مجله علوم آماری، پاییز و زمستان ۱۳۹۶
جلد ۱۱، شماره ۲، ص ۲۱۹ – ۲۴۰
DOI: 10.18869/acadpub.jss.11.2.219

مباحث تشخیصی در مدل‌های خطی آمیخته نیمه‌پارامتری با خطای اندازه‌گیری

هادی امامی، پروانه منصوری القناب

گروه آمار، دانشگاه زنجان

تاریخ دریافت: ۱۳۹۵/۱۱/۱۲ تاریخ آخرین بازنگری: ۱۳۹۶/۱۱/۳۰

چکیده: تمام مشاهدات نقش یکسان در مدل‌های آماری ندارند. گاهی برخی از مشاهدات اثرات نامناسبی روی نتایج تحلیل رگرسیونی دارند. بنابراین شناسایی چنین مشاهداتی در تحلیل داده‌ها از اهمیت ویژه‌ای برخوردار است. برای شناسایی چنین مشاهداتی از روش‌های تشخیصی استفاده می شود. در مقاله حاضر با استفاده از روش حذف موردی و مدل انتقال میانگین نقاط دورافتاده، مباحث تشخیصی در مدل خطی آمیخته نیمه‌پارامتری با خطا در اندازه‌گیری مورد بررسی قرار گرفته است. علاوه بر مباحث حذف موردی، مباحث حذف آزمودنی نیز ارایه شده است. در پایان عملکرد مباحث تشخیصی با استفاده از مجموعه داده‌های واقعی و یک مثال شبیه‌سازی نشان داده شده است.

واژه‌های کلیدی: مدلِ خطی آمیخته، خطا در اندازه گیری، فاصله کوک، روش‌های امتیاز تصحیح شده، حذف موردی.

۱ مقدمه

مباحث تشخیصی در یک تحلیل آماری نقش بسزایی دارند. با حذف یک مشاهده موثر از مجموعه داده‌ها تفاوت معنی‌داری در تحلیل داده‌ها به وجود می‌آید. حذف موردی مشاهدات پایه‌ای برای ساختن آماره‌های تشخیصی است. روش‌های تشخیصی در بسیاری از مدل‌های مختلف مطرح شده است از جمله و مدل های خطی آمیخته تعمیم داده شده است. بکمن و همکاران (۱۹۸۷)، لیسافر و وربک (۱۹۹۸) اندازه‌های تاثیر موضعی را در این مدل‌ها مطرح کردند. کریستنسن و همکاران (۱۹۹۲) و بنرجی و فریز (۱۹۹۷) به

آدرس الکترونیکی نویسنده مسئول مقاله: هادی امامی، h.emami@znu.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62J20، 62J12

ترتیب تحلیل حذف موردی[1] و حذف آزمودنی[2] را مورد مطالعه قرار دادند. اخیرا روش حذف موردی توسط راسخ و امامی (۲۰۱۱) و امامی و امامی (۲۰۱۶) در مدل های خطی مقید با خطاهای همبسته تعمیم داده شده است. همچنین امامی و امامی (۲۰۱۴) روش اندازه تاثیر موضعی را در چنین مدل هایی مورد بررسی قرار داده اند. در مدل‌های رگرسیون ناپارامتری مباحث تشخیصی کم‌تر مورد توجه قرار گرفته‌اند. در میان مطالعات انجام شده، کیم (۱۹۹۶)، کیم و همکاران (۲۰۰۲) باقی‌مانده‌ها، داده‌های نافذ و آماره کوک را در مدل‌های رگرسیونی ناپارامتری و نیمه‌پارامتری مورد کنکاش قرار داده‌اند. مشابه پژوهش کیم و همکاران (۲۰۰۲)، فانگ و همکاران (۲۰۰۲) چنین مباحثی را در مدل‌های خطی آمیخته نیمه‌پارامتری تعمیم داده‌اند. ژانگ و همکاران (۱۹۹۸) این موضوع را مطرح کردند که مدل‌های آمیخته نیمه‌پارامتری اغلب در مطالعات طولی به دلیل ترکیب تأثیر زمان ناپارامتری علاوه بر رگرسیون خطی روی متغیرهای کمکی مفید هستند. از طرفی با توجه به پژوهش‌های دیویدیان و گیلتینان (۱۹۹۵) و لیانگ و لی (۲۰۰۹) متغیرهای مستقل در مدل‌های خطی پارامتری و نیمه‌پارامتری ممکن است با یک خطای غیر قابل اغماضی اندازه‌گیری شده باشند. عدم توجه به این مساله باعث بروز نتایج نادرست در یک تحقیق خواهد شد. بنابراین در چنین مواقعی در نظر گرفتن مدل‌های خطی آمیخته نیمه‌پارامتری با خطا در اندازه‌گیری می‌تواند مطلوب واقع شود. یکی از روش های متداول در برآورد پارامترهای مدل خطی با خطا در اندازه‌گیری استفاده از روش امتیازی تصحیح شده ناکامورا (۱۹۹۰) است. اخیرا زارع و همکاران (۲۰۱۱)، قپانی و همکاران (۲۰۱۶) و امامی (۲۰۱۵) چنین روشی را به ترتیب در مدل های آمیخته خطی و مدل های رگرسیون ریدج با خطا در اندازه گیری تعمیم داده اند. با این وجود در زمینه مباحث تشخیصی در مدل‌های خطی آمیخته نیمه‌پارامتری با خطا در اندازه‌گیری مطالعات چندانی در دسترس نیست. لذا در این مقاله با توجه به پژوهش‌های صورت گرفته توسط زارع و راسخ (۲۰۱۱)، مدل حذف موردی و مدل انتقال میانگین نقاط دورافتاده [3] در مدل‌های خطی آمیخته نیمه‌پارامتری با خطای اندازه‌گیری با استفاده از تابع امتیاز تصحیح شده ناکامورا [4] مورد بررسی و کنکاش قرار می‌گیرد. برای کشف نقاط دورافتاده از آماره آزمون امتیازی اصلاح شده بر پایه مدل‌های انتقال میانگین استفاده خواهد شد. علاوه بر این چندین تحلیل حذف موردی و حذف آزمودنی به عنوان ابزاری برای تحلیل مباحث تشخیصی صورت خواهد گرفت.

[1] Case deletion
[2] Subject deletion
[3] Mean shift outlier model
[4] Nakamura corrected score function

۲ تعریف مدل

شکل ماتریسی مدل خطی آمیخته نیمه‌پارامتری با خطای اندازه گیری به صورت

$$\boldsymbol{Y} = \boldsymbol{X}_*\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + f(\boldsymbol{t}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{X} = \boldsymbol{X}_* + \boldsymbol{\Delta} \tag{۱}$$

تعریف می‌شود، که در آن $\boldsymbol{Y} = (\boldsymbol{Y}_۱, \cdots, \boldsymbol{Y}_m)^T$ طوری $\boldsymbol{Y}_i$ بردار $n \times ۱$ از مشاهدات در i مین آزمودنی (خوشه) است. بردار $\boldsymbol{\beta}$ یک بردار پارامتری $p \times ۱$ از ضرایب رگرسیونی، $\boldsymbol{X}_*$ و $\boldsymbol{Z} = [\boldsymbol{Z}_۱^T, \cdots, \boldsymbol{Z}_m^T]^T$ ماتریس‌های پیش‌بین به ترتیب با ابعاد $n \times p$ و $n \times q$ هستند. $\boldsymbol{Z}_i$ ها ماتریس طرح معلوم و $n \times q_i$ از اثرات تصادفی هستند. $\boldsymbol{b}^T = (\boldsymbol{b}_۱^T, \cdots, \boldsymbol{b}_m^T)$ و هرکدام از $\boldsymbol{b}_i$ ها بردار تصادفی نامشخص از اثرات تصادفی با توزیع $N(۰, \sigma_i^۲\boldsymbol{I})$ که $i = ۱, \cdots, m$ است. $f(.)$ یک تابع نامعلوم هموار و دوبار مشتق‌پذیر روی بازه متناهی، $\boldsymbol{t} = (t_۱, \cdots, t_n)$ که t_i ها اسکالر و $a \leq t_۱, \cdots, t_n \leq b$. $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_۱^T, \cdots, \boldsymbol{\varepsilon}_m^T)$ بردار $n \times ۱$ از خطاهای تصادفی با توزیع $N(۰, \sigma^۲\boldsymbol{I})$ و $\boldsymbol{X}$ مقدار مشاهده شده $\boldsymbol{X}_*$ با خطای اندازه‌گیری $\boldsymbol{\Delta}$ است، $\boldsymbol{\Delta}$ ماتریس تصادفی با توزیع $N(۰, \boldsymbol{I} \otimes \boldsymbol{\Lambda})$ و $\boldsymbol{\Lambda}$ نیز یک ماتریس معین مثبت است. در مدل (۱) فرض می‌شود $\boldsymbol{b}_i$، $\boldsymbol{\varepsilon}$ و $\boldsymbol{\Delta}$ دوبه دو از هم مستقلند و بردار تصادفی $\boldsymbol{b}$ دارای توزیع $N(۰, \sigma^۲\boldsymbol{D})$ است. $\boldsymbol{D}$ ماتریس قطریست که i امین بلوک قطری آن $\gamma_i\boldsymbol{I}$ با $\gamma_i = \sigma_i^۲/\sigma^۲$ مشخص می‌شود. بنابراین توزیع $\boldsymbol{Y}$ نرمال $N(\boldsymbol{X}_*\boldsymbol{\beta} + f(\boldsymbol{t}), \sigma^۲\boldsymbol{V})$ با $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T = \boldsymbol{I} + \sum_{i=۱}^m \gamma_i \boldsymbol{Z}_i\boldsymbol{Z}_i^T$ است. فرض کنید مقادیر گسسته مرتب شده $t_۱ \cdots t_n$ را با $s_۱ \cdots s_q$ نشان داده شود. ارتباط بین $t_۱ \cdots t_n$ با $s_۱ \cdots s_q$ را ماتریس مجاورت $\boldsymbol{N}$ که $n \times q$ هست مشخص می‌کند؛ طوری که $N_{ij} = ۱$ اگر $s_j = t_i$ ؛ در غیر این‌صورت برابر صفر است. فرض کنید $\boldsymbol{f}$ برداری است که مقدار $a_i = f(s_j)$ را شامل شود. بنابراین مدل (۱) را می‌توان به صورت

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{N}\boldsymbol{f} + \boldsymbol{\varepsilon} \tag{۲}$$

بازنویسی کرد. تابع لگاریتم درستنمایی تاوانیده شده برای مدل (۲) به صورت

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{f}; \boldsymbol{X}_*, \boldsymbol{Y}) \;=\; & C_{\sigma^۲} - \frac{۱}{۲\sigma^۲}(\boldsymbol{Y} - \boldsymbol{X}_*\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{N}\boldsymbol{f})^T(\boldsymbol{Y} - \boldsymbol{X}_*\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{N}\boldsymbol{f}) \\ & - \frac{۱}{۲\sigma^۲}\boldsymbol{b}^T\boldsymbol{D}^{-۱}\boldsymbol{b} - \frac{\lambda}{۲\sigma^۲}\int f''(t)^۲ dt, \end{aligned} \tag{۳}$$

بدست می آید، که در آن $C_{\sigma^2} = -\frac{1}{2}\log(2\Pi\sigma^2)^{n+q} - \frac{1}{2}\log|\boldsymbol{D}|$. طبق روش هارویل (۱۹۷۷) با حل معادله $\frac{\partial \ell(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{f}; \boldsymbol{X}_*, \boldsymbol{Y})}{\partial \boldsymbol{b}} = 0$ براورد $\boldsymbol{b}$ به صورت

$$\tilde{b}_{\boldsymbol{\beta},\boldsymbol{f}}(\boldsymbol{X}_*) = (\boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{D}^{-1})^{-1}\boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{X}_*\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f})$$

بدست می آید. اکنون با جایگذاری برآورد اخیر در معادله (۳) داریم:

$$\ell_p(\boldsymbol{\beta}, \boldsymbol{f}; \boldsymbol{X}_*, \boldsymbol{Y}) = C_{\sigma^2} - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}_*\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}_*\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f}) - \frac{\lambda}{2\sigma^2}\boldsymbol{f}^T\boldsymbol{K}\boldsymbol{f},$$

که در آن λ پارامترهموارسازی است و با مینیمم کردن معیار اعتبارسنجی متقابل تعمیم‌یافته تعیین می‌شود و K ماتریس هموارساز متناهی نامنفی است.

۳ برآورد پارامترها

همان‌طور که بیان شد $\boldsymbol{X}_*$ با خطا اندازه‌گیری شده است. با توجه به ناکامورا (۱۹۹۹) اگر $\boldsymbol{X}_*$ با $\boldsymbol{X}$ جایگزین شود برآوردهایی که از توابع امتیاز به‌دست می آید سازگار نیستند. لذا باید خطای اندازه گیری در مدل تصحیح شود. از جمله روش‌های معمول جهت تصحیح خطا در مدل‌های با خطا در اندازه‌گیری استفاده از روش ناکامورا (۱۹۹۹) است. در این روش از تابع امتیاز تصحیح شده‌ای استفاده می شود که امیدریاضی آن نسبت به توزیع با خطا در اندازه‌گیری با تابع امتیاز بر حسب متغیر مستقل یعنی $\boldsymbol{X}_*$ برابر باشد. باید روابط

$$\begin{aligned} E^*[\partial \ell^*(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{f}; \boldsymbol{X}, \boldsymbol{y})/\partial \boldsymbol{b}] &= \partial \ell(\boldsymbol{\beta}, \boldsymbol{f}; \boldsymbol{X}_*, \boldsymbol{y})/\partial \boldsymbol{b}. \\ E^*[\partial \ell_p^*(\boldsymbol{\beta}, \boldsymbol{f}; \boldsymbol{X}, \boldsymbol{y})/\partial \boldsymbol{\beta}] &= \partial \ell(\boldsymbol{\beta}, \boldsymbol{f}; \boldsymbol{X}_*, \boldsymbol{y})/\partial \boldsymbol{\beta}. \end{aligned} \qquad (۴)$$

برای تابع درستنمایی تاوانیده تصحیح شده $\ell^*(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{f}; \boldsymbol{X})$ برقرار باشند. با $\boldsymbol{\Lambda}$ معلوم $\ell^*(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{f}; \boldsymbol{X}, \boldsymbol{Y})$ به صورت

$$\ell^*(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{f}; \boldsymbol{X}, \boldsymbol{Y}) = C_{\sigma^2} - \frac{1}{2\sigma^2}\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{N}\boldsymbol{f})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b} - \boldsymbol{N}\boldsymbol{f})$$

$$-tr(\boldsymbol{V}^{-1})\boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta}\} - \frac{1}{2\sigma^2}\boldsymbol{b}^T\boldsymbol{D}^{-1}\boldsymbol{b} - \frac{\lambda}{2\sigma^2}\boldsymbol{f}^T\boldsymbol{K}\boldsymbol{f}$$

تغییر می‌کند. بنابراین برآورد $\boldsymbol{b}$ با حل معادله $\frac{\partial \ell^*(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{f};\boldsymbol{X},\boldsymbol{Y})}{\partial \boldsymbol{b}} = 0$ به‌صورت

$$\begin{aligned}\tilde{\boldsymbol{b}}_{\boldsymbol{\beta},\boldsymbol{f}}(\boldsymbol{X}) &= (\boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{D}^{-1})^{-1}\boldsymbol{Z}^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f}) \\ &= \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f})\end{aligned}$$

به دست می‌آید. با وارد کردن $\tilde{\boldsymbol{b}}_{\boldsymbol{\beta},\boldsymbol{f}}(\boldsymbol{X})$ در تابع درستنمایی ℓ^* تابع درستنمایی

$$\begin{aligned}\ell_p^*(\boldsymbol{\beta},\boldsymbol{f};\boldsymbol{X},\boldsymbol{Y}) &= \ell^*(\boldsymbol{\beta},\tilde{\boldsymbol{b}}_{\boldsymbol{\beta},\boldsymbol{f}}(\boldsymbol{X}),\boldsymbol{f};\boldsymbol{X},\boldsymbol{Y}) \\ &= C_{\sigma^2} - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f}) \\ &\quad + \frac{1}{2\sigma^2}tr(\boldsymbol{V}^{-1})\boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta} - \frac{\lambda}{2\sigma^2}\boldsymbol{f}^T\boldsymbol{K}\boldsymbol{f}. \qquad (۵)\end{aligned}$$

حاصل می شود. اکنون با مشتق‌گیری از (۵) نسبت به $\boldsymbol{\beta}$ و $\boldsymbol{f}$ و σ^2 برآوردهای ماکسیمم درستنمایی تاوانیده تصحیح شده[5] (MCPLE) پارامترها به صورت

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} - tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$$

$$\hat{\boldsymbol{f}} = (\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{N} + \lambda\boldsymbol{K})^{-1}\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{Y}$$

$$\hat{\sigma}^2 = \frac{1}{n+q}[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{N}\boldsymbol{f})^T\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{N}\boldsymbol{f}) - tr(\boldsymbol{V}^{-1})\boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta} - \lambda\boldsymbol{f}^T\boldsymbol{K}\boldsymbol{f}]$$

به‌دست می‌آیند، که در آن

$$\boldsymbol{W} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{N}(\boldsymbol{N}^T\boldsymbol{V}^{-1}\boldsymbol{N} + \lambda\boldsymbol{K})^{-1}\boldsymbol{N}^T\boldsymbol{V}^{-1}$$

$$\boldsymbol{W}_* = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} - tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}.$$

[5] Maximum corrected penalized likelihood estimators

در ادامه برآورد پارامتر مرتبط با اثرات تصادفی مدل به صورت

$$\begin{aligned}\hat{\boldsymbol{b}} &= (\boldsymbol{Z}^T\boldsymbol{Z}+\boldsymbol{D}^{-1})^{-1}\boldsymbol{Z}^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{N}\hat{\boldsymbol{f}}) \\ &= \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{N}\hat{\boldsymbol{f}}) \qquad (۶)\end{aligned}$$

حاصل می شود. مقادیر برازش شده مدل را نیز می‌توان به صورت

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{N}\hat{\boldsymbol{f}}+\boldsymbol{Z}\hat{\boldsymbol{b}} = \boldsymbol{H}\boldsymbol{Y}$$

نوشت، که درآن $\boldsymbol{H} = \boldsymbol{I}-\boldsymbol{V}^{-1}+\boldsymbol{V}^{-1}\bar{\boldsymbol{H}}$ و

$$\bar{\boldsymbol{H}} = \begin{bmatrix}\boldsymbol{X} & \boldsymbol{N}\end{bmatrix}\begin{bmatrix}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}-tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda} & \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{N} \\ \boldsymbol{N}^T\boldsymbol{V}^{-1}\boldsymbol{X} & \boldsymbol{N}^T\boldsymbol{V}^{-1}\boldsymbol{N}+\lambda\boldsymbol{K}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{X}^T \\ \boldsymbol{N}^T\end{bmatrix}\boldsymbol{V}^{-1}. \qquad (۷)$$

هم‌چنین بردار باقی‌مانده مدل از $\tilde{\boldsymbol{e}} = \boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{N}\hat{\boldsymbol{f}} = (\boldsymbol{I}-\bar{\boldsymbol{H}})\boldsymbol{Y}$ بدست می آید.

۴ مباحث تشخیصی حذف موردی

یک مشاهده که به طور قابل توجهی متفاوت‌تر از سایر مشاهدات است می‌تواند تفاوت زیادی در نتایج تحلیل رگرسیونی ایجاد کند. مباحث تشخیصی تکنیکی برای تشخیص این مشاهدات است. در مباحث تشخیصی از دو مدل متداول حذف موردی و مدل انتقال میانگین نقاط دورافتاده استفاده می‌شود.

۱.۴ مدل حذف موردی

حذف موردی مشاهدات پایه‌ای برای ساختن آماره‌های تشخیصی است. در مدل‌های رگرسیونی خطی معمولاً از روش حذف موردی به منظور بررسی تأثیر تك مشاهدات روی برآورد پارامترها استفاده می‌شود. قضیه زیر روشی برای برآورد پارامترها در مدل حذف موردی ارایه می‌کند.

قضیه ۱: برای مدل (۲) روابط

$$\hat{\boldsymbol{\beta}}_{(ij)} = \hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} - tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{d}_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{e}}}{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{I} - \bar{\boldsymbol{H}})\boldsymbol{d}_c}$$

$$\hat{\boldsymbol{f}}_{(ij)} = \hat{\boldsymbol{f}} - \frac{(\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{N} + \lambda\boldsymbol{K})^{-1}\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{d}_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{e}}}{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{I} - \bar{\boldsymbol{H}})\boldsymbol{d}_c}$$

برقرار است، که در آن (i,j)امین مشاهده شماره $c = n_1 + \cdots + n_{i-1} + j$ را شامل می شود. $\boldsymbol{d}_c$ یک بردار $n \times 1$ است که مولفه c ام آن برابر یک و بقیه مولفه‌ها صفر هستند.

برهان: فرض کنید $c = 1$، ماتریس افراز شده

$$\boldsymbol{V} = \begin{bmatrix} v_{cc} & \mathbf{v}_c^T \\ \mathbf{v}_c & \boldsymbol{V}_{(c)} \end{bmatrix}$$

را در نظر بگیرد و فرض کنید $\boldsymbol{Y}_c$، $\boldsymbol{N}_c$ و $\boldsymbol{X}_c$ ردیف c ام ماتریس‌های $\boldsymbol{Y}$، $\boldsymbol{N}$ و $\boldsymbol{X}$ باشند و $\boldsymbol{Y}_{(c)}$، $\boldsymbol{N}_{(c)}$ و $\boldsymbol{X}_{(c)}$ ماتریس‌هایی باشند که ردیف cام آن‌ها حذف شده است. اکنون با تعریف روابط

$$\begin{aligned}
\boldsymbol{Y}_i^* &= (Y_{i1}, \cdots, Y_{ij-1}, Y_{ij}^*, Y_{ij+1}, \cdots, Y_{in_i})^T \\
\boldsymbol{Y}^* &= (\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_{i-1}^T, \boldsymbol{Y}_i^{*T}, \boldsymbol{Y}_{i+1}^T, \cdots, \boldsymbol{Y}_m^T)^T \\
Y_{ij}^* &= \boldsymbol{X}_{ij}^T + \hat{f}_{(ij)}(t_{ij}) + \mathbf{v}_c^T\boldsymbol{V}_{(c)}^{-1}(\boldsymbol{Y}_{(c)} - \boldsymbol{X}_{(c)}\hat{\boldsymbol{\beta}}_{(ij)} - \boldsymbol{N}_{(c)}\hat{\boldsymbol{f}}_{(ij)})
\end{aligned}$$

با توجه به فانگ و همکاران (۲۰۰۲) برای پارامتر $\boldsymbol{\beta}$ و منحنی هموارساز $\boldsymbol{f}$ خواهیم داشت:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(ij)} &= (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} - tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}^* \\
&= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} - tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{d}_c(Y_{ij} - Y_{ij}^*), \\
\hat{\boldsymbol{f}}_{(ij)} &= (\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{N} + \lambda\boldsymbol{K})^{-1}\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{Y}^* \\
&= \hat{\boldsymbol{f}} - (\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{N} + \lambda\boldsymbol{K})^{-1}\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{d}_c(Y_{ij} - Y_{ij}^*).
\end{aligned}$$

از طرفی با توجه به تعاریف

$$\boldsymbol{Y}_{(c)} = Y_{ij}, \qquad \boldsymbol{N}_{(c)}\boldsymbol{f} = f(t_{ij}), \qquad \boldsymbol{d}_c^T\boldsymbol{V}^{-1} = s_c^{-1}(1, -\mathbf{v}_c^T\boldsymbol{V}_{(c)}) \qquad (8)$$

و $\boldsymbol{X}\hat{\boldsymbol{\beta}}_{(ij)} + \boldsymbol{N}\hat{\boldsymbol{f}}_{(ij)} = \bar{\boldsymbol{H}}\boldsymbol{Y}^*$ داریم

$$\begin{aligned} Y_{ij} - Y_{ij}^* &= s_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(ij)} - \boldsymbol{N}\hat{\boldsymbol{f}}_{(ij)}) \\ &= s_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \bar{\boldsymbol{H}}\boldsymbol{Y}^*) \\ &= s_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\{\boldsymbol{Y} - \bar{\boldsymbol{H}}\boldsymbol{Y} + \bar{\boldsymbol{H}}(\boldsymbol{Y} - \boldsymbol{Y}^*)\} \\ &= s_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \bar{\boldsymbol{H}}\boldsymbol{Y}) + s_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\bar{\boldsymbol{H}}\boldsymbol{d}_c(Y_{ij} - Y_{ij}^*). \end{aligned}$$

که در آن

$$Y_{ij} - Y_{ij}^* = \frac{s_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \bar{\boldsymbol{H}}\boldsymbol{Y})}{1 - s_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\bar{\boldsymbol{H}}\boldsymbol{d}_c} = \frac{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{e}}}{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{I} - \bar{\boldsymbol{H}})\boldsymbol{d}_c}.$$

۲.۴ مدل انتقال میانگین نقاط دورافتاده

با توجه به این‌که داده‌های دورافتاده در تمام مراحل مربوط به تحلیل و تفسیر اطلاعات تاثیرگذار هستند در این بخش مدل تشخیصی انتقال میانگین نقاط دورافتاده مورد بررسی قرار می گیرد. مدل انتقال میانگین نقاط دورافتاده برای مدل (۲) به صورت

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{N}\boldsymbol{f} + \boldsymbol{\phi} + \boldsymbol{\varepsilon}_{ij} \qquad (9)$$

تعریف می‌شود، که در آن $\boldsymbol{\phi}$ نمادی برای نشان دادن نقطه دورافتاده در مدل است. از آزمون فرض $\boldsymbol{\phi} = 0$می‌توان به عنوان آزمون نقاط دورافتاده استفاده کرد. برای مدل (۹) برآوردهای درستنمایی تاوانیده تصحیح شده پارامترهای $\boldsymbol{\beta}$، $\boldsymbol{\phi}$ و $\boldsymbol{f}$ را با علائم $\hat{\boldsymbol{\beta}}_{mij}$، $\hat{\boldsymbol{\phi}}_{mij}$ و $\hat{\boldsymbol{f}}_{mij}$ نشان داده می‌شود.

قضیه ۲: برای مدل انتقال میانگین نقاط دورافتاده داریم:

$$\hat{\boldsymbol{\beta}}_{ij} = \hat{\boldsymbol{\beta}}_{mij} \qquad \hat{\boldsymbol{f}}_{ij} = \hat{\boldsymbol{f}}_{mij}$$

برهان: با توجه به فرض های قضیه ۱ می‌توان گفت $(\hat{\boldsymbol{\beta}}_{(ij)}, \hat{\boldsymbol{f}}_{(ij)})$ تابع درستنمایی تاوانیده تصحیح شده

$$\begin{aligned} L_{ij}(\boldsymbol{\beta}, \boldsymbol{f}) &= (\boldsymbol{Y}_{(c)} - \boldsymbol{X}_{(c)}\boldsymbol{\beta} - \boldsymbol{N}_{(c)}\boldsymbol{f})^T \boldsymbol{V}_{(c)}^{-1} (\boldsymbol{Y}_{(c)} - \boldsymbol{X}_{(c)}\boldsymbol{\beta} - \boldsymbol{N}_{(c)}\boldsymbol{f}) \quad (۱۰) \\ &\quad + \lambda \left(\int f''(t)^2 dt \right) - tr(\boldsymbol{V}^{-1}) \boldsymbol{\beta}^T \boldsymbol{\Lambda} \boldsymbol{\beta} \end{aligned}$$

را مینیمم می‌سازد و $(\hat{\boldsymbol{\beta}}_{(mij)}, \hat{\boldsymbol{f}}_{(mij)}, \hat{\phi})$ تابع

$$\begin{aligned} L_{mij}(\boldsymbol{\beta}, \boldsymbol{f}, \phi) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f} - \phi \boldsymbol{d}_c)^T V^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f} - \phi \boldsymbol{d}_c) \quad (۱۱) \\ &\quad + \lambda \int f''(t)^2 dt - tr(\boldsymbol{V}^{-1}) \boldsymbol{\beta}^T \boldsymbol{\Lambda} \boldsymbol{\beta} \end{aligned}$$

رامینیم می‌کند. با مشتق گیری از تابع درستنمایی $L_{mij}(\boldsymbol{\beta}, \boldsymbol{f}, \phi)$ نسبت به ϕ رابطه

$$\hat{\phi}_{mij}(\boldsymbol{\beta}, \boldsymbol{f}) = (\boldsymbol{d}_c^T \boldsymbol{V}^{-1} \boldsymbol{d}_c)^{-1} \boldsymbol{d}_c^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{N}\boldsymbol{f})$$

می‌شود که با جایگذاری برآورد ϕ در L_{mij} خواهیم داشت:

$$L_{ij} = (\boldsymbol{\beta}, \boldsymbol{f}) = L_{mij}\{\boldsymbol{\beta}, \boldsymbol{f}, \hat{\phi}_{mij}(\boldsymbol{\beta}, \boldsymbol{f})\}.$$

قضیه ۳: آماره آزمون امتیازی برای آزمون فرض $\phi = 0$ در مدل انتقال میانگین نقاط دورافتاده عبارت است از:

$$SC_{ij} = \frac{(\boldsymbol{V}^{-1} \boldsymbol{d}_c \boldsymbol{d}_c^T \boldsymbol{V}^{-1}) \tilde{e}^2}{S_c}$$

برهان: ابتدا برای محاسبه ماتریس اطلاع فیشر تصحیح شده به روش تابع درستنمایی تصحیح شده داریم:

$$\begin{aligned} L_{mij}(\boldsymbol{\beta},\boldsymbol{f},\boldsymbol{\phi}) &= (\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{Nf}-\boldsymbol{\phi d}_c)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{Nf}-\boldsymbol{\phi d}_c) \\ &\quad +\lambda(\int f''(t)^2dt)-tr(\boldsymbol{V}^{-1})\boldsymbol{\beta}^T\boldsymbol{\Lambda\beta} \end{aligned}$$

بنابراین ماتریس اطلاع فیشر تصحیح شده به صورت

$$\boldsymbol{J}(\boldsymbol{\beta},\boldsymbol{\phi})=\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}-tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda} & \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{d}_c \\ \boldsymbol{d}_c^T\boldsymbol{V}^{-1}\boldsymbol{X} & \boldsymbol{d}_c^T\boldsymbol{V}^{-1}\boldsymbol{d}_c \end{bmatrix}$$

خواهد بود. با توجه به کوک و ویزبرگ (۱۹۸۲) آماره آزمون امتیازی تحت فرض $\boldsymbol{\phi}=0$ به صورت

$$SC_{ij}=[\frac{\partial}{\partial\boldsymbol{\phi}}L_{mij}(\boldsymbol{\beta},\boldsymbol{f},\boldsymbol{\phi})]^T\boldsymbol{J}^{\boldsymbol{\phi\phi}}[\frac{\partial}{\partial\boldsymbol{\phi}}L_{mij}(\boldsymbol{\beta},\boldsymbol{f},\boldsymbol{\phi})]|_{(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{f}})}$$

محاسبه می‌شود که در آن $J^{\phi\phi}$ گوشه سمت راست پایینی ماتریس $\boldsymbol{J}^{-1}(\boldsymbol{\beta},\boldsymbol{\phi})$ است. در نتیجه آماره آزمون امتیازی برای مدل (۹) عبارت است از:

$$SC_{ij}=(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{N}\hat{\boldsymbol{f}})^T(\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\boldsymbol{d}_c)^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{N}\hat{\boldsymbol{f}})$$

با استفاده از رابطه $(\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\boldsymbol{d}_c)^{-1}=S_c$ و $\tilde{\boldsymbol{e}}=\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{N}\hat{\boldsymbol{f}}$ نتیجه‌ی مورد نظر حاصل می‌شود یعنی:

$$SC_{ij}=\frac{(\boldsymbol{V}^{-1}\boldsymbol{d}_c\boldsymbol{d}_c^T\boldsymbol{V}^{-1})\tilde{\boldsymbol{e}}^2}{S_c}.$$

۳.۴ فاصله کوک

روش‌های زیادی برای بررسی تاثیر حذف مشاهده بر جوانب مختلف مدل برازش شده وجود دارد. یکی از این روش‌ها استفاده از فاصله کوک است، که نخستین بار توسط کوک (۱۹۷۷) معرفی شد. فاصله کوک تعمیم یافته با استفاده از توان دوم تغییرات وزنی حاصل از حذف مشاهده (ij) ام در براورد پارامترها

به‌صورت

$$CD_{ij}(\boldsymbol{\beta},\boldsymbol{f})=\frac{(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{(ij)})^T\boldsymbol{C}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{(ij)})}{\hat{\sigma}^2}$$

تعریف می‌شود، که در آن

$$\boldsymbol{C}=\begin{bmatrix}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}-tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda} & \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{N}\\ \boldsymbol{N}^T\boldsymbol{V}^{-1}\boldsymbol{X} & \boldsymbol{N}^T\boldsymbol{V}^{-1}\boldsymbol{N}+\lambda\boldsymbol{K}\end{bmatrix}. \qquad (۱۲)$$

اکنون با استفاده از روابط (۷) ، (۸) و قضیه ۱ داریم:

$$CD_{ij}(\boldsymbol{\beta},\boldsymbol{f})=\frac{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\bar{\boldsymbol{H}}\boldsymbol{d}_c}{\hat{\sigma}^2}\{\frac{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{e}}}{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{I}-\bar{\boldsymbol{H}})\boldsymbol{d}_c}\}^2=\frac{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\bar{\boldsymbol{H}}\boldsymbol{d}_c}{\hat{\sigma}^2\{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{I}-\bar{\boldsymbol{H}})\boldsymbol{d}_c\}}t_c^2$$

که در آن

$$t_c=\frac{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{e}}}{\sqrt{\{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{d}_c\}}}$$

باقی‌مانده استیودنت شده مورد c ام است. فاصله کوک برای جزء ناپارمتری $\boldsymbol{\beta}$ نیز به صورت

$$CD_{ij}(\boldsymbol{\beta})=\frac{(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(ij)})^T\{(\boldsymbol{I}_p,0)\boldsymbol{C}^{-1}(\boldsymbol{I}_p,0)^T\}^{-1}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(ij)})}{\hat{\sigma}^2}$$

تعریف می‌شود. از رابطه (۱۲) داریم $(\boldsymbol{I}_p,0)\boldsymbol{C}^{-1}(\boldsymbol{I}_p,0)^T=\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}-tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda}$، آن‌گاه می‌توان نوشت:

$$CD_{ij}(\boldsymbol{\beta})=\frac{[\boldsymbol{d}_c^T\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}-tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{d}_c](\boldsymbol{d}_c^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{e}})^2}{\hat{\sigma}^2\{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{d}_c\}^2}$$

یا به عبارتی دیگر

$$CD_{ij}(\boldsymbol{\beta})=\frac{\boldsymbol{d}_c^T\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}-tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{d}_c}{\hat{\sigma}^2\{\boldsymbol{d}_c^T\boldsymbol{V}^{-1}(\boldsymbol{I}-\tilde{\boldsymbol{H}})\boldsymbol{d}_c\}}t_c^2.$$

مقادیر بزرگ برای $CD_{ij}(\boldsymbol{\beta})$ نشان دهنده تاثیر مشاهده (i,j) ام روی برآورد $\boldsymbol{\beta}$ است.

۴.۴ فاصله درستنمایی

معیار تشخیص تعمیم یافته دیگری که بر پایه لگاریتم تابع درستنمایی به‌دست می‌آید فاصله درستنمایی است (کوک و ویزبرگ، ۱۹۸۲). فرض کنید $L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \boldsymbol{X}, \boldsymbol{Y})$ تابع لگاریتم درستنمایی تصحیح شده روی مجموعه داده‌ها باشد. در این‌صورت فاصله درستنمایی به صورت

$$LD_{ij}(\boldsymbol{\beta}) = ۲[L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}, \boldsymbol{X}, \boldsymbol{Y}) - L(\hat{\boldsymbol{\beta}}_{(ij)}, \hat{\boldsymbol{f}}, \boldsymbol{X}, \boldsymbol{Y})]$$

تعریف می‌شود. با بسط تیلور $L(\hat{\boldsymbol{\beta}}_{(ij)}, \hat{\boldsymbol{f}}, \boldsymbol{X}, \boldsymbol{Y})$ حول $\hat{\boldsymbol{\beta}}$ خواهیم داشت:

$$LD_{ij}(\boldsymbol{\beta}) = ۲[\dot{J}^{*T}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{ij}) + \frac{۱}{۲}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{ij})^T\{-\ddot{J}(\hat{\boldsymbol{\beta}})\}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{ij})]$$

که در آن $\dot{J}^{*T}(\hat{\boldsymbol{\beta}}) = \frac{\partial \ell_p^*(\boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{X}, \boldsymbol{Y})}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}, \hat{\boldsymbol{f}}=\boldsymbol{f}}$ و $\ddot{J} = \frac{\partial^۲ l_p^*(\boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{X}, \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}, \hat{\boldsymbol{f}}=\boldsymbol{f}}$.
این نتایج دقیق است زیرا مشتق سوم برابر صفر است. چون $\dot{J}^{*T}(\hat{\boldsymbol{\beta}}) = ۰$، فاصله درستنمایی مورد نظر برابر

$$LD_{ij}(\boldsymbol{\beta}) = \frac{\boldsymbol{d}_c^T \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} - tr(\boldsymbol{V}^{-۱})\boldsymbol{\Lambda}^{-۱}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{d}_c}{\hat{\sigma}^۲\{\boldsymbol{d}_c^T\boldsymbol{V}^{-۱}(\boldsymbol{I} - \tilde{\boldsymbol{H}})\boldsymbol{d}_c\}}$$

است. همان‌طور که انتظار می رود $LD_{ij}(\boldsymbol{\beta}) = CD_{ij}(\boldsymbol{\beta})$ است.

۵ مباحث تشخیصی حذف آزمودنی

در این قسمت تاثیر حذف آزمودنی روی برآورد $\boldsymbol{\beta}$ و $\boldsymbol{f}$ را بررسی می‌شود. فرض کنید $\hat{\boldsymbol{\theta}}_{[i]} = (\hat{\boldsymbol{\beta}}_{[i]}^T, \hat{\boldsymbol{f}}_{[i]}^T)^T$ برآورد $\boldsymbol{\beta}$ و $\boldsymbol{f}$ با حذف آزمودنی i ام باشد در این صورت داریم:

$$\hat{\boldsymbol{\theta}}_{[i]} = \hat{\boldsymbol{\theta}} - \boldsymbol{C}^{-۱}\begin{bmatrix}\boldsymbol{X}^T \\ \boldsymbol{N}^T\end{bmatrix}\boldsymbol{V}^{-۱}\boldsymbol{E}_i(\boldsymbol{I}_{n_i} - \bar{\boldsymbol{H}}_i)^{-۱}\boldsymbol{E}_i^T\tilde{\boldsymbol{e}} \qquad (۱۳)$$

که در آن

$$\bar{\boldsymbol{H}}_i = \begin{bmatrix} \boldsymbol{X}_i & \boldsymbol{N}_i \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} - tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda} & \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{N} \\ \boldsymbol{N}^T\boldsymbol{V}^{-1}\boldsymbol{X} & \boldsymbol{N}^T\boldsymbol{V}^{-1}\boldsymbol{N} + \lambda\boldsymbol{K} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_i^T \\ \boldsymbol{N}_i^T \end{bmatrix} \boldsymbol{V}_i^{-1}.$$

با استفاده از (۱۳) آماره‌ی کوک تعمیم یافته برای $(\boldsymbol{\beta}, \boldsymbol{f})$ به صورت

$$CD_{[i]}(\boldsymbol{\beta}, \boldsymbol{f}) = \frac{\tilde{\boldsymbol{e}}^T\boldsymbol{E}_i(\boldsymbol{I}_{n_i} - \bar{\boldsymbol{H}}_i^T)^{-1}\boldsymbol{V}_i^{-1}\bar{\boldsymbol{H}}_i(\boldsymbol{I}_{n_i} - \bar{\boldsymbol{H}}_i)^{-1}\boldsymbol{E}_i^T\tilde{\boldsymbol{e}}}{\hat{\sigma}^2}$$

تعریف می‌شود. که در آن $\boldsymbol{E}_i^T\tilde{\boldsymbol{e}}$ بردار باقی مانده $n_i \times 1$ متناظر با آزمودنی i ام است. با استفاده از رابطه $\hat{\boldsymbol{\beta}}_{[i]} - \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} - tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{E}_i(\boldsymbol{I}_{n_i} - \bar{\boldsymbol{H}}_i)^{-1}\boldsymbol{E}_i^T\tilde{\boldsymbol{e}}$ فاصله کوک برای پارامتر $\boldsymbol{\beta}$ به صورت

$$CD_{[i]}(\boldsymbol{\beta}) = \frac{\boldsymbol{R}_i^T\boldsymbol{H}_{\beta,i}\boldsymbol{R}_i}{\hat{\sigma}^2}$$

بیان می‌شود، که در آن $\boldsymbol{R}_i = (\boldsymbol{I}_{n_i} - \bar{\boldsymbol{H}}_i)^{-1}\boldsymbol{E}_i^T\tilde{\boldsymbol{e}}$ و $\boldsymbol{H}_{\beta,i}$ را به صورت

$$\boldsymbol{H}_{\beta,i} = \boldsymbol{E}_i^T\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} - tr(\boldsymbol{V}^{-1})\boldsymbol{\Lambda})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{E}_i \qquad (۱۴)$$

تعریف می‌شود. از آن‌جا که حذف آزمودنی i ام n_i نقطه زمانی را شامل می‌شود بنابراین به محاسبه فاصله کوک برای تعیین تاثیر موضعی روی منحنی $\boldsymbol{f}$ نیاز است. در نتیجه طبق فرمول فاصله کوک خواهیم داشت:

$$CD_{[i]}(\boldsymbol{f}) = \frac{\boldsymbol{R}_i^T\boldsymbol{W}_*\boldsymbol{N}\boldsymbol{S}^{-1}\boldsymbol{N}_i^T(\boldsymbol{N}_i\boldsymbol{S}^{-1}\boldsymbol{N}_i^T)^{-1}\boldsymbol{N}_i\boldsymbol{S}^{-1}\boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{R}_i}{\hat{\sigma}^2}$$

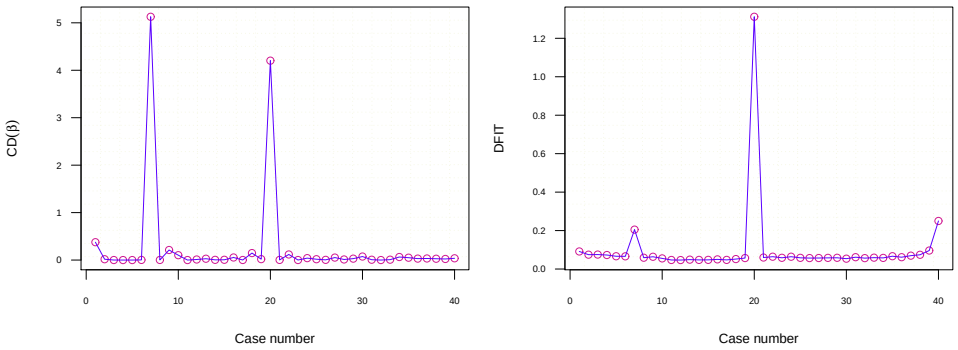
که در آن $\boldsymbol{S} = \boldsymbol{N}^T\boldsymbol{W}_*\boldsymbol{N} + \lambda\boldsymbol{K}$ است.

۶ مطالعه شبیه‌سازی

برای بررسی مباحثی که عنوان شد متغیر پاسخ $\boldsymbol{Y}_{ij}$ از مدل

$$\boldsymbol{Y}_{ij} = \boldsymbol{x}_{ij}^{(۱)}\boldsymbol{\beta}_۱ + \boldsymbol{x}_{ij}^{(۲)}\boldsymbol{\beta}_۲ + \boldsymbol{b}_{۱j} + f(t_{ij}) + \varepsilon_{ij} \qquad i = ۱, \cdots, m; \quad j = ۱, \cdots, q.$$

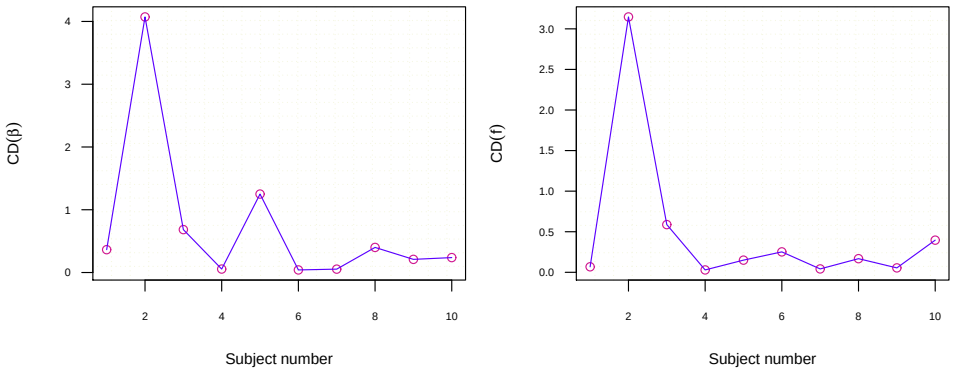
شبیه‌سازی می‌شود، که در آن q تعداد خوشه‌های مستقل و m اندازه خوشه در مطالعات طولی با اندازه نمونه $n = mq$ است (وانگ و همکاران، ۱۹۹۸). فرم ماتریسی این مدل را می توان به صورت $r = ۱$، $\boldsymbol{b}_۱ = (b_{۱۱}, \cdots, b_{۱q})^T$، $\boldsymbol{Y} = (Y_{۱۱}, \cdots, Y_{۱q}, Y_{۲۱}, \cdots, Y_{۲q}, \cdots, Y_{m۱}, \cdots, Y_{mq})^T$ نوشت. $\boldsymbol{X}$، $\boldsymbol{f}$ و ε هم ساختاری مانند $\boldsymbol{Y}$ دارند. در این شبیه‌سازی $q = ۱۰$ و $m = ۴$ است. هم‌چنین در این شبیه‌سازی $\boldsymbol{\beta}_۱ = ۱$، $\boldsymbol{\beta}_۲ = ۲$، $\boldsymbol{x}_{ij}^{(۱)} \sim N(۳۵, ۳)$، $\boldsymbol{x}_{ij}^{(۲)} \sim N(۲۳, ۳)$، $f(t_{ij}) = (t - ۰/۵)^۲$، و $t_{i۱} \sim U(۰, ۰/۲۵)$، $t_{i,j+۱} = t_{ij} + ۰/۲۵ (j = ۲, ۳, ۴)$، $\boldsymbol{b}_{۱j} \sim N(۰, \sigma_۱^۲)$، $\varepsilon_{ij} \sim N(۰, \sigma^۲)$، $\sigma^۲ = ۰/۶^۲$ و $\sigma_۱^۲ = ۰/۵^۲$ انتخاب شد و $\boldsymbol{\Lambda} = diag(۰/۵^۲, ۰/۵^۲)$ است . برای هر ترکیب از پارامترها ۱۰۰۰ بار تکرار انجام شده است. قبل از اجرای شبیه‌سازی تغییراتی در متغیر پاسخ هفتم و بیستم دادیم.



شکل ۱. مباحث تشخیصی حذف موردی برای داده‌های شبیه‌سازی شده: فاصله کوک برای $\boldsymbol{\beta}$ و $\boldsymbol{f}$

هم‌چنین برای متغیر پیش‌بین $\boldsymbol{X}$ در آزمودنی دوم از شماره پنج تا هشت، هفت واحد به آن‌ها اضافه کردیم. شکل ۱ فاصله کوک در برابر مشاهدات برای پارامتر $\boldsymbol{\beta}$ و قسمت ناپارامتری $\boldsymbol{f}$ را نشان می‌دهد. مشاهده هفتم و بیستم تاثیر زیادی روی برآورد $\boldsymbol{\beta}$ و نیز مشاهده بیستم روی برآورد $\boldsymbol{f}$ هم تاثیر زیادی داشته است. مطابق شکل ۲ نتایج حاصل از شبیه‌سازی نشان می‌دهند که آزمودنی دوم یک مشاهده موثر است. بعد از

آزمودنی دوم، آزمودنی پنج تاثیر بالایی در برآورد فاصله کوک روی پارامتر $\boldsymbol{\beta}$ دارد.



شکل ۲. مباحث تشخیصی حذف آزمودنی برای داده‌های شبیه‌سازی شده : فاصله کوک برای $\boldsymbol{\beta}$ و $\boldsymbol{f}$

۷ تحلیل داده‌های فرسایش شدت باران

خاک‌های مناطق خشک و نیمه‌خشک ایران بشدت تحت تاثیر تجمع نمک‌ها به ویژه گچ و کربنات‌ها هستند. خاک‌های این مناطق ذاتا مواد آلی کم و ساختمان ناپایداری دارند. به علاوه پوشش گیاهی طبیعی سطح خاک ها ضعیف و در معرض عوامل فرساینده قرار دارد. اخیرا بلیانی (۲۰۱۷) پژوهشی به منظور بررسی میزان فرسایش پاشمانی در خاک های با خصوصیات مختلف از استان های زنجان و فارس انجام داده است. در پژوهش او ۱۲ نوع بافت خاک مورد مطالعه قرار گرفته است. از بین ۱۲ کلاس بافت خاک ۱۱ بافت خاک از مناطق مختلف استان زنجان و یک بافت خاک سیلتی از استان فارس به تصادف انتخاب شدند. مجموعا تعداد ۱۴۴ جعبه خاک به ابعاد ۲۵ در ۳۵ سانتی متر مورد بررسی قرار گرفته است. در نمونه‌های خاک، درصد آهک ($\boldsymbol{x}_۱$)، درصد شن ($\boldsymbol{x}_۲$) و درصد رس ($\boldsymbol{x}_۳$) به روش هیدرومتری و درصد ماده آلی خاک ($\boldsymbol{x}_۴$) به روش آزمایشگاهی اندازه گیری شده است. از آنجایی که درصد گچ موجود درخاک ($\boldsymbol{x}_۵$) با استفاده از روش اندازه گیری هدایت الکتریکی است و این روش با خطا در اندازه گیری همراه هست میزان این متغیر با ۱۲ تکرار در هربافت معین شده است همچنین میزان سنگ ریزه خاک با جداسازی ذرات اولیه با تعیین نسبت جرمی آن ($\boldsymbol{x}_۶$) مشخص می‌شود . بدون در نظر گرفتن خطای در اندازه گیری بلیانی و همکاران رابطه بین میزان شدت فرسایش باران ($\boldsymbol{y}$) و ویژگی‌های خاک ($\boldsymbol{x}_۶$-$\boldsymbol{x}_۱$)

را با یک مدل رگرسیونی چندگانه مورد مطالعه قرار دادند. در این بخش یک مدل کاملتری همراه با مباحث تشخیصی (تحلیل حساسیت) مربوط به آن ارایه می‌شود. نخست برای تعیین متغیر بخش ناپارامتری با استفاده از آماره‌ی یاتچیو (۲۰۰۳) فرضیه‌ی خطی بودن $f(t)$ بعبارتی فرضیه $H_0 : f(t) = h(t, \boldsymbol{\beta})$ را به ازای یک تابع مشخص $h(.)$ می‌توان آزمود:

$$Z_0 = \frac{n^{1/2}(s^2_{res} - s^2_{diff})}{s^2_{diff}} \xrightarrow{D} N(0, 1)$$

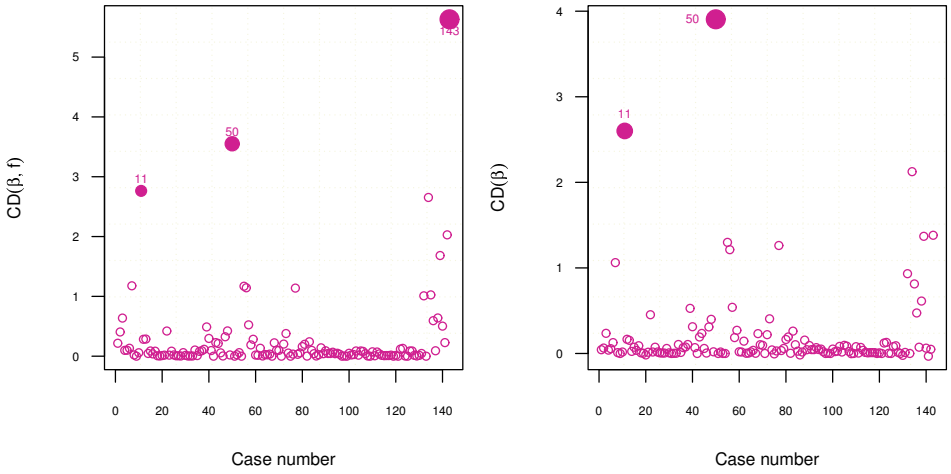
که در آن s^2_{res} برآورد معمول واریانس باقی‌مانده در رگرسیون خطی و $s^2_{diff} = \frac{1}{2n}\sum_{i=2}^{n}(y_i - y_{i-1})^2$ است. برای انجام آزمون فوق برای هریک از متغیرهای $\boldsymbol{x}_j$، $j = 1, \ldots, 6$، درایه های هر متغیر ابتدا باید به شکل صعودی $x_{1j}<\ldots<x_{ij}<\ldots<x_{nj}$ مرتب شود طوری که مقدار y_i در s^2_{res} برابر i امین مشاهده متناظر به $\boldsymbol{x}_{ij}$ مرتب شده است. با انجام این آزمون متغیر $\boldsymbol{x}_6$ را به عنوان بخش ناپارامتری در نظر گرفته می شود زیرا این آماره برای $\boldsymbol{x}_6$ در بین سایر متغیرهای مستقل معنی‌دار است ($Z_0 = 10/531$). با توجه به اطلاعات فوق مدل

$$\boldsymbol{y} = \sum_{j=1}^{4} \boldsymbol{\beta}_j \boldsymbol{x}_j + \boldsymbol{\beta}_5 \boldsymbol{x}_5 + \boldsymbol{Z}_1 \boldsymbol{b}_1 + f(\boldsymbol{x}_6) + \varepsilon, \quad \boldsymbol{x}_5 = \boldsymbol{x}_{*5} + \delta$$

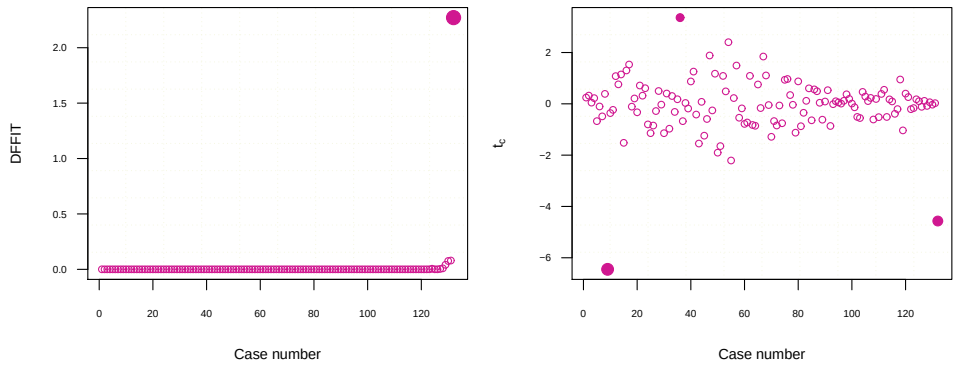
می‌تواند برای برازش داده‌ها مناسب باشد، که در آن $\boldsymbol{Z}_1$ ماتریس طرح اثر نوع بافت خاک ($\boldsymbol{b}_1$) با بعد 12×143 و δ بردار تصادفی خطا با توزیع $N(0, \Lambda)$ است. به دلیل گم شدن اندازه گیری های بعضی متغیرهای مربوط به یک جعبه خاک مدل با $n = 143$ برازش شده است. برای برآورد Λ با در نظر گرفتن $\bar{x}_i$ به عنوان میانگین نمونه‌ای داده‌های تکراری مدل $x_{ij} = x_{*ij} + \delta_{ij}, j = 1, \cdots, m_i, i = 1, \ldots, n$، برآورد ناریب Λ با روش گشتاوری به صورت

$$\hat{\Lambda} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(x_{ij} - \bar{x}_i)^2}{\sum_{i=1}^{n}(m_i - 1)}$$

محاسبه می شود. فاصله کوک تعمیم‌یافته را برای $\boldsymbol{\beta}$ و $(\boldsymbol{\beta}, \boldsymbol{f})$ تحت مدل حذف موردی محاسبه شده و نمودار آن‌ها در شکل ۳ رسم شده است. همانطور که ملاحظه می‌شود مشاهدات ۵۰ و ۱۱ روی $\boldsymbol{\beta}$ و همین مشاهدات با مشاهده ۱۴۳ همزمان روی برآورد $(\boldsymbol{\beta}, \boldsymbol{f})$ موثر هستند. با توجه به نمودار باقی‌مانده‌های

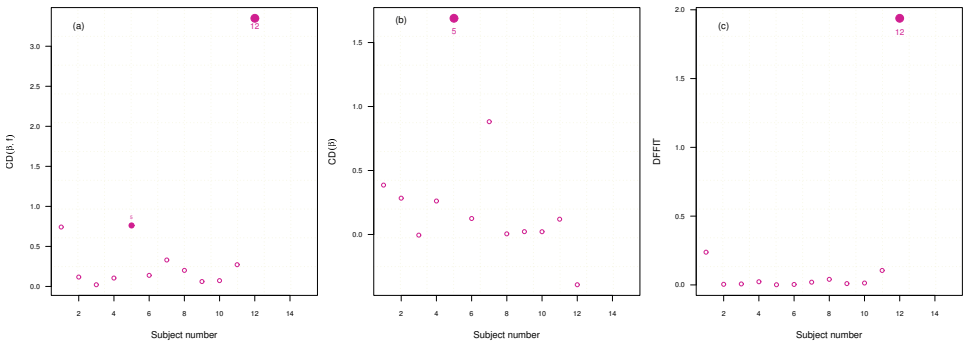


شکل ۳. مباحث تشخیصی حذف موردی برای داده‌های خاکشناسی: فاصله کوک برای $\boldsymbol{\beta}$ و $\boldsymbol{f}$

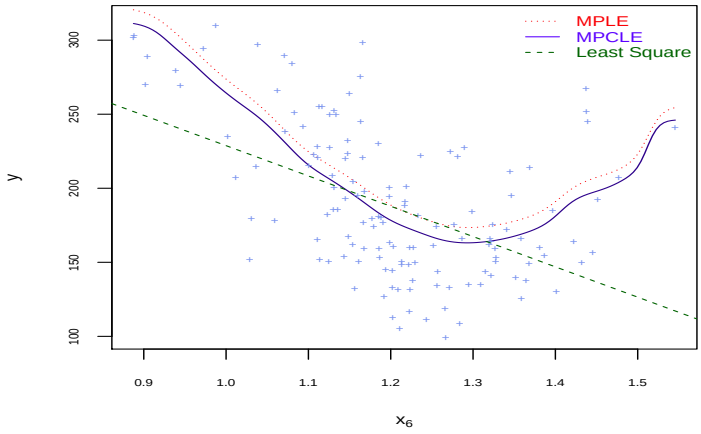


شکل ۴. نمودار باقی‌مانده استیودنت شده و $DFIT$ برای برآورد $\boldsymbol{f}$

استیودنت شده در شکل ۴ سه مشاهده ۵۰، ۱۱ و ۱۳۴ بعنوان مشاهدات دور افتاده شناسایی می‌شوند. هیچکدام از اندازه‌های DFFIT نگرانی جدی ایجاد نمی‌کنند، تنها مشاهده ۱۴۳ نسبت به بقیه مقدار بزرگ‌تری دارد که می‌تواند بعنوان مشاهده موثر روی برآورد $\boldsymbol{f}$در نظر گرفته شود. برای بررسی اثر هر نوع بافت خاک در برآورد پارامترها نمودار $CD_{[i]}(\boldsymbol{\beta}, \boldsymbol{f})$ در شکل ۵ رسم شده است. همانطور که ملاحظه می شود افت خاک از نوع ۵ روی برآورد $\boldsymbol{\beta}$ و بافت خاک از نوع ۱۲ روی برآورد $\boldsymbol{f}$ موثر است.



شکل ۵. مباحث تشخیصی حذف آزمودنی: فاصله کوک برای $\boldsymbol{\beta}$ و $(\boldsymbol{\beta}, \boldsymbol{f})$ و $DFIT$ برای f



شکل ۶. برآورد تابع ناپارامتری (MCPLE) منحنی با خط ممتد و MPLE (منحنی نقطه‌چین)

جدول ۱ برآورد $\boldsymbol{\beta}$، واریانس اثر تصادفی σ_1^2، واریانس خطای مدل σ_1^2 و واریانس خطای اندازه گیری Λ در صورت حذف و عدم حذف مشاهدات موثر را نشان می‌دهد. مقادیر داخل پرانتز مقادیر انحراف استاندارد برآوردها را نشان می‌دهد. درصد تغییرات مربوط به برآورد ضریب هر متغیر پس از حذف مشاهدات موثر در ستون آخر محاسبه شده است. شکل ۶ برآورد قسمت ناپارامتری مدل یعنی $f(x_6)$ را با استفاده از دو روش برآورد ماکسیمم درستنمایی تاوانیده (MPLE) و برآورد ماکسیمم درستنمایی تاوانیده

تصحیح شده (MCPLE) را نشان می‌دهد. پهنای باند که توسط اعتبارسنجی متقابل انتخاب شده برای هر دو روش برآورد به ترتیب ۰/۰۹۸ و ۰/۲۰۱ است. از شکل مشخص است وقتی از روش تصحیح شده ناکامورا استفاده می شود برآورد $f(x_۶)$ کمتر از روشی است که خطاها را نادیده در نظر می‌گیرد. این اختلاف ارتفاع در دو منحنی به دلیل تصحیح خطای اندازه گیری است.

جدول ۱. برآورد ضرایب پارامتری مدل برازش شده به داده های فرسایش شدت باران

ضرایب	کل داده‌ها	حذف مشاهدات ۱۱ و ۵۰	درصد تغییرات
$\hat{\beta}_۱$	۰/۸۸۸(۰/۰۲۱)	۰/۴۷۵(۰/۰۰۹)	۴۶/۵
$\hat{\beta}_۲$	۰/۱۹۲ (۰/۰۰۳)	۰/۱۰۶ (۰/۰۱۱)	۸۱/۱
$\hat{\beta}_۳$	۰/۲۲۰ (۰/۰۰۱)	۰/۳۵۱ (۰/۰۰۸)	۵۹/۵
$\hat{\beta}_۴$	-۶/۴۴۲(۱/۷۹۴)	-۵/۲۱۷(۲/۰۰۳)	۱۸/۹
$\hat{\beta}_۵$	-۲/۶۵۷(۰/۸۴۱)	-۱/۱۰۶(۰/۴۲۲)	۵۸/۳
$\hat{\sigma}_۱^۲$	۰/۰۱۶۹	۰/۰۱۰۳	۰/۶۶۰
$\hat{\sigma}^۲$	۰/۰۸۵	۰/۱۷۹	۱۱۰/۵۸۸
$\hat{\Lambda}$	۰/۱۰۳	-	-

بحث و نتیجه‌گیری

در این مقاله مباحث تشخیصی در مدل خطی آمیخته نیمه‌پارامتری با خطا در اندازه‌گیری با استفاده از تابع امتیاز تصحیح شده ناکامورا بررسی شده است. مباحث تشخیصی حذف موردی و آزمودنی و نیز چندین ابزار تشخیصی برای شناسایی نقاط موثر پیشنهاد شده است. علاوه بر روش نظری با استفاده از شبیه‌سازی نشان داده شد که روش‌های تشخیصی مورد تعمیم یافته برای شناسایی مشاهدات موثر و دورافتاده مناسب هستند. این روش‌ها به تحلیل‌گر داده‌ها در تحلیل مدل خطی آمیخته نیمه‌پارامتری با خطا در اندازه‌گیری کمک شایانی می‌کند. برای مثال در داده‌های فرسایش شدت باران مشاهدات ۵۰ و ۱۱ از بافت خاک‌های استان زنجان به عنوان مشاهدات موثر روی برآورد ضرایب پارامتری مدل شناسایی شدند طوری که این مشاهدات با توجه به جدول ۱ بیشترین درصد تغییر را روی متغیر درصد شن و واریانس مدل ایجاد می

کنند. همچنین مشاهده ۱۴۳ که مربوط به بافت خاک استان فارس است موثرترین مشاهده روی جزء ناپارامتری مدل شناسایی شد. بعنوان یک نکته پایانی می توان گفت مباحث تشخیصی انجام شده در بخش ۷ را می توان مانند یک نوع تحلیل حساسیت درمباحث خاکشناسی در نظر گرفت که امکان سنجش حساسیت نمونه ها و نوع بافت خاک را در یک مدل دقیق را فراهم می آورد.

تشکر و قدردانی

نویسندگان از داوران، سردبیر و ویراستار محترم نشریه که با رهنمودهای ارزنده خود موجب بهتر شدن مقاله شدند و همچنین از دکتر بلیانی بابت در اختیار گذاشتن داده‌ها کمال تشکر را دارند.

مراجع

Balyani, A. (2017), Quantification of Rainfall Erosion in Some Soils of the Semi-arid Regions in North West of Iran, *Ph.D. Thesis in University of Zanjan, Zanjan, Iran* .

Banerjee, M., Frees, L. W. (1997), Influence Diagnostics for Linear Longitudinal Models, *Journal of the American Statistical Association*, **92**, 999–1005.

Beckman, R. J., Nachtsheim, C. J. and Cook, R. D. (1987), Diagnostics for Mixed-models Analysis of Variance, *Technometrics*, **29**, 413–426.

Christensen, R., Pearson, L. M. and Johnson, W. (1992), Case Deletion Diagnostics for Mixed Models, *Technometrics*, **34**, 38–45.

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

Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, London.

Davidian, M, Giltinan, D. M. (1995), *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, London.

Emami, H. (2015), Influence Diagnostics in Ridge Semiparametric Regression Models, *Journal of Statistics and Probablity Letter*, **105**, 106–115.

Emami, H. (2015), Influence Measures in Ridge Linear Measurement Error Models, *Journal of Statistical Research of Iran*, **12**, 39-56.

Emami, H. and Emami, M. (2016), Influence Diagnostics in Constrained General Linear Models. *Communications in Statistics-Theory and Methods*, **45**, 5331-5340.

Emami, H. and Emami, M. (2014), Local Influence in Constrained General Linear Models. *Journal of Data Science*,**12**, 717-726.

Emami, H. and Rasekh, A. (2014), Influence Diagnostics on Testing Linear Hypothesis in Linear Models with Correlated Errors, *Communications in Statistics Theory and Methods*, **4**, 1050-1060.

Fung, W. K., Zhu, Z. Y., Wei, B. C. and He, X. M. (2002), Influence Diagnostics and Outlier Tests for Semiparametric Mixed Models, *Journal of Royal Statitical Society*, **64**, 565- 579.

Ghapani, F., Rasekh, A., and Babadi, B. (2016), The Weighted Ridge Estimator in Stochastic Restricted Linear Measurement Error Models, *Statistical Paper*, DOI 10.1007/s00362-016-0786-3.

Harvill, D. A. (1977), Maximum Likelihood Approaches to Variance Component Estimation and Related Problems (with discussion), *Journal of the American Statistical Association*, **72**, 320-340.

Harrison, D. and Rubinfeld, D. L. (1978), Hedonic Housing Prices And The Demand For Clean Air, *Journal of Environmental Economics and Management*, **5**, 81–102.

Kim, C. (1996), Cook's Distance In Spline Smoothing, *Statistics and Probability Letters*, **31**, 139–144.

Kim, C., Kim, W. and Park, B. U. (2002). Influence Diagnostics in Semiparametric Regression Models, *Statistics and Probability Letters*, **60**, 49-58.

Lesaffre, E. and Verbeke, G. (1998), Local Influence in Linear Mixed Models, *Biometrics*, **54**, 570–582.

Liang, H. and Li, R. (2009), Variable Selection For Partially Linear Models With Measurement Errors, *Journal of the American Statistical Association*, **485**, 234-248.

Nakamura, T. (1990), Corrected Score Function For Errors-in-Variables Models: Methodology and Application to Generalized Linear Models, *Biometrics*, **77**, 127–137.

Wang, N., Lin, X., Gutierrez, R. G. and Carroll, R. J. (1998), Bias Analysis And SIMEX Approach In Generalized Linear Mixed Measurement Error Models, *Journal of American Statistical Association*, **93**, 249–261.

Yatchew, A. (2003), *Semiparametric Regression for the Applied Econometrican*, Cambridge University press , Cambridge.

Zare, K. and Rasekh, A. (2011), Diagnostic Measures for Linear Mixed Measurement Error Models, *SORT*, **35**, 125-144.

Zare, K., Rasekh, A. and Rasekhi, A. (2011), Estimation of Variance Components In Linear Mixed Measurement Error Models, *Statistical Papers*, **53**, 849-863.

Zhong, D., Lin, X. and Sower, M. (1998), Semiparametric Stochastic Mixed models for Longitudinal Data, *Journal of the American Statistical Association*, **93**, 710-719.

Zhong, X. P., Fung, W. K. and Wei, B. C. (2002), Estimation in Linear Models with Random Effects and Errors in Variables, *Annals of the Institute of Statistical Mathematics*, **54**, 595–606.