن‌ها بر متغیر پاسخ است. در هر دو روش QR و PQR در سطح معنی‌داری ۵ درصد تأثیر این دو متغیر معنی‌دار است. برای مقایسه دقت پیش‌بینی سه روش از معیار اعتبارسنجی متقابل ده دسته‌ای[19] استفاده می‌کنیم که معیاری مرسوم در تحلیل داده‌های مقطعی و طولی است [۵]، [۱۶]. برای این منظور مشاهدات مربوط به ۵۰۰ کودک به ۱۰ گروه تقسیم شده (هر گروه شامل ۵۰ کودک است) و سپس در گروه اول، مشاهدات مربوط به ماه‌های ۲٤، ۱۸،۱۲ و ۹ برای هر ۵۰ کودک حذف شده، سپس هر سه مدل به کل داده‌های باقی‌مانده برازش می‌شوند. مقادیر پاسخ برای مشاهدات حذف شده از طریق مدل‌های برازش شده پیش‌بینی شده و میانگین توان دوم انحرافات پاسخ‌های پیش‌بینی شده از مقادیر واقعی آن‌ها[20] (MSE) محاسبه می‌شود. این روند با حذف مشاهدات عنوان شده در گروه ۵۰ تایی دوم ادامه پیدا کرده و بالاخره با گروه دهم پایان می‌پذیرد. در جدول ۳ مقادیر میانگین، میانه و انحراف معیار MSEها که برای ۱۰ گروه در سه چندک مختلف ۰/۵، ۰/۳ و ۰/۷ محاسبه شده، آمده است. چنان‌که از این جدول ملاحظه می‌شود، روش BENQR به لحاظ پیش‌بینی مقایر آینده در طول

19. Ten fold
20. Mean square error

زمان برای چندک‌های متفاوت که یکی از اهداف مهم در تحلیل داده‌های طولی است، عملکرد بهتری نسبت به روش‌های دیگر دارد.

جدول ۲. برآورد پارامترها و انحراف معیار آن‌ها برای سه روش در چندک $\tau = ۰/۵$

پارامترها	BENQR		PQR		QR	
	برآورد	انحراف‌معیار	برآورد	انحراف‌معیار	برآورد	انحراف‌معیار
$\beta_۱$	۰/۱۷۱۷	۰/٤٦۹٦	۰/۰٤۲۸	۰/٦۰٤۲	۰/۰۹٤۰	۰/٦۱٤۳
$\beta_۲$	۰/۳۹۷۷	۰/٦۸۷٥	-۰/٤۳٥۹	۰/۸۷۹۰	۰/٤۰۹۳	۰/۹۲۳۲
$\beta_۳$	۰/۸٥۰۱	۰/۷٤۱۸	۰/۱۰۹٥	۰/۸٥۲۲	۰/۸٥۰۲	۰/۹۷۱۲
$\beta_۴$	-۰/۱۲۳٦	۰/۰۳۳٦	-۰/۱۳٦٦	۰/۰٤۲٥	-۰/۱۲۹۰	۰/۰٥٥۸
$\beta_۵$	۰/۱۱۸۷	۰/۳۸٤٦	۰/۰۰۳۳	۰/٤٥۰۸	۰/۱۳٤۷	۰/٤۷٦٦
$\beta_۶$	-۰/٤٥۱۲	۰/٤٤۰۱	-۰/۲۳۹۷	۰/٥٤۳۱	-۰/٤۹۷۸	۰/٥۷۹۱
$\beta_۷$	۰/۰۱۱٥	۰/۰۲۷۱	-۰/۰۰٦۰	۰/۰۳٤۳	۰/۰۱۰۸	۰/۰۳٥۰
$\beta_۸$	-۰/۰۰۱٦	۰/۰۳٥۰	۰/۰۰۷٤	۰/۰٥٦٦	-۰/۰٤۷	۰/۰٤٤۱
$\beta_۹$	-۰/۰٦۱۷	۰/۱۳۰۳	-۰/۰۱٤۸	۰/۲۱۸۷	-۰/۰۷۹۲	۰/۱٦٥۲
$\beta_{۱۰}$	-۰/۷٥۷۲	۱/٤۱۰۰	-۱/۳۰۹۸	۱/٤۱۱۰	-۰/۹۲٦۸	۱/۰۲۸۸
$\beta_{۱۱}$	۰/۱٥٥۸	۰/۰۲٥٦	۰/۱٤٦٥	۰/۰۲۹٤	۰/۲۰۰٤	۰/۰۱۲۱
$\beta_{۱۲}$	۰/۰۷۱۱	۰/۱۰۷۰	۰/۰٦۷۷	۰/۱۳۱۰	۰/۰٥۷۱۷	۰/۱۳۱٤

جدول ۳. میانگین، میانه و انحراف معیار MSE ها در چندک‌های ۰/۷، ۰/۳ و ۰/۵

چندک	مدل	میانگین	میانه	انحراف معیار
$\tau = ۰/۵$	BENQR	۱٥۱/٦٥۷۹	۱٥۰/۸۷٥٤	٤/۱۸۰۰
	PQR	۱٥۲/۸۷٦٤	۱٥٤/٥۰۸٤	٥/٦۷٥٥
	QR	۱٥۲/٥۲۸٥	۱٥٤/٦٥٤۸	٥/٦۳٥۳
$\tau = ۰/۳$	BENQR	۱٥٤/۲۳٤٦	۱٥۳/۷٤٥۹	٦/۲۰۰۱
	PQR	۱٥٦/۹٤۳۸	۱٥٥/۸۳۷۳	۸/۰٤۸٦
	QR	۱٥٥/٤۸٦۷	۱٥٥/۹۳۲۸	۷/۹۳٤۷
$\tau = ۰/۷$	BENQR	۱٦۱/۱۲۸۷	۱٥۸/۸٤۳٦	٤/٥٦٤۰
	PQR	۱٦٥/٤۳۲٦	۱٦٤/۹٤۳۷	٦/۸۳٤٥
	QR	۱٦٦/٥٦۳٤	۱٦۳/۸٤۳۹	۷/٥٦٤۸

بحث و نتیجه‌گیری

در پژوهش‌های طولی علاوه بر اثرات ثابت که تأثیر هر یک از متغیرهای تبیینی 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 بررسی‌هایی که تعداد مشاهدات از متغیر پاسخ زیاد است، از این رو، 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 ترکیب توابع تاوان لاسوی سازوار و ستیغی و ایجاد تاوان الاستیک‌نت سازوار روی اثرات تصادفی، علاوه بر این‌که اثرات تصادفی را به‌سمت صفر منقبض کرده و اثرات کم اهمیت از مدل حذف شدند، اثرهای تصادفی که تأثیر یک‌سان بر متغیر پاسخ داشته‌اند یا دارای هم‌بستگی زیادی بودند، به‌طور یک‌سان برآورد شده و از این رو، منجر به افزایش دقت مدل شده است. تاوان الاستیک‌نت سازوار با تعریف توزیع پیشینی مناسب روی اثرهای

تصادفی، در مدل لحاظ شده و از دیدگاه آماربیزی بررسی شد. مزیت مدل ارائه شده در برآورد همهٔ پارامترهای مدل با روش نمونه‌گیری گیبس است. نتایج حاصل از شبیه‌سازی نشان داد که روش بیزی ارائه شده در تمامی چندک‌ها نسبت به‌روش‌های متداول دیگر در رگرسیون چندکی برای داده‌های طولی با اثرات تصادفی در برآورد ضرایب رگرسیونی عمل‌کرد بهتری دارد. نتایج حاصل از تحلیل داده‌های واقعی نیز مشخص کرد که روش ارائه شده دقت زیادی نسبت به دو روش دیگر دارد. بنابراین استفاده از مدل ارائه شده در تحلیل داده‌های واقعی توصیه می‌شود.

منابع

1. Koenker R., Bassett G., "Quantile regression", Econometrica, 46 (1978).
2. Koenker R., "Quantile regression for longitudinal data", Journal of Multivariate Analysis, 91 (2004) 74-89.
3. Tibshirani R., "Regression shrinkage and selection via the Lasso", Journal of the Royal Statistical Society, Series B, 58 (1996) 267-288.
4. Zou H., "The adaptive Lasso and its oracle properties", Journal of the American Statistical Association, 101 (2006) 1418-1429.
5. Zou H., Hastie T., "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society Series B, 67 (2005) 301-320.
6. Hoerl A.E., Kennard R., "Ridge regression-applications to non orthogonal problems", Technometrics, 12 (1970) 61-93.
7. Li Q., Lin N., "The Bayesian elastic net", Bayesian Analysis, 5 (2010) 151-170.
8. Chen M., Carlson D., Zaas A., Woods C., Ginsburg G.S., Lucas J., Carin L., "Detection of viruses via statistical gene expression analysis", IEEE Transactions on Biomedical Engineering, 58 (2011) 468-479.
9. Yu K., Moyeed R.A., "Bayesian quantile regression", Statistics and Probability Letters, 54 (2001) 437-447.
10. Geraci M., Bottai M., "Quantile regression for longitudinal data using the asymmetric Laplace distribution", Biostatistics, 8 (2007) 140-154.
11. Kozumi H., Kabayashi G., "Gibbs sampling methods for Bayesian quantile regression", Journal of Statistical Computation and Simulation, 81 (2011) 1565-1578.
12. Polson N.G., Scott J.G., Windle J., "The Bayesian Bridge", Journal of the Royal Statistical Society, Series B, 76 (2014) 713-733.

13. Breymann W., Luthi D., "ghyp: A package on the generalized hyperbolic Distribution and its special cases, R Package Version 1.5.6 URL http:// www.r-project.org (2014).
14. Luo Y., Lian H., Tian M., "Bayesian quantile regression for longitudinal data model", Journal of Statistical Computation and Simulation, 82 (2012) 1635-1649.
15. Gelman A., Carlin J.B., Stern H. S., Rubin D.B., "Bayesian data analysis", Chapman and Hall, London 1995.
16. Lin T.I., Lee J.C., "Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data", Statistics in Medicine, 27 (2008) 1490-1507.