Journal of Statistical Sciences, Autumn and Winter, 2024
Vol. 18, No. 2, pp 457-479
DOI: 10.61186/jss.18.2.11

Piecewise Regression Model Based on Scale Mixture Normal Distribution

Hashemi, F.
Department of Statistics, University of Kashan, Kashan, Iran.

Corresponding author: F. Hashemi, farzane.hashemi1367@kashanu.ac.ir

Received: 23/5/2023 **Revised:** 18/10/2024 **Accepted and Published Online:** 18/10/2024.

Introduction

In many applied regression problems, the relationship between the response variable and the independent variables often changes suddenly in some places, called breakpoints or change points. Sudden changes in these places cause deviations in the analysis, and Data is analyzed. Although non-linear regression models can be used to examine such data, the importance of breakpoints needs to be addressed, and parameters may be estimated with bias. Hence, piecewise regression models show the importance of breakpoints and appropriate estimation of model parameters. In particular, identifying breakpoints is critical to understanding when and how to change the model. This article obtains breakpoints simultaneously by classifying the data into homogeneous clusters. Therefore, clustering methods can be used. Hybrid models for data clustering are powerful and popular models that estimate parameters and detect breakpoints using the Expectation-Maximum (EM) algorithm and its generalizations. Most of these methods focus on the average and variance changes in the range of two consecutive breakpoints.

Material and Methods

One of the valuable and practical tools to obtain MLE model parameters when the form of the density function is complex is the EM algorithm. This method, introduced by Dempster et al. (1977), is used when the data is only partially observed. The incompleteness of the data may be due to various reasons, such as missing observations or truncated or censored observations. The EM algorithm's convergence speed to obtain the optimal solution may be low compared to other methods. Nevertheless, the simplicity of imple-

mentation and obtaining trusted optimal answers can be the main reason for this method. We will use the EM algorithm to estimate the model parameters, i.e., the vector. To implement this algorithm, it is assumed that there are also hidden observations for the observed data.

Results and Discussion
According to the observations in the tables, the average values of TPR and FPR in the CN-BP model are respectively higher and lower than in the t-BP model. Also, the value of TPR decreases in both models as the sample size increases. With the increase in the sample size, the value of FPR increases in both models. Therefore, the CN-BP model is better than the t-BP model in detecting outliers.

Conclusion
In this article, we proposed a method to solve piecewise regression problems with errors from scale mixture normal (SMN) distribution by using finite mixed models. Finding change points is similar to categorizing similar observations into groups in mixed models. We created regression lines to estimate change points. Also, using the shown simulations, outlier locations are identified using some special modes of the SMN model. In addition, this numerical study, with actual and simulated data sets, shows the performance of the superior model proposed. As a result, the proposed SMN breakpoint (SMN-BP) model usually gives accurate estimates of change points and regression lines. Therefore, the SMN-BP model is suitable for solving piecewise regression problems. Piecemeal regression can be applied to problems with an important change point, such as quality control, medicine, economics, and decision-making decision-making.

Keywords: Break point, EM algorithm, Piecewise regression.

Mathematics Subject Classification (2010): 62J02, 62J05.

مجله علوم آماری، پاییز و زمستان ۱۴۰۳
جلد ۱۸، شماره ۲، ص ۴۵۷ -- ۴۷۹
DOI: 10.61186/jss.18.2.11
مقاله پژوهشی

مدل بندی رگرسیون تکه‌ای همراه با خطاهای مخلوط مقیاسی نرمال

فرزانه هاشمی

گروه آمار، دانشگاه کاشان

نویسنده مسئول : فرزانه هاشمی، farzane.hashemi1367@yahoo.com

تاریخ دریافت: ۱۴۰۲/۳/۲ تاریخ بازنگری: ۱۴۰۳/۷/۲۶ تاریخ پذیرش و انتشار: ۱۴۰۳/۷/۲۷

چکیده: یکی از پرکاربردترین مباحث آماری، مسایل رگرسیونی است. در مسایل رگرسیونی فرض اساسی بر روی خطاها، نرمال بودن آنهاست که این فرض در برخی موارد به سبب وجود ویژگی‌های عدم تقارن یا مکان‌های شکست در داده‌ها برقرار نمی‌باشد. مدل رگرسیون تکه‌ای یکی از راه‌های برون رفت در شرایط نرمال نبودن خطاهاست که به‌طور گسترده در حوزه‌های مختلفی به کار گرفته شده‌اند، که در آن‌ها تشخیص نقطه شکست مهم است و مکان‌های شکست در مدل‌های رگرسیون تکه‌ای برای دانستن زمان و چگونگی تغییر الگوی ساختار داده ضروری است. یکی از مشکلات عمده در این داده‌ها وجود دم سنگینی است که با استفاده از برخی توزیع‌ها که به عنوان تعمیمی از توزیع نرمال هستند این مشکل برطرف شده است. در این مقاله بر اساس توزیع مخلوط مقیاسی نرمال، مدل رگرسیونی تکه‌ای مورد بررسی قرار خواهد گرفت که می‌توان به جای نرمال با به کار گیری تعمیم‌هایی از توزیع نرمال این مشکل را برطرف نمود. همچنین این مدل با مدل رگرسیون تکه‌ای استاندارد که برگرفته از خطاهای نرمال است مورد مقایسه قرار خواهد گرفت.

واژه‌های کلیدی: الگوریتم EM، توزیع مخلوط مقیاسی نرمال، رگرسیون تکه‌ای، نقطه شکست.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62J02، 62J05.

۱ مقدمه

در بسیاری از مسایل کاربردی رگرسیون، رابطه بین متغیر پاسخ و متغیرهای مستقل اغلب به طور ناگهانی در برخی مکان‌ها تغییر می‌کند که به آن‌ها نقاط شکست یا نقاط تغییر[1] می‌گویند. تغییرات ناگهانی که در این مکان‌ها اتفاق می‌افتد باعث وجود انحراف در تجزیه و تحلیل داده‌ها می‌شود. اگرچه می‌توان از الگوهای رگرسیون غیرخطی برای بررسی این گونه داده‌ها استفاده کرد، باز هم اهمیت نقاط شکست اغلب نادیده گرفته می‌شود و امکان دارد پارامترها با اریبی تخمین زده شوند. از این رو، مدل‌های رگرسیون تکه‌ای برای نشان دادن اهمیت نقاط شکست و برآورد مناسب پارامترهای مدل مورد استفاده قرار می‌گیرند. به طور خاص، شناسایی نقاط شکست برای درک زمان و چگونگی تغییر مدل بسیار مهم هستند.

در مقالات، مدل‌های رگرسیونی که شامل نقاط شکست هستند به رگرسیون تکه‌ای[2] موسوم‌اند. استفاده از رگرسیون تکه‌ای به جای رگرسیون غیر خطی بسیار رایج است زیرا رگرسیون تکه‌ای به طور گسترده در زمینه‌های مختلف مانند پزشکی، اقتصاد سنجی و هواشناسی به کار گرفته شده‌اند. به عنوان مثال در تحقیقات پزشکی، خطر تولد نوزاد مبتلا به سندرم داون برای مادران جوان کم است، اما این خطر به طور قابل توجهی با افزایش سن مادر پس از یک آستانه سنی خاص افزایش می‌یابد (موجئو ، ۲۰۰۸) . در اقتصاد سنجی، نقاط شکست نشانه‌های مهمی از زمانی هستند که ساختار بازار به دلیل مداخلات دولت در بازارهای مالی دچار تغییر شده‌اند (لوشی و همکاران ، ۲۰۱۰). یکی دیگر از کاربردهای رگرسیون تکه‌ای، دسته‌بندی اشیا به خوشه‌های همگن است. به عنوان مثال، چانگ و همکاران (۲۰۱۵) وابستگی میزان مصرف سوخت (اندازه گیری شده با مایل در هر گالن) به وزن ماشین و قدرت آن را مورد بررسی قرار دادند. آن‌ها دو خوشه جدا مشخص کردند که میزان مصرف سوخت به طور ناگهانی پس از فراتر رفتن وزن خودرو و سرعت آن از یک آستانه مشخص تغییر می‌کند.

مساله اصلی در مدل‌های رگرسیون تکه‌ای، مشخص کردن تعداد بخش‌ها و مرزهای تقسیم‌بندی شده است که این مساله اغلب به طور جداگانه در تحقیقات رگرسیون تکه‌ای مورد بحث قرار می‌گیرند. به عنوان مثال، سیوپرکا (۲۰۱۱، ۲۰۰۹) ابتدا یک برآوردگر را برای مشخص کردن نقاط شکست، پیشنهاد کرد و سپس برآورد پارامترهای مدل را تحت این نقاط شکست انجام داد. روش‌های متعددی برای تخمین نقاط شکست و برآورد پارامترها در مدل رگرسیون تکه‌ای پیشنهاد شده است که برای اطلاعات بیشتر می‌توان به سیوپرکا (۲۰۰۴)؛ سیوپرکا و داپزل (۲۰۰۸)؛ فرنهاد (۲۰۰۶)؛ چن و همکاران (۲۰۱۱) مراجعه کرد. جلیلی و همکاران (۱۳۹۵) نمودار کنترل جدیدی مبتنی بر رویکرد آزمون خطی تعمیم یافته و رگرسیون تکه ای، به منظور پایش پروفایل های خطی چندگانه با اثرات متقابل در فاز ۲ ارائه دادند. آنها نشان دادند نمودار کنترل پیشنهادی عملکرد بسیار بهتری نسبت به نمودار کنترل مبتنی بر روش حداقل مربعات خطا دارا بودند. یکی از معایب مربوط به این روش‌ها تشخیص نقطه شکست واحد است. مشکل عمده دیگر مربوط به داده‌های پرت است. برای غلبه بر این مشکل لو و چانگ (۲۰۲۱) مدل رگرسیون

[1] Break-points or Change-points
[2] Piecewise regressions

تکه‌ای را بر پایه توزیع تی معرفی کردند و برآورد پارامترهای مدل را بر اساس روش درستنمایی انجام دادند. آنها نشان دادند که این مدل برای پوشش داده پرت مناسب‌تر از مدل بر پایه توزیع نرمال است. اخیرا، اتنبرت و بین (۲۰۲۴) روشی هوشمند را برای تخمین نقاط شکست معرفی کردند. آنها نشان دادند که نقاط شکست در این روش در یک مرحله بدست می‌آیند. در این مقاله، ما از روشی برای تخمین چندین نقطه شکست و پارامترهای رگرسیون در حالتی که خطای‌های مدل از توزیع مخلوط مقیاسی نرمال[1] (SMN) پیروی می‌کند ارائه می‌دهیم که به طور همزمان با استفاده از مدل‌های آمیخته بیان شده‌اند. درباره اهمیت استفاده از توزیع SMN می‌توان به خواصی از این توزیع در بعضی حالت‌های خاص آن برای شناسایی نقاط پرت نام برد. در این مقاله به کمک شبیه‌سازی به این خاصیت اشاره شده است. این موضوع از انگیزه‌های معرفی این مدل است. از دیگر موارد انگیزشی و نوآوری می‌توان به براورد پارامترهای مدل به صورت فرم بسته در برخی از حالت‌های خاص توزیع SMN اشاره کرد.

در این مقاله بدست آوردن نقاط شکست همزمان با طبقه‌بندی داده‌ها به داخل خوشه‌های همگن انجام می‌شود. بنابراین می‌توان از روش‌های خوشه‌بندی استفاده کرد. مدل‌های آمیخته برای خوشه‌بندی داده‌ها مدل‌هایی قوی و محبوب هستند، که با استفاده از الگوریتم امید–ماکسیمم[2] (EM) و تعمیم‌های آن به برآورد پارامترها و تشخیص نقاط شکست می‌پردازند. برای بررسی بیشتر این مطالب منابع یانگ (۲۰۱۴)؛ یلدیریم (۲۰۱۴)؛ کشاورز و هانگ (a،b۲۰۱۴)؛ چانگ و همکاران (۲۰۱۶)؛ جعفری و گلعلی‌زاده (۱۴۰۲) می‌تواند مورد مطالعه قرار گیرند. بیشتر این روش‌ها بر روی تغییرات میانگین و واریانس در محدوده دو نقطه شکست متوالی تمرکز دارند. تفاوت‌های اساسی این مقاله در مقابل مقالات گذشته استفاده از مدل‌های آمیخته برای شناسایی نقاط شکست، برآورد پارامترها با استفاده از الگوریتم EM و بدست آوردن فرم بسته برای برخی پارامترهای مدل، شناسایی خودکار داده‌های پرت و استفاده از یک خانواده از توزیع‌ها برای خطاهای مدل است.

با توجه به مزیت مدل‌های آمیخته برای خوشه‌بندی داده‌ها و تشخیص نقطه تغییر و مرز بین خوشه‌ها، با استفاده از الگوریتم EM، یک الگوریتم قوی برای خوشه‌بندی، تخمین نقاط تغییر و پارامترهای مدل رگرسیونی ارائه می‌دهیم. در بخش ۲ به طور خلاصه توزیع SMN را بررسی و برخی از ویژگی‌های مرتبط با آن را تشریح می‌کنیم. در بخش ۳ مدل پیشنهادی را بیان و برآورد پارامترها و همچنین تخمین تعداد نقاط شکست را ارائه می‌کنیم. دو مطالعه شبیه‌سازی در بخش ۴ و دو مثال واقعی در بخش ۵ برای بررسی اثربخشی روش پیشنهادی انجام شده است.

[1] Scale mixture normal distribution
[2] Expectation and maximization

۲ مروری بر توزیع مخلوط مقیاسی نرمال

متغیر تصادفی X دارای توزیع SMN با پارامتر مکان μ، پارامتر مقیاس $\sigma^۲$ و پارامتر $\boldsymbol{\nu}$ است، و با نماد $\text{SMN}(\mu, \sigma^۲, \boldsymbol{\nu})$ نشان داده می‌شود، هرگاه نمایش تصادفی آن به صورت

$$X \stackrel{d}{=} \mu + W^{-۱/۲} X_۱, \quad X_۱ \perp W, \tag{۱}$$

باشد، که در آن $\perp$ نماد استقلال دو متغیر، $X_۱$ یک متغیر تصادفی دارای توزیع نرمال با میانگین ۰ و واریانس $\sigma^۲$ و W یک متغیر تصادفی مثبت با تابع توزیع دلخواه $H(.;\boldsymbol{\nu})$ و مستقل از $X_۱$ است. با توجه به نمایش تصادفی (۱) به سادگی می‌توان نتیجه گرفت که $X|W = w \sim N(\mu, w^{-۱}\sigma^۲)$ است. بنابراین می‌توان تابع چگالی بردار تصادفی X را به صورت

$$f_{\text{SMN}}(x; \mu, \sigma^۲, \nu) = \int_۰^\infty \phi(x; \mu, w^{-۱}\sigma^۲) dH(w; \boldsymbol{\nu}) \quad x \in \mathbb{R},$$

بدست آورد، که در آن $\phi(.;\mu,\sigma^۲)$ تابع چگالی نرمال با میانگین μ و واریانس $\sigma^۲$ است. با توزیع‌های مختلفی که برای W می‌توان در نظر گرفت، موارد خاص از توزیع کلی SMN به دست می‌آیند. در این مقاله روی چند مورد رایج از موارد خاص توزیع SMN متمرکز هستیم.

الف- توزیع نرمال (N) : اگر توزیع W تباهیده در نقطه ۱ باشد، یعنی W با احتمال ۱ برابر ۱ باشد، آنگاه توزیع SMN همان توزیع نرمال $\text{N}(\mu, \sigma^۲)$ است.

ب- توزیع تی (t) : اگر W دارای توزیع گاما با پارامتر مقیاس $\nu/۲$ و پارامتر شکل $\nu/۲$ باشد، آنگاه توزیع SMN همان توزیع تی $X \sim t(\mu, \sigma^۲, \nu)$ با تابع چگالی

$$f_{\text{t}}(x; \mu, \sigma^۲, \nu) = \frac{(\pi\nu)^{-۱/۲}\Gamma\big(\frac{\nu+۱}{۲}\big)\sigma}{\Gamma\big(\nu/۲\big)\,[۱ + (x-\mu)/(\sigma^۲\nu)]^{(\nu+۱)/۲}}, \tag{۲}$$

است. باید توجه داشت که در توزیع تی اگر $\nu = ۱$ باشد، آنگاه توزیع کوشی بدست می‌آید. فرض کنید $W \sim \text{Gamma}(\nu/۲, \nu/۲)$ و $X \sim \text{t}(\mu, \sigma^۲, \nu)$ باشد. در این صورت W به شرط $X = x$ دارای توزیع گاما $\text{Gamma}(\nu + ۱, \nu + \frac{(x-\mu)^۲}{\sigma^۲})$ است. به عبارت دیگر داریم:

$$f(w \mid x) = \frac{w^{\frac{\nu+۱}{۲}-۱}}{\Gamma\big(\frac{\nu+۱}{۲}\big)}\Big(\frac{\nu + \frac{(x-\mu)^۲}{\sigma^۲}}{۲}\Big)^{\frac{\nu+۱}{۲}} \times \exp\Big\{-\frac{w}{۲}\big(\nu + \frac{(x-\mu)^۲}{\sigma^۲}\big)\Big\}.$$

ج- توزیع اسلش[1] (SL): اگر توزیع W بتا با پارامتر ν و ۱ باشد، آنگاه توزیع SMN همان توزیع اسلش $X \sim SL(\mu, \sigma^2, \nu)$ با تابع چگالی

$$f_{SL}(x; \mu, \sigma^2, \nu) = \int_0^1 \nu w^{\nu} \phi(wx; w\mu, \sigma^2) dw \quad x \in \mathbb{R}.$$

است. در صورتی‌که توزیع W بتا با پارامتر ν و ۱ باشد و $\boldsymbol{X} \sim \text{SL}(\mu, \sigma^2, \nu)$ باشد، آنگاه W به شرط $X = x$ دارای توزیع بریده شده گاما $\text{Gamma}(\nu + 2, \nu + \frac{(x-\mu)^2}{\sigma^2}))I_{[0,1]}$ است، که در آن $I_{[0,1]}(a) = \begin{cases} 1 & a \in [0,1] \\ 0 & o.w. \end{cases}$ تابع نشانگر است.

د- توزیع نرمال آلوده شده[2] (CN): اگر توزیع W دارای تابع چگالی به صورت

$$h(w; \boldsymbol{\nu}) = \nu_1 I_{\{\nu_2\}}(w) + (1 - \nu_1) I_{\{1\}}(w), \qquad (3)$$

باشد، که در آن $\boldsymbol{\nu} = (\nu_1, \nu_2)$، آنگاه توزیع SMN همان توزیع نرمال آلوده شده $X \sim \text{CN}(\mu, \sigma^2, \boldsymbol{\nu})$ با تابع چگالی

$$f_{CN}(x; \mu, \sigma^2, \boldsymbol{\nu}) = \nu_1 \phi(x; \mu, \sigma^2) + (1 - \nu_1)\phi(x; \mu, \nu_2 \sigma^2),$$

است. حال اگر $X \sim \text{CN}(\mu, \sigma^2, \boldsymbol{\nu})$ باشد. آنگاه W به شرط $X = x$ دارای تابع چگالی

$$f(w \mid x) = \frac{(\nu_1 I_{\{\nu_2\}}(w) + (1 - \nu_1) I_{\{1\}}(w))\phi(x; \mu, w\sigma^2)}{\nu_1 \phi(x; \mu, \sigma^2) + (1 - \nu_1)\phi(x; \mu, \nu_2 \sigma^2)},$$

است. در این مدل تابع چگالی رابطه (۳) را به صورت

$$h(w; \boldsymbol{\nu}) = \nu_1^{\frac{w - 1/\nu_2}{1 - 1/\nu_2}} (1 - \nu_1)^{\frac{1-w}{1-1/\nu_2}}. \qquad (4)$$

بازنوسی می‌کنیم. حال متغیر تصادفی $V = \frac{W - 1/\nu_2}{1 - 1/\nu_2}$ دارای توزیع برنولی با پارامتر ν_2 است. از این روش برای بدست آوردن برآورد پارامترهای مدل به صورت فرم بسته استفاده می‌کنیم.

---

[1] Slash distribution
[2] Contaminated-normal distribution

۳ مدل رگرسیون تکه‌ای با خطاهای مخلوط مقیاسی نرمال

در این بخش مدل‌های رگرسیون تکه‌ای براساس توزیع نرمال[۱](N-BP)، توزیع تی[۲](t-BP)، توزیع اسلش[۳](SL-BP) و توزیع نرمال آلوده شده [۴](CN-BP) در این مقاله مورد بررسی قرار می‌گیرند. در حالت کلی این مدل را با نماد SMN-BP[۵] نمایش می‌دهیم. فرض کنید $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ یک نمونه تصادفی nتایی از مدل رگرسیونی با متغیر مستقل $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jp})$ و متغیر وابسته $y_j \in \mathbb{R}$ برای $j = 1, \ldots, n$ باشد. حال متغیر n-بعدی $\boldsymbol{x}_d^* = (x_{1d}, \ldots, x_{nd})^\top$ را در نظر می‌گیریم که در آن $d = 1, \ldots, p$ است. بر این اساس مشاهدات $(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_n, y_n)$ را می توان به ترتیب افزایش در هر یک از $\boldsymbol{x}_d^*$ قرار داد بطوری‌که نقاط شکست را می‌توان با رتبه مشاهدات در مجموعه مرتب شده از $\boldsymbol{x}_d^*$ توصیف کرد. فرض کنید مدل رگرسیون تکه‌ای $c-1$ نقطه شکست با مقادیر $\tau_1, \ldots, \tau_{c-1}$ داشته که $1 \leq \tau_1 < \ldots < \tau_{c-1} \leq n-1$ است. همچنین $\tau_0 = 1$ و $\tau_c = n$ در نظر می‌گیریم. یک مدل رگرسیون تکه‌ای شامل $c-1$ نقطه شکست را می‌توان به صورت

$$\begin{aligned} &y_{ij} = \beta_{i0} + \beta_{i1} x_{j1} + \ldots + \beta_{ip} x_{jp} + \epsilon_{ij}, \\ &\tau_{i-1} \leq y_j < \tau_i, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n, \end{aligned} \tag{۵}$$

بیان کرد، بطوری که ϵ_{ij} از هم مستقل و دارای توزیع $\text{SMN}(0, \sigma_i, \nu_i)$ هستند. به دلیل اینکه $c-1$ نقطه تغییر در $n-1$ مشاهده ابتدایی صورت می‌گیرد، بنابراین $\binom{n-1}{c-1}$ حالت برای $c-1$ نقطه شکست ایجاد می‌شود. برای هر ترکیب $\tau_1, \ldots, \tau_{c-1}$ متغیر تصادفی $\omega_{\tau_1 \ldots \tau_{c-1}}$ را به صورت

$$\omega_{\tau_1 \ldots \tau_{c-1}} = \begin{cases} 1 & \text{اگر } 1 \leq \tau_1 < \ldots < \tau_{c-1} \leq n-1 \text{ نقاط شکست باشند،} \\ 0 & \text{در غیر این صورت.} \end{cases}$$

تعریف می‌کنیم. حال $\gamma_{\tau_1 \ldots \tau_{c-1}} = P(\omega_{\tau_1 \ldots \tau_{c-1}} = 1)$ را در نظر می‌گیریم، در نتیجه

$$\sum_{1 \leq \tau_1 < \ldots < \tau_{c-1} \leq n-1} \cdots \sum \gamma_{\tau_1 \ldots \tau_{c-1}} = 1.$$

---

[1] Normal Break Point
[2] Student's t Break Point
[3] Slash Break Point
[4] Contaminated-normal Break Point
[5] Scale mixture normal break point

حال متغیر Z_{ij} را به صورت

$$\sum_{1\le\tau_1<\ldots<\tau_{c-1}\le n-1,\ \tau_{i-1}\le j<\tau_i}\cdots\sum \omega_{\tau_1\ldots\tau_{c-1}} \quad 1\le i\le c,\ \ 1\le j\le n,$$

تعریف می‌کنیم و برای سایر نقاط ۰ در نظر می‌گیریم. بنابراین می‌توان نتیجه گرفت که Z_{ij} یک متغیر برنولی با مقدار ۱ برای حالتی که $(\boldsymbol{x}_j, y_j)$ متعلق به بخش iام و ۰ برای سایر نقاط است. در نتیجه مدل رگرسیون تکه‌ای به یک مدل کلاس‌بندی واضح با متغیرهای مشاهده شده $(\boldsymbol{x}, \boldsymbol{y}) = \big((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\big)^\top$ و متغیرهای غیر مشاهده شده $\boldsymbol{Z} = (\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top)^\top$ که در آن $\boldsymbol{Z}_k = (Z_{k1}, \ldots, Z_{kp})^\top$ تبدیل می‌شود. با توجه به مطالب بیان شده شکل تابع درستنمایی پیچیده است و بدست آوردن برآورد پارامترها به روش مستقیم کار دشواری است. برای به دست آوردن برآوردگرهای ماکسیمم درستنمایی[1]،(MLE) الگوریتم EM یک روش مناسب است. برای اجرای این الگوریتم روی مدل رگرسیون تکه‌ای با خطاهای مقیاسی نرمال کافی است از نمایش تصادفی

$$\begin{aligned} &Y_j|W_j = w_j, Z_{ij} = 1 \sim \mathrm{N}(\boldsymbol{x}_j^\top \boldsymbol{\beta}_i, w_j^{-1}\sigma_i^2),\\ &W_j|Z_{ij} = 1 \sim H(.;\nu),\\ &\boldsymbol{Z}_j \sim \mathrm{Multi}(1, \pi_1, \ldots, \pi_c) \end{aligned} \qquad (۶)$$

استفاده شود، که در آن Multi نماد توزیع چندجمله‌ای و π_iها وزن‌های آمیخته هستند به طوری که $\sum_{i=1}^{c} \pi_i = 1$ و مقدار آنها بر اساس سایر پارامترها بدست می‌آید و نیاز به برآورد ندارند.

۳.۱ برآورد ماکسیمم درستنمایی با استفاده از الگوریتم EM

یکی از ابزارهای مفید و کاربردی جهت به دست آوردن MLE پارامترهای مدل وقتی که فرم تابع چگالی پیچیده باشد، الگوریتم EM است. این روش که نخستین بار توسط دمپستر و همکاران (۱۹۷۷) معرفی شد، زمانی کاربرد دارد که داده‌ها بطورکامل مشاهده نشده باشند. کامل نبودن داده‌ها ممکن است به دلایل مختلف همانند وجود مشاهدات گمشده یا مشاهدات بریده شده یا سانسور شده، باشد. می‌دانیم سرعت همگرایی الگوریتم EM برای بدست آوردن جواب بهینه امکان دارد نسبت به سایر روش‌ها کم باشد. با این وجود، سادگی اجرا و بدست آوردن جواب‌های بهینه مورد اعتماد می‌تواند دلیل اصلی استفاده از این روش باشد. برای برآورد پارامترهای مدل یعنی بردار $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{c-1})$، $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_c)$، $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_c^2)$ و $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_c)$ که بطور کلی با $\boldsymbol{\theta} = \big(\boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\nu}\big)$ نمایش می‌دهیم از الگوریتم EM استفاده خواهیم کرد. برای اجرای این الگوریتم، فرض می‌شود برای داده‌های مشاهده شده‌ی $(\boldsymbol{x}, \boldsymbol{y}) = \big((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\big)^\top$، مشاهدات پنهان $\boldsymbol{W} = (W_1, \ldots, W_n)^\top$ و

[1] Maximum Likelihood Estimator

$\boldsymbol{Z} = (\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top)^\top$ نیز وجود دارند. در اینصورت لگاریتم تابع درستنمایی داده‌های کامل به صورت

$$\ell_c(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{W}, \boldsymbol{Z}) = \sum_{i=1}^{c}\sum_{j=1}^{n} Z_{ij}\Big\{ \log h(W_j;\nu_i) + \frac{1}{2}\log W_j - \log\sigma_i - \frac{W_j}{2\sigma_i^2}\big(y_j - \boldsymbol{x}_j^\top\boldsymbol{\beta}_i\big)\Big\} \qquad (7)$$

خواهد بود. در این صورت الگوریتم EM در دو گام تکرار می‌شود.

الف- گام E: برای هر تکرار k، تابع Q که به صورت امید ریاضی تابع (7) تحت شرط مشاهدات و $\hat{\boldsymbol{\theta}}^{(k)}$ محاسبه شده است، را به صورت

$$Q(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(k)}) = \sum_{i=1}^{c}\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}\Big\{\hat{\Upsilon}_{ij}^{(k)}(\nu_i) + \frac{1}{2}\widehat{t}_{ij}^{(k)} - \log\sigma_i - \frac{\widehat{w}_{ij}^{(k)}}{2\sigma_i^2}\big(y_j - \boldsymbol{x}_j^\top\boldsymbol{\beta}_i\big)\Big\}, \qquad (8)$$

به دست می‌آوریم، که در آن $\widehat{w}_j^{(k)} = E\big(W_j \mid \cdot, \hat{\Upsilon}_{ij}^{(k)}(\nu_i) = E\big(\log h(W_j;\nu_i) \mid y_j, Z_{ij} = 1, \hat{\boldsymbol{\theta}}^{(k)}\big)$ $y_i, Z_{ij} = 1, \widehat{\boldsymbol{\theta}}^{(k)}\big)$ و $\widehat{t}_{ij}^{(k)} = E\big(\log W_j \mid y_i, Z_{ij} = 1, \widehat{\boldsymbol{\theta}}^{(k)}\big)$ که با استفاده از بخش ۲ به دست می‌آیند. برآورد $\hat{z}_{ij}^{(k)}$ از طریق امید ریاضی شرطی Z_{ij} به شرط y_j و مقدار برآورد پارامترها در مرحله kام به صورت

$$\begin{aligned}\hat{z}_{ij}^{(k)} = E\big(Z_{ij} \mid y_j, \hat{\boldsymbol{\theta}}^{(k)}\big) &= \sum_{\substack{1\le\tau_1<\ldots<\tau_{c-1}\le n-1\\ \tau_{i-1}\le j<\tau_i}}\cdots\sum E\big(\omega_{\tau_1\ldots\tau_{c-1}} \mid y_j, \hat{\boldsymbol{\theta}}^{(k)}\big)\\ &= \sum_{\substack{1\le\tau_1<\ldots<\tau_{c-1}\le n-1\\ \tau_{i-1}\le j<\tau_i}}\cdots\sum P\big(\omega_{\tau_1\ldots\tau_{c-1}} = 1 \mid y_j, \hat{\boldsymbol{\theta}}^{(k)}\big),\end{aligned}$$

بدست می‌آیند. یک برآورد برای $P\big(\omega_{\tau_1\ldots\tau_{c-1}} = 1 \mid y_j, \hat{\boldsymbol{\theta}}^{(k)}\big)$ به صورت

$$\hat{\gamma}^{*}_{\tau_1\ldots\tau_{c-1}} = \Big(\prod_{j=1}^{\tau_1} f_{\text{SMN}}(y_j;\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}_1^{(k)}, \hat{\sigma}_1^{(k)^2}, \hat{\nu}_1^{(k)})\ldots\prod_{j=\tau_{c-1}+1}^{n} f_{\text{SMN}}(y_j;\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}_c^{(k)}, \hat{\sigma}_c^{(k)^2}, \hat{\nu}_c^{(k)})\Big)\hat{\gamma}^{(k)}_{\tau_1\ldots\tau_{c-1}}\Big/$$
$$\sum_{1\le g_1<\ldots<g_{c-1}\le n-1}\cdots\sum\prod_{j=1}^{g_1} f_{\text{SMN}}(y_j;\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}_1^{(k)}, \hat{\sigma}_1^{(k)^2}, \hat{\nu}_1^{(k)})\ldots\prod_{j=g_{c-1}+1}^{n} f_{\text{SMN}}(y_j;\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}_c^{(k)}, \hat{\sigma}_c^{(k)^2}, \hat{\nu}_c^{(k)})\hat{\gamma}^{(k)}_{g_1\ldots g_{c-1}},$$

حاصل می‌شود، که در آن $\hat{\gamma}_{\ldots}^{(k)}$ برآورد $\gamma_{\ldots}$ تحت $\hat{\boldsymbol{\theta}}^{(k)}$ است. این برآورد تحت $1 \le \tau_1 < \ldots < \tau_{c-1} \le n-1$ است که یکی از حالات نقاط شکست است و صورت کسر بالا از این حالت حاصل می‌شود و مخرج این کسر تمام

$\binom{n-1}{c-1}$ حالت‌های ممکن برای نقاط شکست است که به صورت جمع در مخرج آمده است. بنابراین برآورد Z_{ij} به صورت

$$\hat{z}_{ij}^{(k)} = \begin{cases} \sum \cdots \sum_{\substack{1 \le \tau_1 < \ldots < \tau_{c-1} \le n-1 \\ \tau_{i-1} \le j < \tau_i}} \gamma^*_{\tau_1 \ldots \tau_{c-1}} & \text{اگر } 1 \le i \le c, \quad 1 \le j \le n, \\ 0 & \text{سایر نقاط} \end{cases}$$

حاصل شده است. حال برآورد پارامترهای $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{c-1})$ را به صورت زیر بروزرسانی می‌کنیم:

$$\widehat{\boldsymbol{\tau}}^{(k+1)} = \arg\max_{1 \le \tau_1 < \ldots < \tau_{c-1} \le n-1} \ \gamma_{\tau_1 \ldots \tau_{c-1}}|_{\boldsymbol{\tau}} \tag{۹}$$

ب- گام CM: در این گام برای هر تکرار k، پارامتر مجهول $\boldsymbol{\theta}$ به صورت $\widehat{\boldsymbol{\theta}}^{(k+1)}$ به روزرسانی می‌شود. با ماکسیمم‌سازی (۸) نسبت به $\boldsymbol{\theta}$ برآوردهای به روز شده به صورت

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_i^{(k+1)} &= \Big[\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{w}_{ij}^{(k)} \boldsymbol{x}_j \boldsymbol{x}_j^\top\Big]^{-1} \Big[\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{w}_{ij}^{(k)} y_i \boldsymbol{x}_i\Big], \\ \widehat{\sigma}^{2(k+1)} &= \frac{1}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}} \sum_{j=1}^{n} \widehat{w}_{ij}^{(k)} \big(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}^{(k+1)})\big)^2 \end{aligned}$$

به دست می‌آیند. چون فرم بسته‌ای برای برآورد پارامتر ν وجود ندارد، آن را بوسیله ماکسیمم‌سازی تابع Q و به صورت $\widehat{\nu}^{(k+1)} = \arg\max_\nu Q(\nu, \widehat{\boldsymbol{\beta}}^{(k+1)}, \widehat{\sigma}^{2(k+1)})$ به روز رسانی می‌کنیم.

این فرآیند تا زمانی ادامه دارد که یک شرط همگرایی مانند $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})| < \varepsilon$ برقرار باشد که در آن $\ell(\widehat{\boldsymbol{\theta}}^{(k)})$ مقدار تابع درستنمایی در مرحله kام است.

$$\ell(\widehat{\boldsymbol{\theta}}) = \sum_{j=1}^{n} \log\Big(\sum_{i=c}^{n} f_{\text{SMN}}(y_j; \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}, \widehat{\sigma}_i^2, \widehat{\nu})\Big).$$

در این طرح برای جلوگیری از عدم همگرایی، از معیار همگرایی ایتکن [1] با میزان خطا $\varepsilon = 10^{-5}$، برای توقف فرایند استفاده شده است.

[1] Aickten

۳.۲ برآورد پارامتر ν در مدل CN

با توجه به مطالب ارایه شده در پایان بخش ۲، متغیر W را به V تبدیل می‌کنیم. حال داریم:

$$\begin{aligned}\widehat{v}_{ij}^{(k)} =& E\big(V_j \mid y_j, Z_{ij}=۱, \widehat{\boldsymbol{\theta}}^{(k)}\big) = P\big(V_j=۱ \mid y_i, Z_{ij}=۱, \widehat{\boldsymbol{\theta}}^{(k)}\big)\\ =& \frac{\nu_{۱i}\phi(y_j; \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_i^{(k)}, \hat{\sigma}_i^{(k)^۲})}{f_{\text{CN}}(y_j; \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_i^{(k)}, \hat{\sigma}_i^{(k)^۲}, \hat{\boldsymbol{\nu}}_i^{(k)})}.\end{aligned} \tag{۱۰}$$

بنابراین برآورد پارامتر $\boldsymbol{\nu}$ به صورت

$$\begin{aligned}\hat{\nu}_{۱i}^{(k+۱)} &= \max\Big\{\nu_{i۱}^*, \frac{۱}{n_i}\sum_{j=۱}^{n}\hat{z}_{ij}^{(k)}\hat{v}_{ij}^{(k)}\Big\},\\ \hat{\nu}_{۲i}^{(k+۱)} &= \max\Big\{\nu_{i۲}^*, \frac{\sum_{j=۱}^{n}\hat{z}_{ij}^{(k)}(۱-\hat{v}_{ij}^{(k)})\big(\frac{y_j-\boldsymbol{x}_j^\top\hat{\boldsymbol{\beta}}_i^{(k)}}{\hat{\sigma}_i^{(k)}}\big)^۲}{\sum_{j=۱}^{n}\hat{z}_{ij}^{(k)}(۱-\hat{v}_{ij}^{(k)})}\Big\},\end{aligned}$$

حاصل می‌شود، که در آن n_i تعداد مشاهداتی است که در فاصله $(\tau_{i-۱}, \tau_i]$ قرار دارند، $\nu_{i۱}^* = ۰/۵$ و $\nu_{i۲}^* = ۰/۰۰۱$.

۳.۳ تشخیص خودکار نقاط پرت

در این بخش تشخیص خودکار نقاط پرت را برای مدل t-BP و CN-BP ارایه می‌دهیم. برای مدل t-BP از فاصله ماهالانوبیس[1] استفاده می‌کنیم. ابتدا فرض می‌کنیم y_j متعلق به فاصله دو نقطه شکست $(\tau_{i-۱}, \tau_i]$ باشد. در این صورت با استفاده از رابطه

$$\delta(y_j; \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_i^{(k)}, \sigma_i^{(k)^۲}) = \big(\frac{y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_i^{(k)}}{\hat{\sigma}_i^{(k)}}\big)^۲$$

که مجانبا دارای توزیع خی دو با ۱ درجه آزادی است، نقاط پرت را شناسایی می‌کنیم. در مدل CN-BP ابتدا فرض می‌کنیم y_j متعلق به فاصله دو نقطه شکست $(\tau_{i-۱}, \tau_i]$ باشد. سپس مقدار $\widehat{v}_{ij}$ را محاسبه می‌کنیم و اگر مقدار $\widehat{v}_{ij} > ۰/۵$ باشد، آنگاه y_j یک نقطه خوب است و در صورتیکه $\widehat{v}_{ij} \leq ۰/۵$ باشد y_j یک نقطه پرت است.

[1] Mahalanobis

۳.۴ انتخاب پارامتر اولیه و انتخاب مدل مناسب

در ادامه برخی از فنون اجرای الگوریتم EM بر روی مدل رگرسیون تکه‌ای را مورد بررسی قرار می‌دهیم. این فنون شامل مقادیر اولیه و معیار مناسب بودن مدل است.

الف- مقادیر اولیه: به دلیل آنکه مدل‌های آمیخته ساده منجر به تولید توابع لگاریتم درستنمایی چند مدی می‌شوند، استفاده از الگوریتم EM بدون در نظر گرفتن مقدار اولیه ($\hat{\boldsymbol{\theta}}^{(۰)}$) مناسب ممکن است به جواب‌های بهینه منجر نشود. همچنین مقادیر اولیه مناسب سبب افزایش سرعت همگرایی الگوریتم EM خواهد شد. در نتیجه انتخاب مقدار اولیه مناسب در این مدل‌ها بسیار با اهمیت است. روش‌های متفاوتی برای انتخاب مقادیر اولیه مناسب موجود است. در زیر به یکی از این روش‌ها اشاره می‌شود.

- ابتدا برای نمونه‌ی تحت مطالعه تعداد نقاط شکست یعنی c را مشخص می‌نماییم.
- مقادیر اولیه برای $\hat{\gamma}^{(۰)}_{\tau_۱\ldots\tau_{c-۱}}$ را در هر گروه هم شانس و برابر $\frac{۱}{\binom{n-۱}{c-۱}}$ در نظر می‌گیریم.
- برای هر i، بر اساس مقادیر $\hat{\gamma}^{(۰)}_{\tau_۱\ldots\tau_{c-۱}}$، مقادیر اولیه $\hat{\sigma}_i^{(۰)۲} = ۱$ برای هر گروه در نظر می‌گیریم. حال با در نظر گرفتن مقادیر داده شده $\hat{\gamma}^{(۰)}_{\tau_۱\ldots\tau_{c-۱}}$، $\hat{\sigma}_i^{(۰)۲} = ۱$، مقدار $\hat{\boldsymbol{\beta}}_i^{(۰)}$ را با استفاده از مرحله CM محاسبه می‌کنیم.

ب- معیار مناسبت بودن مدل: برای به دست آوردن تعداد مناسب نقاط شکست c، باید الگویتم EM را برای مقادیر متفاوت c اجرا کرد و برای هر c، یکی از معیارهای مناسب بودن مدل را محاسبه کرده و براساس آن c مناسب را اختیار کرد. به منظور بررسی و مقایسه‌ی مدل‌های آماری مطرح شده در این مقاله، از دو معیار AIC[۱] و BIC[۲] که به صورت زیر محاسبه می‌شوند، استفاده خواهد شد:

$$\text{AIC} = ۲m - ۲\ell_{\max} \quad \text{و} \quad \text{BIC} = m\log n - ۲\ell_{\max},$$

که در آن m تعداد پارامترهای برآورد شده و $\ell_{\max}$ لگاریتم تابع ماکسیمم درستنمایی است.

۴ مطالعه شبیه‌سازی

در این شبیه‌سازی ویژگی‌های اریبی (Bias) و میانگین مربعات خطای (MSE) برآوردگرهای ماکسیمم درستنمایی بر اساس الگوریتم EM را بررسی خواهیم کرد. برای این شبیه‌سازی، به تعداد ۵۰۰ بار نمونه‌هایی به حجم ۵۰،۱۰۰، ۲۰۰ و ۴۰۰ از مدل‌های بیان شده با پارامترهای $\boldsymbol{\beta}_۱ = (۰, ۰/۵)$، $\boldsymbol{\beta}_۲ = (۱۰, -۰/۵)$، $\sigma_۱ = ۱$، $\sigma_۲ = ۱/۵$ و

[1] Akaike information criterion
[2] Bayesian information criterion

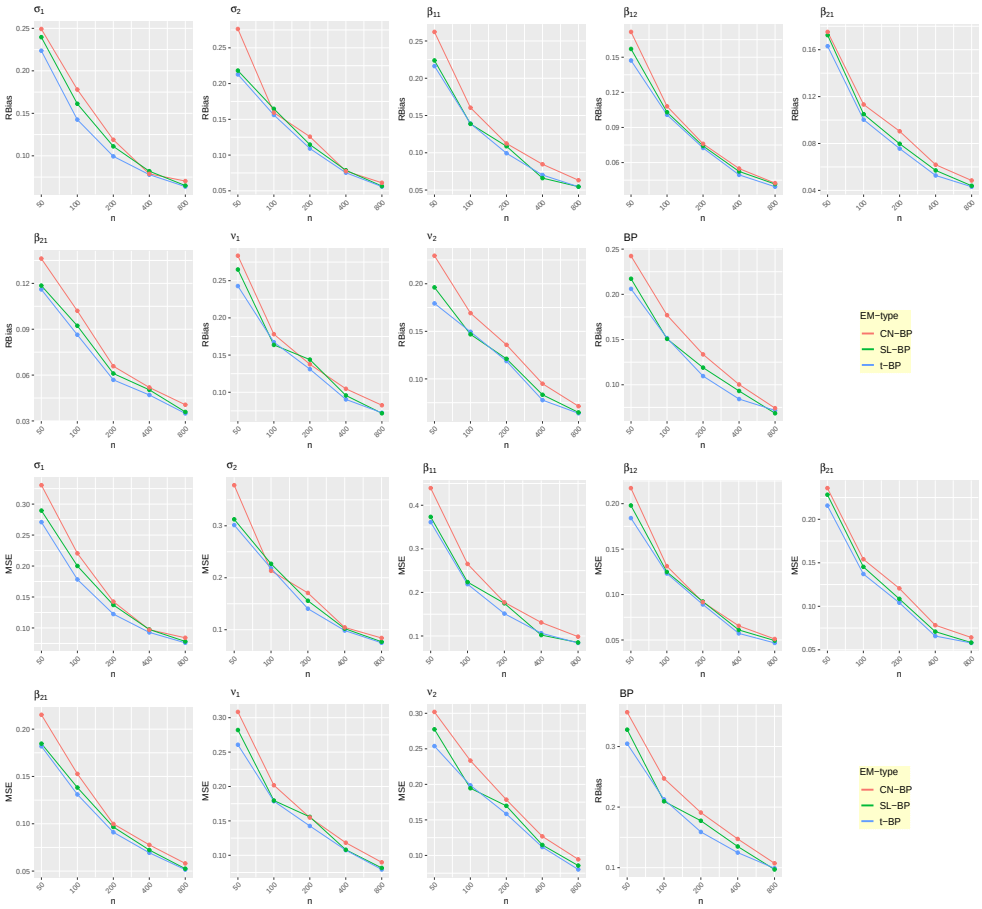
برای مدل t-BP پارامتر $\nu = ۳$، برای مدل SL-BP پارامتر $\nu = ۵$ و برای مدل CN-BP پارامترهای $\nu_۱ = ۰/۴$ و $\nu_۲ = ۱/۲$ تولید شده است. همچنین مقدار نقطه شکست برای همه مدل‌ها به ترتیب ۲۵،۵۰، ۱۰۰ و ۲۰۰ برای حجم نمونه‌های ۵۰،۱۰۰، ۲۰۰ و ۴۰۰ است. همچنین متغیر پیشگوی $\boldsymbol{x} = (۱, x_{i۱})^\top$ نیز برای هر n یعنی $x_{i۱}$ از فاصله‌ی $(۰, ۳)$ بطور یکنواخت تولید شده و ثابت در نظر گرفته شده است. در هر تکرار پس از برآورد پارامترها مدل، اریبی و مربع خطای آنها محاسبه شده است. در شکل ۱ میانگین اریبی و مربعات خطای که به صورت زیر محاسبه می‌شوند، گزارش شده است:

$$\text{Bias} = \frac{۱}{۵۰۰}\sum_{i=۱}^{۵۰۰}\left(\hat{\theta}_i - \theta\right) \quad \text{و} \quad \text{MSE} = \frac{۱}{۵۰۰}\sum_{i=۱}^{۵۰۰}\left(\hat{\theta}_i - \theta\right)^۲.$$

با توجه به این شکل مشاهده می‌شود که اریبی و میانگین مربعات خطای برآوردگرهای ماکسیمم درستنمایی با افزایش حجم نمونه به سمت صفر میل می‌کنند. بنابراین برآوردگرهای ML در مدل SMN-BP دارای ویژگی مطلوب مجانبی هستند. همچنین مشاهده شده که تخمین نقاط شکست با افزایش حجم نمونه به خوبی صورت گرفته است. این نقاط با افزایش حجم نمونه به مقدار واقعی نزدیک شده‌اند.

در ادامه، کارایی و عملکرد مدل SMN-BP را از طریق شبیه‌سازی مورد بررسی قرار می‌دهیم. در این شبیه‌سازی به بررسی مقایسه مدل‌ها تحت داده‌های تولید شده از دو فرضیه می‌پردازیم. در فرضیه اول $(S۱)$ خطاهای مدل رگرسیون تکه‌ای از توزیع SMN که در آن $W^{-۱} \sim E(۰/۵)$ ($E(۰/۵)$ یعنی توزیع نمایی با میانگین ۲) و در فرضیه دوم $(S۲)$ $W^{-۱} \sim BS(\alpha, ۱)$ ($W^{-۱} \sim BS(\alpha, ۱)$ یعنی توزیع بیرنبام-سادرز (بیرنبام و ساندرز ، ۱۹۶۹) با پارامتر شکل α و پارامتر مقیاس ۱) تولید می‌شوند. این موضوع باعث ایجاد دم سنگینی در داده‌های تولید شده می‌شود. داده‌های تولید شده از مدل‌های بیان شده برای نمونه‌های $n = ۵۰, ۱۰۰, ۲۰۰$ با دو نقطه شکست که برای هر نمونه تولید شده‌اند. پارامتر مشترک آنها $\sigma_۱ = ۱$، $\sigma_۲ = ۱/۵$، $\sigma_۳ = ۲$، $\boldsymbol{\beta}_۱ = (۰, ۰/۵)$، $\boldsymbol{\beta}_۲ = (۳, -۰/۵)$، $\boldsymbol{\beta}_۳ = (۲, ۰/۵)$ و متغیر پیشگوی $\boldsymbol{x} = (۱, x_{i۱})^\top$ نیز برای هر n یعنی $x_{i۱}$ از فاصله $(۰, ۵)$ بطور یکنواخت تولید شده و ثابت در نظر گرفته شده است. همچنین پارامتر $\alpha_۱ = ۱$، $\alpha_۲ = ۱$ و $\alpha_۳ = ۲$ را برای توزیع بیرنبام ساندرز درنظر می‌گیریم. در هر تکرار از شبیه‌سازی مدل‌های N-BP، t-BP، SL-BP و CN-BP بر روی داده‌های تولید شده از دو فرضیه برازش داده می‌شوند و شاخص‌های AIC، BIC، مجموع مربعات باقیمانده‌ها[1] (RSS) و مقادیر نقطه شکست (BP) اندازه‌گیری می‌شود. در انتها میانگین این شاخص‌ها در جدول ۱ گزارش شده است. با دقت در جدول ۱ ملاحظه می‌شود که تاثیر داده‌های پرت به ترتیب در فرضیه $S۱$ و $S۲$ در برآوردهای مدل CN-BP و t-BP در مقابل سایر مدل‌ها کمتر است. همچنین مشاهده می‌شود که مدل N-BP بیشترین تاثیر را در وجود داده‌های پرت دارد که نشان از عدم توانایی آن در مدل‌سازی داده‌های پرت دارد. این موضوع نشان می‌دهد که استفاده از سایر حالت‌های خاص مدل SMN-BP یعنی CN-BP و t-BP می‌تواند مناسبتر باشد. این موضوع می‌تواند

[1] Residual Sum of Squares



شکل ۱. نتایج شبیه‌سازی برای بررسی اریبی و میانگین مربعات خطای برآوردگرها.

به دلیل حساسیت دو مدل بیان شده در برابر داده پرت باشد. همچنین مشاهده شده است که با افزایش حجم نمونه تخمین دقیق‌تری از نقاط شکست در مدل SMN−BP صورت می‌پذیرد. بنابراین می‌توان نتیجه گرفت که در مقابل داده‌هایی که دارای دم سنگینی هستند استفاده از مدل‌های CN−BP و t−BP می‌تواند مناسب‌تر باشد. همچنین با توجه به مقادیر بیان شده برای معیار RSS می‌توان نتیجه گرفت که تحت فرضیه $S1$ مدل CN−BP نسبت به سایر مدل‌ها بهتر است. در مورد فرضیه $S2$، مدل t−BP بهتر است. بنابراین می‌توان نتیجه گرفت که دو مدل t−BP و CN−BP در مقابل داده پرت حساسیت مناسبی دارند. در ادامه برای بررسی شناسایی خودکار داده پرت شبیه‌سازی بعد را طراحی کردیم.

همان طور که نشان داده شد عملکرد مدل‌های CN−BP و t−BP در برابر داده‌های دم سنگین بهتر است.

جدول ۰۱. نتایج شبیه‌سازی برای عملکرد مدل SMN–BP در برخورد با داده‌های بسیار دم سنگین

فرضیه	نقاط تغیر نقاط	n	معیار	N–BP	t–BP	SL–BP	CN–BP
S_1	(۱۵, ۳۲)	۵۰	AIC	۲۹۸٫۵	۲۵۵٫۱	۲۶۸٫۲	۲۵۲٫۵
			BIC	۳۰۵٫۶	۲۹۰٫۷	۲۹۵٫۸	۲۸۵٫۴
			RSS	۹۷٫۵۳	۳۲٫۱۴	۵۵٫۳۲	۲۲٫۸۹
			BP	(۱۰٫۱۲, ۴۲٫۱۲)	(۱۴٫۷۳, ۳۴٫۱۷)	(۱۲٫۹۸, ۳۶٫۰۵)	(۱۵٫۱۲, ۳۲٫۵۰)
	(۲۵, ۶۷)	۱۰۰	AIC	۶۰۱٫۱۷	۵۳۰٫۴۲	۵۸۰٫۶۴	۵۱۵٫۷۷
			BIC	۶۴۰٫۱۲	۵۹۰٫۳۰	۵۹۷٫۴۲	۵۸۸٫۹۱
			RSS	۲۱۱٫۵۵	۶۹٫۱۷	۱۲۰٫۱۴	۵۰٫۲۳
			BP	(۱۴٫۱۴, ۷۵٫۱۴)	(۲۲٫۴۳, ۶۹٫۸۱)	(۲۰٫۶۱, ۷۳٫۵۵)	(۲۶٫۰۸, ۶۶٫۸۳)
	(۵۵, ۱۲۵)	۲۰۰	AIC	۱۲۵۰٫۰۸	۱۱۰۲٫۵۲	۱۱۹۰٫۸۶	۱۰۶۷٫۴۵
			BIC	۱۳۰۲٫۸۵	۱۲۰۲٫۴۱	۱۲۴۰٫۷۶	۱۱۹۰٫۴۵
			RSS	۴۴۰٫۸۶	۱۵۰٫۴۰	۲۵۰٫۷۵	۱۲۸٫۶۲
			BP	(۶۳٫۸۲, ۱۰۶٫۴۱)	(۵۱٫۸۶, ۱۲۸٫۴۱)	(۶۰٫۰۷, ۱۱۵٫۶۱)	(۵۳٫۹۷, ۱۲۶٫۰۸)
S_2	(۱۵, ۳۲)	۵۰	AIC	۳۱۷٫۹	۲۸۸٫۰	۲۹۵٫۴	۲۹۰٫۶
			BIC	۳۶۹٫۰	۳۴۷٫۲	۳۵۵٫۳	۳۵۰٫۶
			RSS	۹۰٫۴۰	۱۱٫۵۷	۴۳٫۸۰	۳۳٫۱۵
			BP	(۹٫۴۵, ۴۳٫۵۲)	(۱۵٫۰۹, ۳۲٫۱۲)	(۱۳٫۵۶, ۳۷٫۸۸)	(۱۴٫۵۲, ۳۵٫۶۳)
	(۲۵, ۶۷)	۱۰۰	AIC	۶۴۰٫۰۸	۵۸۰٫۱۲	۶۰۸٫۳۳	۵۹۹٫۰۵
			BIC	۷۸۸٫۱۲	۷۱۰٫۲۳	۷۲۹٫۸۶	۷۲۰٫۵۴
			RSS	۱۸۸٫۱۷	۵۵٫۱۴	۹۰٫۵۲	۷۰٫۸۶
			BP	(۱۶٫۳۳, ۷۳٫۱۸)	(۲۵٫۱۸, ۶۷٫۵۹)	(۲۱٫۵۲, ۷۲٫۴۷)	(۲۳٫۹۲, ۷۱٫۳۳)
	(۵۵, ۱۲۵)	۲۰۰	AIC	۱۳۱۰٫۳۶	۱۱۸۲٫۳۶	۱۲۴۰٫۱۷	۱۲۰۵٫۱۵
			BIC	۱۵۹۰٫۷۷	۱۴۵۰٫۱۲	۱۴۹۷٫۶۴	۱۴۷۷٫۱۰
			RSS	۴۰۰٫۶۳	۱۱۰٫۲۹	۱۹۲٫۴۲	۱۴۹٫۳۲
			BP	(۶۵٫۴۲, ۱۰۷٫۹۲)	(۵۶٫۴۴, ۱۲۹٫۸۳)	(۶۱٫۳۹, ۱۱۲٫۸۸)	(۵۰٫۶۹, ۱۳۰٫۴۱)

بنابراین در ادامه قصد داریم با استفاده از شبیه‌سازی، کارایی مدل‌های بیان شده در شناسایی داده پرت را مورد بررسی قرار دهیم. داده‌های شبیه‌سازی بخش ؟؟ را در نظر می‌گیریم. برای هر مجموعه داده شبیه‌سازی شده، نقاط پرت با اندازه ۲۰ را که به طور یکنواخت در محدوده داده‌های شبیه سازی تولید می‌شوند را به داده‌ها اضافه می‌کنیم. در داده‌های شبیه‌سازی شده دو نقطه تغییر وجود دارد و در نتیجه داده‌ها به سه دسته تقسیم می‌شوند. داده‌ها در محدوده اول را با y_{j1} و در محدوده سوم را با y_{j3} نشان می‌دهیم. در هر یک از ۵۰۰ تکرار به داده‌های تولید شده در بازه‌های $(\min(y_{j1}) - 5, \min(y_{j1}))$ و $(\max(y_{j3}), \max(y_{j3}) + 5)$ به‌ترتیب به دسته اول و سوم داده پرت اضافه می‌کنیم. بنابراین برای هر نمونه تولید شده ۲۰ داده پرت وجود دارد. در جدول ۲ میانگین نسبت نقاط پرتی را که به درستی به عنوان نقاط پرت شناسایی شده‌اند[1] (TPR) و نسبت نقاط خوب را که به عنوان نقاط پرت[2] (FPR) مشخص نشده اند را برای ارزیابی عملکرد تشخیص نقاط پرت ارائه می‌دهیم. از معیارهای TPR و FPR برای نشان دادن کارایی مدل در شناسایی نقاط پرت استفاده می‌کنیم.

با توجه به جدول ۲ مشاهده می‌شود که میانگین مقادیر TPR و FPR در مدل CN–BP به ترتیب بیشتر و

[1] True positive rate
[2] False positive rate

جدول ۲. میانگین مقادیر TPR و FPR برای ۵۰۰ تکرار برای عملکردهای تشخیص نقاط پرت.

۲S		۱S				
t-BP	CN-BP	t-BP	CN-BP		n	تغییر نقاط
۰٫۸۲۰	۰٫۸۸۲	۰٫۸۱۵	۰٫۸۷۱	TPR	۵۰	$(۱۵, ۳۲)$
۰٫۰۸۲	۰٫۱۷۰	۰٫۰۸۱	۰٫۱۶۹	FPR		
۰٫۸۰۲	۰٫۸۶۴	۰٫۸۱۳	۰٫۸۵۶	TPR	۱۰۰	$(۲۵, ۶۷)$
۰٫۰۵۹	۰٫۱۴۶	۰٫۰۷۵	۰٫۱۶۷	FPR		
۰٫۷۹۳	۰٫۸۳۱	۰٫۸۱۲	۰٫۸۴۱	TPR	۲۰۰	$(۵۵, ۱۲۵)$
۰٫۰۷۳	۰٫۱۳۳	۰٫۰۶۷	۰٫۱۴۳	FPR		

کمتر از مدل t-BP هستند. در ادامه با افزایش حجم نمونه مقدار TPR در هر دو مدل کاهش پیدا می‌کند. همچنین مشاهده شده است که با افزایش حجم نمونه مقدار FPR در هر دو مدل افزایش پیدا می‌کند. بنابراین مدل CN-BP در شناسایی نقاط پرت بهتر از مدل t-BP است.

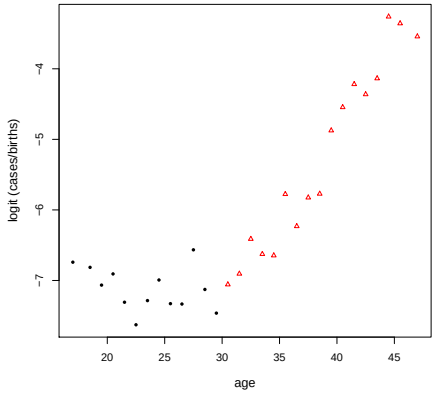
۵ مثال‌های کاربردی

مثال ۱. خطر ابتلا به سندروم داون برای مادران جوان‌تر کم است، اما با افزایش سن مادر این خطر به یک‌باره افزایش می‌یابد. مهم است که بدانیم این خطر چه زمانی و چگونه با سن مادر تغییر می‌کند. در بسته segmented در نرم افزار R داده‌هایی برای بررسی رابطه بین سن مادر و خطر ابتلاء به سندروم داون وجود دارد. این داده ها شامل ۳۰ نمونه هستند که در آنها سن مادر و همچنین آزمایشاتی مربوط به سندروم داون در مادر باردار انجام شده است. نمودارهای پراکندگی در شکل ۲ یک تغییر ناگهانی در الگوی خطر ابتلا به سندروم داون را در جنین در حدود ۳۰ سالگی مادر را نشان می‌دهد. با در نظر گرفتن یک نقطه شکست در این داده‌ها، $p-$مقدار برای آزمون اندرسون دارلینگ[1] به ترتیب برای دو دسته برابر ۰٫۸۵۸ و ۰٫۲۳۶ است. این مقادیر نشان دهنده نرمال بودن داده‌ها هستند که فرض استفاده از توزیع SMN را توجیه می‌کنند. نتایج برازش مدل‌های مختلف شامل مقدار تابع درستنمایی و معیارهای مناسبت توزیع یعنی AIC و BIC در جدول ۳ آورده شده است. برای نشان دادن مناسب بودن مدل پیشنهادی برای این داده‌ها از یک مدل رگرسیون غیر خطی نیز استفاده می‌کنیم. مدل غیر خطی بیان شده به‌صورت

$$y_j = \frac{\beta_0}{1 + \beta_1 \exp\{-\beta_2 x_j\}} + \varepsilon_j \qquad (۱۱)$$

در نظر گرفته شده است. خطاهای مدل (۱۱) از توزیع SMN پیروی می‌کنند. نتایج مربوط به مقدار تابع درستنمایی، AIC و BIC نیز در جدول ۳ آورده شده است. همان طور که مشاهد می‌شود در تمامی حالتهای خاص توزیع SMN، مدل رگرسیون تکه‌ای مقادیر AIC و BIC کمتری دارد. این موضوع نشان می‌دهد مدل رگرسیون تکه‌ای بهتر از مدل

[1] Anderson-Darling test



شکل ۲. نمودار پراکندگی سن در مقابل تابع لوجیت موارد تقسیم بر تولد در داده های سندروم داون.

غیر خطی بر این داده‌ها برازش شده است.

جدول ۳. مقادیر تابع درستنمایی و معیارهای AIC و BIC برای داده‌های سندروم داون.

معیار	N-BP		t-BP		SL-BP		CN-BP	
	تکه‌ای	غیرخطی	تکه‌ای	غیرخطی	تکه‌ای	غیرخطی	تکه‌ای	غیرخطی
$-\ell_{\max}$	−۴۵٫۴۱۸	−۷۱٫۳۳۲	−۳۱٫۸۷۹	−۵۶٫۴۳۹	−۳۵٫۴۸۳	−۶۲٫۷۵۱	−۲۷٫۲۹۰	−۵۰٫۴۰۹
AIC	۱۰۴٫۸۳۶	۱۵۰٫۶۶۴	۷۹٫۷۵۸	۱۲۲٫۸۷۸	۸۶٫۹۶۶	۱۳۵٫۵۰۲	۷۲٫۵۸۰	۱۱۲٫۸۵۸
BIC	۱۱۴٫۶۴۴	۱۵۶٫۲۶۸	۹۰٫۹۶۷	۱۲۹٫۸۸۴	۹۸٫۱۷۵	۱۴۲٫۵۰۸	۸۵٫۱۹۰	۱۲۱٫۲۶۵

حال با توجه به مطالب بیان شده، نتایج مربوط به برآورد پارامترهای مدل رگرسیون تکه‌ای در جدول ۴ آورده شده است. نتایج برازش مدل‌های مختلف شامل مقدار برآورد شده‌ی پارامترها همراه معیار RSS در جدول ۴ آورده شده است. با توجه به مقادیر AIC و BIC در جدول ۳ دیده می‌شود که مدل CN-BP نسبت به سایر مدل‌ها به داده‌ها بهتر برازش داده شده است. نتایج به دست آمده در مورد تمامی مدل‌ها (به جدول ۴ مراجعه کنید) نشان می‌دهد که خطر ابتلاء جنین به سندروم داون به طور چشمگیری پس از ۲۸٫۵ سالگی سن مادر افزایش می‌یابد. حال با توجه به معیار BIC بهترین مدل CN-BP است، که در آن برآورد شیب خط مدل رگرسیون تکه‌ای از ۰٫۰۰۹− به ۰٫۳۰۸ می‌رسد. این تغییر شیب ناگهانی در سن ۲۸٫۷ سالگی مادران باردار است. همچنین نتیجه دیگر این است که خطر ابتلاء به سندروم داون به طور قابل توجهی با سن مادر قبل از ۲۸٫۷ سالگی در مدل CN-BP مرتبط نیست، زیرا شیب خط رگرسیونی در این بازه زیاد نمی‌باشد.

مثال ۲. تجزیه و تحلیل روند تغییرات آب و هوا به ویژه برای بررسی مشکل گرمایش جهانی بسیار مهم است. چندین تحقیق (کارل و همکاران ، ۲۰۰۰؛ من ، ۲۰۰۵) به روش رگرسیونی برای مطالعه تغییرات روند دما صورت گرفته

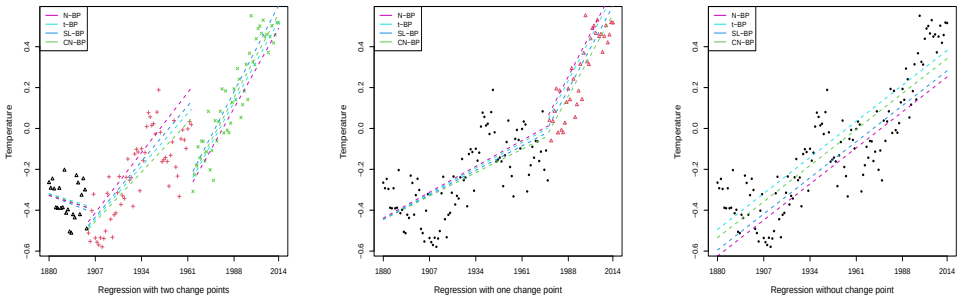
جدول ۴. برآورد پارامترها و برآورد مقدار نقطه تغییر برای داده‌های سندروم داون .

CN-BP	SL-BP	t-BP	N-BP	پارامتر
−۶٫۷۰۱	−۶٫۷۰۱	−۶٫۷۰۱	−۶٫۷۰۱	β_{11}
−۰٫۰۰۹	−۰٫۰۱۵	−۰٫۰۱۱	−۰٫۰۰۲	β_{12}
−۱۴٫۵۲۹	−۱۵٫۱۲۸	−۱۴٫۰۱۸	−۱۴٫۷۹۷	β_{21}
۰٫۳۰۸	۰٫۲۱۵	۰٫۲۵۵	۰٫۲۴۷	β_{22}
۱٫۰۱۸	۱٫۲۳۴	۱٫۰۸۲	۰٫۹۲۷	σ_1
۱٫۱۹۷	۰٫۹۹۶	۱٫۱۷۳	۱٫۲۸۶	σ_2
(۰٫۴۲۹, ۱٫۳۹۸)	۲٫۹۵۷	۴٫۸۲۳	−−	ν
۱۲٫۱۵	۱۲٫۲۴	۱۲٫۰۲	۱۲٫۴۱	BP
۲٫۲۰۱	۲٫۳۸۷	۲٫۱۵۸	۲٫۵۵۱	RSS

است و بسیاری از این محققان گزارش کردند که قابل توجه‌ترین نقطه تغییر در گرمایش زمین در اواسط دهه ۱۹۷۰ بوده است. در این مثال داده‌های مربوط به گرمایش زمین را که توسط مرکز هادلی [1] جمع آوری شده‌اند را مورد بررسی قرار می‌دهیم. این داده‌ها در بعضی نقاط به صورت گمشده بودند که از آنها صرف نظر شده است. این دماها در بازه زمانی ۱۹۸۰–۲۰۱۴ مورد بررسی قرار گرفته‌اند که میانگین این دماها در سال به عنوان مبنا در نظر گرفته شده است. شکل ۳ نتایج به دست آمده توسط مدل‌های مختلف SMN-BP را با استفاده از مدل‌هایی شامل ۰، ۱، و ۲ نقطه تغییر نشان می‌دهد. برازش داده‌ها با مدل SMN-BP، تغییر قابل توجهی را در سال ۱۹۷۶ نشان می‌دهد که با نتایج مطالعات گذشته مطابقت دارد. با توجه به جدول ۵ در تمامی مدل‌ها، مدلی با ۲ نقطه تغییر به دلیل اینکه BIC کمتری دارد بهتر برازش شده است. حال در بین مدل‌هایی با ۲ نقطه تغییر مدل CN-BP را ترجیح می‌دهیم زیرا هم BIC کمتری دارد و علاوه بر آن مقدار RSS کمتری دارد. این موضوع نشان می‌هد که مدل بیان شده در برابر سایر مدل‌های رقیب از مجموع مربع باقیمانده کمتری برخوردار است. از دلایل دیگر رد مدل با سه نقطه تغییر می‌توان به این موضوع اشاره کرد که در دسته سوم فقط ۱۰ داده قرار گرفته است که این موضوع نیز می‌تواند باعث اریبی برآوردها در این دسته شده باشد. در مدل با ۲ نقطه تغییر قبل از سال ۱۹۰۲، هیچ ارتباط معنی‌داری بین گرمایش جهانی و زمان وجود نداشت زیرا برآورد شیب خط $\beta_{11} = ۰٫۰۰۴۳$ است، که این موضوع می‌تواند به دلیل پیشرفت نچندان زیاد اکثر کشورها در صنعتی شدن باشد. در بازه زمانی ۱۹۰۲ تا ۱۹۶۴ برآورد شیب خط $\beta_{21} = ۰٫۰۲۱$ است که یک شیب ملایم برای افزایش دمای جهانی است که از دلایل آن صنعتی شدن برخی از کشورها و تولید گازهای گلخانه‌ای زیادی در سطح جهان است که موجب افزایش دمای کره زمین شده است. اما در بازه زمانی ۱۹۶۳ تا ۲۰۱۴ این شیب به شدت افزایش پیدا کرده است که مقدار آن $\beta_{31} = ۰٫۰۴۵$ است که دلیل آن می‌تواند افزایش خارج از قاعده مراکز صنعتی در تمامی کشورهای دنیا باشد. این مراکز صنعتی با تولید گازهای گلخانه‌ای زیاد باعث بالا رفتن بی‌رویه دمای سطح زمین شده‌اند. برای تکمیل و صحت نتایج بدست آمده آزمون اندرسون–دارلینگ برای نرمال بودن داده‌ها انجام شده است. با توجه به دو نقطه شکست و سه دسته داده، $p-$مقدار برای آزمون اندرسون دارلینگ به ترتیب برای سه دسته برابر ۰٫۴۲۵، ۰٫۳۲۸ و ۰٫۷۸۲ است. این مقادیر نشان دهنده نرمال بودن داده‌ها هستند که

[1] The Met Office Hadley Center

فرض استفاده از توزیع SMN را توجیه می‌کنند.



(ج) (ب) (الف)

شکل ۳. کاربرد SMN-BP برای داده‌های گرمایش سطح زمین.

جدول ۵. مقادیر بدست آمده برای شاخص‌های BIC و RSS در داده‌های مربوط به گرمایش زمین.

تعداد نقاط تغییر	شاخص	N-BP	t-BP	SL-BP	CN-BP
۰	BIC	۱۲۸۲٫۹۴۶	۹۹۸٫۵۲۳	۱۰۵۹٫۴۰۱	۹۸۹٫۴۲۷
	RSS	۳٫۰۴۹	۲٫۸۶۰	۲٫۹۶۸	۲٫۵۶۰
۱	BIC	۱۱۰۹٫۵۰۶	۹۵۶٫۴۸۶	۱۰۰۲٫۴۸۰	۹۲۱٫۷۵۰
	RSS	۲٫۸۸۵	۲٫۴۷۶	۲٫۶۸۵	۲٫۳۰۸
۲	BIC	۱۰۹۶٫۶۰۴	۹۰۹٫۹۶۷	۹۷۸٫۱۷۵	۸۹۰٫۷۰۹
	RSS	۲٫۴۵۲	۲٫۰۱۷	۲٫۲۲۵	۱٫۹۸۵

۶ بحث و نتیجه‌گیری

در این مقاله، با استفاده از مدل‌های آمیخته متناهی روشی را برای حل مسائل رگرسیون تکه‌ای با خطاهایی از توزیع SMN پیشنهاد کردیم. پیداکردن نقاط تغییر فرآیندی مشابه دسته‌بندی داده‌ها به گروه‌هایی از مشاهدات مشابه در مدل‌های آمیخته است. با در نظر گرفتن نیرومندی مدل‌های آمیخته در خوشه‌بندی داده‌ها، یک مدل SMN-BP برای تخمین نقاط تغییر و پارامترهای رگرسیونی ایجاد کردیم. همچنین با استفاده از شبیه‌سازی نشان داده شد که شناسایی نقاط پرت با استفاده از برخی حالت‌های خاص مدل SMN-BP به خوبی صورت می‌پذیرد. علاوه بر این چندین بررسی عددی با مجموعه داده‌های واقعی و شبیه سازی شده کارایی و برتری مدل پیشنهادی را نشان می‌دهند. در نتیجه، مدل پیشنهادی SMN-BP معمولاً برآوردهای دقیقی از نقاط تغییر و پارامترهای رگرسیون را ارائه می‌دهد.

از این رو، مدل SMN-BP یک رویکرد مطلوب برای حل مشکلات رگرسیون تکه‌ای است. رگرسیون تکه‌ای را می‌توان در مسائل مختلفی که در آنها نقطه تغییر مهم است مانند کنترل کیفیت، پزشکی، اقتصاد و پردازش سیگنال اعمال کرد. در بیشتر موارد این داده‌ها دارای کجی و چولگی هستند. در تحقیقات آینده، سعی می‌کنیم از مدل‌ها دارای چولگی و کجی برای برآورد نقطه تغییر قویتر برای مجموعه‌های داده با نقاط پرت استفاده کنیم. همچنین قصد داریم از خوشه‌بندی فازی برای برآورد پارامترهای مدل رگرسیون تکه‌ای و نقاط تغییر با داده‌هایی دارای کجی استفاده کنیم.

تقدیر و تشکر

نویسنده بر خود لازم می‌داند؛ مراتب سپاس خود را از سردبیر محترم، داوران گرامی و ویراستار مجله که با نظرات ارزشمند خود موجبات ارتقای این مقاله را فراهم کردند، اعلام نماید.

مراجع

جلیلی، م.، بشیری، م.، منطقی، م. و توفیق، ع. (۱۳۹۵)، توسعه رویکردی مبتنی بر رگرسیون تکه ای برای پایش پروفایل های خطی چندگانه با اثرات متقابل در فاز، مهندسی و مدیریت کیفیت، ۶، ۲۳۷-۲۴۹.

جعفری، ف.، و گلعلی‌زاده، م. (۱۴۰۲)، مدل‌بندی رگرسیونی چندکی آمیخته تاوانیده دوگانه از طریق رویکرد درستنمایی، مجله علوم آماری، ۱۷، ۳۸۹-۴۰۵.

Birnbaum, Z. W., and Saunders, S. C. (1969), A New Family of Life Distributions, *Journal of Applied Probability*, **6** (2), 319–327.

Chang, S. T. and Lu, K. P. (2016), Change-Point Detection for Shifts in Control Charts Using EM Change- Point Algorithms, *Quality and Reliability Engineering International*, **32**, 889–900.

Chang, S. T, Lu, K. P. and Yang, M. S. (2015), Fuzzy Change-Point Algorithms for Regression Models, *IEEE Transactions on Fuzzy Systems*, **23** (6), 2343–2357.

Chen, C. W. S., Chan, J. S., Gerlach K. R., and Hsieh, W. Y. L. (2011), A Comparison of Estimators for Regression Models with Change Points, *Statistics and Computing*, **21**, 395–414.

Ciuperca, G. (2009), The M-Estimation in a Multi-Phase Random Nonlinear Model, *Statistics & Probability Letters*, **75**, 573–580.

Ciuperca, G. (2011), A General Criterion to Determine the Number of Change-Points, *Statistics & Probability Letters*, **81**, 1267–1275.

Ciuperca, G. (2004), Maximum Likelihood Estimator in a Two-Phase Nonlinear Regression Model, *Statistics & Decisions*, **22**, 335–349.

Ciuperca, G. and Dapzol, N. (2008), Maximum Likelihood Estimator in a Multi-Phase Random Regression Model, *Statistics*, **42**, 363–381

Dempster, A. P., Laird N. M., and Rubin D. B. (1977), Maximum Likelihood From Incomplete Data via the EM-Algorithm, *Journal of the Royal Statistical Society. Series B (methodological)*, **39**, 1–38.

Fearnhead, P. (2006), Exact and Efficient Bayesian Inference for Multiple Change-Point Problems, *Statistics and Computing*, **16**, 203–213.

Julious, S. A. (2001). Inference and Estimation in a Change Point Regression Problem, *Journal of the Royal Statistical Society Series D*, **50**, 51–61.

Karl, T.R., Knight, R.W. and Baker, B. (2000), The Record Breaking Global Temperatures of 1997 and 1998: Evidence for an Increase in the Rate of Global Warming?, *Geophysical Research Letters*, **27** (5), 719–722

Keshavarz, M. and Huang, B. (2014a), Bayesian and Expectation Maximization Methods for Multivariate Change-Point Detection, *Computers & Chemical Engineering*, **60**, 339–353.

Keshavarz M. and Huang, B. (2014b), Expectation Maximization Method for Multivariate Change-Point Detection in Presence of Unknown and Changing Covariance, *Computers & Chemical Engineering*, **69**, 128–146.

Loschi, R. H., Pontel, J. G. and Cruz F. R. B. (2010), Multiple Change-Point Analysis for Linear Regression models, *Chilean Journal of Statistics*, **1**, 93–112.

Lu, K. P., and Chang, S. T. (2021), Robust Algorithms for Change-Point Regressions Using the t-Distribution, *Mathematics*, **9**, 2394.

Menne, J. M. (2005), Abrupt Global Temperature Change and the Instrumental Record, *In: Record, 18th Conference on Climate Variability and Change*.

Muggeo, V. M. R. (2008), Segmented: an R package to Fit Regression Models with Broken-Line Relationships, *R News*, **8**, 20–25.

von Ottenbreit, M., and De Bin, R. (2024), Automatic Piecewise Linear Regression, *Computational Statistics*, 1-41.

Yang, F. (2014), Robust Mean Change-Point Detecting Through Laplace Linear Regression Using EM Algorithm, *Journal of Applied Mathematics*, 2014 (1): 856350.

Yildirim S., Singh S.S., and Doucet A. (2014), An Online Expectation-Maximization Algorithm for Change-Point Models, *Journal of Computational and Graphical Statistics*, **22**(4), 906–26.