

SID



سرویس های ویژه



سرویس ترجمه تخصصی



کارگاه های آموزشی



بلاگ مرکز اطلاعات علمی



عضویت در خبرنامه



فیلم های آموزشی

کارگاه های آموزشی مرکز اطلاعات علمی جهاد دانشگاهی



مباحث پیشرفته یادگیری عمیق؛
شبکه های توجه گرافی
(Graph Attention Networks)



کارگاه آنلاین آموزش استفاده از
وب آو ساینس



کارگاه آنلاین مقاله روزمره انگلیسی

تشخيص نقاط پرت در مدل رگرسيوني ليو

فروغ حاجي بافري فروشاني، عبدالرحمن راسخ و محمدرضا آخوند

گروه آمار، دانشگاه شهيد چمران اهواز

تاريخ دريافت: ۱۳۹۲/۱۲/۳ تاريخ آخرين بازنگري: ۱۳۹۳/۳/۱۹

چکيده: در حضور هم خطي با ناپايدار بودن برآورد کمترين توان هاي دوم پارامترها، انتظار مي رود که باقي مانده ها هم ناپايدار باشند و در اين صورت ممکن است که یک باقي مانده بزرگ از برازش کمترين توان هاي دوم نمايان گر یک مشاهده پرت نباشد و برعکس. در اين صورت لزوم بررسي نقاط پرت هنگامی که از روش هاي معمول برآورد غير از کمترين توان هاي دوم از جمله برآورد گر ليو استفاده مي شود ضروري به نظر مي رسد. در اين مقاله با استفاده از روش انتقال ميانگين نقاط پرت، آماره آزمون لازم براي شناسايي اين نقاط به هنگام استفاده از برآورد گر ليو تعميم داده مي شود. در ادامه با استفاده از مجموعه داده اي واقعي کاربرد اين روش مورد ارزيابي قرار مي گيرد.

واژه هاي کليدي: برآورد گر ليو، نقاط پرت، هم خطي، روش انتقال ميانگين نقاط پرت.

۱ مقدمه

وجود هم خطي در ميان متغيرهاي پيشگو اثری جدی بر برآوردها و پيش بينی مي گذارد و ممکن است باعث شود که برآوردهاي کمترين توان هاي دوم از

آدرس الكترونيک مسئول مقاله: عبدالرحمن راسخ، rasekh_a@scu.ac.ir
کد موضوع بندي رياضي (۲۰۱۰): ۶۲J۲۰

مقادیر واقعی دور شوند. بنابراین برآوردگرهای اریب از جمله برآوردگرهای ریج و لیو^۱ برای کاهش این اثرات پیشنهاد شده‌اند (هورل و کنارد، ۱۹۷۰؛ لیو، ۱۹۹۳ و آکدیز و کسیرنلر، ۱۹۹۵).

از سوی دیگر بررسی وجود مشاهدات موثر^۲ و پرت به دلیل اثرات بسیار جدی آن‌ها بر برآوردگرها از سال‌ها قبل مورد توجه آماردانان قرار گرفته و مطالعه روش‌های تشخیص این مشاهدات در مدل‌های مختلف آماری انجام شده است (بلسلی و همکاران، ۱۹۸۰؛ کوک، ۱۹۸۶؛ چاترجی و هادی، ۱۹۸۶؛ والکر و برچ، ۱۹۸۸؛ شی، ۱۹۹۷ و شی و وانگ، ۱۹۹۹). در مقالات زیادی به این نکته توجه شده است که مشاهدات موثر و پرت به نوع برآوردگر پارامترها بستگی دارد. یعنی ممکن است مشاهدات مذکور تحت برآوردگر ریج از برآوردگر کمترین توان‌های دوم متفاوت باشند. تروسکی و همکاران (۱۹۸۰) مشاهدات پرت را در رگرسیون ریج با استفاده از روش انتقال میانگین نقاط پرت^۳ مورد مطالعه قرار دادند. والکر و برچ (۱۹۸۸) مشاهدات موثر را در برآوردگر رگرسیونی ریج معمولی بررسی کردند. بلسلی (۱۹۹۱) مشاهدات پرنفوذ^۴ را در رگرسیون ریج بررسی کرد. تروسکی و همکاران (۱۹۹۴) به تشخیص نقاط پرت در رگرسیون ریج و مؤلفه‌های اصلی در حضور هم‌خطی پرداختند. جاهوفر و چن (۲۰۰۹) مشاهدات موثر را در برآوردگر ریج اصلاح‌شده^۵ مطالعه کردند. به‌علاوه جاهوفر و چن (۲۰۱۱) مشاهدات موثر مکانی را در برآوردگر ریج اصلاح‌شده مطالعه کردند. همچنین آن‌ها در سال ۲۰۱۲ مشاهدات موثر مکانی را در برآوردگر لیو بررسی کردند. ارتاس و همکاران (۲۰۱۳) مشاهدات موثر را در برآوردگرهای لیو و لیو اصلاح‌شده به‌طور مختصر مورد مطالعه قرار دادند. همچنین جاهوفر (۲۰۱۳) به بررسی مشاهدات موثر در مدل رگرسیونی لیو پرداخته است.

۱ Ridge and Liu estimators
 ۲ Influential observations
 ۳ Mean shift outlier method
 ۴ High leverage observations
 ۵ Modified

هدف اصلی این مقاله مطالعه روش انتقال میانگین نقاط پرت در مدل رگرسیونی تحت برآوردگر لیو است. بر این اساس در بخش ۲ به معرفی مدل و برآوردگر لیو پرداخته می‌شود. در بخش ۳ روش انتقال میانگین نقاط پرت برای برآوردگر لیو تعمیم داده می‌شود و آماره مناسب برای آزمون نقاط پرت ارائه می‌گردد. در بخش ۴ پس از معرفی مجموعه داده‌ای واقعی مباحث نظری روی این داده‌ها پیاده می‌شود و در نهایت به بحث و نتیجه‌گیری پرداخته خواهد شد.

۲ مدل و برآوردگر لیو

مدل رگرسیون خطی

$$y = Z\gamma + \varepsilon \quad (1)$$

را در نظر بگیرید، که در آن بردار y بردار $n \times 1$ مشاهدات متغیر پاسخ، Z ماتریس $n \times p$ متغیرهای پیش‌گو، γ بردار $p \times 1$ ضرایب رگرسیونی و ε بردار $n \times 1$ خطاهای تصادفی با $E(\varepsilon) = 0$ و $Var(\varepsilon) = \sigma^2 I_n$ هستند. همچنین ممکن است فرض شود $\varepsilon \sim N(0, \sigma^2 I_n)$. فرم متعارف^۶ مدل (۱) به صورت

$$y = X\beta + \varepsilon \quad (2)$$

است، که در آن $X = ZT$ ، $\beta = T'\gamma$ و T ماتریس متعامدی است که ستون‌های آن بردارهای ویژه متعامد یک‌که ماتریس $Z'Z$ را تشکیل می‌دهد؛ بنابراین $X'X = T'Z'ZT = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ که $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ مقادیر ویژه $Z'Z$ هستند. در این صورت برآوردگر کمترین توان‌های دوم معمولی β برابر $\hat{\beta} = \Lambda^{-1} X'y$ است. همان‌طور که اشاره شد راهبردهای مختلفی از جمله برآوردگرهای ریج و لیو برای مواجهه با مسئله هم‌خطی پیشنهاد شده است. هورل و کنارد (۱۹۷۰) برآوردگر رگرسیونی ریج را به صورت

$$\hat{\beta}_k = (\Lambda + kI)^{-1} X'y$$

^۶ Canonical form

پیشنهاد دادند، که در آن k پارامتر اریبی این برآوردگر است.

لیو (۱۹۹۳) یک برآوردگر اریب جدیدی با استفاده از مزایای برآوردگرهای ریج و استاین معرفی کرد (هورل و کنار، ۱۹۷۰ و استاین، ۱۹۵۶). بدین معنی که به جای افزودن محدودیت^۷ ریج به فرم $\beta = k\beta + \varepsilon_0$ به مدل رگرسیونی، محدودیت جدید $d\beta = \beta + \varepsilon_0$ را با شرایط $E(\varepsilon_0) = 0$ ، $Var(\varepsilon_0) = \sigma^2 I_n$ و $E(\varepsilon_0 \varepsilon_0') = 0$ به صورت

$$\begin{pmatrix} y \\ d\hat{\beta} \end{pmatrix} = \begin{pmatrix} X \\ I \end{pmatrix} \beta + \begin{pmatrix} \varepsilon \\ \varepsilon_0 \end{pmatrix} \quad (۳)$$

اضافه کرد. برآوردگر لیو $\hat{\beta}_d$ با توجه به مدل تعمیم یافته (۳) به صورت

$$\hat{\beta}_d = (\Lambda + I)^{-1} (\Lambda + dI) \beta$$

پیشنهاد گردید، که در آن $0 < d < 1$ پارامتر اریبی برآوردگر لیو است. لیو (۱۹۹۳) برآوردهایی برای d با استفاده از روش‌های برآورد k در برآوردگر ریج ارائه داد. مزیت برآوردگر لیو به برآوردگر ریج این است که برآوردگر لیو تابعی خطی بر حسب پارامتر اریبی d است. بنابراین انتخاب این پارامتر راحت است. در مقالات اخیر به ویژه در زمینه اقتصاد، مهندسی و دیگر زمینه‌های آماری، برآوردگر لیو روش‌ها و ایده‌های جدیدی را ارائه می‌دهد (کسیرنلر و همکاران، ۱۹۹۹؛ آکدنیز و کسیرنلر، ۲۰۰۱؛ کسیرنلر و ساکالی‌اگلو، ۲۰۰۱؛ هابرت و ویجیکون، ۲۰۰۶؛ توریگو و یوجی، ۲۰۰۶؛ الهیتی و کبریا، ۲۰۰۹؛ لیو، ۲۰۱۱؛ لی و یانگ، ۲۰۱۲ و مانسون و همکاران، ۲۰۱۲).

۳ تشخیص نقاط پرت در گرسون خطی

این حقیقت که اگر مشاهده‌ای نقطه پرت باشد به این معنی است که وقتی مدل انتخابی به داده‌ها برازش داده می‌شود، مشاهده مذکور دارای باقی‌مانده بزرگی است؛ اما لزوماً این نکته استنباط نمی‌شود که مشاهده موثری نسبت به معادله برازش شده باشد (دراپر و جان، ۱۹۸۱). از سوی دیگر یک نقطه پرت ممکن است نتایج برازش

^۷ Restriction

مدل را به مقدار زیادی تحت تاثیر قرار دهد، به گونه‌ای که حذف آن از مجموعه داده‌ها نتایج کاملاً متفاوتی به بار آورد. به همین جهت شناسایی و تشخیص نقاط پرت و بررسی اثر آن‌ها بر روی جنبه‌های متفاوت یک تحلیل، برای یک تحلیل‌گر از اهمیت ویژه‌ای برخوردار است.

۱۳ آزمون نقاط پرت در روش کمترین توان‌های دوم

برای اینکه m مشاهده مشکوک از داده‌ها به‌عنوان مشاهده پرت مورد آزمون معنی‌داری قرار گیرند، ابتدا داده‌ها طوری مرتب می‌شوند که $n - m$ مشاهده پاک در ابتدا و m مشاهده مشکوک در انتها قرار گیرند. روش‌های مختلفی به‌منظور یافتن مقدار m و مشاهدات موجود در این مجموعه توسط آماردانان مختلف از جمله هادی (۱۹۹۲) و ریانی و اتکینسن (۲۰۰۰) پیشنهاد شده است. به این ترتیب می‌توان مدل انتقال میانگین نقاط پرت را برای رگرسیون خطی به صورت

$$y = X\beta + D_m\theta + \varepsilon \quad (۴)$$

تعریف کرد، که در آن $D_m = (0, I_m)'$ و I_m ماتریسی واحد متناظر با m مشاهده حذف شده، 0 ماتریس $(n - m) \times m$ با درایه‌های صفر و θ یک بردار m بعدی حاوی تغییرات مربوط به مشاهدات مشکوک است. $H = X\Lambda^{-1}X'$ ماتریس کلاسه‌دار^۸ مدل (۲) است که تحت فرض $\theta = 0$ به صورت

$$H = \begin{pmatrix} H_{n-m \times n-m} & H_{m \times n-m} \\ H_{n-m \times m} & H_{m \times m} \end{pmatrix}$$

افراز می‌شود. بر این اساس مانده‌های مدل (۲)، $(e = (I_n - H)y)$ مطابق با ماتریس کلاسه‌دار به صورت $e = (e'_{n-m}, e'_m)'$ افراز می‌شود. آماره آزمون فرضیه $\theta = 0 : H_0$ (یا $E(y) = X\beta$) در برابر $H_1 : \theta \neq 0$ (یا $E(y) = X\beta + D_m\theta$) توسط رانو و توتنبرگ (۱۹۹۹) و سیر و لی (۲۰۰۳) به صورت

$$F_m = \frac{e'_m (I_m - H_m)^{-1} e_m / m}{s_m^2 / (n - p - m)} \quad (۵)$$

^۸ Hat matrix

پیشنهاد شده است، که در آن $s_m^2 = e'e - e'_m(I_m - H_m)^{-1}e_m$ آماره (۵) دارای توزیع F با درجه‌های آزادی m ، $n - p - m$ و رد آزمون دلیلی بر وجود نقاط پرت در بین m مشاهده است. آماره آزمون پرت بودن مشاهده i ام، به‌ازای $m = 1$ به صورت

$$F_1 = \frac{(n - p - 1)(1 - h_i)^{-1}e_i^2}{s_i^2} \quad (6)$$

خواهد بود. وقتی e_i^2 بزرگ باشد، F_1 یک مقدار معنی دار بزرگ را ارائه می‌کند (رائو و توتنبرگ، ۱۹۹۹ و سبیر و لی، ۲۰۰۳).

۲.۳ آزمون نقاط پرت در روش رگرسیونی لیو

حضور هم‌خطی همراه با مشاهدات پرت دو مسئله‌ای است که باید به صورت هم‌زمان بررسی شوند. ناپایداری برآورد کمترین توان‌های دوم پارامترها تحت هم‌خطی ممکن است ناپایداری باقی مانده‌ها را به همراه داشته باشد. همچنین ممکن است یک مانده بزرگ حاصل از روش کمترین توان‌های دوم نمایان گر یک مشاهده پرت نباشد و برعکس (بلسلی و همکاران، ۱۹۸۰ و والکر و بروج، ۱۹۸۸). بنابراین حضور این مشاهدات همراه با وجود هم‌خطی میان متغیرهای پیش‌گو بسیار پیچیده می‌شود. بر این اساس بلسلی و همکاران (۱۹۸۰) تاکید کردند که هم‌خطی باید قبل از تلاش برای شناسایی مشاهدات پرت بررسی و کنترل شود. تروسکی و همکاران (۱۹۹۴ و ۱۹۸۰) تشخیص نقاط پرت را در حضور هم‌خطی با استفاده از برآوردگرهای ریج و مؤلفه‌های اصلی بررسی کردند و آماره لازم را ارائه دادند.

در این مقاله تشخیص مشاهدات پرت در رگرسیون لیو با استفاده از مدل انتقال میانگین بررسی می‌شود. به این صورت که همانند مورد مشابه در تحلیل کمترین توان‌های دوم، فرض کنید m مشاهده مشکوک هستند. این مشاهدات را در یک مجموعه گنجانده به صورتی که $n - m$ مشاهده پاک در ابتدا و m مشاهده مشکوک در ادامه آن‌ها قرار گیرند. در این صورت از مدل انتقال میانگین نقاط پرت (۴) در رگرسیون لیو استفاده می‌شود. ابتدا با روش دارپر و جان (۱۹۸۱) مدل (۳) تحت

فرض H_0 ، یعنی مشاهده پرت در بین m مشاهده مشکوک وجود نداشته باشد
($\theta = 0$)، به صورت

$$\begin{pmatrix} y_{n-m} \\ y_m \\ d\hat{\beta} \end{pmatrix} = \begin{pmatrix} X_{n-m} \\ X_m \\ I_p \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_{n-m} \\ \varepsilon_m \\ \varepsilon_0 \end{pmatrix} \quad (7)$$

نوشته می شود. اگر مدل (7) به صورتی بازنویسی شود که m مشاهده مشکوک در
آخر مشاهدات قرار گیرند، در این صورت مدل به صورت

$$\begin{pmatrix} y_d \\ y_m \end{pmatrix} = \begin{pmatrix} X_d \\ X_m \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_d \\ \varepsilon_m \end{pmatrix} \quad (8)$$

خواهد شد، که در آن $\varepsilon_d = (y'_{n-m}, d\hat{\beta})'$ و $X_d = (X'_{n-m}, I_p)'$ و $\varepsilon_0 = (\varepsilon'_{n-m}, \varepsilon'_0)'$

با در نظر گرفتن مدل (8) به صورت مدل آمیخته

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon} \quad (9)$$

که در آن $\tilde{y} = (y'_d, y'_m)'$ ، $\tilde{X} = (X'_d, X'_m)'$ ، $\tilde{\varepsilon} = (\varepsilon'_d, \varepsilon'_m)'$ و برآوردگر کمترین
توان های دوم تعمیم یافته برابر با $\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$ است که همان برآوردگر لیو را
نتیجه می دهد و بردار باقی مانده های تعمیم یافته به صورت

$$\begin{aligned} \tilde{\varepsilon} &= (I_{n+p} - \tilde{H})\tilde{y} \\ &= \begin{pmatrix} (I_{n+p-m} - \tilde{H}_d) & -\tilde{H}_{dm} \\ -\tilde{H}_{md} & (I_m - \tilde{H}_m) \end{pmatrix} \begin{pmatrix} y_d \\ y_m \end{pmatrix} = \begin{pmatrix} \tilde{\varepsilon}_d \\ \tilde{\varepsilon}_m \end{pmatrix} \quad (10) \end{aligned}$$

حاصل می شود، که در آن $\tilde{H}_{ab} = \tilde{X}_a(\tilde{X}'\tilde{X})^{-1}\tilde{X}'_b$ یک زیر ماتریس از ماتریس
کلاه دار تعمیم یافته $\tilde{H} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$ است. توجه شود که \tilde{H} و $\tilde{H}_{aa} = \tilde{H}_a$
ماتریسی متقارن و خودتوان است. بنابراین مجموع توان های دوم باقی مانده های
تعمیم یافته تحت فرض H_0 برابر است با:

$$\begin{aligned} \tilde{\varepsilon}'\tilde{\varepsilon} &= \tilde{y}'(I_{n+p} - \tilde{H})\tilde{y} \\ &= y'_d(I_{n+p-m} - \tilde{H}_d)y_d - y'_m\tilde{H}_{md}y_d - y'_d\tilde{H}_{dm}y_m + y'_m(I_m - \tilde{H}_m)y_m. \end{aligned}$$

از سوی دیگر اگر فرض شود که مشاهده پرت در بین m مشاهده مشکوک وجود دارد ($\theta \neq 0$) در این صورت با فرض H_1 مواجه هستیم، بنابراین مدل (۳) با در نظر گرفتن این فرض به صورت

$$\begin{pmatrix} y_{n-m} \\ y_m \\ d\hat{\beta} \end{pmatrix} = \begin{pmatrix} X_{n-m} & 0 \\ X_m & I_m \\ I_p & 0 \end{pmatrix} \begin{pmatrix} \beta \\ \theta \end{pmatrix} + \begin{pmatrix} \varepsilon_{n-m} \\ \varepsilon_m \\ \varepsilon_0 \end{pmatrix} \quad (11)$$

مطرح می شود، اگر مدل (۱۱) به گونه ای بیان شود که m مشاهده مشکوک در آخر مشاهدات نمونه قرار گیرند، مدل

$$\begin{pmatrix} y_d \\ y_m \end{pmatrix} = \begin{pmatrix} X_d & 0 \\ X_m & I_m \end{pmatrix} \begin{pmatrix} \beta \\ \theta \end{pmatrix} + \begin{pmatrix} \varepsilon_d \\ \varepsilon_m \end{pmatrix} \quad (12)$$

حاصل می شود. مدل (۱۲) به عنوان مدل انتقال میانگین نقاط پرت تحت محدودیت لیو به صورت

$$\tilde{y} = \tilde{X}\beta + \tilde{D}_m\theta + \tilde{\varepsilon} \quad (13)$$

در نظر بگیرید، که در آن $\tilde{y} = (y'_d, y'_m)'$ ، $\tilde{X} = (X'_d, X'_m)'$ ، $\tilde{\varepsilon} = (\varepsilon'_d, \varepsilon'_m)'$ ، $\tilde{D}_m = \begin{pmatrix} 0 \\ I_m \end{pmatrix}$ و 0 ماتریس صفر با بعد $(n+p-m) \times m$ است.

لم ۱: برآوردهای β ، θ و مجموع توان های دوم باقی مانده ها در مدل انتقال میانگین نقاط پرت (۱۳) به ترتیب عبارتند از

$$\tilde{\beta} = (X'_d X_d)^{-1} X'_d y_d \quad (1)$$

$$\tilde{\theta} = (I_m - \tilde{H}_m)^{-1} \tilde{e}_m \quad (2)$$

$$\tilde{r}' \tilde{r} = y'_d (I_{n+p-m} - \tilde{H}_d) y_d - y'_d \tilde{H}_{dm} (I_m - \tilde{H}_m)^{-1} \tilde{H}_{md} y_d \quad (3)$$

برهان: با روش کمترین توان های دوم داریم:

$$\begin{pmatrix} \tilde{\beta} \\ \tilde{\theta} \end{pmatrix} = \left(\begin{pmatrix} \tilde{X} \\ \tilde{D}_m \end{pmatrix}' (\tilde{X}, \tilde{D}_m) \right)^{-1} \begin{pmatrix} \tilde{X} \\ \tilde{D}_m \end{pmatrix}' \tilde{y}$$

$$\begin{aligned}
 &= \begin{pmatrix} \tilde{X}'\tilde{X} & \tilde{X}'\tilde{D}_m \\ \tilde{D}'_m\tilde{X} & \tilde{D}'_m\tilde{D}_m \end{pmatrix}^{-1} \begin{pmatrix} X'_d y_d + X'_m y_m \\ y_m \end{pmatrix} \\
 &= \begin{pmatrix} (X'_d X_d)^{-1} X'_d y_d \\ (I_m - \tilde{H}_m)^{-1} \tilde{e}_m \end{pmatrix}
 \end{aligned}$$

توجه شود که

$$\begin{aligned}
 (X'_d X_d)^{-1} &= (\tilde{X}'\tilde{X} - X'_m X_m)^{-1} \\
 &= (\tilde{X}'\tilde{X})^{-1} + (\tilde{X}'\tilde{X})^{-1} X'_m (I_m - \tilde{H}_m)^{-1} X_m (\tilde{X}'\tilde{X})^{-1}
 \end{aligned}$$

بنابراین

$$\begin{aligned}
 &(X'_d X_d)^{-1} X'_m y_m - (\tilde{X}'\tilde{X})^{-1} X'_m (I_m - \tilde{H}_m)^{-1} y_m \\
 &= (\tilde{X}'\tilde{X})^{-1} X'_m y_m + (\tilde{X}'\tilde{X})^{-1} X'_m (I_m - \tilde{H}_m)^{-1} X_m (\tilde{X}'\tilde{X})^{-1} X'_m y_m \\
 &- (\tilde{X}'\tilde{X})^{-1} X'_m (I_m - \tilde{H}_m)^{-1} y_m \\
 &= (\tilde{X}'\tilde{X})^{-1} X'_m y_m - (\tilde{X}'\tilde{X})^{-1} X'_m (I_m - \tilde{H}_m)^{-1} (I_m - \tilde{H}_m) y_m = 0.
 \end{aligned}$$

در نتیجه $\tilde{\beta} = (X'_d X_d)^{-1} X'_d y_d$ از سوی دیگر

$$\begin{aligned}
 \tilde{\theta} &= -(I_m - \tilde{H}_m)^{-1} X_m (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \tilde{y} + (I_m - \tilde{H}_m)^{-1} y_m \\
 &= -(I_m - \tilde{H}_m)^{-1} \left[X_m (\tilde{X}'\tilde{X})^{-1} X_d y_d + X_m (\tilde{X}'\tilde{X})^{-1} X_m y_m - y_m \right] \\
 &= (I_m - \tilde{H}_m)^{-1} \left[(I_m - \tilde{H}_m) y_m - \tilde{H}_m y_d \right] \\
 &= (I_m - \tilde{H}_m)^{-1} \tilde{e}_m.
 \end{aligned}$$

از طرفی با استفاده از برآوردگرهای β و θ ، بردار مقادیر برازش شده مدل انتقال میانگین نقاط پرت تحت محدودیت لیو با استفاده از مدل (۱۲) برابر است با

$$\begin{aligned}
 \begin{pmatrix} \hat{y}_d \\ \hat{y}_m \end{pmatrix} &= \begin{pmatrix} X_d & 0 \\ X_m & I_m \end{pmatrix} \begin{pmatrix} \tilde{\beta} \\ \tilde{\theta} \end{pmatrix} \\
 &= \begin{pmatrix} X_d (X'_d X_d)^{-1} X'_d & 0 \\ 0 & I_m \end{pmatrix} \begin{pmatrix} y_d \\ y_m \end{pmatrix}.
 \end{aligned}$$

بنابراین بردار باقی مانده‌های تعمیم‌یافته مدل (۱۲) به صورت

$$\tilde{r} = \begin{pmatrix} \tilde{r}_d \\ \tilde{r}_m \end{pmatrix} = \begin{pmatrix} I_{n+p-m} - X_d(X'_d X_d)^{-1} X'_d \\ 0 \end{pmatrix} \begin{pmatrix} y_d \\ y_m \end{pmatrix}. \quad (14)$$

حاصل می‌شود و مجموع توان‌های دوم باقی مانده‌ها در مدل انتقال میانگین نقاط پرت تحت محدودیت لیو به راحتی از رابطه (۱۴) به دست می‌آید.

حال اختلاف مجموع توان‌های دوم باقی مانده‌های تعمیم‌یافته در دو مدل (۹) و (۱۳) که به ترتیب تحت فرض‌های H_0 و H_1 مبنی بر عدم حضور مشاهده پرت در بین m مشاهده مشکوک و حضور آن برابر است با

$$\tilde{e}'_m \tilde{e}_m - \tilde{r}'_m \tilde{r}_m = \tilde{e}'_m (I_m - \tilde{H}_m)^{-1} \tilde{e}_m$$

لم ۲: تحت مدل (۱۲) با بردار خطاهای دارای توزیع نرمال و باقی مانده‌های \tilde{e}_m در رابطه (۱۰)، فرم درجه دوم $\tilde{e}'_m (I_m - \tilde{H}_m)^{-1} \tilde{e}_m / \sigma^2$ دارای توزیع χ^2 با m درجه آزادی است.

برهان: باقی مانده‌های تعمیم‌یافته تحت محدودیت لیو برابر است با:

$$\tilde{e} = (I_{n+p} - \tilde{H})\tilde{y} = (I_{n+p} - \tilde{H})(\tilde{X}\beta + \tilde{\varepsilon}) = (I_{n+p} - \tilde{H})\tilde{\varepsilon}. \quad (15)$$

در این صورت واضح است که $\tilde{e} \sim N_{n+p} \left(0, \sigma^2 (I_{n+p} - \tilde{H}) \right)$ حال بدون از دست رفتن کلیت مسئله، زیرمجموعه \tilde{e}_m که شامل m عضو انتهایی \tilde{e} است با توجه به (۱۵) به صورت

$$\tilde{e}_m = (0, I_m) \tilde{e} = (0, I_m) (I_{n+p} - \tilde{H})\tilde{\varepsilon} \quad (16)$$

نشان داده می‌شود، بنابراین $\tilde{e}_m \sim N_m \left(0, \sigma^2 (I_m - \tilde{H}_m) \right)$ از طرفی با در نظر گرفتن (۱۶) می‌توان نوشت:

$$\begin{aligned} \tilde{e}'_m (I_m - \tilde{H}_m)^{-1} \tilde{e}_m &= \tilde{\varepsilon}' (I_{n+p} - \tilde{H})' \begin{pmatrix} 0 \\ I_m \end{pmatrix} (I_m - \tilde{H}_m)^{-1} (0, I_m) \\ &\times (I_{n+p} - \tilde{H})\tilde{\varepsilon} = \tilde{\varepsilon}' M' M^* M \tilde{\varepsilon} = \tilde{\varepsilon}' E \tilde{\varepsilon}. \end{aligned}$$

که در آن $M = (I_{n+p} - \tilde{H})^{-1}$ ، $M^* = \begin{pmatrix} \circ & \circ \\ \circ & (I_m - \tilde{H}_m)^{-1} \end{pmatrix}$ و $E = M^* M M^*$ از سوی دیگر چون \tilde{H} خودتوان و متقارن است، M نیز خودتوان و متقارن است، بنابراین ماتریس E نیز خودتوان و دارای رتبه m است. در نتیجه با توجه به قضیه توزیع فرم‌های درجه دوم (سیبر و لی، ۲۰۰۳) اثبات کامل می‌شود.

لم ۳: اگر $\tilde{s}_m^2 = \tilde{e}'\tilde{e} - \tilde{e}'_m(I_m - \tilde{H}_m)^{-1}\tilde{e}_m$ آنگاه \tilde{s}_m^2/σ^2 دارای توزیع χ^2 با $n - m$ درجه آزادی است.

برهان: تعریف کنیم $\tilde{E} = \tilde{e}'E\tilde{e} = \tilde{e}'E^*\tilde{e}$ که در آن $E^* = M - MM^*M$ واضح است که E^* خودتوان و دارای رتبه $n - m$ است. بنابراین با استدلال مشابه با لم ۲ اثبات کامل می‌شود.

لم ۴: \tilde{s}_m^2 و $\tilde{e}'_m(I_m - \tilde{H}_m)^{-1}\tilde{e}_m$ مستقل هستند.

برهان: واضح است که $EE^* = 0$ ، بنابراین فرم‌های درجه دوم \tilde{s}_m^2 و $\tilde{e}'_m(I_m - \tilde{H}_m)^{-1}\tilde{e}_m$ مستقلند و با استفاده از قضیه استقلال آماری (سیبر و لی، ۲۰۰۳، قضیه ۲-۵) اثبات کامل می‌شود.

بنابراین آماره آزمون برای فرض $\theta = 0$ تحت محدودیت لیو به صورت

$$\tilde{F}_m = \left(\frac{n-m}{m} \right) \frac{\tilde{e}'_m(I_m - \tilde{H}_m)^{-1}\tilde{e}_m}{\tilde{s}_m^2} \quad (17)$$

پیشنهاد می‌شود که دارای توزیع F با درجه‌های آزادی m و $n - m$ است. رد آزمون دلیلی بر وجود نقاط پرت در بین m مشاهده هست. برای آزمون پرت بودن مشاهده i ام، هنگامی که $m = 1$ ، آماره (۱۷) به صورت

$$\tilde{F}_1 = \frac{\tilde{e}_i^2}{\tilde{s}_i^2 m_{ii}} \quad (18)$$

خواهد بود، که در آن m_{ii} ، i امین عضو روی قطر ماتریس $M = (I_{n+p} - \tilde{H})$ و $\tilde{s}_i^2 = (\tilde{e}'\tilde{e} - (\tilde{e}_i^2/m_{ii}))/ (n - 1)$ هستند.

۴ مثال کاربردی

هدف در این بخش صرفاً پیاده‌سازی مباحث نظری ارائه شده در بخش‌های قبیل بر روی داده‌هایی است که دارای شرایط مورد نیاز از جمله هم‌خطی هستند. لذا از داده‌های سیمان پورتلند (وودز و همکاران، ۱۹۳۲) مندرج در جدول ۱ استفاده شده است که علی‌رغم قدیمی بودن اغلب در مورد مسئله هم‌خطی توسط محققان مورد استفاده قرار گرفته است. این مجموعه داده محصول یک تحقیق تجربی از حرارت تکامل یافته در طول سخت شدن سیمان پورتلند و وابستگی این گرما به درصد چهار کلینکری که از آن سیمان تولید می‌شود، است.

جدول ۱: داده‌های سیمان پورتلند

y	x_4	x_3	x_2	x_1	
۷۸/۵	۶۰	۶	۲۶	۷	۱
۷۴/۳	۵۲	۱۵	۲۹	۱	۲
۱۰۴/۳	۲۰	۸	۵۶	۱۱	۳
۸۷/۶	۴۷	۸	۳۱	۱۱	۴
۵۹/۹	۳۳	۶	۵۲	۷	۵
۱۰۹/۲	۲۲	۹	۵۵	۱۱	۶
۱۰۲/۷	۶	۱۷	۷۱	۳	۷
۷۲/۵	۴۴	۲۲	۳۱	۱	۸
۹۳/۱	۲۲	۱۸	۵۴	۲	۹
۱۱۵/۹	۲۶	۴	۴۷	۲۱	۱۰
۸۳/۸	۳۴	۲۳	۴۰	۱	۱۱
۱۱۳/۳	۱۲	۹	۶۶	۱۱	۱۲
۱۰۹/۴	۱۲	۸	۶۸	۱۰	۱۳

وودز و همکاران (۱۹۳۲) مدل خطی بدون عرض از مبدأ را به این داده‌ها برازش دادند. دیگر محققان همچون هالد (۱۹۵۲)، گورمان و توماس (۱۹۶۶)، دانیل و وود (۱۹۸۰)، کسیرنلر و همکاران (۱۹۹۹) و ساکالی‌اگلو و کسیرنلر (۲۰۰۸) مدل خطی با عرض از مبدأ را به داده‌ها برازش دادند. تحت این مدل $n = ۱۳$ مشاهده اما $p = ۵$ ضریب رگرسیونی نامعلوم وجود دارد.

برای تشخیص وجود هم‌خطی از شاخص عدد شرطی استفاده می‌شود که از نسبت ریشه دوم بزرگ‌ترین مقدار ویژه (λ_{max}) ماتریس $Z'Z$ بر کوچک‌ترین مقدار ویژه (λ_{min}) آن به صورت $\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ به دست می‌آید (بلسلی و همکاران، ۱۹۸۰). در مثال حاضر عدد شرطی برابر $6053/3434$ است که نشان‌دهنده وجود هم‌خطی قوی است (بلسلی و همکاران، ۱۹۸۰)، لذا برآوردگر کمترین توان‌های دوم غیرقابل اطمینان و در نتیجه ناپایدار است. بنابراین لزوم استفاده از برآوردگرهای جایگزین از جمله برآوردگر لیو آشکار می‌شود. مقادیر مختلفی از برآورد پارامتر اریبی برآوردگر لیو برای این مجموعه داده توسط محققان پیشنهاد شده اما ساکالی اگلو و کسیرنلر (۲۰۰۸) پارامتر اریبی بهینه را $d = 0/61$ برآورد کردند. در جدول ۲ برآورد کمترین توان‌های دوم و لیو پارامترهای مدل همراه با انحراف معیار متناظرشان و همچنین مقدار احتمال محاسبه شده‌اند.

جدول ۲: برآورد کمترین توان‌های دوم و لیو پارامترها

مقدار احتمال	انحراف استاندارد	برآورد	پارامتر	روش برآورد
$<0/0001$	$0/0116$	$-1/6371$	β_0	کمترین توان‌های دوم
$0/0002$	$0/0317$	$-0/2099$	β_1	
$<0/0001$	$0/0859$	$-0/9160$	β_2	
$<0/0001$	$0/0238$	$-1/8401$	β_3	
$0/3995$	$70/0858$	$62/3713$	β_4	لیو
$0/0001$	$0/0116$	$-1/6371$	β_0	
$0/0002$	$0/0317$	$-0/2098$	β_1	
$0/0000$	$0/0859$	$-0/9156$	β_2	
$0/0001$	$0/0237$	$-1/8334$	β_3	
$0/3995$	$42/7856$	$38/0761$	β_4	

برای بررسی اینکه i آمین مشاهده پرت است یا خیر، به ترتیب از آماره‌های (۶) و (۱۸) تحت روش‌های کمترین توان‌های دوم و لیو استفاده می‌شود. همان‌طور که در جدول ۳ ملاحظه می‌شود در سطح ۵ درصد در روش کمترین توان‌های دوم با استفاده از آماره (۶) هیچ نقطه پرتی تشخیص داده نشد اما در روش رگرسیون لیو، با آماره (۱۸) مشاهدات ۶ و ۸ به‌عنوان نقطه پرت شناسایی شدند.

جدول ۳: آماره آزمون نقاط پرت در دو روش کمترین توان‌های دوم و لیو

برآورد			برآورد		
لیو	کمترین توان‌های دوم	مشاهده	لیو	کمترین توان‌های دوم	مشاهده
۶/۵۷	۳/۸۷	۸	۰/۰۳	۷/۴ * ۱۰ ^{-۷}	۱
۰/۸۶	۰/۴۱	۹	۱/۰۴	۰/۵۴	۲
۰/۰۳	۰/۰۴	۱۰	۰/۳۴	۰/۱۲	۳
۱/۴۲	۱/۱۸	۱۱	۰/۷۶	۰/۶۸	۴
۰/۱۷	۰/۱۹	۱۲	۰/۰۱	۰/۰۱	۵
۲/۴۲	۱/۳۱	۱۳	۶/۹۳	۴/۰۶	۶
			۰/۹۳	۰/۵۲	۷

بحث و نتیجه‌گیری

در این مقاله روش انتقال میانگین نقاط پرت برای شناسایی نقاط پرت پیشنهاد گردید. بر این اساس مشاهدات پرت متفاوتی در داده‌های سیمان پورتلند در دو روش کمترین توان‌های دوم و لیو بر اساس معیارهای ارائه‌شده تشخیص داده شد. به این صورت که با کاهش اثر هم‌خطی با به‌کارگیری برآورد لیو مشاهداتی پرت شناسایی شدند که در روش کمترین توان‌های دوم پرت تشخیص داده نشدند. بر پایه این حقیقت هم‌خطی باید قبل از بررسی و مطالعه مباحث تشخیصی کنترل شود. در قسمت کاربردی یک مقدار برای d تعیین شد و بر اساس آن برآورد لیو ضرایب رگرسیونی داده‌ها به دست آمد و با استفاده از این مقدار d مشاهدات پرت شناسایی شدند؛ اما با توجه به وابستگی این مقدار به هر مشاهده، ممکن است لازم باشد که هر مشاهده پرت برای مقادیر مختلف d به دست آید که می‌تواند در مطالعات بعدی مورد بررسی قرار گیرد.

تقدیر و تشکر

نویسندگان از پیشنهادهای ارزشمند داوران محترم که موجب ارتقای سطح کیفی مقاله شد، کمال تشکر را دارند.

مراجع

- Akdeniz, F. and Kaciranlar, S. (1995), On the Almost Unbiased Generalized Liu Estimator and Unbiased Estimation of the Bias and MSE, *Communications in Statistics-Theory and Methods*, **24**, 1789-1797.
- Akdeniz, F. and Kaciranlar, S. (2001), More on the New Biased Estimator in Linear Regression, *Indian Journal of Statistics*, **63**, 321-325.
- Alheety, M. I. and Kibria, B. M. G. (2009), On the Liu and Almost Unbiased Liu Estimators in the Presence of Multicollinearity with Heteroscedastic or Correlated Errors, *Mathematics and its Applications*, **4**, 155-167.
- Belsley, D. A. (1991), *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, John Wiley, New York.
- Belsley, D. A., Kuh, E. and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influence Data and Source of Collinearity*, John Wiley, New York.
- Cook, R. D. (1986), Assessment of Local Influence, *Royal Statistical Society*, **48B**, 133-169.
- Chatterjee, S. and Hadi, A. S. (1986), Influential Observations, High-leverage Points and Outliers in Linear Regression, *Statistical Science*, **1**, 379-416.
- Daniel, C. and Wood, S. (1980), *Fitting Equations to Data: Computer Analysis of Multifactor Data*, John Wiley, New York.
- Draper, N. R. and John, J. A. (1981), Influential Observations and Outliers in Regression, *Technometrics*, **23**, 21-26.

Ertas, A., Erisoglu, M. and Kaciranlar, S. (2013), Detecting Influential Observations in Liu and Modified Liu Estimators, *Applied Statistics*, **40**, 1735-1745.

Gorman, J. W. and Toman, R. J. (1966), Selection of Variables for Fitting Equations to Data, *Technometrics*, **8**, 27-51.

Hadi, A. S. (1992), Identifying Multiple Outliers in Multivariate Data, *Journal of Royal Statistical Society, B*, **54**, 761-771.

Hald, A. (1952), *Statistical Theory with Engineering Applications*, John Wiley, New York.

Hoerl, A. and Kennard, R. (1970), Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, **12**, 55-67.

Hubert, M. H. and Wijekoon, P. (2006), Improvement of the Liu Estimator in Linear Regression Model, *Statistical Papers*, **47**, 471-479.

Jahufer, A. (2013), Detecting Global Influential Observations in Liu Regression Model, *Open Journal of Statistics*, **3**, 5-11.

Jahufer, A. and Chen, J. B. (2009), Assessing Global Influential Observations in Modified Ridge Regression, *Statistics and Probability Letters*, **79**, 513-518.

Jahufer, A. and Chen, J. (2011), Measuring Local Influential Observations in Modified Ridge Regression, *Data Science*, **9**, 359-372.

Jahufer, A. and Chen, J. B. (2012), Identifying Local Influential Observations in Liu Estimator, *Metrika*, **75**, 425-438.

- Kaciranlar, S. and Sakallioğlu, S. (2001), Combining the Liu Estimator and the Principal Component Regression Estimator, *Communications in Statistics- Theory and Methods*, **30**, 2699-2706.
- Kaciranlar, S., Sakallioğlu, S., Akdeniz, F., Styan, G. P. H. and Werner, H. J. (1999), A New Biased Estimator in Linear Regression and a Detailed Analysis of the Widely-Analysed Dataset on Portland Cement, *The Indian Journal of Statistics, B*, **61**, 443-459.
- Li, Y. and Yang, H. (2012), A New Liu-Type Estimator in Linear Regression Model, *Statistical Papers*, **53**, 427-437.
- Liu, K. (1993), A New Class of Biased Estimate in Linear Regression, *Communications in Statistics-Theory and Methods*, **22**, 393-402.
- Liu, X. Q. (2011), Improved Liu Estimator in a Linear Regression Model, *Statistical Planning and Inference*, **141**, 189-196.
- Mansson, K., Kibria, B. M. and Shukur, G. (2012), On Liu Estimators for the Logit Regression Model, *Economic Modelling*, **29**, 1483-1488.
- Rao, C. A. and Toutenburg, H. (1999), *Linear Models: Least Squares and Alternative*, Springer, New York.
- Riani, M. and Atkinson, A. C. (2000), Robust Diagnostic Data Analysis: Transformations in Regression, *Technometrics*, **42**, 384-398.
- Sakallioğlu, S. and Kaciranlar, S. (2008), A New Biased Estimator Based on Ridge Estimator, *Statistical Papers*, **49**, 669-689.
- Seber, G. A. F. and Lee, A. J. (2003), *Linear Regression Analysis*, John Wiley, New Jersey.

Shi, L. (1997), Local Influence in Principle Component Analysis, *Biometrika*, **84**, 175-186.

Shi, L. and Wang, X. (1999), Local Influence in Ridge Regression, *Computational Statistics and Data Analysis*, **31**, 341-353.

Stein, C. (1956), Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 197-206.

Torigoe, N. and Ujiie, K. (2006), On the Restricted Liu Estimator in the Gauss-Markov Model, *Communications in Statistics-Theory and Methods*, **32**, 1713-1722.

Troskie, C. G., Chalton, D. O., Stewart, T. J. and Jacobs, M. (1994), Detection of Outliers and Influential Observations in Regression Analysis Using Stochastic Prior Information, *Communications in Statistics-Theory and Methods*, **23**, 3453-3476.

Troskie, C. G., Coutsourides, D. and Jacobs, M. (1980), Detection of Outliers in the Presence of Multicollinearity, *Technical Report No. 7, Department of Mathematical Statistics, University of Cape Town*.

Walker, E. and Birch, J. B. (1988), Influence Measure in Ridge Regression, *Technometrics*, **30**, 221-227.

Woods, H., Steinour, H. H. and Starke, H. R. (1932), Effect of Composition of Portland Cement on Heat Evolved During Hardening, *Industrial and Engineering Chemistry*, **24**, 1207-1214.

SID



سرویس های
ویژه



سرویس ترجمه
تخصصی



کارگاه های
آموزشی



بلاگ
مرکز اطلاعات علمی



عضویت در
خبرنامه



فیلم های
آموزشی

کارگاه های آموزشی مرکز اطلاعات علمی جهاد دانشگاهی



مباحث پیشرفته یادگیری عمیق؛
شبکه های توجه گرافی
(Graph Attention Networks)



کارگاه آنلاین آموزش استفاده از
وب آوساینس



کارگاه آنلاین مقاله روزمره انگلیسی