روش‌های شناسایی نقاط تأثیرگذار برای روش کمترین توان‌های دوم

منیره معنوی[۱]، مهدی روزبه[۲]

تاریخ دریافت: ۱۳۹۹/۰۱/۱۷
تاریخ پذیرش: ۱۳۹۹/۱۰/۰۴

چکیده:
روش کمترین توان‌های دوم برای برآورد ضرایب رگرسیونی مدل‌های خطی روشی بسیار ساده، کاربردی و مفید است. این روش آماری توسط کاربران رشته‌های مختلف به سبب ارائه بهترین برآوردگر خطی نااریب با کمترین واریانس مورداستفاده قرار می‌گیرد. متأسفانه این روش در شرایطی که مشاهده (مشاهدات) دورافتاده در مجموعه داده حضورداشته باشند، خروجی قابل‌اطمینانی نخواهد داشت، زیرا نقطه فروریزش (معیار استواری برآوردگر) این روش ۰٪ است. به همین سبب شناسایی این مشاهدات امری حائز اهمیت است. تاکنون روش‌های مختلفی برای شناسایی این مشاهدات پیشنهادشده است. در این مقاله به‌مرور و بحث در مورد جزئیات روش‌های معرفی‌شده پرداخته می‌شود. در انتها با ارائه یک مثال شبیه‌سازی به بررسی هر یک از روش‌های معرفی‌شده می‌پردازیم.

واژه‌های کلیدی: کمترین توان‌های دوم، نقطه اهرمی، نقطه دورافتاده، شناسایی نقاط دورافتاده.

۱ مقدمه

بدون شک رگرسیون یکی از مهم‌ترین شاخه‌های علم آمار است که اهدافی نظیر درک چگونگی روابط بین متغیرها و پیش‌بینی را دنبال می‌کند. امروزه رگرسیون در تمام شاخه‌های علوم نظیر اقتصاد، پزشکی، مهندسی، فیزیک، مدیریت، بازرگانی، هواشناسی، بیولوژی، علوم زیستی و علوم اجتماعی جایگاه خود را یافته و مورد استقبال فراوان افراد قرارگرفته است.

اولین گام در تجزیه‌وتحلیل رگرسیون خطی برآورد ضرایب رگرسیونی می‌باشد. اولین روشی که بدین منظور معرفی شد روش کمترین قدرمطلق انحراف‌ها[۳] بود که در سال ۱۷۵۷ توسط روگر-ژوزف پسکویچ ابداع شد. سی سال پس از او پیرسیمون لاپلاس این روش را مورداستفاده قرار داد [۱]. ولی این روش در برابر مزایای بسیار خوب روش کمترین توان‌های دوم تاب نیاورد و منسوخ شد. کارل فردریک گاوس [۱۶] روش کمترین توان‌های دوم برای برآورد ضرایب رگرسیونی معرفی نمود. به اعتقاد او یک برآورد خوب، برآوردی است که خطا را کمینه نماید و بدین‌سان از تابع زیان توان دوم خطا استفاده نمود. وی با انتشار قضیه گاوس مارکوف در سال ۱۸۲۳ گام نهایی را به‌منظور تکمیل این روش برداشت [۱۷]. این روش، روشی بسیار رایج، ساده و کاربردی است که در صورت برقراری پذیره‌های کلاسیک طبق قضیه گاوس مارکوف[۴] برآوردگر کمترین توان‌های دوم منجر به بهترین برآوردگر خطی نااریب با کمترین واریانس می‌شود. این روش محبوب در کنار سادگی محاسبات و تمامی محاسنش در برخی شرایط نظیر حضور هم خطی[۵] و نقاط دورافتاده[۶] در مجموعه داده‌ها و همچنین در حضور داده‌های بعد بالا عملکرد بسیار ضعیفی داشته و نتایج آن بسیار گمراه‌کننده و دور از واقعیت است.

در این قسمت لازم است که به معرفی مدل رگرسیونی بپردازیم. مدل رگرسیونی خطی چندگانه به‌صورت

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (۱)$$

است، به‌طوری‌که در آن $\mathbf{y} = (y_1, \ldots, y_n)^\top$، بردار متغیر پاسخ، $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)_{n\times p}$ ماتریس مشاهدات متغیرهای توضیحی با $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top$, $j = 1, \ldots, p$، بردار $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ پارامترها و $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ بردار خطا است.

[۱] دانش‌آموخته کارشناسی ارشد آمار، دانشگاه سمنان، سمنان، ایران.
[۲] هیئت‌علمی گروه آمار، دانشگاه سمنان، سمنان، ایران. (نویسنده مسئول: mahdi.roozbeh@semnan.ac.ir)

[3] Least absolute deviation
[4] Gauss–Markov theorem
[5] Collinearity
[6] Outliers

نقاط دورافتاده و تأثیرگذار یکی از مسائل چالش‌برانگیزی است که کشمکش‌های فراوانی را در علم آمار به وجود آورده است. مردم عادی در مواجهه با این نقاط به‌ندرت از اصطلاح دورافتاده استفاده می‌کنند. در عوض از مواردی همانند:

- این عدد بسیار بزرگ است، احتمالاً اشتباهی رخ‌داده است،
- مقدار زیاد درآمد ماه گذشته تنها از روی شانس و اقبال است و به آن اهمیتی نمی‌دهم،
- مقدارهای بسیار کوچک به دلیل خرابی دستگاه اندازه‌گیری حاصل‌شده است و اندازه‌گیری‌ها دوباره باید تکرار شوند و یا اندازه‌ی بسیار کوچک را نادیده می‌گیرم،
- زمانی که وزن افراد نمونه‌ی موردبررسی بین ۵۰ تا ۹۰ کیلوگرم است، چرا وزن فردی ۲۲۰ کیلوگرم است؟ شاید دستگاه اندازه‌گیری خراب بوده است و یا این عدد به‌درستی ثبت‌نشده است،

استفاده می‌کنند و در بسیاری موارد آن را نادیده می‌گیرند. اما یک آماردان به‌آسانی از این الفاظ استفاده نمی‌کند و از دیدگاه او ممکن است، درآمد زیاد و یا وزن زیاد و یا مقدار بسیار کوچک واقعی باشد. در حقیقت نقاط دورافتاده در نظر اغلب افراد، به‌عنوان خطا محسوب شده اما ممکن است که اطلاعات زیادی را به دوش کشیده باشد. یک کارشناس آمار در برخورد با این موارد، به‌کارگیری آزمونی را برای شناسایی نقاط دورافتاده پیشنهاد کرده و پس از شناسایی نقاط به دنبال علل وقوع این موارد و راهکارهایی برای مقابله با آن‌ها می‌گردد. زیرا حضور این نقاط در مدل منجر به تعیین اشتباه مدل، برآوردهای اریب برای پارامترها و نتایج اشتباه خواهند شد ([۲۴]، [۴۱]). نقاط دورافتاده ممکن است به دلایلی اعم از اخلال در فرآیند استخراج، خطای ثبت، خطاهای انسانی و یا موارد غیرطبیعی واقعی رخ دهند [۳]. با ظهور و پیدایش نقاط دورافتاده و با توجه به اهمیت آن محققان بررسی‌های خود را روی این موضوع شروع و هرکدام تعاریف و راهکارهای متفاوتی برای شناسایی و مقابله با آن ارائه دادند. برخی از این تعاریف عبارت‌اند از

- یک نقطه دورافتاده، مشاهده‌ای است که به‌قدری از سایر مشاهدات منحرف‌شده تا تردیدهایی را به وجود آورد که آن مشاهده توسط مکانیسم دیگری ایجادشده است [۱۹].
- نقطه دورافتاده به‌عنوان مشاهده‌ای تعریف می‌شود که مقدار آن هماهنگ با الگوی تولیدشده به‌وسیله سایر داده‌ها نیست [۵].
- نقاطی که از انبوه داده‌ها فاصله دارد و از طرح کلی داده‌ها تبعیت نمی‌کند نقاط دورافتاده نامیده می‌شوند. این نقاط ممکن است، هم در متغیر توضیحی (نقاط اهرمی بد) و هم در متغیر پاسخ رخ دهند [۳۴]. تورکان و همکاران [۳۸] نیز به همین نکته اشاره‌کرده و به‌طورکلی نقاط دورافتاده را به دو نوع نقاط X دورافتاده[7] و نقاط y دورافتاده[8] تقسیم نمودند.
- یک نقطه دورافتاده، مشاهده (یا مجموعه‌ای از مشاهدات) است که به نظر می‌رسد دارای مانده‌ای متفاوت با مانده‌ی سایر نقاط مجموعه داده است [۴].
- یک نقطه دورافتاده مشاهده‌ای است که خارج از الگوی کلی توزیع است [۲۶].
- بلسی و همکارانش [۷] یک مشاهده تأثیرگذار را یکی از مواردی که به‌صورت جداگانه و یا همراه با چندین مشاهده دیگر تأثیر بزرگ و قابل‌توجهی روی مقادیر محاسبه‌شده از برآوردهای مختلف (ضرایب، خطای استاندارد آماره t و ...) نسبت به سایر مشاهدات داراست، تعریف نمودند.
- داس و گوگی [۱۱] یک مشاهده تأثیرگذار را مشاهده‌ای تعریف نمودند که پس از حذف آن تغییر چشمگیری در مدل ایجاد شود. نقطه دورافتاده، یک مشاهده تأثیرگذار است اما عکس آن لزوماً درست نیست [۲۵].

روسو و ون دریزن [۳۳] اظهار داشتند که به‌طورکلی یک مجموعه داده می‌تواند چهار نوع متفاوت از نقاط را درون خود جای دهد که شامل مشاهدات عادی[9]، نقاط دورافتاده عمودی[10]، نقاط اهرمی خوب[11] و نقاط اهرمی بد[12] است. در ادامه به تعاریف این نوع نقاط پرداخته می‌شود.

تعریف ۱.۱. *مشاهدات عادی: مشاهداتی که از طرح کلی داده‌ها تبعیت می‌کنند و حول خط برازش کمترین توان‌های دوم پراکنده‌شده و روی آن تأثیر منفی نمی‌گذارند.*

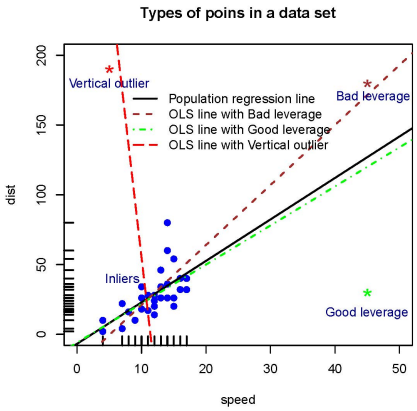
[7] X-outliers
[8] y-outliers
[9] Regular observations Or Inliers
[10] Vertical outliers
[11] Good leverage points
[12] Bad leverage points

تعریف ۲.۱. *نقاط دورافتاده عمودی: نقطه‌ی i ام $(\mathbf{x}_i, y_i)$ نقطه‌ی دورافتاده عمودی نامیده می‌شود اگر مؤلفه y آن از طرح کلی مؤلفه y سایر نقاط پیروی نکند و از توده آن‌ها دور باشد اما مؤلفه $\mathbf{x}$ آن از طرح کلی مؤلفه $\mathbf{x}$ سایر نقاط پیروی کند.*

تعریف ۳.۱. *نقطه اهرمی خوب: نقطه i ام $(\mathbf{x}_i, y_i)$ نقطه اهرمی خوب نامیده می‌شود اگر مؤلفه $\mathbf{x}$ آن از طرح کلی مؤلفه $\mathbf{x}$ سایر نقاط پیروی نکند و از آن‌ها دور باشد، اما مؤلفه y آن از طرح کلی مؤلفه‌ی y سایر نقاط پیروی کند. حضور نقطه اهرمی خوب روی خط برازش کمترین توان‌های دوم معمولی اثر بدی نمی‌گذارد.*

تعریف ۴.۱. *نقطه اهرمی بد (دورافتاده افقی): نقطه i ام $(\mathbf{x}_i, y_i)$ نقطه اهرمی بد نامیده می‌شود اگر مؤلفه $\mathbf{x}$ آن از طرح کلی مؤلفه $\mathbf{x}$ سایر نقاط پیروی نکند و از آن‌ها دور باشد و همچنین مؤلفه y آن از طرح کلی مؤلفه y سایر نقاط پیروی نکند. حضور نقطه اهرمی بد روی خط برازش کمترین توان‌های دوم معمولی اثر مخربی می‌گذارد.*

از تعاریف ۱.۱ تا ۴.۱ می‌توان به اهمیت موقعیت و جایگاه مشاهدات در فضای $\mathbf{X}$ و $\mathbf{y}$ پی برد. این نقاط در شکل شماره ۱ نشان داده‌شده است. این شکل با استفاده از مجموعه داده "Cars" موجود در نرم‌افزار R رسم شده است. ۳۰ مشاهده اول این مجموعه داده، مشاهدات عادی نمودار هستند و خط مشکی خط رگرسیونی برازش داده‌شده به روش کمترین توان‌های دوم، تنها با در نظر گرفتن این مشاهدات (نقاط آبی رنگ) است. در ادامه، نقطه اهرمی خوب، نقطه اهرمی بد و نقطه دورافتاده عمودی نیز تولید شد و خط برازش به روش کمترین توان‌های دوم با در نظر گرفتن مشاهدات عادی و هرکدام از این سه مورد مجدداً به‌صورت مجزا رسم شد. پرواضح است که نقاط اهرمی بد و نقاط دورافتاده عمودی تأثیر بسیار مخربی روی روش کمترین توان‌های دوم داشته‌اند.



شکل ۱: انواع نقاط در یک مجموعه داده و تأثیر آن‌ها روی خط برازش کمترین توان‌های دوم.

روش کمترین توان‌های دوم معمولی به‌شدت تحت تأثیر این نقاط قرار می‌گیرد. زیرا در روش کمترین توان‌های دوم معمولی مسئله بهینه‌سازی کمیت $\sum_{i=1}^{n} e_i^2$ است. به‌طوری‌که در آن e_i مانده‌های مدل است و مانده‌ها در آن با وزن یکسان به توان دوم می‌رسند. به‌عبارت‌دیگر تمام مشاهدات در این روش از اهمیت یکسانی برخوردار هستند. همین امر موجب تأثیرپذیری شدید روش کمترین توان‌های دوم از نقاط دورافتاده می‌شود.

۲ شناسایی نقاط دورافتاده و تأثیرگذار

همان‌طور که بیان شد اثربخشی و اهمیت نقاط دورافتاده و تأثیرگذار روی روش کمترین توان‌های دوم تحلیل‌گر را وادار به شناسایی این مشاهدات و تعیین اثرشان روی جنبه‌های مختلف تحلیل می‌نماید. این امر نقش اساسی در صحت نتایج، ایفا می‌نماید زیرا در صورت تشخیص اشتباه عدم حضور نقاط دورافتاده در مجموعه داده و به دنبال آن استفاده از روش‌های کلاسیک نتایج نادرست را به همراه خواهند داشت. همچنین تشخیص نادرست نقاط دورافتاده و استفاده از روش‌های استوار، ممکن است منجر به از دست رفتن اطلاعات شود،

زیرا ممکن است نقاطی که به‌عنوان نقاط دورافتاده شناسایی‌شده‌اند، در حقیقت دورافتاده نبوده بلکه سیگنال‌های حاکی از لغزش در انتخاب روش صحیح برآورد باشند. اگر بگوییم که شناسایی این نقاط یکی از مهم‌ترین شاخه‌های مباحث تشخیصی و عیب‌شناسی رگرسیون[۱۳] است، بدون شک اغراق نکرده‌ایم. ویزبرگ [۳۹] روش‌ها و شاخص‌های ارائه‌شده برای شناسایی نقاط دورافتاده را تحلیل تأثیر نام گذارد. مقصود اصلی تحلیل تأثیر، اندازه‌گیری میزان تغییرات به وجود آمده از دیدگاه‌های مختلف در شرایط وجود مجموعه‌ای از مشاهدات تأثیرگذار است. سؤال بسیار مهمی که در این قسمت مطرح می‌شود این است که مشاهدات تأثیرگذار دقیقاً روی چه چیزی تأثیر می‌گذارند و این تأثیر به چه میزان نتایج تحلیل‌گر را تخریب و تحریف می‌نماید؟ برای پاسخ به این سؤال در قدم اول می‌بایست هدف اصلی تحلیل به‌طور دقیق توسط تحلیل‌گر تعیین شود. برای مثال اگر برآورد $\boldsymbol{\beta}$ موردعلاقه تحلیل‌گر باشد اندازه‌گیری تأثیر مشاهدات روی $\hat{\boldsymbol{\beta}}$ موردنظر خواهد بود و درصورتی‌که پیش‌بینی مقصود تحلیل‌گر باشد، آنگاه بررسی تأثیر روی مقادیر برازش داده‌شده، مناسب‌تر از تأثیر روی $\hat{\boldsymbol{\beta}}$ است.

روش‌های شناسایی نقاط دورافتاده را می‌توان به دو دسته‌ی روش‌های تک‌متغیره و روش‌های چندمتغیره تقسیم نمود. این تقسیم‌بندی حتی در مدل‌های پارامتری و ناپارامتری نیز انجام‌پذیر است. در مدل‌های رگرسیونی ناپارامتری بررسی مسائل عیب‌شناسی رگرسیون بسیار پیچیده‌تر است. محققانی نظیر ایوبانک [۱۳]، سیلورمن [۳۵]، توماس [۳۷] و کیم [۲۲] مانده‌ها، اهرم‌ها و نوعی از فاصله‌ی کوک در اسپلاین‌های هموارسازی در مدل‌های ناپارامتری را مطالعه نمودند. معیارهای تأثیر بی‌شماری توسط محققان مختلف معرفی شدند. اما به‌طورکلی می‌توان شناسایی نقاط را به دو گروه روش‌های عددی و روش‌های شهودی نیز تقسیم نمود که در ادامه به بررسی روش‌های شناسایی نقاط دورافتاده و معرفی معیارهای تأثیر پرداخته می‌شود.

۳ روش‌های عددی

استفاده از آزمون‌ها و روش‌های عددی مختلف برای شناسایی نقاط دورافتاده بسیار سودمند خواهد بود. در این قسمت به معرفی برخی از این آزمون‌ها و معیارها می‌پردازیم.

۱.۳ تحلیل مانده‌ها

یکی از قدیمی‌ترین روش‌ها برای تعیین و بررسی مشکلات ایجادشده در مدل، بررسی مانده‌های کمترین توان‌های دوم است که به‌صورت زیر تعریف می‌شوند.

$$e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = ۱, \ldots, n \qquad (۲)$$

که در آن $\hat{y}_i$ مقدار برازش داده‌شده است. مشاهداتی که دارای مانده‌های بزرگ (از لحاظ قدرمطلقی) باشند، مشکوک هستند. یک ضعف بزرگ مانده‌ها ناهمواریانسی آن‌ها بوده و همین مسئله موجب سوءظن محقق نسبت به مانده‌ها و به‌کارگیری مانده‌های استیودنت‌شده است. مانده‌های استیودنت‌شده[۱۴] به‌صورت

$$r_i = \frac{e_i}{\hat{\sigma}_{\text{ols}}\sqrt{۱ - h_{ii}}}, \quad i = ۱, \ldots, n,$$

است. که در آن $\hat{\sigma}^۲_{\text{ols}} = \text{MSE}$ و h_{ii} اهرم مشاهده i ام (i امین عنصر قطری ماتریس هت که در بخش‌های آتی تعریف می‌شود) است. درصورتی‌که $|r_i| > ۱.۹۶$ یا $|r_i| > ۲$ باشد، مشاهده‌ی i ام دورافتاده است [۱۴]. النبرگ [۱۲] نشان داد که $\frac{r_i^۲}{n-p}$ دارای توزیع بتا است. بکمن و تروسل [۶] مانده استیودنت‌شده را به‌صورت

$$t_i = \frac{e_i}{\hat{\sigma}_{\text{ols}(-i)}\sqrt{۱ - h_{ii}}}, \quad i = ۱, \ldots, n,$$

معرفی نمودند. به‌طوری‌که $\hat{\sigma}_{\text{ols}(-i)}$ میانگین مجموع توان دوم مانده‌ها با حذف i امین مشاهده است. همچنین می‌توان نوشت

$$\begin{aligned}\hat{\sigma}^۲_{\text{ols}(-i)} &= \frac{y^\top_{(-i)}(\mathbf{I} - \mathbf{H}_{(-i)})y_{(-i)}}{(n-p-۱)} \\ &= \frac{(n-p)\hat{\sigma}^۲_{\text{ols}}}{n-p-۱} - \frac{e_i^۲}{(n-p-۱)(۱-h_{ii})},\end{aligned}$$

که در آن $\mathbf{H}_{(-i)}$ ماتریس هت[۱۵] (که در بخش بعدی به‌تفصیل توضیح داده می‌شود) بدون در نظر گرفتن مشاهده i ام است. هولگین و ولسچ [۲۰] این مانده‌ها را برای شناسایی نقاط دورافتاده توصیه کردند. درصورتی‌که $|t_i| > ۱.۹۶$ و یا $|t_i| > ۲$ باشد، مشاهده‌ی i ام دورافتاده است [۱۴]. بلسی و همکارانش [۷]، و بسیاری از نویسندگان دیگر مانده استیودنت‌شده با حذف مشاهدات را به مانده‌های استیودنت‌شده به دلایل زیر ترجیح می‌دادند.

- t_i ها در حقیقت آماره‌های t به‌منظور آزمون معنی‌داری ضریب i امین بردار واحد u_i در مدل نقطه دورافتاده میانگین $= \mathbf{y}$ $\mathbf{X}\boldsymbol{\beta} + u_i\delta + \boldsymbol{\varepsilon}$ است (برای مشاهده جزئیات بیشتر به [۷]

[13] Diagnostics
[14] Student residuals
[15] Hat matrix

مراجعه نمایید). [۶] بکمن و تروسل اظهار داشتند که تحت فرض نرمال این آماره دارای توزیع t - استیودنت با درجه آزادی $n-p-1$ است.

- با عملیات جبری مختصری می‌توان رابطه میان r_i و t_i را به‌صورت زیر آشکار نمود.

$$t_i = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}, \qquad (3)$$

با توجه به عبارت (۳) کاملاً واضح است که t_i یک تبدیل یکنوا از r_i است. به‌عبارت‌دیگر،

$$\lim_{t_i^2\to\infty} r_i^2 = (n-p),$$

بنابراین t_i انحرافات بزرگ را به طرز چشم‌گیری نسبت به r_i نمایش می‌دهد.

- $\hat{\sigma}^2_{(-i)}$ برای مسائلی که خطاهای بزرگی در i امین مشاهده دارد، مقاوم و استوار است.

۲.۳ تحلیل ماتریس هت

ماتریس هت نقش بسزایی در محاسبه‌ی مقادیر برازش داده‌شده و مانده‌ها و ماتریس واریانس کوواریانس آن‌ها دارد. مقدار برازش داده‌شده در رگرسیون به‌صورت

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{H}\mathbf{y},$$

محاسبه می‌شود که در آن $\mathbf{H}$ ماتریس هت نام دارد. عناصر قطری این ماتریس با h_{ii} نمایش داده می‌شود و اهرم نقطه i ام نامیده می‌شود.

ماتریس هت درواقع واریانس-کواریانس‌های $\hat{\mathbf{y}}$ و $\hat{\boldsymbol{\varepsilon}} = \mathbf{e} = \mathbf{y}-\hat{\mathbf{y}}$ را تعیین می‌کند. زیرا $\mathrm{Var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$ و $\mathrm{Var}(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})$ به همین سبب نقاط اهرمی نقش مهمی را در شناسایی مشاهدات تأثیرگذار ایفا می‌کند. اگر مشاهده i ام اهرم سنگینی (بزرگ) داشته باشد، نقطه‌ی اهرمی نامیده می‌شود. مقادیر بزرگ h_{ii} منجر به کوچک شدن $\mathrm{Var}(e_i)$ شده و منجر به برازش نزدیک‌تر $\hat{\mathbf{y}}$ به $\mathbf{y}$ خواهد شد. هولگین و ولسچ [۲۰] بررسی اهرم را برای شناسایی نقاط اهرم بالا توصیه نمودند. به‌طورکلی $\frac{1}{n} \le h_{ii} < 1$ با مقدار متوسط $\bar{h} = \frac{\mathrm{trace}(\mathbf{H})}{n} = \frac{(p+1)}{n}$ است (منظور از trace اثر ماتریس است). اگر $h_{ii} \ge 2\bar{h}$ مشاهده i ام به‌عنوان یک مشاهده اهرمی در نظر گرفته می‌شود ([۲۳]. و [۲۰]). بنابراین به‌منظور تعیین نقاط دورافتاده استفاده هم‌زمان عناصر قطری ماتریس هت و مانده‌های استیودنت‌شده یا استیودنت‌شده با حذف مشاهدات مفید است.

چاترجی و هادی [۸] ماتریس هت تقویت‌شده[16] را با فرض $\mathbf{X}^* = (\mathbf{X}:\mathbf{y})$ به‌صورت

$$\mathbf{H}^* = \mathbf{H} + \frac{\big(\mathbf{I}-\mathbf{H}\big)\mathbf{y}^\top\mathbf{y}\big(\mathbf{I}-\mathbf{H}\big)}{\mathbf{y}^\top\big(\mathbf{I}-\mathbf{H}\big)\mathbf{y}},$$

تعریف نمودند. به دلیل این‌که $\mathbf{X}^*$ به‌طور توأم دارای اطلاعاتی هم در مورد $\mathbf{X}$ و هم در مورد $\mathbf{y}$ است، معیار

$$h^*_{ii} = [\mathbf{x}_i : y_i]\begin{bmatrix}\mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{y}\\ \mathbf{y}^\top\mathbf{X} & \mathbf{y}^\top\mathbf{y}\end{bmatrix}\begin{bmatrix}\mathbf{x}_i^\top\\ y_i,\end{bmatrix} \qquad (4)$$

برای اندازه‌گیری تأثیر نقاط استفاده‌شده و اطلاعات جامع‌تری را در اختیار کاربر قرار می‌دهد.

۳.۳ پتانسیل‌ها

هادی [۱۸] معتقد بود که اگر یک نقطه با اهرم بالا در مدل حضورداشته باشد، ممکن است که ماتریس اطلاعات (ماتریس طرح) را بشکند و در نتیجه سایر مشاهدات اهرم مناسبی نداشته باشند. او یک مورد را از مشاهدات با اهرم معلوم حذف کرد و پتانسیل[17] را به‌صورت

$$\mathrm{p}_{ii} = \mathbf{x}_i\Big(\mathbf{X}^\top_{(-i)}\mathbf{X}_{(-i)}\Big)^{-1}\mathbf{x}_i^\top,$$

تعریف کرد به‌طوری‌که $\mathbf{X}_{(-i)}$ ماتریس داده با حذف سطر i ام و $\mathbf{x}_i$ سطر i ام ماتریس طرح است. رابطه این معیار با اهرم نقاط به‌صورت $p_{ii} = \frac{h_{ii}}{1-h_{ii}}$ است. مشاهدات با پتانسیل بالا به‌عنوان نقاط اهرم بالا شناخته می‌شوند [۲۱]. هادی [۱۸] استفاده از $\mathrm{mean}(p_{ii}) + c\times$ s.td.(p_{ii}) را به‌عنوان نقطه بحرانی پیشنهاد نمود. (منظور از s.td. انحراف استاندارد می‌باشد.) که در آن c مقدار ثابتی است که هادی [۱۸] مقادیر ۲ یا ۳ را برای این ثابت مناسب دانست. با توجه به این حقیقت که میانگین و انحراف استاندارد غیراستوار هستند، هادی [۱۸] به‌منظور شناسایی نقاط بسیار دور، استفاده از میانه و انحراف قدرمطلق میانه را توصیه نمود. یعنی $\mathrm{median}(\mathrm{p}_{ii}) + c\times\mathrm{MAD}(\mathrm{p}_{ii})$ که در آن

$$\mathrm{MAD}(\mathrm{p}_{ii}) = \frac{\mathrm{median}\Big\{\big|\mathrm{p}_{ii} - \mathrm{median}(\mathrm{p}_{ii})\big|\Big\}}{0.674}.$$

[16] Augment Hat matrix

[17] Potentials

۴.۳ آماره اندریز-پرگیبون

اندریز و پرگیبون برای بررسی تغییرات $|\mathbf{X}^{*\top}\mathbf{X}^{*}|$ در حضور مشاهده i ام و بدون آن استفاده از آماره اندریز-پرگیبون[18] را مفید دانستند. این آماره به‌صورت

$$\mathrm{AP}_i = \frac{|\mathbf{X}^{*\top}_{(-i)}\mathbf{X}^{*}_{(-i)}|}{|\mathbf{X}^{*\top}\mathbf{X}^{*}|},$$

تعریف می‌شود. مقادیر کوچک این آماره زنگ خطری برای تحلیل‌گر است. متأسفانه این روش بین نقاط اهرم بالا (در فضای $\mathbf{X}$) و نقطه دورافتاده (در فضای $\mathbf{y}$) تمایزی قائل نمی‌شود و همین مورد منجر به بی‌رغبتی محققان در به‌کارگیری آن و فراموشی این روش شده است [۸].

۵.۳ فاصله درستنمایی

این روش، یک روش اندازه‌گیری تأثیر نقاط مبتنی بر فاصله بین لگاریتم درستنمایی $\hat{\boldsymbol{\beta}}$ و $\hat{\boldsymbol{\beta}}_{(-i)}$ برآوردگر کمترین توان‌های دوم بدون حضور مشاهده i ام) است. [۹] فاصله درستنمایی[19] را به‌صورت

$$\begin{aligned}\mathrm{LD}_i &= ۲\Big(\mathrm{L}(\hat{\boldsymbol{\beta}} - \mathrm{L}(\hat{\boldsymbol{\beta}}_{(-i)})\Big)\\ &= n \times \log\Big(\frac{n}{n-۱}.\frac{n-p-۱}{t_i^{۲}+n-p-۱}\Big) + \frac{t_i^{۲}(n-۱)}{(۱-h_{ii})(n-p-۱)} - ۱,\end{aligned}$$

تعریف کردند. که در آن $\mathrm{L}(\hat{\boldsymbol{\beta}})$ لگاریتم درستنمایی $\hat{\boldsymbol{\beta}}$، $\mathrm{L}(\hat{\boldsymbol{\beta}}_{(-i)})$ لگاریتم درستنمایی $\hat{\boldsymbol{\beta}}_{(-i)}$، است. ناحیه اطمینان متقارن این فاصله $\Big\{\boldsymbol{\beta}: ۲(\mathrm{L}(\hat{\boldsymbol{\beta}}) - \mathrm{L}(\hat{\boldsymbol{\beta}}_{(-i)})) \le \mathfrak{X}^{۲}_{\alpha}(p+۱)|\Big\}$ است که در آن $\mathfrak{X}^{۲}$ توزیع کای‌دو با $(p+۱)$ درجه آزادی است. این روش بر اساس توزیع داده‌ها حاصل شد درصورتی‌که سایر روش‌های بیان‌شده کاملاً عددی بودند.

۶.۳ آماره نسبت کوواریانس

آماره نسبت کوواریانس[20] از تقسیم دترمینان ماتریس-واریانس کوواریانس محاسبه‌شده با حذف مشاهده i ام به ماتریس واریانس-کوواریانس محاسبه‌شده با همه مشاهدات فراهم می‌شود. یعنی

$$\mathrm{COVRATIO}_i = \frac{\Big|(\mathbf{X}^{\top}_{(-i)}\mathbf{X}_{(-i)})^{-۱}\hat{\sigma}^{۲}_{\mathrm{ols}(-i)}\Big|}{\Big|(\mathbf{X}^{\top}\mathbf{X})^{-۱}\hat{\sigma}^{۲}_{\mathrm{ols}}\Big|}, i = ۱,\ldots,n, \quad (۵)$$

به‌طوری‌که $\mathbf{X}_{(-i)}$ ماتریس طرح حاصل از حذف i امین سطر است. کوک و ویزبرگ [۹] این آماره را COVRATIO نامیدند. اگر $|\mathrm{COVRATIO}_i| > ۱$ باشد، احتساب i امین مشاهده دقت برآورد را بهبود می‌بخشد و اگر $|\mathrm{COVRATIO}_i| > ۱$ باشد، در نظر گرفتن مشاهده مربوطه دقت برآورد را تخریب می‌نماید. همچنین اگر این نسبت نزدیک به ۱ باشد، تأثیر مشاهده i ام روی ضرایب رگرسیونی کم بوده و برخلاف آن هر چه که این نسبت از ۱ دور باشد، تأثیر مشاهده i روی ضرایب رگرسیونی بزرگ و بزرگ‌تر می‌شود. مقدار بحرانی این آماره برای نمونه‌های بزرگ که به عقیده چاترجی و هادی [۸] یک درجه‌بندی سخت‌گیرانه برای مشاهدات است، به‌صورت

$$\mathrm{COVRATIO}_i > ۱ + \frac{۳p}{n} \quad \text{یا} \quad \mathrm{COVRATIO}_i < ۱ - \frac{۳p}{n},$$

است. البته برخی عقیده دارند که کران پایین تنها درصورتی‌که $n > ۳p$ باشد، مناسب است. یکی از محاسن این آماره در مقایسه با آماره‌های دیگری که در این مقاله مرور خواهند شد، این است که اطلاعاتی در مورد دقت کلی برآورد فراهم می‌آورد، زیرا $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^{۲}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-۱}$.

۷.۳ آماره کوک-ویزبرگ

کوک و ویزبرگ [۹] لگاریتم نسبت حجم بیضی‌های اطمینان $(۱-\alpha)$۱۰۰٪ در حضور مشاهده i ام و بدون آن را به‌عنوان معیار تأثیر به‌صورت

$$\begin{aligned}\mathrm{CW}_i &= \frac{۱}{۲}\log(۱-h_{ii}) + \frac{p}{۲}\log\Big(\frac{(n-p-۱)F_{\alpha}(p,n-p)}{(n-p-r_i^{۲})F_{\alpha}(p,n-p-۱)}\Big)\\ &= -\frac{۱}{۲}\mathrm{COVRATIO}_i + \frac{p}{۲}\log\Big(\frac{F_{\alpha}(p,n-p)}{F_{\alpha}(p,n-p-۱)}\Big),\end{aligned}$$

پیشنهاد دادند که در آن F_{α} چندک بالای توزیع فیشر در سطح α است. کوک و ویزبرگ [۹] بر این عقیده بودند که اگر این کمیت بزرگ و مثبت باشد، حذف i امین مشاهده، منجر به افزایش قابل‌توجه حجم خواهد شد و درصورتی‌که کمیت بزرگ و منفی باشد، منجر به کاهش قابل‌توجه حجم خواهد بود.

جدا از مقادیر ثابت (نسبت مقادیر F) این آماره معادل آماره نسبت کوواریانس خواهد بود [۸].

[18] Andrews and Pregibon
[19] Likelihood distance
[20] Covariane ratio statistics
[21] Cook's distance
[22] Ralph Dennis Cook

۸.۳ آماره کوک

آماره فاصله کوک[21] توسط رالف دنیز کوک[22] دانشمند آمریکایی به‌منظور اندازه‌گیری تأثیر هر مشاهده به‌تنهایی روی مدل برازش داده‌شده معرفی شد [۹].

در این روش فاصله بین مقدار برازش داده‌شده به روش کمترین توان‌های دوم با حضور یک نقطه و بدون آن برای تمامی n مشاهده اندازه‌گیری می‌شود. به‌طور طبیعی هر چه که این فاصله بیشتر شود نقطه موردبررسی تأثیر بیشتری روی مدل برازش داده‌شده خواهد داشت. این آماره به‌صورت

$$\mathrm{D}_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})^\top (\mathbf{X}^\top \mathbf{X})^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})}{(p+1)\hat{\sigma}^2_{\text{ols}}}, \qquad (۶)$$

است. به‌طوری‌که $\hat{\sigma}^2_{\text{ols}}$ برآورد واریانس مانده‌ها است. در عمل استفاده از (۶) بسیار دشوار است، به‌خصوص زمانی که حجم نمونه بزرگ باشد. زیرا برای محاسبه آماره فاصله کوک برای هر مشاهده می‌بایست مشاهده‌ی موردنظر حذف‌شده و آماره فاصله کوک برای آن محاسبه شود و این عمل باید به تعداد مشاهدات تکرار شود که امری بسیار زمان‌بر خواهد بود. به همین سبب می‌توان با ساده‌سازی رابطه (۶) به‌صورت

$$\mathrm{D}_i = \frac{r_i^2}{2} \cdot \frac{h_{ii}}{1 - h_{ii}}, \qquad (۷)$$

عبارت ساده‌تری یافت. راولینگس و همکاران [۲۹] مقدار $\frac{4}{n-p-1}$ را به‌عنوان مقدار بحرانی برای این آماره در رگرسیون خطی چندگانه معرفی نمود. به‌عبارت‌دیگر اگر $\mathrm{D}_i > \frac{4}{n-p-1}$ باشد، یعنی مشاهده i ام مشاهده مؤثری است. البته کوک و ویزبرگ [۹] مقدار ۱ را به‌عنوان مقدار بحرانی برای این آماره معرفی کردند. همچنین آن‌ها پیشنهاد کردند که D_i ها با توزیع فیشر مرکزی با p و $n-p$ درجه آزادی مقایسه شوند. اما تورکان و همکاران [۳۸] بر این باورند که این روش مقادیر بحرانی اغراق‌آمیز و بزرگی را نشان می‌دهد. آن‌ها استفاده از مقدار بحرانی پیشنهادشده توسط راولینگس و همکاران [۲۹] را منطقی‌تر و معقول‌تر می‌دانند. به‌طور ایدئال، مقدار بحرانی باید تفسیر پایداری را در فرآیند تحلیل رگرسیون ارائه دهد. استیونز [۳۶] به‌کارگیری فاصله کوک را پس از داشتن برخی علائم نظیر اهرم سنگین و بزرگ بودن مانده استیودنت‌شده توصیه نمود.

۹.۳ آماره تفاوت در برازش‌ها

روش تفاوت در برازش[23] توسط بلسی و همکاران [۷] معرفی شد. البته برخی آن را تحت عنوان فاصله ولسچ-کوه[24] نیز می‌شناسند. این آماره برای مشاهده i ام به‌صورت

$$\mathrm{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(-i)}}{\sqrt{\hat{\sigma}^2_{\text{ols}(-i)} h_{ii}}}, \quad i = 1, \ldots, n, \qquad (۸)$$

نیز محاسبه می‌شود. که در آن $\hat{y}_{(-i)}$ مقدار برازش داده‌شده بدون حضور مشاهده i است. به عقیده بلسی و همکاران [۷] درصورتی‌که $|\mathrm{DFFITS}_i| > \sqrt{\frac{p}{n}}$ باشد، مشاهده i ام تأثیرگذار است. همچنین برخی محققین بر این باورند که اگر قدرمطلق آماره تفاوت در برازش‌ها بین ۱ تا ۲ و بزرگ‌تر از ۲ باشد مشاهده مربوطه تأثیرگذار است. البته تعداد انگشت‌شماری از افراد نیز مقدار $|\mathrm{DFFITS}_i| > 2\frac{\sqrt{(p+1)}}{n}$ و برخی $|\mathrm{DFFITS}_i| > 2\frac{\sqrt{(p+1)}}{n-p-1}$ را مرزهای تعیین‌کننده نقاط بحرانی برای آماره تفاوت در برازش‌ها می‌دانند [۱۱]. به سبب ارتباط تنگاتنگی که میان این آماره و آماره کوک وجود دارد، نتایج این دو آماره مشابه خواهد بود.

۱۰.۳ فاصله ولسچ

فاصله ولسچ[25] توسط ولسچ [۴۰] به‌عنوان ابزاری برای عیب‌شناسی رگرسیون و شناسایی نقاط تأثیرگذار برای نمونه‌هایی با حجم بزرگ‌تر از ۱۵ معرفی شد. فاصله ولسچ به‌صورت

$$\mathrm{W}_i^2 = (n-1) t_i^2 \frac{h_{ii}}{(1-h_{ii})^2}, \qquad (۹)$$

محاسبه می‌شود. ولسچ [۴۰] مقدار $3\sqrt{p}$ را به‌عنوان نقطه بحرانی این فاصله معرفی نمود.

۱۱.۳ آماره کوک اصلاح‌شده

آماره کوک اصلاح‌شده[26] با تلاش‌های آتکینسون معرفی شد. به همین جهت به آماره آتکینسون[27] نیز مشهور است. این آماره با جایگزینی σ^2 با $\sigma^2_{(-i)}$ و $\frac{n-p-1}{p-1}$ به‌جای $\frac{1}{p+1}$ و در نظر گرفتن توان $\frac{1}{2}$ در آماره کوک به دست می‌آید. به‌عبارت‌دیگر،

$$\mathrm{A}_i = \Big(\frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})^\top (\mathbf{X}^\top \mathbf{X})^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})}{(p+1)\hat{\sigma}^2_{\text{ols}(-i)}} \cdot \frac{n-p-1}{p} \Big)^{\frac{1}{2}}, \qquad (۱۰)$$

[23] Difference in Fits(DFFITS)
[24] Welsch-Kuh Distance
[25] Welsch's distance
[26] Modified Cook's square distance
[27] Atkinson

که با ساده‌سازی رابطه (۱۰) می‌توان آن را به‌صورت

$$\begin{aligned} \mathrm{A}_i &= \sqrt{\frac{h_{ii}}{۱-h_{ii}}}.\left|\frac{e_i}{\hat{\sigma}_{\mathrm{ols}(-i)}\sqrt{۱-h_{ii}}}\right| \\ &= |t_i|\sqrt{\frac{h_{ii}}{۱-h_{ii}}}.\Big(\frac{n-p}{p}\Big)^{\frac{۱}{۲}} = \mathrm{DFFITS}_i.\Big(\frac{n-p-۱}{p}\Big)^{\frac{۱}{۲}}, \end{aligned}$$

نوشت. همچنین می‌توان رابطه (۹) را به‌صورت

$$\mathrm{A}_i = \sqrt{\mathrm{D}_i \frac{(n-p-۱)\hat{\sigma}^{۲}_{\mathrm{ols}}}{\hat{\sigma}^{۲}_{\mathrm{ols}(-i)}}},$$

نیز نوشت. مقدار بحرانی برای این آماره ۱ است. یعنی اگر $\mathrm{A}_i > ۱$ باشد، مشاهده i ام تأثیرگذار است.

۱۲.۳ آماره هادی

این آماره به‌منظور شناسایی تأثیر پنهانی هر مشاهده به‌صورت

$$\mathrm{H}^{۲}_i = \frac{p}{۱-h_{ii}}\frac{d^{۲}_i}{۱-d^{۲}_i} + \frac{h_{ii}}{۱-h_{ii}}, \quad i = ۱,\ldots,n,$$

تعریف‌شده است به‌طوری‌که $d_i = \frac{e^{۲}_i}{\mathbf{e}^\top\mathbf{e}}$ توان دوم i امین مانده استانداردشده است. هادی [۱۸] استفاده از رابطه $\mathrm{mean}(\mathrm{H}^{۲}_i) + c\sqrt{\mathrm{Var}(\mathrm{H}^{۲}_i)}$ را به‌عنوان نقطه بحرانی برای این آماره توصیه نمود که در آن c مقدار ثابتی است که مقادیر مناسب نظیر ۲ یا ۳ برای آن انتخاب می‌شود. از سوی دیگر هادی [۱۸] به سبب استوار نبودن میانگین و انحراف استاندارد، استفاده از $\mathrm{median}(\mathrm{H}^{۲}_i)+c\mathrm{MAD}(\mathrm{H}^{۲}_i)$ را به‌عنوان نقطه بحرانی برای آماره هادی سفارش می‌کند به‌طوری‌که

$$\mathrm{MAD}(\mathrm{H}^{۲}_i) = \frac{\mathrm{median}\Big\{\Big|\mathrm{H}^{۲}_i - \mathrm{median}(\mathrm{H}^{۲}_i)\Big|\Big\}}{۰/۶۷۴},$$

است.

۱۳.۳ آماره پنا

آماره پنا[۲۸] تأثیر مشاهده i ام را با استفاده از سایر مشاهدات اندازه‌گیری می‌نماید. پنا [۲۸] پیشنهاد کرد که به‌جای بررسی تأثیر کلی حذف مشاهده i ام روی تمام عناصر بردار $\hat{\mathbf{y}}$ تأثیر حذف این مشاهده روی هر نقطه نمونه بر بردار $\hat{\mathbf{y}}$ به‌طور جداگانه بررسی شود. آماره پنا به‌صورت

$$\mathrm{S}_i = \frac{a_i^\top a_i}{(p+۱)\widehat{\mathrm{Var}}(\hat{y}_i)}, \qquad (۱۱)$$

تعریف می‌شود که در آن $\mathbf{a}_i = \Big(\hat{y}_i - \hat{y}_{i(-۱)}, \ldots \hat{y}_i - \hat{y}_{i(-n)},\Big)^\top$ و $\widehat{\mathrm{Var}}(\hat{y}_i) = \hat{\sigma}^{۲}_{ols}h_{ii}$ است. این آماره ویژگی‌های زیر را داراست [۱۱]:

- در صورت عدم وجود نقطه دورافتاده در داده و مقادیر با اهرم کوچک، $\mathrm{E}(\mathrm{S}_i) = \frac{۱}{(p+۱)}$،
- S_i به‌طور تقریبی دارای توزیع نرمال است،
- در حضور نقاط دورافتاده با اهرم بالا و یکسان، حساسیت آماره پنا برای شناسایی آن‌ها نسبت به شرایط عادی کمتر خواهد شد.

۱۴.۳ تأثیر جزئی

معیارهایی که تاکنون موردبحث قرارگرفت، با فرض این‌که تمام ضرایب رگرسیونی سهم یکسانی در مدل دارند، عمل می‌کنند که همین امر ممکن است منجر به گمراهی روش از مسیر یافتن پاسخ صحیح شود. بدین سبب اطلاعات مربوط به یک ضریب رگرسیونی موردعلاقه است. زیرا یک مشاهده می‌تواند تنها در یک بعد و یا حتی چندین بعد تأثیرگذار باشد. در این بخش معیارهایی که تأثیر مشاهدات را روی تک‌تک ضرایب رگرسیونی بررسی می‌کنند، معرفی می‌شود.

۱.۱۴.۳ تأثیر یک مشاهده روی یک ضریب تنها

تأثیر i امین مشاهده روی j امین ضریب را می‌توان به‌صورت

$$\mathrm{D}_{i,j} = \frac{t^{۲}_i(h_{ii} - h_{ii(-j)})}{۱-h_{ii}}, \qquad (۱۲)$$

یافت به‌طوری‌که در آن $h_{ii(-j)}$ اهرم مشاهده i ام حاصل از برازش مدل زمانی است که متغیر j ام از مدل حذف می‌شود. همچنین می‌توان عبارت (۱۲) را با انجام محاسباتی به‌صورت

$$\mathrm{D}_{i,j} = \frac{r^{۲}_i}{۱-h_{ii}}\frac{\omega_{ij}}{W_j^\top W_j},$$

بیان نمود که در آن ω_{ij} ، i امین عنصر بردار مانده زمانی که متغیر j ام $(\mathbf{x}_j)$ روی $\mathbf{X}_{(-j)}$ (تمامی متغیرها به‌جز متغیر j ام) رگرسیون شده باشند. یعنی i امین عنصر $W_j = (\mathbf{I} - \mathbf{H}_{(-j)})\mathbf{x}_j$ است.

۲.۱۴.۳ آماره تفاوت در بتاها

آماره تفاوت در بتاها[۲۹] به‌منظور محاسبه تغییرات مشاهده‌شده در پارامترها با توجه به معادله رگرسیون جدید ایجادشده بعد از حذف i امین مشاهده از مجموعه داده‌ها استفاده می‌شود. در حقیقت این آماره نشان‌دهنده میزان تغییر تنها یک ضریب رگرسیون با حذف یک مشاهده است که تأثیر آن مشاهده را روی هر ضریب رگرسیون با استفاده از فاصله استانداردشده میان برآورد پارامتر با حضور مشاهده مربوطه و بدون آن

[28] Peña's statistic

[29] Difference in Betas (DFBETAS)

می‌یابد. آماره تفاوت در پارامترها توسط بلسی و همکاران [۷] پیشنهاد شد. این آماره به‌صورت

$$\text{DFBETAS}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\sqrt{\hat{\sigma}^2_{\text{ols}(-i)} C_{jj}}}, j = 1, \ldots, p \ , i = 1, \ldots, n, \quad (۱۳)$$

تعریف می‌شود. به‌طوری‌که C_{jj}، j امین عنصر قطری ماتریس $(\mathbf{X}^\top \mathbf{X})^{-1}$، $\hat{\beta}_j$ برآورد پارامتر j ام با حضور تمام مشاهدات، $\hat{\beta}_{j(-i)}$ برآورد پارامتر j ام بدون حضور مشاهده i ام است. لازم به ذکر است که مخرج رابطه (۱۳) برآورد واریانس ضریب رگرسیونی برآورد شده بدون حضور i امین مشاهده است. واضح است که اگر مقدار این آماره بزرگ شود یعنی مشاهده موردنظر تأثیر زیادی روی برازش مدل خواهد داشت. مقدار بحرانی برای این آماره $\frac{2}{\sqrt{n}}$ است، یعنی اگر $|\text{DFBETAS}_{j,i}| > \frac{2}{\sqrt{n}}$ باشد، آنگاه i امین مشاهده تأثیرگذار است [۷]. البته تعداد انگشت‌شماری از محققان مقدار بزرگ‌تر از ۲ را برای مقدار بحرانی پیشنهاد می‌نمایند [۲۳]. همچنین کازینی و چارتیر [۲۷] اعتقاد دارند که مقدار بحرانی برای این آماره درصورتی‌که حجم نمونه کمتر از ۳۰ باشد، ۱ و در غیر این صورت $\frac{2}{\sqrt{n}}$ است، زیرا مقدار بحرانی توصیه‌شده تابعی از حجم نمونه است و مقادیر بحرانی (به‌ویژه برای حجم‌های کوچک) ممکن است تمایل به شناسایی نقاط بیشتری از مشاهدات نسبت به تحلیل‌گر داشته باشد. بدین‌سان وقتی‌که حجم نمونه کوچک است استفاده توأم این آماره در کنار مباحث تشخیصی دیداری پیشنهاد می‌شود. یک نقص بزرگ این آماره این است که به تعداد پارامترها و مشاهدات آماره تولید می‌شود که منجر به تولید مجموعه‌ی عظیمی از نتایج برای ارزیابی خواهد شد که تا حدی گیج‌کننده است ([۳۰]، [۳۱] و [۳۲]).

۳.۱۴.۳ اهرم جزئی

برای محاسبه تغییر در i امین عنصر قطری ماتریس هت، زمانی که متغیر j ام به مدل رگرسیونی اضافه (یا حذف) می‌شود. به‌کارگیری اهرم جزئی[30] به‌صورت

$$h^P_{ii} = h_{ii(-j)} + \delta^2_{ij}$$

مفید است. که در آن $\delta^2_{ij} = \omega^2_{ij}/W_j^\top W_j$ نشان‌دهنده سهم متغیر j ام در اهرم i ام است. که در آن ω_{ij}، i امین عنصر بردار مانده زمانی که متغیر j ام $(\mathbf{x}_j)$ روی $\mathbf{X}_{(-j)}$ (تمامی متغیرها به‌جز متغیر j ام) رگرسیون شده باشند. یعنی i امین عنصر $W_j = (\mathbf{I} - \mathbf{H}_{(-j)})\mathbf{x}_j$ است.

در این قسمت بیان یک نکته بسیار حائز اهمیت است. به‌طور طبیعی پس از حذف نقطه (نقاط تأثیرگذار) انتظار ایجاد تفاوت چشمگیری در نتایج را خواهیم داشت. هر یک از آماره‌های معرفی‌شده مقادیر بحرانی متفاوتی دارند که نقاط تأثیرگذار در مجموعه داده را شناسایی می‌نمایند. در بعضی مواقع به‌کارگیری تنها یک آماره به‌منظور فراهم نمودن اطلاعات لازم راجع به نقاط تأثیرگذار کافی است. اما غالباً لازم است بیشتر از یک آماره موردبررسی قرار گیرد. دلیل این امر این است که مقادیر بحرانی یا تابعی از حجم نمونه و یا تعداد متغیرهای توضیحی و یا هر دو هستند. حتی در شرایطی اعتبار مقادیر بحرانی مورداستفاده، در هاله‌ای از ابهام قرار می‌گیرد و تنها در صورت وجود بعضی شرایط، این مقادیر بحرانی قابل‌اطمینان خواهند بود. به همین سبب در استفاده از نقاط بحرانی بسیار باید محتاط بود. از سوی دیگر این روش‌ها به‌گونه‌ای طراحی نشده‌اند که آزمون رسمی یک فرضیه باشند، بلکه برای شناسایی مشاهداتی که روی نتایج رگرسیون بیشتر از سایر مشاهدات در مجموعه داده‌ها اثر می‌گذارند، طراحی‌شده‌اند. بنابراین مقادیر تشخیص داده‌شده باید با یکدیگر مقایسه شوند. بهترین راه برای مقایسه استفاده هم‌زمان روش‌های عددی و نمایش‌های گرافیکی نظیر نمودار ساقه برگ[31]، نمودار شاخص[32] و نمودار P-R[33] است [۱۸].

۴ روش‌های شهودی

روش‌های شهودی به‌منظور قدرت بالای سیستم بصری انسان در شناسایی الگوها ظریف و پیچیده بسیار سودمند خواهند بود و نگاه جامع و کاملی در کسری از ثانیه برای فرد فراهم می‌آورند. علاوه بر آن نمودارهای متنوع و شکیل موجب جذب توجه و علاقه خواننده شده و مطالب درک شده تا زمان زیادی در ذهن او باقی می‌ماند. نمودارها به‌سادگی و در زمان کوتاهی روابط بسیار پیچیده و مشکل را آشکار ساخته و بدین‌سان باعث صرفه‌جویی در زمان خواهند شد. بنابراین روش‌های شهودی بسیار محبوب بوده و در صورت معتبر بودن روش مورداستفاده، قابل‌اطمینان هم خواهند بود. بدین‌سان استفاده از روش‌های شهودی در کنار آزمون‌ها و روش‌های عددی بسیار راهگشا

[30] Partial leverage
[31] Stem-and-leave display
[32] Index plot
[33] P-R plot

خواهد بود. در تحلیل‌های تک‌متغیره می‌توان به روش‌هایی نظیر نمودار جعبه‌ای دومتغیره[۳۴] ، پوسته محدب [۳۵] نمودارهای عیب‌شناسی رگرسیون، نمودار مانده‌های استیودنت شده و ... اشاره نمود.

۱.۴ نمودار متغیر افزوده

نمودار متغیر افزوده[۳۶] نیز یکی از روش‌های شهودی و دیداری است که برای تعیین روابط خطی بین متغیرها و شناسایی تأثیرها به‌وفور مورداستفاده قرار می‌گیرد.

با فرض این‌که هدف برازش مدل

$$\mathbf{y} = \mathbf{X}_{(-j)}\boldsymbol{\beta}^{\star} + \theta_j \mathbf{x}_j + \boldsymbol{\varepsilon} \qquad (۱۴)$$

باشد به‌طوری‌که $\boldsymbol{\beta}^{\star}$ بردار ضرایب رگرسیونی بدون حضور ضریب رگرسیونی مربوط به متغیر j ام، $\mathbf{X}_{(-j)}$ ماتریس طرح بدون ستون j ام و θ_j ضریب رگرسیونی مربوط به متغیر j ام است. عبارت $(\mathbf{I}-\mathbf{H}_{(-j)})$ را در طرفین مدل (۱۴) را ضرب نموده با توجه به این‌که

$$(\mathbf{I} - \mathbf{H}_{(-j)})\mathbf{X}_{(-j)} = \circ$$

نتیجه می‌شود

$$(\mathbf{I} - \mathbf{H}_{(-j)})\mathbf{y} = (\mathbf{I} - \mathbf{H}_{(-j)})\mathbf{x}_j\theta_j + (\mathbf{I} - \mathbf{H}_{(-j)})\boldsymbol{\varepsilon}$$

با بازنویسی رابطه (۱۴) می‌توان نوشت

$$R_j = W_j\theta_j + \varepsilon^{*} \qquad (۱۵)$$

به‌طوری‌که $R_j = (\mathbf{I} - \mathbf{H}_{(-j)})\mathbf{y}$ و $W_j = (\mathbf{I} - \mathbf{H}_{(-j)})\mathbf{x}_j$ است. به‌بیان‌دیگر، R_j و W_j بردار مانده‌ها زمانی که $\mathbf{y}$ و $\mathbf{x}_j$ به ترتیب روی $\mathbf{X}_{(-j)}$ رگرسیون شده باشند، هستند. از سوی دیگر با توجه به اینکه $E(R_j) = W_j\theta_j$، رسم نمودار R_j در برابر W_j توسط موستلر و توکی [۲۷] پیشنهاد شد.

ویژگی‌های این نمودار عبارت است از [۸]:

- این نمودار ممکن است به‌صورت خطی مستقیم از مبدأ با شیب $\hat{\theta}_j$ ظاهر شود،
- مانده‌های مدل رگرسیون چندگانه (۱۴) معادل با مانده‌های مدل رگرسیون خطی ساده (۱۵) هستند،
- پراکنش نقاط به‌صورت بصری نشان خواهد داد که کدام‌یک از نقاط در تعیین $\hat{\theta}_j$ بیشترین تأثیر را دارند.

بلسی و همکاران [۷] این نمودار را نمودار اهرم رگرسیون جزئی نامیدند اما کوک و ویزبرگ [۹] این نمودار را با عنوان متغیر افزوده معرفی نمودند.

یکی از کاربردهای بسیار دلپذیر و خوشایند این نمودار، استفاده از آن به‌منظور شناسایی ارتباطات غیرخطی میان متغیرهای پاسخ و توضیحی است که لزوم به‌کارگیری روش‌های ناپارامتری و نیمه پارامتری را نشان می‌دهد [۳۰].

۵ مطالعه شبیه‌سازی

در این بخش در یک مطالعه شبیه‌سازی به بررسی و کاربرد آماره‌ها و روش‌های بیان‌شده می‌پردازیم. در این مطالعه بردار پارامترها به‌صورت $\boldsymbol{\beta} = (-۲, ۳, ۵, -۳, ۴)^{\top}$ در نظر گرفته‌شده است. برای تولید ماتریس طرح از

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{I}), \quad \boldsymbol{\mu} = (\circ, \ldots, \circ)^{\top}_{p\times ۱}$$

استفاده شد. همچنین به‌منظور ایجاد نقاط دورافتاده در مجموعه داده‌ها، درصدی از خطاها از توزیع $-t$ استیودنت غیرمرکزی و مابقی از توزیع نرمال استاندارد تولید می‌شود، یعنی

$$\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_۱^{\top}, \boldsymbol{\varepsilon}_۲^{\top})^{\top}, \quad \boldsymbol{\varepsilon}_۱ \sim N(\circ, \mathbf{I}), \quad \boldsymbol{\varepsilon}_۲ \sim t_۲(۸),$$

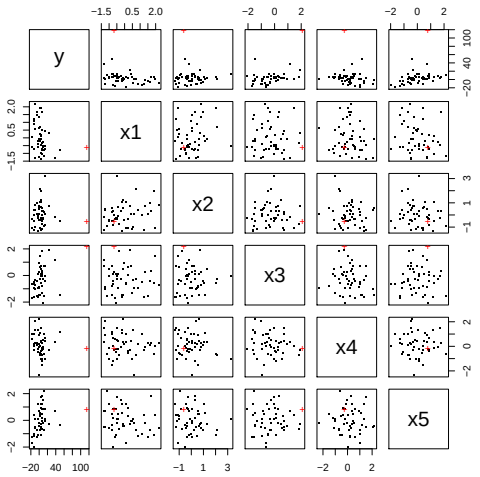
به‌طوری‌که $t_\nu(\delta)$ نشان‌دهنده توزیع $-t$ استیودنت غیر مرکزی با ν درجه آزادی و پارامتر غیرمرکزی δ است. در نهایت متغیر پاسخ با $n = ۵\circ$ مشاهده از مدل $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ تولید می‌شود. در ابتدا به‌منظور بررسی کلی نحوه پراکندگی نقاط دورافتاده نمودار پراکنش داده‌ها در شکل ۲ نشان داده‌شده است. نقاطی که با " + " نشان داده‌شده‌اند، نقاط دورافتاده هستند. انواع نقاط در مجموعه داده شبیه‌سازی‌شده در شکل ۳ مشهود است. مشاهده ۱۴ نقطه اهرمی و مشاهدات ۴۹ و ۴۰ دورافتاده هستند. نتایج تمامی روش‌های عددی مطرح‌شده در مقاله بر روی نمودار با استفاده از بسته نرم‌افزاری "ggplot2" در نرم‌افزار R رسم شده‌اند. نمودار شکل ۴ نیز همانند نمودار ۳ است منتها میزان تأثیر هر مشاهده روی تحلیل با دایره‌هایی پیرامون هر مشاهده به نمایش درآمده است. به‌منظور بررسی مانده‌ها، نمودار مانده‌ها، مانده‌های استانداردشده و مانده‌های استیودنت‌شده در شکل ۵ رسم شده است. مشاهدات ۴۰ و ۴۹ در این نمودارها به‌عنوان نقاط دورافتاده شناسایی‌شده‌اند. نمودار مانده‌های استیودنت‌شده در برابر چندک‌های توزیع t -استیودنت در

[34] Bivariate boxplot
[35] Convex hull
[36] Add variable plot

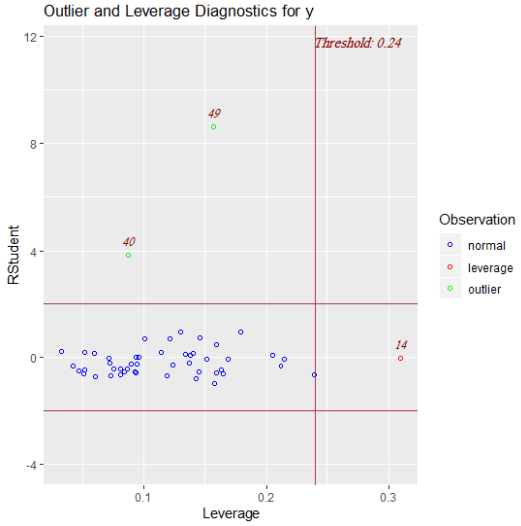
شکل ۶ قابل‌مشاهده است. در شکل ۷ نمودار مربوط به نتایج ماتریس هت و ماتریس هت تقویت‌شده رسم شده است. مشاهده ۱۴ در نمودار اول و مشاهدات ۴۹، ۴۰ و ۱۴ در نمودار دوم مشکوک شناسایی‌شده است. با توجه به نمودار شکل ۸ واضح است که آماره اندرز-پرگیبون مشاهدات ۱۴، ۴۰ و ۴۹ را مشکوک اعلام می‌کند. نتایج آماره پتانسیل در شکل ۹ قابل‌مشاهده است. فاصله درستنمایی که در شکل ۱۰ مشهود است، نقاط ۴۰ و ۴۹ را تأثیرگذار می‌شناسد. نتایج آماره نسبت کوواریانس با استفاده از دو آماره بیان‌شده در شکل ۱۱ نمایش داده‌شده است. آماره کوک-ویزبرگ و آماره کوک نیز مشابه روش درستنمایی مشاهدات ۴۰ و ۴۹ را به‌عنوان مشاهدات مشکوک می‌شناسند، نتایج این دو آماره در شکل‌های ۱۲ و ۱۳ خلاصه‌شده است. نتایج آماره تفاوت در برازش‌ها به ازای سه مقدار بهینه معرفی‌شده در متن در شکل ۱۴، فاصله ولسچ در شکل ۱۵، آماره کوک اصلاح‌شده در شکل ۱۶، آماره هادی در شکل ۱۷ و نمودار نتایج آماره پنا در شکل ۱۸ قابل‌مشاهده است. همچنین تأثیر هر مشاهده روی یک ضریب تنها در شکل ۱۹، آماره تفاوت در پارامترها در شکل ۲۰، و اهرم جزئی در شکل ۲۱ نشان داده‌شده است.

بدون اغراق می‌توان گفت که نمودار جعبه‌ای یکی از ساده‌ترین و کاربردی‌ترین روش‌ها برای کشف و شناسایی نقاط دورافتاده هست، نمودار جعبه‌ای مجموعه داده موردبررسی در شکل ۲۲ مشهود است. نقاط دورافتاده و تأثیرگذار بارنگ قرمز مشخص‌شده‌اند. نمودار متغیر افزوده (از روش‌های شهودی) در شکل ۲۳ نمایش داده‌شده است. یکی دیگر از نمودارهای مفید برای شناسایی نقاط دورافتاده نمودار Cell Map است. این نمودار در شکل ۲۴ قابل‌مشاهده است. در این نمودار به ازای هر مشاهده یک سطر و به ازای هر متغیر یک ستون رسم می‌کند. سلول‌های زرد رنگ نشان‌دهنده مشاهدات عادی، سلول‌های قرمز رنگ نشان‌دهنده نقاط دورافتاده، مشاهدات نارنجی نشان‌دهنده نقاط اهرمی و مشاهدات سفید رنگ نشان‌دهنده مشاهدات گمشده است.

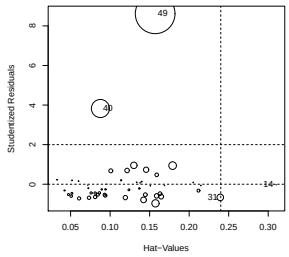
همان‌طورکه بیان شد به سبب وابستگی مقدار بحرانی آماره‌ها به n، p یا هر دو استفاده از تنها یک آماره به‌منظور شناسایی نقاط دورافتاده کار درستی نیست و باید به‌طور هم‌زمان از چند آماره استفاده نمود. البته استفاده از آماره‌ها به همراه روش‌های شهودی بسیار مفید خواهد بود. با توجه به شکل‌ها کاملاً روشن است که روش‌های مختلف جواب‌های متفاوتی داشتند اما تقریباً تمامی روش‌ها (عددی و شهودی) عجیب‌ترین مشاهده، یعنی مشاهده ۴۹ را شناسایی نمودند.



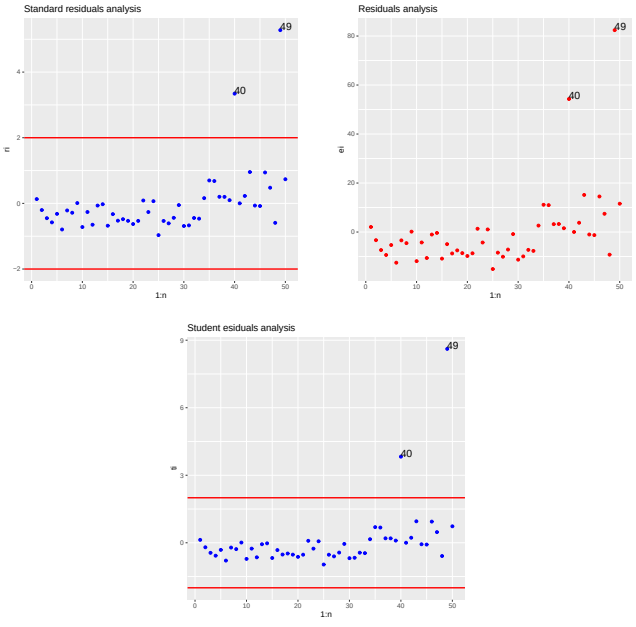
شکل ۲: نمودار پراکنش.



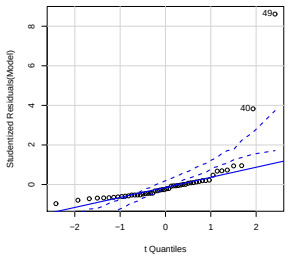
شکل ۳: انواع نقاط در مجموعه داده‌ها.



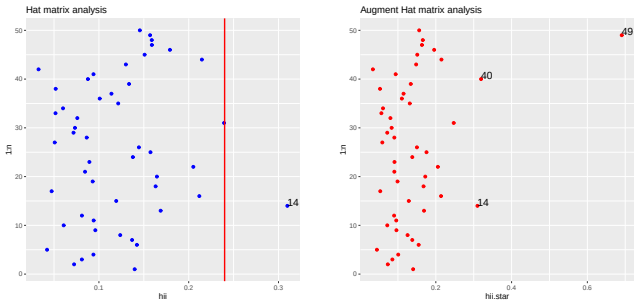
شکل ۴: نمودار مانده‌های استیودنت‌شده در برابر اهرم نقاط.



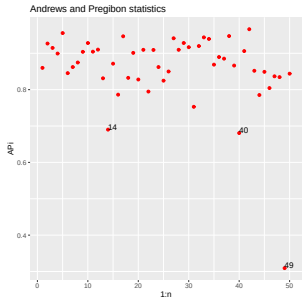
شکل ۵: نمودار مانده‌ها.



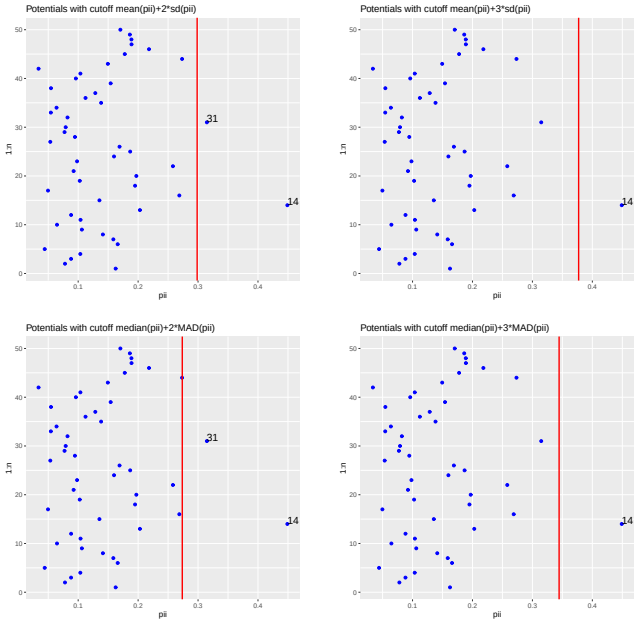
شکل ۶: نمودار مانده‌های استیودنت‌شده در برابر اهرم نقاط.



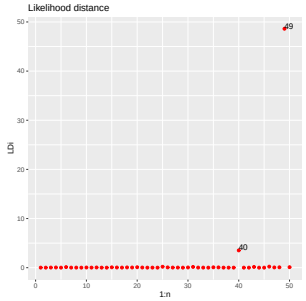
شکل ۷: نمودار نتایج ماتریس هت و ماتریس هت تقویت‌شده.



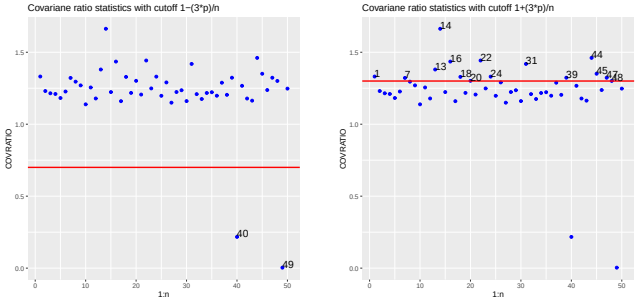
شکل ۸: نمودار نتایج آماره اندزیر و پرگیبون.



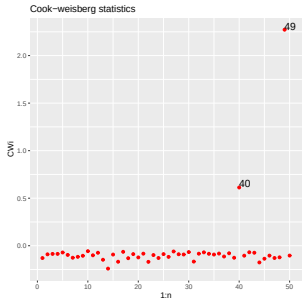
شکل ۹: نمودار نتایج پتانسیل.



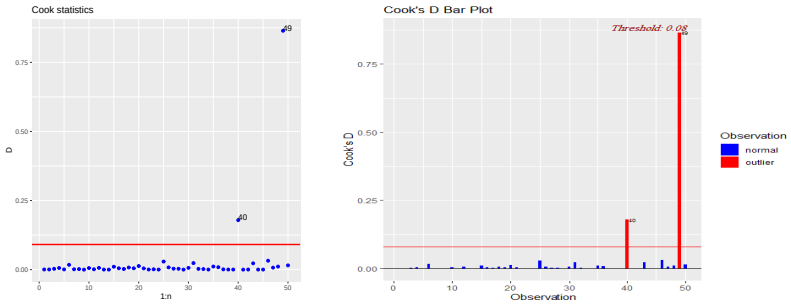
شکل ۱۰: نمودار نتایج فاصله درستنمایی.



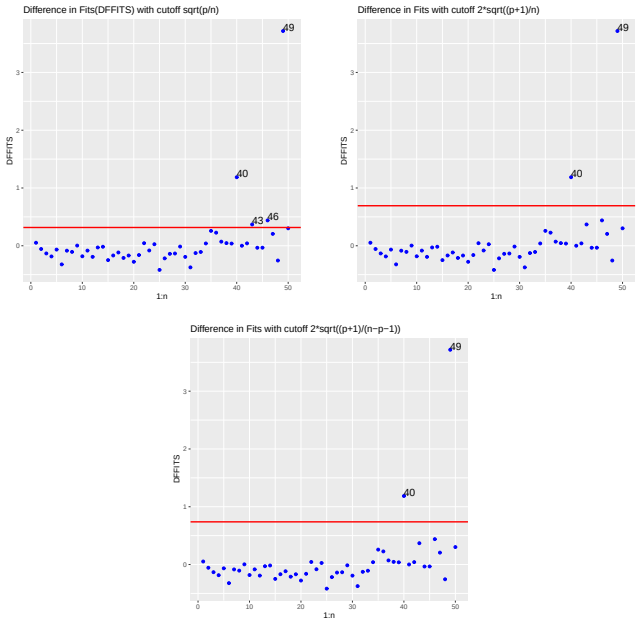
شکل ۱۱: نتایج آماره نسبت کوواریانس.



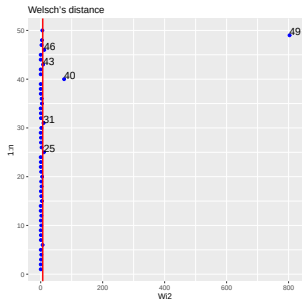
شکل ۱۲: نتایج آماره کوک-ویزبرگ.



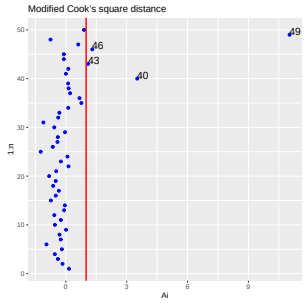
شکل ۱۳: نتایج آماره کوک.



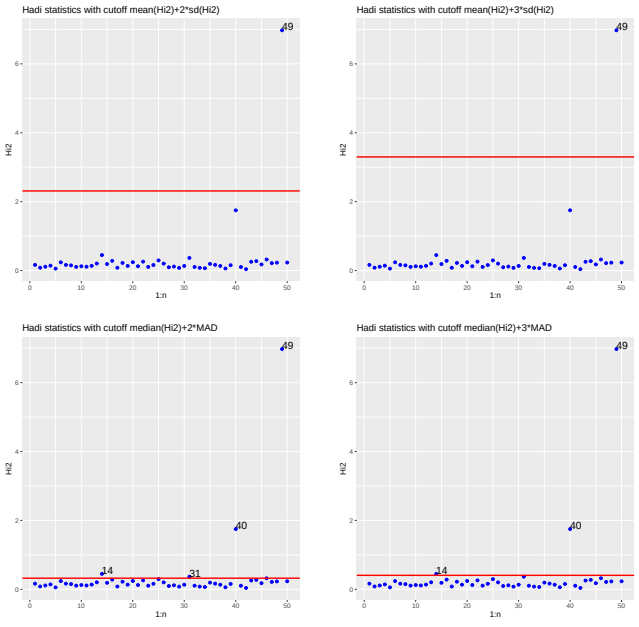
شکل ۱۴: نمودار نتایج آماره تفاوت در برازش‌ها.



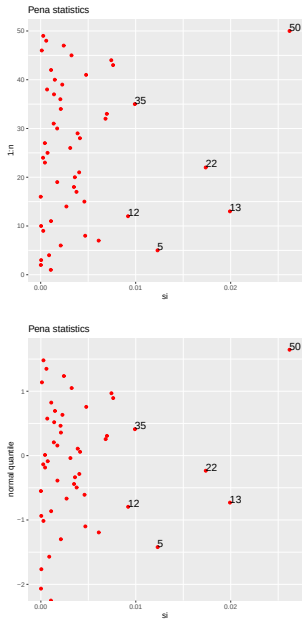
شکل ۱۵: نمودار نتایج فاصله ولسچ.



شکل ۱۶: نمودار نتایج آماره کوک اصلاح‌شده.



شکل ۱۷: نمودار نتایج آماره هادی.



شکل ۱۸: نمودار نتایج آماره پنا.

۶ بحث و نتیجه‌گیری

نقاط دورافتاده، نتایج و خروجی‌های مدل را به‌طور قابل‌توجهی منحرف می‌کنند. بنابراین قبل از هر تحلیل باید این نقاط به‌درستی شناسایی‌شده و موردبررسی قرار بگیرند. در شرایطی ممکن است که روش مدل‌بندی، نادرست انتخاب‌شده باشد و نقاط دورافتاده شناسایی‌شده، واقعاً دورافتاده نباشند بلکه نشانه‌هایی برای اعلام نادرستی مدل باشند. روش‌های مطرح‌شده در این مقاله، برای شناسایی انواع نقاط موجود در نمونه بسیار مفید و قابل‌استفاده می‌باشند. به‌کارگیری هم‌زمان چندین روش بسیار راهگشا خواهد بود و دید جامعی نسبت به انواع مشاهدات در اختیار کاربر قرار می‌دهد. لازم به توضیح است که روش‌های مطرح‌شده در این مقاله قابل‌تعمیم به مدل‌های نیمه پارامتری نیز هست [۲].

تقدیر و تشکر

نویسندگان مقاله ضمن تشکر از اعضای محترم هیئت تحریریه مجله، از پیشنهادها و نظرات ارزشمند داوران و ویراستار محترم مقاله که موجب ارتقاء سطح آن گردید کمال تشکر و قدردانی را دارند.

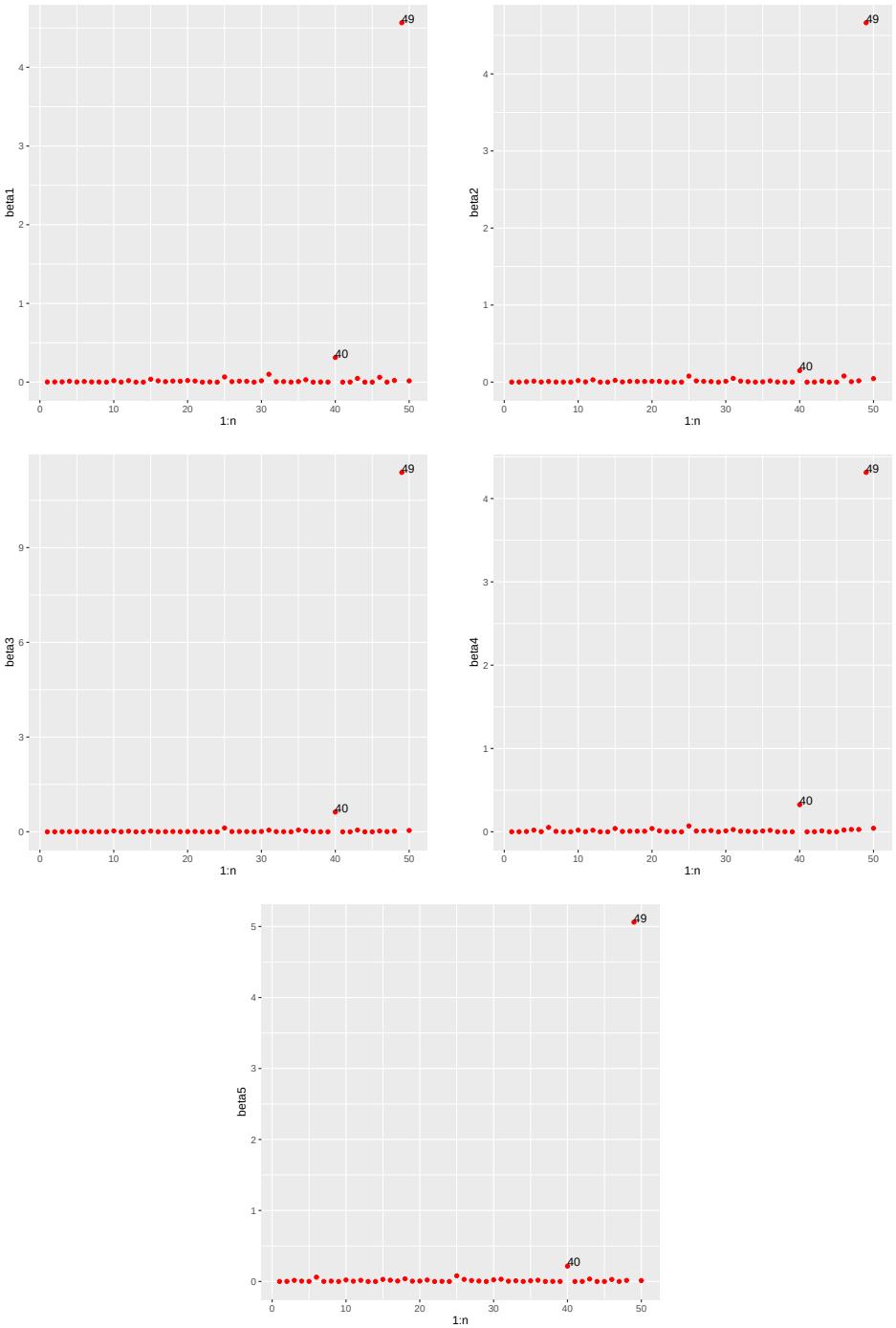
مراجع

[۱] نیرومند، ح.، (۱۳۸۷). *تحلیل رگرسیونی خطی*. انتشارات ارسلان، مشهد.

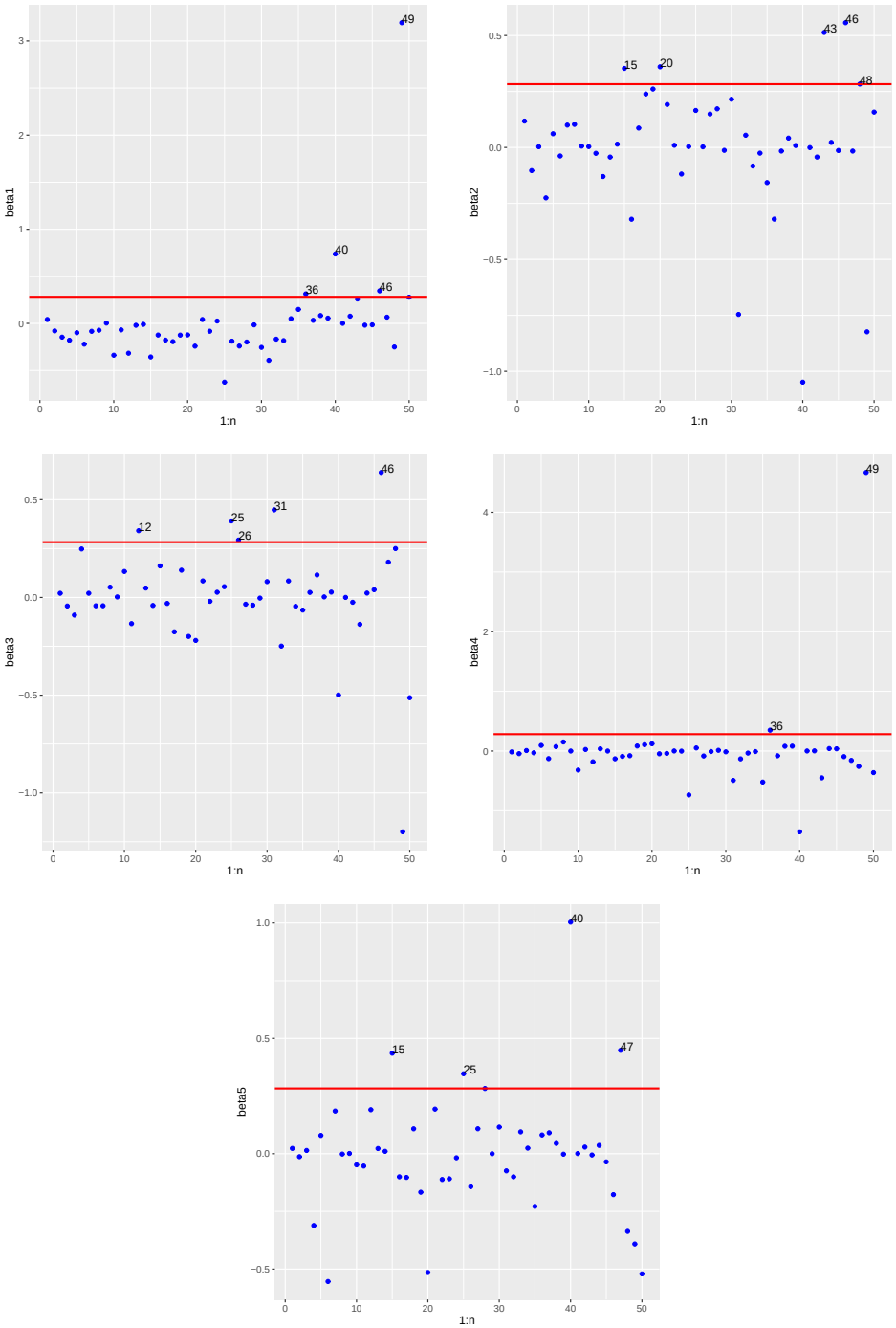
[2] Amini, M. and Roozbeh, M. (2015). Optimal partial ridge estimation in restricted semiparametric regression models, *Journal of Multivariate Analysis*, **136**, 26-40.

[3] Bahadir, B., Inki, H. and Karadavut, U. (2014). Determination of outlier in live-weight performance data of Japanese quails (coturnix x Japonica) by dfbeta and dfbetas techniques, *Italian Journal of Animal Science*. **13(1)**, 151-154.

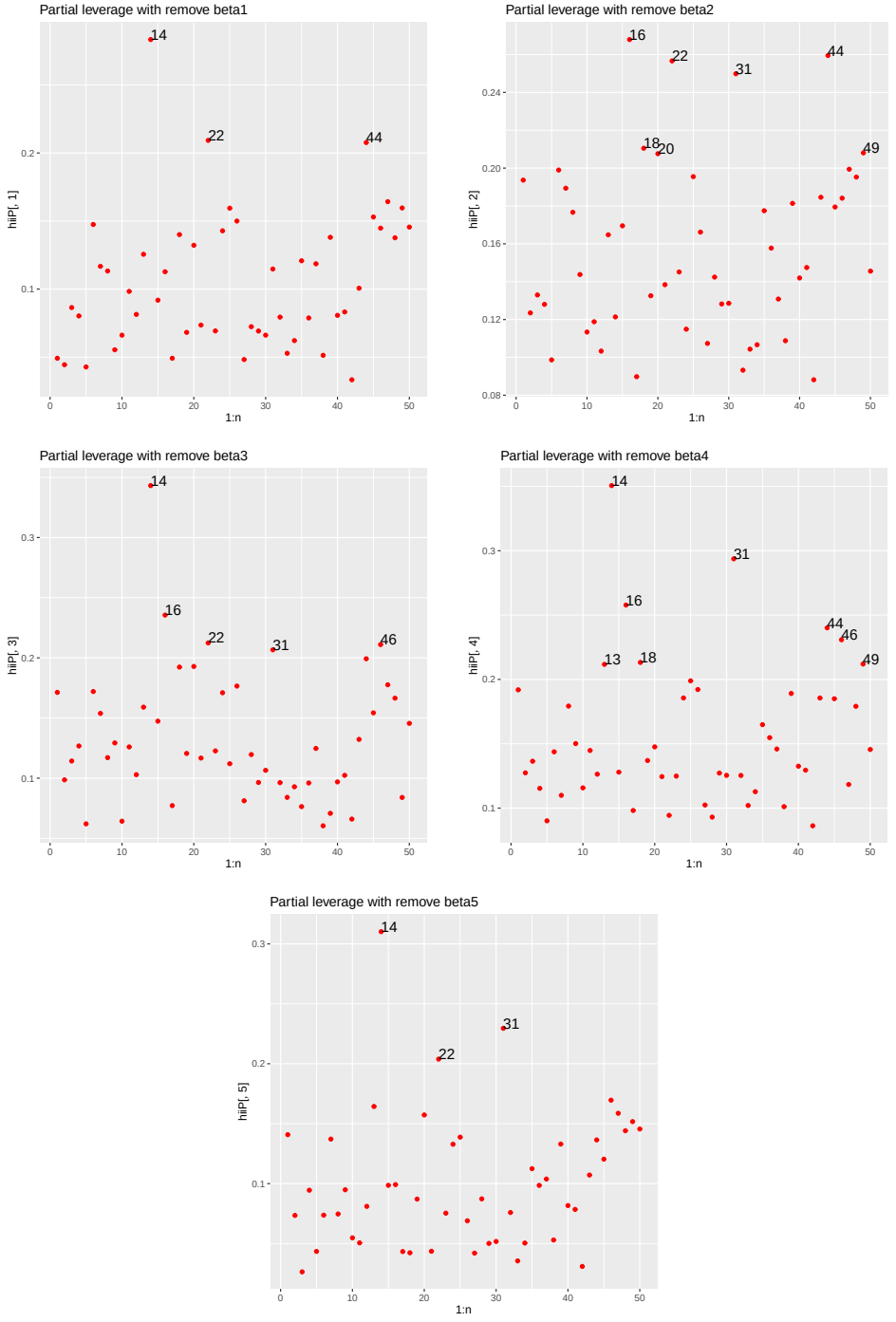
[4] Barnett, V. and Lewis, T. (1994). *Outliers in statistical sata*. JohnWiley ad Sons, New York.



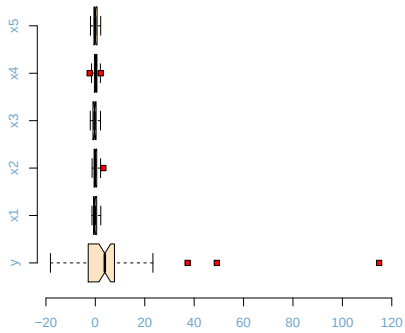
شکل ۱۹: نمودار نتایج تأثیر جزئی.



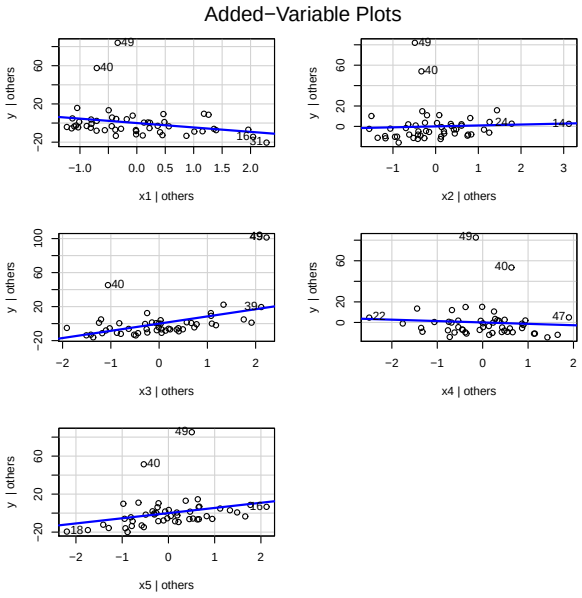
شکل ۲۰: نمودار نتایج تفاوت در پارامترها.



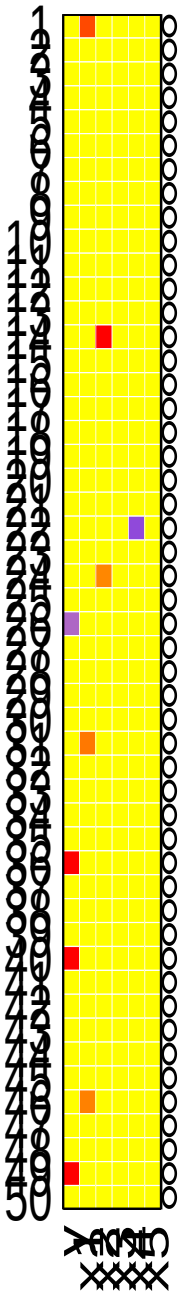
شکل ۲۱: نمودار نتایج اهرم جزئی.



شکل ۲۲: نمودار جعبه‌ای.



شکل ۲۳: نمودارهای متغیر افزوده.



شکل ۲۴: نمودار Cell Map.

[5] Beckman, R.J. and Cook, R.D. (1983). Outlier..... s, *Technometrics*. **25(2)**, 119-149.

[6] Beckman, R.J. and Trussell, H.J. (1974). The distribution of an arbitrary and the effects of updating in multiple regression, *Journal of the American Statistical Association*. **69(345)**, 199-201.

[7] Belsey, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. John Wiley, New Jersey.

[8] Chatterjee, S. and Hadi, A. S. (1986). Influential observation, high leverage points, and outliers in linear regression, *Journal Statistical Science*. **1(3)**, 379-416.

[9] Cook, R. D., and Weisberg (1982). *Residuals and influence in regression*. Chapman and Hall, New York.

[10] Cousineau, D. and Chartier, S. (2010). Outliers detection and treatment: a review. Detección y tratamiento de valores extremos: una revisión, *International Journal of Psychological Research*, **3(1)**, 58-67.

[11] Das, M.K. and Gogoi, B. (2015). Influential observations and cutoffs of different influence measures in multiple linear regression, *International Journal of Computational and Theoretical Statistics*. **2(2)**, 79-85.

[12] Ellenberg, J.H. (1973). The joint distribution of the standardized least squares residuals from a general linear regression, *Journal of the American Statistical Association*. **68(344)**, 941-943.

[13] Eubank, R.L. (1985). Diagnostics for smoothing splines, *Journal of the Royal Statistical Society*. **47(2)**, 332–341.

[14] Eyduran, E., Ozdemir, T. and Alarslan, E. (2005). Important of diagnostics in multiple regression analysis, *Journal of Appled Sciences*. **8(10)**, 1792-1796.

[15] Fox, J. (2002). *An R and S-PLUS companion to applied regression*, Sage, California.

[16] Gauss, C. F. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambien-tium, perthes et besser, Hamburg. Werke, **7**, 1–280.

[17] Gauss, C. F. (1823). *Theoria combinationis observationum erroribus minimum obnoxiae*, translated by G.W. Stewart as theory of the combination of observations least subject to errors: part one, part two, supplement: supplement Pt. 1 and Pt. 2, New York: SIAM.

[18] Hadi, A.S. (1992). A new measure of overall potential influence in linear regression, *Computational Statistics and Data Analysis*. **14(1)**, 1–27.

[19] Hawkins, D. (1980). *Identification of outliers*, Chapman and Hall, London.

[20] Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA, *Journal The American Statistician*, **32(1)**, 17-22.

[21] Imon, A.H.M.R. (2005). Identifying multiple influential observations in linear regression, *Journal of Applied Statistics*, **32(9)**, 929-946.

[22] Kim, C. (1996). Cook's distance in spline smoothing, *Statistics and Probability Letters*. **31(2)**, 139–144.

[23] Kutner, M. H., Nachtesim, C. J., Neter, J., and Li, W. (2004). *Applied linear statistical models*. McGraw-Hill, New York.

[24] Liu H., Shah S. and Jiang W. (2004). On-line outlier detection and data cleaning, *Computers and Chemical Engineering*. **28(9)**, 1635–1647.

[25] Montgomersy, D.C., PECK, E.A. and Vining, G.G. (2007). *Introduction to linear regression*. Analysis,3rd edn, Wiley series in probability and statistics, Hoboken.

[26] Moore, D.S. and Mccabe G.P. (1999). *Introduction to the practice of statistics*. Science and Math, New York.

[27] Mosteller, F. and Tukey, J.W. (1977) *Data analysis and regression*. Science and Math, Tepi.

[28] Pena, D. A. (2005). New Statistic for Influence in Linear Regression, *Technometrics*, **47(1)**, 1-12.

[29] Rawlings, J.O., Pantula, S.G. and Dickey, D.A. (1998). *Applied regression analysis: a research tool*. Springer, New York.

[30] Roozbeh, M. (2016). Robust ridge estimator in restricted semiparametric regression models, *Journal of Multivariate Analysis*. **147(5)**, 127-144.

[31] Roozbeh, M. and Arashi, M. (2017). Least-trimmed squares: asymptotic normality of robust estimator in semiparametric regression models, *Journal of Statistical Computation & Simulation*. **147**, 1130-1147.

[32] Roozbeh, M. and Babaie-Kafaki, S. (2016). Extended least trimmed squares estimator in semiparametric regression models with correlated errors, *Journal of Statistical Computation and Simulation*. **186**, 357-372.

[33] Rousseeuw, P. J., and Van Driessen, K. (2006), Computing its regression for large data sets, *Data mining and Knowledge Discovery*. **12(1)**, 29–45.

[34] Rousseeuw, P.J. and Van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*. **85(411)**, 633-639.

[35] Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion), *Journal of the Royal Statistical Society: Series B (Methodological)*. **47(1)**, 1–52.

[36] Stevens, J.P. (1984). Outliers and influential data points in regression analysis, *Psychological Bulletin*, **95(2)**, 334-344.

[37] Thomas, W. (1991). Influence diagnostics for the cross-validated smoothing parameter in spline smoothing, *Journal of the American Statistical Association*. **86(415)**, 693-698.

[38] Turkan, S., Cetin, C.M. and Toktamis, O. (2012). Outliers detection by regression diagnostics based on robust parameter estimates, *Journal of mathematics and statistics*. **41(1)**, 147-155.

[39] Weisberg, S. (1983). Some principles for regression diagnostics and influence analysis, *Technometrics*. **25(3)**, 240-244.

[40] Welsch, R.E. (1982). *Influence functions and regression diagnostics*. In modern data analysis, New York.

[41] Williams, G. J., Baxter, R. A., He, H. X., Hawkins, S. and Gu, L. (2002). A Comparative Study of RNN for Outlier Detection in Data Mining, *2002 IEEE International Conference on Data-mining*, Proceedings, 709-712 ,Maebashi, Japan.

Influential points Detection Methods for the Least Squares Method

Monireh Manavi[1] Mahdi Roozbeh[2]

Abstract:

The method of least squares is a very simple, practical and useful approach for estimating regression coefficients of the linear models. This statistical method is used by users of different fields to provide the best unbiased linear estimator with the least variance. Unfortunately, this method will not have reliable output if outliers are present in the dataset, as the collapse point (estimator consistency criterion) of this method is 0%. It is therefore important to identify these observations. Until now, the various methods have been proposed to identify these observations. In this article, the proposed methods are reviewed and discussed in details. Finally, by presenting a simulation example, we examine each of the proposed methods.

Keywords: Least Squares, Leverage point, Outliers, Outlier Detection.

[1]**Master's degree graduated, statistics and Computer science, Semnan university, Semnan, Iran.**

[2]**Faculty of mathematics, Semnan university, Semnan, Iran.**