سیزدهمین کنفرانس آمار ایران
۴-۲ شهریور ۱۳۹۵

تحلیل حساسیت و آزمون نقاط دورافتاده در مدل های آمیخته نیمه پارامتری

هادی امامی[۱]، پروانه منصوری[۲]

دانشگاه زنجان

چکیده: مدل های آمیخته نیمه پارامتری کاربرد گسترده ای در مباحث اقتصاد سنجی و زیست سنجی به ویژه داده های طولی دارند. این مقاله تاثیر تحلیل حساسیت روی برآوردگرهای ماکسیمم درستنمایی تاوانیده را در مدل های حذف موردی و حذف آزمودنی بررسی می کند. هم چنین معادل بودن مدل حذف موردی و حذف آزمودنی با مدل انتقال میانگین نقاط دورافتاده از مباحث دیگر این مقاله می باشد. در ادامه با استفاده از یک مثال شبیه سازی شده کاربرد تحلیل حساسیت ارزیابی می شود.

واژه‌های کلیدی: فاصله کوک، تابع درستنمایی تاوانیده، رگرسیون نیمه پارامتری، اسپلاین هموارسازی

کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲j۲۰ ، ۶۲j۰۲ .

۱ مقدمه

تحلیل حساسیت یکی از بخش های مهم تحلیل آماری به شمار می رود. آماره کوک (۱۹۷۷) یک مبحث مهم تحلیل حساسیت می باشد. امروزه آماره کوک کاربرد گسترده ای در مدل های خطی پیدا کرده است و محاسبه آن با نرم افزارهای معروف آماری مانند SPSS و SAS کمک زیادی به محققان می کند. در مدل حذف موردی، آماره کوک بدون نیاز به برآورد دوباره مدل به ازای مشاهدات حذف شده محاسبه می شود. زارع و راسخ (۲۰۱۱) تحلیل حساسیت در مدل های آمیخته خطی با خطا در اندازه گیری را بررسی کرده اند. در این مقاله تاثیر تحلیل حساسیت برای برآوردگرهای ماکسیمم درستنمایی تاوانیده (MPLE) بر پایه حذف موردی و حذف آزمودنی بررسی می شود. هدف ما محاسبه مستقیم معیار حساسیت برای اثرات ثابت پارامتری و برآورد تابع ناپارامتری است که مشابه با آماره کوک و آماره DFFITS بلسلی و همکاران (۱۹۸۰) در مدل رگرسیون خطی است. در این مقاله جهت تمایز بین مفاهیم مدل رگرسیون خطی با مدل آمیخته نیمه پارامتری از DFIT به جای DFFITS استفاده می شود. آماره کوک (۱۹۷۷) به عنوان معیار حساسیت در برآورد پارامتر خطی و DFIT برای اندازه گیری تغییرات یک برازش ناپارامتری کاربرد دارد. هم چنین می توان از آن دو برای شناسایی موارد یا آزمودنی

[۲] پروانه منصوری : p.mansoori@znu.ac.ir

های موثر برای MPLE در مدل های آمیخته نیمه پارامتری استفاده کرد.

۲ مدل و برآورد پارامترها

آزمایشی با m آزمودنی و n_i مشاهده طوری که $i = ۱, \ldots, m$ را در نظر بگیرید. Y_{ij} متغیر پاسخ i امین آزمودنی در زمان t_{ij} به صورت زیر مدل بندی می شود:

$$Y_{ij} = \mathbf{X}_{ij}^T\beta + f(t_{ij}) + \mathbf{Z}_{ij}^T\mathbf{b}_i + U_i(t_{ij}) + \varepsilon_{ij} \qquad (۱.۲)$$

که β یک بردار $p \times ۱$ از ضرایب رگرسیونی، $\mathbf{X}_{ij}$ ماتریس متغیرهای توصیفی، $f(t)$ تابع هموار، b_i بردار مستقل $q_i \times ۱$ اثرات تصادفی با ماتریس متغیرهای تصادفی $\mathbf{Z}_{ij}$، $\mathbf{U_i}(t)$ فرایند تصادفی مستقل و ε_{ij} بردار خطاهای تصادفی می باشد. بردارهای ویژه آزمودنی را به این صورت نامگذاری می کنیم: $\mathbf{Y_i} = (Y_{i۱}, \ldots, Y_{in_i})^T$ ، به همین ترتیب $\mathbf{X}_i$ ، $\mathbf{U}_i$ ، $\mathbf{Z}_i$ و $\varepsilon_\mathbf{i}$ نیز به صورت مشابه تعریف می شوند. فرض کنید $t^\circ = (t_۱^\circ, \ldots, t_r^\circ)^T$ بردار مقادیر گسسته مرتب شده نقاط زمانی t_{ij} ، $(i = ۱, \ldots, m, j = ۱, \ldots, n_i)$ ، $\mathbf{N_i}$ ماتریس رخداد $n_i \times r$ آزمودنی i ام باشد، طوری که $t_{ij} = t_l^\circ$ اگر عنصر (j,l)ام آن برابر ۱ باشد. متغیرهای ε،Y،X،N،U و b بردارهایی هستند که از جمع کردن m بردار ویژه آزمودنی روی هم به دست می آیند. در تمام مقاله ماتریس ها و بردارها به صورت پررنگ نمایش داده می شوند. بنابراین مدل ۱.۲ را به صورت ماتریس و بردارب صورت زیر بازنویسی می کنیم:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{N}\mathbf{f} + \mathbf{Z}\mathbf{b} + \mathbf{U} + \varepsilon \qquad (۲.۲)$$

خطاها (ε) دارای توزیع نرمال $N(۰, \sigma^۲)$ ، اثر تصادفی $\mathbf{b} \sim N(۰, \mathbf{D})$ که $D = diag(\mathbf{D}_۱, \ldots, \mathbf{D_m})$ و $\mathbf{U}$ دارای توزیع $N(۰, \Gamma)$ که $\Gamma = diag(\Gamma_۱, \ldots, \Gamma_m)$ است. ماتریس کواریانس $\mathbf{Y}$برابر است با: $cov(\mathbf{y}) = \mathbf{V} = diag(\mathbf{V}_۱, \ldots, \mathbf{V_m})$ طوری که $\mathbf{V_i} = \mathbf{Z_i}\mathbf{D_i}\mathbf{Z_i^T} + \sigma^۲\mathbf{I_i} + \mathbf{\Gamma_i}$.

تابع درستنمایی (β, f) را در زیر در نظر بگیرید:

$$l(\beta, \mathbf{f}; \mathbf{Y}) = \frac{-۱}{۲}log|\mathbf{V}| - \frac{۱}{۲}(\mathbf{Y} - \mathbf{X}\beta - \mathbf{N}\mathbf{f})^\mathbf{T}\mathbf{V}^{-۱}(\mathbf{Y} - \mathbf{X}\beta - \mathbf{N}\mathbf{f})$$

با توجه به تابع فوق، با ماکسیمم کردن تابع زیر برآوردگرهای ماکسیمم درستنمایی تاوانیده حاصل می شوند:

$$\mathbf{L}(\beta, \mathbf{f}) = l(\beta, \mathbf{f}; \mathbf{Y}) - \frac{\lambda}{۲}\int f^{"}(t)^۲dt = l(\beta, \mathbf{f}; \mathbf{Y}) - \frac{\lambda}{۲}\mathbf{f^T}\mathbf{K}\mathbf{f} \qquad (۳.۲)$$

λ پارامتر هموارسازی و $\mathbf{K}$ ماتریس هموارساز متناهی نامنفی می باشد. در این مقاله فرض می شود $\mathbf{V}$و λ پارامترهای معلوم هستند؛ در این صورت رابطه ۳.۲ یک تابع درجه دوم است و MPLE از رابطه خطی زیر به دست می آید:

$$\mathbf{C}\begin{pmatrix}\beta\\ \mathbf{f}\end{pmatrix} = \begin{pmatrix}\mathbf{X^T}\mathbf{V}^{-۱}\\ \mathbf{N^T}\mathbf{V}^{-۱}\end{pmatrix}\mathbf{Y} \qquad (۴.۲)$$

$$\mathbf{C} = \begin{pmatrix}\mathbf{X^T}\mathbf{V}^{-۱}\mathbf{X} & \mathbf{X^T}\mathbf{V}^{-۱}\mathbf{N}\\ \mathbf{N^T}\mathbf{V}^{-۱}\mathbf{X} & \mathbf{N^T}\mathbf{V}^{-۱}\mathbf{N} + \lambda\mathbf{K}\end{pmatrix}$$

در تمام مقاله فرض می شود $\mathbf{C}$ پررتبه $p+r$ است. این مطلب این اطمینان را ایجاد می کند که ماتریس$(\mathbf{X},\mathbf{NT})$ هم پر رتبه است طوری که $\mathbf{T}=(\mathbf{1},\mathbf{t}^{\circ})$ و $\mathbf{1}$ یک بردار یکه $r\times 1$ می باشد. طبق رابطه ۴.۲ برآورد پارامترها به صورت زیر می باشد:

$$\hat{\beta}=(\mathbf{X^T W_x X})^{-1}\mathbf{X^T W_x Y} \tag{۵.۲}$$

$$\hat{f}=(\mathbf{N^T W_f N}+\lambda\mathbf{K})^{-1}\mathbf{N^T W_f Y} \tag{۶.۲}$$

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

$$\mathbf{W_f}=\mathbf{V^{-1}}-\mathbf{V^{-1}X}(\mathbf{X^T V^{-1}X})^{-1}\mathbf{X^T V^{-1}}$$

هم چنین برآورد اثرات ثابت با امید ریاضی شرطی به شکل زیر بدست می آید:

$$\hat{\mathbf{b_i}}=\mathbf{D_i Z_i^T V_i^{-1}}(\mathbf{Y_i}-\mathbf{X_i}\hat{\beta}-\mathbf{N_i}\hat{\mathbf{f}}) \tag{۷.۲}$$

$$\hat{\mathbf{b}}=(\hat{\mathbf{b}}_{\mathbf{1}}^{\mathbf{T}},\ldots,\hat{\mathbf{b}}_{\mathbf{m}}^{\mathbf{T}})^T$$

با کمی محاسبه از دو رابطه ۵.۲ و ۷.۲ مقادیر برازش شده برابر $\hat{\mathbf{Y}}=\mathbf{X}\hat{\beta}+\mathbf{N}\hat{\mathbf{f}}+\mathbf{Z}\hat{\mathbf{b}}=\mathbf{HY}$ به دست می آید که ماتریس های $\mathbf{H}$ و $\bar{\mathbf{H}}$ برابر با : $\mathbf{H}=\mathbf{I}-\mathbf{\Sigma V^{-1}}+\mathbf{\Sigma V^{-1}}\bar{\mathbf{H}}$ و

$$\bar{\mathbf{H}}=\begin{pmatrix}\mathbf{X} & \mathbf{N}\end{pmatrix}\mathbf{C}^{-1}\begin{pmatrix}\mathbf{X^T}\\ \mathbf{N^T}\end{pmatrix}\mathbf{V^{-1}} \tag{۸.۲}$$

می باشند. $\mathbf{I}$ یک ماتریس همانی است و $\mathbf{\Sigma}=\sigma^{\mathbf{2}}\mathbf{I}+\mathbf{\Gamma}$. ماتریس $\mathbf{H}$ نقش ماتریس برازشی در رگرسیون خطی را ایفا می کند؛ اما در اینجا داده های نافذ اثرات ثابت وابسته به ماتریس $\bar{\mathbf{H}}$ می باشند، این وابستگی در مورد باقی مانده ها هم صدق می کند:

$$\tilde{\mathbf{e}}=\mathbf{Y}-\mathbf{X}\hat{\beta}-\mathbf{N}\hat{\mathbf{f}}=(\mathbf{I}-\bar{\mathbf{H}})\mathbf{Y} \tag{۹.۲}$$

۳ تحلیل حذف موردی

حذف موردی مشاهدات پایه ای برای ساختن آماره های تحلیل حساسیت می باشد. در مدلهای رگرسیونی خطی معمولاً از روش حذف موردی به منظور بررسی تأثیر تك مشاهدات بر روی برآورد پارامترها استفاده می شود. در این بخش تاثیر حذف یک تک مشاهده روی برآورد (β,f) بررسی می شود.

۱.۳ برآوردها تحت حذف موردی

فرض کنید $\hat{\theta}_{(\mathbf{ij})}=(\hat{\beta}_{(\mathbf{ij})}^{\mathbf{T}},\hat{\mathbf{f}}_{(\mathbf{ij})}^{\mathbf{T}})^{\mathbf{T}}$ براورد θ بدون (i,j) امین مشاهده باشد. محاسبه $\hat{\theta}_{(ij)}$ به ازای تمام (i,j) ها ومقایسه آن با $\hat{\theta}$ مخصوصا زمانی که اندازه نمونه بزرگ باشد زمان زیادی را می طلبد. با استفاده از قضیه ۱.۳ می توان برآورد پارامترها را تحت حذف مشاهده بدون بررسی کل نمونه بدست آورد. برای سادگی مشاهدات شماره گذاری می شوند؛ فرض کنید (i,j)امین مشاهده شماره $c=n_1+\cdots+n_{i-1}+j$ را به خود می گیرد. هم چنین فرض کنید d_c یک بردار $n\times 1$می باشد طوری که در مرتبه c ام مقدار

یک و در بقیه جاها مقدار صفر را خواهد داشت.

قضیه ۱.۳ : با استفاده از نمادگزاری های بالا نتیجه می شود که:

$$\hat{\beta}_{(\mathbf{ij})} = \hat{\beta} - \frac{(\mathbf{X^T W_x X})^{-1}\mathbf{X^T W_x d_c d_c^T V^{-1}\tilde{e}}}{\mathbf{d_c^T V^{-1}(I-\bar{H})d_c}} \tag{۱.۳}$$

$$\hat{\mathbf{f}}_{(\mathbf{ij})} = \hat{\mathbf{f}} - \frac{(\mathbf{N^T W_f N} + \lambda\mathbf{K})^{-1}\mathbf{N^T W_f d_c d_c^T V^{-1}\tilde{e}}}{\mathbf{d_c^T V^{-1}(I-\bar{H})d_c}} \tag{۲.۳}$$

برای داده های مستقل با $\mathbf{V} = \sigma^2\mathbf{I}$ رابطه ۱.۳ به شکل زیر تغییر می کند:

$$\hat{\beta}_{(\mathbf{c})} = \hat{\beta} - \frac{(\mathbf{X^T W_x X})^{-1}\mathbf{X^T W_x d_c \tilde{e}_c}}{1 - \mathbf{\bar{h}_{cc}}} \tag{۳.۳}$$

$\tilde{e}_c$ مولفه c ام e در ۹.۲ و $\bar{h}_{cc}$ ، c ا مین عنصر قطری $\bar{H}$ در ۸.۲ می باشد. اگر جزء ناپارامتری یعنی f در مدل نباشد ۳.۳ همان فرمول کوک (۱۹۷۷) در رگرسیون خطی می باشد.

۲.۳ آماره کوک و DFIT

آماره کوک تعمیم یافته به صورت مربع آماره مقیاس شده $\hat{\theta}$ از θ به شکل زیر تعریف می شود:

$$\mathbf{CD_{ij}}(\beta, \mathbf{f}) = (\hat{\theta} - \hat{\theta}_{(\mathbf{ij})})^{\mathbf{T}}\mathbf{C}(\hat{\theta} - \hat{\theta}_{(\mathbf{ij})})$$

واضح است که هرچقدر این فاصله بزرگتر باشد احتمال اینکه مشاهده (i,j)ام یک مشاهده موثر باشد بیشتر است. این آماره را می توان با جایگذاری از قضیه ۱.۳ به صورت مستقیم به دست آورد:

$$\mathbf{CD_{ij}}(\beta, \mathbf{f}) = \mathbf{d_c^T V^{-1}\bar{H}d_c}\left\{\frac{\mathbf{d_c^T V^{-1}\tilde{e}}}{\mathbf{d_c^T V^{-1}(I-\bar{H})d_c}}\right\}^2 = \frac{\mathbf{d_c^T V^{-1}\bar{H}d_c}}{\mathbf{d_c^T V^{-1}(I-\bar{H})d_c}} t_c^2$$

$$t_c = \frac{\mathbf{d_c^T V^{-1}\tilde{e}}}{\sqrt{\{\mathbf{d_c^T V^{-1}(I-\bar{H})d_c}\}}}$$

t_c باقی مانده استیودنت شده مورد c ام می باشد. در مدل های نیمه پارامتری معمولا تاثیر جزئی روی برآورد β, f را به صورت مجزا بررسی می کنند. بنابراین آماره کوک برای جزء پارامتری مدل یعنی β به شکل زیر به دست خواهد آمد:

$$\mathbf{CD_{ij}}(\beta) = (\hat{\beta} - \hat{\beta}_{(\mathbf{ij})})^{\mathbf{T}}\{(\mathbf{I_p}, 0)\mathbf{C}^{-1}(\mathbf{I_p}, 0)^{\mathbf{T}}\}^{-1}(\hat{\beta} - \hat{\beta}_{(\mathbf{ij})})$$

به بیانی ساده تر و با استفاده از قضیه ۱.۳ نتیجه می شود که:

$$\mathbf{CD_{ij}}(\beta) = \frac{\mathbf{d_c^T W_x X(X^T W_x X)^{-1}X^T W_x d_c}}{\mathbf{d_c^T V^{-1}(I-\bar{H})d_c}} t_c^2$$

همان طور که مشاهده می شود تاثیر جزئی موردcام روی $\hat{\beta}$ به مقادیر نافذ یا به اندازه t_c بستگی دارد. مقادیرنافذ جزء ناپارامتری، عناصر قطری $\mathbf{H}_\beta = \mathbf{W_x X(X^T W_x X)^{-1}X^T W_x}$ می باشد. پیرو کوک (۱۹۷۷) جهت معناداری مشاهدات موثر این معیار را با مقادیر توزیع $\chi^2_{\mathbf{p},\alpha}$ می سنجند.

$$(\hat{\beta} - \beta)^{\mathbf{T}}\{(\mathbf{I_p}, 0)\mathbf{C}^{-1}(\mathbf{I_p}, 0)^{\mathbf{T}}\}^{-1}(\hat{\beta} - \beta) \leqslant \chi^2_{p,\alpha}$$

که یک بیضی اطمینان در سطح $\alpha - ۱$ است. از کوک (۱۹۷۷) علاقه مند به $\hat{\beta}_{(ij)}$ های هستیم که در ناحیه اطمینان ۱۰درصد باشد. از آن جایی که ممکن است اثر مشاهدات به دلیل هم پوشانی در یک همسایگی موضعی بر روی برآوردهای اسپلاین، معیار حساسیت را کاهش دهد؛ می توان از اثر جزئی با استفاده از رابطه ۴.۳ استفاده کرد.

$$\mathbf{DFIT_{ij}} = \frac{\mid \mathbf{d_c^T N}(\hat{\mathbf{f}}_{(ij)} - \hat{\mathbf{f}}) \mid}{\mathbf{s_{ij}}} \qquad (۴.۳)$$

طوری که $\mathbf{s_{ij}^۲}$ عنصر قطری c ام $\mathbf{N}(۰, \mathbf{I_r})\mathbf{C}^{-۱}(۰, \mathbf{I_r})^{\mathbf{T}}\mathbf{N^T}$ می باشد.

۳.۳ مدل انتقال میانگین نقاط دورافتاده و آزمون نقاط دورافتاده

از مدل انتقال میانگین نقاط دورافتاده اغلب برای آزمون اینکه آیا مشاهده مورد بررسی یک نقطه دورافتاده برای مدل محسوب می شود استفاده می شود. به ازای مورد (i, j) ام این مدل به فرم زیر می باشد:

$$\mathbf{Y_{kl}} = \mathbf{X_{kl}^T}\beta + f(t_{kl}) + \mathbf{Z_{kl}^T}b_i + U_k(t_{kl}) + \varepsilon_{kl} \qquad k \neq i, l \neq j,$$

$$\mathbf{Y_{ij}} = \mathbf{X_{ij}^T}\beta + f(t_{ij}) + \mathbf{Z_{ij}^T}b_i + U_i(t_{ij}) + \gamma + \varepsilon_{ij} \qquad (۵.۳)$$

که γ یک پارامتر اضافی برای نشان دادن وجود نقطه دورافتاده می باشد. بنابراین آزمون نقاط دورافتاده را می توان به صورت آزمون فرض $\gamma = ۰$ بیان کرد. فرض کنید $\tilde{\beta}_{\mathbf{ij}}$ و $\tilde{\mathbf{f}}_{\mathbf{ij}}$ و $\tilde{\gamma}_{\mathbf{ij}}$ ؛ MPLE مدل ۵.۳ باشد. قضیه زیر برقرار است:

قضیه ۳ .۳ : برآوردگرهای ماکسیمم درستنمایی تحت مدل ۲.۲ برابر برآوردگرهای ماکسیمم درستنمایی مدل ۵.۳ با مجموعه تمام مشاهدات می باشد. یعنی :

$$\hat{\mathbf{f}}_{(\mathbf{ij})} = \tilde{\mathbf{f}}_{\mathbf{ij}} \qquad \hat{\beta}_{(\mathbf{ij})} = \tilde{\beta}_{\mathbf{ij}}$$

معادل بودن مدل حذف موردی و مدل نقاط دورافتاده در رگرسیون خطی را در کوک و ویزبرگ (۱۹۸۲) می توان دید. این مساله توسط وی (۱۹۹۸)در کلاس گسترده تری از مدل های پارامتری بسط داده شده است. قضیه ۳.۳ برابری این دو مدل را در مدل های آمیخته نیمه پارامتری تائید می کند. هم چنین MPLE برای γ برابر است با:

$$\tilde{\gamma}_{\mathbf{ij}} = (\mathbf{d_c^T V^{-۱} d_c})^{-۱}\mathbf{d_c^T V^{-۱}}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{(\mathbf{ij})} - \mathbf{N}\hat{\mathbf{f}}_{\mathbf{ij}}) = \frac{\mathbf{d_c^T V^{-۱}\tilde{e}}}{\mathbf{d_c^T V^{-۱}(I - \bar{H})d_c}}$$

و واریانس آن تحت فرض صفر برابر است با:

$$var(\tilde{\gamma}_{\mathbf{ij}}) = \frac{\mathbf{d_c^T V^{-۱}(I - \bar{H})V(I - \bar{H})V^{-۱}d_c}}{\{\mathbf{d_c^T V^{-۱}(I - \bar{H})d_c}\}^{۲}}$$

که مجانبا برابر با واریانس بیزی آن یعنی:

$$var_{\mathbf{B}}(\tilde{\gamma}_{\mathbf{ij}}) = \frac{۱}{\mathbf{d_c^T V^{-۱}(I - \bar{H})d_c}}$$

می باشد. مقدار استاندارد شده $\tilde{\gamma}_{\mathbf{ij}}$ با واریانس بیزی آن برابر با باقی مانده های استیودنت شده می باشد و از آن می توان برای بررسی این که آیا مشاهده (i, j) ام در مدل یک نقطه پرت محسوب می شود استفاده کرد. همانطور که در مدل رگرسیون خطی ساده از باقی مانده های مقیاس شده برای شناسایی نقاط دورافتاده استفاده می شود؛ در این مدل از نمودار t_c که نمودار باقی مانده های استیودنت شده می باشد برای غربالگری نقاط دور افتاده می توان استفاده نمود.

۴ تحلیل حذف آزمودنی

در مطالعات طولی مشاهدات آزمودنی های مشابه اغلب مقادیر مشابهی در متغیرهای کمکی می گیرند. بنابراین شناسایی آزمودنی های موثر در مقابل مشاهدات موثر می تواند مفید واقع شود. در این بخش تاثیر حذف آزمودنی i ام روی $\hat{\beta}$ و $\hat{\mathbf{f}}$ در مدل بررسی می شود. فرض کنید $\hat{\theta}_{[\mathbf{i}]} = (\hat{\beta}^{\mathbf{T}}_{[\mathbf{i}]}, \hat{\mathbf{f}}^{\mathbf{T}}_{[\mathbf{i}]})^{\mathbf{T}}$ برآورد β و f با حذف آزمودنی i ام باشد. قضیه زیر فرمولی برای برآورد ارائه می کند.

قضیه ۴ : فرض کنید $\mathbf{E_i} = (۰, \cdots, ۰, \mathbf{I_{ni}}, ۰, \cdots, ۰)^{\mathbf{T}}$ یک ماتریس $n \times n_i$ باشد. بنابراین:

$$\hat{\theta}_{[\mathbf{i}]} = \hat{\theta} - \mathbf{C}^{-۱} \begin{pmatrix} X^T \\ N^T \end{pmatrix} \mathbf{V}^{-۱}\mathbf{E_i}(\mathbf{I_{n_i}} - \bar{\mathbf{H}}_{\mathbf{i}})^{-۱}\mathbf{E_i^T}\tilde{\mathbf{e}}$$

طوری که

$$\bar{\mathbf{H}}_{\mathbf{i}} = \begin{pmatrix} \mathbf{X_i} & \mathbf{N_i} \end{pmatrix} \mathbf{C}^{-۱} \begin{pmatrix} \mathbf{X_i^T} \\ \mathbf{N_i^T} \end{pmatrix} \mathbf{V_i}^{-۱}$$

آماره کوک تعمیم یافته تحت مدل حذف آزمودنی به شکل زیر تعریف می شود:

$$\mathbf{CD}_{[\mathbf{i}]}(\beta, \mathbf{f}) = (\hat{\theta}_{[\mathbf{i}]} - \theta)^{\mathbf{T}}\mathbf{C}(\hat{\theta}_{[\mathbf{i}]} - \theta)$$

با استفاده از قضیه ۴ فرمول را می توان بازنویسی کرد:

$$\mathbf{CD}_{[\mathbf{i}]}(\beta, \mathbf{f}) = \tilde{\mathbf{e}}^{\mathbf{T}}\mathbf{E_i}(\mathbf{I_{n_i}} - \bar{\mathbf{H}}_{\mathbf{i}}^{\mathbf{T}})^{-۱}\mathbf{V_i}^{-۱}\bar{\mathbf{H}}_{\mathbf{i}}(\mathbf{I_{n_i}} - \bar{\mathbf{H}}_{\mathbf{i}})^{-۱}\mathbf{E_i^T}\tilde{\mathbf{e}}$$

که $\mathbf{E_i^T}\tilde{\mathbf{e}}$ بردار باقی مانده $n_i \times ۱$ متناظر با آزمودنی i ام می باشد. همانند بخش قبل برای شناسایی تاثیر آزمودنی i ام روی برآورد β باید $\mathbf{CD}_{[\mathbf{i}]}(\beta)$ را با توزیع $\chi^{۲}_{p,\alpha}$ ارزیابی کرد. از آنجایی که حذف آزمودنی i ام شامل n_i نقطه زمانی می باشد؛ به محاسبه $DFIT$ در n_i نقطه زمانی برای شناسایی تاثیر موضعی روی منحنی برازش شده f نیاز هست. هم چنین از مدل انتقال میانگین نقاط دورافتاده برای i امین مورد می توان استفاده کرد:

$$Y_i = \mathbf{X_i}\beta + \mathbf{N_i}\mathbf{f} + \mathbf{Z_i}\mathbf{b_i} + \mathbf{\Delta} + \mathbf{U_i} + \varepsilon_{\mathbf{i}}$$

$$Y_j = \mathbf{X_j}\beta + \mathbf{N_j}\mathbf{f} + \mathbf{Z_j}\mathbf{b_j} + \mathbf{U_j} + \varepsilon_{\mathbf{j}} \qquad j \neq i \tag{۱.۴}$$

که در آن $\mathbf{\Delta}$ یک بردار $n_i \times ۱$ و نشان دهنده آزمودنی های دورافتاده می باشد. فرض کنید $\tilde{\beta}_{[\mathbf{i}]}$ ، $\tilde{\mathbf{f}}_{[\mathbf{i}]}$ و $\tilde{\mathbf{\Delta}}_{[\mathbf{i}]}$ نمایانگر MPLE برای β ، $\mathbf{\Delta}$ و f مدل ۱.۴ باشد. مشابه قضیه ۳.۳ می توانیم نشان دهیم:

$$\tilde{\beta}_{[\mathbf{i}]} = \hat{\beta}_{[\mathbf{i}]} \qquad \tilde{\mathbf{f}}_{[\mathbf{i}]} = \hat{\mathbf{f}}_{[\mathbf{i}]} \qquad \tilde{\mathbf{\Delta}}_{[\mathbf{i}]} = \mathbf{R_i}$$

۵ تحلیل داده شبیه سازی شده

برای ارزیابی معیار های تحلیل حساسیت در این مقاله داده شبیه سازی شده با دو نقطه دور افتاده را بررسی می کنیم. متغیر کمکی X_i اندازه های غیر نرمال BMI از ۲۰ آزمودنی اول داده های پروژسترون سوئرز وهمکاران (۱۹۹۸) می باشد. مقادیر آن از ۲۱,۵۳۲۸ تا

۳۸,۰۱۶۵ مرتب شده اند. مدل به صورت زیر است:

$$Y_{ij} = ۰/۵(X_i - ۲۶) + f(t_{ij}) + b_i + U_i(t_{ij}) + \varepsilon_{ij} \qquad i = ۱, \cdots, ۲۰ \qquad j = ۱, \cdots, ۴$$

که در آن $f(t) = (t - ۰/۵)^۲$ ، $t_{i۱} \sim U(۰, ۰/۲۵)$ ، $t_{i,j+۱} = t_{ij} + ۰/۲۵ \quad (j = ۲, ۳, ۴)$ ، $b_i \sim N(۰, ۰/۱۲)$، $U_i(t)$ فرآیند اتورگرسیو مرتبه اول با واریانس ۰,۲۲ و خودهمبستگی مرتبه اول ۰,۳ و $\varepsilon_{ij} \sim N(۰, ۰/۱)$ می باشد. آزمودنی ها از هم مستقل هستند. متغیر پاسخ Y_{ij} از مقدار ۲,۸- تا ۶,۹ مرتب شده است. الگوریتم MPLE این نتایج را به دنبال دارد: $\hat{\beta} = ۰/۴۵۲$ و $\lambda = ۰$.شکل ۱ ، $\mathbf{CD}(\beta)$ و $\mathbf{DFIT}$ برای برآورد f را نشان می دهد. مورد ۲۰ اثر زیادی روی برآورد β و مورد ۶۰ اثر زیادی روی برآورد f دارد. از آنجایی که مورد ۶۰ نقطه نفوذ ضعیفی محسوب می شود اثر آن روی برآورد β خیلی بزرگ نیست. با توجه به بحث حذف مشاهدات، همان طوری که در شکل ۲ مشاهده می کنیم آزمودنی پنجم تاثیر زیادی روی برآورد β دارد. هم چنین با توجه به شکل ۲ آزمودنی ۵ و ۱۵ به عنوان نقاط دورافتاده شناسایی شده اند. در این مثال دونقطه دورافتاده با اندازه های مباحث تشخیصی که بحث شد شناسایی شدند. این نتایج کاملا رضایت بخش می باشد.

بحث و نتیجه گیری

در این مقاله مدل های حذف موردی و حذف آزمودنی را در مدل های آمیخته نیمه پارامتری بررسی کردیم که بسیار با اهمیت هستند و نقش به سزایی در تحلیل داده ها دارند .

مراجع

Belsley D. A. , Kuh, E. and Welsh, R. E. (1980) *Regression Diagnostics*, Wiley, New York.

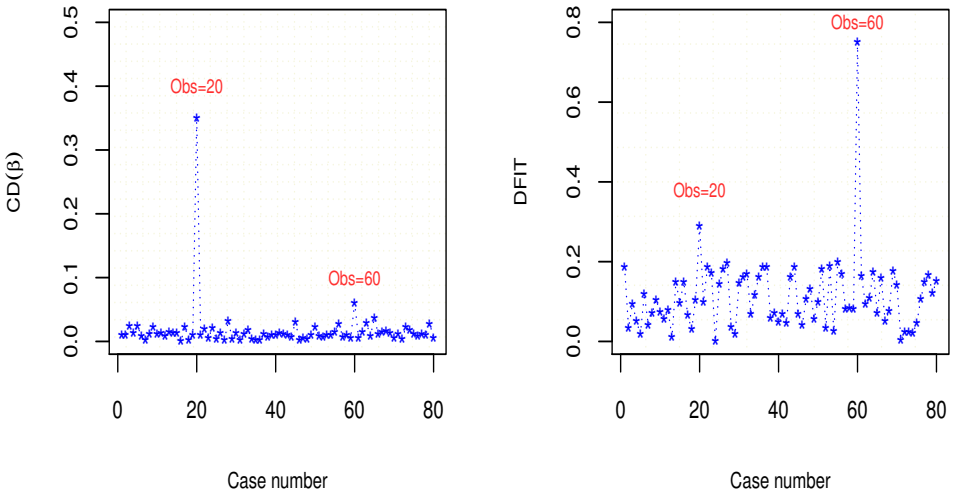
Cook R. D. (1977)Detection of influential observations in linear regresion. *Technometrics*19, 15-18.

Cook R. D. and Weisberg S. (1982) , *Residual and influence in Regression*, Chapman & Hall, New York.

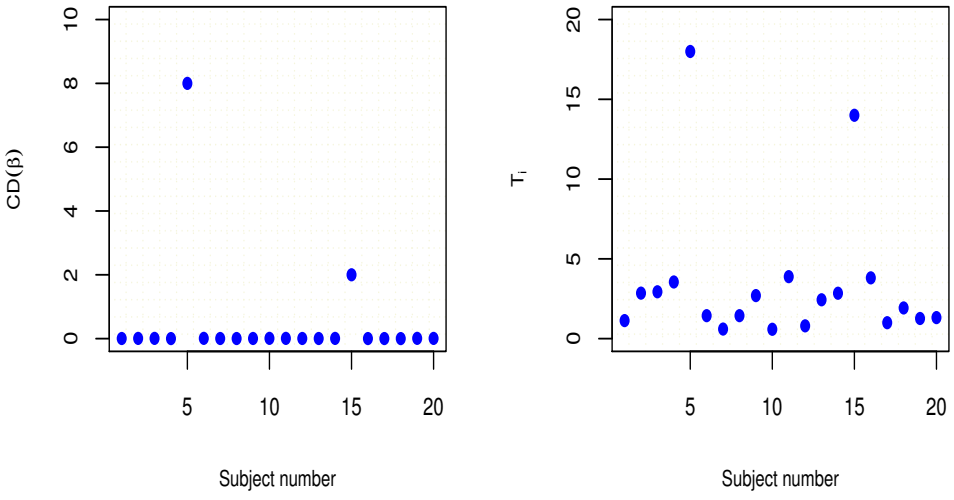
Sowers, M. F., Crutchfield, M., Randolph, J. F., Shiapo,B., Zhang, B. , Pietra, M.L., Schork, M.A. (1998),Urinaryovarian and gonadotrophin hormone levels in premeopausal women with low bone mass.*J.Bone Min.statist.Ress*.,13, 1191-1202.

Wei B. C. (1998) *Exponential Family Nonlinear Models*, Springer, Singapore.

Zare K. and Rasekh A. (2011) ,Diagnostics measure for linear mixed measurment error models. *Sort35*.,13, 125-144.



شکل ۱: تاثیر حذف موردی روی داده های شبیه سازی شده



شکل ۲: تاثیر حذف آزمودنی روی داده های شبیه سازی شده