روش های شناسایی و تحلیل داده های پرت در تحلیل رگرسیون

فریبا اسدی، حسین فلاح زاده، الهام خالق پناه نوقابی
دانشکده بهداشت و پیراپزشکی فردوس، دانشگاه علوم پزشکی بیرجند
گروه آمار و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی شهید صدوقی یزد
گروه آمار، دانشکده علوم ریاضی و آمار، دانشگاه بیرجند

چکیده

در تمامی مطالعات وجود داده پرت و دور افتاده از مسائل مشکل ساز در تجزیه و تحلیل نتایج می باشد. داده پرت یا به عبارتی نقطه دورافتاده نقطه ای است که از سایر نقاط فاصله زیادی داشته باشد. این گونه داده ها به دلایل مختلفی از جمله اشتباه در جمع آوری داده ها، ابزار اندازه گیری نادرست، وجود افراد غیرمعمول در نمونه و... به وجود می آیند. در این مقاله هدف شناساندن راه های تشخیصی داده های پرت و درمان آنها در تحلیل رگرسیون می باشد. در این مقاله روش های تشخیصی داده های دور افتاده بر روی ۲۰ زن ۲۵-۳۴ ساله بررسی شد که هدف بررسی ارتباط توده چربی بدن با ۲ متغیر قطر ران و قطر بازو می باشد. تجزیه و تحلیل داده ها با استفاده از نرم افزار SAS صورت پذیرفت. با استفاده از روشهای معرفی شده داده سوم به عنوان داده پرت تشخیص داده شد و پس از انجام رگرسیون با حضور داده پرت و بدون حضور آن داده، مشخص شد که این داده بی تاثیر بوده است.

واژه‌های کلیدی: داده پرت، فاصله کوک، باقیمانده های حذف شده استیودنت شده.

۱ مقدمه

داده پرت یا به عبارتی نقطه دورافتاده نقطه ای است که از سایر نقاط فاصله زیادی داشته باشد. در رگرسیون می توانیم تعاریف دیگری نیز برای نقطه دور افتاده ارائه دهیم که عبارتند از [۴]:

۱. نقطه دورافتاده رگرسیونی: نقطه ای است که از خطی که با $n-1$ نقطه دیگر برازش می شود فاصله زیادی داشته باشد.

۲. نقطه دورافتاده باقیمانده ای: نقطه ای است که دارای قدرمطلق باقیمانده استیودنت شده یا استاندارد شده بزرگی باشد.

۳. نقطه دورافتاده از فضای (x, y): نقطه ای است که مختصات x و y آن از دیگر نقاط فاصله زیادی دارد.

این گونه داده ها به دلایل مختلفی به وجود می آیند:

- ابزار اندازه گیری نادرست،

- خطاهای انسانی از قبیل خطا در نوشتن داده ها یا وارد کردن اشتباه داده ها در رایانه،
- جمع آوری داده ها از جوامع مختلف،
- وجود افراد غیر معمول در نمونه [۱، ۹].

با توجه به اینکه داده های پرت در تمام مراحل مربوط به آنالیز و تفسیر اطلاعات چه از لحاظ ساختاری و چه از لحاظ مفهومی تأثیر گذار هستند و بعضی موارد امکان نتیجه گیری منطقی از اطلاعات جمع آوری شده وجود ندارد و دچار خطاهای علمی آماری از لحاظ پایایی و روایی می شویم [۲]. بنابراین شناسایی و تشخیص داده های پرت و اقدام در جهت اصلاح آن ها امری ضروری در تمامی تحقیقات می باشد.
آماردانان روش های مختلفی را برای شناسایی داده های پرت پیشنهاد کرده اند. از جمله روش های شناسایی داده های پرت در رگرسیون عبارتند از:

۱. استفاده از نمودار جعبه ای
۲. استفاده از نمودار هیستوگرام
۳. استفاده از نمودار پراکنش
۴. استفاده از فاصله ی کوک[۱]
۵. استفاده از باقیمانده های حذف شده استیودنت شده
۶. استفاده از ماتریس H
۷. استفاده ازمجموع مربعات موزون فاصله امین-i نقطه از مرکز داده ها $(WSSD_i)$
۸. استفاده از تفاوت در مقدار برازش شده ی استاندارد شده $(DFBETAS_{j,i})$

باقیمانده های حذف شده ی استیودنت شده به باقیمانده هایی گفته می شود که در مدل های رگرسیونی به صورت زیر تعریف شده اند [۷]:
در یک مجموعه از داده ها که داده ی i-ام پرت به نظر می رسد با حذف مورد i-ام باقیمانده ها محاسبه می شود. بنابراین برای پیش بینی مقدار مشاهده از فرمول زیر استفاده می کنیم:

$$\hat{Y}_{(-i)}(x_{i,1},...,x_{i,n}) = \hat{\beta}_{0(-i)} + \hat{\beta}_{(-i)}x_{i,1} +... + \hat{\beta}_{(-i)}x_{i,k}.$$

اگر تفاوت این مقدار برازش شده را از مقدار واقعی آن بدست بیاوریم و بر خطای استاندارد آن تقسیم کنیم باقیمانده های حذف شده استیودنت شده بدست می آید. پس داریم:

$$T = \frac{y_i - \hat{Y}_{(-i)}(x_{i,1},...,x_{i,n})}{\frac{\hat{\sigma}_{(-i)}}{\sqrt{1-h_{i,i}}}},$$

که در آن $h_{i,i}$، همان عامل i-ام در ماتریس $H = X(X^TX)^{-1}X^T$ است که تمام مشاهدات در ماتریس X وجود دارند. T_i در اینجا دارای توزیع t-استیودنت با $(n-1)-k-1 = n-k-2$ درجه آزادی می باشد، به همین دلیل به باقیمانده های حذف شده استیودنت شده[۲] گاهی اوقات t-باقیمانده گویند. اگر مقدار T_i از ۲ (بعضی محققان ۳ در نظر می گیرند.) بزرگتر باشد، نشان می دهد که آن داده پرت است و احتیاج به بررسی علت دارد. وقتی که برای تشخیص داده های پرت از باقیمانده های حذف شده استیودنت شده استفاده می کنیم واضح تر از

[۱] Cook's

[۲] Studentized deleted residuals

زمانی هستند که از باقیمانده های استاندارد شده استفاده کنیم، می توان ثابت کرد که مجموع مقادیر ماتریس H در رگرسیون خطی معمولی برابر با p می باشد، که p در اینجا همان تعداد ضرایب رگرسیون (β) است. متوسط این مقادیر در یک مجموعه n عضوی برابر با $\frac{p}{n}$است. لذا برای هر مشاهده ای که $h_{i,i} > \frac{۲p}{n}$ یا $\frac{۳p}{n}$ آن باشد. مشاهده i-ام یک داده پرت به حساب می آید. با رسم نمودار باقیمانده های استاندارد شده در برابر مقادیر برازش شده و نمودار باقیمانده های حذف شده استیودنت شده در برابر مقادیر برازش شده در تمام مدل های رگرسیون هر داده ای که در محدوده ی داده های دیگر نباشد به عنوان داده ی پرت می توان به حساب آورد. برای تشخیص داده های پرت با استفاده از نمودارها می توان گفت که درنمودارجعبه ای برای هر مجموعه داده ای، مشاهداتی که خارج از دامنه ی نمودار قرارگیرد جزء داده های پرت هستند. در نمودار پراکنش نیز می توان داده ای را که در خارج از محدوده ی سایر داده ها قرار گرفته باشد به عنوان داده ی دور افتاده به حساب آوریم. در تشخیص فاصله کوک استفاده از اندازه مربع فاصله بین $\hat{\beta}$ برآورد حداقل مربعات روی تمام نقاط و $\hat{\beta}_{(-i)}$ برآورد بدست آمده براساس حذف i-امین نقطه را پیشنهاد کرده است [۳]. این اندازه فاصله ای می تواند به شکل کلی زیر بیان شود:

$$D_i(M,c) = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})' M (\hat{\beta}_{(-i)} - \hat{\beta})}{c},\ i = ۱,...,n,$$

که در آن M و c به ترتیب $M = X'X$ و $c = pMSE$ تعریف می شوند. معمولا نقاطی که برای آنها $D_i > ۱$ است به عنوان نقاط پرت شناسایی می شوند [۱].
همچنین می توان تأثیر i-امین مشاهده را بر مقدار پیش بینی شده یا برازش نشده را مورد رسیدگی قرار داد. یک شاخص قابل قبول عبارتست از:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(-i)}}{\sqrt{s^۲_{(-i)} h_{i,i}}},\ i = ۱,...,n,$$

که در آن مقدار برازش شده ی $\hat{y}_{(-i)}$ بدون به کار گرفتن i-امین مشاهده است. در یک مجموعه داده کوچک یا متوسط اگر برای مشاهده ای $|DFFITS_i| > ۱$ باشد آن را به عنوان مشاهده ی پرت در نظر می گیریم ولی برای یک مجموعه داده بزرگ به طور کلی هر مشاهده ای که برای آن داشته باشیم $|DFFITS_i| > ۲\sqrt{\frac{p}{n}}$ بایستی مورد توجه قرارگیرد [۶]. در شاخص $DFFITS_{j,i}$ بر این مطلب اشاره می شود که اگر i-امین مشاهده کنار گذاشته شود، ضریب رگرسیونی $\hat{\beta}_j$ بر حسب واحد انحراف معیار چقدر تغییر می کند. این آماره به شکل روبرو تعریف می شود:

$$DFFITS_{j,i} = \frac{\hat{\beta}_j - \beta_{j(i)}}{\sqrt{s^۲_{(i)} c_{jj}}},$$

که در آن c_{jj} عضو j-ام قطر $(X'X)^{-۱}$ و $\beta_{j(i)}$ ضریب جمله j-ام رگرسیون بدون به حساب آوردن i-امین مشاهده است. به طورکلی می گوئیم که اگر $|DFFITS_i| > \frac{۲}{\sqrt{n}}$ مشاهده ی i-ام تأثیر قابل توجهی روی ضریب رگرسیون j-ام دارد.

۲ روش بررسی

داده های استفاده شده مربوط به ۲۰ زن ۲۵-۳۴ ساله است که هدف بررسی ارتباط توده چربی بدن با ۲ متغیر

قطر ران و قطر بازو می باشد. این داده ها در گریبل و لیر [۵] و کاتنر و همکاران [۸] مورد استفاده قرار گرفته است. اولین و ساده ترین روش بررسی داده پرت استفاده از نمودارها می باشد که در این مطالعه با توجه به مبحث مورد بررسی از آوردن آنها خودداری کرده و سایر روشها را توسط نرم افزار SAS بررسی نمودیم.

۳ یافته ها

همانطورکه در جدول ۱ (به پیوست مراجعه شود) مشاهده می کنیم، قدرمطلق مقدار $DFFITS_i$ برای داده سوم از مقدار ۱ تجاوز کرده و همچنین مقدار $DFFITS_{j,i}$ داده سوم برای هر سه پارامتر بیشتر از مقدار $\frac{۲}{\sqrt{n}} = ۰/۴۴۷$ شده است. پرت بودن این داده از طریق مقادیر H نیز تایید می شود. ذکر این نکته مهم است که داده های پرت به دو دسته تأثیر گذار و بی تأثیر تقسیم می شوند. وقتی داده ای تأثیرگذار است، که تمامی ضرایب و تحلیل ها در نبود آن تغییر معنی داری کنند.

در این سوال پس از بررسی ماتریس کواریانس متوجه شدیم بین دو متغیر پیشگو هم خطی وجود دارد به همین دلیل متغیر قطر بازو را که همبستگی کمتری با متغیر پاسخ داشت حذف کردیم و رگرسیون را با متغیر قطر ران یکبار با حضور داده پرت (مدل ۱) و یکبار با حذف داده پرت (مدل ۲) انجام دادیم. خروجی در جدول ۲ (به پیوست مراجعه شود) نشان داده شده است.

با توجه به جدول ۳ (به پیوست مراجعه شود) همانطور که می بینید تغییر بسیار ناچیزی در ضرایب و انحراف استاندارد ضرایب صورت گرفته است پس نتیجه می گیریم که داده پرت تشخیص داده شده از آزمونها و نمودارها تاثیرگذار نبوده و احتیاجی به حذف آن نداریم.

۴ نتیجه گیری

پس از تشخیص داده های پرت اولین قدم بررسی علت دور افتاده بودن آن داده است که باید با دقت و جدیت پیگیری شود. بعضی اوقات متوجه می شویم که داده پرت به یکی از دلایل ذکر شده در این مقاله است. در چنین مواردی در صورت امکان داده مذکور اصلاح و در غیر این صورت از داده ها حذف میکنیم. گاهی اوقات نیز این داده، مشاهده ای کاملاً پذیرفتنی و قابل قبول است. حذف چنین نقاطی نیز می توانند خطرناک باشد. گاهی در می یابیم که داده دورافتاده از بقیه داده ها اهمیت بیشتری دارد، زیرا ممکن است خواص کلیدی مدل را کنترل کند. گاهی نیز داده دور افتاده ممکن است به عدم مناسبتها در مدل نیز اشاره داشته باشد. گاهی نیز هیچ دلیل خاصی برای دورافتاده بودن آن پیدا نمی کنیم. در این مواقع یکبار رگرسیون را بدون آن داده و یکبار با حضور آن انجام می دهیم. اگر ضرایب بدست آمده از این دو مدل تفاوت چندانی نداشته باشد پس آن داده غیر موثر است و تحلیل را فقط با حضور آن داده انجام می دهیم ولی اگر متوجه تغییر شدیدی شدیم می فهمیم آن داده تاثیر گذار است و باید تحلیل را یکبار با حضور آن داده و یکبار بدون آن حتما انجام دهیم.

مراجع

[۱] بابایی، غ.، امانی، ف.، بیگلریان، الف. و کشاورز، م. (۱۳۸۶). روش های تعیین داده های پرت در مطالعات پزشکی. مجله دانشکده پزشکی دانشگاه علوم پزشکی تهران، دوره ۶۵، ۲۴-۲۷.

[۲] رضوی پاریزی، الف. (۱۳۸۲). مقدمه ای بر تحلیل رگرسیون خطی. ترجمه، مؤلفین: مونتگمری. د. و پک، الف. انتشارات دانشگاه شهید باهنر کرمان، ۱۹۳-۳۰۷.

[3] Cook, R. D. and Weisberg, S. (1982). Residuals and influence in regression. Chapman and Hall.

[4] Evans, V. P. (1991). Strategies for detecting outliers in regression analysis: an introductory primer. Advances in Social Science Methodology, 271–286.

[5] Grybill, F. A. and Lyer, H. K. (1994). Regression Analysis: Concepts and Applications. 84-352.

[6] Holmes, F. W. (2012). Distribution of variables by method of outlier detection. Frontiers in Psychology, Quantitative Psychology and Measurement, 3:211.
URL: http://journal.frontiersin.org/article/10.3389/fpsyg.2012.00211/full.

[7] Kendall, M. G. and Buckland, W. R. (1957). A dictionary of statistical terms. Oliver and Boyd, Edinburgh.

[8] Kutner, M. H., Nachtsheim, J. C. and William, L. J. N. (2005). Applied Linear Statistical Models. Gordon. B, McGraw-Hill, rwin, 384-413.

[9] Rousseeuw, P. J. and Leroy, A. M. (1987). Robust regression and outlier detection. John Wiley and Sons, New York.

۵ پیوست

در ادامه جداول مربوط به مقاله ارائه شده است.

جدول ۱: نتایج حاصل از بررسی داده های پرت با استفاده از مقادیر $DFBETA$، $DFFIT$ و ماتریس H

	$DFFBETA$				Diag Hat			
n	r	$Rstd$	H	$DFFITS$	C	Int	X_1	X_2
۱	−۱/۶۸۲۷	−۰/۷۳	۰/۲۰۱	−۰/۳۶۶۱	۰/۴۵۹	−۰/۳۰۵۲	−۰/۱۳۱۵	۰/۲۳۲
۲	۳/۶۴۲۹	۱/۵۳۴۳	۰/۰۵۸۹	۰/۳۸۳۸	۰/۰۴۵۴۸	۰/۱۷۲۶	۰/۱۱۵۰	−۰/۱۴۲۶
۳	−۳/۱۷۶۰	−۱/۶۵۴۳	۰/۳۷۱۹	−۱/۲۷۳۱	۰/۴۹۰۱۶	−۰/۸۴۷۱	−۱/۱۸۲۵	۱/۰۶۶۹
۴	−۳/۱۵۸۵	−۱/۳۴۸۵	۰/۱۱۰۹	−۰/۴۷۶۳	۰/۰۷۲۱۶	−۰/۱۰۱۶	−۰/۲۹۳۵	۰/۱۹۶۱
۵	−۰/۰۰۰۲	−۰/۰۰۰۱۲	۰/۲۴۸۰	−۰/۰۰۰۱	۰	−۰/۰۰۰۱	۰	۰/۰۰۰۱
۶	−۰/۳۶۰۸	−۰/۱۴۷۵	۰/۱۲۸۶	−۰/۰۵۶۷	۰/۰۰۱۱۴	۰/۰۳۹۷	۰/۰۴۰۱	−۰/۰۴۴۳
۷	۰/۷۱۶۲	۰/۲۹۸۱	۰/۱۵۵۵	۰/۱۲۷۹	۰/۰۰۵۷۶	−۰/۰۷۷۵	−۰/۰۱۵۶	۰/۰۵۴۳
۸	۴/۰۱۴۷	۱/۷۶۰۱	۰/۰۹۶۳	۰/۵۷۴۵	۰/۰۹۷۹۴	۰/۲۶۱۴	۰/۳۹۱۱	−۰/۳۳۲۵
۹	۲/۶۵۵۱	۱/۱۱۷۶	۰/۱۱۴۶	۰/۴۰۲۲	۰/۰۵۳۱۳	−۰/۱۵۱۴	−۰/۲۹۴۷	۰/۲۴۶۹
۱۰	−۲/۴۷۴۸	−۱/۰۳۳۷	۰/۱۱۰۲	−۰/۳۶۳۹	۰/۰۴۳۹۶	۰/۲۳۷۷	۰/۲۴۴۶	−۰/۲۶۸۸
۱۱	۰/۳۳۵۸	۰/۱۳۶۷	۰/۱۲۰۳	۰/۰۵۰۵	۰/۰۰۰۹۰	−۰/۰۰۹۰	۰/۰۱۷۱	−۰/۰۰۲۵
۱۲	۲/۲۲۵۵	۰/۹۲۳۲	۰/۱۰۹۳	۰/۳۲۳۳	۰/۰۳۵۱۵	−۰/۱۳۰۵	۰/۰۲۲۵	۰/۰۷۰
۱۳	−۳/۹۴۶۹	−۱/۸۲۵۹	۰/۱۷۸۴	−۰/۸۵۰۸	۰/۲۱۲۱۵	۰/۱۱۹۴	۰/۵۹۲۴	−۰/۳۸۹۵
۱۴	۳/۴۴۷۵	۱/۵۲۴۸	۰/۱۴۸۰	۰/۶۳۵۵	۰/۱۲۴۸۹	۰/۴۵۱۷	۰/۱۱۳۲	−۰/۲۹۷۷
۱۵	۰/۵۷۰۶	۰/۲۶۷۲	۰/۳۳۳۲	۰/۱۸۸۹	۰/۰۱۲۵۸	−۰/۰۰۳۰	−۰/۱۲۴۸	۰/۰۶۸۸
۱۶	۰/۶۴۲۳	۰/۲۵۸۱	۰/۰۹۵۳	۰/۰۸۳۸	۰/۰۰۲۴۷	۰/۰۰۹۳	۰/۰۴۳۱	−۰/۰۲۵۱
۱۷	−۰/۸۵۰۹	−۰/۳۴۴۵	۰/۱۰۵۶	−۰/۱۱۸۴	۰/۰۰۴۹۳	۰/۰۷۹۵	۰/۰۵۵۰	−۰/۰۷۶۱
۱۸	−۰/۷۸۲۹	−۰/۳۳۴۴	۰/۱۹۶۸	−۰/۱۶۵۵	۰/۰۰۹۴۶	۰/۱۳۲۱	۰/۰۷۵۳	−۰/۱۱۶۱
۱۹	−۲/۸۵۷۳	−۱/۱۷۶۲	۰/۰۶۷۰	−۰/۳۱۵۱	۰/۰۳۲۳۶	−۰/۱۲۹۶	−۰/۰۰۴۱	۰/۰۶۴۴
۲۰	۱/۰۴۰۴	۰/۴۰۹۴	۰/۰۵۰۱	۰/۰۹۴۰	۰/۰۰۳۱۰	۰/۰۱۰۲	۰/۰۰۲۳	−۰/۰۰۳۳

جدول ۲: ماتریس همبستگی

		قطر بازو	قطر ران	توده چربی
قطر بازو	ضریب همبستگی پیرسون	۱	۰/۹۲۴	۰/۸۴۳
	p-value	–	۰/۰۰	۰/۰۰
قطر ران	ضریب همبستگی پیرسون	۰/۹۲۴	۱	۰/۸۷۸
	p-value	۰/۰۰	–	۰/۰۰
توده چربی	ضریب همبستگی پیرسون	۰/۸۴۳	۰/۸۷۸	۱
	p-value	۰/۰۰	۰/۰۰	–

جدول ۳: نتایج حاصل از برازش مدل رگرسیونی با حضور داده پرت (مدل ۱) و مدل رگرسیونی بدون حضور داده پرت(مدل۲)

		ضریب	انحراف استاندارد	آماره t	p-value
مدل (۱)	عرض از مبدا	−۲۳/۶۳۴	۵/۶۵۷	−۴/۱۷۸	۰/۰۰۱
	قطر ران	۰/۸۵۷	۰/۱۱۰	۷/۷۸۶	۰/۰۰۰
مدل (۲)	عرض از مبدا	−۲۳/۶۸۳	۵/۶۹۹	−۴/۱۵۶	۰/۰۰۱
	قطر ران	۰/۸۶۰	۰/۱۱۱	۷/۷۵۴	۰/۰۰