برآورد استوار در مدل رگرسیون خطی: روش واگرایی توان چگالی

سمیه گله [۱]، روح الله روزگار [۲]

تاریخ دریافت: ۹۸/۶/۲۶

تاریخ پذیرش: ۹۹/۳/۱

چکیده:

روش کمترین واگرایی توان چگالی یک برآورد استوار در مواجهه با موقعیت‌هایی که داده‌ها شامل تعدادی داده پرت هستند ارائه می‌دهد. در این پژوهش به معرفی و استفاده از برآوردگر استوار کمترین واگرایی توان چگالی برای برآورد پارامترهای مدل رگرسیون خطی پرداخته و در ادامه با چند مثال عددی از رگرسیون خطی، استواری این برآوردگر را در مواجهه با مجموعه داده‌هایی که شامل تعدادی داده پرت هستند نشان می‌دهیم.

واژه‌های کلیدی: برآوردیابی، برآوردگر کمترین واگرایی توان چگالی، برآوردگر ماکسیمم درستنمایی، داده پرت، داده‌های تلفن بلژیک، رگرسیون استوار.

۱ مقدمه

استفاده از روش کمترین فاصله، یکی از روش‌های مهم در استنباط آماری به شمار می‌رود. در این روش، پارامتر موردعلاقه خود را به وسیله مینیمم کردن یک اندازه مناسب بین داده‌ها و مدل در نظر گرفته‌شده برآورد می‌کنیم[۴]. در بیشتر موارد برای برآورد پارامترهای مجهول با موقعیت‌هایی روبه‌رو می‌شویم که داده‌ها شامل تعدادی داده پرت هستند و اغلب اوقات نیز نمی‌توانیم به‌سادگی از این نقاط پرت صرف‌نظر کنیم و حتی گاهی در داده‌های چندمتغیره تشخیص این داده‌های پرت بسیار مشکل به نظر می‌رسد. بنابراین حذف کردن داده‌های پرت از مجموعه داده‌ها همیشه درست‌ترین راه‌حل نمی‌باشد و بهتر است به جای حذف این داده‌های پرت از روش‌های استوار برآورد استفاده کنیم. برآوردگرهای استوار در مواجهه با داده‌های پرت حساسیت کمتری از خود نشان داده و در سایر مواقع که داده پرتی وجود ندارد برآوردی شبیه سایر برآوردگرهای معمول ارائه می‌دهد. به دلیل اینکه اغلب برآوردگرهای کمترین واگرایی مبتنی بر چگالی از جمله خانواده واگرایی توان چگالی دارای خصوصیت استواری هستند [۴] بنابراین می‌توان از آن‌ها برای برآورد پارامترهای موردنظر زمانی که داده‌ها شامل تعدادی داده پرت هستند استفاده کرد. در خانواده کمترین واگرایی توان چگالی [۳]، پارامتر موردعلاقه خود را از طریق مینیمم کردن فاصله بین تابع چگالی واقعی داده‌ها و تابع چگالی مدل در نظر گرفته برای آن‌ها برآورد می‌کنیم. خانواده *MDPD* برای اولین بار در سال ۱۹۹۸ توسط باسو [۱] معرفی شد و محققانی چون گوش و باسو [۴] و [۳]، دوریو و ایسایا [۲] و...در تحقیقات خود خواص مجانبی، سازگاری، استواری، کارایی و همچنین موارد استفاده این برآوردگر را در زمینه برآورد نقطه‌ای، فاصله اطمینان‌های متفاوت و آزمون‌های فرضیه موردبررسی قرار داده‌اند. تنها پارامتر تأثیرگذار در این خانواده، پارامتر نامنفی α است که تعادل بین کارایی و استواری این برآوردگر را برقرار می‌کند.

۲ برآوردگر کمترین واگرایی توان چگالی

تعریف ۱.۲. فرض کنید که G یک توزیع با تابع چگالی g و تکیه‌گاه χ باشد. همچنین فرض کنید که $\mathscr{F}_\theta = \{f_\theta(x) : \theta \in \Theta, x \in \chi\}$ خانواده‌ای از توابع چگالی و $p \geq 1$ و $\Theta \subseteq R^p$ باشد. آنگاه اندازه واگرایی توان چگالی بین f یکی از چگالی‌های درون خانواده چگالی و g به صورت زیر تعریف می‌شود:

$$d_\alpha(g,f) = \int_\chi \{f^{1+\alpha}(x) - (1+\frac{1}{\alpha})f^\alpha(x)g(x) + \frac{1}{\alpha}g^{1+\alpha}(x)\}dx. \qquad \alpha > 0 \tag{1}$$

[۱] دانش‌آموخته کارشناسی ارشد آمار، گروه ریاضی دانشگاه یاسوج، یاسوج، ایران. somaie.galleh@gmail.com

[۲] هیئت‌علمی گروه ریاضی، دانشگاه یاسوج، یاسوج، ایران. roozegar@yu.ac.ir

[3] Minimum Density Power Divergence(MDPD)

[4] Kullback-Leibler

در حالت خاص $\alpha = 0$ داریم:

$$d_0(g,f) = \lim_{\alpha \to 0} d_\alpha(g,f) = \int_\chi g(x) \ln(\frac{g(x)}{f(x)}) dx.$$

این فاصله آماری خاص در واقع همان اندازه واگرایی کولبک-لیبلر [۴] است و برآوردگری که از مینیمم سازی این عبارت به دست می‌آید در واقع همان برآوردگر ماکسیمم درستنمایی است. اگر برآوردگر $MDPD$ بردار پارامتری θ را به وسیله $T_\alpha(G)$ نشان دهیم، آنگاه

$$d_\alpha(f_\alpha, f_{T_\alpha(G)}) = \min_{\theta \subseteq \Theta} d_\alpha(g, f_\theta). \quad (۲)$$

توجه داشته باشید که در رابطه (۱) عبارت سوم مستقل از θ است و در فرآیند مینیمم سازی از این عبارت استفاده نمی‌شود.

۱.۲ برآوردگر کمترین واگرایی توان چگالی برای مشاهدات مستقل و هم توزیع

فرض کنید $x_1, ..., x_n$ یک نمونه تصادفی مستقل و هم توزیع از توزیع واقعی g باشد. همان‌طور که اشاره کردیم برآوردگر $MDPD$ برای برآورد پارامتر مجهول θ را از مینیمم کردن عبارت زیر به دست می‌آوریم:

$$\int f_\theta^{1+\alpha}(x) dx - (1 + \frac{1}{\alpha}) \int f_\theta^\alpha(x) dG(x),$$

به طوری که $dG(x)$ دیفرانسیل تابع توزیع $G(x)$ است. در اکثر اوقات توزیع واقعی G ناشناخته است و بنابراین از تقریب آن یعنی توزیع تجربی $G_n(.)$ استفاده می‌شود. می‌دانیم $G_n(x)$ یک اندازه گسسته روی نقاط $x_1, ..., x_n$ با اندازه یکسان $\frac{1}{n}$ است، بنابراین داریم $\int f_\theta^\alpha dG_n(x) = \sum_{i=1}^n \frac{1}{n} f_\theta^\alpha(X_i)$، در نتیجه برآوردگر $MDPD$ پارامتر θ را می‌توان به وسیله مینیمم کردن عبارت زیر محاسبه کرد.

$$H_\alpha(\theta) = \int f_\theta^{1+\alpha}(x) dx - (1 + \frac{1}{\alpha}) \frac{1}{n} \sum_{i=1}^n f_\theta^\alpha(X_i). \quad (۳)$$

اگر مشتق $H_\alpha(\theta)$ نسبت به θ را $U_\alpha(\theta)$ بنامیم، آنگاه با مشتق گرفتن از رابطه (۳) نسبت به θ معادله برآوردی به صورت زیر خواهد بود.

$$U_\alpha(\theta) = \int u_\theta^{1+\alpha}(x) f_\theta^{1+\alpha}(x) - \frac{1}{n} \sum_{i=1}^n u_\theta(X_i) f_\theta^\alpha(X_i) = 0.$$

هنگامی که $u_\theta(x) = \frac{\partial}{\partial \theta} \log f_\theta(x)$. این معادله به صورت مدل وزنی معادله ماکسیمم درستنمایی در نظر گرفته می‌شود. با توجه به اینکه برای تمام $\alpha > 0$ این معادله برای مشاهدات پرت، وزن پایین‌تر و به مشاهداتی که به نقطه مد داده‌ها نزدیک‌تر باشند وزن بیشتری اختصاص می‌دهد. بنابراین هر چه مقدار α به یک نزدیک‌تر باشد استواری برآوردگر $MDPD$ بیشتر خواهد شد. در مقابل در حالتی که $\alpha = 0$ (در این حالت برآوردگر $MDPD$ همان برآوردگر غیر استوار ماکسیمم درستنمایی است) تمام مشاهدات حتی آن‌هایی که به عنوان داده پرت می‌شناسیم وزنی برابر یک به خود می‌گیرند. عملاً انتخاب α نزدیک به صفر، باعث بهبود کارایی مجانبی برآوردگر در مقایسه با روش کمترین فاصله می‌شود. بنابراین برآوردگر $MDPD$ یک توازن بین کارایی و استواری برقرار می‌کند.

۲.۲ برآوردگر کمترین واگرایی توان چگالی برای مشاهدات مستقل و غیر هم توزیع

فرض کنید $Y_1, ..., Y_n$ یک نمونه تصادفی مستقل به حجم n و برای هر $i = 1, ..., n$، $Y_i \sim g_i$ باشد. توجه کنید که این نمونه تصادفی اگرچه ممکن است توزیع‌های متفاوتی داشته باشند ولی همگی آن‌ها در پارامتر θ مشترک هستند. در این حالت برای برآورد پارامتر موردنظر ما باید اختلاف بین داده‌ها و مدل را به‌طور جداگانه برای هر نقطه محاسبه کرده و سپس سعی کنیم تمامی این فواصل را مینیمم کنیم. یک راه‌حل ساده‌تر این است که برای مینیمم کردن تک‌تک فواصل بین هر چگالی واقعی با مدل در نظر گرفته برای آن، از تمام فواصل بین چگالی‌های واقعی و مدل در نظر گرفته برای هر نقطه، میانگین بگیریم و سپس میانگین حاصل را مینیمم کنیم. اگر به جای چگالی مجهول واقعی داده‌ها از برآورد آن $\hat{g}_i$ استفاده کرده و فاصله بین هر چگالی f_i و $\hat{g}_i$ را با $d_\alpha(\hat{g}_i, f_i(.;\theta)), \theta \in \Theta$ نشان دهیم آنگاه از مینیمم کردن عبارت

$$\frac{1}{n} \sum_{i=1}^n d_\alpha(\hat{g}_i, f_i(.;\theta)), \theta \in \Theta, \quad i = 1, ..., n,$$

پارامتر موردنظر را برآورد می‌کنیم. به دلیل اینکه در این حالت از هر تابع چگالی واقعی g_i، تنها یک نقطه Y_i در دسترس داریم بنابراین مجبوریم که در تخمین چگالی g_i از توزیع تباهیده شده در نقطه Y_i استفاده کنیم و بنابراین

$$\begin{aligned} d_\alpha(\hat{g}_i, f_i(.;\theta)) &= \\ &\int f_i(y;\theta)^{1+\alpha} dy - (1 + \frac{1}{\alpha}) f_i(Y_i;\theta)^\alpha + K, \end{aligned} \quad (۴)$$

که K یک ثابت و مستقل از پارامتر θ است. بنابراین برای برآورد θ کافی است که تابع هدف زیر را مینیمم کنیم.

$$\begin{aligned} H_n(\theta) &= \frac{1}{n} \sum_{i=1}^n \{ \int f_i(y;\theta)^{1+\alpha} dy - (1 + \frac{1}{\alpha}) f_i(Y_i;\theta)^\alpha \} \\ &= \frac{1}{n} \sum_{i=1}^n V_i(Y_i;\theta), \end{aligned} \quad (۵)$$

که $V_i(Y_i;\theta) = \int f_i(y;\theta)^{1+\alpha} dy - (1 + \frac{1}{\alpha}) f_i(Y_i;\theta)^\alpha$ است. با مشتق گرفتن از رابطه (۵) نسبت به θ، معادله برآورد به صورت زیر در می‌آید.

$$\begin{aligned} \nabla \sum_{i=1}^n V_i(Y_i;\theta) &= \sum_{i=1}^n \{ f_i(Y_i;\theta)^\alpha u_i(Y_i;\theta) \\ &- \int f_i(y;\theta)^{1+\alpha} u_i(y;\theta) dy \} = 0, \end{aligned} \quad (۶)$$

که منظور از $\bigtriangledown$ در عبارات بالا مشتق جزئی نسبت به θ است و داریم $u_i(y;\theta) = \bigtriangledown \ln f_i(y;\theta)$. معادله برآوردی بالا زمانی ناریب است که چگالی مولد (چگالی واقعی که داده‌ها از آن پیروی می‌کنند) g_i به ازای $i = ۱, ۲, ...$ متعلق به خانواده چگالی مدل در نظر گرفته‌شده یعنی $\mathscr{F}_{i,\theta} = \{f_i(.;\theta) : \theta \in \Theta\}$ باشد. اگر $\alpha \to ۰$ تابع هدف موردنظر ما زمانی مینیمم می‌شود که عبارت $\frac{۱}{n}\sum_{i=۱}^{n}\{-\ln(f_i(Y_i;\theta))\}$ مینیمم شود که معادل با ماکسیمم کردن عبارت $\frac{۱}{n}\sum_{i=۱}^{n}\{\ln(f_i(Y_i;\theta))\}$ می‌باشد و لذا محاسبات منجر به برآوردگر ماکسیمم درستنمایی می‌شود. به‌طورکلی برآوردگر کمترین واگرایی توان چگالی $T_\alpha(G_۱,...,G_n)$ برای مشاهدات مستقل ولی غیر هم توزیع آن برآوردگری است که در عبارت زیر صدق کند:

$$\frac{۱}{n}\sum_{i=۱}^{n} d_\alpha(g_i(.), f_i(.;T_\alpha(G_۱,...,G_n))) = \min_{\theta\in\Theta}\frac{۱}{n}\sum_{i=۱}^{n} d_\alpha(g_i(.), f_i(.;\theta)). \quad (۷)$$

یعنی برآوردگر $MDPD$ آن مقداری از پارامتر θ در فضای پارامتری Θ است که به ازای آن میانگین تمام فواصل بین g_i ها و f_iها مینیمم شود.

۳.۲ برآورد ضرایب رگرسیون خطی با استفاده از برآوردگر کمترین واگرایی توان چگالی

یک مثال آشنا از داده‌های مستقل و غیر هم توزیع، بحث رگرسیون خطی است. باسو [1] (بخش ۳-۵) به‌طور خلاصه در مورد کاربرد برآوردگر $MDPD$ در برآورد پارامترهای رگرسیون خطی توضیحاتی ارائه داده است. مدل رگرسیون خطی به صورت زیر تعریف می‌شود:

$$y_i = x_i^T\beta + \varepsilon_i, \quad i = ۱,...,n,$$

وقتی‌که ε_i(خطاها) متغیرهای تصادفی نرمال با میانگین صفر و واریانس مجهول $\sigma^۲$ هستند. همچنین $x_i^T = (x_{i۱},...,x_{ip})$بردار متغیرهای مستقل مربوط به iاُمین مشاهده و $\beta = (\beta_۱,...,\beta_p)^T$ بردار ضرایب رگرسیونی می‌باشند. بنابراین به ازای هر i و $i = ۱,...,n$ ، $y_i \sim N(x_i^T\beta, \sigma^۲)$. یعنی y_iها متغیرهای تصادفی مستقل و غیر هم توزیع هستند. هدف ما برآورد بردار پارامتری $\theta = (\beta^T, \sigma^۲)^T$ است. نشان دادیم که در حالتی که داده‌ها مستقل و غیر هم توزیع هستند می‌توانیم برآورد $MDPD$ بردار θ را به وسیله مینیمم کردن عبارت (۵) به دست بیاوریم. با جایگذاری f_i در عبارت (۵) داریم:

$$\begin{aligned} H_n(\theta) &= \frac{۱}{n}\sum_{i=۱}^{n}\left\{ \frac{۱}{(\sqrt{۲\pi\sigma^۲})^\alpha}\cdot\frac{۱}{\sqrt{۱+\alpha}} - (\frac{۱+\alpha}{\alpha})(\frac{۱}{(\sqrt{۲\pi\sigma^۲})^\alpha})e^{\frac{-\alpha}{۲\sigma^۲}(y_i - x_i^T\beta)^۲} \right\} \\ &= \frac{۱}{n}\sum_{i=۱}^{n} V_i(y_i;\theta,x_i). \end{aligned}$$

اگر $\nabla_j, j = ۱,...,p$ و $\nabla_{p+۱}$ را به ترتیب مشتق جزئی نسبت به β_j و $\sigma^۲$ تعریف کنیم آنگاه داریم:

$$\begin{aligned} \nabla_j(V_i(y_i;\theta,\mathbf{x}_i)) &= -(\frac{۱+\alpha}{\alpha}).\frac{(-(-\alpha))}{\sigma^{۲+\alpha}(۲\pi)^{\frac{\alpha}{۲}}} x_{ij}(y_i - \mathbf{x}_i^T\beta)e^{\frac{-\alpha}{۲\sigma^۲}(y_i-\mathbf{x}_i^T\beta)^۲} \\ &= -\frac{۱+\alpha}{(۲\pi)^{\frac{\alpha}{۲}}\sigma^{\alpha+۲}} x_{ij}(y_i - \mathbf{x}_i^T\beta)e^{\frac{-\alpha}{۲\sigma^۲}(y_i-\mathbf{x}_i^T\beta)^۲}, \end{aligned} \quad (۸)$$

و همچنین

$$\begin{aligned} \nabla_{p+۱}(V_i(y_i;\theta,\mathbf{x}_i)) &= \frac{-\alpha}{۲(۲\pi)^{\frac{\alpha}{۲}}\sqrt{۱+\alpha}\sigma^{\alpha+۲}} \\ &\quad + (\frac{۱+\alpha}{\alpha})(\frac{\alpha}{۲(۲\pi)^{\frac{\alpha}{۲}}\sigma^{\alpha+۲}})e^{\frac{-\alpha}{۲\sigma^۲}(y_i-\mathbf{x}_i^T\beta)^۲} \\ &\quad + \frac{\alpha(y_i-\mathbf{x}_i^T\beta)^۲(۱+\alpha)}{۲\alpha\,\sigma^{\alpha+۴}(۲\pi)^{\frac{\alpha}{۲}}}e^{\frac{-\alpha}{۲\sigma^۲}(y_i-\mathbf{x}_i^T\beta)^۲} \\ &= -\frac{۱}{(۲\pi)^{\frac{\alpha}{۲}}} \\ &\quad \times\left[\frac{\alpha}{۲\sigma^{\alpha+۲}\sqrt{۱+\alpha}} - \frac{(۱+\alpha)e^{-\alpha(y_i-\mathbf{x}_i^T\beta)^۲}}{۲\sigma^{\alpha+۲}}(۱-\frac{(y_i-\mathbf{x}_i^T\beta)^۲}{\sigma^۲}) \right]. \end{aligned} \quad (۹)$$

بنابراین برای به دست آوردن برآورد ضرایب رگرسیونی مشتقات گفته‌شده را برابر صفر قرار می‌دهیم.

$$\Longrightarrow \sum_{i=۱}^{n}(x_{ij}(y_i - \mathbf{x}_i^T\beta))e^{\frac{-\alpha}{۲\sigma^۲}(y_i-\mathbf{x}_i^T\beta)^۲} = ۰, j = ۱,...,p. \quad (۱۰)$$

$$\Longrightarrow \sum_{i=۱}^{n}(۱-\frac{(y_i-\mathbf{x}_i^T\beta)^۲}{\sigma^۲})e^{-\alpha(y_i-\mathbf{x}_i^T\beta)^۲} = \frac{\alpha}{(۱+\alpha)^{\frac{۳}{۲}}}. \quad (۱۱)$$

با توجه به تعداد متغیرهای مستقل به‌کاررفته در مدل رگرسیون(p) با حل کردن معادلات (۱۰) و (۱۱) و به دست آوردن ریشه‌های آن‌ها برآورد بردار پارامتری θ یعنی $\hat{\theta}^T = (\hat{\beta}^T, \hat{\sigma}^۲)$ به دست می‌آید. در بخش‌های قبل نحوه محاسبه برآوردگرهای $MDPD$ برای داده‌های مستقل و غیر هم توزیع دارای چگالی نرمال ارائه شد، توجه کنید که برای به دست آوردن برآوردگر $MDPD$ برای هر تابع چگالی، می‌توانیم با جایگذاری تابع چگالی موردنظر خودمان و سپس حل کردن معادلات به‌دست‌آمده برآوردهای موردنظر را از این روش محاسبه کنیم.

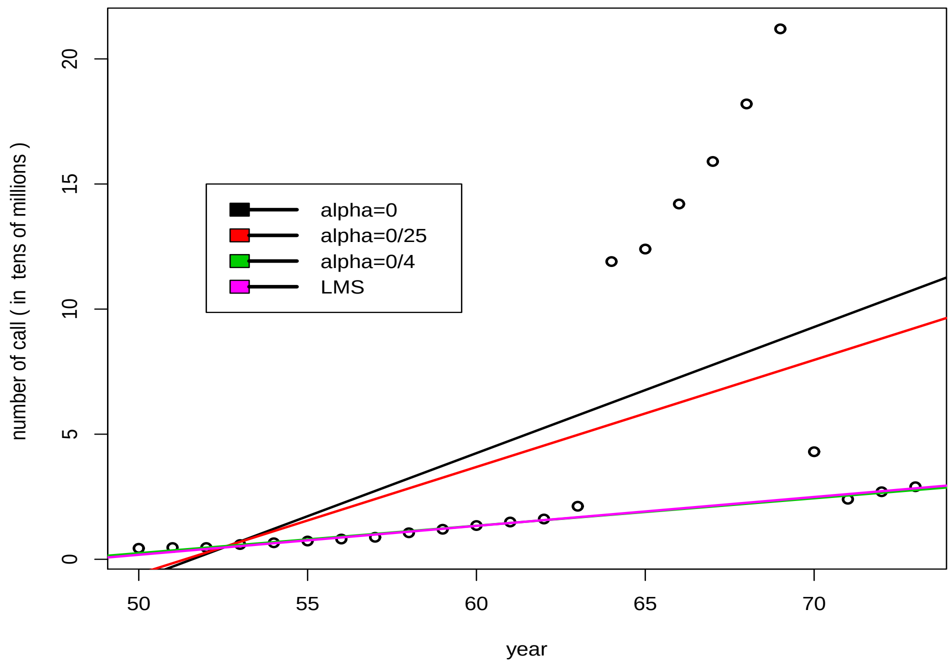
[5] Least Median Square(LMS)

۴.۲ مطالعات عددی

در مثال واقعی زیر برای مقایسه هر چه بهتر خصوصیت استواری روش $MDPD$، برآوردها را از سه روش $MDPD$، روش ماکسیمم درستنمایی و روش برآورد استوار میانه [۵] ارائه کرده و نتایج را مقایسه کرده‌ایم.

مثال ۲.۲. مجموعه داده‌های تلفن بلژیک به عنوان یک مثال واقعی از رگرسیون خطی در نظر گرفته‌شده است [۵] (جدول ۲، فصل ۲). این مجموعه داده‌ها توسط وزارت اقتصاد کشور بلژیک تهیه و شامل تعداد مکالمات بین‌المللی از سال ۱۹۵۰ تا سال ۱۹۷۳ است. در این داده‌ها سال به عنوان متغیر مستقل و تعداد مکالمات به عنوان متغیر وابسته در نظر گرفته‌شده است. برای این داده‌ها برآورد از روش $MDPD$ به ازای چندین مقدار مختلف α و برآورد از روش رگرسیون میانه در جدول (۱) و نمودار پراکنش این داده‌ها به همراه بعضی خط‌های برازش شده از روش $MDPD$ به ازای چندین مقدار مختلف α به همراه خط برازش شده از روش میانه در شکل (۱) نشان داده‌شده است. خط برازش شده به داده‌ها به ازای $\alpha = ۰$ که متناظر با برآورد ماکسیمم درستنمایی است به‌شدت تحت تأثیر داده‌های پرت قرارگرفته و برآورد مناسبی نیست، همچنین خط برازش شده از روش $MDPD$ به ازای $\alpha = ۰/۴$ و روش میانه کاملاً بر یکدیگر منطبق بوده، که این نکته استواری روش $MDPD$ را نشان می‌دهد. مقادیر برآورد شده در جدول (۱) نیز به خوبی نشان می‌دهد که روش $MDPD$ با $\alpha \geq ۰/۴$ همانند روش رگرسیون میانه، برآوردی بسیار استوار در مواجهه با داده‌های پرت از خود نشان می‌دهد. لازم به ذکر است که برآوردهای حاصل از روش $MDPD$ از کارایی بالاتری نسبت به برآوردگر LMS برخوردار هستند.

مثال ۳.۲. در این شبیه‌سازی قصد داریم رفتار استوار برآوردگر $MDPD$ را در مواجهه با افزایش داده‌های پرت نشان دهیم. اگر نمونه اصلی را که شامل ۲۰۰ داده تولیدشده از مدل $Y = X_{n_i} + \varepsilon$ باشد به طوری که $X_{n_1=180} \sim \mathscr{U}(۰, ۰/۵)$، $X_{n_2=20} \sim \mathscr{U}(۰/۵, ۱)$ و $\varepsilon \sim \mathscr{N}(۰, ۰/۱)$ در نظر بگیریم و در هر مرحله به ترتیب تعداد $m = ۵(۱۰, ۲۰)$ داده پرت از مدل $Y = ۰/۵X + \varepsilon$ که $X \sim \mathscr{U}(۰/۷, ۱)$ و $\varepsilon \sim \mathscr{N}(۰, ۰/۰۵)$ باشند، را تولید و به داده‌ها اضافه کنیم و برای هر مرحله برآورد ضرایب رگرسیونی را از روش $MDPD$ محاسبه می‌کنیم. نتایج به‌دست‌آمده از این شبیه‌سازی و برای ۱۰۰ بار تکرار در جدول (۲) ارائه‌شده، که نشان می‌دهد برآوردگر ماکسیمم درستنمایی (متناظر با $\alpha = ۰$) حتی برای تعداد کم داده پرت به‌شدت تحت تأثیر قرار می‌گیرد. در صورتی که روش $MDPD$ با مقادیر $\alpha > ۰$ استواری بالایی در مواجهه با افزایش این داده‌های پرت از خود نشان داده و برآوردی استوار ارائه داده است.



شکل ۱: برآورد ضرایب رگرسیونی داده‌های تلفن بلژیک از روش میانه، روش ماکسیمم درستنمایی و روش $MDPD$ با چندین مقدار مختلف α

α	۰(MLE)	۰/۰۵	۰/۱	۰/۲۵	۰/۴	۰/۵	۰/۸	۱	LMS
$\beta_۰$	−۲۶/۰۱	−۲۵/۵۳	−۲۴/۹۴	−۲۱/۹۷	−۵/۲۴	−۵/۲۶	−۵/۳۱	−۵/۳۶	−۵/۶۱
$\beta_۱$	۰/۵۰	۰/۵۰	۰/۴۸	۰/۴۳	۰/۱۱	۰/۱۱	۰/۱۱	۰/۱۱	۰/۱۲
σ	۵/۳۸	۵/۴۰	۵/۴۱	۵/۲۹	۰/۱۱	۰/۱۱	۰/۱۲	۰/۱۲	− − −

جدول ۱: برآورد ضرایب رگرسیونی داده‌های تلفن بلژیک از روش $MDPD$ برای مقادیر مختلف α و برآورد از روش میانه

		$m=۵$	$m=۱۰$	$m=۲۰$
$MLE(\alpha=۰)$	$\beta_۰$	$۰/۰۲۱(۵/۷\times ۱۰^{-۴})$	$۰/۰۴۱(۱/۸\times ۱۰^{-۳})$	$۰/۰۶۶(۴/۵\times ۱۰^{-۳})$
	$\beta_۱$	$۰/۸۹۳(۱/۳\times ۱۰^{-۱})$	$۰/۷۹۹(۴/۱۳\times ۱۰^{-۲})$	$۰/۶۹۴(۹۴/۶\times ۱۰^{-۳})$
	$\sigma^۲$	$۱/۰۰(۸/۱\times ۱۰^{-۱})$	$۱/۰۰(۸/۱\times ۱۰^{-۱})$	$۱/۰۰(۸/۱\times ۱۰^{-۱})$
$\alpha=۰/۵$	$\beta_۰$	$۰/۰۰۰(۲/۱۲\times ۱۰^{-۴})$	$۰/۰۰۴(۲/۰۴\times ۱۰^{-۴})$	$۰/۰۱۲(۶/۹\times ۱۰^{-۴})$
	$\beta_۱$	$۰/۹۹۵(۱۸/۳\times ۱۰^{-۴})$	$۰/۹۸۵(۲/۲۹\times ۱۰^{-۳})$	$۰/۹۴۶(۱/۱\times ۱۰^{-۲})$
	$\sigma^۲$	$۰/۱۰۱(۰/۴۰\times ۱۰^{-۴})$	$۰/۱۰۲(۰/۶۵\times ۱۰^{-۴})$	$۰/۱۱۱(۲/۲\times ۱۰^{-۴})$
$\alpha=۱$	$\beta_۰$	$۰/۰۰۰(۲/۱۷\times ۱۰^{-۴})$	$-۰/۰۰۱(۲/۳۶\times ۱۰^{-۴})$	$۰/۰۰۲(۲/۵\times ۱۰^{-۴})$
	$\beta_۱$	$۰/۹۹۹(۱۶/۰\times ۱۰^{-۴})$	$۱/۰۰۱(۲۰/۳\times ۱۰^{-۴})$	$۰/۹۹۱(۲۲/۷\times ۱۰^{-۴})$
	$\sigma^۲$	$۰/۱۰۰(۰/۵۶\times ۱۰^{-۴})$	$۰/۱۰۲(۰/۵۲\times ۱۰^{-۴})$	$۰/۱۰۶(۰/۸۳\times ۱۰^{-۴})$

جدول ۲: برآورد ضرایب رگرسیونی و MSE(داخل پرانتز) از روش $MDPD$ با مقادیر مختلف داده پرت

بحث و نتیجه‌گیری

یکی از روش‌های متداول و کلاسیک در برآورد پارامترهای مدل رگرسیون خطی روش ماکسیمم درستنمایی می‌باشد. اما یکی از مشکلات مهم این روش این است که برآوردگرهایی که با استفاده از این روش به دست می‌آید تحت تأثیر داده‌های پرتی که احیاناً در داده‌ها وجود دارند قرارگرفته و استواری کمی از خود نشان می‌دهند. برای غلبه بر این مشکل در سال‌های اخیر روش‌های مختلفی مانند روش میانه و برآوردگرهای M معرفی گردیدند. این برآوردگرها هرچند استواری بالایی دارند اما در مقابل کارایی چندانی ندارند. در این پژوهش هدف معرفی نوع جدیدی از برآوردگرها به نام برآوردگرهای $MDPD$ در مدل‌های رگرسیون خطی است. برآوردگرهای $MDPD$ علاوه بر اینکه استواری مناسبی در مواجهه با داده‌های پرت از خود نشان می‌دهند کارایی بالایی دارند. در واقع برآوردگرهای $MDPD$ از بین برآوردگرهای استوار برآوردگری را انتخاب می‌کند که کارایی بالاتری از خود نشان دهد.

مراجع

[1] Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and Efficient Estimation by Minimizing a Density Power Divergence, *Biometrika*, **85**(3), 549-559.

[2] Durio, A. and Isaia, E. D. (2011). The Minimum Density Power Divergence Approach in Building Robust Regression Models, *Informatica*, **22** (1), 43-56.

[3] Ghosh, A. and Basu, A. (2013). Robust Estimation for Independent Non-homogeneous Observations Using Density Power Divergence with Applications to Linear Regression, *Electronic Journal of Statistics*, **7**, 2420-2456.

[4] Ghosh, A. and Basu, A. (2015). Robust Estimation for Non-homogeneous Data and the Selection of the Optimal Tuning Parameter: the Density Power Divergence Approach, *Journal of Applied Statistics*, **42**(9), 2056-2072.

[5] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, John Wiley and Sons, New York.