Journal of Statistical Sciences, Autumn and Winter, 2023
Vol. 17, No. 2, pp 389-405
DOI: 10.52547/jss.17.2.09

Double Penalized Mixed Effects Quantile Regression Modeling Using the Maximum Likelihood Approach

Jafari Maryaki, F., Golalizadeh, M.
Department of Statistics, Tarbiat Modares University, Tehran , Iran.

Corresponding author: M. Golalizadeh, golalizadeh@modares.ac.ir

Received: 31/10/2022 **Revised:** 16/11/2023 **Accepted and Published Online:** 18/11/2023.

Introduction

Mixed effects is one of the powerful statistical approaches used to model the relationship between the response variable and some predictors in analyzing data with a hierarchical structure. Estimating parameters in these models is often done via either the least squares error or maximum likelihood approaches. The estimated parameters obtained through either of these approaches could be more efficient if the error distributions are non-normal. In such cases, using the mixed effects quantile regression is preferred alternatively. Moreover, when the number of variables studied increases, the penalized mixed effects quantile regression is one of the best methods to gain prediction accuracy and the model's interpretability. We propose a new tool to model the hierarchical structural data by introducing the penalty, a function of random effects and fixed effects. This new idea can simultaneously estimate the parameters and predict random effects. We can then make further statistical inferences on the parameters using the likelihood function built based on modeling such as a new penalized mixed effects quantile regression model.

Material and Methods

In this paper, under the assumption of an asymmetric Laplace distribution for random effects, we proposed a double penalized model in which both the random and fixed effects are independently penalized. Then, invoking a standard algorithm to estimate the parameters using the likelihood approach constructs our subsequent step in modeling the data. Since the likelihood function had no tractable form, we considered the EM algorithm to approx-

imate the integrals. To do so, we also extended the previously proposed algorithm to derive the maximum likelihood estimate of the parameters.

Results and Discussion
The performance of the new method proposed in this paper is evaluated in the simulation studies, and a discussion of the results is presented along with a comparison with some competing models. Our simulation experiments show that in addition to the robustness of the presented method, it performs better compared to the types of penalty in the model based on different penalty parameters. Applying the proposed model in real data analysis also showed that the new idea works well compared with some standard tools. The exciting feature of the double-penalized model is on confining the computation procedure to estimate only one tune parameter because the other one is optimized, using the samples and updated components while implementing the algorithm. Our results showed that considering a lower value for the tuning parameter of penalizing the fixed effects is preferred. We cannot recommend choosing a specific distribution for the random effects while using our proposed model. However, we can assure that our model does the same task as other competing models do in this regard.

Conclusion
The model presented in this paper is a double penalized model, which is a function of random effects and model parameters. In this paper, we achieved the optimal shrinkage for the effects using the simulated and real data. Similar to some standard methods, the degree of shrinkage for the random effects in our model depends on the tuning parameter selection. We aim to study the efficiency of our proposed model while one prefers to follow a Bayesian approach. Also, correctly choosing the penalty parameters from a theoretical viewpoint is another proposal for future research.

Keywords: Mixed Effects Quantile Regression, Laplace Distribution, Penalty Function, Shrinkage Approach, High Dimensional.

Mathematics Subject Classification (2010): 62G08, 62J05.

مجله علوم آماری، پاییز و زمستان ۱۴۰۲
جلد ۱۷، شماره ۲، ص ۳۸۹ -- ۴۰۵
DOI: 10.52547/jss.17.2.09
مقاله پژوهشی

مدل‌بندی رگرسیونی چندکی آمیخته تاوانیده دوگانه از طریق رویکرد درستنمایی

فروزان جعفری مریکی، موسی گلعلی‌زاده

گروه آمار، دانشگاه تربیت مدرس

نویسنده مسئول : موسی گلعلی‌زاده، golalizadeh@modares.ac.ir

تاریخ دریافت: ۱۴۰۱/۸/۹ تاریخ بازنگری: ۱۴۰۲/۸/۲۵ تاریخ پذیرش و انتشار: ۱۴۰۲/۸/۲۷

چکیده: مدل اثرهای آمیخته از جمله ابزارهای قوی آماری است که برای مدل‌بندی ارتباط بین متغیر پاسخ و متغیرهای تبیینی در تحلیل داده‌هایی با ساختار سلسله مراتبی به‌کار می‌رود. برآوردگرهای به‌دست آمده در این مدل‌ها با استفاده از هر یک از روش‌های کمترین توان دوم خطاها و ماکسیمم درستنمایی، زمانی‌که توزیع خطاها غیر نرمال باشد، از کارایی لازم برخوردار نیستند. در این‌گونه مواقع می‌توان از مدل رگرسیون چندکی آمیخته به‌عنوان جایگزین استفاده کرد. زمانی‌که تعداد متغیرهای مورد بررسی در این نوع مدل‌بندی افزایش می‌یابد، رگرسیون چندکی آمیخته تاوانیده یکی از بهترین روش‌ها برای افزایش دقت پیشگویی و تفسیرپذیری مدل است. در این مقاله با در نظر گرفتن توزیع لاپلاس نامتقارن برای اثرهای تصادفی، یک مدل تاوانیده دوگانه به عنوان تابعی همزمان از اثرهای تصادفی و پارامترهای مدل پیشنهاد می‌شود. سپس، عملکرد روش پیشنهادی با استفاده از مطالعه شبیه‌سازی آماری مورد ارزیابی قرار گرفته و بحث راجع به نتایج حاصل به همراه مقایسه با برخی مدل‌های رقیب ارائه می‌شود. به‌علاوه، کاربستی از آن در تحلیل یک مثال واقعی نمایش داده خواهد شد.

واژه‌های کلیدی: رگرسیون چندکی آمیخته، توزیع لاپلاس، تابع تاوان، رویکرد انقباضی، بُعد بالا.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62G08. ، 62J05.

۱ مقدمه

مدل‌های رگرسیونی میانگین بیشترین کاربرد را برای ایجاد ارتباط بین میانگین متغیر پاسخ و هر تابعی از متغیرهای تبیینی دارند. اما در مواردی محقق علاقه‌مند است تحلیل داده‌ها را براساس مدل‌بندی معیارهای دیگری از توزیع متغیر پاسخ غیر از میانگین انجام دهد. به عنوان یک فعالیت علمی در راستای چنین نیازی، رگرسیون چندکی که حالت توسعه یافته‌یی از رگرسیون میانگین است، توسط کوئنکر و باست (۱۹۷۸) معرفی شد. در چند دهه اخیر، القای نقش اثرهای تصادفی در مدل‌های رگرسیون چندکی برای تحلیل داده‌های طولی مورد توجه آماردانان و محققین فعال در حوزه کاربرد آمار قرار گرفته است. به عنوان شاهد مثالی از این موضوع می‌توان به فعالیت گراسی و بوتای (۲۰۰۷) اشاره کرد. لازم به توضیح است که ماهیت داده‌های طولی به‌گونه‌ای است که بین مشاهدات هر آزمودنی در زمان‌های متفاوت یک تغییرپذیری درون آزمودنی رخ می‌دهد که اگر در مدل لحاظ نشود منجر برآوردهای اریب برای پارامترهای مدل می‌شود (دیگل و همکاران ، ۲۰۰۲).

وقتی تعداد متغیرهای مورد بررسی در مدل افزایش می‌یابد، روش‌های انتخاب متغیر، با تاوانیده کردن مدل مورد نظر، باعث افزایش دقت پیشگویی و تفسیرپذیری مدل می‌شوند. در این راستا روش‌های انقباضی متفاوتی مانند لاسو (تیبشیرانی ، ۱۹۹۶)، لاسوی سازوار (زو ، ۲۰۰۶)، انحراف مطلق برش صاف[1] $SCAD$ (فن و لی، ۲۰۰۱) و گاروت نامنفی[2] (بریمن ، ۱۹۹۵) گسترش یافت. لازم به اشاره است که با استفاده از این روش‌ها برآورد و انتخاب متغیر به‌طور همزمان انجام می‌شود. محققان زیادی از رویکردهای انقباضی برای تحلیل داده‌های طولی در رگرسیون چندکی استفاده کردند. منابع علمی نشان می‌دهند که کوئنکر (۲۰۰۴) برای اولین بار مدل رگرسیونی چندکی L_1 تاوانیده را برای تحلیل داده‌های طولی پیشنهاد داد به‌طوری‌که برای برآورد چندک‌های توزیع با اثرهای ویژه هر آزمودنی از روش‌های تاوانیده کردن اثرهای تصادفی در تابع زیان چندکی استفاده می‌شود. سپس، گراسی و بوتای (۲۰۰۷) به‌جای ارزیابی تابع زیان تاوانیده از رویکرد تابع درستنمایی مبتنی بر تابع چگالی لاپلاس نامتقارن بهره بردند. آنها، با درنظر گرفتن توزیع نرمال برای اثرهای تصادفی، رویکردی معادل با روش ارائه‌شده در کوئنکر (۲۰۰۴) معرفی کرده و نشان دادند روش آنها در مقایسه با روش اضافه کردن تابع تاوان به تابع زیان بهتر عمل می‌کند. جالب است که این رویکرد به‌صورت خودکار سطح بهینه‌ای از پارامتر تاوان را به‌دست می‌آورد. باندل و همکاران (۲۰۱۰) روشی برای انتخاب همزمان اثرهای تصادفی و ثابت با استفاده از رویکرد لاسوی سازوار برای مدل‌های رگرسیونی مبتنی بر میانگین پیشنهاد دادند. لی و همکاران (۲۰۲۰) با تلفیق روش ارائه شده در باندل و همکاران (۲۰۱۰) و کوئنکر (۲۰۰۴) الگوریتمی برای به دست آوردن همزمان برآوردهای اثرهای ثابت و تصادفی ارائه دادند. در واقع، آنها، بر اساس تاوان لاسو ، در مدل رگرسیونی چندکی آمیخته تاوان را هم برای اثرهای تصادفی و هم اثرهای ثابت پیشنهاد کردند. الگوریتم ارائه شده توسط آنها رویکردی غیر درستنمایی و بر اساس الگوریتم‌های معرفی شده در رگرسیون چندکی آمیخته تاوانیده لاسو است. به نظر می‌رسد با استفاده از رویکرد درستنمایی، در مدل رگرسیونی

[1] Smoothly Clipped Absolute Deviation
[2] Non-negative garotte

چندکی آمیخته تاوانیده که تاوان اعمال شده همزمان بر روی اثرهای تصادفی و اثرهای ثابت اعمال می‌شود رویکرد بهتری برای استنباط راجع به پارامترها فراهم شود. با چنین نگاهی، آقامحمدی و محمدی (۱۳۹۴) با کمک رویکرد بیزی و فراخوانی تاوان لاسو و لاسوی تطبیق پذیر به تحلیل داده‌های طولی توسط مدل رگرسیون چندکی پرداختند.

در مقاله حاضر با پیروی از رویکرد انتخاب متغیر معرفی شده توسط تیبشیرانی (۱۹۹۶)، با تلفیق مدل ارائه شده در کوئنکر (۲۰۰۴) و رویکرد درستنمایی بحث شده در گراسی و بوتای (۲۰۰۷)، رویکردی جدید در انتخاب متغیر در مدل رگرسیون چندکی آمیخته پیشنهاد می‌شود که می‌توان به‌طور همزمان پارامترها را برآورد و اثرهای تصادفی را پیشگویی کرد. این رویکرد طوری است که به طور همزمان علاوه بر انقباض پارامترها، به‌طور خودکار با استفاده از داده‌ها و تابع درستنمایی آنها به درجه‌ای از انقباض اثرهای تصادفی دست می‌یابد. برای بررسی عملکرد این مدل تلفیقی، آزمایش شبیه‌سازی برای مقایسه مدل‌های متنوع در حالت‌های متفاوت انجام و نتایج مورد بحث قرار می‌گیرد. تحلیل یک مجموعه داده واقعی برای نمایش کاربست بهتر مدل پیشنهادی نیز ارائه می‌شود.

به منظور ارائه مطالب مرتبط با موضوع این مقاله، در بخش ۲ به معرفی مدل رگرسیون چندکی آمیخته پرداخته می‌شود. در بخش ۳، پس از بررسی رگرسیون چندکی آمیخته تاوانیده، مدل پیشنهادی این مقاله نیز معرفی می‌شود. نتایج حاصل از انجام شبیه‌سازی و تحلیل نتایج مرتبط با بررسی یک مثال واقعی پایان بخش مقاله حاضر است.

۲ خلاصه‌ای از مدل رگرسیونی چندکی آمیخته

مدل رگرسیونی چندکی آمیخته معمولاً به‌صورت

$$y_{ij} = \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta}_\tau + u_i + \varepsilon_{ij\tau}, \qquad j = 1,\ldots,n_i,\ \ i = 1,\ldots,N$$

نوشته می‌شود به‌طوری‌که $\boldsymbol{x}_{ij}{}^T$ بردار p بعدی از سطرهای ماتریس طرح معلوم $\boldsymbol{X}_i$ و y_{ij} ، jامین مشاهده متغیر تصادفی پیوسته در آزمودنی iام، $\boldsymbol{\beta}_\tau$ برداری p بعدی از پارامترها و $\varepsilon_{ij\tau}$ خطای تصادفی مدل است که مستقل از هم و به‌علاوه مستقل از اثرهای تصادفی u_i درنظر گرفته می‌شوند. تابع چندکی شرطی خطی آمیخته به‌صورت

$$G_{y_{ij}|u_i}(\tau|\boldsymbol{x}_{ij},u_i) = \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta}_\tau + u_i, \qquad j = 1,\ldots,n_i,\ \ i = 1,\ldots,N \qquad (1)$$

تعریف می‌شود، که در آن $G_{y_{ij}|u_i}(\tau|\boldsymbol{x}_{ij},u_i) = F^{-1}_{y_{ij}|u_i}(\cdot)$ معکوس تابع توزیع متغیر پاسخ به شرط اثرهای تصادفی u_iها است و $0 < \tau < 1$. می‌توان انتظار داشت که مقدار برآورد به‌دست آمده از معادله (۱) به τ امین چندک توزیع y_{ij} وابسته است. اما، در ادامه برای سادگی مباحث، اندیس τ حذف می‌شود. گراسی و بوتای (۲۰۰۷)

پیشنهاد دادند که اگر اثرهای تصادفی معلوم باشند، برآورد $\boldsymbol{\beta} \in \mathbb{R}^p$ کمیتی است که عبارت

$$\tilde{V}_i(\boldsymbol{\beta}) = \sum_{j \in \{j:\tilde{y}_{ij} \geq \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta}\}}^{n_i} \tau \left|\tilde{y}_{ij} - \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta}\right| + \sum_{j \in \{j:\tilde{y}_{ij} < \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta}\}}^{n_i} (1-\tau) \left|\tilde{y}_{ij} - \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta}\right| \quad (۲)$$

را مینیم کند، که در آن $\tilde{y}_{ij} = y_{ij} - u_i$. آن‌ها نشان دادند که با در نظر گرفتن توزیع لاپلاس نامتقارن برای متغیر پاسخ، کمینه‌کننده رابطه (۲) از طریق رویکرد تابع درستنمایی به‌دست می‌آید. لازم به اشاره است که توزیع y_{ij} به شرط معلوم بودن u_iها، برای هر $i = 1, \ldots, N, \;\; j = 1, \ldots, n_i$ به‌طور مستقل توزیع لاپلاس نامتقارن است اگر

$$f\left(y_{ij}|\boldsymbol{\beta}, u_i, \sigma\right) = \frac{\tau(1-\tau)}{\sigma} \exp\{-\rho_\tau(\frac{y_{ij} - \mu_{ij}}{\sigma})\},$$

یا به صورت نمادین $y_{ij}|u_i \sim ALD(\mu_{ij}, \phi, \tau)$, که در آن $\mu_{ij} = \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} + u_i$ به عنوان پارامتر مکانی، معرف پیشگوی خطی امین چندک بوده، $\sigma \in (0, \infty)$ پارامتر مقیاس و $\tau \in (0, 1)$ پارامتر چولگی، $\rho_\tau(\nu) = \nu\left(\tau - I\left(\nu \leq 0\right)\right)$ تابع چک (به دلیل نمایش هندسی آن) و $I(\cdot)$ تابع نشانگر است. اگر بردار مشاهدات هر آزمودنی به‌صورت $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{in_i})$ نوشته شود، آنگاه به‌شرط معلوم بودن اثرهای تصادفی داریم $f\left(\boldsymbol{y}_i|u_i, \boldsymbol{\beta}, \sigma\right) = \prod_{j=1}^{n_i} f(y_{ij}|u_i, \boldsymbol{\beta}, \sigma)$. به دلیل نامعلوم بودن u_iها تابع چگالی توام $(\boldsymbol{y}_i, u_i)$ برای $i = 1, \ldots, N$ به‌صورت $f\left(\boldsymbol{y}_i, u_i|\eta\right) = f\left(\boldsymbol{y}_i|u_i, \boldsymbol{\beta}, \sigma\right) f\left(u_i|\varphi\right)$ نوشته می‌شود، که در آن $\eta = (\boldsymbol{\beta}, \sigma, \varphi)$ پارامترهای مدل بوده و $f\left(u_i|\varphi\right)$ تابع چگالی اثرهای تصادفی است. تابع چگالی توام $(\boldsymbol{y}, \boldsymbol{u})$ برای N آزمودنی به‌صورت $f\left(\boldsymbol{y}, \boldsymbol{u}|\eta\right) = \prod_{i=1}^{N} f\left(\boldsymbol{y}_i|u_i, \boldsymbol{\beta}, \sigma\right) f\left(u_i|\varphi\right)$ است که در آن $\boldsymbol{y} = (y_1, \ldots, y_N)$ و $\boldsymbol{u} = (u_1, \ldots, u_N)$. در نتیجه تابع چگالی حاشیه‌ای $\boldsymbol{y}$ با انتگرال‌گیری نسبت به اثرهای تصادفی به صورت $f\left(\boldsymbol{y}|\eta\right) = \int_{R^N} f\left(\boldsymbol{y}, \boldsymbol{u}|\eta\right) d\boldsymbol{u}$ بدست می‌آید، که منجر به تابع درستنمایی به صورت $L(\eta; \boldsymbol{y}) = \sum_{i=1}^{N} L(\eta; \boldsymbol{y}_i)$ می‌شود. این تابع مبنای مناسبی برای انجام استنباط آماری راجع به پارامترهای مدل است که در بخش بعد به آن پرداخته می‌شود.

۳ برآورد درستنمایی مدل رگرسیونی چندکی آمیخته تاوانیده

رویکرد رگرسیون چندکی تاوانیده توسط کوئنکر (۲۰۰۴) با اضافه کردن تابع تاوان لاسو به اثرهای نصادفی مدل به‌صورت

$$\min_{u,\boldsymbol{\beta}} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \omega_{ij} \rho_\tau\left(y_{ij} - \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} - u_i\right) + \lambda \sum_{i=1}^{N} |u_i|$$

معرفی شد، که در آن ω_{ij} وزنی است که برای کم کردن اثر داده‌های دورافتاده و هم‌چنین چندک τام در پیشگویی u_i استفاده می‌شود تا برآورد حاصل به استواری[1] برسد و پارامتر λ به‌عنوان پارامتر تاوان درنظر گرفته می‌شود. همانطور که ملاحظه می‌شود در این رویکرد تاوان به‌کار برده شده تنها تابعی از اثرهای تصادفی است و انقباضی روی پارامتر اثرهای ثابت، یعنی β صورت نمی‌گیرد. به علاوه، می‌توان ملاحظه کرد که رویکرد پیشنهادی کوئنکر (۲۰۰۴) توانایی مدل‌بندی داده های بُعد بالا را ندارد چرا که مدل ایشان کوششی در انتخاب متغیر ندارد. بنابراین، پیشنهاد مقاله حاضر آن است که مدنظر قرار دادن عبارت

$$\min_{u,\boldsymbol{\beta}} \sum_{i=1}^{N}\sum_{j=1}^{n_i}\omega_{ij}\rho_\tau\left(y_{ij}-\boldsymbol{x}_{ij}{}^T\boldsymbol{\beta}-u_i\right)+\lambda_1\sum_{i=1}^{N}|u_i|+\lambda_2\sum_{l=1}^{p}|\beta_l| \qquad (۳)$$

که در آن $\lambda_i \geq 0$ها، به ازای $i = 1, 2$ پارامترهای تاوان هستند، رویکرد مناسبی برای مدل بندی داده های بُعد بالا از طریق مدل رگرسیون چندکی آمیخته است. با قبول چنین پیشنهادی، هدف اولیه ما این است که استنباط در مورد پارامترهای مدل را براساس تابع درستنمایی مرتبط با این رویکرد انجام دهیم. می توان انتظار داشت که تابع درستنمایی حاصل شامل انتگرال یا انتگرال‌هایی است که شکل صریحی برای حل آن‌ها وجود ندارد. نکته امیدوار کننده این است که، معمولاً، برای تقریب اینگونه انتگرال(ها) روش الگوریتم مونت کارلوی EM (بوث و هابرت ، ۱۹۹۹) مورد استفاده قرار می‌گیرد. به علاوه، توزیع اثرهای تصادفی نقش مهم و اساسی را در برآورد درستنمایی پارامترهای مدل بازی می‌کند. بنابراین در ادامه به بررسی نقش توزیع‌های متفاوت برای اثرهای تصادفی پرداخته می‌شود.

اگر اثرهای تصادفی دارای توزیع لاپلاس نامتقارن با پارامتر مکان صفر، پارامتر مقیاس ϕ و پارامتر چولگی ۰/۵ نمایش داده شده با نماد اختصاری $u_i \sim ALD(0, \phi, 0.5)$ به ازای $i = 1, \ldots, N$ باشد. آنگاه، تابع چگالی توام $(\boldsymbol{y}_i, u_i)$ برای iامین آزمودنی به‌صورت

$$\begin{aligned} f\left(\boldsymbol{y}_i, u_i|\eta\right) &= f\left(u_i|\varphi\right)\prod_{j=1}^{n_i} f(y_{ij}|u_i,\boldsymbol{\beta},\sigma) \\ &= \left\{\frac{\tau(1-\tau)}{\sigma}\right\}^{n_i}\frac{1}{4\varphi}\exp\left\{-\sum_{j=1}^{n_i}\left\{-\rho_\tau(\frac{y_{ij}-\mu_{ij}}{\sigma})\right\}-\frac{|u_i|}{2\varphi}\right\} \end{aligned} \qquad (۴)$$

به دست می‌آید. با توجه به این که توزیع شرطی $f\left(u_i\,|\boldsymbol{y}_i,\eta\right) \propto f\left(\boldsymbol{y}_i, u_i|\eta\right)$ تابعی لگ- مقعر (گیلکز ، ۱۹۹۵) است، می‌توان از از طریق روش نمونه‌گیری گیبز یا الگوریتم نمونه‌گیری ردی تطبیقی برای تولید نمونه‌های تصادفی به عنوان تحقق‌هایی از اثرهای تصادفی استفاده کرد. لذا، اگر فرض شود نمونه m_i تایی $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{im_i})$ رخدادهایی از اثرهای تصادفی تولید شده از تابع چگالی شرطی اثرهای تصادفی به شرط متغیر پاسخ باشند برآورد

[1] Robust

ماکسیمم درستنمایی ϕ به‌صورت

$$\hat{\phi} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{k=1}^{m_i}\rho_{0.5}\left(v_{ik}\right) \qquad (۵)$$

به دست می‌آید. توجه شود که تابع چگالی توام (۴) را می‌توان به صورت

$$f\left(\boldsymbol{y}_i, u_i|\eta\right) = \frac{\tau^{n_i}(1-\tau)^{n_i}}{\left(8\varphi^2\right)^{n_i}{\lambda_1}^{n_i}}\exp\left\{-\frac{1}{\sigma}\left[\sum_{j=1}^{n_i} c_{ij}\left|y_{ij}-\mu_{ij}\right| + \lambda_1\left|u_i\right|\right]\right\} \qquad (۶)$$

بازنویسی کرد، که در آن $c_{ij} = \left[(1-\tau)\, I\left(y_{ij} \le \mu_{ij}\right) + \tau I\left(y_{ij} > \mu_{ij}\right)\right]$ و $\lambda_1 = \frac{\sigma}{2\varphi}$. توجه شود که در این حالت همین مقدار λ_1 در رابطه (۳) هم استفاده می‌شود. پس از آن، برای به‌دست آوردن برآورد ماکسیمم درستنمایی پارامتر η برای τ امین چندک مدل، از الگوریتم معرفی شده در (گراسی و بوتای ، ۲۰۰۷) که به صورت زیر توسعه داده شد، استفاده می‌شود.

الگوریتم ۱. الگوریتم برآورد ماکسیمم درستنمایی پارامتر η برای τامین چندک:

گام ۱- ابتدا قرار دهید $t = 0$ و برای پارامترهای $\eta^{(t)} = \left(\boldsymbol{\beta}^{(t)}, \sigma^{(t)}, \varphi^{(t)}\right)$ مقادیر اولیه‌ای در نظر بگیرید و آن‌ها را در تابع چگالی $f\left(u_i\,|\boldsymbol{y}_i, \eta^{(t)}\right)$ جای‌گذاری کنید.

گام ۲- با استفاده از روش‌های نمونه‌گیر به‌طور مثال نمونه‌گیر گیبز *(گیلکز و همکاران ، ۱۹۹۵)* از تابع چگالی، $f\left(u_i\,|\boldsymbol{y}_i, \eta^{(t)}\right)$ برای هر آزمودنی $i = 1, \ldots, N$ به‌طور مستقل نمونه‌هایی تولید کنید و آن‌ها را در بردار ${\boldsymbol{v}_i}^{(t)} = \left({v_{i1}}^{(t)}, \ldots, {v_{im_i}}^{(t)}\right)$ قرار دهید.

گام ۳- عبارت $\sum_{i=1}^{N}\frac{1}{m_i}\sum_{k=1}^{m_i}\sum_{j=1}^{n_i}\rho_\tau\left({\tilde{y}^{(t)}}_{ikj} - {\boldsymbol{x}_{ij}}^T\boldsymbol{\beta}\right) + \lambda_2\sum_{l=1}^{p}|\beta_l|$ را نسبت به β مینیمم کنید، که در آن ${\tilde{y}^{(t)}}_{ikj} = y_{ij} - {v_{ik}}^{(t)}$ و $\sum_{l=1}^{p}|\beta_l|$ تاوان لاسو در تابع هدف است. توجه شود که بعد از انجام این عمل $\boldsymbol{\beta}^{(t+1)}$ها به‌دست می‌آیند. در نتیجه می‌توان بلافاصله واریانس خطاها را از طریق رابطه

$$\sigma^{(t+1)} = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N}\frac{1}{m_i}\sum_{k=1}^{m_i}\sum_{j=1}^{n_i}\rho_\tau\left({\tilde{y}^{(t)}}_{ikj} - {\boldsymbol{x}_{ij}}^T\boldsymbol{\beta}^{(t+1)}\right)$$

محاسبه کرد. در نهایت قرار دهید $\varphi^{(t+1)} = \hat{\phi}({\boldsymbol{v}_1}^{(t)}, \ldots, {\boldsymbol{v}_N}^{(t)})$، که در آن $\hat{\phi}$ برآورد ماکسیمم درستنمایی ϕ به شرط $\boldsymbol{y}$ است که از طریق عبارت (۵) قابل محاسبه است. شایان ذکر است که، در این مرحله، λ_1 در عبارت (۳) به صورت ${\lambda_1}^{(t+1)} = \frac{\sigma^{(t+1)}}{2\varphi^{(t+1)}}$ قابل محاسبه است.

گام ۴- حال قرار دهید $t = t + 1$ و گام ۱ تا ۳ را تکرار کنید.

گام ۵- الگوریتم را تا جایی تکرار کنید که برآوردهای $\eta^{(t)}$ به ازای دنباله متناهی از t به همگرایی قابل قبولی برسد.

در الگوریتم ذکر شده ، در هر تکرار $\eta^{(t+1)}$ برآورد ماکسیمم درستنمایی η به شرط $\eta^{(t)}$ است. مرحله نمونه‌گیری در اجرای این الگوریتم مستلزم محاسبه همگرایی زنجیره مارکوف است (کولز و کارلین ، ۱۹۹۶) و بنابراین برآورد به‌دست آمده با استفاده از این روش به تعداد نمونه‌ای که از تابع چگالی تولید می‌شود و تعداد دورریز بستگی دارد. ذکر این نکته ضروری است که فرآیند اجرای الگوریتم طوری است که شخص به‌طور خودکار به درجه بهینه‌ای از انقباض اثرهای تصادفی دست می‌یابد. بنابراین، برای اجرای این الگوریتم تنها لازم است پارامتر تاوان λ_2 از پیش تعیین شود. این بدان معنی است که مقدار بهینه پارامتر λ_1 هنگام اجرای الگوریتم خود به خود به دست می‌آید.

۴ آزمایش شبیه‌سازی و تحلیل نتایج آن

در این بخش قصد داریم کارکرد مدل پیشنهادی را از طریق آزمایش‌های شبیه‌سازی بررسی کنیم. واضح است که می‌توان رویکرد شبیه‌سازی پیشنهادی را برای چندک‌های متفاوت توزیع متغیر پاسخ بکار گرفت. اما، به دلیل محدودیت فضای مقاله، خود را محدود به چندک خاص از توزیع متغیر پاسخ خواهیم کرد. شایان ذکر است که در تمام موارد شبیه‌سازی، ستون‌های ماتریس طرح مرکزی شده‌اند و در نتیجه مدل نهائی شامل عرض از مبداء نیست. همچنین توجه شود که برای تولید نمونه تصادفی با استفاده از روش نمونه‌گیر گیبز در این بخش و بخش مثال کاربردی از کتابخانه $armspp$ در نرم افزار R و تابع $arms()$ استفاده شده است. برای ارزیابی مدل پیشنهاد شده در بخش قبل، آزمایش شبیه‌سازی با تعداد آزمودنی $N = ۱۰$ انجام می‌شود که هر یک دارای $n = ۲۳$ مشاهده هستند. همچنین فرض می‌کنیم $\tau = ۰/۲۵$. برای تولید داده از مدل

$$y_{ij} = \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} + u_i + \varepsilon_{ij} \quad i = ۱, \ldots, ۱۰, \quad j = ۱, \ldots, ۲۳ \tag{۷}$$

به قسمی که $\boldsymbol{\beta} = (۳, ۱/۵, ۰, ۰, ۲, ۰, ۰, ۰)^T$ و $\boldsymbol{x}_{ij}^T = (x_{ij۱}, \ldots, x_{ij۸})$ استفاده می‌شود. به علاوه، بردار $\boldsymbol{x}_{ij}$ها از توزیع نرمال ۸ متغیره استاندارد تولید شده‌اند طوری‌که همبستگی بین هر x_{ijl} و x_{ijk} به صورت

$$\mathrm{Corr}(x_{ij\ell}, x_{ijk}) = (۰/۵)^{|\ell - k|}, \quad ۱ \le \ell, k \le ۸$$

تعریف شده است. تعداد تکرار در شبیه‌سازی برای هر یک از حالت‌های مورد مطالعه را برابر ۱۰۰در نظر گرفتیم. در هر تکرار تعداد نمونه تولیدشده در نمونه گیر گیبز برابر $m_i = ۱۰۰۰۰$ به ازای $i = ۱, \ldots, ۱۰$ اختیار شد و سپس تعداد گام ۳ و تعداد دورریز در هر تکرار برابر با ۲۵۰۰ مدنظر گرفته شده است. لازم به اشاره است که نحوه انتخاب توزیع آماری برای ε_{ij}ها و u_iها در ادامه بحث تشریح می‌شوند. برای مقایسه آماری مناسب مدل‌های رقیب از نرم

$L_۱$ و $L_۲$ استفاده شده است که به ترتیب با پانویس ۱و ۲ برای ارزیابی بردار شیب به‌صورت

$$\|\hat{\beta}-\beta\|_۱=\sqrt{\sum_{\ell=۱}^{p}|\hat{\beta}_\ell-\beta_\ell|},\quad \|\hat{\beta}-\beta\|_۲=\sqrt{\sum_{\ell=۱}^{p}(\hat{\beta}_\ell-\beta_\ell)^۲}$$

تعریف می‌شوند، که در آن $\hat{\beta}_l$ برآورد چندکی ضریب رگرسیونی و β_l مقدار واقعی آن پارامتر متناظر است. ذکر این نکته ضروری است که با توجه به روش ارائه شده در این مقاله، پارامتر تاوان برای اثرهای تصادفی ($\lambda_۱$) در الگوریتم معرفی شده به‌صورت بهینه انتخاب می‌شود. به عبارتی دیگر، این پارامتر تاوان به‌طور خودکار از طریق نمونه‌های تولید شده و تابع درستنمایی متناظر با آن نمونه‌ها به‌دست می‌آید و نیازی به اعلام مقدار آن قبل از شروع شبیه‌سازی نیست. لذا، باتوجه به روش ارائه شده در این مقاله کافیست برای تولید نمونه‌ها پارامتر تاوان برای ضرایب رگرسیونی ($\lambda_۲$) از پیش تعیین شود. انتظار می‌رود پس از مدل‌بندی داده‌ها با مدل‌های مورد مطالعه و با استفاده از معیارهای موجود مقدار بهینه آن توسط روش بهینه‌سازی مورد محاسبه قرار گیرد. برای تولید نمونه‌های شبیه‌سازی، پنج مقدار متفاوت ۰٫۰۱, ۰٫۰۲, ۰٫۰۴, ۰٫۰۵, ۰٫۱۰ برای پارامتر تاوان $\lambda_۲$ در نظر گرفته شده است. تجربه محاسباتی ما نشان داد که برای مقادیر بزرگتر، به طور مثال $\lambda_۲=(۰٫۵, ۱, ۵)$ تمام ضرایب معادله (۷) برابر صفر برآورد می‌شود که مطلوب ما در انجام شبیه‌سازی نیست. در نهایت، برای تکمیل شبیه‌سازی، توزیع اثرهای تصادفی به صورت $ALD(۰, \phi, ۰٫۵)$ اختیار شده است. نتایج حاصل از آزمایش شبیه‌سازی در این حالت در جدول ۱ آمده است که در ادامه به تشریح نتایج پرداخته خواهد شد. طبق جدول ۱ شبیه‌سازی برای سه توزیع تصادفی برای خطای مدل به صورت تاوان دوگانه رگرسیون چندکی آمیخته[1] ($DLQR$)، رگرسیون ضرایب رگرسیونی تاوانیده[2] ($PFQR$)، رگرسیون اثرهای تصادفی تاوانیده[3] ($PRQR$) و رگرسیونی چندکی آمیخته معمولی[4] (QR) انجام شده است.

با توجه به نتایج به دست آمده ملاحظه می‌شود که انتخاب توزیع خطاها در عملکرد نهایی هر یک از روش‌های به کاربرده شده تاثیرگذار است. برای پیگیری سرراست نتایج، ابتدا به ازای هر روش مورد استفاده، مدل با $\lambda_۲$های متفاوت با هم و سپس مدل‌ها بر اساس توزیع‌های متفاوت درنظر گرفته شده با هم مقایسه می‌شوند. می‌توان ملاحظه کرد، زمانی که $\lambda_۲$ برابر با ۰/۰۲ و ۰/۰۴ باشد، وقتی که توزیع خطا نرمال و $t(۴)$ اختیار شوند مدل پیشنهادی عملکرد مناسبی نداشته است. البته باید اشاره شود که اگر چه در تمامی تکرارهای مرتبط با شبیه‌سازی برآورد پارامترهایی که واقعاً صفر بودند دقیقا صفر شد اما میزان $L_۱$ و $L_۲$ به دست آمده در این دو حالت تفاوت چشم‌گیری نسبت به حالت $\lambda_۲$های دیگر داشتند. اتفاق چنین امری به این دلیل است که برآوردهای غیر صفر پارامترها تفاوت معنی‌داری با مقدار واقعی پارامترهای متناظر غیر صفر در شروع تولید نمونه‌های شبیه‌سازی داشتند. به عبارتی دیگر، مدل مورد اشاره توانسته است پارامترهایی که دقیقاً صفر بوده‌اند را به صورت دقیق صفر برآورد کند. اما به همان اندازه

[1] Double Lasso Penalized Quantile Regression
[2] Penalized Fixed Effect Quantile Regression
[3] Penalized Random Effect Quantile Regression
[4] General Linear Quantile Regression

برآوردهای پارامترهای غیرصفر را بسیار منقبض کرد، و می‌توان گفت مدل به حالت کم برازش رسید. باید اشاره شود که وقتی برای مولفه خطا توزیع $ALD(۰, ۱, ۰/۲۵)$ اختیار شد، تفاوت مشاهده شده خیلی محسوس نبود. طبق نتایج به دست آمده به نظر می‌رسد که برای $\lambda_۲$های متفاوت، نتایج مربوط به $\lambda_۲$=۰/۰۵ برای توزیع نرمال در مدل $DLQR$ عملکرد قابل قبولی داشته است. حال آن‌که، در صورت اختیار $t(۴)$ برای توزیع خطا، اگر شخص خود را محدود به $\lambda_۲$=۰/۰۱ کند نتایج حاصل نسبت به $\lambda_۲$های دیگر رضایت بخش‌تر است. نباید از این نکته غافل شد که هنگام برازش مدل $PFQR$، انتخاب متغیر بدون اعمال تاوان روی اثرهای تصادفی صورت می‌گیرد. این در حالی است که مدل $PRQR$ تاوان را فقط روی اثرهای تصادفی اعمال می‌کند و مدل QR چیزی جز برازش یک مدل رگرسیون چندکی آمیخته عادی نیست. اگر شخص خود را محدود به مدل $PFQR$ کند، به نظر می‌رسد اختیار $\lambda_۲$=۰/۰۴ منجر به عملکرد بهتری نسبت به بقیه $\lambda_۲$ها خواهد شد. ملاحظه می‌شود دو مدل $PRQR$ و QR هیچ انتخاب متغیری نداشتند. نکته حائز اهمیت این است که در حالت $PRQR$ که به درجه بهینه از پارامتر تاوان برای اثرهای تصادفی رسیدیم نتایج حاصل از اجرای آن زمانی که توزیع خطاها $t(۴)$ و $ALD(۰, ۱, ۰/۲۵)$ اختیار شود با نتایج حاصل از اجرای مدل QR تقریباً برابر است.

طبق نتایج به دست آمده در این بخش و در ادامه پژوهش مرتبط با موضوع این مقاله می‌توان امیدوار بود که هنگام برازش مدل $DLQR$ بتوان $\lambda_۲$ای به دست آورد که عملکرد بهتری نسبت به بقیه مدل‌های دیگر داشته باشد. این امر نیاز به بررسی عمیق‌تر (از هر دو دیدگاه نظری و کاربردی) دارد و در فعالیت‌های آتی دنبال خواهد شد.

۵ تحلیل مثال واقعی

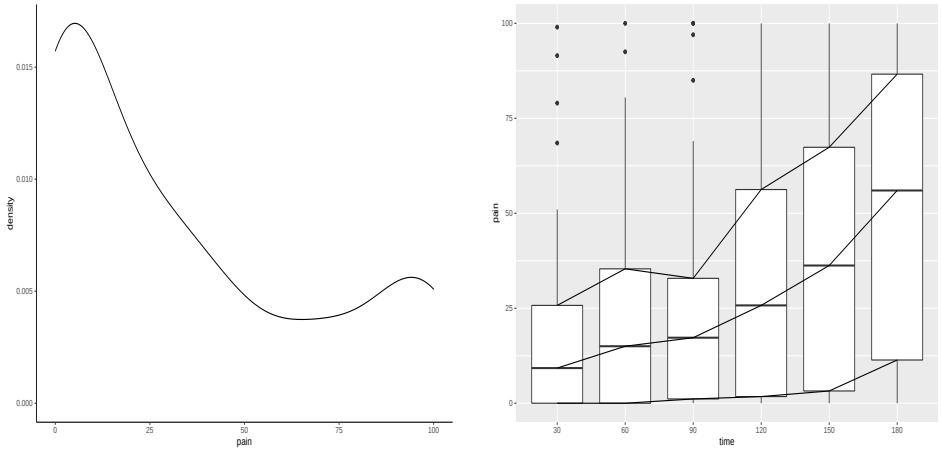
داده‌های آزمایش درد زایمان که توسط داویس (۱۹۹۱) گزارش و توسط جانگ (۱۹۹۶) تحلیل شده است، شامل گزارشات ثبت شده از میزان درد زایمان در زمان‌های متفاوت از آزمودنی‌های مورد بررسی است. تعداد آزمودنی مورد بررسی برابر با N=۳۴ زن در آزمایشگاه است. نحوه ثبت داده‌ها به‌گونه‌ای است که ۱۵ آزمودنی به تصادف انتخاب می‌شوند و در گروه بیماران مصرف کننده داروهای درد قرار می‌گیرند و مابقی از آنها دارونما مصرف می‌کنند. میزان درد برای هر آزمودنی، هر ۳۰ دقیقه یک بار اندازه گرفته می‌شود طوری که در نهایت ۶ بار میزان درد ثبت می‌شود. میزان درد (به عنوان متغیر پاسخ) اعدادی بین صفر تا ۱۰۰ اختیار می‌کند که صفر نمایانگر عدم وجود درد و ۱۰۰ نشان‌دهنده میزان درد شدید است. میزان چارک اول، میانه، میانگین و چارک سوم متغیر پاسخ برای گروه اول به ترتیب برابر با ۱۸/۱۲، ۴۲، ۵۰/۹۹، ۹۲/۳۸ و برای گروه دوم که دارونما مصرف کرده اند به ترتیب برابر با ۹/۵،۰، ۱۷/۵۹، ۲۶ است. نمودار جعبه‌ای و برآورد تجربی تابع چگالی برای کل داده‌ها صرف نظر از نوع داروی مصرفی در شکل ۱ رسم شده است. به علاوه، چارک‌های اول، دوم و سوم برای همه زمان‌ها در نمودار جعبه‌ای رسم شد تا نمایش مناسب از توزیع داده‌ها ارائه شود. مشابه توضیحات ارائه شده برای شکل ۱، در شکل ۲ نیز نمایش تغییرات متغیر پاسخ به تفکیک هر دو گروه رسم شده است. همانطور که ملاحظه می‌شود در دوره زمانی اندازه گرفته شده رفتار توزیع متغیر پاسخ (میزان درد) در زمان‌های اندازه گرفته شده تا حد زیادی نامتقارن است. به زبان علمی، در برآورد

جدول ۱. نتایج شبیه‌سازی براساس مدل‌های متفاوت و توزیع‌های مختلف خطا.

توزیع خطا	مدل	$\lambda_۲$	معیارهای ارزیابی مدل	
			$\|\|\hat{\beta}-\beta\|\|_۱$	$\|\|\hat{\beta}-\beta\|\|_۲$
$N(۰,۱)$	$DLQR$	۰/۰۱	۲/۹۹۵۶	۲/۳۴۵۵
		۰/۰۲	۷/۵۶۲۰	۱۴/۶۲۸۷
		۰/۰۴	۷/۵۹۸۷	۱۵/۲۸۶۲
		۰/۰۵	۰/۵۳۰۹	۰/۱۵۸۵
	$PRQR$	۰/۱	۱/۲۰۴۵	۰/۸۶۶۳
		–	۰/۷۶۶۲	۰/۳۸۷۱
	$PFQR$	۰/۰۱	۰/۷۴۳۳	۰/۳۱۲۰
		۰/۰۲	۱/۹۰۳۳	۱/۲۴۴۱
		۰/۰۴	۰/۵۱۱۱	۰/۱۱۶۳
		۰/۰۵	۰/۷۲۳۰	۰/۳۹۳۵
	QR	۰/۱	۱/۱۸۸۸	۰/۸۴۶۸
		–	۰/۷۶۴۹	۰/۳۸۵۳
$t(۴)$	$DLQR$	۰/۰۱	۰/۸۱۲۰	۰/۲۹۹۹
		۰/۰۲	۷/۱۹۱۲	۱۲/۸۴۲۸
		۰/۰۴	۷/۱۰۶۲	۱۴/۳۷۹۹
		۰/۰۵	۱/۰۲۰۳	۰/۵۹۹۰
	$PRQR$	۰/۱	۱/۵۳۵۷	۱/۳۶۴۲
		–	۱/۲۳۶۳	۰/۵۷۷۷
	$PFQR$	۰/۰۱	۰/۸۱۹۴	۰/۳۰۷۱
		۰/۰۲	۲/۶۲۴۹	۲/۲۲۱۹
		۰/۰۴	۰/۸۰۷۲	۰/۲۱۳۰
		۰/۰۵	۱/۰۰۳۸	۰/۵۸۳۹
	QR	۰/۱	۱/۵۰۵۳	۱/۳۰۹۹
		–	۱/۲۴۲۱	۰/۶۱۹۱
$ALD(۰,۱,۰/۲۵)$	$DLQR$	۰/۰۱	۰/۵۸۹۶	۰/۰۸۸۷
		۰/۰۲	۰/۵۹۲۶	۰/۰۸۶۲
		۰/۰۴	۰/۷۲۳۹	۰/۱۸۱۲
		۰/۰۵	۰/۸۹۹۱	۰/۲۷۸۲
	$PRQR$	۰/۱	۱/۴۵۰۲	۱/۲۳۸۱
		–	۰/۸۶۸۱	۰/۴۵۶۲
	$PFQR$	۰/۰۱	۰/۷۶۷۴	۰/۲۹۳۵
		۰/۰۲	۱/۳۰۱۰	۰/۴۹۰۸
		۰/۰۴	۰/۷۳۶۶	۰/۱۸۶۴
		۰/۰۵	۰/۹۵۵۷	۰/۵۳۷۶
	QR	۰/۱	۱/۴۵۹۳	۱/۲۵۲۰
		–	۰/۸۷۲۴	۰/۴۵۴۸

تجربی چگالی‌ها، چوله بودن توزیع میزان درد قابل ملاحظه است. چنین استنباطی از طریق بررسی چندک‌های متغیر پاسخ نیز قابل درک است. نمایش نمودار برای دو گروه نشان می‌دهند که رفتار مورد اشاره برای آزمودنی‌هایی که دارو مصرف کردند مشهودتر است و این یعنی برآورد تجربی چگالی توزیع متغیر پاسخ چوله است. برای داشتن معیار عددی برای میزان چولگی، مقدار عددی آن بدون تفکیک گروه‌ها محاسبه و عدد ۰/۸۳ به دست آمد که نشان‌دهنده چوله به راست بودن توزیع داده‌ها است. علاوه بر آن، میزان چولگی به تفکیک دو گروه نشان داد که این معیار برای گروه مصرف کننده دارونما ۰/۰۵ و برای مصرف دارو ۱/۴۸ شد. اعداد گزارش شده نشان‌دهنده این واقعیت هستند

که در گروه اول چولگی توزیع متغیر پاسخ مشاهده نشده است، اما در گروه دوم (مصرف کننده دارو) توزیع متغیر پاسخ به شدت چوله است. این موضوع نشان می‌دهد که رگرسیون چندکی گزینه مناسب‌تری نسبت به رگرسیون مبتنی بر میانگین برای نمایش تغییرات متغیر پاسخ براساس تغییرات ناشی از متغیرهای تبیینی است.



(ب) (الف)

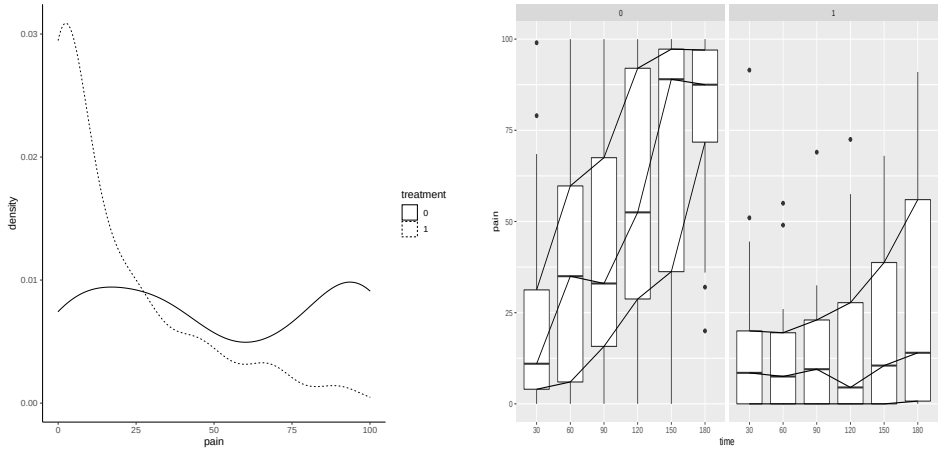
شکل ۱. الف- نمودار جعبه‌ای بمیزان درد آزمودنی ها به تفکیک زمان اندازه‌گیری و ب- برآورد تجربی چگالی میزان درد تمام مشاهدات بدون تفکیک گروه‌بندی.

با توجه به این که برای یک آزمودنی در زمان‌های متفاوتی متغیر پاسخ اندازه گرفته شده است، بنابراین در داده‌های مورد بررسی بین مشاهدات هر آزمودنی یک همبستگی ذاتی وجود دارد که با اضافه کردن اثرهای تصادفی به مدل این همبستگی لحاظ می‌شود. لذا، برای لحاظ این واقعیت، از مدل رگرسیونی پیشنهادی توسط گراسی و بوتای (۲۰۰۷) به صورت

$$G_{y_{ij}|u_i}(\tau|\boldsymbol{x}_{ij}, u_i) = \boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} + u_i, \qquad j = ۱,\ldots,n_i, \;\; i = ۱,\ldots,۳۴ \qquad (۸)$$

استفاده شد. با پیروی از ایشان، متغیرهای تبیینی و پاسخ و پارامترهای مدل در ساختاربندی مبتنی بر رابطه (۸) به صورت

$$\boldsymbol{X_i} = \begin{bmatrix} ۱ & R_i & T_{i۱} & R_iT_{i۱} \\ \vdots & \vdots & \vdots & \vdots \\ ۱ & R_i & T_{in_i} & R_iT_{in_i} \end{bmatrix}, \; \boldsymbol{\beta} = \beta_۰,\ldots,\beta_۳)^T, \; \boldsymbol{y_i} = (y_{i۱},\ldots,y_{in_i})^T$$



(ب) (الف)

شکل ۲. الف– نمودار جعبه‌ای و ب– برآورد تجربی چگالی میزان درد دو گروه آزمودنی دریافت کننده دارونما (۰) و دریافت کننده دارو (۱).

هستند، که در آن y_{ij} میزان درد iامین بیمار در زمان j است. توجه شود که R_i نشان‌دهنده مصرف دارو یا دارونما است. در پیاده‌سازی مدل طوری عمل شد که اگر آزمودنی دارو مصرف کند R_i مقدار ۰ و اگر دارونما مصرف کند مقدار ۱ را اختیار می‌کند. به علاوه، T_{ij} حاصل تقسیم زمان اندازه‌گیری متغیر پاسخ برای هر آزمودنی بر ۳۰ دقیقه است و n_i=۶ زمان آخرین اندازه‌گیری برای iامین آزمودنی است. بنا به دلایلی که در تحلیل اکتشافی (توزیع) داده‌ها مطرح شد یکی از روش‌های پیشنهادی برای تحلیل این داده‌ها مدل رگرسیون چندکی است. از طرفی دیگر، همانطور که در مقدمه تشریح شد برای برازش اولیه مدل رگرسیون چندکی نیازی به لحاظ فرض نرمال بودن توزیع خطاها نیست. برای این منظور، مدل رگرسیونی چندکی (۱) مد نظر قرار می‌گیرد.

برای تحلیل مدل (۱)، گراسی و بوتای (۲۰۰۷) توزیع نرمال را برای توزیع اثرهای تصادفی در نظر گرفتند. بنا به رویکرد مطرح شده در این مقاله، ما فرض می‌کنیم: $u_i \sim ALD(۰, \phi, ۰/۵)$. سپس از رویکردهای $DLQR$ و $PFQR$ برای تاوانیده‌کردن پارامترهای شیب استفاده می‌شود. علاوه بر آن، رویکرد $PRQR$ و QR نیز در برازش مدل (۸) دنبال خواهد شد. نتایج حاصل از برازش مدل‌های مورد اشاره در جدول ۲ آمده است. شایان ذکر است که معیار مناسب برای انتخاب مدل در اینجا $-۲\log L\left(\hat{\eta}\,|\boldsymbol{y},\boldsymbol{u}\right)$ است. طبق نتایج جدول ۲، گزارش برازش مدل برای سه چندک $\tau = (۰/۲۵, ۰/۵, ۰/۷۵)$ ارائه شده است. همانطور که ملاحظه می‌شود هنگام برازش مدل‌های $DLQR$ و $PFQR$ انتخاب متغیر صورت گرفته است و برای $\tau = (۰/۲۵, ۰/۷۵)$ طبق معیار در نظر گرفته شده به نظر می رسد $\lambda_۲$=۵ نتیجه بهتری نسبت به $\lambda_۲$های دیگر دارد. در این مدل‌ها برای این مقدار $\lambda_۲$ میزان $\hat{\beta}_۱$ صفر برآورد شده است و این یعنی مصرف دارو به تنهایی در میزان درد تاثیر ندارد. اما برای $\tau = ۰/۵$، مقدار $\hat{\beta}_۴$ برابر با صفر برآورد شده است که نشان می‌دهد اثر متقابل داروی مصرفی و زمان اندازه گیری در میزان درد اثری ندارد.

این نتیجه را می‌توان اینگونه تفسیر کرد که شخص بیمار دارو را در چه زمانی از ملاحظه درد مصرف کند تاثیر قابل ملاحظه‌ای در میزان کاهش درد او ندارد. درصورت برازش مدل $PRQR$، به نظر می‌رسد برای وقتی که $\tau = ۰/۲۵$ برآوردهای حاصل مناسب‌تر از برآوردهای به دست آمده با انتخاب چندک‌های دیگر باشد. در عوض، هنگام برازش مدل QR، نتایج مرتبط با $\tau = ۰/۷۵$ بهتر از چندک‌های دیگر است. اما نکته حائز اهمیت این است که خواه تاوان بر روی اثرات تصادفی اعمال شود خواه نشود، انتخاب متغیر در مدل‌های $PRQR$ و QR صورت نمی‌گیرد. در حالت کلی به نظر می‌رسد که داروی مصرفی روی میزان درد اثر معناداری ندارد. به عبارت دیگر بین دو گروه تفاوت معناداری بر اساس میزان درد وجود ندارد. اما زمان اندازه‌گیری میزان درد و اثر متقابل آن با دارو می‌تواند در میزان درد اثر معناداری داشته باشد.

جدول ۲. نتایج برازش مدل‌های متفاوت در تحلیل مثال واقعی.

			برآورد پارامتر				
مدل	τ	$\lambda_۲$	$-۲\log L(\hat{\eta}\|\boldsymbol{y},\boldsymbol{u}.)$	$\hat{\beta}_۰$	$\hat{\beta}_۱$	$\hat{\beta}_۲$	$\hat{\beta}_۳$
		۰/۰۱	۶۹۵۲۶/۴۷	۲/۴۹۷۲	−۰/۷۰۴۲	۰/۰۰۴۲	۰/۰۰۰۰
		۰/۵	۶۹۵۴۷/۹۵	۲/۴۱۵۴	۰/۰۰۰۵	−۰/۰۰۰۱	−۰/۰۰۳۰
	۰/۲۵	۱	۶۶۹۵۵/۵۳	۲/۴۱۵۰	۰/۰۰۰۰	۰/۰۰۰۰	۰/۰۰۰۰
		۵	۶۹۳۳۲/۳۵	۲/۳۳۷۶	۰/۰۰۰۰	−۰/۰۱۴۰	۰/۰۰۷۵
		۰/۰۱	۳۰۷۵۹/۹۸	۱/۹۶۳۵	۰/۰۰۰۱	۰/۰۲۶۹	−۰/۰۰۹۳
		۰/۵	۳۰۷۹۴/۶۶	۲/۰۳۲۸	−۰/۰۰۰۱	۰/۰۱۳۰	−۰/۰۰۰۱
$DLQR$	۰/۵	۱	۳۰۸۰۲/۹۸	۲/۰۶۶۶	−۰/۰۰۰۱	۰/۰۰۸۰	۰/۰۰۰۰
		۵	۳۰۷۹۲/۲۹	۲/۱۴۴۸	−۰/۰۱۵	−۰/۰۰۷۷	۰/۰۰۰۰
		۰/۰۱	۲۳۳۵۱/۲۴	−۶/۶۰۵۸	۰/۶۰۳۶	۶/۳۹۹۵	۰/۰۰۰۰
		۰/۵	۸۵۵۳۲/۹۳	−۵/۳۳۹۲	۰/۰۰۰۰	۶/۳۳۵۱	−۰/۱۶۶۷
	۰/۷۵	۱	۲۳۰۹۳/۰۱	−۵/۳۳۹۲	۰/۰۰۰۰	۶/۳۳۵۱	−۰/۱۶۶۷
		۵	۸۸۸۳۷/۸	−۵/۳۳۹۲	۰/۰۰۰۰	۶/۳۳۵۱	−۰/۱۶۶۷
		۰/۰۱	۷۲۵۲۹/۹۱	۲/۴۹۱۴	۰/۰۰۰۰	۰/۰۰۴۱	−۰/۹۸۷۴
		۰/۵	۷۰۱۰۹/۶۶	۲/۳۳۵۶	۰/۰۰۰۰	۰/۰۰۱۵	−۰/۰۴۲۴
	۰/۲۵	۱	۷۰۲۰۵/۸۸	۲/۳۹۸۷	۰/۰۰۰۰	۰/۰۰۷۸	۰/۳۶۳۷
		۵	۷۰۰۲۲/۰۶	۲/۳۲۸۹	۰/۰۰۰۰	۰/۰۰۰۵	−۰/۰۲۰۵
		۰/۰۱	۳۱۴۰۹/۲۳	۱/۹۵۷۴	۰/۰۰۰۱	۰/۰۱۸۸	۰/۱۰۹۴
		۰/۵	۳۱۴۱۶/۸۹	۲/۰۲۴۴	−۰/۱۷۳۶	۰/۰۱۴۴	۰/۰۰۰۰
$PFQR$	۰/۵	۱	۳۱۴۲۶/۲۱	۲/۰۴۷۰	۰/۰۰۰۰	۰/۰۰۵۱	۰/۰۰۰۰
		۵	۳۱۴۲۸/۴۸	۲/۱۱۶۲	−۰/۳۹۵۹	۰/۰۰۰۰	۰/۰۰۰۰
		۰/۰۱	۲۴۱۲۴/۶۰	−۶/۶۱۳۴	−۰/۲۳۷۴	۶/۳۹۸۳۱	۰/۰۰۰۰
		۰/۵	۲۴۰۶۴/۸۵	−۵/۵۶۹۸	۰/۰۰۰۰	۶/۳۸۶۴۵	−۰/۰۴۱۳
	۰/۷۵	۱	۲۴۰۴۲/۰۱	−۶/۶۶۱۰	۰/۰۰۰۰	۶/۳۵۷۲۳	۰/۰۰۰۰
		۵	۱۸۱۶۹/۷۰	۱/۸۷۶۹	−۰/۰۶۸۷	۰/۰۰۰۱۴	۰/۰۰۰۲
	۰/۲۵	−	۱۶۷۲۵۳/۷۴	۲/۳۳۱۸	−۰/۹۳۰۵	۰/۰۰۳۷۹	−۰/۸۳۹۵
$PRQR$	۰/۵	−	۳۱۲۵۵/۵۷	۱/۹۸۳۰	۱/۸۹۳۷	۰/۰۱۹۳۴	−۰/۵۲۵۵
	۰/۷۵	−	۳۵۵۷۴/۹۲	۱/۹۸۳۰	۱/۸۹۳۷	۰/۰۱۹۳۴	−۰/۵۲۵۵
	۰/۲۵	−	۷۳۱۳۰/۸۵	۲/۴۹۶۰	−۰/۲۷۴۷	۰/۰۰۴۵	−۱/۱۹۹۰
QR	۰/۵	−	۳۲۱۲۷/۸۰	۱/۹۶۱۹	۲/۴۰۲۹	۰/۰۲۰۰	−۰/۷۵۱۰
	۰/۷۵	−	۲۶۰۷۶/۴۹	−۶/۶۱۳۳	−۲/۹۱۱۵	۶/۴۰۰۱	۲/۵۰۹۹

بحث و نتیجه‌گیری

مدل ارائه شده در این مقاله یک مدل تاوانیده دوگانه است که تاوان علاوه بر اینکه تابعی از اثرهای تصادفی است، تابعی از پارامترهای مدل هم هست. در مقایسه با روش ارائه شده در کوئنکر (۲۰۰۴) که درجه انقباض اثرها وابسته به انتخاب پارامتر تاوان بود در این مقاله به درجه بهینه‌ای از انقباض اثرها با استفاده از داده‌های شبیه‌سازی شده و داده‌های واقعی دست یافتیم که در نوع خود پژوهش جدیدی در حوزه مورد مطالعه مرتبط با مدل بندی رگرسیونی چندکی آمیخته است. آنچه در ادامه پژوهش مطالعه خواهد شد بررسی نحوه انتخاب پارامتر تاوان مناسب از هر دو دیدگاه نظری و کاربردی است.

تقدیر و تشکر

نویسندگان مقاله از پیشنهادات و نظرات داوران و ویراستار محترم مجله که در بهبود کیفیت مقاله بسیار مؤثر واقع شد، تقدیر و تشکر می‌نمایند.

مراجع

آقامحمدی، ع. و محمدی، س. (۱۳۹۴)، رگرسیون چندکی بیزی با تاوان لاسو و لاسوی تطبیق‌پذیر برای داده‌های طولی دودویی. مجله علوم آماری، ۹، ۱۶۷-۱۴۹.

Booth, J. G. and Hobert, J. P. (1999), Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm. *Journal of the Royal Statistical Society*, **61**, 265–285.

Bondell, H. D, Krishna, A. and Ghosh, S. K. (2010), Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effect Models, *Biometrics*, **66**, 1069–1077.

Breiman, L. (1995), Better Subset Selection Using Nonnegative Garrote, *Technonometrics*, **37**, 373–384.

Cowles, M. K. , and Carlin, B. P. (1996), Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, *Journal of the American Statistical Association*, **91**, 883–904.

Davis, C. S. (1991), Semi-Parametric and Non-Parametric Methods for the Analysis of Repeated Measurements with Applications to Clinical Trials., *Statistics in Medicine*,**10**, 1959–1980.

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2002), *Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press.

Fan, J. and Li, R. (2001), Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Geraci, M. , and Bottai, M. (2007), Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution. *Biostatistics*, **8**, 140-154.

Gilks, W. R. (1995), Derivative-Free Adaptive Rejection Sampling for Gibbs Sampling. *Bayesian Statistics*, **4**, 641–649.

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995), Adaptive Rejection Metropolis Sampling within Gibbs Sampling, *Applied Statistics*, **44**, 455-472.

JUNG, S. (1996), Quasi-Likelihood for Median Regression Models, *Journal of the American Statistical Association*, **91**, 251–257.

Koenker, R., and Bassett, G. (1978), Regression Quantiles. *Econometrica*, **46**, 33-50.

Koenker, R. (2004), Quantile Regression for Longitudinal Data. *Journal of Multivariate Analysis*, **91**, 74-89.

Li, H., Liu, Y. and Luo, Y. (2020), Double Penalized Quantile Regression for the Linear Mixed Effects Model, *Journal of Systems Science and Complexity*, **33**, 2080–2102.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, **58**, 267–88.

Zou, H. (2006), The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association*, **101**, 1418–1429.