مجله مدل‌سازی پیشرفته ریاضی
سال ۱۴۰۴، دوره ۱۵، شماره ۳، ص ۴۵–۵۴

شاپا چاپی: ۲۲۵۱-۸۰۸۸
DOI: 10.22055/jamm.2025.48126.2320

شناسایی نقاط پرت به روش انتقال واریانس در یک مدل خطی تحت محدودیت خطی تصادفی

مریم عطوان، بابک بابادی ×، عبدالرحمن راسخ

دانشگاه شهید چمران اهواز، دانشکده علوم ریاضی و کامپیوتر، گروه آمار
تاریخ دریافت: ۱۴۰۳/۰۷/۲۰ تاریخ پذیرش: ۱۴۰۴/۰۷/۲۷
دبیر مسئول: بهزاد منصوری

چکیده: در مدل‌های خطی روش‌های مختلفی برای شناسایی نقاط پرت پیشنهاد شده است. از سوی دیگر وجود هم‌خطی در میان متغیرهای توضیحی می‌تواند منجر به ناپایداری برآوردها گردد. از جمله روش‌های مواجه با هم‌خطی می‌توان به استفاده از محدودیت‌های خطی تصادفی اشاره نمود. در این مقاله روش انتقال واریانس در مدل‌های خطی تحت محدودیت خطی تصادفی مورد مطالعه قرار می‌گیرد. این روش بر این پایه استوار است که یک نقطه پرت در اثر تورم در واریانس آن رخ می‌دهد. ابتدا برآورد پارامترها در این مدل به‌دست می‌آید و سپس آماره‌های آزمون امتیاز و نسبت درستنمایی برای شناسایی نقطه پرت مطرح می‌شوند. در ادامه یک مطالعه‌ی شبیه‌سازی برای نشان دادن توانایی آزمون‌ها مطرح و با استفاده از یک روش بوت استرپ پارامتری توزیع تجربی آماره‌های مرتبط با این آزمون‌ها محاسبه می‌گردد. در پایان نیز مدل پیشنهادی به یک مجموعه از داده‌های واقعی برازش شده است.

واژه‌های کلیدی: آزمون امتیاز، آزمون نسبت درستنمایی، روش بوت استرپ.

رده‌بندی ریاضی: 62j05

۱ مقدمه

نقاط پرت مشاهداتی هستند که به نظر می‌رسد با سایر داده‌های مجموعه سازگار نیستند و می‌توانند تأثیر مخربی بر تحلیل آماری داشته باشند. برای تشخیص این نوع مشاهدات در مدل‌های خطی، روش‌های مختلفی پیشنهاد شده است. در میان این روش‌ها می‌توان به حذف موردی و مدل انتقال واریانس اشاره کرد. اساس روش اول مبتنی بر این فرض است که نقاط پرت ناشی از تغییر در میانگین این مشاهدات می‌باشند و دومین روش این فرض را مطرح می‌کند که یک نقطه‌ی پرت از یک عبارت خطا با واریانس متورم به‌دست می‌آید. کوک و همکاران (۱۹۸۲) نشان دادند که برآوردهای ماکسیمم درستنمایی برای موقعیت نقطه پرت تحت این دو روش می‌تواند متفاوت باشد؛ مگر این‌که بزرگ‌ترین قدرمطلق باقیمانده استیودنت شده، مربوط به باقیمانده‌ی همان مشاهده باشد. با استفاده از باقیمانده ماکسیمم درستنمایی محدود شده (REML)، تامپسون (۱۹۸۵) نشان داد که واریانس باقیمانده و موقعیت نقطه پرت تحت دو روش فوق یکسان است. گومز و همکاران (۲۰۱۰) با استفاده از برآورد حداکثر درستنمایی محدود شده این روش را به مدل‌های خطی آمیخته تعمیم داده و آن را مدل انتقال واریانس برای داده‌های پرت نامیدند. همچنین بابادی و همکاران (۲۰۱۶) روش انتقال واریانس را در مدل‌های خطی آمیخته با

×نویسنده مسئول مقاله
رایانامه: b.babadi@scu.ac.ir (B. Babadi)

خطا در اندازه‌گیری جهت یافتن نقاط پرت معرفی کردند. در رگرسیون خطی، مستقل بودن متغیرها از اهمیت خاصی برخوردار است به‌گونه‌ای که اگر استقلال بین متغیرها از بین برود، تجزیه و تحلیل رگرسیونی تحت تأثیر قرار می‌گیرد. درصورت وجود همبستگی بین متغیرهای مدل با پدیده‌ای به نام هم‌خطی مواجه هستیم. هم‌خطی سبب بروزمشکلاتی مانند: ناپایدار بودن برآوردها، فواصل اطمینان بزرگ (به‌دلیل افزایش انحراف معیار)، عدم اعتبار برآورد کمترین توان دوم و ... می‌شود. روش‌های مختلفی جهت مقابله با هم‌خطی وجود دارد از جمله: استفاده از اطلاعات پیشین و اطلاعات کمکی درقالب محدودیت خطی تصادفی بر روی ضرایب رگرسیونی، استفاده از برآوردگرهای اُریب (مانند برآوردگر لیو، برآوردگر ریج)، استفاده‌ی هم‌زمان از برآوردگر اُریب و محدودیت خطی تصادفی، حذف متغیرهای دچار هم‌خطی و افزایش حجم نمونه.
برای رفع مشکل هم‌خطی برآوردگرهای مختلفی توسط استین (۱۹۵۶)، هورل و کنارد (۱۹۷۰) و لیو (۱۹۹۳) پیشنهاد شده است. بنا به گفته‌ی رائو و همکاران (۲۰۰۸) یک روش برای کاهش مشکلات هم‌خطی، استفاده از اطلاعات کمکی در خصوص پارامترهای مدل رگرسیونی است و از آن تحت عنوان محدودیت خطی یاد می‌شود. دراین مقاله روش انتقال واریانس برای حالتی که متغیرهای مستقل دارای هم‌خطی باشند، مطرح شده است. ابتدا برآورد پارامترها با استفاده از تابع درستنمایی محدود شده و اضافه کردن اطلاعات کمکی به مدل به‌دست می‌آیند. سپس آزمون‌های امتیاز و نسبت درستنمایی برای این مدل محاسبه و در پایان یک مطالعه شبیه‌سازی به‌منظور توانایی آزمون‌ها ارائه و مدل پیشنهادی به یک مجموعه از داده‌های واقعی برازش می‌شود.

۲ مدل انتقال واریانس

یک مدل رگرسیون خطی را به‌صورت زیر در نظر بگیرید:

$$y = X\beta + \varepsilon \tag{۱.۲}$$

که در آن y بردار ستونی n بعدی از مشاهدات، X یک ماتریس پر رتبه ستونی $n \times p$ از مقادیر معلوم، β یک بردار $p \times ۱$ بعدی از پارمترهای مجهول و ε یک بردار $n \times ۱$ از خطاهای تصادفی و دارای توزیع $N(۰, \sigma^۲ I_n)$ می‌باشد.
فرض کنید مشاهده‌ی i-ام دارای تورم در واریانس باشد. یک مدل انتقال واریانس برای این مشاهده به‌صورت زیر تعریف می‌شود:

$$y = X\beta + \delta_i d_i + \varepsilon \tag{۲.۲}$$

عامل تورم واریانس یعنی δ_i دارای توزیع نرمال با میانگین صفر و واریانس α_i است که به شکل یک اثر تصادفی و نامشخص نمایان می‌شود. همچنین d_i برداری است که i-امین عضو آن یک و مابقی اعضای آن صفر می‌باشد.
با فرض اینکه ε و δ_i مستقل از هم باشند، ماتریس واریانس-کوواریانس مشاهدات به‌صورت زیر نوشته می‌شود:

$$Var(y) = \alpha_i d_i d_i' + \sigma^۲ I_n = \sigma^۲(\omega_i d_i d_i' + I_n) = \sigma^۲ H_i$$

که در آن $\omega_i = \dfrac{\alpha_i}{\sigma^۲}$ است و واریانس مشاهده‌ی i-ام به اندازه‌ی $\sigma^۲\omega_i$ متورم می‌شود.
فرض کنید بردار پارامتر β تحت محدودیت‌های خطی تصادفی به فرم زیر باشد:

$$r = R\beta + e \quad ; \quad e \sim N(0, \sigma^۲ I_s)$$

که r یک بردار $s \times ۱$ و R یک ماتریس معلوم $s \times p$ با رتبه s است. حال با اضافه کردن محدودیت خطی تصادفی به مدل (۲.۲)، مدل به‌صورت زیر بازنویسی می‌شود:

$$\begin{pmatrix} y \\ r \end{pmatrix} = \begin{pmatrix} X \\ R \end{pmatrix} \beta + \begin{pmatrix} d_i \\ 0 \end{pmatrix} \delta_i + \begin{pmatrix} \varepsilon \\ e \end{pmatrix},$$
$$y_r = X_r\beta + d_{ir}\delta_i + e_r \tag{۳.۲}$$

با فرض این که ε و e ناهمبسته باشند، ماتریس واریانس -کوواریانس y_r برابر با $Var(y_r) = \sigma^۲ H_{ir} = \sigma^۲ \begin{bmatrix} H_i & 0 \\ 0 & I_s \end{bmatrix}$ است.

۳ تابع درستنمایی محدود شده برای مدل انتقال واریانس

تابع لگاریتم درستنمایی محدود شده برای مدل (۳.۲) به‌صورت زیر نوشته می‌شود:

$$RL(\omega_i,\sigma^{۲};y_r)=\frac{۱}{۲}\Big\{C(X_r)+(N-p)\log\sigma^{۲}+\log|H_{ir}|+\log|X_r^{'}H_{ir}^{-۱}X_r|$$
$$+\frac{۱}{\sigma^{۲}}\left(y_r-X_r\hat{\beta}_r(\omega_i)\right)^{'}H_{ir}^{-۱}\left(y_r-X_r\hat{\beta}_r(\omega_i)\right)\Big\}.$$

که

$$\hat{\beta}_r(\omega_i)=(X^{'}H_i^{-۱}X+R^{'}R)^{-۱}(X^{'}H_i^{-۱}y+R^{'}r)$$
$$=\hat{\beta}_r-\frac{\omega_i}{۱+\omega_i(۱-a_{ii})}(X^{'}X+R^{'}R)^{-۱}x_i\hat{e}_{ir}$$

در تساوی فوق $\hat{\beta}_r=(X^{'}X+R^{'}R)^{-۱}(X^{'}y+R^{'}r)$، $a_{ii}=d_i^{'}X(X^{'}X+R^{'}R)^{-۱}X^{'}d_i$، $H_i^{-۱}=I_n-\frac{\omega_i}{\omega_i+۱}d_id_i^{'}$، $\hat{e}_{ir}=y_i-x_i^{'}\hat{\beta}_r$ و $N=n+s$ است.
با توجه به رابطه

$$\left(y_r-X_r\hat{\beta}_r(\omega_i)\right)^{'}H_{ir}^{-۱}\left(y_r-X_r\hat{\beta}_r(\omega_i)\right)=(y_r-X_r\hat{\beta}_r)^{'}(y_r-X_r\hat{\beta}_r)-\frac{\omega_i\hat{e}_{ir}^{۲}}{۱+\omega_i(۱-a_{ii})}$$
$$=(y_r-X_r\hat{\beta}_r)^{'}(y_r-X_r\hat{\beta}_r)-\frac{\hat{e}_{ir}^{۲}}{(۱-a_{ii})}+\frac{\hat{e}_{ir}^{۲}}{(۱-a_{ii})(۱+\omega_i(۱-a_{ii}))}$$

یک شکل جایگزین برای تابع لگاریتم درستنمایی محدودشده به‌صورت زیر می‌توان به‌دست آورد که از آن برای برآورد پارامترهای ω_i و $\sigma^{۲}$ استفاده می‌شود:

$$RL_i(\omega_i;\sigma^{۲};y_r)=-\frac{۱}{۲}\Big\{C(X_r)+(N-p-۱)\log\sigma^{۲}+\log(۱+\omega_i(۱-a_{ii}))\sigma^{۲}$$
$$+\frac{(y_r-X_r\hat{\beta}_r)^{'}(y_r-X_r\hat{\beta}_r)}{\sigma^{۲}}-\frac{\hat{e}_{ir}^{۲}}{\sigma^{۲}(۱-a_{ii})}+\frac{\hat{e}_{ir}^{۲}}{\sigma^{۲}(۱-a_{ii})(۱+\omega_i(۱-a_{ii}))}\Big\}$$

با مشتق گرفتن از $RL_i(\omega_i;\sigma^{۲};y_r)$ نسبت به $\sigma^{۲}$ و $\sigma^{۲}(۱+\omega_i(۱-a_{ii}))$ برآورد $\sigma^{۲}$ و ω_i به‌صورت زیر به‌دست می‌آید:

$$\hat{\sigma}_{ir}^{۲}=\begin{cases}\dfrac{(N-p-t_{ir}^{۲})\hat{\sigma}_r^{۲}}{(N-p-۱)} & t_{ir}^{۲}>۱\\ \hat{\sigma}_r^{۲} & \text{در غیر این‌صورت}\end{cases}$$

$$\hat{\omega}_{ir}=\begin{cases}\dfrac{(N-p)(t_{ir}^{۲}-۱)}{(۱-a_{ii})(N-p-t_{ir}^{۲})} & t_{ir}^{۲}>۱\\ ۰ & \text{در غیر این‌صورت}\end{cases}$$

جایی که $t_{ir}^{۲}=\dfrac{\hat{e}_{ir}^{۲}}{a_{ii}\hat{\sigma}_r^{۲}}$ و $\hat{\sigma}_r^{۲}=\dfrac{۱}{N-p}\left[y^{'}y-\hat{\beta}_r^{'}X^{'}y)+(r^{'}r-\hat{\beta}_r^{'}R^{'}r)\right]$ است.

۴ آزمون امتیاز و آزمون نسبت درستنمایی

شناسایی نقاط تأثیرگذار مدل با توجه به نقش هر مشاهده (ممکن است مشاهده‌ی پرت و یا مشاهده‌ای موثر و پرنفوذ باشد) در آنالیز داده‌ها نقش مهمی دارد. به‌همین دلیل شناسایی این نقاط در مدل همیشه مورد بحث و بررسی قرار می‌گیرد. در این بخش آزمون امتیاز و آزمون نسبت درستنمایی را برای مدل انتقال واریانس معرفی و نشان داده می‌شود که این دو آزمون تابعی از باقیمانده‌های استیودنت شده‌ی مشاهده‌ای هستند که احتمالاً به‌عنوان نقطه‌ی پرت معرفی خواهد شد.

۱.۴ آزمون امتیاز (SC):

برای آزمون این فرضیه که مشاهده i-ام واریانس متورمی ندارد $(H_۰ : \omega_i = 0)$ در مقابل این فرضیه که مشاهده دارای واریانس متورم باشد $(H_۱ : \omega_i > ۰)$، از آماره آزمون امتیاز استفاده می‌کنیم. کاکس و هینکلی (۱۹۷۴) شکل کلی آماره آزمون امتیاز را به‌صورت زیر معرفی کردند:

$$SC_i = \left[\frac{\partial}{\partial \omega_i} RL_i(\omega_i; \sigma^۲; y_r)\right]^۲ J_{\omega\omega}$$

که $J_{\omega\omega}$ درایه پایین گوشه سمت راست معکوس ماتریس اطلاع فیشر مورد انتظار $J(\sigma^۲, \omega_i)$ می‌باشد. تحت فرض $H_۰$ این مقدار برابر است با:

$$J_{\omega\omega} = \frac{۲(N-p)}{(۱-a_{ii})^۲(N-p-۱)}$$

در نتیجه می‌توان آماره‌ی آزمون امتیاز برای مشاهده‌ی i-ام را به‌صورت زیر به‌دست آورد:

$$SC_i(\omega_i = 0) = \frac{N-p}{۲(N-p-۱)}(t_{ir}^۲ - ۱)^۲ \qquad t_{ir}^۲ > ۱$$

۲.۴ آزمون نسبت درستنمایی (LRT)

اگر $LR_۰(\hat{\sigma}_r^۲; y_r)$ مقدار لگاریتم تابع درستنمایی تحت فرض $H_۰$ و $LR_i(\hat{\omega}_{ir}, \hat{\sigma}_{ir}^۲; y_r)$ مقدار لگاریتم تابع درستنمایی تحت فرض $H_۱$ باشد، آماره آزمون نسبت درستنمایی به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned} LRT_i &= -۲(LR_۰(\hat{\sigma}_r^۲; y_r) - LR_i(\hat{\omega}_{ir}, \hat{\sigma}_{ir}^۲; y_r)) \\ &= (N-p-۱)\log\frac{(N-p-۱)}{(N-p-t_{ir}^۲)} - \log t_{ir}^۲ \qquad t_{ir}^۲ > ۱ \end{aligned}$$

به‌راحتی می‌توان نشان داد که آماره آزمون امتیاز و آماره آزمون درستنمایی توابعی صعودی از $t_{ir}^۲$ می‌باشند.

۵ مطالعه شبیه‌سازی

در این بخش با استفاده از مطالعه شبیه‌سازی، توانایی آزمون‌های امتیاز و نسبت درستنمایی را براساس خطای نوع اول و توان آزمون مورد بررسی قرار می‌دهیم. همانند مک دونالد و گلانئو (۱۹۷۵) جهت تولید متغیرهای توضیحی با درجات مختلف هم‌خطی از فرمول زیر استفاده می‌کنیم:

$$x_{ij} = (۱-\rho^۲)^{\frac{۱}{۲}} z_{ij} + \rho z_{ip}; \quad i = ۱, ۲, ..., n \;\;,\;\; j = ۱, ۲, ..., p-۱$$

z_{ij} و $\rho^۲$ به ترتیب اعداد مستقل از توزیع نرمال استاندارد و همبستگی میان دو متغیر توضیحی را نشان می‌دهند. برای همبستگی سه مقدار $\rho = ۰/۴۲, ۰/۶, ۰/۹۸$، $n = ۱۰, ۲۰, ۶۰$، $p = ۳$، $s = ۲$ و $\sigma^۲ = ۲$ را در نظر می‌گیریم. همچنین در این شبیه‌سازی مقادیر $\beta = (\beta_۰, \beta_۱, \beta_۲) = (۵, ۰/۲, ۰/۵)$ در نظر گرفته می‌شود. سپس l-امین مجموعه از داده شبیه‌سازی را به‌فرم زیر تولید می‌کنیم:

$$\begin{aligned} y_l &= X\beta + \varepsilon_l \\ r_l &= R\beta + e_l; \quad l = ۱, ۲, ..., ۱۰۰۰ \end{aligned}$$

که $y_l = (y_{1l}, y_{2l}, ..., y_{nl})'$، $X = (1_n, X^{(1)}, X^{(2)})$ و 1_n یک بردار n بعدی با درایه‌های یک، $X^{(j)} = (x_{1j}, ..., x_{nj})'$; $j = 1, 2$، $r_l = (r_{1l}, ..., r_{sl})'$ و همانند y_l بازنویسی می‌شود. همچنین $\varepsilon_l \sim N(0, \sigma^2 I_n)$، $R^{(j)} = (R_{1j}, ..., R_{sj})'$; $j = 1, 2$، $R = (1_s, R^{(1)}, R^{(2)})$ و همانند r_l بازنویسی می‌شود. $e_l \sim N(0, \sigma^2 I_s)$، R_{ij} نیز از توزیع نرمال استاندارد تولید می‌شود. به‌طور کلی با اضافه کردن محدودیت تصادفی داده‌ها از مدلی به فرم زیر تولید می‌شود:

$$y_{l(r)} = X_r \beta + e_{l(r)}$$

برای انجام مطالعه شبیه‌سازی از نرم‌افزار R استفاده شده است. برای هر ترکیب از پارامترها، ۱۰۰۰ تکرار درنظر گرفته و در هر تکرار آماره‌ی آزمون امتیاز و آزمون نسبت درستنمایی برای مشاهده‌ی اول محاسبه می‌شود (انتخاب این مشاهده کاملاً اختیاری می‌باشد). همچنین با توجه به گومز و همکاران (۲۰۱۰) از روش بوت استرپ پارامتری برای به‌دست آوردن توزیع تجربی برای آماره آزمون نسبت درستنمایی و آماره آزمون امتیاز، تحت شرایطی که هیچ داده‌ی پرتی وجود نداشته باشد، استفاده شده است.
برای تولید توزیع تجربی آماره‌های آزمون مورد بحث، ابتدا $g = 2500$ مجموعه داده از مدل زیر تولید می‌شود:

$$y^*_{lg(r)} = X_r \hat{\beta}_{lr} + e^*_{lg(r)}$$

که $e^*_{lg(r)}$ دارای توزیع $N(0, \hat{\sigma}^2_{lr} I_N)$ می‌باشد. $\hat{\beta}_{lr}$ و $\hat{\sigma}^2_{lr}$ برآوردهای مدل رگرسیونی محدود شده حاصل از بردار مشاهدات $y_{l(r)}$ تحت محدودیت خطی تصادفی هستند. در هر بردار از مشاهدات y^*_{lg} آماره‌های آزمون امتیاز و نسبت درستنمایی برای یک مدل انتقال واریانس محدود شده متناظر با مشاهده اول محاسبه می‌شود. همچنین صدک ۹۵ هر آماره آزمون را با استفاده از توزیع تجربی، به‌عنوان یک مقدار آستانه‌ای برای آماره آزمون محاسبه شده از داده‌های y_l به‌دست می‌آوریم. برای به‌دست آوردن تخمینی مناسب از خطای نوع اول در یک آماره‌ی آزمون با $\alpha = 0.05$، کافیست تعداد مجموعه داده‌هایی که در آن‌ها آماره‌های آزمون از صدک ۹۵ حاصل از توزیع تجربی بیشتر باشد را بر ۱۰۰۰ تقسیم کنیم.
همچنین با درنظر گرفتن اثر انتقال واریانس و مقادیر $\delta = 2, 4, 6$، می‌توان حساسیت نسبی آزمون‌ها نسبت به تغییر در واریانس مشاهده اول را محاسبه کرد. مجدداً برای هر ترکیب از پارامترها ۱۰۰۰ مجموعه داده از مدل زیر تولید می‌کنیم.

$$y_l = X\beta + \delta d_1 + \varepsilon_l$$

که d_1 برداری n بعدی با مقدار یک در عضو اول و مقدار صفر در سایر درایه‌ها است.
مانند آن‌چه در خطای نوع اول صورت گرفت، آماره‌های آزمون امتیاز و نسبت درستنمایی و توزیع تجربی متناظر با هر کدام از آن‌ها محاسبه شده و توان آزمون‌ها به‌دست می‌آید.
خطای نوع اول تجربی متناظر با آماره‌های آزمون امتیاز و نسبت درستنمایی برای مقادیر به‌دست آمده از توزیع تجربی تحت فرض صفر و توان آزمون آن‌ها به‌ترتیب در جداول (۱) و (۲) ارائه شده است.

جدول ۱: احتمال خطای نوع اول آزمون امتیاز و آزمون نسبت درستنمایی در سطح $\alpha = 0.05$، $\delta = 2$

LRT	Scor	n	ρ
۰٫۰۴۹	۰٫۰۶۲	۱۰	
۰٫۰۴۷	۰٫۰۵۳	۲۰	۰٫۴۲
۰٫۰۵۸	۰٫۰۴۵	۶۰	
۰٫۰۴۹	۰٫۰۶۰	۱۰	
۰٫۰۴۷	۰٫۰۵۲	۲۰	۰٫۶
۰٫۰۵۸	۰٫۰۴۵	۶۰	
۰٫۰۴۴	۰٫۰۵۶	۱۰	
۰٫۰۴۷	۰٫۰۵۳	۲۰	۰٫۹۸
۰٫۰۵۹	۰٫۰۴۵	۶۰	

جدول ۲: توان آزمون مربوط به آزمون امتیاز و آزمون نسبت درستنمایی

$\delta = ۶$		$\delta = ۴$		$\delta = ۲$			
LRT	Scor	LRT	Scor	LRT	Scor	n	ρ
۰٫۹۳۹	۰٫۸۷۹	۰٫۶۶۸	۰٫۶۰۸	۰٫۲۳۴	۰٫۲۰۵	۱۰	
۰٫۹۷۳	۰٫۹۷۳	۰٫۷۲۹	۰٫۶۹۴	۰٫۲۵۳	۰٫۲۲۶	۲۰	۰٫۴۲
۰٫۹۸۴	۰٫۹۹۰	۰٫۷۸۸	۰٫۷۶۹	۰٫۲۹۳	۰٫۲۴۴	۶۰	
۰٫۹۳۶	۰٫۸۷۸	۰٫۶۶۶	۰٫۶۰۰	۰٫۲۳۶	۰٫۲۰۵	۱۰	
۰٫۹۷۳	۰٫۹۷۲	۰٫۷۲۸	۰٫۶۹۴	۰٫۲۵۳	۰٫۲۲۷	۲۰	۰٫۶
۰٫۹۸۴	۰٫۹۹۰	۰٫۷۸۸	۰٫۷۶۸	۰٫۲۹۵	۰٫۲۴۴	۶۰	
۰٫۹۴۳	۰٫۸۸۴	۰٫۶۷۳	۰٫۵۹۷	۰٫۲۳۶	۰٫۲۰۱	۱۰	
۰٫۹۷۳	۰٫۹۷۳	۰٫۷۲۹	۰٫۶۹۶	۰٫۲۴۴	۰٫۲۲۴	۲۰	۰٫۹۸
۰٫۹۸۴	۰٫۹۹۱	۰٫۷۸۷	۰٫۷۶۸	۰٫۲۹۲	۰٫۲۴۱	۶۰	

با مشاهده‌ی مقادیر موجود در جدول (۱) نتیجه می‌گیریم که در حالت کلی خطای نوع اول تجربی برای هر دو آماره‌ی آزمون امتیاز و آزمون نسبت درستنمایی به مقدار مورد نظر $\alpha = ۰٫۰۵$ نزدیک است. از طرفی مقادیر جدول (۲) نشان می‌دهد با افزایش عامل انتقال واریانس δ و حجم نمونه توان آزمون‌ها افزایش می‌یابد. این جداول همچنین نشان می‌دهند که نتایج حاصل از خطای نوع اول و توان تجربی دو آزمون به یکدیگر نزدیک است. برای سطوح دیگری از همبستگی و واریانس $\sigma^۲$ نتایج مشابهی حاصل می‌شود که از ارائه آن در این مقاله خودداری نموده‌ایم.

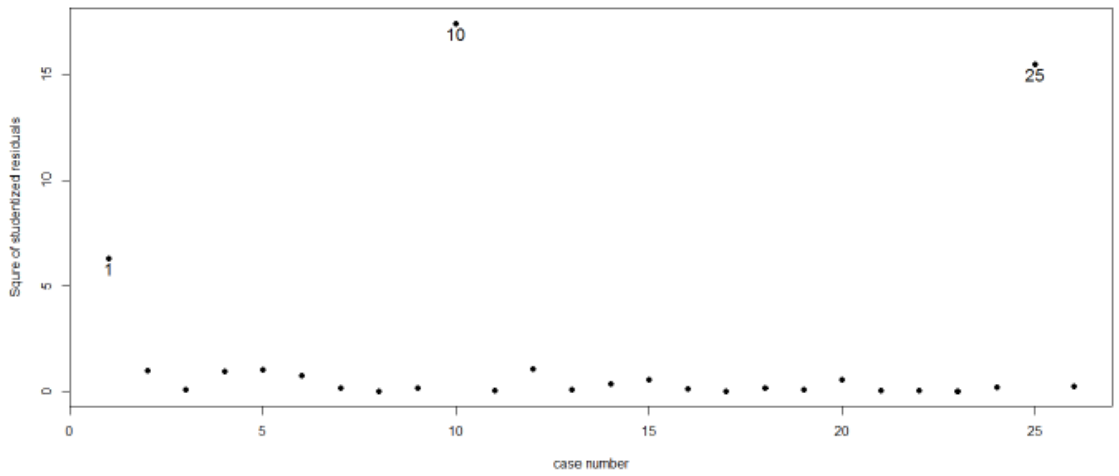
۶ مثال عددی

فیرن (۱۹۸۳) نتایج آزمایش واسنجی اشعه نزدیک به مادون قرمز NIR را برای دو سری داده‌ی گندم، به‌منظور اندازه‌گیری محتوای پروتئین (متغیر پاسخ) ارائه داد. متغیرهای توضیحی بازتاب اشعه NIR بر گندم با شش طول موج مختلف هستند. مقادیر مختلف طول موج با استفاده از فرمول $\log(\frac{۱}{C})$ ساخته شده، که پارامتر C در آن بازتاب اشعه می‌باشد. هدف از انجام واسنجی یافتن یک ترکیب خطی از متغیرهای توضیحی به‌منظور پیش‌بینی محتوای پروتئین موجود در گندم است. مقدار عدد شرطی ۲۷۰۳٫۰۴۱ نشان‌دهنده‌ی وجود هم‌خطی بین متغیرهای اول، چهارم و پنجم می‌باشد.

جهت کاهش اثر هم‌خطی از سری اول داده‌های جمع‌آوری شده چهار مشاهده به تصادف در قالب محدودیت خطی تصادفی به فرم زیر مورد استفاده قرار می‌گیرد.

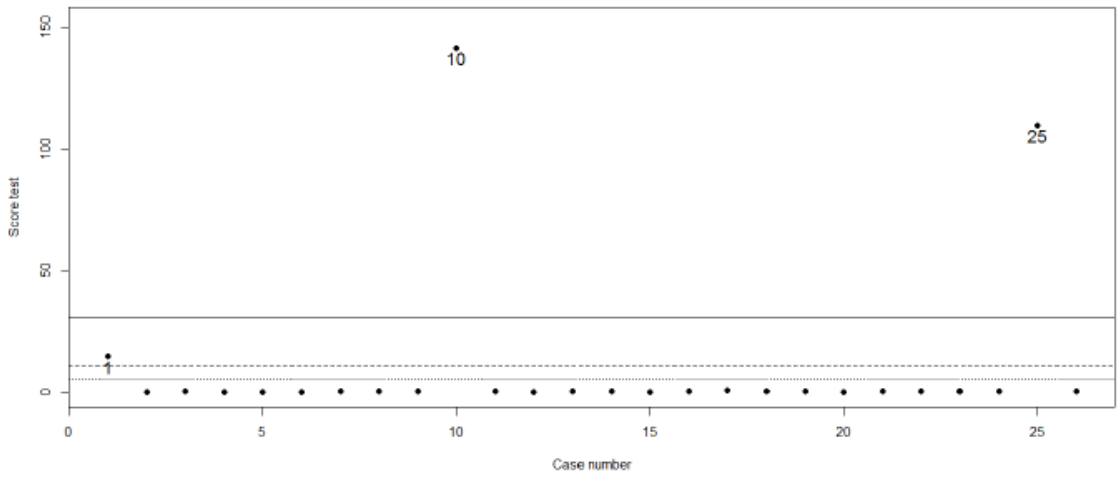
$$r = \begin{bmatrix} ۱۱٫۶۷ \\ ۱۱٫۳۵ \\ ۱۲٫۳۵ \\ ۸٫۰۵ \end{bmatrix} \qquad R = \begin{bmatrix} ۴۵۰ & ۱۱۵ & ۲۳۶ & ۳۵۲ & ۳۴۰ & -۱۵ \\ ۴۹۱ & ۱۴۷ & ۲۶۹ & ۳۸۹ & ۳۹۱ & ۱ \\ ۵۹۲ & ۲۲۹ & ۳۶۰ & ۴۸۴ & ۵۲۴ & ۵۱ \\ ۴۸۸ & ۱۳۴ & ۲۵۸ & ۳۹۳ & ۳۷۷ & -۵ \end{bmatrix}$$

برای تعیین مشاهدات پرت، با استفاده از باقیمانده‌های استیودنت شده مشاهدات ۲۵ و ۱۰ و ۱ به‌عنوان نقاط پرت احتمالی معرفی شده است (شکل (۱) را ملاحظه کنید).

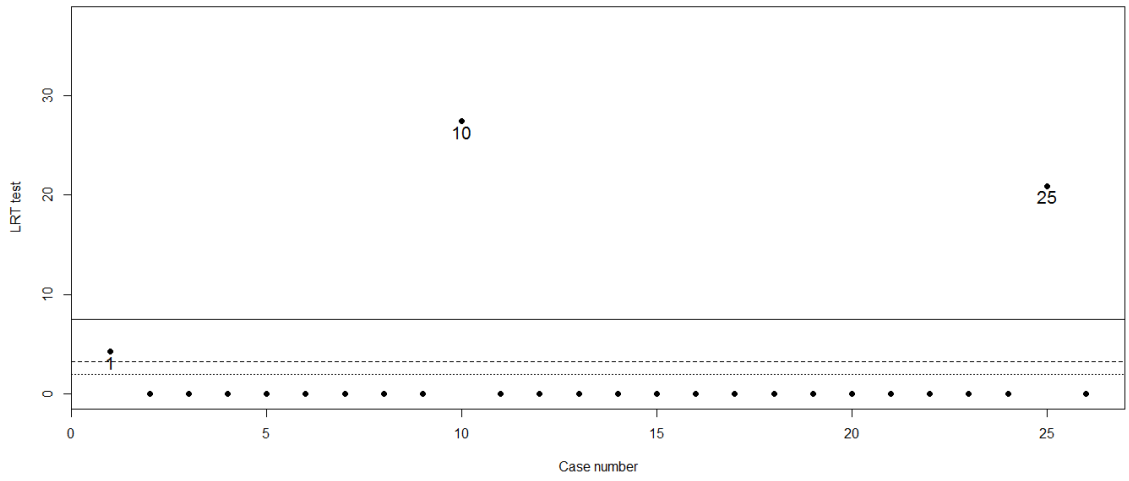


شکل ۱: نمودار باقی‌مانده استیودنت شده $(t_{ir}^{۲})$ در برابر ترتیب مشاهدات

ابتدا برای هر مشاهده آماره‌های آزمون امتیاز و نسبت درستنمایی محاسبه شده است. سپس ۲۵۰۰ مجموعه داده شبیه‌سازی از مدلی که نقطه پرت در آن وجود ندارد، تولید و برای هر مجموعه داده‌ی شبیه‌سازی آماره‌های آزمون امتیاز و نسبت درستنمایی را محاسبه و به‌ترتیب نزولی مرتب شده است. در آخر نیز توزیع تجربی آماره‌های مرتب محاسبه می‌شود. آماره‌های آزمون امتیاز و نسبت درستنمایی برای هر کدام از داده‌های گندم به‌همراه صدک‌های ۹۰ از توزیع تجربی اولین، دومین و سومین آماره‌های مرتب متناظر با این آماره‌های آزمون در شکل‌های (۲) و (۳) نمایش داده شده است. از آنجا که آماره‌های آزمون مشاهدات ۲۵، ۱۰ و ۱ از صدک‌های ۹۰ آماره‌ی مرتب متناظرشان بیشتر است، در نتیجه این نقاط به‌عنوان نقاط پرت معرفی می‌شود.



شکل ۲: نمودار آماره‌ی آزمون امتیاز برای هر مشاهده، به همراه صدک ۹۰ توزیع تجربی برای k آماره‌ی مرتب اول از آماره‌ی SC تحت فرض صفر. خط توپر $(k = ۱)$، خط بریده بریده $(k = ۲)$، نقطه‌چین $(k = ۳)$



شکل ۳: نمودار آماره‌ی آزمون نسبت درستنمایی برای هر مشاهده، به همراه صدک ۹۰ توزیع تجربی برای k آماره‌ی مرتب اول از آماره‌ی LRT تحت فرض صفر. خط توپر $(k = ۱)$، خط بریده بریده $(k = ۲)$، نقطه‌چین $(k = ۳)$

۷ نتیجه‌گیری

در این مقاله ابتدا یک مدل انتقال واریانس تحت محدودیت خطی تصادفی معرفی و با استفاده ازتابع لگاریتم درستنمایی محدود شده برآورد پارامترها تحت این مدل محاسبه می‌شود. سپس آماره آزمون امتیاز و آماره آزمون نسبت درستنمایی محاسبه و نشان داده شد که هر دو آماره توابعی صعودی از باقی‌مانده استیودنت شده مشاهده‌ای هستند که دارای تورم در واریانس است. در ادامه جهت نشان دادن توانایی این آزمون‌ها یک مطالعه‌ی شبیه‌سازی ارائه و با استفاده از یک روش بوت استرپ پارامتری توزیع تجربی آماره‌های مرتب برای این آزمون‌ها محاسبه و نشان داده شد با افزایش اثر انتقال واریانس و اندازه نمونه توان آزمون‌ها افزایش پیدا می‌کند.

فهرست منابع

Babadi, B., et al. (2016). “A variance shift model for detection of outliers in the linear mixed measurement error models.” *Communications in Statistics-Theory and Methods* **45** (24): 7350-7366.

Cook, R.D., Holschuh, N. and Weisberg, S. (1982). A note on an alternative outlier model, *J. R. Statist. Soc. Ser*. B **44**, pp. 370-376.

Cox, D. and D. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall, London.

Fearn, T. (1983). Collaborative evaluation of near infrared reflectance analysis for the determination of protein, moisture and hardness in wheat, *Journal of the Science of Food and Agriculture*, **34**, 1011-1017.

Gumedze, F.N., Welham, S.J., Gogel, B.J. and Thompson, R. (2010). A variance shift model for detection of outliers in the linear mixed model, *Computational Statistics and Data Analysis*, **54**, 2128-2144.

Hoerl, A.E. and Kennard, R.W. (1970), Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometric*., **12** (1), 55–67.

Liu, K. (1993), A New Class of Biased Estimate in Linear Regression, *Communications in Statistics-Theory and Methods*, **22**, 393–402.

McDonald MC, Galarneau DI. (1975). A Monte Carlo evaluation of some ridge-type estimators, *J. Amer. Statist. Assoc.* 70:407–416

Rao, C.R., Toutenburg, H., Shalabh and Heumann, C. (2008). *Linear Models and Generalizations: Least Squares and Alternatives. Third ed*, Springer Berlin.

Stein, C. (1956), Inadmissibility of the Usual Estimator for Mean of Multivariate Normal Distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **31** (1), 197–206.

Thompson, R. (1985). A note on restricted maximum likelihood estimation with an alternative outlier model, *J. R. Statist. Soc. Ser.* B **47**, pp. 53-55.

Identifing outliers using variance shift method in linear model under linear stochastic constraint

M. Atvan, B. Babadi†, A. R. Rasekh

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

Received: 11 October 2024 Accepted: 19 October 2025

Communicated by: Behzad Mansouri

Abstract: In linear models, various methods have been proposed to identify outliers. On the other hand, the existence of collinearity among the explanatory variables can lead to instability of estimates. To deal with collinearity, among different methods, one can refer to the stochastic linear constraints. This paper examines the variance shift method in linear models under these stochastic linear constraints. The method is based on the premise that an outlier arises due to an inflation in its variance. First, we obtain the parameter estimates for this model. Then, we propose score and likelihood ratio test statistics to identify outliers. Next, we conduct a simulation study to demonstrate the effectiveness of the proposed test statistics. Additionally, we calculate the empirical distributions of the test statistics using a parametric bootstrap method. Finally, we apply the proposed model to a set of real data.

Keywords: Score test, Likelihood ratio test, Bootstrap procedure.

©2024 Shahid Chamran University of Ahvaz, Ahvaz, Iran. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-NonComertial 4.0 International (CC BY-NC 4.0 license) (http://creativecommons.org/licenses/by-nc/4.0/).

†**Corresponding author.**
E-mail addresses: **(B. Babadi)** b.babadi@scu.ac.ir .