شناسایی مشاهدات پرت در مدل رگرسیونی ریج تحت محدودیت‌های خطی تصادفی

عبدالرحمن راسخ، بهزاد منصوری و نرگس هدایت پور

گروه آمار، دانشکده علوم ریاضی و کامپیوتر، دانشگاه شهید چمران اهواز

چکیده: در تحلیل رگرسیونی مطالعه مباحث تشخیصی شامل تعیین مشاهدات مؤثر و نقاط پرت از اهمیت ویژه‌ای برخوردار است. حساسیت روش کمترین توان‌های دوم نسبت به حضور مشاهدات مؤثر و داده‌های پرت در مدل موجب شد که گامی در جهت توسعه مباحث تشخیصی به منظور ارائه معیارهایی برای اندازه‌گیری تأثیر و شدت وابستگی به این مشاهدات برداشته شود. تعیین مشاهدات مؤثر و نقاط پرت در داده‌ها، زمانی که متغیرهای مستقل همخطی داشته باشند، بسیار پیچیده و مشکل است و خصوصاً اینکه حضور همخطی می‌تواند برخی از داده‌های غیرعادی را پوشش دهد. یکی از روش‌های مورد توجه برای تعیین مشاهدات پرت، روش انتقال میانگین است. در این مقاله، روش انتقال میانگین را برای برآوردگر ریج تحت محدودیت‌های خطی تصادفی؛ که به منظور کاهش اثر همخطی استفاده شده، تعمیم داده و برای این برآوردگر آماره آزمون جهت شناسایی مشاهدات پرت ارائه خواهد شد. در نهایت توانایی این روش را با استفاده از یک مثال کاربردی از داده‌های واقعی نشان داده می‌شود.

واژه‌های کلیدی: همخطی، رگرسیون ریج، رگرسیون ریج تحت محدودیت‌های خطی تصادفی، مشاهدات پرت، روش انتقال میانگین.

° آدرس الکترونیک مسئول مقاله: بهزاد منصوری b.mansouri@scu.ac.ir

°کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲J۰۷، ۶۲J۲۰.

۱ مقدمه

وجود همخطی در مدل مشکلاتی چون ناپایداری و عدم کیفیت در برآوردهای آماری را به همراه دارد و همچنین دلیلی بر عدم اعتبار برآورد کمترین توان‌های دوم است (بلزلی و همکاران، ۲۰۰۴)، مطالعات وسیعی راجع به مشکل همخطی و راهکارهای مواجه با آن انجام شده است (مونتوگومری و همکاران، ۲۰۰۱)، برای برخورد با اثرات نامطلوب همخطی راه‌های مختلفی چون به‌کارگیری برآوردگرهای آمیخته، ریج و ریج آمیخته پیشنهاد شده است (بلزلی و همکاران، ۲۰۰۴)، علاوه بر همخطی حضور مشاهدات پرت نیز می‌توانند اثرات بسیار جدی بر برآوردگرها اعمال کنند. گاهی اوقات زیرمجموعه کوچکی از داده‌ها تأثیر نامناسب و قابل ملاحظه‌ای را بر نتایج حاصل از تحلیل رگرسیونی اعمال می‌کنند. تعیین و ارزیابی میزان تأثیر این مشاهدات غیرعادی در مقوله مباحث تشخیصی، گنجانده می‌شود. یک عمل متعارف در تجزیه و تحلیل رگرسیونی، آزمون داده‌های پرت و تعیین این مشاهدات است. وجود هم‌زمان مشاهدات پرت و همخطی در مدل، مسئله‌ای نامعمول و غیرعادی نیست؛ بلکه حضور توأم این دو موضوع شرایط را پیچیده و مشکل می‌کند (لارنس و مارش، ۱۹۸۴)، حتی مسئله همخطی می‌تواند حالت هم‌پوشانی براي مشاهدات غیرعادی ایجاد کند (بلزلی و همکاران، ۲۰۰۴)، همچنین در مقالات زیادی به این نکته که ممکن است مشاهدات تأثیرگذار در برآورد ریج متمایز با برآورد کمترین توان‌های دوم باشد، توجه شده است (مونتوگومری و همکاران (۲۰۰۱)، والکر و بیرچ (۱۹۸۸)، استیک(۱۹۸۶))، روش‌های تشخیص این مشاهدات برای رگرسیون ریج براساس رویکرد حذف موردی توسط والکر و بیرچ (۱۹۸۸) ارائه شده است. استیک (۱۹۸۶) نشان داد که اثر مشاهدات با نفوذ بر برآورد ریج کمتر از اثری است که همان مشاهدات بر برآورد کمترین توان‌های دوم دارند. جیانژین و هایان (۱۹۹۵) روش انتقال میانگین را با هدف تعیین نقاط پرت برای رگرسیون ریج مورد مطالعه قرار دادند. به دنبال آن آماره آزمون جهت بررسی پرت بودن مشاهدات، تحت شرایط استفاده از رگرسیون ریج به صورت اطلاعات کمکی توسط تروسکی و همکاران (۱۹۹۴) ارائه گردید. شی (۱۹۹۷) و شی و وانگ (۱۹۹۹) به ترتیب به بررسی تحلیل تأثیر مکانی در روش مؤلفه‌های اصلی و روش رگرسیون ریج پرداختند. روش‌های تشخیصی موردی برای رگرسیون ریج اصلاح‌شده به وسیله جاهوفر و چن (۲۰۰۹) بسط داده شده است. همچنین جاهوفر و چن (۲۰۰۱) به مطالعه تحلیل تأثیر مکانی در رگرسیون ریج اصلاح‌شده نیز پرداختند. قپانی و همکاران (۲۰۱۵) نیز مباحث تشخیصی را در مدل‌های با خطا در اندازه‌گیری تحت برآورد ریج مورد مطالعه قرار دادند. در زمینه مباحث تشخیصی در مدل‌های رگرسیونی با برآورد ریج تحت محدودیت‌های خطی مطالعه‌ای انجام

نشده است. بر همین اساس، در این مقاله مباحث تشخیصی و به طور ویژه، آماره آزمون پرت بودن مشاهدات به مدل‌های رگرسیونی با برآورد ریج تحت محدودیت خطی تصادفی تعمیم‌یافته است. بر این اساس، در بخش ۲ پیش‌زمینه و تعریفی از برآورد ریج و ریج آمیخته ارائه داده می‌شود. در بخش ۳ روش تشخیصی انتقال میانگین در رگرسیون کمترین توان‌های دوم و ریج مطرح می‌شود و در پی آن این روش برای برآورد ریج آمیخته تعمیم داده خواهد شد. در بخش ۴ عملکرد روش پیشنهاد شده با استفاده از داده‌های مربوط به سیمان پورتلند نشان داده خواهد شد. در نهایت در بخش ۵ بحث و نتیجه‌گیری مطرح می‌شود.

۲ پیش‌زمینه و تعاریف اولیه

مدل رگرسیونی

$$Y = X\beta + \varepsilon \ , \ \ \varepsilon \sim \mathrm{N}(\circ, \sigma^{۲} I) \qquad (۱)$$

را در نظر بگیرید، که در آن Y یک بردار $n \times ۱$ از مشاهدات، X یک ماتریس $n \times p$ با رتبه p از متغیرهای مستقل، β یک بردار $p \times ۱$ از ضرایب رگرسیونی نامعلوم و ε یک بردار $n \times ۱$ از خطاهاست. برآورد ضرایب رگرسیونی با استفاده از روش متداول کمترین توان‌های دوم بر مبنای کمینه کردن مجموع توان دوم خطا به صورت $\hat{\beta} = (X'X)^{-۱}X'Y$, محاسبه می‌شود. بردار مانده‌های رگرسیون که به صورت تفاضل مقدار مشاهده شده از مقدار برآورد شده تعریف شده عبارت است از:

$$e = Y - X\hat{\beta} = Y - X(X'X)^{-۱}X'Y = (I_n - H)Y.$$

موقعیتی را که در آن بین متغیرهای مستقل وابستگی خطی وجود داشته باشد، تحت عنوان مسئله همخطی مطرح می‌کنند. به منظور رفع اثرات همخطی بر برآورد پارامترها راهکارهایی از جمله استفاده از برآوردهای اریب شامل برآوردهای ریج، استاین و لیو ارائه گردیده است. برآورد ریج توسط هورل و کنارد (۱۹۷۰) به صورت

$$\hat{\beta}(k) = (X'X + kI)^{-۱}X'Y \quad k > \circ, \qquad (۲)$$

پیشنهاد شد، که در آن k پارامتر اریبی است. رگرسیون ریج، توسط بسیاری از محققان مانند سویندل (۱۹۷۹)، ترنکلر (۱۹۸۴)، سینگ و همکاران (۱۹۸۶) و فلاح و سلام (۲۰۱۱) توسعه داده شده است. تروسکی و

همکاران (۱۹۹۴) به این موضوع که برآوردهای اریب از جمله برآورد ریج نوع خاصی از اطلاعات کمکی هستند اشاره کرده است. بر همین اساس برآورد ریج را از طریق افزودن نوع خاصی از محدودیت‌های تصادفی به مدل (۱)؛ تحت عنوان محدودیت ریج، به صورت

$$\circ = k^{\frac{1}{2}} I_p \beta + u \quad , \quad u \sim N(\circ, \sigma^2 I_p), \tag{۳}$$

پیشنهاد شده است که در آن $\circ$ و $k^{\frac{1}{2}} I_p$ به ترتیب بردار $p \times 1$ و ماتریس $p \times p$ و u نیز یک بردار تصادفی $p \times 1$ از خطاها با فرض $E(\varepsilon u') = \circ$ هستند. ساکالی‌اغلو و کسیرنلر (۲۰۰۸) بیان کردند که از طریق ادغام محدودیت (۳) با مدل (۱) مدل آمیخته

$$\begin{pmatrix} Y \\ \circ \end{pmatrix} = \begin{pmatrix} X \\ k^{1/2} I_p \end{pmatrix} \beta + \begin{pmatrix} \varepsilon \\ u \end{pmatrix},$$

را خواهیم داشت. با استفاده از روش کمترین توان‌های دوم تعمیم‌یافته، برآورد ضرایب که معادل با برآورد ریج است حاصل می‌شود. راه دیگر پاسخ‌گویی به مسئله همخطی، استفاده از اطلاعات کمکی موجود، افزون بر نمونه است. این روش توسط بلزلی و همکاران (۲۰۰۴) مطرح شده است. اغلب افزون بر مشاهدات نمونه (Y, X)، اطلاعات کمکی راجع به برخی از ضرایب رگرسیونی موجود است. باید توجه داشت که گاهی این اطلاعات دقیق نبوده و با خطا همراه هستند. در این حالت اطلاعات کمکی در قالب محدودیت‌های خطی تصادفی بر ضرایب رگرسیونی به صورت

$$r = R\beta + e \quad , \quad e \sim N(\circ, \sigma^2 I) \tag{۴}$$

تعریف می‌شوند، که در آن r و R به ترتیب بردار تصادفی $j \times 1$ و ماتریس $j \times p$ با رتبه و معلوم هستند و e بردار تصادفی از خطاها به طوری که $E(\varepsilon u') = \circ$ است. تیل و گولدبرگر (۱۹۶۱) و تیل (۱۹۶۳) بر مبنای نگرش دوربین (۱۹۵۳) راهبردی به منظور برآورد ضرایب رگرسیونی تحت محدودیت‌های خطی تصادفی ارائه دادند. تدبیر اندیشیده شده در راهبرد پیشنهادی، آمیختن اطلاعات غیرنمونه‌ای (۴) با مدل (۱) است. بر اساس این رویکرد آن‌ها برآورد آمیخته را به صورت

$$\hat{\beta}_R = \hat{\beta} + (X'X)^{-1} R' [I + R(X'X)^{-1} R']^{-1} (r - R\hat{\beta}),$$

پیشنهاد کردند. استفاده از روش رگرسیون ریج و اعمال محدودیت‌های خطی تصادفی به طور هم‌زمان، توسط لی و یانگ (۲۰۱۰) و اوزکال (۲۰۰۹) و مورد مطالعه قرار گرفته است. اوزکال (۲۰۰۹) از طریق ادغام محدودیت‌های خطی تصادفی (۴) با مدل (۱) و تحت محدودیت حداقل کردن فاصله برآورد از مقدار واقعی، برآورد ریج آمیخته را به صورت

$$\hat{\beta}_M(k) = (X'X + R'R + kI_P)^{-1}(X'Y + R'r) \quad k > 0, \tag{۵}$$

پیشنهاد داد.

۳ تشخیص نقاط پرت در مدل‌های رگرسیونی خطی

تشخیص مشاهدات غیرعادی یکی از مفاهیم با اهمیت و قابل توجه است. ویزبرگ (۱۹۸۳) روش‌ها و معیارهای ارائه شده برای تشخیص این نوع مشاهدات را به عنوان آنالیز تأثیر معرفی کرد. هدف اصلی آنالیز تأثیر، اندازه‌گیری میزان تغییرات به وجود آمده در جنبه‌های مختلف تحلیل در شرایط وجود مجموعه‌ای آشفته از مشاهدات است. زمانی که یك مشاهده تفاوت اساسی با دیگر مشاهدات داشته باشد، می‌تواند یك انحراف اساسی در نتایج تحلیل رگرسیونی ایجاد کند. علاقمندیم که این مجموعه از داده‌ها را، که به آنها نقاط پرت گفته می‌شود، تعیین و تأثیر آن‌ها را بر جنبه‌های مختلف تحلیل رگرسیونی ارزیابی می‌شود. نقاط پرت ممکن است نتایج برازش مدل را تحت تأثیر قرار دهند، به گونه‌ای که حذف آن‌ها از مجموعه داده‌ها نتایج کاملاً متفاوتی به بار آورد. روش‌ها و معیارهای مختلفی به منظور شناسایی چنین مشاهداتی پیشنهاد شده است. از جمله این روش‌ها می‌توان به روش حذف یک به یک مشاهدات اشاره نمود. برای بررسی دیگر روش‌ها می‌توان به ویزبرگ (۲۰۰۹)، بلزلی و همکاران (۲۰۰۴)، چاترجی و هادی (۱۹۸۶) و چاترجی و هادی (۱۹۸۶) مراجعه کرد.

۱.۳ روش انتقال میانگین در رگرسیون کمترین توان‌های دوم

روش انتقال میانگین برای تعیین مشاهدات پرت به کار می‌رود. اصول اولیه این روش مبتنی بر اضافه کردن متغیرهای کمکی به مدل است. در این روش با اضافه کردن متغیرهای کمکی به مشاهدات مشکوک و آزمون معنی‌دار بودن ضرایب رگرسیونی این متغیرها پرت بودن مشاهدات ارزیابی می‌شود. برای اینکه K مشاهده از داده‌ها به عنوان مشاهدات پرت مورد آزمون معنی‌داری قرار گیرند، ابتدا داده‌ها طوری مرتب می‌شوند که

$n-K$ مشاهده پاک در ابتدا و K مشاهده پرت ممکن در انتها قرار بگیرند. مقصود از مشاهدات پاک مجموعه مشاهدات مانده از n مشاهده پس از حذف داده‌هایی است که مشکوک به پرت بودن هستند. با هدف تعیین K و مشاهدات موجود در این مجموعه آماردانانی از جمله هادی (۲۰۰۰) و ریانی و اتکینسن (۱۹۹۲) راهکارهای مختلفی را پیشنهاد داده‌اند. به این ترتیب می‌توان مدل انتقال میانگین نقاط پرت را برای رگرسیون خطی به صورت

$$Y = X\beta + Q\gamma + \varepsilon, \qquad (۶)$$

تعریف کرد، که در آن $Q = \begin{pmatrix} 0 \\ I_K \end{pmatrix}$ و در آن I_K ماتریس همانی متناظر با K مشاهده حذف شده، O ماتریس $(n-K)\times K$ و γ یک بردار $K\times 1$ حاوی تغییرات مربوط به مشاهدات پرت ممکن است. به این ترتیب فرض‌های

$$H_0 : \gamma = 0 (E(Y) = X\beta) \text{ در مقابل } H_1 : \gamma \neq 0 (E(Y) = X\beta + Q\gamma), \qquad (۷)$$

برای آزمون پرت بودن K مشاهده در نظر گرفته می‌شوند، با توجه به اینکه مجموعه مشاهدات به دو بخش تفکیک شده است، ماتریس پیش‌بینی $H = X(X'X)^{-1}X'$ را می‌توان برای مدل رگرسیونی (۱) و تحت فرض صفر به صورت

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

افراز کرد. بردار مانده را می‌توان با استناد به رابطه $e = (I_n - H)Y$ به صورت $e = (e_{n-K}', e_K')'$ افراز کرد. به این ترتیب، آماره F به منظور آزمون فرض پرت بودن K مشاهده به صورت

$$F_K = \frac{e_K'(I_K - H_K)^{-1}e_K/K}{S_K^2/(n-p-K)} \qquad (۸)$$

پیشنهاد شده است (سبر و لی، ۲۰۰۲)، که در آن $S_K^2 = e'e - e_K'(I_K - H_{K\times K})^{-1}e_K$ آماره فوق دارای توزیع F با درجه‌های آزادی K و $(n-p-K)$ است. با هدف آزمون فرض بیان شده مبنی بر پرت بودن مشاهده iام، وقتی $K = 1$ است؛ آماره فوق به صورت

$$F_i = \frac{e_i^2}{(1-h_i)S_{(i)}^2} \qquad (۹)$$

در می‌آید. این آماره دارای توزیع F با درجه آزادی ۱ و $n-p-1$ است. وقتی e_i^2 بزرگ باشد، F_i یک مقدار معنی‌دار بزرگ را ارائه می‌کند (سبر و لی، ۲۰۰۲؛ و رائو و توتنبرگ، ۱۹۹۵)،

۲.۳ روش انتقال میانگین در رگرسیون ریج

تروسکی و همکاران (۱۹۹۴) بر مبنای اصول روش انتقال میانگین شیوه‌ای مطلوب باهدف شناسایی مشاهدات پرت تحت شرایط اعمال محدودیت‌های خطی تصادفی پیشنهاد کردند، که در آن مجموع توان دوم مانده‌های تعمیم‌یافته جایگزین مجموع توان دوم مانده‌ها می‌شود. جیانژین و هایان (۱۹۹۵) نیز روش انتقال میانگین تحت رگرسیون ریج را مورد مطالعه قرار دادند. فرض کنید K مشاهده مظنون به پرت بودن هستند (هادی، ۱۹۹۲)، با استناد بر روش‌های پیشنهادی، مجموعه مشاهدات به دو بخش K و $n-K$ مشاهده‌ای تقسیم می‌شوند. با فرض وجود همخطی در مدل، برای مقابله با آن، روش رگرسیون ریج به صورت محدودیت‌های (۳) اعمال می‌شود. بنابراین مدل آمیخته

$$\begin{pmatrix} Y_{n-K} \\ Y_K \\ 0 \end{pmatrix} = \begin{pmatrix} X_{n-K} \\ X_K \\ k^{\frac{1}{2}} I_p \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_{n-K} \\ \varepsilon_K \\ u \end{pmatrix} \qquad (۱۰)$$

را خواهیم داشت. اگر امکان پرت بودن K مشاهده موجود در بخش دوم داده‌ها وجود داشته باشد، با قرار دادن $n-K$ مشاهده پاک در ابتدا و K مشاهده مشکوک در انتهای مجموعه داده‌ها، مدل (۱۰) به صورت

$$\begin{pmatrix} Y_a \\ Y_K \end{pmatrix} = \begin{pmatrix} X_a \\ X_K \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_a \\ \varepsilon_K \end{pmatrix} \qquad (۱۱)$$

نگاشته می‌شود، که در آن $Y_a = \begin{pmatrix} Y_{n-K} \\ 0 \end{pmatrix}$ ، $X_a = \begin{pmatrix} X_{n-K} \\ k^{\frac{1}{2}} I_p \end{pmatrix}$ و $\varepsilon_a = \begin{pmatrix} \varepsilon_{n-K} \\ u \end{pmatrix}$ است. مدل (۱۱) را می‌توان به صورت

$$Y^* = X^* \beta + \varepsilon^* \qquad (۱۲)$$

بازنویسی کرد، که در آن $Y^* = \begin{pmatrix} Y_a \\ Y_K \end{pmatrix}$، $X^* = \begin{pmatrix} X_a \\ X_K \end{pmatrix}$ و $\varepsilon^* = \begin{pmatrix} \varepsilon_a \\ \varepsilon_K \end{pmatrix}$.

مدل تحت فرض پرت بودن K مشاهده که به عنوان مدل انتقال میانگین معرفی شده، عبارت است از:

$$\begin{pmatrix} Y_{n-K} \\ Y_K \\ \circ \end{pmatrix} = \begin{pmatrix} X_{n-K} & \circ \\ X_K & I_K \\ k^{\frac{1}{2}} I_p & \circ \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \varepsilon_{n-K} \\ \varepsilon_K \\ u \end{pmatrix}, \qquad (۱۳)$$

مدل آمیخته (۱۳) را می‌توان به صورت

$$Y^* = X^*\beta + Q_r\gamma + \varepsilon^* \qquad (۱۴)$$

بازنویسی کرد، که در آن $Q_r = \begin{pmatrix} \circ \\ I_K \end{pmatrix}$ و γ بردار پارامترهای انتقال هستند. تحت فرض صفر (پرت نبودن K مشاهده) تروسکی و همکاران (۱۹۹۴) مجموع توان‌های دوم بردار مانده‌های تعمیم‌یافته برای مدل‌های (۱۲) و (۱۴) تحت نادرست بودن فرض پرت بودن K مشاهده را به ترتیب با روابط (۱۵) و (۱۶) به صورت

$$\tilde{r}_n{}'\tilde{r}_n = Y^{*'}[I_{n+p} - X^*(X^{*'}X^*)^{-1}X^*]Y^*, \qquad (۱۵)$$

$$\tilde{r}'\tilde{r} = Y_a{}'[I_{n+p-K} - X_a(X_a{}'X_a)^{-1}X_a']Y_a, \qquad (۱۶)$$

محاسبه کردند. آن‌ها به منظور آزمون فرض $H_\circ : \gamma = \circ$ در مقابل $H_1 : \gamma \neq \circ$ تفاوت بین مجموع توان دوم مانده‌های تعمیم‌یافته تحت هر دو مدل (۱۲) و (۱۴) را به صورت

$$\tilde{r}_n{}'\tilde{r}_n - \tilde{r}'\tilde{r} = \tilde{r}_K{}'(I_K - Z_{KK})^{-1}\tilde{r}_K$$

تعیین کردند، که در آن

$$I - Z = \begin{bmatrix} I_{n+p-K} - X_a\left(X^{*'}X^*\right)^{-1}X_a' & -X_a\left(X^{*'}X^*\right)^{-1}X_K' \\ -X_K\left(X^{*'}X^*\right)^{-1}X_a' & I_K - X_K\left(X^{*'}X^*\right)^{-1}X_K' \end{bmatrix}.$$

و $Z_{ij} = X_i(X'X + kI_p)^{-1}X_j' \quad i,j = a,K$ زیرماتریس‌هایی از ماتریس Z هستند. آنها همچنین آماره

$$F_K^* = \frac{\tilde{r}_K{}'(I_K - Z_{KK})^{-1}\tilde{r}_K(n-K)}{KS_{-K}^2} \qquad (۱۷)$$

را برای بررسی فرض پرت بودن K مشاهده ارائه کردند، که در آن

$$S^{۲}_{-K} = \tilde{r}_n'(I - Z)^{-۱}\tilde{r}_n - \tilde{r}_K'(I_K - Z_{KK})^{-۱}\tilde{r}_K.$$

آماره مذکور دارای توزیع F با درجه آزادی K و $n - K$ است.

۳.۳ روش انتقال میانگین در رگرسیون ریج تحت محدودیت‌های خطی تصادفی

حضور همخطی موجب ناکارآمدی و ناپایداری برآورد ضرایب گشته و همچنین انتظار می‌رود که ناپایداری مانده‌ها را نیز در پی داشته باشد. بر همین اساس، در چنین مواقعی مانده‌ها در مباحث تشخیصی از کارایی لازم برخوردار نخواهد بود و این امکان وجود دارد که نتایج به دست آمده از معیارهای تشخیصی و شناسایی نقاط پرت بر مبنای مانده‌ها نادرست یا گمراه کننده باشد. به این ترتیب، بررسی مشاهدات پرت و بسط روش‌های تشخیص این نقاط در شرایط به‌کارگیری برآورد ریج آمیخته به عنوان یک گام اولیه برای کاهش این پیامدها ضرورت پیدا می‌کند. در این قسمت روش مذکور برای تشخیص نقاط پرت برای رگرسیون ریج تحت محدودیت‌های خطی تصادفی بسط داده می‌شود. حال با افزودن محدودیت‌های خطی تصادفی (۴) به مدل رگرسیونی (۱۰)، مدل آمیخته

$$\begin{pmatrix} Y_{n-K} \\ Y_K \\ ۰ \\ r \end{pmatrix} = \begin{pmatrix} X_{n-K} \\ X_K \\ k^{\frac{۱}{۲}} I_p \\ R \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_{n-K} \\ \varepsilon_K \\ u \\ e \end{pmatrix} \qquad (۱۸)$$

را خواهیم داشت. مشابه قبل اگر بخواهیم مدل (۱۸) به گونه‌ای بازآرایی می‌شود که K مشاهده مشکوک در انتهای مجموعه داده‌ها قرار گیرند، آنگاه با مدل

$$\begin{pmatrix} Y_r \\ Y_K \end{pmatrix} = \begin{pmatrix} X_r \\ X_K \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_r \\ \varepsilon_K \end{pmatrix} \qquad (۱۹)$$

مواجه هستیم، که در آن $Y_r = \begin{pmatrix} Y_{n-K} \\ r \\ ۰ \end{pmatrix}$ ، $X_r = \begin{pmatrix} X_{n-K} \\ R \\ k^{\frac{۱}{۲}} I_p \end{pmatrix}$ و $\varepsilon_r = \begin{pmatrix} \varepsilon_{n-K} \\ e \\ u \end{pmatrix}$.

مدل (۱۹) را می‌توان به صورت

$$Y^{**} = X^{**}\beta + \varepsilon^{**} \qquad (۲۰)$$

بازنویسی کرد. برآورد کمترین توان‌های دوم تعمیم‌یافته تحت مدل آمیخته (۲۰) برابر

$$\hat{\beta}^{**} = (X^{**'}X^{**})^{-۱}X^{**'}Y^{**}$$

است، که معادل با برآورد ریج آمیخته در رابطه (۵) است. بردار مانده تعمیم‌یافته برای مدل (۲۰) براساس برآورد ریج تحت محدودیت‌های تصادفی به صورت

$$\begin{aligned} \tilde{r}_n^* &= Y^{**} - X^{**}\hat{\beta}_M(k) = (I_{n+p+j} - Z^*)Y^{**} \\ &= \begin{bmatrix} (I_{n+p+j-K} - Z_{rr}^*)\, Y_r - Z_{rK}^* Y_K \\ -Z_{Kr}^* Y_r + (I_K - Z_{KK}^*)\, Y_K \end{bmatrix} = \begin{pmatrix} \tilde{r}_{n-K}^* \\ \tilde{r}_K^* \end{pmatrix} \end{aligned}$$

حاصل می‌شود، که در آن $Z^* = X^{**}(X^{**'}X^{**})^{-۱}X^{**'}$ ماتریس برازش کمترین توان‌های دوم تحت مدل (۲۰) است و زیرماتریس‌های Z_{ij}^* به صورت

$$Z_{ij}^* = X_i^{**}(X^{**'}X^{**})^{-۱}X_j^{**'} \qquad i,j = r,K,$$

تعریف می‌شوند. شایان ذکر است که ماتریس‌های Z^* و Z_{rr}^* خودتوان هستند. مجموع توان دوم مانده‌های تعمیم‌یافته برای مدل (۲۰) عبارت است از:

$$\begin{aligned} \tilde{r}_n^{*'}\tilde{r}_n^* &= Y^{**'}(I_{n+p+j} - Z^*)Y^{**} \\ &= Y_r'(I_{n+j+p-k} - Z_{rr}^*)Y_r - Y_K'Z_{Kr}^*Y_r - Y_r'Z_{rK}^*Y_K + Y_K'(I_K - Z_{KK}^*)Y_K. \end{aligned}$$

اکنون اگر فرض شود که به پرت بودن K مشاهده مظنون هستیم. آنگاه مدل انتقال میانگین تحت برآورد ریج آمیخته و فرض H_1 به صورت

$$\begin{pmatrix} Y_{n-K} \\ Y_K \\ \circ \\ r \end{pmatrix} = \begin{pmatrix} X_{n-K} & \circ \\ X_K & I_K \\ k^{\frac{1}{2}} I_p & \circ \\ R & \circ \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \varepsilon_{n-K} \\ \varepsilon_K \\ u \\ e \end{pmatrix},$$

نگاشته می‌شود. اگر مدل فوق را به گونه‌ای بنویسیم که $n-K$ مشاهده پاک در ابتدا و K مشاهده مشکوک در آخر مشاهدات نمونه قرار گیرند، مدل

$$\begin{pmatrix} Y_r \\ Y_K \end{pmatrix} = \begin{pmatrix} X_r & \circ \\ X_K & I_K \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \varepsilon_r \\ \varepsilon_K \end{pmatrix} \qquad (۲۱)$$

حاصل می‌شود. مدل (۲۱) را می‌توان به صورت

$$Y^{**} = X^{**}\beta + Q_r^*\gamma + \varepsilon^{**} \qquad (۲۲)$$

بازنویسی کرد. در اینجا $Q_r^* = \begin{pmatrix} \circ \\ I_K \end{pmatrix}$ و γ به عنوان بردار پارامتر انتقال معرفی می‌شود. به منظور محاسبه برآورد کمترین توان‌های دوم تعمیم‌یافته ضرایب و مجموع توان دوم مانده تعمیم‌یافته تحت مدل (۲۲) لم زیر بیان و اثبات می‌شود.

لم ۱. برآوردهای β و γ، مجموع توان دوم مانده‌ها در مدل انتقال میانگین نقاط پرت تحت محدودیت تصادفی ریج در رابطه (۲۲) به ترتیب عبارتند با

$$\begin{aligned} \hat{\beta}^* &= (X_r{}'X_r)^{-1}X_r{}'Y_r, \\ \hat{\gamma}^* &= (I_K - Z_{KK})^{-1}\tilde{r}_K^*, \\ \tilde{r}^{*'}\tilde{r}^* &= Y_r{}'[I_{n+j+p-k} - X_r(X_r{}'X_r)^{-1}X_r{}']Y_r \\ &= Y_r{}'(I_{n+j+p-k} - Z_{rr}^*)Y_r - Y_r{}'Z_{rK}^*(I_K - Z_{KK}^*)^{-1}Z_{Kr}^*Y_r. \end{aligned}$$

برهان. با استفاده از روش کمترین توان‌های دوم داریم:

$$\begin{pmatrix} \hat{\beta}^* \\ \hat{\gamma}^* \end{pmatrix} = \left[\begin{pmatrix} X^{**} \\ Q_r^* \end{pmatrix}' \begin{pmatrix} X^{**} & Q_r^* \end{pmatrix} \right]^{-1} \begin{pmatrix} X^{**} \\ Q_r^* \end{pmatrix}' Y^{**}$$
$$= \begin{bmatrix} (X_r' X_r)^{-1} X_r' Y_r \\ (I_K - Z_{KK})^{-1} \tilde{r}_K^* \end{bmatrix}.$$

با توجه به اینکه

$$(X_r{}' X_r)^{-1} = (X^{**'} X^{**} - X_K{}' X_K)^{-1}$$
$$= (X^{**'} X^{**})^{-1} + (X^{**'} X^{**})^{-1} X_K{}' (I_K - Z_{KK}^*)^{-1} X_K (X^{**'} X^{**})^{-1},$$

می‌توان نتیجه گرفت که:

$$(X_r{}' X_r)^{-1} X_K{}' Y_K - (X^{**'} X^{**})^{-1} X_K{}' (I_K - Z_{KK}^*)^{-1} Y_K = 0$$

به منظور محاسبه بردار مانده تعمیم‌یافته، محاسبه بردار مقادیر برازش شده ضرورت می‌یابد. به این ترتیب بر طبق محاسبات جبری و با جایگذاری مقادیر $\hat{\beta}^*$ و $\hat{\gamma}^*$ داریم:

$$\hat{Y}^{**} = \begin{pmatrix} \hat{Y}_r \\ \hat{Y}_K \end{pmatrix} = \begin{pmatrix} X_r & 0 \\ X_K & I_K \end{pmatrix} \begin{pmatrix} \hat{\beta}^* \\ \hat{\gamma}^* \end{pmatrix}$$
$$= \begin{bmatrix} X_r (X_r' X_r)^{-1} X_r' & 0 \\ 0 & I_K \end{bmatrix} \begin{pmatrix} Y_r \\ Y_K \end{pmatrix},$$

بنابراین بردار مانده تعمیم‌یافته عبارت است از:

$$\tilde{r}^* = \begin{bmatrix} I_{n+j+p-K} - X_r (X_r{}' X_r)^{-1} X_r' & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} Y_r \\ Y_K \end{pmatrix} = \begin{pmatrix} \tilde{r}_1^* \\ \tilde{r}_2^* \end{pmatrix},$$

همچنین مجموع توان دوم بردار مانده تعمیم‌یافته مطابق آنچه که ذکر شد، محاسبه می‌شود. به منظور انجام این آزمون، تعیین تفاضل بین مجموع توان دوم مانده‌های تعمیم‌یافته تحت فرض صفر که مبنی بر عدم وجود

مشاهده پرت است و مجموع توان‌های دوم مانده‌های تعمیم‌یافته تحت فرض مقابل آن، ضرورت دارد. بر همین اساس خواهیم داشت:

$$\tilde{r}_n^{*'}\tilde{r}_n^* - \tilde{r}^{*'}\tilde{r}^* = \tilde{r}_K^{*'}(I_K - Z_{KK}^*)^{-1}\tilde{r}_K^*,$$

لم ۲. اگر تعریف شود

$$S_{-K}^{*2} = \tilde{r}_n^{*'}(I - Z^*)^{-1}\tilde{r}_n^* - \tilde{r}_K^{*'}(I_K - Z_{KK}^*)^{-1}\tilde{r}_K^*,$$

آنگاه S_{-K}^{*2}/σ^2 و $\tilde{r}_K^{*'}(I_K - Z_{KK}^*)^{-1}\tilde{r}_K^*/\sigma^2$ دارای توزیع کای دو به ترتیب با درجات آزادی $(n+j-K)$ و K هستند.

برهان. فرض کنید $\tilde{r}_n^* = (I - Z^*)(X^{**}\beta + \varepsilon^{**}) = (I - Z^*)\varepsilon^{**}$ آنگاه خواهیم داشت:

$$\tilde{r}_n^* \sim N_{n+j+p}(0, \sigma^2 I_{n+j+p}(I - Z^*)),$$

همچنین با توجه به این موضوع که فرض کرده‌ایم K مشاهده پرت در انتهای مجموعه داده‌ها قرار دارد و بدون از دست رفتن کلیت مسئله می‌توان نوشت:

$$\tilde{r}_K^* = (0, I_K)\tilde{r}_n^* = (0, I_K)(I_{n+j+p} - Z^*)\varepsilon^{**}. \qquad (۲۳)$$

در نتیجه مشابه قبل استنباط می‌شود که:

$$\tilde{r}_K^* \sim N_K(0, \sigma^2 I_K(I_K - Z_{KK}^*)),$$

از سوی دیگر با استناد به (۲۳) فرم درجه دوم $\tilde{r}_K^*$ به صورت

$$\begin{aligned}
&\tilde{r}_K^{*'}(I_K - Z_K^*)^{-1}\tilde{r}_K^*/\sigma^2 \\
&= \sigma^{-2}[\varepsilon^{**'}(I - Z^*)' \begin{pmatrix} 0 \\ I_K \end{pmatrix} (I_K - Z_{KK}^*)^{-1}(0\ ,\ I_K)(I - Z^*)\varepsilon^{**}] \\
&= \sigma^{-2}\varepsilon^{**\prime}\tilde{M}'\tilde{M}^*\tilde{M}\varepsilon^{**} = \sigma^{-2}\varepsilon^{**\prime}\tilde{N}\varepsilon^{**},
\end{aligned}$$

تعریف می‌شود. همچنین S_{-K}^{*2} با توجه به (۲۳) به فرم

$$\begin{aligned}
S_{-K}^{*2} &= \tilde{r}_n^{*'}(I - Z^*)^{-1}\tilde{r}_n^* - \tilde{r}_K^{*'}(I_K - Z_{KK}^*)^{-1}\tilde{r}_K^* \\
&= \varepsilon^{**\prime}(\tilde{M} - \tilde{M}'\tilde{M}^*\tilde{M})\varepsilon^{**} = \tilde{\varepsilon}_r{}'\tilde{N}^*\tilde{\varepsilon}_r,
\end{aligned}$$

نگاشته می‌شود، که در آن $\tilde{M} = (I - Z^*)$، $\tilde{M}^* = \begin{bmatrix} 0 & 0 \\ 0 & (I_K - Z^*_{KK})^{-1} \end{bmatrix}$، $\tilde{N} = \tilde{M}'\tilde{M}^*\tilde{M}$ و $\tilde{N}^* = (\tilde{M} - \tilde{M}'\tilde{M}^*\tilde{M})$. با استناد به خودتوان و متقارن بودن ماتریس Z^* استنتاج می‌شود که $\tilde{N}\tilde{N} = \tilde{N}$ و $\tilde{N}^*\tilde{N}^* = \tilde{N}^*$. این نتیجه دلالت بر خودتوان بودن ماتریس‌های $\tilde{N}$ و $\tilde{N}^*$ داشته و با توجه به اینکه $tr\left(\tilde{N}\right) = K$ و $tr\left(\tilde{N}^*\right) = n + j - K$، براساس قضیه ۷-۲ صفحه ۲۸ کتاب سبر و لی (۲۰۰۲) استنباط می‌شود که $\tilde{r}^{*'}_K(I_K - Z^*_{KK})^{-1}\tilde{r}^*_K/\sigma^2$ و S^{*2}_{-K}/σ^2 به ترتیب دارای توزیع χ^2_K و χ^2_{n+j-K} هستند.

لم ۳. متغیرهای تصادفی $\tilde{r}^{*'}_K(I_K - Z^*_{22})^{-1}\tilde{r}^*_K$ و S^{*2}_{-K}/σ^2 از هم مستقل هستند.

برهان. با توجه به این نکته که $\tilde{N}^*\tilde{N} = O$ استقلال توزیع $\tilde{r}^{*'}_K(I_K - Z^*_{KK})^{-1}\tilde{r}^*_K$ و S^{*2}_{-K} براساس مثال ۱۲-۲ صفحه ۲۹ کتاب سبر و لی (۲۰۰۲) استنباط می‌شود.

بر پایه نتایج حاصل از لم‌های مذکور

$$F^{**}_K = \frac{\tilde{r}^{*'}_K(I_K - Z^*_{KK})^{-1}\tilde{r}^*_K(n + j - K)}{KS^{*2}_{-K}} \qquad (۲۴)$$

آماره را برای آزمون فرض بیان شده مبتنی بر بررسی پرت بودن K مشاهده ارائه می‌دهیم. آماره F^{**}_K دارای توزیع F با درجه آزادی K و $(n + j - K)$ است. رد این آزمون بر پرت بودن مشاهدات تحت بررسی دلالت می‌کند. در شرایطی که قصد داشته باشیم فرض پرت بودن مشاهده iام یعنی $K = 1$ آزمون شود، آنگاه آماره (۲۴) به صورت

$$F^{**}_i = \frac{\tilde{r}^{*2}_i(n + j - 1)}{S^{*2}_{-i}m^*_{ii}}$$

درمی‌آید، که در آن $S^{*2}_{-i} = (\tilde{r}^{*'}_n(I - Z^*)^{-1}\tilde{r}^*_n) - (\tilde{r}^{*2}_i/m^*_{ii})$ و m^*_{ii} درایه iام روی قطر ماتریس $(I_{n+j+p} - Z^*)$ است.

۴ مثال کاربردی

به منظور نشان دادن چگونگی کاربرد روش ارائه شده، مباحث نظری روی داده‌هایی که قبلاً مدل‌بندی شده‌اند، پیاده خواهد شد. لذا داده‌های مربوط به سیمان پورتلند معروف به داده‌های هالد (۱۹۵۲) را مورد استفاده قرار می‌دهیم. این داده‌ها به طور گسترده توسط آماردانان زیادی از جمله وود و همکاران (۱۹۳۲)، کسیرنلر

و همکاران (۱۹۹۹) و ساکالواغلو و کسیرنلر (۲۰۰۸) مورد استفاده قرار گرفته‌اند. همچنین در مقالات و مطالعات مربوط به مباحث تشخیصی به دلیل داشتن شرایط مورد نظر از این داده‌ها به دفعات بهره‌برداری شده است. این مجموعه از داده‌ها حاصل یک تحقیق تجربی مربوط به گرمای تولید شده در طول فرآیند سفت شدن سیمان پرتلند است. گرمای حاصل شده وابسته به درصد چهار ماده مرکب موجود در سیمان تولید شده است. ارتباط این چهار ماده مرکب و تأثیر آن‌ها بر میزان گرمای حاصل شده از اهمیت ویژه‌ای برخوردار بوده و در فرایند تولید سیمان تأثیرات قابل ملاحظه‌ای دارد. در ابتدا با هدف تحلیل مدل رگرسیونی، اقدام به محاسبه برآورد ضرایب براساس روش کمترین توان‌های دوم می‌شود. نتایج در جدول ۱ ارائه شده است.

جدول ۱: نتایج حاصل از آنالیز کمترین توان‌های ممکن.

متغیرهای مستقل	برآورد ضرایب	خطای استاندارد	t آماره‌ای	مقدار احتمال
عرض از مبدأ	۶۲٫۴۰۵۴	۷۰٫۰۷۱۰	۰٫۸۹۱	۰٫۳۹۹
X_1	۹٫۱۲۴۲	۴٫۳۸۱۰	۲٫۰۸۳	۰٫۰۷۱
X_2	۷٫۹۳۸۷	۱۱٫۲۶۲۸	۰٫۷۰۵	۰٫۵۰۱
X_3	۰٫۶۵۲۷	۴٫۸۳۴۰	۰٫۱۳۵	۰٫۸۹۶
X_4	−۲٫۴۱۱۳	۱۱٫۸۶۸۲	−۰٫۲۰۳	۰٫۸۴۴

با توجه به نتایج موجود در جدول ۱، بدیهی است که کوچک بودن مقدار آماره آزمون و بزرگ بودن مقدار احتمال در سطح $\alpha = ۰٫۰۱$ برای کلیه ضرایب رگرسیونی نشان دهنده معنی‌دار نبودن، عدم حضور و تأثیر ملموس هر یک از متغیرهای مستقل است. ظهور مقادیر کوچک غیر منتظره‌ای برای آماره t می‌تواند دلیلی بر وجود همخطی در میان متغیرهای مستقل تلقی شود. علیرغم آنکه سایر محققین وجود همخطی بین این داده‌ها را گزارش کرده‌اند، اما در اینجا نیز ابتدا همخطی در بین متغیرهای مستقل بررسی می‌شود. نتیجه این بررسی در جدول ۲ آمده است. براساس این جدول مشاهده می‌شود که یک همخطی بین متغیرهای مستقل X_1 ، X_2، X_3 و X_4 وجود دارد. برای آشنایی بیشتر در خصوص شاخص عدد شرطی و نسبت‌های واریانس می‌توان به بلزلی و همکاران (۲۰۰۴) صفحه ۱۰۴ تا ۱۰۷ مراجعه کرد.

به دلیل وجود هم‌خطی قوی، برآوردگر کمترین توان‌های دوم قابل اطمینان نیست و در نتیجه ناپایدار است. بنابراین با هدف تقلیل اثرات همخطی لزوم استفاده از برآوردگرهای ریج و ریج تحت محدودیت تصادفی آشکار

جدول ۲: نتایج بررسی وجود همخطی بین متغیرها در مجموعه داده‌های سیمان پورتلند.

مقادیر	شاخص	نسبت‌های واریانس				
ویژه	شرطی	عرض از مبدأ	$X_۱$	$X_۲$	$X_۳$	$X_۴$
۴٫۱۲۰	۱٫۰۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰
۰٫۵۵۴	۲٫۷۲۷	۰٫۰۰	۰٫۰۱	۰٫۰۰	۰٫۰۰	۰٫۰۰
۰٫۲۸۹	۳٫۷۷۸	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰
۰٫۰۳۸	۱۰٫۴۶۲	۰٫۰۰	۰٫۰۶	۰٫۰۰	۰٫۰۵	۰٫۰۰
۰٫۰۰۰۶۶	۲۴۹٫۵۷۸	۱٫۰۰	۰٫۹۳	۱٫۰۰	۰٫۹۵	۱٫۰۰

می‌شود. با استفاده از روش اثر ریج، یک مقدار بهینه برای پارامتر ریج به طور تقریبی $K = ۰٫۰۷۵$ تعیین می‌شود.

به منظور کاهش هر چه بیشتر همخطی، کسیرنلر و همکاران (۱۹۹۹) پیشنهاد دادند که یک محدودیت خطی تصادفی به صورت $r = R\beta + e$ که در آن $R = (۰, ۱, -۱, ۱, ۰)$، $r = ۰$ و $e\left(۰, \sigma^۲ I\right)$ بر مجموعه داده‌ها اعمال شود.

اکنون به منظور بررسی پرت بودن مشاهده i در مجموعه داده‌های سیمان پورتلند و ارزیابی و مقایسه نتایج حاصل، آماره‌های آزمون روش انتقال میانگین تحت هر یک از برآوردگرهای کمترین توان دوم ریج و ریج آمیخته محاسبه می‌شود (جدول ۳)، با توجه به جدول ۳ ملاحظه می‌شود که مشاهدات ۶ و ۸ به این دلیل که مقدار آماره آزمون بیان شده در روابط (۸) و (۱۷) مربوط به این دو مشاهده بزرگ است، پرت محسوب شده و مورد توجه هستند. در حالی که مقدار آماره آزمون تعریف شده در رابطه (۲۴) مربوط به مشاهدات ۶، ۸ و ۱۳ بزرگ بوده؛ که دلالت بر پرت بودن این مشاهدات دارد. به این ترتیب وجود تمایز بین نتایج حاصله از روش انتقال میانگین در رگرسیون ریج آمیخته نسبت به رگرسیون ریج و کمترین توان‌های دوم آشکار می‌شود.

جدول ۳: آماره آزمون نقاط پرت در سه روش کمترین توان‌های دوم، ریج و ریج تحت محدودیت تصادفی آزمون در سطح معنی‌داری $\alpha = ۰/۱$.

مشاهدات	F_i	F_i^*	F_i^{**}
۱	۰/۰۰۰۰۰۸	۰/۱۳۹۷	۰/۱۸۸۶
۲	۰/۶۱۶۶	۱/۳۸۸۱	۱/۳۳۷۲
۳	۱/۲۷۹۵	۰/۰۱۳۵	۰/۰۱۴۴
۴	۰/۷۷۶۰	۰/۴۴۹۴	۰/۴۲۹۹
۵	۰/۰۱۶۴	۰/۰۰۰۲	۰/۰۰۸۲
۶	۴/۶۴۹۷	۶/۷۹۴۱	۶/۸۱۶۰
۷	۰/۵۹۵۵	۰/۹۵۵۴	۱/۰۲۱۰
۸	۴/۴۲۴۰	۷/۳۰۸۹	۶/۶۷۴۷
۹	۰/۴۷۶۸	۱/۳۳۱۴	۱/۳۴۱۴
۱۰	۰/۰۴۴۵	۰/۰۰۰۴	۰/۰۱۲۰
۱۱	۱/۳۴۷۵	۱/۸۷۸۴	۱/۰۰۸۱
۱۲	۰/۲۲۰۶	۰/۰۴۶۱	۰/۰۴۶۱
۱۳	۱/۵۰۰۶	۳/۰۲۱۵	۳/۱۷۹۲
آماره	۳/۵۹	۳/۱۸	۳/۱۴

بحث و نتیجه‌گیری

در این مقاله نشان دادیم که آماره آزمون روش انتقال میانگین که با هدف شناسایی مشاهدات پرت تحت هر یک از برآوردگرهای ریج و ریج تحت محدودیت تصادفی و کمترین توان‌های دوم استفاده شده؛ نتایج متمایزی را در اختیار تحلیل‌گر قرار می‌دهد. بنابراین نمی‌توان به مشاهداتی که توسط آماره آزمون بر مبنای روش کمترین توان‌های دوم تحت شرایط وجود همخطی در مدل پرت شناخته می‌شوند، اعتماد کرد. بر پایه این حقیقت، همان‌گونه که بلزلی و همکاران (۲۰۰۴) نیز در خصوص مدل‌های رگرسیونی با برآورد کمترین توان‌های دوم اشاره کرده است، همخطی باید قبل از بررسی و مطالعه مباحث تشخیصی کنترل شود. بنابراین بسط مباحث

تشخیصی به رگرسیون ریج تحت محدودیت‌های خطی تصادفی ضرورت دارد. مقابله با مشکل همخطی و تشخیص مشاهدات تأثیرگذار و پرت از جمله مفاهیم مهم و قابل توجه در مدل‌های رگرسیونی هستند که تعمیم و گسترش این مفاهیم گامی در جهت ترقی و توسعه آنالیز رگرسیونی خواهد بود. شایان ذکر است که موضوع مورد مطالعه در این مقاله قابلیت تعمیم دارد. به این منظور پیشنهاد می‌شود که مباحث تشخیصی را برای سایر برآوردگرهای اریب که تعمیمی از برآورد ریج هستند، بسط و توسعه داد. همچنین امکان بسط مباحث تشخیصی برای برآورد ریج و ریج آمیخته با خطای ناهمبسته و ناهمگی وجود دارد.

تقدیر و تشکر

نویسندگان مقاله از داوران و ویراستار محترم مجله برای ارزیابی و ویرایش این مقاله کمال قدردانی و تشکر را دارند.

مراجع

[1] Belsley, D. A., Kuh, E. and Welsch, R. E. (2004), *Regression Diagnostics*, New York: John Wiley.

[2] Chatterjee, S. and Hadi, A. S. (1986), Influential Observations, High Leverage Points and Outliers in Linear Regression, *Statistical Science*, **1**, 379–416.

[3] Chatterjee, S. and Hadi, A. S. (1988), *Sensitivity Analysis of Linear Regression*, New York: John Wiley.

[4] Durbin, J. (1953), A Note on Regression When There is Extraneous Information About One of the Coefficients, *Journal of American Statistical Association*, **48**, 799–808.

[5] Fallah, M. and Salam, A. (2011), A Modification of the Ridge Type Regression Estimators, *American Journal of Applied Sciences*, **8**,97–102 ISSN 1546-9239.

[6] Ghapani, F. Rasekh, A. R., Akhond, M.R. and Babadi, B. (2015), Detection of Outliers and Inflectional Observations in Linear Ridge Measurement Errors Models with Stochastic Linear Restrictions, *Journal of Sciences, Islamic Republic of Iran*, **26**, 355–366.

[7] Hadi, A. S. (1992), Identifying Multiple Outliers in Multivariate Data, *Journal of Royal Statistics Society*, **54**, 761–771.

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

[9] Hoerl, A. E. and Kennard, R. W. (1970), Ridge Regression: Biased Estimation for Non-Orthogonal Problems, *Technometrics*, **12**, 55–67.

[10] Jahufer, A. and Chen, J. (2009), Assessing Global Influential Observations in Modified Ridge Regression, *Statistical and Probability Letters*,**79**, 513–518.

[11] Jahufer, A. and Chen, J. (2011), Measuring Local Influential Observations in Modified Ridge Regression, *Journal of Data Science*, **9**, 359–372.

[12] Jianxin, P. and Haiyan, X. (1995), Outliers and Influential Observations in a Ridge Mean Shift Regression, *Journal of Mathematical Sciences*, **8**, 12–26.

[13] Kaciranlar, S., Sakallioglu, 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 Datased on Portland Cement, *Sankhya Indian Journal Statistic, Ser (B)*, **61**, 443–459.

[14] Lawrence, K. D. and Marsh, L. C. (1984), Robust Ridge Regression Methods for Predicting U.S. Coal Mining Fatalities, *Communication in Statistics: Theory and Methods*, **13**, 139–149.

[15] Li, Y. and Yang, H. (2010), A New Stochastic Mixed Ridge Estimator in Linear Regression Model, *Statistical Papers*, **51**, 315–323.

[16] Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001), *Introduction to Linear Regression Analysis*, 3rd edition, New York: John Wiley.

[17] Özkale, M. R. (2009), A Stochastic Restricted Ridge Regression Estimator, *Journal of Multivariate Analysis*, **100**, 1706–1716.

[18] Rao, C. R. and Toutenburg, H. (1995), *Linear Models: Least Squares and Alternatives*, New York: Springer.

[19] Riani, M. and Atkinson, A. C. (2000), Robust Diagnostic Data Analysis, *Tranformations in Regression Technometrics*, **42**, 384–398.

[20] Sakallioğlu, S. and Kaciranlar, S. (2008), A New Biased Estimator Based on Ridge Estimation, *Statistical Papers*, **49**, 669–689.

[21] Seber, G. A. F. and Lee, A. J. (2002), *Linear Regression Analysis*, New York: John Wiley.

[22] Shi, L. (1997), Local Influence in Principal Component Analysis, *Biometrika*,**84**, 175–186.

[23] Shi, L. and Wang, X. (1999), Local Influence in Ridge Regression, *Sankhya. Series (B)*, **48**, 342–36.

[24] Singh, B., Chaubey, Y. P. and Dwivedi, T. D. (1986), An Almost Unbiased Ridge Estimator, *Computational Statistics and Data Analysis*, **52**, 879–895.

[25] Steece, B. M. (1986), Regressor Space Outliers in Ridge Regression, *Communication in Statistics: Theory and Methods*, **15**, 3599–3605.

[26] Swindel, B. F. (1976), Good Estimators Based on Prior Information, *Communication in Statistics: Theory and Methods*, **5**, 1065–1075.

[27] Theil, H. (1963), On the Use of Incomplete Prior Information in Regression Analysis, *Journal of American Statistical Association*, **58**, 401–414.

[28] Theil, H. and Goldberger, A. S. (1961), On Pure and Mixed Statistical Estimation in Economics, *Internnational Economic Reviw*, **2** (1), 65–78.

[29] Trenkler, G. (1984), On the Performance of Biased Estimators in the Linear Regression Model with Correlated or Heteroscedastic Errors, *Journal of Econometrics*, **25**, 179–190.

[30] 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, *Communication in Statistics: Theory and Methods*,**23**, 3453–3476.

[31] Walker, E. and Birch, J. B. (1988), Influence Measures in Ridge Regression, *Technometrics*,**30**(2), 221–227.

[32] Weisberg, S. (1983), Some Principles for Regression Diagnostics and Influence Analysis, *Technometrics*, **25**, 240–244.

[33] Wood, H., Steinour, H. H. and Strake, H.R. (1932), Effect of Composition of Portland Cement on Heat Evolved During Hardening, *Journal of Industrial and Engineering Chemistry*, **24**, 1207–1214.

Outlier Detection in Ridge Regression Model Under Stochastic Linear Restrictions

Rasekh, A., Mansouri, B. and Hedaiatpoor, N.

Faculty of Mathematical Sciences and Computer, Shahid Chamran University of Ahvaz, Ahvaz , Iran.

Abstract: The study of regression diagnostic, including identification of the influential observations and outliers, is of particular importance. The sensitivity of least squares estimators to the outliers and influential observations lead to extending the regression diagnostic in order to provide criteria to assess the anomalous observations. Detecting influential observations and outliers in the presence of collinearity is a complicated task, in the sense that collinearity may cover some of the unusual data. One of the considerable methods to identify outliers is the mean shift outliers method. In this article, we extend the mean shift outliers method to the ridge estimates under linear stochastic restrictions, which is used to reduce the effect of collinearity, and to provide the test statistic to identify the outliers in these estimators. Finally, we show the ability of our proposed method using a practical example of real data.

Keywords: Collinearity, Ridg Regression, Ridg Regression under linear restriction, Outlier, Mean shift method

Mathematics Subject Classification (2010): $62J07$, $62J20$.