

رگرسیون چندکی بیزی با توان الاستیکنت سازوار برای داده‌های طولی

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

پذیرش ۹۴/۱۰/۶

دریافت ۹۳/۱۰/۱۳

چکیده

بررسی داده‌های طولی قسمت مهمی از پژوهش‌های 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 توان 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 مقایسه کارایی روش ارائه شده با روش‌های متداول، بررسی شبیه‌سازی انجام شده و در پایان نحوه کاربست مدل‌ها در قالب مثال کاربردی شرح داده می‌شود.

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

مقدمه

در بسیاری از تحقیقات مربوط به علوم پزشکی و علوم اجتماعی 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، به عبارت دیگر یک واحد آزمایشی تحت اندازه‌گیری مکرر در طول زمان قرار می‌گیرد. زیرا اندازه 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، مطالعه طولی^۱ و به داده‌های جمع‌آوری شده در این نوع پژوهش‌ها، داده‌های طولی^۲ گویند. هدف اولیه این تحقیقات تشخیص عواملی است که متغیرهای 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^۳ است. این مدل رگرسیونی را اولین بار کوئنکر و باست [۱] ارائه کرد. این روش نسبت به روش‌های دیگر دارای دو مزیت مهم است. نخست این‌که رگرسیون چندکی نسبت به ناپایداری واریانس و داده‌های دور افتاده حساس نیست و دوم این‌که اطلاعات جزئی‌تری نسبت به تأثیر متغیرهای تبیینی در چندک‌های مختلف توزیع متغیر پاسخ ارائه می‌دهد.

در بررسی داده‌های طولی علاوه بر اثرهای ثابت که تغییرات درون گروه‌ها (تغییرات در طول زمان) را کنترل می‌کند، اثرات تصادفی^۴ 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^۵ (تاوان با نرم L_1) روی اثرات تصادفی بررسی کرد. در واقع این روش با ایجاد تاوان، اثرات تصادفی کم‌اهمیت را به سمت صفر ترنجش^۶ کرده و مدلی تنک^۷ ایجاد می‌کند.

اولین بار تیشیرانی [۳] رگرسیون لاسو که با ایجاد تاوان لاسو روی پارامترها انجام می‌شود را ارائه کرد. سپس زو [۴] رگرسیون لاسوی سازوار^۸ را به‌منظور گسترش رگرسیون لاسو پیشنهاد کرد. در این روش بر خلاف تاوان لاسو که برای همه ضرایب اندازه تاوان را یکسان در نظر می‌گیرد، اندازه تاوان‌های مختلفی برای ضرایب رگرسیونی متفاوت لحاظ می‌شود. بررسی‌های زو [۴] 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. در غیر این صورت عملکرد این روش مانند روش کمترین توان‌های دوم عادی^۹ (OLS) خواهد بود. در واقع دقت روش‌های رگرسیونی لاسو به انتخاب کران بسیار حساس است [۵]. از طرف دیگر در حالت $n > k$ 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 ستیگی^{۱۰} به لحاظ پیش‌بینی نسبت به روش‌های لاسو عملکرد بهتری دارد [۳]. از این‌رو، زو و هستی [۵]، روش رگرسیونی الاستیکنت^{۱۱} که تابع تاوان آن ترکیبی اکیداً محدب از توابع تاوان لاسو و ستیگی است، را معرفی کردند. چنان‌که دیدیم رگرسیون ستیگی را اولین بار هورل و کنارد [۶]، 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 ستیگی، مدل رگرسیونی با تاوان الاستیکنت علاوه بر ترنجش ضرایب کم‌اهمیت مدل به سمت 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 [۵]. 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^{۱۲}

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 ستیخی 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 ولین [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 ولین [۷] و چن و همکاران [۸] در مدل ارائه شده برای پارامترهای تنظیم در هر دو تاوان لاسو و ستیخی 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 $y_i = x_i' \beta + \varepsilon_i, i = 1, \dots, n$ را در نظر بگیرید که در آن بردار مربوط به متغیرهای تبیینی، β یک بردار $k \times 1$ بعدی از پارامترها و ε_i مؤلفه خطا است. در مدل رگرسیونی میانگین، هدف استنباط دربارهٔ متوسط مقدار پاسخ به‌ازای مقادیر متفاوت متغیرهای تبیینی است، از این‌رو، به $E(y_i | x_i) = x_i' \beta$ توجه شده است. اما در رگرسیون چندکی، به این تابع چندک شرطی توجه شده است:

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

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 $x_i' \beta$ چندک τ -ام شرطی متغیر y_i به شرط معلوم بودن بردار x_i را نشان می‌دهد. در رگرسیون معمولی میانگین توزیع خطاها صفر در نظر گرفته می‌شود، اما در رگرسیون چندکی، چندک τ -ام توزیع ε_i ‌ها برابر صفر هستند، یعنی $\int_{-\infty}^{\tau} f_{\varepsilon_i}(t) dt = \tau$. در رگرسیون چندکی، ضرایب رگرسیونی یعنی β از مینیم کردن این عبارت نسبت به β برآورد می‌شود:

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

که در آن $\rho_\tau(\cdot)$ نشان‌دهنده تابع زیان و به صورت $\rho_\tau(u) = \frac{|u| + (2\tau - 1)u}{2}$ تعریف می‌شود [۱].

چون رابطه (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 موئید [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:

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

که در آن $0 < \tau < 1$ 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 $y = (y_1, \dots, y_n)$ بدین‌صورت به‌دست می‌آید:

$$L(\beta, \sigma | y, x, \tau) = \left(\frac{\tau(1-\tau)}{\sigma} \right)^n \exp \left\{ -\sum_{i=1}^n \rho_{\tau} \left(\frac{y_i - x'_i\beta}{\sigma} \right) \right\}, \quad (2)$$

بنابراین ماکسیم کردن رابطه (۲) در حضور پارامتر مزاحم σ با مینیم کردن رابطه (۱) نسبت به β معادل است. به‌همین دلیل در بسیاری از موارد، تحلیل رگرسیون چندکی با استفاده از این توزیع بررسی می‌شود. یو و موئید [۹] با فرض $\sigma = 1$ و تعریف توزیع پیشینی ناآگاهی‌بخش برای β ($\pi(\beta) \propto 1$) رگرسیون چندکی بیزی را در داده‌های مقطعی معرفی و بررسی کردند.

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

$$y_{ij} = x'_{ij}\beta + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, n_i, \quad (3)$$

است که در آن y_{ij} ، j -امین پاسخ اندازه‌گیری شده روی i -امین واحد آزمایشی و x_{ij} بردار $1 \times k$ بعدی مربوط به متغیرهای تبیینی، β یک بردار $1 \times k$ بعدی از پارامترها و α_i اثر تصادفی مربوط به پاسخ y_{ij} را نشان می‌دهد. توجه شود که در این مدل α_i ها 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 توان لاسو روی بردار $\alpha = (\alpha_1, \dots, \alpha_n)$ ($p(\alpha) = \lambda \sum_{i=1}^n |\alpha_i|$) پارامترهای مدل یعنی

α و β را از مینیم کردن تابع زیان

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

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 توانیده فوق نسبت به‌روش‌های غیرتوانیده 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_i ها لاپلاس متقارن هستند، مدل کوننکر [۲] به‌دست می‌آید. با توجه به مشکلات تاوان لاسو که در بخش اول به آن‌ها اشاره شد، هدف این مقاله بررسی مدل رگرسیون چندکی در داده‌های 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

$$\sum_{i=1}^n \sum_{j=1}^{n_i} \rho_{\tau} (y_{ij} - x'_{ij} \beta - \alpha_i) + \sum_{i=1}^n \nu_i |\alpha_i| + \lambda \sum_{i=1}^n \alpha_i^2 \quad (4)$$

به‌دست می‌آید که در آن مؤلفه $\sum_{i=1}^n \nu_i |\alpha_i| + \lambda \sum_{i=1}^n \alpha_i^2$ را تاوان الاستیکنت سازوار گویند. در این تاوان به

عبارت $\sum_{i=1}^n \nu_i |\alpha_i|$ 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 مینیم کردن عبارت رابطه (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 رابطه (۴) است. اگر فرض کنیم در رابطه (۳)، $\varepsilon_{ij} \sim_{iid} ALD(\cdot, \sigma, \tau)$ و توزیع پیشینی برای α_i را مانند چن و همکاران [۸] بدین‌صورت در نظر بگیریم:

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

آن‌گاه توزیع پسینی بردار α بدین‌صورت به‌دست می‌آید:

$$\pi(\alpha | y, x, \sigma, \lambda, \nu, \beta) \propto \sigma^N \exp \left\{ -\sigma^{-1} \left(\sum_{i=1}^n \sum_{j=1}^{n_i} \rho_{\tau} (y_{ij} - x'_{ij} \beta - \alpha_i) - \sum_{i=1}^n \nu_i |\alpha_i| \right) - \lambda \sum_{i=1}^n \alpha_i^2 \right\}, \quad (5)$$

که در آن $N = \sum_{i=1}^n n_i$ تعداد کل مشاهدات را نشان می‌دهد. از این‌رو، مینیم کردن تابع هدف (۴) معادل

ماکسیم کردن توزیع پسینی در رابطه (۵) در حضور پارامتر مزاحم σ است. بنابراین می‌توان توزیع لاپلاس نامتقارن را برای این مدل استفاده کرد. با توجه به رابطه (۵) توزیع پسینی بردار α توزیع شناخته شده‌ای نیست و نمی‌توان برای استنباط از آن استفاده کرد. از این‌رو، برای سادگی مدل و استفاده از روش نمونه‌گیری گیبس، ابتدا توزیع لاپلاس نامتقارن و عبارت $-(\sigma)^{-1} \nu_i |\alpha_i|$ را به‌صورت توزیع‌های آمیخته در

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

$$u = k_{\tau} e + \sqrt{\tau\sigma} z$$

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

$$y_{ij} = x'_{ij}\beta + \alpha_i + k_{\tau} e_{ij} + \sqrt{\tau\sigma} z_{ij}, \quad (۴)$$

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 $|\alpha_i| \nu_i (\sigma)^{-1} \exp(-(\sigma)^{-1} \nu_i)$ که همان توزیع لاپلاس متقارن است را می‌توان به‌صورت توزیعی آمیخته از هسته‌های فیجر-بارتلت با توزیع آمیختگی گاما بدین‌صورت نوشت [۱۲]:

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

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 $\pi(\lambda | \phi_{\tau}) = \phi_{\tau} \exp\{-\phi_{\tau} \lambda\}$ و $\pi(\nu_i | \phi_{\lambda}) = \phi_{\lambda} \exp\{-\phi_{\lambda} \nu_i\}$ 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_{λ} کوچک باشد، آنگاه ترجیح به‌سمت صفر α_i ‌ها بیشتر و توانیدن آن‌ها نیز بزرگ است. چون تعداد α_i ‌ها زیاد است، از این‌رو، برای پارامتر ϕ_{λ} توزیع پیشینی را به‌صورت پخ^{۱۴} یعنی $\pi(\phi_{\lambda}) \propto \phi_{\lambda}^{-1}$ در نظر می‌گیریم. برای ϕ_{τ} نیز توزیع پیشینی را به‌صورت $\pi(\phi_{\tau}) \propto \phi_{\tau}^{-1}$ قرار می‌دهیم. حال با توجه به مدل روابط (۵)، (۶) و توزیع‌های پیشینی تعریف شده، مدل سلسله مراتبی را می‌توان بدین‌صورت خلاصه کرد:

$$y_{ij} | x_{ij}, \beta, \alpha_i, \sigma, e_{ij} \sim N(x'_{ij}\beta + \alpha_i + k_{\tau} e_{ij}, \tau\sigma e_{ij})$$

$$e_{ij} | \sigma \sim \text{Exp}\left(\frac{\tau(1-\tau)}{\sigma}\right), \quad \beta | b, B. \sim N_k(b, B),$$

$$\pi(\alpha_i | \nu_i, \lambda, \sigma, w_i) \propto \exp\{-\lambda \alpha_i\} \times \frac{\nu_i}{\sigma} \left\{1 - \frac{\nu_i |\alpha_i|}{w_i \sigma}\right\}_+, \quad w_i \sim \text{Gamma}(\tau, 1)$$

$$\nu_i | \phi_{\lambda} \sim \text{Exp}(\phi_{\lambda}), \quad \lambda | \phi_{\tau} \sim \text{Exp}(\phi_{\tau}), \quad \pi(\phi_{\lambda}) \propto \frac{1}{\phi_{\lambda}},$$

$$\pi(\phi_{\tau}) \propto \frac{1}{\phi_{\tau}}, \quad \sigma | c, d. \sim \text{IGamma}(c, d),$$

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

پارامترها و ابرپارامترها به صورت بسته قابل حصول نیست. اما می‌توان توزیع‌های پسینی شرطی کامل همه پارامترها و ابرپارامترها را بدین صورت محاسبه کرد:

هسته توزیع پسینی α_i بدین صورت به دست می‌آید:

$$\pi(\alpha_i | \cdot) \propto \exp \left\{ -\frac{1}{\sigma^2} \sum_{j=1}^{n_i} \frac{(y_{ij} - x'_{ij}\beta - \alpha_i - k_1 e_{ij})^2}{e_{ij}} \right\} \times \exp \{-\lambda \alpha_i^2\} \times \left\{ 1 - \frac{\nu_i |\alpha_i|}{w_i \sigma} \right\}_+.$$

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 $u_1, \dots, u_n$ که دارای توزیع یکنواخت

در بازه $(0, 1)$ هستند، توزیع پسینی α_i را به صورت توزیع توأم

$$\begin{aligned} \pi(\alpha_i, u_i | \cdot) &\propto \exp \left\{ -\frac{1}{\sigma^2} \sum_{j=1}^{n_i} \frac{(y_{ij} - x'_{ij}\beta - \alpha_i - k_1 e_{ij})^2}{e_{ij}} \right\} \\ &\times \exp \{-\lambda \alpha_i^2\} \times I(0 < u_i < 1 - \frac{\nu_i |\alpha_i|}{w_i \sigma}), \end{aligned}$$

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

$$(u_i | \alpha_i, \cdot) \sim \text{Unif}(\cdot, 1 - \frac{\nu_i |\alpha_i|}{w_i \sigma}), i = 1, \dots, n,$$

که در آن $\text{Unif}(a, b)$ توزیع یکنواخت در بازه (a, b) است. همچنین

$$(\alpha_i | u_i, \cdot) \sim N(\bar{\mu}_i, \tilde{\sigma}_i) \times I(|\alpha_i| \leq \frac{(1 - u_i)\sigma w_i}{\nu_i}), i = 1, \dots, n$$

که در آن

$$\bar{\mu}_i = \tilde{\sigma}_i \left[\frac{1}{\sigma^2} \sum_{j=1}^{n_i} \frac{(y_{ij} - x'_{ij}\beta - k_1 e_{ij})}{e_{ij}} \right] \text{ و } \tilde{\sigma}_i = (\lambda + \frac{1}{\sigma^2} \sum_{j=1}^{n_i} \frac{1}{e_{ij}})$$

است. بنابراین توزیع پسینی شرطی کامل α_i نرمال بریده شده در نقاط $\frac{(1 - u_i)\sigma w_i}{\nu_i}$ و $-\frac{(1 - u_i)\sigma w_i}{\nu_i}$

با میانگین $\bar{\mu}_i$ و واریانس $\tilde{\sigma}_i$ است.

$$\pi(w_i | \cdot) \propto w_i \exp\{-w_i\} \times I(w_i \geq \frac{\nu_i |\alpha_i|}{(1 - u_i)\sigma}), i = 1, \dots, n.$$

اگر $a_i = \nu_i |\alpha_i| [(1 - u_i)\sigma]^{-1}$ ، آنگاه توزیع پسینی w_i توزیع گامای بریده شده در نقطه a_i است. برای سهولت در نمونه‌گیری از این توزیع، فرض می‌کنیم $q_i = w_i - a_i$ ، آنگاه توزیع q_i بدین صورت به دست می‌آید:

$$\begin{aligned} \pi(q_i | \cdot) &= c(q_i + a_i) \exp \{-(q_i + a_i)\} \times I(q_i \geq 0) \\ &= c q_i \exp\{-q_i\} \exp\{-a_i\} + c a_i \exp \{-q_i\} \exp\{-a_i\}, \quad (V) \end{aligned}$$

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

$$\pi(q_i | \cdot) = \frac{1}{1 + a_i} \text{Gamma}(r, 1) + \frac{a_i}{1 + a_i} \text{Gamma}(1, 1),$$

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

$$\begin{aligned} \pi(e_{ij} | \cdot) &\propto (e_{ij})^{-1/r} \exp \left\{ -\frac{1}{4\sigma e_{ij}} (y_{ij} - x'_{ij}\beta - \alpha_i - k_i e_{ij})^r - \frac{\tau(1-\tau)e_{ij}}{\sigma} \right\} \\ &\propto (e_{ij})^{-1/r} \exp \left\{ -\frac{1}{r} (\hat{\delta}_{ij}^r e_{ij}^{-1} + \hat{\gamma}_{ij}^r e_{ij}) \right\}, \end{aligned}$$

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

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

که در آن

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

نشان‌دهنده توزیع گاوسی وارون تعمیم یافته^{۱۶} با تابع چگالی

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

است. برای تولید نمونه تصادفی از این توزیع، از تابع $\text{rgig}()$ که در بسته ghyp در نرم‌افزار R موجود است، می‌توان استفاده کرد [۱۳].

$$\begin{aligned} \pi(\sigma | \cdot) &\propto \sigma^{-N/r} \exp \left\{ -\frac{1}{4\sigma} \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{(y_{ij} - x'_{ij}\beta - \alpha_i - k_i e_{ij})^r}{e_{ij}} \right\} \\ &\propto \sigma^{-N} \exp \left\{ -\sum_{i=1}^n \sum_{j=1}^{n_i} \frac{\tau(1-\tau)e_{ij}}{\sigma} \right\} \times \sigma^{-n} \exp \left\{ -\sum_{i=1}^n \frac{\nu_i |\alpha_i|}{\sigma} \right\} \times \left(\frac{1}{\sigma} \right)^{c+1} \exp \left\{ -\frac{d}{\sigma} \right\} \end{aligned}$$

بنابراین توزیع پسینی شرطی کامل σ بدین‌صورت به‌دست می‌آید:

$$(\sigma | \cdot) \sim \text{IGamma}(a, b)$$

که در آن $a = r^{-1}(rN) + n + c$ و

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

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

$$(\nu_i | \cdot) \sim \text{Gamma}(\tau, \frac{|\alpha_i|}{\sigma} + \phi_i), i = 1, \dots, n, \quad (\phi_i | \cdot) \sim \text{Gamma}(n, \sum_{i=1}^n \nu_i),$$

$$(\lambda | \cdot) \sim \text{Exp}(\phi_\tau + \sum_{i=1}^n \alpha_i^\tau), \quad (\phi_\tau | \cdot) \sim \text{Exp}(\lambda).$$

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

$$(\beta | \cdot) \sim N_k(Bb, B),$$

که در آن

$$b = \frac{1}{\tau\sigma} \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{x_{ij}(y_{ij} - \alpha_i - k e_{ij})}{e_{ij}} + B^{-1}b. \quad \text{و} \quad B^{-1} = \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{x_{ij}x'_{ij}}{e_{ij}} + B^{-1}.$$

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:

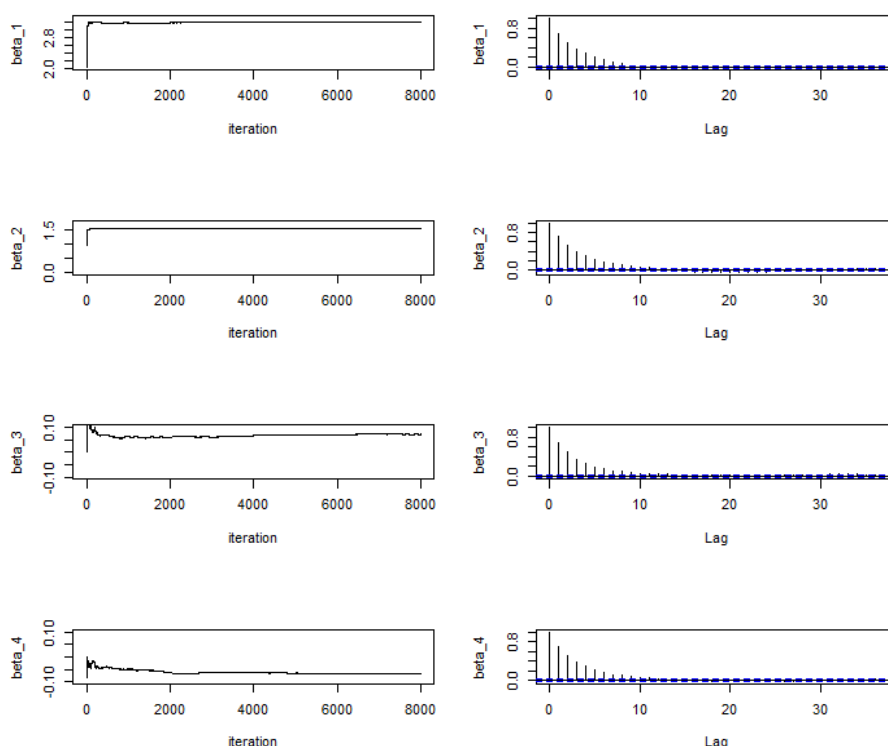
$$y_{ij} = x_{ij}^1\beta_1 + x_{ij}^2\beta_2 + \dots + x_{ij}^v\beta_v + x_{ij}^8\beta_8 + \alpha_i + \varepsilon_{ij},$$

که در آن $i = 1, \dots, 100$ و $j = 1, \dots, 5$ است. توجه کنیم که تعداد متغیرهای پاسخ در مقایسه با متغیرهای تبیینی زیاد است. از این‌رو، استفاده از توان 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 $x_{ij}^k \sim N(0, 1)$ برای $k = 1, \dots, 8$ تولید کرده و توزیع اثرهای تصادفی را نیز برابر $\alpha_i \sim N(0, 4)$ فرض می‌کنیم. برای β نیز بردارهای

$$\beta = 5, 0, 0, 0, 0, 0, 0, 0, \quad \beta = 3, 1 / 5, 0, 0, 2, 0, 0, 0, \\ \beta = 0 / 85, 0 / 85, 0 / 85, 0 / 85, 0 / 85, 0 / 85, 0 / 85$$

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 مؤلفه خطا نیز توزیع‌های نرمال استاندارد، تی با سه درجه آزادی و توزیع لاپلاس متقارن در نظر می‌گیریم. به‌منظور کاهش حساسیت مدل به توزیع‌های پیشینی، برای پارامتر β توزیع پیشینی را $\beta \sim N_8(0, 100I)$ و برای ابرپارامتر σ نیز توزیع پیشینی پخ به‌صورت $IGamma(0 / 0.1, 0 / 0.1)$ فرض می‌شود. مدل ارائه شده یعنی رگرسیون چندکی با توان 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 توان 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 به عنوان دوره همگرایی مطلوب کنار می‌گذاریم. برای ارزیابی همگرایی زنجیر و تعیین تقریبی تعداد نمونه‌ها به عنوان همگرایی مطلوب از آماره $\hat{R}$ که گلن و همکاران [۱۵] ارائه کرده‌اند، استفاده می‌کنیم. برای همه 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 $\beta = ۳,۱ / ۵,۰,۰,۲,۰,۰,۰$ نشان می‌دهند. برای اجتناب از طولانی شدن نمودارها، فقط برای چهار مؤلفه اول بردار β ، نمودارها نمایش داده شده‌اند.



شکل ۱. نمودار اثر و تابع خود همبستگی در روش رگرسیون چندکی با توان الاستیکنت در چندک $\tau = ۰/۵$ هر دو نمودار به خوبی 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 سه روش 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 به عنوان برآورد در نظر گرفته شده است). برای مقایسه کارایی روش‌ها و به منظور لحاظ کردن تغییرپذیری نتایج حاصل از شبیه‌سازی، روند شبیه‌سازی را ۱۰۰ مرتبه تکرار کرده و دو معیار ارزیابی^{۱۷} پارامترها و جذر میانگین توان دوم خطا^{۱۸} را طبق این روابط محاسبه می‌کنیم:

17. Bias

18. Root mean square error

$$Bias(\hat{\beta}_k) = \left| \frac{1}{100} \sum_{i=1}^{100} (\hat{\beta}_{ik} - \beta_{ik}) \right|, k = 1, \dots, 8 \quad RMSE(\hat{\beta}_k) = \sqrt{\frac{1}{100} \sum_{i=1}^{100} (\hat{\beta}_{ik} - \beta_{ik})^2}$$

جدول ۱. برآورد پارامترها و انحراف معیار آن‌ها زمانی که توزیع خطاها نرمال استاندارد و $\tau = 0.5$ است

مدل	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
β_{True}	۰	۰	۰	۰	۰	۰	۰	۰
BENQR	۰/۰۹۷	-۰/۰۷۲	۰/۰۰۲	-۰/۰۲۹	-۰/۰۴۰	۰/۰۴۱	۰/۰۰۹	۰/۰۰۶
انحراف معیار	۰/۱۱۴	۰/۱۲۴	۰/۱۲۱	۰/۱۱۰	۰/۱۲۶	۰/۱۱۸	۰/۱۳۵	۰/۱۱۹
PQR	۰/۱۷۳	-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	۰/۱۲۰	-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	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	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	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	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	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	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

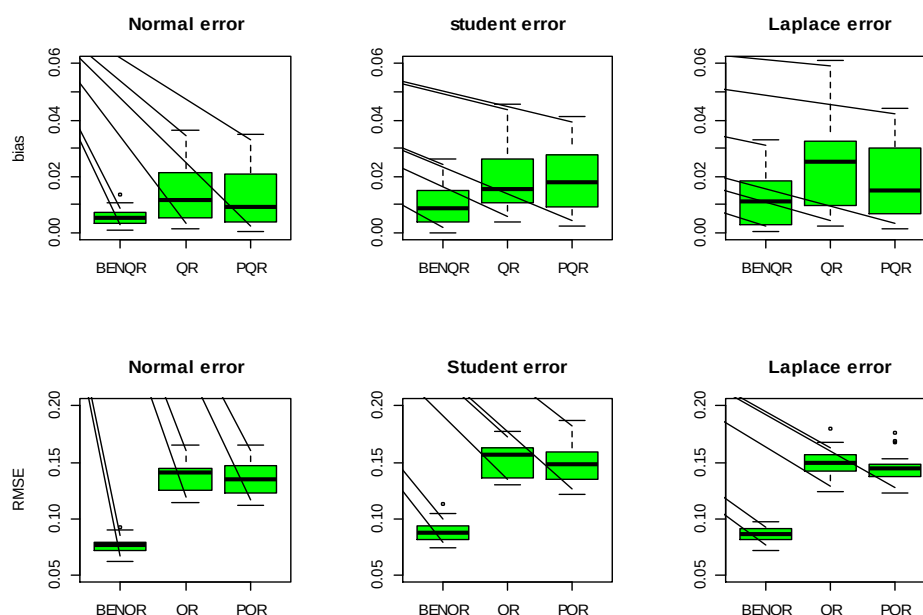
هرچه اندازه این معیارها کوچک باشد، حاکی از کارایی زیاد مدل است. چون نوشتن تمام مقادیر با توجه به وجود سه مدل، دو روش برآورد برای سه توزیع خطا و بیست و چهار پارامتر در قالب جدول امکان‌پذیر نیست، بنابراین نتایج در قالب نمودارهای جعبه‌ای برای چندک‌های (۰/۲۵، ۰/۵، ۰/۷۵) $\tau =$ ، خلاصه شده‌اند که نتایج آن در شکل‌های ۲، ۳ و ۴ ارائه شده است. هر چه نمودارهای جعبه‌ای کوچکتر و نزدیک به صفر باشند، نشان‌دهنده آریبی کمتر و برآورد بهتر پارامترها است. در شکل ۲ عملکرد روش‌ها از نقطه نظر آریبی و RMSE در چندک $\tau = 0.25$ مقایسه می‌شود. با توجه به این شکل روش BENQR در تمامی توزیع‌های فرض شده برای خطاها عملکرد بهتری نسبت به سایر روش‌ها دارد. شکل‌های ۳ و ۴ نیز عملکرد روش‌ها را از نظر معیارهای ذکر شده در چندک‌های $\tau = 0.5$ و $\tau = 0.75$ 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

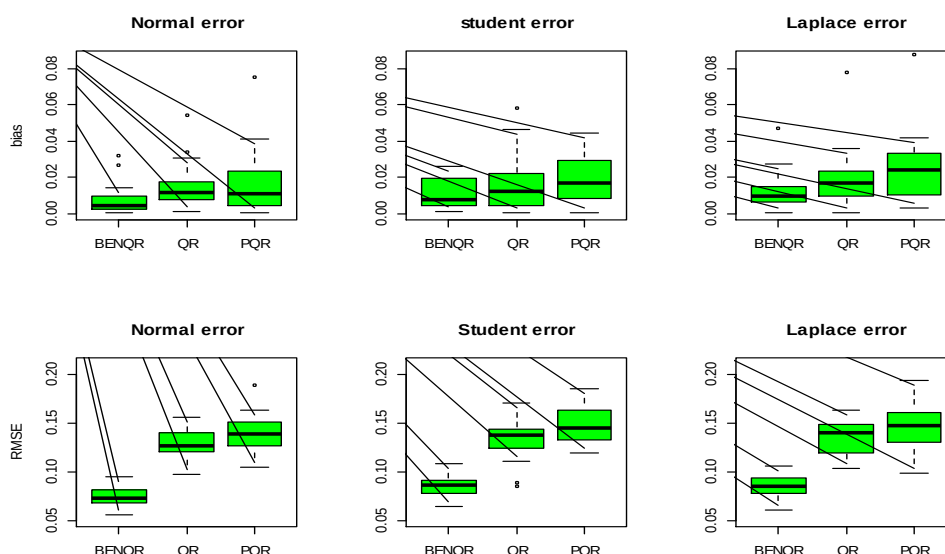
در این بخش به بررسی و تحلیل داده‌های واقعی می‌پردازیم. این داده‌ها شامل رشد اندازه قد ۵۰۰ کودک از بدو تولد تا دوسالگی است که از سال ۱۳۹۱ تا سال ۱۳۹۲ (به مدت دو سال) از درمان‌گاه‌های شهرستان زنجان به تصادف انتخاب و جمع‌آوری شده است. هدف بررسی تأثیر برخی عوامل بر میزان رشد قد کودکان است. این

عوامل یعنی متغیرهای تبیینی عبارت‌اند از: جنسیت کودک $x^{(۱)}$ ، فاصله 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 اندازه‌قد کودک از بدو تولد تا دو سالگی y است که در فاصله‌های زمانی بدو تولد، ۱، ۲، ۴، ۶، ۹، ۱۲، ۱۸ و ۲۴ ماهگی اندازه‌گیری شده است. بنابراین مدل با اثرات تصادفی برای این داده‌ها را بدین‌صورت در نظر می‌گیریم:

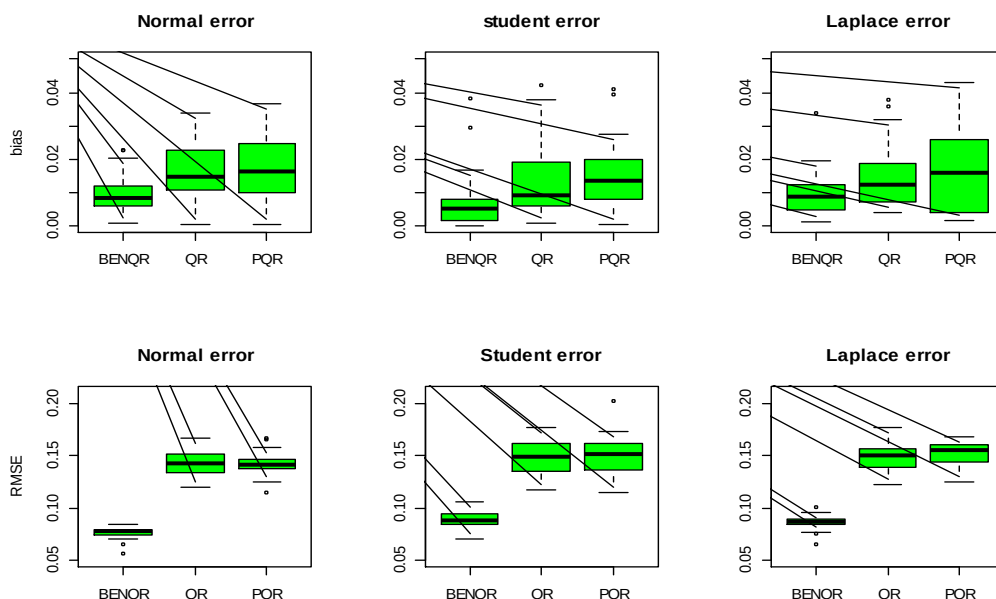
$$y_{ij} = x_{ij}^{(۱)}\beta_1 + x_{ij}^{(۲)}\beta_2 + \dots + x_{ij}^{(۱۲)}\beta_{12} + \alpha_i + \varepsilon_{ij},$$



شکل ۲. عملکرد روش‌ها به لحاظ معیار ارزیابی و RMSE در چندک ۰/۲۵ $\tau = 0.25$



شکل ۳. عملکرد روش‌ها به لحاظ معیار ارزیابی و RMSE در چندک ۰/۵ $\tau = 0.5$



شکل ۴. عملکرد روش‌ها به لحاظ معیار اریبی و RMSE در چندک $\tau = 0.75$

که در آن $i = 1, \dots, 50, j = 0, 1, 2, 4, 6, 9, 12, 18, 24$ است. در برازش مدل به روش BENQR توزیع مؤلفه‌های خطا ε_{ij} را لاپلاس نامتقارن، $ALD(\cdot, \sigma, \tau)$ 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 به عنوان دوره همگرایی مطلوب کنار می‌گذاریم. برای پارامترهای β و σ ، توزیع‌های پیشینی را ناآگاهی‌بخش به صورت $\beta \sim N_{\lambda}(\cdot, 100I)$ و $\sigma \sim IGamma(0 / 0.1 / 0)$ در نظر می‌گیریم. جدول ۲ برآورد پارامترهای مدل و انحراف معیار آن‌ها را برای سه روش ذکر شده در چندک $\tau = 0.5$ نشان می‌دهد. ملاحظه می‌شود که هر سه روش ضرایب مدل را تقریباً یکسان برآورد کرده‌اند. اندازه انحراف معیار برآوردها در هر سه روش برای متغیرهای تبیینی $x^{(4)}$ و $x^{(11)}$ در مقایسه با اندازه 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^{۱۹} 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 ۲۴، 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^{۲۰} (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 مقایره 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 $\tau = 0.5$

پارامترها	BENQR		PQR		QR	
	برآورد	انحراف معیار	برآورد	انحراف معیار	برآورد	انحراف معیار
β_1	۰/۱۷۱۷	۰/۴۶۹۶	۰/۰۴۲۸	۰/۶۰۴۲	۰/۰۹۴۰	۰/۶۱۴۳
β_2	۰/۳۹۷۷	۰/۶۸۷۵	-۰/۴۳۵۹	۰/۸۷۹۰	۰/۴۰۹۳	۰/۹۲۳۲
β_3	۰/۸۵۰۱	۰/۷۴۱۸	۰/۱۰۹۵	۰/۸۵۲۲	۰/۸۵۰۲	۰/۹۷۱۲
β_4	-۰/۱۲۳۶	۰/۰۳۳۶	-۰/۱۳۶۶	۰/۰۴۲۵	-۰/۱۲۹۰	۰/۰۵۵۸
β_5	۰/۱۱۸۷	۰/۳۸۴۶	۰/۰۰۳۳	۰/۴۵۰۸	۰/۱۳۴۷	۰/۴۷۶۶
β_6	-۰/۴۵۱۲	۰/۴۴۰۱	-۰/۲۳۹۷	۰/۵۴۳۱	-۰/۴۹۷۸	۰/۵۷۹۱
β_7	۰/۰۱۱۵	۰/۰۲۷۱	-۰/۰۰۶۰	۰/۰۳۴۳	۰/۰۱۰۸	۰/۰۳۵۰
β_8	-۰/۰۰۱۶	۰/۰۳۵۰	۰/۰۰۷۴	۰/۰۵۶۶	-۰/۰۰۴۷	۰/۰۴۴۱
β_9	-۰/۰۶۱۷	۰/۱۳۰۳	-۰/۰۱۴۸	۰/۲۱۸۷	-۰/۰۷۹۲	۰/۱۶۵۲
β_{10}	-۰/۷۵۷۲	۱/۴۱۰۰	-۱/۳۰۹۸	۱/۴۱۱۰	-۰/۹۲۶۸	۱/۰۲۸۸
β_{11}	۰/۱۵۵۸	۰/۰۲۵۶	۰/۱۴۶۵	۰/۰۲۹۴	۰/۲۰۰۴	۰/۰۱۲۱
β_{12}	۰/۰۷۱۱	۰/۱۰۷۰	۰/۰۶۷۷	۰/۱۳۱۰	۰/۰۵۷۱۷	۰/۱۳۱۴

جدول ۳. میانگین، میانه و انحراف معیار MSE ها در چندک‌های 0.5 ، 0.7 ، 0.3 و 0.5

چندک	مدل	میانگین	میانه	انحراف معیار
$\tau = 0.5$	BENQR	۱۵۱/۶۵۷۹	۱۵۰/۸۷۵۴	۴/۱۸۰۰
	PQR	۱۵۲/۸۷۶۴	۱۵۴/۵۰۸۴	۵/۶۷۵۵
	QR	۱۵۲/۵۲۸۵	۱۵۴/۶۵۴۸	۵/۶۳۵۳
$\tau = 0.3$	BENQR	۱۵۴/۲۳۴۶	۱۵۳/۷۴۵۹	۶/۲۰۰۱
	PQR	۱۵۶/۹۴۳۸	۱۵۵/۸۳۷۳	۸/۰۴۸۶
	QR	۱۵۵/۴۸۶۷	۱۵۵/۹۳۲۸	۷/۹۳۴۷
$\tau = 0.7$	BENQR	۱۶۱/۱۲۸۷	۱۵۸/۸۴۳۶	۴/۵۶۴۰
	PQR	۱۶۵/۴۳۲۶	۱۶۴/۹۴۳۷	۶/۸۳۴۵
	QR	۱۶۶/۵۶۳۴	۱۶۳/۸۴۳۹	۷/۵۶۴۸

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 روش، 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 به سمت 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 روش، 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 همه پارامترهای مدل با روش نمونه‌گیری گیبس است. نتایج حاصل از شبیه‌سازی نشان داد که روش بیزی ارائه شده در تمامی چندک‌ها نسبت به روش‌های 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](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.