روش‌های تشخیصی در مدل‌های خطی ریج نیمه‌پارامتری با خطا در اندازه‌گیری

هادی امامی[۱]

گروه آمار، دانشگاه زنجان

تاریخ دریافت: تاریخ پذیرش:

چکیده: در این مقاله، مباحث تشخیصی در مدل‌های خطی نیمه‌پارامتری با خطا در اندازه‌گیری باوجود همخطی چندگانه مطالعه می‌شود. در ابتدا برای برطرف نمودن اثرات همخطی چندگانه، با استفاده از روش ماکزیمم درستنمایی تصحیح‌شده تاوانیده برآوردگرهای ریج پیشنهاد می‌شود و سپس براساس رویکرد حذف موردی، آماره‌های تشخیصی برای شناسایی و ارزیابی مشاهدات مؤثر در برآوردگرهای پیشنهادی معرفی می‌شوند. در ادامه نشان داده می‌شود که این آماره‌ها تابعی از مقادیر باقی‌مانده و داده‌های نافذ هستند. علاوه بر رویکرد حذف موردی، رویکرد تأثیر موضعی بر پایه لگاریتم درستنمایی تصحیح‌شده تاوانیده جهت شناسایی و ارزیابی مشاهدات دورافتاده و مؤثر در مدل موردبحث تعمیم داده‌شده است. سپس با استفاده از مطالعه شبیه‌سازی کارایی برآوردگرهای پیشنهادی با توجه به معیار میانگین خطای مجانبی ارزیابی می‌شود و در پایان روش‌های تشخیصی مطرح‌شده به یک مجموعه داده واقعی مورد تحلیل و بررسی قرارگرفته است.

واژه‌های کلیدی: اسپلاین هموارساز، برآوردگر ریج، خطا در اندازه‌گیری، معیار اعتبارسنجی، مدل‌های خطی نیمه پارامتری، هم خطی چندگانه.

رده‌بندی ریاضی (۲۰۱۰): ۶۲J۰۵، ۶۲J۰۷

۱- مقدمه

در سه دهه اخیر مباحث تشخیصی ازجمله شناسایی مشاهدات مؤثر و دورافتاده و ارزیابی مناسبت مدل، توجه زیادی را به خود معطوف نموده است طوری که این مقوله بیش از همه در مدل‌های خطی پارامتری موردتوجه قرارگرفته است. در این زمینه کوک [۱]، کوک و ویزبرگ [۲] و بلسلی

[۱]- آدرس الکترونیکی نویسنده مسئول مقاله:h.emami@znu.ac.ir

و همکاران [۳] مطالعات وسیعی انجام داده‌اند. در مدل‌های رگرسیون نا پارامتری و نیمه‌پارامتری مباحث تشخیصی به‌ندرت موردمطالعه قرارگرفته‌اند، از میان مطالعات پایه‌ای انجام‌شده کیم [۴] و کیم و همکاران [۵] باقیمانده‌ها، داده‌های نافذ و آماره کوک را در مدل‌های خطی نا پارامتری و نیمه‌پارامتری مورد کنکاش قرار داده‌اند. مشابه پژوهش آن‌ها فونگ و همکاران [۶] چنین مباحثی را به مدل‌های آمیخته خطی نیمه‌پارامتری تعمیم داده‌اند. از طرف دیگر وجود همخطی چندگانه در مدل‌های خطی نیمه‌پارامتری دور از تصور نیست. شبیه مدل‌های خطی پارامتری وجود همخطی در مدل‌های خطی نیمه‌پارامتری مسئله‌ساز است و اثرات منفی آن روی اجزای مختلف مدل شناخته‌شده است. به‌عنوان مثال حضور همخطی در این مدل‌ها تورم واریانس و یا حذف متغیرهای مهم از مدل را به همراه دارد. برای جزئیات بیشتر درباره اثرات همخطی می‌توان به بلسلی [۷] مراجعه نمود. در صورت وجود همخطی بین متغیرهای توضیحی برآوردگرهای جایگزین دیگری پیشنهادشده‌اند که عموماً اریب می‌باشند. از میان این برآوردگرها، برآوردگر ریج هورل و کنارد [۸] نگاه‌های گسترده‌ای را به خود معطوف داشته است. دیدگاه رگرسیون ریج اخیراً در مدل‌های خطی نیمه‌پارامتری معمولی(بدون خطا در اندازه‌گیری) توسط بسیاری از پژوهشگران ازجمله دوران و آکدنیس[۹]، روزبه و آرشی [۱۰] و روزبه [۱۱] توسعه‌یافته است. با گسترش مدل‌های خطی نیمه‌پارامتری ریج، امامی ([۱۲] و[۱۳]) تحلیل مباحث تشخیصی را به ترتیب از دو رویکرد تحلیل حذف موردی و تحلیل تأثیر موضعی موردتحقیق و کنکاش قرار داده است.

یکی دیگر از مهم‌ترین فرضیه‌های پایه تحلیل‌های آماری این است که مشاهدات مدل درست اندازه‌گیری شده‌اند. نقض چنین فرضی حاکی از نفوذ خطای اندازه‌گیری در داده‌هاست که در صورت وجود موجب کم اعتباری آزمون‌های آماری است. در رابطه با این موضوع به فولر [۱۴] رجوع شود.در مواجه با نقض چنین فرضیه‌ای دو روش در پژوهش هانفلت و لیانگ [۱۵] اشاره‌شده است. روش نخست استفاده از درستنمایی تصحیح‌شده ناکامورا [۱۶] است که به‌خوبی خطاهای اندازه‌گیری را در مدل‌های نرمال، پوآسن و گاما تصحیح می‌کند. به نظر می‌رسد استفاده از این روش برای محاسبات کاربردی راحت‌تر است. روش دوم توسط استفانسکی و کارول[۱۷] مطرح‌شده است که بیشتر براساس یک تابع از برآوردگر ناریب است تا تقریب مناسبی از تابع درستنمایی. برای رویارویی با مشکل وجود توأم همخطی چندگانه و خطا در اندازه‌گیری در مدل‌های خطی پارامتری، راسخ [۱۸]، صالح و شالب [۱۹] (با استفاده از روش دوم) و قپانی و همکاران[۲۰] (با استفاده از روش نخست) استفاده از برآوردگرهای ریج را پیشنهاد داده‌اند. ازآنجایی‌که برآوردگرهای مطرح‌شده در مقابل مشاهدات مؤثر و دورافتاده حساس هستند راسخ [۲۱] و قپانی و همکاران[۲۲] به ترتیب با روش تحلیل تأثیر موضعی و روش انتقال میانگین

مباحث تشخیصی در مدل‌های خطی پارامتری ریج با خطا در اندازه‌گیری را موردمطالعه و بررسی قرار داده‌اند.

با توجه به مطالب بالا حضور توأم همخطی و مشاهدات مؤثر در مدل‌های خطی نیمه‌پارامتری با خطا در اندازه‌گیری دور از امکان نیست. ازنظر نویسندگان تاکنون پژوهشی درزمینهٔ رویارویی با همخطی و طرح مباحث تشخیصی در مدل‌های خطی نیمه‌پارامتری با خطا در اندازه‌گیری با مشکل همخطی موردتوجه قرار نگرفته است. بنابراین در این مقاله ما برای رویارویی با مشکل همخطی ابتدا برآوردگر ریج را در مدل‌های خطی نیمه‌پارامتری با خطا در اندازه‌گیری و با استفاده از ماکزیمم درستنمایی تصحیح‌شده پیشنهاد می‌دهیم و سپس با استفاده از روش حذف موردی و تحلیل تأثیر موضعی مباحث تحلیل تشخیصی جهت شناسایی مشاهدات دورافتاده را توسعه‌خواهیم داد.لذا در بخش بعدی مدل خطی نیمه پارامتری با خطا در اندازه‌گیری تعریف می‌شود. در بخش ۳ برآوردگرهای ریج با استفاده از مدل اسپلاین جزئی به دست می‌آیند و خواص مجانبی بخش پارامتری مدل بررسی می‌شود سپس برآورد ریج بخش پارامتری مدل با استفاده از معیار میانگین مربعات خطا با برآوردگر غیر ریج مقایسه می‌شود. در بخش ۴ شاخص‌های تشخیصی با استفاده از دو رویکرد حذف موردی و تأثیر موضعی بحث شده‌اند. در بخش ۵ داده‌های شبیه‌سازی و داده‌های سفال مصری مورد تحلیل و بررسی قرار می‌گیرد و در بخش ۶ نتایج کلی مطرح می‌شود.

۲-پیش‌زمینه و تعاریف

متغیرهای توضیحی در مدل‌های خطی نیمه‌پارامتری ممکن است با یک خطای غیرقابل اغماضی اندازه‌گیری شده باشند. در چنین مواقعی در نظر گرفتن مدل‌های خطی نیمه‌پارامتری با اندازه-گیری در خطا به شکل زیر می‌تواند مطلوب باشد.

$$\begin{aligned} y &= \mathbf{Z}\beta + g(t) + \epsilon \\ \mathbf{X} &= \mathbf{Z} + \boldsymbol{\delta} \end{aligned} \qquad (۱)$$

در مدل بالا $y=(y_1,\ldots,y_n)'$ بردار n تایی مشاهدات و $\mathbf{Z}=(z_1,\ldots,z_n)'$ ماتریس $n\times p$ از متغیرهای توضیحی پنهان است. بردار $g(\mathrm{t})=(g(t_1),\ldots,g(t_n))'$ مقادیر تابع نامعلوم نا پارامتری است. معمولاً t_iها روی یک بازه واحد کراندار هستند و به شکل $t_1\le t_2\le\ldots\le t_n$ مرتب می‌شوند. ϵ یک بردار با n مؤلفه از خطاهای مستقل و دارای توزیع نرمال با میانگین صفر و واریانس $\sigma^2\mathbf{I}_n$است که در آن $\mathbf{I}_n$ ماتریس همانی از مرتبه n است. ماتریس $\mathbf{X}=(x_1,\ldots,x_n)'$، ماتریس مشاهده‌شده $\mathbf{Z}$ با خطای اندازه‌گیری $\boldsymbol{\delta}$ است. $\boldsymbol{\delta}$ یک ماتریس تصادفی $n\times p$ مستقل از مجموعه $(y,\mathbf{X},t,\varepsilon)$ است که سطرهای آن از هم مستقل‌اند طوری

که هر سطر دارای توزیع نرمال با میانگین صفر و کوواریانس ماتریس $\mathbf{\Lambda}$ (بعد سطری و ستونی p) است.در واقع $\boldsymbol{\delta}\sim N\left(\circ,\mathbf{I}_n\otimes\mathbf{\Lambda}\right)$ که در آن $\otimes$ علامت ضرب کرونکر است.

۲-۱- روش درستنمایی تصحیح‌شده

در این بخش به‌اختصار روش درستنمایی تصحیح‌شده در مدل‌های خطی با خطا در اندازه‌گیری توضیح داده می‌شود.

فرض کنید تابع لگاریتم درستنمایی $l\left(\beta,\mathbf{Z},y\right)$، تابع امتیاز $U\left(\beta,\mathbf{Z},y\right)$، تابع فیشر $I\left(\beta,\mathbf{Z}\right)$ و تابع اطلاع مشاهده‌شده $J\left(\beta,\mathbf{Z},y\right)=-\partial U\left(\beta,\mathbf{Z},y\right)/\partial\beta$، توابعی نسبت به β با شرط $\mathbf{Z}$ و y باشند. اگر $\beta_\circ$ مقدار پارامتر درست و E^+ امید نسبت به متغیر تصادفی y باشد بنابراین

$$E^+\{U\left(\beta_\circ,\mathbf{Z},y\right)\}=\circ,E^+\{J\left(\beta,\mathbf{Z},y\right)\}=\mathbf{I}\left(\beta,\mathbf{Z}\right). \qquad (۲)$$

وقتی‌که $\mathbf{Z}$ منوط به خطا و $\mathbf{X}$ مقدار مشاهده‌شده $\mathbf{Z}$ است بنابراین رابطه $E^+\{U\left(\beta_\circ,\mathbf{X},y\right)\}=\circ$ الزاماً برقرار نیست و برآوردی که از رابطه $U\left(\beta,\mathbf{X},y\right)=\circ$ به دست می‌آید به‌طور حتم سازگار نخواهد بود. برای تصحیح این رابطه ناکامورا (۱۹۹۰) تابع درستنمایی تصحیح‌شده $l^*\left(\beta,\mathbf{X},y\right)$ را پیشنهاد داد که در رابطه زیر صدق می‌کند.

$$E^*\left\{l^*\left(\beta,\mathbf{X},y\right)\right\}=l\left(\beta,\mathbf{Z},y\right)$$

که در آن E^* میانگین شرطی نسبت به $\mathbf{X}$ به‌شرط $\mathbf{Z}$ و y داده‌شده است.

فرض کنید $U^*\left(\beta,\mathbf{X},y\right)=\partial l^*\left(\beta,\mathbf{X},y\right)/\partial\beta$ و $J^*\left(\beta,\mathbf{X},y\right)=-\partial U^*\left(\beta,\mathbf{X},y\right)/\partial\beta$ به ترتیب تابع امتیاز و تابع اطلاع تصحیح‌شده باشند. اگر E^* و β قابل تعویض باشند خواهیم داشت:

$$E^*\left\{U^*\left(\beta,\mathbf{X},y\right)\right\}=U\left(\beta,\mathbf{Z},y\right),E^*\left\{J^*\left(\beta,\mathbf{X},y\right)\right\}=J\left(\beta,\mathbf{Z},y\right) \quad (۳)$$

مقدار $\hat{\beta}$ که از رابطه $U^*\left(\hat{\beta},\mathbf{X},y\right)=\circ$ حاصل می‌شود یک برآوردگر تصحیح‌شده درستنمایی از β نامیده می‌شود.با توجه به توضیحات ناکامورا [۱۶]تابع درستنمایی تصحیح‌شده l^* و برآورد تصحیح‌شده درستنمایی $\hat{\beta}$ دارای ویژگی‌های خوبی است (هانفلت و لیانگ [۱۵]). فرض کنید $E=E^+E^*$ امید کامل را نشان دهد آنگاه از رابطه (۲) و (۳) خواهیم داشت:

$$E\left\{U^*\left(\beta_0,\mathbf{X},y\right)\right\}=E^+E^*\left\{U^*\left(\beta_0,\mathbf{X},y\right)\right\}=E^+\left\{U^*\left(\beta_0,\mathbf{Z},y\right)\right\}=0 \quad (۴)$$

رابطه (۴) نشان می‌دهد که تابع امتیاز تصحیح‌شده ناریب است. می‌توان نشان داد که تحت برخی شرایط نظم $\hat{\beta}$ دارای خواص سازگاری و توزیع نرمال مجانبی است (ناکامورا [۱۶]).

۳-مدل‌های اسپلاین جزئی و برآورد ریج

مشکل همخطی در داده‌ها برآورد پارامترهای رگرسیون را به‌شدت تحت تأثیر قرار می‌دهد. در مواجه با همخطی در مدل‌های رگرسیونی به‌جای استفاده از برآوردگرهای ناریب معمولی استفاده از برآوردگرهای اریب مناسب پیشنهاد می‌شود. از میان این برآوردگرها، برآورد رگرسیون ریج هورل و کنارد [۲۳] حائز اهمیت است.همان‌طور که قبلاً گفته شد امکان بروز همخطی در مدل-های خطی نیمه‌پارامتری دور از انتظار نیست. بنابراین در این بخش، با توجه به پژوهش‌های روزبه و آرشی [۱۰] و دوران و همکاران [۲۴] روش‌های رگرسیون ریج با استفاده از تابع درستنمایی تصحیح‌شده تاوانیده به مدل‌های خطی نیمه‌پارامتری با خطا در اندازه‌گیری تعمیم می‌یابد تا برآوردگرهای ریج مؤلفه‌های پارامتری و نا پارامتری حاصل شوند. بدین منظور ابتدا باید تابع ماکزیمم درستنمایی تاوانیده تصحیح‌شده به دست آید. لذا ابتدا: تابع لگاریتم درستنمایی توان-های دوم تاوانیده ریج به شکل زیر تعریف می‌شود:

$$l(\theta,\mathbf{Z},y)=\frac{-n}{2}\log\left(2\pi\sigma^2\right) -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{y_i-z_i'\beta-g(t_i)\right\}^2-\frac{\lambda}{2\sigma^2}\int g''(t)^2dt-\frac{k}{2\sigma^2}(\beta'\beta-d) \quad (۵)$$

که $k>0$ ضریب لاگرانژ و $\theta=(\beta,g)'$ است. در این مقاله انتخاب پارامتر هموارساز $\lambda>0$ با مینیمم کردن معیار اعتبار سنجی تقاطعی $GCV(\lambda)$ به دست می‌آید. برآورد پارامترها را می‌توان شبیه به مدل اسپلاین جزئی به دست آورد. توضیحات بیشتر در مورد مدل‌های اسپلاین جزئی و کاربرد آماری آن‌ها در گرین و سیلورمن (۱۹۹۴) بیان‌شده است.فرض کنید مقادیر مرتب‌شده و مجزای $t_1,\ldots,t_n$ با $s_1,\ldots s_q$ نشان داده شود. رابطه بین $t_1,\ldots,t_n$ و $s_1,\ldots s_q$ توسط میانگین ماتریس $n\times q$ با نام $\mathbf{N}$ با درایه‌های $\mathbf{N}_{ij}=1$ اگر $t_i=s_j$ و در غیر این صورت $\mathbf{N}_{ij}=0$، توضیح داده می‌شود. ازآنجایی‌که فرض شده است t_iها همه یکسان نیستند بنابراین $q\geq 2$ خواهد بود. اکنون می‌توان g را به شکل برداری از مقادیر $g(s_j)=a_i$ نوشت.با توجه به گرین و سیلورمن [۲۵]ماتریسی مانند $\mathbf{M}$ وجود دارد که تنها به گره‌های $\{s_j\}$ وابسته است طوری که مینیمم مقادیر $\int g''(t)^2dt$ مساوی $g'\mathbf{M}g$ است. بنابراین تابع لگاریتم درستنمایی توان‌های دوم تاوانیده ریج به شکل $l(\beta,\mathbf{Z},y)$

$$l(\theta,\mathbf{Z},y)=\frac{-n}{۲}\log\left(۲\pi\sigma^{۲}\right)-\frac{۱}{۲\sigma^{۲}}\|y-\mathbf{Z}\beta-\mathbf{N}g\|^{۲}$$
$$-\frac{\lambda}{۲\sigma^{۲}}g'\mathbf{M}g-\frac{k}{۲\sigma^{۲}}(\beta'\beta-d). \qquad (۶)$$

بازنویسی می‌شود. چون متغیر $\boldsymbol{Z}$ قابل‌مشاهده نیست برای برآورد پارامترها از روش درستنمایی تصحیح‌شده استفاده می‌کنیم. برای این کار لگاریتم درستنمایی تصحیح‌شده جمع توان‌های دوم تاوانیده ریج به شکل

$$l^{*}(\theta,\mathbf{X},y)=\frac{-n}{۲}\log\log\left(۲\pi\sigma^{۲}\right)$$
$$-\frac{۱}{۲\sigma^{۲}}\left\{\|y-\mathbf{X}\beta-\mathbf{N}g\|^{۲}-n\beta'\mathbf{\Lambda}\beta\right\}-\frac{\lambda}{۲\sigma^{۲}}g'\mathbf{M}g-\frac{k}{۲\sigma^{۲}}(\beta'\beta-d) \qquad (۷)$$

خواهد بود، که با مشتق‌گیری (۷) نسبت به β و g، برآوردگرهای ریج $\hat{\beta}_k$ و $\hat{g}_k$ به ترتیب به‌صورت

$$\hat{\beta}_k=\{\mathbf{X}'(\mathbf{I_n}-\mathbf{S})\mathbf{X}+k\mathbf{I_p}-n\mathbf{\Lambda}\}^{-۱}\mathbf{X}'(\mathbf{I_n}-\mathbf{S})y \qquad (۸)$$

$$\hat{g}_k=(\mathbf{N}'\mathbf{N}+\lambda\mathbf{M})^{-۱}\mathbf{N}'\left(y-\mathbf{X}\hat{\beta}_k\right) \qquad (۹)$$

به دست می‌آیند (ضمیمه رابطه (۱) را ببینید)، که در آن $\mathbf{S}$ یک ماتریس هموارساز و برابر $\mathbf{S}=\mathbf{N}(\mathbf{N}'\mathbf{N}+\lambda\mathbf{M})^{-۱}\mathbf{N}'$است. با قرار دادن $k=۰$ در رابطه (۸) و (۹) برآوردگرهای معمولی β و g برابر است با:

$$\hat{\beta}=\{\mathbf{X}'(\mathbf{I}_n-\mathbf{S})\mathbf{X}-n\mathbf{\Lambda}\}^{-۱}\mathbf{X}'(\mathbf{I}_n-\mathbf{S})y$$
$$\hat{g}=(\mathbf{N}'\mathbf{N}+\lambda\mathbf{M})^{-۱}\mathbf{N}'\left(y-\mathbf{X}\hat{\beta}\right).$$

در این حالت برآورد ماکزیمم درستنمایی تصحیح‌شده برای $\sigma^{۲}$ برابر است با:

$$\hat{\sigma}^{۲}=\frac{۱}{n}\left\{\|y-\mathbf{X}\hat{\beta}-\mathbf{N}\hat{g}\|^{۲}-n\beta'\mathbf{\Lambda}\beta+\hat{\lambda}\hat{g}'\mathbf{M}\hat{g}\right\}$$

از رابطه (۸) داریم:

$$\mathbf{X}\hat{\beta}_k=\mathbf{X}\{\mathbf{X}'(\mathbf{I}_n-\mathbf{S})\mathbf{X}+k\mathbf{I}_p-n\mathbf{\Lambda}\}^{-۱}\mathbf{X}'(\mathbf{I}_n-\mathbf{S})y=\tilde{\mathbf{H}}_k y$$

و از رابطه (۹) خواهیم داشت:

$$\mathbf{N}\hat{\mathrm{g}}_k = \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{M})^{-1}\mathbf{N}'\left(y - \mathbf{X}\hat{\beta}_k\right) = \mathbf{S}\left(\mathbf{I}_n - \tilde{\mathbf{H}}\right)y = \mathbf{H}_k^{*}y.$$

از روابط (۸) و (۹) مقادیر تقریبی $\hat{y}$ را می‌توان به شکل

$$\hat{y} = \mathbf{X}\beta_k + \mathbf{N}g_k = \mathbf{H}y$$

نوشت. که در آن

$$\mathbf{H} = \mathbf{S} + \left(\mathbf{I}_n - \mathbf{S}\right)\mathbf{X}\{\mathbf{X}'\left(\mathbf{I}_n - \mathbf{S}\right)\mathbf{X} + k\mathbf{I}_p - \mathrm{n}\mathbf{\Lambda}\}^{-1}\mathbf{X}'\left(\mathbf{I}_n - \mathbf{S}\right) = \tilde{\mathbf{H}}_k + \mathbf{H}_k^{*},$$

ماتریس مقادیر نافذ می‌باشد. بردار باقیمانده‌ها از رابطه

$$e = \left(\mathbf{I}_n - \mathbf{H}\right)y,$$

به دست خواهد. یکی از نکات کلیدی در رگرسیون ریج انتخاب پارامتر k است. روش‌های متفاوتی برای انتخاب این پارامتر ارائه‌شده است ازجمله روش تکراری هورل و کنارد [۲۳]، معیار آماره C_p و معیار GCV روش‌های عمومی هستند.

۳-۱- بررسی خواص مجانبی برآورد ریج پارامتر مدل

در این بخش ابتدا توزیع مجانبی برآورد $\hat{\beta}_k$ در قالب یک قضیه به‌اختصار بررسی می‌شود. سپس با توجه به نتایج این قضیه دو برآوردگر $\hat{\beta}_k$ و $\hat{\beta}$ طبق معیار میانگین مربعات خطای مجانبی باهم مقایسه می‌شوند.فرض می‌شود که مقادیر حدی، $\lim_{n\to\infty} n^{-1}\mathbf{Z}'\left(\mathbf{I}_n - \mathbf{S}\right)^{2}\mathbf{Z}$ $\lim_{n\to\infty} n^{-1}\mathbf{Z}'\left(\mathbf{I}_n - \mathbf{S}\right)\mathbf{Z}$ و $\lim_{n\to\infty} n^{-1}\mathbf{D}$ وجود دارند، که $\mathbf{D} = \left(\mathbf{Z}\beta + \mathbf{N}g\right)'\left(\mathbf{I}_{\mathrm{n}} - \mathbf{S}\right)^{2}\left(\mathbf{Z}\beta + \mathbf{N}g\right)$ است.

قضیه ۱: تحت فرضیات قبل، $\hat{\beta}_k$ دارای توزیع نرمال مجانبی با امید ریاضی و واریانس مجانبی

$$avar\left(\hat{\beta}_k\right) = \mathbf{C}_k^{-1}\mathbf{\Phi}\mathbf{C}_k^{-1}E\left(\hat{\beta}_k\right) = \mathbf{C}_{\circ}\mathbf{C}_k^{-1}\beta$$

است که در آن $\mathbf{C}_k = \mathbf{Z}'\left(\mathbf{I}_n - \mathbf{S}\right)\mathbf{Z} + k\mathbf{I}_p$، $\mathbf{C}_{\circ} = \mathbf{Z}'\left(\mathbf{I}_n - \mathbf{S}\right)\mathbf{Z}$ و $\mathbf{\Phi} = \mathbf{D} + \mathbf{Z}'\left(\mathbf{I}_n - \mathbf{S}\right)^{2}\mathbf{Z}\sigma^{2}$ هستند.

اثبات: ضمیمه را ببینید.

از قضیه **۱** می‌توان میانگین مربعات خطای مجانبی برآوردگر $\hat{\beta}_k$ را به شکل

$$amse\left(\hat{\beta}_k\right) = avar\left(\hat{\beta}_k\right) + bias\left(\hat{\beta}_k\right)bias\left(\hat{\beta}_k\right)'$$
$$= \mathbf{C}_k^{-1}\mathbf{\Phi}\mathbf{C}_k^{-1} + k^2\mathbf{C}_k^{-1}\beta\beta'\mathbf{C}_k^{-1}$$

به دست آورد. در ابتدا برای مقایسه $amse$ دو برآوردگر $\hat{\beta}_k$ و $\hat{\beta}$ برای $k > 0$ رابطه

$$\mathbf{\Pi} = avar\left(\hat{\beta}\right) - avar\left(\hat{\beta}_k\right)$$
$$= \mathbf{C}_0^{-1}\mathbf{\Phi}\mathbf{C}_{0\pm}^{-1} - \mathbf{C}_k^{-1}\mathbf{\Phi}\mathbf{C}_k^{-1} \succ 0$$

برقرار است. ماتریس $\mathbf{\Pi}$ معین مثبت است(طبق تعریف ۱ ضمیمه) چون $\mathbf{C}_k \succ \mathbf{C}_0$ بنابراین $\mathbf{C}_0^{-1} \succ \mathbf{C}_k^{-1}$ و درنتیجه $\mathbf{C}_0^{-1}\mathbf{\Phi}\mathbf{C}_0^{-1} \succ \mathbf{C}_k^{-1}\mathbf{\Phi}\mathbf{C}_k^{-1}$. اکنون با توجه به $\mathbf{\Pi}$ اختلاف بین ماتریس‌های میانگین مربعات خطای مجانبی $\hat{\beta}$ و $\hat{\beta}_k$ برابر با

$$\Delta_k = amse\left(\hat{\beta}\right) - amse\left(\hat{\beta}_k\right) = \Pi - k^2\mathbf{C}_k^{-1}\beta\beta'\mathbf{C}_k^{-1}$$

است. با توجه به لم۲ ضمیمه $\Delta_k \succ 0$ اگر و تنها اگر

$$d = k - \frac{1}{\sqrt{\beta'\mathbf{C}_k^{-1}\mathbf{\Pi}^{-1}\mathbf{C}_k^{-1}\beta}} \leq 0. \qquad (10)$$

۳-۲- برآورد k و λ

از رابطه (۱۰) مشخص هست که مقدار k مناسب را به‌راحتی نمی‌توان یافت. یک روش برای پیدا کردن k رسم Δ_k در برابر یک بازه پیوسته از k است. مقدار k متناظر به ماکزیمم Δ_k می‌تواند به‌عنوان k بهینه در نظر گرفته شود [۱۰]. برای پیدا کردن پارامتر ریج و پارامتر هموارساز مناسب از معیار اعتبار سنجی متقابل تعمیم‌یافته (GCV) استفاده می‌شود. اخیراً روزبه [۲۶] و آرشی و روزبه [۲۷] این روش را برای برآورد هم‌زمان پارامترهای k و λ به‌کاربرده‌اند. فرض کنید با حذف iامین سطر از داده‌ها مقدارهای $\hat{\beta}_{k(i)}$ و $\hat{g}_{k(i)}$ برآوردهای ریج β و g باشند. آنگاه برآورد مناسب k و λ مقدارهایی هستند که به ازای آن‌ها مؤلفه $\mathbf{W}\hat{\beta}_{k(i)} + \mathbf{N}\hat{g}_{k(i)}$ یعنی $\left[\mathbf{W}\hat{\beta}_{k(i)} + \mathbf{N}\hat{g}_{k(i)}\right]_i$ پیش‌بینی خوبی برای y_i است، به عبارتی مقدارهای مناسب برای k و λ مقدارهایی هستند که تابع زیر را مینیمم کنند:

$$GCV(k,\lambda) = n^{-1}\sum_{i=1}^{n}\left(y_i - \left[\mathbf{W}\hat{\beta}_{k(i)} + \mathbf{N}\hat{g}_{k(i)}\right]_i\right)^2$$

همچنین λ را می‌توان به‌جز روش هم‌زمان برآورد از طریق دومرحله‌ای نیز به دست آورد برای این کار با قرار دادن k معلوم مناسب در تابع بالا معیار اعتبارسنجی تعمیم‌یافته را می‌توان تنها برحسب λ نوشت و λ را با مینیمم کردن تابع حاصل به دست آورد [۱۰].

۴- شاخص‌های مؤثر در برآورد لگاریتم درستنمایی تصحیح‌شده جمع توان‌های دوم تاوانیده ریج

در مباحث ویژه برای شناسایی مشاهدات مؤثر که به شکل جدی نتایج تحلیل آماری را تحت تأثیر قرار می‌دهند دو رویکرد اصلی وجود دارد. رویکرد نخست براساس روش حذف موردی است که در آن اثر حذف یک مشاهده در برآوردها به‌طور مستقیم توسط برخی از معیارهای معتبر نظیر آماره کوک [۱] ارزیابی می‌شود. رویکرد دوم تأثیر موضعی است که توسط کوک [۲۸] مطرح‌شده است.در این رویکرد با واردکردن اغتشاش‌های جزئی به اجزای مختلف مدل به بررسی ثبات برآوردهای مدل پس از اغتشاش می‌پردازد. در ادامه ابتدا رویکرد حذف موردی و سپس تأثیر موضعی بحث می‌شود.

۴-۱- حذف موردی

فرض کنید وقتی مشاهده iاز مجموعه داده حذف‌شده است مقادیر $\hat{\beta}_{k(i)}$ و $\hat{g}_{k(i)}$ برآوردگرهای لگاریتم درستنمایی تصحیح‌شده تاوانیده ریج β و g هستند.

قضیه۲: برآورد پارامترهای θ بعد از حذف مشاهده i ام به شکل زیر به دست می‌آیند:

$$\hat{\beta}_{k(i)} = \hat{\beta}_k - \frac{\{\mathbf{X}'(\mathbf{I_n}-\mathbf{S})\mathbf{X}+k\mathbf{I_p}-n\mathbf{\Lambda}\}^{-1}\mathbf{X}'(\mathbf{I_n}-\mathbf{S})e_i\xi_i}{1-h_{ii}}$$

$$\hat{g}_{k(i)} = \hat{g}_k - \frac{(\mathbf{N}'\mathbf{N}+\lambda\mathbf{M})^{-1}\mathbf{N}'[\mathbf{I_n}-\mathbf{X}\{\mathbf{X}'(\mathbf{I_n}-\mathbf{S})\mathbf{X}+k\mathbf{I_p}-n\mathbf{\Lambda}\}^{-1}\mathbf{X}'(\mathbf{I_n}-\mathbf{S})e_i\xi_i]}{1-h_{ii}},$$

که ξ_i برداری با بعد n است که i- اُمین درایه آن ۱ و سایر درایه‌های آن مقدار صفر می‌گیرند. e_i نیز i-اُمین مؤلفه از بردار باقیمانده تاوانیده ریج e است.

اثبات: فرض کنید $y^* = (y_1^*,\ldots,y_n^*)$ و $y_j^* = \begin{cases} y_i^*, j=i \\ y_j, j\neq i \end{cases}$ طوری که $y_i^* = x_i'\hat{\beta}_{k(i)} + \hat{g}_{k(i)}(t_i)$ است. با توجه به تعریف $\hat{\beta}_{k(i)}$ و $\hat{g}_{k(i)}$ مقادیری هستند که تابع زیر را برای هر β و تابع هموار g مینیمم کنند:

$$\| y^* - \mathbf{X}\beta - \mathbf{N}g \|^2 + \lambda \int g''(t)^2 dt + k\left(\beta'\beta - d\right) - n\beta'\mathbf{\Lambda}\beta$$

$$\geq \sum_{j \neq i} (y_j - x_j'\beta - g(t_i))^2 + \lambda \int g''(t)^2 dt + k\left(\beta'\beta - d\right) - n\beta'\mathbf{\Lambda}\beta \qquad (11)$$

$$\geq \| y^* - \mathbf{X}\hat{\beta}_{k(i)} - \mathbf{N}\hat{g}_{k(i)} \|^2 + \lambda \int \hat{g}_{k(i)}''(t)^2 dt + k\left(\hat{\beta}_{k(i)}'\hat{\beta}_{k(i)} - d\right) - n\hat{\beta}_{k(i)}'\mathbf{\Lambda}\hat{\beta}_{k(i)}$$

با توجه به کران پایین تابع (۱۱)، واضح است مقدار مینیمم تابع با جایگزینی $\beta_{k(i)} = \hat{\beta}_{k(i)}$ و $g_{k(i)} = \hat{g}_{k(i)}$ در رابطه

$$\| y^* - \mathbf{X}\beta_{k(i)} - \mathbf{N}g_{k(i)} \|^2 + \lambda g_{k(i)}'\mathbf{M}g_{k(i)} + k\left(\beta_{k(i)}'\beta_{k(i)} - d\right) - n\beta_{k(i)}'\mathbf{\Lambda}\beta_{k(i)}$$

حاصل می‌شود. بنابراین با مشتق‌گیری از رابطه بالا نسبت به $\beta_{k(i)}$ و $g_{k(i)}$ برآوردها برابر

$$\hat{\beta}_{k(i)} = \{\mathbf{X}'(\mathbf{I}_n - \mathbf{S})\mathbf{X} + k\mathbf{I}_p - n\mathbf{\Lambda}\}^{-1}\mathbf{X}'(\mathbf{I}_n - \mathbf{S})y^*$$

$$\hat{g}_{k(i)} = (\mathbf{N}'\mathbf{N} + \lambda\mathbf{M})^{-1}\mathbf{N}'\left(y^* - \mathbf{X}\hat{\beta}_{k(i)}\right)$$

به دست می‌آیند (کافی است در رابطه (۱) ضمیمه به‌جای y مقدار y^* و به‌جای β و g مقادیر $\beta_{k(i)}$ و $g_{k(i)}$ قرار گیرند). با قرار دادن $y^* = y - (y - y^*)$ در روابط بالا و ذکر این نکته که $y - y^* = \xi_i(y_i - y_i^*)$، رابطه

$$\hat{\beta}_{k(i)} = \hat{\beta}_k - \{\mathbf{X}'(\mathbf{I}_n - \mathbf{S})\mathbf{X} + k\mathbf{I}_p - n\mathbf{\Lambda}\}^{-1}\mathbf{X}'(\mathbf{I}_n - \mathbf{S})\xi_i(y_i - y_i^*)$$

حاصل می‌شود. در ادامه برای $\hat{g}_{k(i)}$ رابطه زیر به دست می‌آید:

$$\hat{g}_{k(i)} = \hat{g}_k - (\mathbf{N}'\mathbf{N} + \lambda\mathbf{M})^{-1}\mathbf{N}'[\mathbf{I}_n - \mathbf{X}\{\mathbf{X}'(\mathbf{I}_n - \mathbf{S})\mathbf{X} + k\mathbf{I}_p - n\mathbf{\Lambda}\}\mathbf{X}'(\mathbf{I}_n - \mathbf{S})]\xi_i(y_i - y_i^*)$$

از طرفی داریم:

$$\begin{aligned} y_i - y_i^* &= \xi_i'\left(y - \mathbf{X}\hat{\beta}_{k(i)} - \mathbf{N}\hat{g}_{k(i)}\right) = \xi_i'\left(y - \mathbf{S}y^* - (\mathbf{I}_n - \mathbf{S})\mathbf{X}\hat{\beta}_{k(i)}\right) \\ &= \xi_i'\left(y - \mathbf{H}y^*\right) = \xi_i'\left[y - \mathbf{H}y + \mathbf{H}\left(y - y^*\right)\right] \\ &= \xi_i'\left(y - \mathbf{H}y\right) + \xi_i'\mathbf{H}\xi_i\left(y_i - y_i^*\right) = e_i + h_{ii}\left(y_i - y_i^*\right). \end{aligned}$$

بنابراین $y_i - y_i^* = \frac{e_i}{1 - h_{ii}}$ به دست می‌آید و اثبات کامل می‌شود.

۴-۱-۱ تأثیر روی $\hat{\beta}_k$

حداقل دو نوع آماره کوک به شکل زیر می‌تواند برای برآورد لگاریتم درستنمایی تصحیح‌شده کمترین توان‌های دوم تاوانیده ریج می‌توان مطرح شود.

$$D_{\beta i} = \frac{(\hat{\beta}_k - \hat{\beta}_{k(i)})' \mathbf{X}'\mathbf{X}\left(\hat{\beta}_k - \hat{\beta}_{k(i)}\right)}{\sigma^2 tr\left(\tilde{\mathbf{H}}_{\mathbf{k}}\right)}$$

$$D^*_{\beta i} = \frac{(\hat{\beta}_k - \hat{\beta}_{k(i)})' \{\mathbf{X}'(\mathbf{I}_n - \mathbf{S})\mathbf{X} + k\mathbf{I}_p - n\mathbf{\Lambda}\}(\mathbf{X}'(\mathbf{I}_n - \mathbf{S})\mathbf{X})^{-1} \{\mathbf{X}'(\mathbf{I}_n - \mathbf{S})\mathbf{X} + k\mathbf{I}_p - n\mathbf{\Lambda}\}\left(\hat{\beta}_k - \hat{\beta}_{k(i)}\right)}{\sigma^2 tr\left(\tilde{\mathbf{H}}_k\right)}$$

که در آن $\tilde{\mathbf{H}}_k = \mathbf{X}\{\mathbf{X}'(\mathbf{I}_n - \mathbf{S})\mathbf{X} + k\mathbf{I}_p - n\mathbf{\Lambda}\}^{-1}\mathbf{X}'(\mathbf{I}_n - \mathbf{S})$ با درایه $\tilde{h}_{ij}$ است. $D_{\beta i}$ را می‌توان با استفاده از روابط قضیه ۱ به‌صورت تابعی از باقیمانده‌ها و داده‌های نافذ نیز نوشت (ضمیمه را ببینید)

$$D_{\beta i} = \frac{\left[\sum_{j=1}^{n} \tilde{h}_{ij}^2\right] e_i^2}{\sigma^2 tr\left(\tilde{\mathbf{H}}_k\right)(1 - h_{ii})^2}$$

۴-۱-۲ تأثیر روی $\hat{g}_k$

برای سنجش تأثیر i- اُمین مشاهده روی $\hat{g}_k$ فاصله کوک زیر تعریف می‌شود:

$$D_{gi} = \frac{(\hat{g}_{k(i)} - \hat{g}_k)' \mathbf{N}'\mathbf{N}\left(\hat{g}_{k(i)} - \hat{g}_k\right)}{\sigma^2 tr\left(\mathbf{H}^*_k\right)(1 - h_{ii})^2}$$

شبیه $D_{\beta i}$، D_{gi} را می‌توان به‌صورت باقیمانده‌ها و مقادیر نافذ به شکل زیر نوشت:

$$D_{gi} = \frac{\left[\sum_{j=1}^{n} h^{*2}_{ij}\right] e_i^2}{\sigma^2 tr\left(\mathbf{H}^*_k\right)(1 - h_{ii})^2}$$

که h^*_{ij} درایه ماتریس $\mathbf{H}^*_k = \mathbf{S}\left(\mathbf{I}_n - \tilde{\mathbf{H}}_k\right)$ است.

۴-۱-۳ تأثیر روی $\hat{y}_k$

معیار سنجش برای میزان تأثیر i اُمین مشاهده روی بردار مقادیر برازش به شکل

$$D_{yi}=\frac{(\hat{y}_k-\hat{y}_{k(i)})'\left(\hat{y}_k-\hat{y}_{k(i)}\right)}{\sigma^2 tr(\mathbf{H})}$$

تعریف می‌شود. D_{yi} را می‌توان به‌صورت تقریبی تابعی از باقیمانده‌ها و داده‌های نافذ نیز نوشت:

$$D_{yi}\approx\frac{h_{ii}e_i^2}{\sigma^2 tr(\mathbf{H})(1-h_{ii})^2}$$

۴-۱-۴ روابط بین فاصله کوک‌ها

برای بررسی رابطه نظری بین D_{yi}، D_{gi}، $D_{\beta i}$ با توجه به $\hat{y}=\mathbf{X}\hat{\beta}_k+\mathbf{N}\hat{g}_k$ و محاسبات ۴-۱-۱ و ۴-۱-۲ خواهیم داشت.

$$\begin{aligned}\sigma^2 tr(\mathbf{H})D_{yi}&=(\hat{y}_k-\hat{y}_{k(i)})'\left(\hat{y}_k-\hat{y}_{k(i)}\right)\\&=\{\mathbf{X}\left(\hat{\beta}_k-\hat{\beta}_{k(i)}\right)+\mathbf{N}\left(\hat{g}_{k(i)}-\hat{g}_k\right)\}'\left\{\mathbf{X}\left(\hat{\beta}_k-\hat{\beta}_{k(i)}\right)+\mathbf{N}\left(\hat{g}_{k(i)}-\hat{g}_k\right)\right\}\\&=(\hat{\beta}_k-\hat{\beta}_{k(i)})'\mathbf{X}'\mathbf{X}\left(\hat{\beta}_k-\hat{\beta}_{k(i)}\right)+(\hat{g}_{k(i)}-\hat{g}_k)'\mathbf{N}'\mathbf{N}\left(\hat{g}_{k(i)}-\hat{g}_k\right)\\&+2(\hat{\beta}_k-\hat{\beta}_{k(i)})'\mathbf{X}'\mathbf{N}\left(\hat{g}_{k(i)}-\hat{g}_k\right)\\&=\sigma^2 tr\left(\tilde{\mathbf{H}}_k\right)D_{\beta i}+\sigma^2 tr\left(\mathbf{H}_k^*\right)D_{gi}+2(\hat{\beta}_k-\hat{\beta}_{k(i)})'\mathbf{X}'\mathbf{N}\left(\hat{g}_{k(i)}-\hat{g}_k\right)\end{aligned}$$

۴-۲- رویکرد تأثیر موضعی

رویکرد تأثیر موضعی ابتدا توسط کوک [۲۸] معرفی شد و سپس سایر افراد ازجمله توماس و کوک [۲۹] و کوان و فونگ [۳۰] آن را موردبررسی و کنکاش بیشتری قراردادند. در این بخش ابتدا روابط پایه‌ای تأثیر موضعی مرور می‌شود و سپس این رویکرد به مدل‌های خطی ریج نیمه-پارامتری با خطا در اندازه‌گیری تعمیم داده می‌شود. شایان‌ذکر است که این تعمیم بر پایه لگاریتم درستنمایی تصحیح‌شده تاوانیده است.مطابق روش کوک [۲۸]، ابتدا انحراف ω را که ماهیت آن در ادامه بیشتر مشخص می‌شود بررسی می‌کنیم. ابتدا فرض می‌شود که مقدار ω برای داده‌ها و مدل اصلی برابر صفر است. $\hat{\theta}_\omega$ برآورد θ تحت انحراف ω است. لگاریتم درستنمایی تصحیح‌شده تاوانیده تحت انحراف داده‌شده با $l^*(\theta,\omega)$ نشان داده می‌شود طوری که $l^*(\theta,0)$ برابر با $l^*(\theta)$ در رابطه (۷) است و برای هر ω، $l^*(\theta,\omega)$ در $\theta=\hat{\theta}_\omega$ مینیمم می‌شود. با

هر بردار واحد سویه d در فضای ω، انحنا نرمال نگاشت از ω به $l^*\left(\hat{\theta}_\omega\right)$ در $\omega=0$ و در جهت d برابر $d'\ddot{\mathbf{F}}d$ تعریف می‌شود که در آن $\ddot{\mathbf{F}}$ برابر

$$\ddot{\mathbf{F}}=\frac{\partial^2 l^*\left(\hat{\theta}_\omega\right)}{\partial\omega\partial\omega'} \qquad (۱۲)$$

است. در نقطه $\omega=0$، $\ddot{\mathbf{F}}$ ماتریس تأثیر برای θ نامیده می‌شود. ازآنجایی‌که $\frac{\partial l^*\left(\hat{\theta}_\omega,\omega\right)}{\partial\theta}=0$ است محاسبات مستقیم با استفاده از قاعده زنجیره‌ای رابطه

$$\ddot{\mathbf{F}}=\Delta'\ddot{\mathbf{L}}^{-1}\Delta \qquad (۱۳)$$

را نتیجه می‌دهد. که در آن $\Delta=\frac{\partial^2 l^*(\theta,\omega)}{\partial\theta\partial\omega'}$ در نقطه $\theta=\hat{\theta}$ و $\omega=0$ است و عبارت میانی $\ddot{\mathbf{L}}=\frac{\partial^2 l^*(\theta)}{\partial\theta\partial\theta'}$ در $\theta=\hat{\theta}$ در نظر گرفته می‌شود. کوک [۲۸] نشان داد که ساختار ویژه $\ddot{\mathbf{F}}$ برای اهداف تشخیصی می‌تواند مفید باشد. عناصر قطری $\ddot{\mathbf{F}}$، حساسیت موضعی به لگاریتم درستنمایی را توسط انحرافات مؤلفه-محور اندازه‌گیری می‌کند. برای مثال، اگر پاسخ انحراف داده‌شده $y+\omega$ است لذا مقدار نسبتاً بزرگی از اولین عنصر قطری ماتریس تأثیر، اثر نسبتاً بزرگی از اولین مشاهده روی θ را نشان می‌دهد. هرچند تحلیل‌های تأثیر معمولاً برای اندازه‌های نمونه کوچک تعریف می‌شوند لکن انتظار می‌رود که اندازه‌های تأثیر روی g هنگامی‌که n بزرگ است، ماتریس تأثیر $\ddot{\mathbf{F}}$ را دربر داشته باشد.لذا توصیه می‌شود که تأثیر جزئی را با تمرکز روی جزئی از θ یعنی β در نظر بگیریم.بدین منظور، $\hat{g}$ را به‌عنوان تابعی از $\hat{\beta}$ در نظر می‌گیریم. و ماتریس تأثیر جزئی را از انحنای نرمال نگاشت شده از ω به $l^*\left(\hat{\beta}_\omega,\hat{g}\left(\hat{\beta}_\omega\right)\right)$ تعریف می‌کنیم. ماتریس تأثیر جزئی برای β از فرمول

$$\ddot{\mathbf{F}}(\beta)=\Delta'\left(\ddot{\mathbf{L}}^{-1}-\begin{pmatrix}0 & 0\\ 0 & \mathbf{L}_g^{-1}\end{pmatrix}\right)\Delta, \qquad (۱۴)$$

به دست می‌آید که $\mathbf{L}_g$ زیر ماتریسی از $\ddot{\mathbf{L}}$ می‌باشد که متناظر با g است. تأثیر جزئی برای بخش دیگر (یا معادل آن) از همان طریق ارائه می‌شود. برای جزئیات بیشتر به کوک ([۲۸]،ص۱۴) مراجعه نمایید.

۴-۲-۱ انحراف وزن‌های نمونه

نخست داده‌ها را با تغییر وزنی هر نمونه در معیارهای حداقل مربعات انحراف می‌دهیم. این کار معادل انحراف واریانس ϵ_i در مدل می‌باشد. با وزن‌های $\omega_i + ۱$ داده‌شده به i-اُمین مشاهده، رابطه زیر را داریم:

$$l^*(\theta,\omega) = \frac{-۱}{۲\sigma^۲}\sum_{i=۱}^{n}(۱+\omega_i)(y_i - x_i'\beta - g(t_i))^۲ + \frac{n}{۲\sigma^۲}\beta'\Lambda\beta - \frac{\lambda}{۲\sigma^۲}\int g''(t)^۲ dt$$
$$-\frac{k}{۲\sigma^۲}(\beta'\beta - d) - \frac{n}{۲}\log ۲\pi\sigma^۲$$

با $\omega = (\omega_۱,\ldots,\omega_n)' \in \mathbf{R}^n$. با محاسبه مستقیم رابطه‌های زیر

$$\Delta = \hat{\sigma}^۲(\mathbf{X},\mathbf{N})'\hat{\mathbf{E}} \quad \text{و} \quad \ddot{\mathbf{L}} = \begin{pmatrix} \mathbf{X}'\mathbf{X} + k\mathbf{I}_p - n\Lambda & \mathbf{X}'\mathbf{N} \\ \mathbf{N}'\mathbf{X} & \mathbf{N}'\mathbf{N} + \lambda\mathbf{M} \end{pmatrix}$$

به دست می‌آیند.که $\hat{\mathbf{E}} = diag\left(y - \mathbf{X}\hat{\beta}_k - \mathbf{N}\hat{g}_k\right)$، ماتریس قطری است. بنابراین ماتریس تأثیر

$$\ddot{\mathbf{F}}_\omega = \hat{\mathbf{E}}\mathbf{H}\hat{\mathbf{E}} \qquad (۱۵)$$

به دست می‌آید. عناصر قطری $\ddot{\mathbf{F}}_\omega$ کاملاً به منحنی‌های تأثیر نمونه بر پایه حذف نمونه مرتبط هستند (برای مثال کوک و ویزبرگ[۲] را ببینید). همچنین ماتریس‌های تأثیر جزئی برای مؤلفه‌های پارامتری به شکل زیر به دست می‌آیند:

$$\ddot{\mathbf{F}}_\omega(\beta) = \hat{\mathbf{E}}(\mathbf{H} - \mathbf{S})\hat{\mathbf{E}} \quad \text{و} \quad \ddot{\mathbf{F}}_\omega(g) = \hat{\mathbf{E}}(\mathbf{H} - \mathbf{X}(\mathbf{X}'\mathbf{X} - n\Lambda + k\mathbf{I}_\mathrm{p})^{-۱}\mathbf{X}')\hat{\mathbf{E}} \qquad (۱۶)$$

محاسبه ماتریس تأثیر (۱۵) آسان است، ولی همانند فاصله‌ی کوک طبق حذف موردی، وابسته به باقیمانده‌ها و مقادیر نافذ است. یک تفاوت آشکار رابطه (۱۵) بافاصله کوک این است که از باقیمانده‌های غیراستاندارد استفاده می‌کند. این کار می‌تواند به انتخاب انحرافات نسبت داده شود. اگر انحراف ω خود به‌طور صحیح مقیاس بندی شود، طوری که وزن σ_ω باشد، باقیمانده‌ها نیز در رابطه (۱۴) توسط σ استانداردسازی می‌شوند.

۴-۲-۲ انحراف متغیر پاسخ

اکنون با جایگزینی y با $y + \omega$ وقتی‌که $\omega \in R^n$ انحراف متغیر پاسخ را بررسی می‌کنیم، در این مورد خواهیم داشت:

$$l^*(\theta,\omega)=\frac{-1}{2\sigma^2}\| y+\omega-\mathbf{X}\beta-\mathbf{N}g\|^2 + \frac{n}{2\sigma^2}\beta'\mathbf{\Lambda}\beta-\frac{\lambda}{2\sigma^2}g'\mathbf{M}g-\frac{k}{2\sigma^2}(\beta'\beta-d)-\frac{n}{2}\log 2\pi\sigma^2.$$

آنگاه ماتریس تأثیر به شکل ساده

$$\ddot{\mathbf{F}}_y=\mathbf{H} \qquad (17)$$

به دست خواهد آمد و ماتریس‌های تأثیر جزئی برابر $\ddot{\mathbf{F}}_y(g)=\mathbf{H}-\mathbf{X}(\mathbf{X}'\mathbf{X}-n\mathbf{\Lambda}+k\mathbf{I}_p)^{-1}\mathbf{X}'$ است که در آن $\ddot{\mathbf{F}}_y(\beta)=\mathbf{H}-\mathbf{S}$ است. اینکه تأثیر موضعی تحت انحراف پاسخ متناظر به داده‌های نافذ است تعجب‌آور نیست. اینجا تأثیر جزئی ابزار درک مقادیر نافذ شرطی و حاشیه‌ای در مدل‌های خطی نیمه پارامتری را فراهم می‌کند. می‌توان $\mathbf{H}-\mathbf{S}$ را به‌عنوان ماتریس مقادیر نافذ جهت برآورد β بعد از تطبیق برآورد g و $\mathbf{H}-\mathbf{X}(\mathbf{X}'\mathbf{X}-n\mathbf{\Lambda}+k\mathbf{I}_p)^{-1}\mathbf{X}'$ را به‌عنوان ماتریس مقادیر نافذ برای برآورد g بعد از تطبیق برآورد β در نظر بگیریم. به‌طور مشابه می‌توان $\mathbf{X}(\mathbf{X}'\mathbf{X}-n\mathbf{\Lambda}+k\mathbf{I}_p)^{-1}\mathbf{X}'$ و $\mathbf{S}$ را به‌عنوان ماتریس‌های مقادیر نافذ حاشیه‌ای برای مؤلفه‌های پارامتری و نا پارامتری در نظر گرفت. به‌طور خلاصه داریم:

مقادیر نافذ شرطی + مقادیر نافذ حاشیه‌ای = مقادیر نافذ توأم

۳-۲-۴ انحراف متغیرهای کمکی

انحراف متغیرهای کمکی تأثیر بسیار پیچیده‌ای روی برآورد دارد. واضح است که خطاهای اندازه-گیری روی متغیرهای کمکی می‌تواند منجر به اریبی نسبتاً جدی در برآورد ضرایب رگرسیون خطی شود [۱۴]. در مدل‌های خطی نیمه پارامتری پژوهشگران اندکی روی مبحث اریبی کار کردند. تحلیل تأثیر موضعی تحت انحراف متغیرهای کمکی ممکن است یک دید متناوب به مدل‌های خطا در اندازه‌گیری فراهم نماید.

انحراف $\mathbf{X}$ به $\mathbf{X}_\omega=\mathbf{X}+\omega l_i'$ را که $\omega\in R^n$ و $l_i\in R^p$ یک بردار یکه با i-اُمین عنصر برابر با یک است، را در نظر بگیرید، طوری که تنها i-اُمین متغیر کمکی انحراف داده می‌شود. در این مورد رابطه

$$l^*(\theta,\omega)=\frac{-1}{2\sigma^2}\| \mathbf{Y}-\mathbf{X}\beta-\mathbf{N}g-\omega l_i'\beta\|^2+\frac{n}{2\sigma^2}\beta'\mathbf{\Lambda}\beta-\frac{\lambda}{2\sigma^2}g'\mathbf{M}g - \frac{k}{2\sigma^2}(\beta'\beta-d)-\frac{n}{2}\log 2\pi\sigma^2.$$

برقرار است. با توجه به محاسبات در ضمیمه، ماتریس تأثیر رابطه

$$\ddot{\mathbf{F}}_{x,i} = (l_i'\hat{\beta}_k)^2\mathbf{S} + \mathbf{A}'(\mathbf{X}'(\mathbf{I}-\mathbf{S})\mathbf{X} - n\mathbf{\Lambda} + k\mathbf{I}_p)^{-1}\mathbf{A} \qquad (۱۸)$$

را نتیجه می‌دهد، که $\mathbf{A} = l_i e' - \left(l_i'\hat{\beta}_k\right)\mathbf{X}'(\mathbf{I}-\mathbf{S})$ است. برای ماتریس‌های تأثیر جزئی β و g به ترتیب رابطه‌های:

$$\ddot{\mathbf{F}}_{x,i}(\beta) = \ddot{\mathbf{F}}_{x,i} - (l_i'\hat{\beta}_k)^2\mathbf{S}$$
$$\ddot{\mathbf{F}}_{x,i}(g) = \ddot{\mathbf{F}}_{x,i} - \left(\left(l_i'\hat{\beta}_k\right)\mathbf{X} - el_i'\right)(\mathbf{X}'\mathbf{X})^{-1}\left(\left(l_i'\hat{\beta}_k\right)\mathbf{X} - el_i'\right)'$$

حاصل می‌شود. توجه شود که رابطه (۱۸) به دو بخش تقسیم‌شده است. اولین بخش رابطه (۱۷)، $(l_i'\hat{\beta}_k)^2\mathbf{S}$، بخشی از برآورد نا پارامتری است. باقیمانده فرمول (۱۸) همان شکلی را داراست که ماتریس تأثیر برای مدل‌های خطی دارد. برای تفسیر و بحث مفصل راجع این بخش به کوک [۲۸] بخش ۵ رجوع کنید.

۵-۱- شبیه‌سازی

در این بخش برای ارزیابی تئوری‌های بیان‌شده در بخش ۴ یک مجموعه داده‌های شبیه‌سازی شده از مدل (۱) در نظر گرفته‌شده است. در واقع مدل خطی نیمه پارامتری با خطا در اندازه‌گیری موردنظر برابر

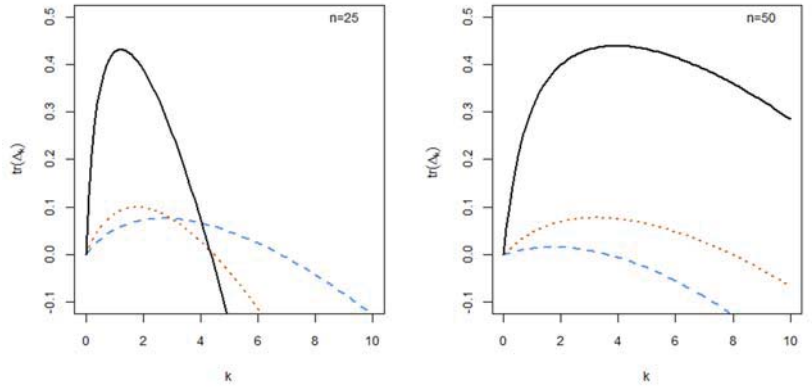
$$y = \mathbf{Z}\beta + \sin 2\pi t + \varepsilon$$
$$\mathbf{X} = \mathbf{Z} + \boldsymbol{\delta}$$

است که در آن درایه‌های مربوط به متغیرهای $\mathbf{Z}$ برای داشتن همخطی چندگانه در ستون‌ها به شکل زیر تولید می‌شوند:

$$z_{ij} = \left(1-\rho^2\right)^{\frac{1}{2}} u_{ij} + \rho u_{ij} \qquad j = 1,2,3 \qquad i = 1,\ldots,n$$

که در آن u_{ij} از توزیع نرمال استاندارد تولید می‌شوند و ضریب ρ اختصاص داده‌شده میزان همبستگی بین ستون‌های ماتریس $\mathbf{Z}$ را نشان می‌دهد. در هر تکرار بردار β برابر بردار ویژه متناظر به بزرگ‌ترین مقدار ویژه‌ماتریس $\mathbf{Z}'(\mathbf{I}_n - \mathbf{S})\mathbf{Z}$ انتخاب‌شده است و بردار t از توزیع یکنواخت صفر و یک تصادفی به دست می‌آید. عناصر ε از هم مستقل هستند و به‌طور تصادفی از توزیع نرمال استاندارد تولید می‌شوند. کوواریانس ماتریس $\mathbf{\Lambda}$ به شکل $diag\,\mathbf{\Lambda} = (0/5, 0/75, 1)$ انتخاب‌شده است. برای تولید داده‌هایی با میزان همبستگی متفاوت، مقادیر ρ برابر ۰/۷۵، ۰/۸۵ و ۰/۹۵ در نمونه‌هایی با اندازه‌های $n = 25$ و $n = 50$ در نظر گرفته‌شده است. داده‌ها در نرم‌افزار

R و با ۱۰۰۰ تکرار شبیه‌سازی شده‌اند. نتیجه حاصل از یک نمونه تصادفی از این ۱۰۰۰ تکرار برای نمونه‌های با اندازه متفاوت در شکل **۱** و جدول **۱** نشان داده‌شده است.



شکل (۱): رسم اثر Δ_k در مقابل k به ازای ρ های متفاوت (خط ممتد $\rho = ۰/۹۵$، خط نقطه‌ای $\rho = ۰/۸۵$ و خط بریده $\rho = ۰/۷۵$)

شکل **۱** رسم مقادیر Δ_k را به ازای k متفاوت نشان می‌دهد از هرکدام از دو تصویر شکل **۱** مشخص است که به ازای هر ρ متفاوت مقادیری از k وجود دارند که به ازای آن‌ها مقدار اثر $amse\left(\hat{\beta}_k\right)$ از اثر $amse\left(\hat{\beta}\right)$ کوچک‌تر است به عبارتی Δ_k مثبت است. جدول **۱** مقادیر برآورد k، λ اثر $amse\left(\hat{\beta}_k\right)$، $amse\left(\hat{\beta}\right)$ و d را نشان می‌دهد. لازم به ذکر برآورد λ در دو مرحله به‌دست‌آمده است. ابتدا با محاسبه Δ_k به ازای مقادیر مختلف k، kی متناظر به ماکزیمم Δ_k مقدار kمناسب را می‌دهد. سپس با قرار دادن k مناسب در $GCV\left(k,\lambda\right)$ این تابع نسبت به λ مینیمم می‌شود و λ مشخص می‌شود. با توجه به جدول مقادیر منفی d رابطه (۱۰) شرط مثبت بودن Δ_k برای kهای بهینه را تائید می‌کنند. در ادامه برای بررسی تحلیل تشخیصی به روش حذف موردی مدل شبیه‌سازی شده بالا را دوباره با اندازه نمونه‌های متفاوت در نظر می‌گیریم. این بار سه مجموعه داده با اندازه‌های ۲۵، ۵۰ و ۱۰۰ تولید می‌شوند.مقدار ضریب همبستگی ρ و بردار پارامتر β به ترتیب مقدار اختیاری $\rho = ۰/۹۵$ و $\beta' = (۱,-۳,۵)$ در نظر گرفته می‌شود. در هر مجموعه داده برای داشتن مقادیر دورافتاده و مؤثر مشاهدات ۵، ۱۰ و ۲۵ با یک مجموعه داده‌های دورافتاده مؤثر جایگزین می‌شود. به این شکل که $y_{۵}$ و $y_{۲۵}$ را برابر میانگین y بعلاوه سه برابر انحراف معیار y و $sin۲\pi t_{۱۰}$ را برابر ۱ قرار می‌دهیم. جدول

۲ مقادیر آماره‌های کوک را با ۵۰۰۰ تکرار نشان می‌دهد. با توجه به جدول مشخص است که هر سه آماره به‌درستی مشاهدات مؤثر را شناسایی می‌کنند.

جدول (۱): نتایج مربوط به داده‌های شبیه‌سازی

n	ρ	k	λ	$amse\left(\hat{\beta}\right)$	$amse\left(\hat{\beta}_k\right)$	d
	۰/۷۵	۲/۷	۰/۱۹	۰/۳۲۳	۰/۲۴۳	−۰/۶۳
۲۵	۰/۸۵	۱/۸	۰/۱۱	۰/۲۲۱	۰/۱۲۰	−۰/۵۶
	۰/۹۵	۱/۲	۰/۰۸	۰/۶۳۸	۰/۲۰۱	−۰/۸۱
	۰/۷۵	۱/۷	۰/۱۳	۰/۱۰۰	۰/۰۵۱	−۰/۴۰
۵۰	۰/۸۵	۳/۲	۰/۰۹	۰/۱۸۱	۰/۱۰۳	−۰/۲۲
	۰/۹۵	۳/۹	۰/۰۵	۰/۵۶۲	۰/۱۵۲	−۰/۱۳

جدول (۲): مقادیر آماره‌های D_{yi}، D_{gi} و $D_{\beta i}$ برای داده‌های شبیه‌سازی شده. مقادیر داخل پرانتز مقادیر انحراف استاندارد می‌باشند.

اندازه نمونه	مشاهدات	$D_{\beta i}$	D_{gi}	D_{yi}
۲۵	۵	۰/۰۹۱ (۰/۰۰۹۶)	۰/۰۵۵ (۰/۰۰۰۲۳)	۰/۱۰۵ (۰/۰۰۸۲)
	۱۰	۰/۰۷۷ (۰/۰۰۹۹)	۰/۱۵۱ (۰/۰۰۲۶)	۰/۱۴۷ (۰/۰۰۷۷)
	۲۵	۰/۱۰۴ (۰/۰۰۸۴)	۰/۰۸۱ (۰/۰۰۳۵)	۰/۲۱۸ (۰/۰۰۳۹)
	دیگر مشاهدات	≤۰/۰۲۵ (۰/۰۰۹۱)	≤۰/۰۴۳ (۰/۰۰۵۰)	۰/۰۷۸ (۰/۰۰۸۵)
۵۰	۵	۰/۱۱۹ (۰/۰۰۷۱)	۰/۱۰۶ (۰/۰۰۴۱)	۰/۲۹۳ (۰/۰۰۹۲)
	۱۰	۰/۱۰۱ (۰/۰۰۱۹)	۰/۳۱۰ (۰/۰۰۵۷)	۰/۴۰۰ (۰/۰۰۶۱)
	۲۵	۰/۲۱۸ (۰/۰۰۷۸)	۰/۰۹۸ (۰/۰۰۷۱)	۰/۳۴۷ (۰/۰۰۳۱)
	دیگر مشاهدات	≤۰/۰۴۳ (۰/۰۰۶۶)	≤۰/۰۶۵ (۰/۰۰۸۸)	۰/۰۹۰ (۰/۰۰۷۱)
۱۰۰	۵	۰/۴۱۷ (۰/۰۰۲۱)	۰/۰۰۹ (۰/۰۰۱۳)	۰/۳۷۹ (۰/۰۰۱۱)
	۱۰	۰/۲۱۶ (۰/۰۰۵۳)	۰/۶۲۰ (۰/۰۰۱۱)	۰/۶۶۰ (۰/۰۰۲۳)
	۲۵	۰/۵۳۳ (۰/۰۰۱۲)	۰/۱۰۳ (۰/۰۰۳۱)	۰/۷۲۷ (۰/۰۰۱۱)
	دیگر مشاهدات	≤۰/۰۶۴ (۰/۰۰۱۰)	≤۰/۰۴۳ (۰/۰۰۱۷)	۰/۰۷۵ (۰/۰۰۰۱)

از جدول **۲** مشخص است که با افزایش حجم نمونه مشاهدات دورافتاده با توجه به‌اندازه‌ای که می‌گیرند آشکارتر از سایر مشاهدات ظاهر می‌شوند. این در حالی است که در نمونه‌های کوچک‌تر مشاهدات دورافتاده (مشاهده ۵ و ۱۰ در نمونه با اندازه ۲۵) به‌راحتی آشکار نمی‌شوند.

۵-۲- داده‌های سفال مصری

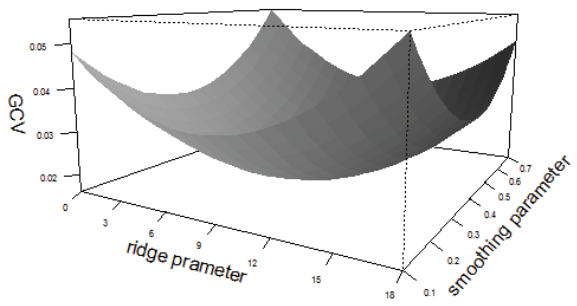
در این بخش برای انجام تحلیل عددی و نموداری مباحث تشخیصی مطرح‌شده در بخش‌های قبل از یک مجموعه داده شناخته‌شده به نام داده‌های سفال مصری استفاده می‌کنیم. به‌طور خلاصه این مجموعه داده مربوط به گزارش حاصل از بررسی‌های باستان‌شناسی گسترده‌ای از تولید و پخش سفال در شهر باستانی الامرنا مصر است. داده‌ها شامل اندازه‌گیری محتویات شیمیایی (عناصر معدنی) موجود در بسیاری از نمونه‌های سفال است که با استفاده از دو روش متفاوت یکی روش تجزیه‌وتحلیل نوترون (NAA) و دیگری روش پلاسمای القایی (ICP) به‌دست‌آمده‌اند. برای توضیحات بیشتر روش اندازه‌گیری اسمیت و همکاران [۳۱] را ببینید. مجموعه سفال‌ها از مکان‌های مختلف اطراف شهر جمع‌آوری‌شده‌اند.در کل دو نوع خاک رس لای و مارن در ساخت سفال‌های مصر باستان استفاده می‌شده است و باستان‌شناسان برای جدا کردن قطعات سفال‌های مشترک (که از کشورهای آفریقایی شمالی وارد می‌شدند) از سفال لای و مارن مصری، آن‌ها را گروه‌بندی کرده‌اند. نوع گروه‌بندی اشیا سفالی با توجه به دو منبع اصلی، یکی کد طرح و دیگری کد مکان سفال است که هرکدام از این دو نوع گروه‌بندی برای باستان‌شناسان حائز اهمیت است. نهایتاً با توجه به این نوع از گروه‌بندی قطعات سفال انتخابی به ۲۸ گروه تقسیم‌بندی شدند که گروه‌های ۶ و ۱۵ و ۱۸ و ۲۳سفال‌های وارداتی بودند. در هر گروه تعداد متفاوتی از سفال‌ها دارای کد طرح و کد مکانی یکسانی وجود دارد که در اصل می‌تواند به‌عنوان مشاهدات تکراری در نظر گرفته شود. از میان تمام عناصر معدنی علاقه‌مند به بررسی رابطه بین عنصر Na به‌عنوان متغیر پاسخ که با روش NAA اندازه‌گیری شده است در برابر شش عنصر Na, Al, K, V, Cr و Mn به‌عنوان متغیرهای توضیحی که با روش NAA اندازه‌گیری شده‌اند هستیم. راسخ [۲۱] این داده‌ها را با برازش یک مدل خطی پارامتری کامل با خطا در اندازه‌گیری موردمطالعه قرار داده است. او نشان داد که یک همخطی با اندازه متوسط بین متغیرهای توضیحی وجود دارد. به‌جای استفاده از انتخاب زیرمجموعه از متغیرها او برآورد ریج پارامترها را با $k = ۰/۱۵$ به دست آورد و نشان داد که این نوع برآوردگر برآورد پارامترها را بهبود می‌بخشد. اخیراً قپانی و همکاران [۲۰] از نگاهی دیگر برآورد پارامترهای ریج وزنی متغیرهای فوق را با الگوی مدل خطی پارامتری محدودشده با اندازه‌گیری در خطا موردبررسی و تحلیل قرار داده‌اند. از طرفی برای بررسی و شناسایی مشاهدات مؤثر راسخ [۲۱] و قپانی و همکاران[۲۲] از رویکرد مباحث تشخیصی به بررسی و مطالعه این داده‌ها پرداخته‌اند .آن‌ها این مجموعه داده را با استفاده

از روش‌های تأثیر موضعی و انتقال میانگین دورافتاده به ترتیب در مدل خطی پارامتری ریج با خطا در اندازه‌گیری و مدل‌های خطی پارامتری با اندازه‌گیری در خطا تحت محدودیت خطی تصادفی مورد تحلیل و بررسی قراردادند و گروه‌های ۱۸، ۶ و ۱۲ (به ترتیب با بیشترین تأثیر) در مدل خطی پارامتری ریج با خطا در اندازه‌گیری و گروه ۱۳ در برازش مدل خطی پارامتری با اندازه‌گیری در خطا تحت محدودیت خطی تصادفی به‌عنوان گروه‌های مؤثر شناسایی کردند.ازآنجایی‌که رابطه بین متغیر (پاسخ) Na و متغیر Mn را می‌توان به شکل غیرخطی فرض کرد (شکل۳) و با توجه به وجود همخطی در ستون‌های ماتریس متغیرهای توضیحی با خطا در اندازه‌گیری (طبق پژوهش‌های پیشین) می‌توان مدل خطی نیمه‌پارامتری (۱) را با درنظر گرفتن $\mathbf{X}=(Al,K,C,V,Cr)$ به‌عنوان متغیرهای مربوط به مؤلفه‌های پارامتری و $g(t)=Mn$ به‌عنوان متغیر مربوط به مؤلفه نا پارامتری مدل برای برازش به این مجموعه از داده‌ها در نظر گرفت. ابتدا از محاسبات مشخص می‌شود نسبت بزرگ‌ترین مقدار ویژه به کوچک‌ترین مقدار ویژه‌ماتریس $\mathbf{X}'(\mathbf{I}-\mathbf{S})\mathbf{X}$ برابر $\frac{\lambda_{max}}{\lambda_{min}}=۹۸۷۱/۵$ می‌باشد که نشان از یک همخطی شدید بین متغیرهای هموار شده است. لذا در اینجا برای برآورد پارامترها از روش رگرسیون ریج با یک k مناسب می‌توان استفاده نمود. با توجه به اینکه در داده‌ها تکرار وجود دارد یک برآورد نااریب برای $\mathbf{\Lambda}$ را می‌توان از روش لیانگ و همکاران [۳۲] به دست آورد. با استفاده از این روش برآورد ماتریس $\mathbf{\Lambda}$ در مدل ما برابر است با:

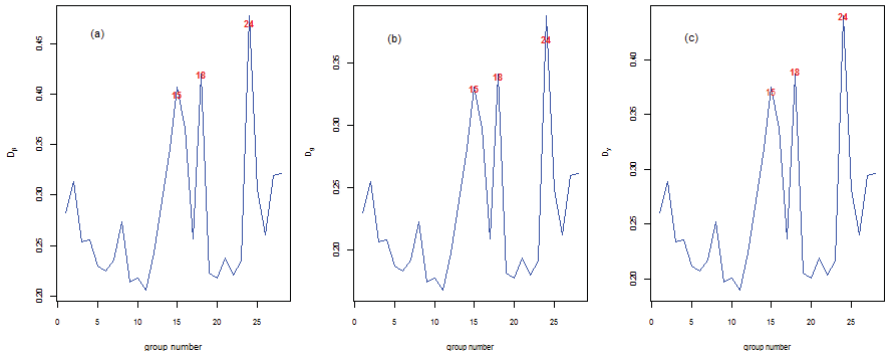
$$\hat{\mathbf{\Lambda}}=\begin{bmatrix} ۲/۳۵۶ & ۱/۳۴۴ & ۰/۰۱۹۰ & ۰/۰۹۴۰ & ۰/۲۲۹۱ \\ ۱۷/۷۲۴ & ۱۸/۸۸۰ & ۰/۱۳۸ & ۱/۳۵۷ & ۰۹۴۰/۰ \\ ۱/۰۹۳ & -۰/۰۹۹ & ۰/۱۵۰ & ۰/۱۳۸۰ & ۰/۰۱۹۰ \\ ۲۹۱/۱۵۹ & ۷۴۷/۲۷۲ & -۰/۰۹۹ & ۱۸/۱۸۰ & ۱/۳۴۴ \\ ۴۴۴/۳۶۸ & ۲۹۱/۱۵۹ & ۱/۰۹۳ & ۱۷/۷۲۴ & ۱/۳۵۶ \end{bmatrix}$$

با توجه به شکل ۲ با مینیمم کردن معیار اعتبارسنجی متقابل هم‌زمان نسبت به k و λ برآورد هم‌زمان پارامتر ریج و پارامتر هموارساز برابر۱۰/۲۳ و ۰/۳۵ به دست می‌آید. نخست با استفاده از رویکرد حذفی به ارزیابی مشاهدات می‌پردازیم. شکل **۳** مقادیر آماره‌های کوک به ترتیب از چپ به راست D_β، D_g و D_y را در برابر شماره گروه‌های سفال نمایش می‌دهد. با توجه به شکل (**۳.a**) مشاهده ۲۴، ۱۸ و ۱۵ به ترتیب به‌عنوان مشاهدات مؤثر با بیشترین تأثیر روی بردار $\hat{\beta}_k$ شناسایی می‌شوند. یادآور می‌شویم چنانکه در ابتدا بیان شد گروه‌های۱۵و ۱۸ جزو گروه‌های وارداتی می‌باشند. از شکل (**۳.b**) و (**۳.c**) روشن است که همین مشاهدات می‌توانند روی مقادیر برازشی مؤلفه نا پارامتری g و مقادیر برازشی کل $\hat{y}_k$ مؤثر واقع شوند. این نتیجه زیاد دور از انتظار نیست زیرا که بخش اصلی آماره‌های D_β، D_g و D_y به باقیمانده‌های کل e_i و مقادیر

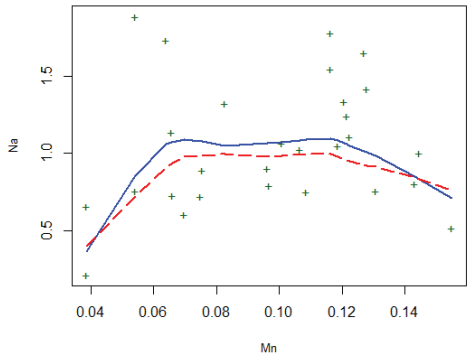
نافذ کل مدل یعنی h_{ii} بستگی دارند. یادآوری می‌شود که همیشه مشاهدات یکسانی روی تمامی اجزای مدل اثر ندارند و ممکن است مشاهدات با توجه به نوع داده متفاوت باشند. تأثیر این گروه‌ها در برازش مؤلفه نا پارامتری g در شکل **۴** نشان داده‌شده است در واقع این شکل مقادیر برازشی $\hat{g}$ قبل از حذف سه مشاهده مؤثر ۱۵، ۱۸ (گروه‌های وارداتی) و ۲۴ را با خط ممتد و بعد از حذف مشاهدات مؤثر را با خط بریده به نمایش داده است. کاملاً مشهود است که وجود این مشاهدات تأثیر بسزایی در برآورد $\hat{g}$ دارد.



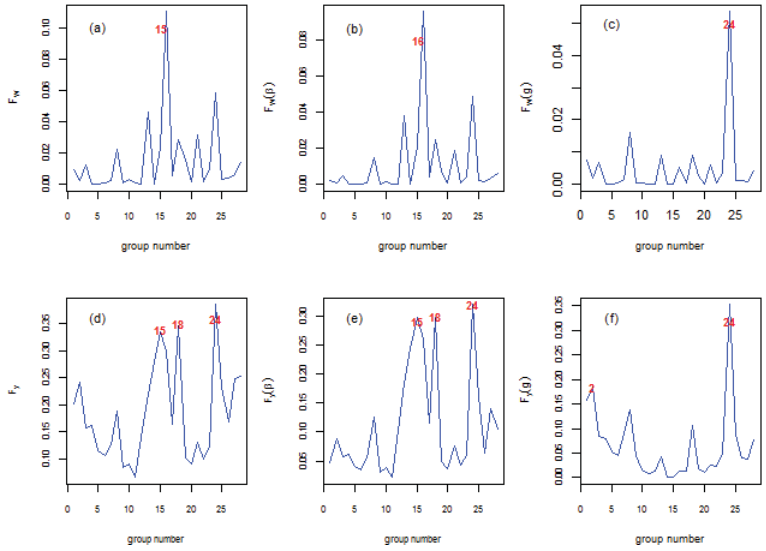
شکل (۲): نمودار GCV در مقابل پارامتر ریج k و پارامتر هموارساز λ برای داده‌های سفال مصری



شکل (۳): مقادیر (a)- آماره D_β (b)- آماره D_g (c)- آماره D_y در مقابل شماره گروه‌ها



شکل (۴): برازش مؤلفه نا پارامتری: خط بریده با حذف سه گروه مؤثر ۱۵،۱۸ و ۲۴ – خط ممتد بدون گروه‌های مؤثر



شکل (۵): تحلیل تأثیر مکانی داده‌های سفال مصری:عناصر روی قطر اصلی ماتریس‌های (a) $\ddot{\mathbf{F}}_w(y)$، (b) $\ddot{\mathbf{F}}_w(\beta)$، (c) $\ddot{\mathbf{F}}_w(g)$، (d) $\ddot{\mathbf{F}}(y)$، (e) $\ddot{\mathbf{F}}(\beta)$ و (f) $\ddot{\mathbf{F}}(g)$ نمایانگر سهم هریک از گروه‌ها در ماتریس‌های تأثیر می‌باشند.

۶- نتیجه‌گیری

شناسایی مشاهدات مؤثر در مدل‌های خطی نیمه پارامتری در حضور همخطی چندگانه و وجود خطا در اندازه‌گیری از دو رویکرد اصلی روش حذف موردی و تأثیر موضعی موردمطالعه قرار گرفت.با توجه به روش حذف موردی اندازه‌های تأثیر توسط مقادیر نافذ و باقیمانده‌ها با استفاده از فرمول‌های به‌دست‌آمده در قضیه ۲ بخش ۴-۱ محاسبه می‌شوند. معتقدیم که فرمول‌های حذف موردی که در این مقاله به آن‌ها رسیدیم به‌عنوان بخشی از هر تحلیل داده‌ی مهم می‌توانند کمک‌کننده باشند. از طرفی هدف از شاخص‌های موضعی شناسایی و درک بهتر مشاهداتی است که ممکن است در تحلیل مرتبط با یک مدل خاص اهمیت داشته باشد. مقاله حاضر پژوهش کوک[۲۸] را به‌منظور سنجش تأثیر موضعی تحت انحراف وزن‌های نمونه، متغیر پاسخ و یا متغیرهای کمکی برای برآوردگرهای درستنمایی تصحیح‌شده تاوانیده در مدل خطی نیمه پارامتری، بسط می‌دهد. بر این باوریم که شاخص‌های تأثیر موضعی که در اینجا به آن می‌رسیم می‌تواند به‌عنوان بخشی از هر تحلیل داده جدی مفید باشد. مخصوصاً که در برگیری یک تابع نا پارامتری می‌تواند تأثیر مهمی روی تحلیل تأثیر برای برآورد ضریب رگرسیون خطی داشته باشد. اندازه‌های تأثیر ارائه‌شده در این مقاله نقشی را در فهم اینکه چگونه هر مؤلفه برای برآورد دیگری کمک می‌کند، ایفا می‌کند. به ویژه‌ماتریس‌های نافذ شرطی و حاشیه‌ای برای دو مؤلفه می‌توانند به‌عنوان ماتریس‌های تأثیر جزئی تحت انحراف متغیر پاسخ در نظر گرفته شوند. آنالیز تأثیر تحت انحراف متغیرهای کمکی نقطه‌نظر متناوبی را برای مدل‌های خطی خطا در اندازه‌گیری ارائه می‌کند. انحراف وزنی مشاهدات نیز نظریه حذف موردی را تعمیم می‌دهد طوری که یک تقریب خوب به شاخص‌های حذف موردی بدون اجبار برای دوباره برآورد کردن پارامترهای هر حذف را فراهم می‌کند. ماتریس تأثیر متناظر می‌تواند در شناسایی مشاهدات مؤثر کمک‌کننده باشد.با توجه به اینکه اندازه تأثیر یک مشاهده تنها با افزایش اندازه نمونه گرایش به تقلیل دارد، شاخص-های تأثیر احتمالاً در نمونه کوچک یا نسبتاً کم محبوب‌ترین می‌باشند.

ضمیمه:

برای رابطه (۸) و (۹) با مشتق‌گیری از رابطه (۷) نسبت به β و g به‌سادگی خواهیم داشت:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} - n\boldsymbol{\Lambda} + k\mathbf{I}_p & \mathbf{X}'\mathbf{N} \\ \mathbf{N}'\mathbf{X} & \mathbf{N}'\mathbf{N} + \lambda\mathbf{M} \end{pmatrix} \begin{pmatrix} \beta \\ g \end{pmatrix} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{N}' \end{pmatrix} y \qquad (۱)$$

از معادله دوم ماتریس داریم:

$$\mathbf{N}'\mathbf{X}\beta + (\mathbf{N}'\mathbf{N} + \lambda\mathbf{M})g = \mathbf{N}'y \rightarrow g = (\mathbf{N}'\mathbf{N} + \lambda\mathbf{M})^{-1}\mathbf{N}'(y - \mathbf{X}\beta)$$

با قرار دادن g در معادله اول ماتریس و تعریف $\mathbf{S} = \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda\mathbf{M})^{-1}\mathbf{N}'$ روابط

$$(\mathbf{X}'\mathbf{X}-n\Lambda+k\mathbf{I}_p)\beta+\mathbf{X}'\mathbf{N}\mathbf{g}=(\mathbf{X}'\mathbf{X}-n\Lambda+k\mathbf{I}_p)\beta+\mathbf{X}'\,\mathbf{S}(y-\mathbf{X}\beta)=\mathbf{X}'y$$

$$\rightarrow(\mathbf{X}'\mathbf{X}-\mathbf{X}'\mathbf{S}\mathbf{X}-n\Lambda+k\mathbf{I}_p)\beta=\mathbf{X}'y-\mathbf{X}'\mathbf{S}y$$

$$\rightarrow(\mathbf{X}'(\mathbf{I}_n-\mathbf{S})\mathbf{X}-n\Lambda+k\mathbf{I}_p)\beta=\mathbf{X}'(\mathbf{I}-\mathbf{S})y$$

به دست می‌آیند. از رابطه اخیر خیلی راحت $\hat{\beta}_k$ حاصل می‌شود که با قرار دادن آن در g مقدار $\hat{g}_k$ نیز نتیجه می‌شود. از حل دو معادله فوق رابطه‌های (۸) و (۹) به دست می‌آیند.

تعریف ۱ : برای دو ماتریس $\mathbf{A}$ و $\mathbf{B}$ گوییم $\mathbf{A}\succ\mathbf{B}$ یا $\mathbf{A}-\mathbf{B}\succ 0$ اگر $\mathbf{A}-\mathbf{B}$ معین مثبت باشد.

اثبات قضیه۱: برای اثبات ابتدا لم زیر تعریف می‌شود:

لم ۱: از فونگ و همکاران [۶] داریم:

$$\mathbf{X}'(\mathbf{I_n}-\mathbf{S})\mathbf{X}=\mathbf{Z}'(\mathbf{I}_n-\mathbf{S})\mathbf{Z}+n\Lambda+O_p\left(n^{\frac{1}{2}}\right)$$

با توجه به لم ۱ رابطه‌های زیر به دست می‌آیند:

$$n^{-1}\left(\mathbf{X}'(\mathbf{I_n}-\mathbf{S})\mathbf{X}-n\Lambda+k\mathbf{I}_p\right)=n^{-1}\left(\mathbf{Z}'(\mathbf{I}_n-\mathbf{S})\mathbf{Z}+\mathbf{k}\mathbf{I}_p\right)+O_p\left(n^{\frac{-1}{2}}\right)$$

$$\sqrt{n}\,\hat{\beta}_k=\left[n^{-1}\left(\mathbf{Z}'(\mathbf{I_n}-\mathbf{S})\mathbf{Z}+k\mathbf{I}_p\right)+O_p\left(n^{\frac{-1}{2}}\right)\right]^{-1}n^{\frac{-1}{2}}\mathbf{X}'(\mathbf{I_n}-\mathbf{S})y$$

$$=\left[\mathbf{I}_p+O_p\left(n^{\frac{-1}{2}}\right)\right]^{-1}\left[n^{-1}\left(\mathbf{Z}'(\mathbf{I}_n-\mathbf{S})\mathbf{Z}+k\mathbf{I}_p\right)\right]^{-1}n^{\frac{-1}{2}}\mathbf{X}'(\mathbf{I}_n-\mathbf{S})y$$

$$=\left[\mathbf{I}_p+O_p\left(n^{\frac{-1}{2}}\right)\right]\left[n^{-1}\left(\mathbf{Z}'(\mathbf{I}_n-\mathbf{S})\mathbf{Z}+k\mathbf{I}_p\right)\right]^{-1}n^{\frac{-1}{2}}\mathbf{X}'(\mathbf{I}_n-\mathbf{S})y$$

با توجه رابطه آخر $\sqrt{n}\hat{\beta}_k$ به شکل زیر خلاصه می‌شود:

$$\sqrt{n}\,\hat{\beta}_k=\mathbf{C}^{-1}\xi+O_p\left(n^{\frac{-1}{2}}\right)$$

می‌توان نشان داد که در آن $\xi = n^{-\frac{1}{2}}\mathbf{X}'(\mathbf{I_n}-\mathbf{S})y$ دارای توزیع مجانبی نرمال هست (به فونگ و همکاران [۶] مراجعه شود).ازآنجایی‌که $E(\xi) = n^{-\frac{1}{2}}\mathbf{C}_{\circ}\beta + O_p\left(n^{-\frac{1}{2}}\right)$ درنتیجه خواهیم داشت:

$$\sqrt{n}\left(\hat{\beta}_k - \mathbf{C}_{\circ}\mathbf{C}^{-1}\beta\right) = \mathbf{C}_k^{-1}\left[\xi - E(\xi)\right] + O_p\left(n^{-\frac{1}{2}}\right)$$

که نشان می‌دهد $\sqrt{n}\left(\hat{\beta}_k - \mathbf{C}_{\circ}\mathbf{C}_k^{-1}\beta\right)$ دارای توزیع مجانبی نرمال با میانگین صفر می‌باشد. از طرفی با استفاده از واریانس شرطی برای ξ داریم:

$$\begin{aligned}
var(\xi) &= E\left(var(\xi|y)\right) + var(E(\xi \mid y)) \\
&= n^{-1}E\left(y'(\mathbf{I}_n - \mathbf{S})^2 y\mathbf{\Lambda}\right) + n^{-1}var\left(\mathbf{X}'(\mathbf{I}_n - \mathbf{S})y\right) \\
&= n^{-1}\left[\left(\mathbf{Z}\beta + \mathbf{N}g\right)'(\mathbf{I}_n - \mathbf{S})^2\left(\mathbf{Z}\beta + \mathbf{N}g\right) + \sigma^2 tr\left(\mathbf{I}_n - \mathbf{S}\right)\right] + n^{-1}\mathbf{Z}'(\mathbf{I}_n - \mathbf{S})^2\mathbf{Z}\sigma^2 \\
&= n^{-1}\left[\mathbf{D} + \mathbf{Z}'(\mathbf{I_n} - \mathbf{S})^2\mathbf{Z}\sigma^2\right] = n^{-1}\mathbf{\Phi}
\end{aligned}$$

ازآنجایی‌که $avar\left(\sqrt{n}\hat{\beta}_k\right) = \mathbf{C}_k^{-1}\,\mathrm{var}(\xi)\mathbf{C}_k^{-1}$ بنابراین داریم:

$$avar\left(\hat{\beta}_k\right) = \mathbf{C}_k^{-1}\left[\mathbf{D} + \mathbf{Z}'(\mathbf{I}_n - \mathbf{S})^2\mathbf{Z}\sigma^2\right]\mathbf{C}_k^{-1}$$

و اثبات قضیه کامل است.

لم ۲: فرض کنید $\mathbf{A}$ یک ماتریس معین نامنفی $n\times n$، a یک بردار $n\times 1$ باشد آنگاه $\mathbf{A} - aa'$ معین نامنفی است اگر و تنها اگر $a'\mathbf{A}^{-}a \le 1$ باشد که در آن $\mathbf{A}^{-}$ وارون تعمیم‌یافته ماتریس $\mathbf{A}$ است. (رائو ۲۰۰۸).

از رابطه اول قضیه ۱ برای $D_{\beta i}$ می‌توان نوشت:

$$D_{\beta i}=\frac{(\hat{\beta}_k-\hat{\beta}_{k(i)})'\mathbf{X}'\mathbf{X}\left(\hat{\beta}_k-\hat{\beta}_{k(i)}\right)}{\sigma^2 tr\left(\tilde{\mathbf{H}}_k\right)}$$

$$=\frac{\xi_i e_i\left(\mathbf{I}_n-\mathbf{S}\right)\mathbf{X}\{\mathbf{X}'\left(\mathbf{I}_n-\mathbf{S}\right)\mathbf{X}+k\mathbf{I}_p-n\mathbf{\Lambda}\}^{-1}\mathbf{X}'\mathbf{X}\{\mathbf{X}'\left(\mathbf{I}_n-\mathbf{S}\right)\mathbf{X}+k\mathbf{I}_p-n\mathbf{\Lambda}\}^{-1}\mathbf{X}'\left(\mathbf{I}_n-\mathbf{S}\right)e_i\xi_i}{\sigma^2 tr\left(\tilde{\mathbf{H}}_k\right)(1-h_{ii})^2}$$

$$=\frac{\xi_i e_i\tilde{\mathbf{H}}_k{}'\tilde{\mathbf{H}}_k\xi_i e_i}{\sigma^2 tr\left(\tilde{\mathbf{H}}_k\right)(1-h_{ii})^2}=\frac{\left[\sum_{j=1}^{n}\tilde{h}_{ij}^2\right]e_i{}^2}{\sigma^2 tr\left(\tilde{\mathbf{H}}_k\right)(1-h_{ii})^2}$$

محاسبات ماتریس تأثیر رابطه (۱۱) را برای همه ماتریس‌های تأثیر استفاده می‌کنیم. ماتریس $\ddot{\mathbf{L}}$ برای همه موارد یکسان است که معکوس آن برابر است با

$$\ddot{\mathbf{L}}^{-1}=\begin{pmatrix}\mathbf{L}^{11} & \mathbf{L}^{12}\\ \mathbf{L}^{21} & \mathbf{L}^{22}\end{pmatrix}$$

$$=\begin{pmatrix}(\mathbf{X}'\left(\mathbf{I}-\mathbf{S}\right)\mathbf{X}-n\mathbf{\Lambda}+k\mathbf{I}_p)^{-1} & -(\mathbf{X}'\left(\mathbf{I}-\mathbf{S}\right)\mathbf{X}-n\mathbf{\Lambda}+k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{N}\mathbf{T}^{-1}\\ -\mathbf{T}^{-1}\mathbf{N}'\mathbf{X}(\mathbf{X}'\left(\mathbf{I}-\mathbf{S}\right)\mathbf{X}-n\mathbf{\Lambda}+k\mathbf{I}_p)^{-1} & \mathbf{T}^{-1}+\mathbf{T}^{-1}\mathbf{N}^{\mathrm{T}}\mathbf{X}(\mathbf{X}'\left(\mathbf{I}-\mathbf{S}\right)\mathbf{X}-n\mathbf{\Lambda}+k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{N}\mathbf{T}^{-1}\end{pmatrix},$$

که $\mathbf{T}=\left(\mathbf{N}'\mathbf{N}+\lambda\mathbf{M}\right)$ است.

برای محاسبه Δ متناظر به انحراف وزن نمونه در بخش ۴-۲-۱ داریم:

$$\frac{\partial l^*(\theta,\omega)}{\partial\omega_i}=\frac{-1}{2\sigma^2}\left[\| y-\mathbf{X}\beta-\mathbf{N}g\|^2\right]_i\rightarrow\frac{\partial l^*(\theta,\omega)}{\partial\omega_i\partial\beta}=\sigma^{-2}\left[\mathbf{X}'\left(y-\mathbf{X}\beta-\mathbf{N}g\right)\right]_i$$

$$\frac{\partial l^*(\theta,\omega)}{\partial\omega_i\partial g}=\sigma^{-2}\left[\mathbf{N}'\left(y-\mathbf{X}\beta-\mathbf{N}g\right)\right]_i$$

که $\left[\mathbf{A}\right]_i$ نشان‌دهنده iامین عضو بردار $\mathbf{A}$ است. با تعریف $\beta=\hat{\beta}$ و $g=\hat{g}$ در مشتقات بالا و نمایش ماتریسی خواهیم داشت:

$$\Delta=\hat{\sigma}^{-2}(\mathbf{X},\mathbf{N})'\,\mathrm{diag}\left(y-\mathbf{X}\hat{\beta}-\mathbf{N}\hat{g}\right)$$

برای Δ تحت انحراف i-اُمین متغیر کمکی، در بخش ۴-۲-۳ داریم:

$$\Delta=-\left(l_i'\hat{\beta}_{ki}\right)\begin{pmatrix}\mathbf{X}'\\ \mathbf{N}'\end{pmatrix}+\begin{pmatrix}l_i e'\\ \circ\end{pmatrix}.$$

فرض کنید $\hat{\beta}_{ki}=\left(l_i'\hat{\beta}_{ki}\right)$.بنابراین از رابطه (۱۱) می‌توان نوشت:

$$\ddot{\mathbf{F}}_{x,i} = \boldsymbol{\Delta}'\ddot{\mathbf{L}}^{-1}\boldsymbol{\Delta} = \hat{\beta}_{ki}^{2}\,(\mathbf{X} \quad \mathbf{N})\ddot{\mathbf{L}}^{-1}(\mathbf{X} \quad \mathbf{N})' - \hat{\beta}_{ki}\,(\mathbf{X} \quad \mathbf{N})\ddot{\mathbf{L}}^{-1}(el_i' \quad \circ)'$$
$$-\hat{\beta}_{ki}\,(el_i' \quad \circ)\ddot{\mathbf{L}}^{-1}(\mathbf{X} \quad \mathbf{N})' + (el_i' \quad \circ)\ddot{\mathbf{L}}^{-1}(el_i' \quad \circ).$$

از تعریف $\mathbf{H}$ و $\ddot{\mathbf{L}}^{-1}$ می‌توان محاسبات زیر را نتیجه گرفت:

$$\ddot{\mathbf{F}}_{x,i} = \hat{\beta}_{ki}^{2}\,\mathbf{H} - \hat{\beta}_{ki}\,(\mathbf{X}\mathbf{L}^{1} + \mathbf{N}\mathbf{L}^{2})l_i e' - \hat{\beta}_{ki}\,el_i'(\mathbf{L}^{1}\mathbf{X}' + \mathbf{L}^{2}\mathbf{N}') + +el_i'\mathbf{L}^{1}l_i e'$$
$$= \hat{\beta}_{ki}^{2}\,(\mathbf{S} + (\mathbf{I}-\mathbf{S})\mathbf{X}(\mathbf{X}'(\mathbf{I}-\mathbf{S})\mathbf{X} - n\boldsymbol{\Lambda} + k\mathbf{I}_{\mathbf{p}})^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S})$$
$$-el_i'(\mathbf{X}'(\mathbf{I}-\mathbf{S})\mathbf{X} - n\boldsymbol{\Lambda} + k\mathbf{I}_p)^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S})\hat{\beta}_{ki}$$
$$-\hat{\beta}_{ki}\,(\mathbf{I}-\mathbf{S})\mathbf{X}(\mathbf{X}'(\mathbf{I}-\mathbf{S})\mathbf{X} - n\boldsymbol{\Lambda} + k\mathbf{I}_p)^{-1}l_i e' + el_i'(\mathbf{X}'(\mathbf{I}-\mathbf{S})\mathbf{X} - n\boldsymbol{\Lambda} + k\mathbf{I}_p)^{-1}l_i e'$$
$$= \hat{\beta}_{ki}^{2}\,\mathbf{S} + \mathbf{A}'(\mathbf{X}'(\mathbf{I}-\mathbf{S})\mathbf{X} - n\boldsymbol{\Lambda} + k\mathbf{I}_p)^{-1}\mathbf{A},$$

که در آن $\mathbf{A} = l_i e' - \hat{\beta}_{ki}\,\mathbf{X}'(\mathbf{I}-\mathbf{S})$. محاسبه ماتریس‌های تأثیر دیگر به همین صورت به‌دست‌آمده است که به خاطر شباهت حذف‌شده است.

تشکر و قدردانی

نویسندگان مقاله از داوران محترم و سردبیر محترم به خاطر نکته‌ها و رهنمودهای ارزشمندشان که موجب ارتقای کیفیت مقاله ازلحاظ نگارشی و علمی شدند، بسیار سپاس گذار هستند.

منابع

[1] Cook, R. D. (1977). Detection of Influential Observations in Linear Regression. *Technometrics*, 19, **15**-18.

[2] Cook, R. D. and Wesiberg, S. (1982). *Residuals and Influence in Regression.* Chapman & Hall, New York

[3] Belsley, D. A., Kuh. E. and Welsch. R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Soureces of Collinearity.* John Wiley & Sons, New York.

[4] Kim, C. (1996). Cook's Distance in Spline Smoothing, Statist *Proba Lett.*, **31**, 139-144.

[5] Kim, C., Park, B. U. and Kim, W. (2002). Influence Diagnostics in Semiparametric Regression Models, *Statist. Probab. Lett,* **60**, 49-58.

[6] Fung, W. K., Zhu, Z. Y., Wei, B. C. and He. X. (2002). Influence Diagnostics and Outlier Tests for Semiparametric Mixed Models. *J. R. Stat. Soc.* Ser. B, **47**, 332-341.

[7] Belsley, D. A. (1991). *Collinearity Diagnostics: Collinearity and Weak Data in Regression.* John Wiley & Sons, New York.

[8] Hoerl, A. E. and Kennard R.W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics* **12**: 69–82

[9] Duran, E. A. and Akdeniz, F. (2012). Efficiency of the Modified Jackknifed Liu-type Estimator. *J. Stat Paper.* **53**(2): 265-280.

[10] Roozbeh, M., and Arashi, M. (2013). Feasible Ridge Estimator in Partially Linear Models, *Multivariate Analysis*, **116**, 35-44.

[11] Roozbeh, M. (2015). Shrinkage Ridge Estimator in Semiparametric Regression Models. *J. Multivariate Anal.* **136**, 56-74.

[12] Emami, H. (2015). Influence Diagnostics in Ridge Semiparametric Regression Models. *J. Stat Prob Lett*, **105**, 106-115.

[13] Emami, H. (2016). Local Influence for Liu Estimators in Semiparametric Linear Models. *Stat Pap*, Doi:10.1007/s00362-016-0775-6.

[14] Fuller, W. A. (1987). Measurements Error Models. John Wiley & Sons, New York.

[15] Hanfelt J. and Liang, K. Y. (1997). Approximate Likelihood for Generalized Linear Errors in Variables Models, *J. Roy. Statist. Soc.* Ser. B, **59**, 627-637.

[16] Nakamura, T. (1990). Corrected score Functions for Error in Variables Models: Methodologe and Application to Generalized Linear Models. *Biometrika*, **77**, 127-137.

[17] Stefanski, L. A. and Carrol, R. J. (1987). Conditional Scores and Optimal Scores for Generalized Linear Measurement reeor Models. *Biometrika*, **74**, 703-716.

[18] Rasekh, A. R. (2001). Ridge Estimation in Functional Measurement Error Models. *Inst. Stat. Univ. Paris*, **45**, 47-59.

[19] Saleh A.K. and Md. Shalabh. (2014). A ridge regression estimation approach to the measurement error model. *J. Mul. Anal*. **123**: 68-84.

[20] Ghapani, F. Rasekh, A. R. and Babadi, B. (2016). The Weighted Ridge Estimator in Stochastic Restricted Linear Measurment Error Models. *J. Stat Paper.* Doi:10.1007/s00362-016-0786-3

[21] Rasekh, A. R. (2006). Local Influence in Measurement Error Models with Ridge Estimate. *Computational Statistic and Data Analysis*, **50**, 2822-2834.

[22] Ghapani, F., Rasekh, A. R. and Babadi, B. (2015). Mean Shift and Influence Diagnostics in linear Mixed Measurement Error Models. *J. Am Math Mange Sci,* **23**, 37-59

[23] Hoerl, A. E. and Kennard, R. W. (1976). Ridge Regression: Iterative Estimation of the Biasing Parameter. *Communications in Statistics- theory and Methods*, **5**, 77-88.

[24] Duran, E. A., Hardel, W. K. and Osipenko, M. (2012). Difference Based Ridge and Liu Type Estimators in Semiparametric Regression Models. *J. Multivariate Anal*, **105**, 164-175

[25] Green, P.J., and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*, London: Chapman and Hall.

[26] Roozbeh, M. (2017). Optimal QR-based Estimation in Partially Linear Regression Models WithCorrelated errors using GCV criterion. *Computational Statistics and Data Analysis*. DOI:http://dx.doi.org/10.1016/j.csda.2017.08.002

[27] Arashi, M. and Roozbeh M. (2015).Some Improved Estimation Strategies in High-Dimensional Semiparametric Regression Models with Application to Riboflavin Production Data. *J. Stat. Paper.* DOI 10.1007/s00362-016-0843-y.

[28] Cook, R. D.(1986). Assissement of Local Influence (with discussion). *J. Roy Statist Soc* Ser. B **48**, 133-169.

[29] Thomas, W. and Cook, R. D. (1989). Assessing Influence On Regression Coefficients in Generalized Linear Models. *Biometrika* **76**, 741–749.

[30] Kwan, C. W. and Fung, W. K. (1998). Assessing local influence for specific restricted likelihood: application to factor analysis. *Psychometrika* **63**, 35–46.

[31] Smith, D.M., Hart, F.A., Symond, R.D. and Walsh, J.N. (1988). Analysis of Roman Pottery From Colchester by Inductively Coupled Plasma Spectrometry. In: Slater, E.A., Tate, J.O. (Eds.), Science and Archaeology Glasgow 1987, vol. 196(I). B.A.R., Oxford, pp. 41–55.

[32] Liang, H. Hardle, W. and Carroll, R.J. (1999). Estimation in a semiparametric partially linear errors-in-variables model, *Ann. Statist*. **27**: 1519–1535.

Influence Diagnostics in Ridge Semiparametric Linear Measurement Error Models

Hadi Emami

Department of Statistics, University of Zanjan, Zanjan, Iran

Abstract

In this paper influence diagnostics in the semi parametric linear measurement error is studied when the problem of multicollinearity is exist. First to combat multicollinearity, some ridge estimators are proposed based the penalized corrected likelihood approach, then using case deletion method the influence measures are defined to detect influential observations. In continue it is shown these measures are function of residuals and leverages. Furtherer more local influence approach is developed based the penalized likelihood function for assessing the influence of points. A simulation study is performed to illustrate the efficiency of proposed ridge estimators based on the asymptotical mean square error. Finally, the influence diagnostic approaches are applied to a real data.

Keyword: Generalized Cross Validation, Spline Smoothing, Measurement Error, Multicollinearity, Ridge Estimator, Semiparametric Linear Model

Mathematics Subject Classification (2010): 62J05, 62J0