آشکارسازی نقاط پرت به روش حداقل مربعات پیراسته

الهام رستمی

گروه آمار دانشگاه شهید چمران اهواز

چکیده: تحلیل رگرسیونی با استفاده از برآوردگر حداقل مربعات یک ابزار آماری مفید است که در اکثر علوم به طور گسترده مورد استفاده قرار می‌گیرد. متأسفانه روش حداقل مربعات، حساسیت بسیاری به نقاط پرت دارد. بر همین اساس روش‌های آماری برای آشکارسازی نقاط پرت توسعه یافته است. از جمله این روش‌ها، روش‌های رگرسیون مقاوم و استوار می‌باشد. در این پژوهش کاربرد رگرسیون مقاوم و برتری آن را نسبت به روش حداقل مربعات در داده‌های اقتصادی مورد ارزیابی قرار می‌دهیم. بدین منظور یکی از روش‌های رگرسیون مقاوم به نام حداقل مربعات پیراسته را با روش حداقل مربعات مقایسه کرده و نتایج حاصله نشان می‌دهد که روش برآورد حداقل مربعات پیراسته بهترعمل‌می‌کند و بی‌جهت تحت تأثیر نقاط پرت قرارنمی‌گیرد. بنابراین استفاده از رگرسیون مقاوم در تحلیل داده‌های اقتصادی به دلیل ارائه اطلاعات دقیق‌تر، توصیه می‌شود.

واژه‌های کلیدی: رگرسیون مقاوم، روش حداقل مربعات پیراسته، نقاط بانفوذ، نقطه فروریزش، رگرسیون استوار

۱ مقدمه

معمولاً کاربران از روش حداقل مربعات در تحلیل داده‌های اقتصادی بهره می‌گیرند. با این وجود، تجربه نشان داده است، انحرافات بسیاری در این نوع داده‌ها به واسطه تغییرات سیاست‌های اقتصادی بوجود می‌آید. این انحرافات باعث ایجاد نقاط پرت و مشاهدات بانفوذ و در نتیجه موجب تحلیل اشتباه می‌شود. لازم به ذکر است نه تنها متغیر وابسته را می‌توان به عنوان نقاط پرت قلمداد کرد، بلکه متغیر مستقل را نیز تحت عنوان نقاط بانفوذ، نوع دیگر نقاط پرت محسوب می‌شوند.

مسأله حساسیت رگرسیون حداقل مربعات نسبت به نقاط پرت ممکن است ناشی از برقرارنبودن برخی‌از فرض‌های مدل ازجمله فرض اساسی ثابت بودن واریانس خطا (هم‌واریانسی) باشد. از این رو در تحلیل داده‌ها لازم است از روش‌هایی استفاده

کنیم که به‌وجود نقاط پرت حساس‌نباشد. روش‌های رگرسیون مقاوم و استوار جایگزین مناسبی برای روش حداقل مربعات خواهد بود. این دو روش تأثیر نقاط پرت را تعدیل می‌کند و برازش بهتری را برای اکثر داده‌ها فراهم می‌نمایند. اغلب افراد گمان می‌برند که روش‌های رگرسیون مقاوم نقاط پرت را پنهان می‌کنند، اما خلاف این واقعیت دارد، زیرا نقاط پرت دورتر از خط برازش قرار می‌گیرند و بنابراین از طریق بزرگ بودن باقیمانده‌های آنها آشکار می‌شوند، در صورتی‌که باقیمانده‌های استاندارد روش حداقل مربعات به هیچ وجه نقاط پرت را نمایش نمی‌دهند.

روش‌های رگرسیون مقاوم و استوار اهداف مشترکی دارند، با این تفاوت که در روش‌های استوار در ابتدا نقاط پرت را شناسایی و حذف می‌نمایند، سپس بوسیله رگرسیون حداقل مربعات داده‌های باقیمانده را برازش می‌دهند. در صورتی‌که در روش های مقاوم در ابتدا یک خط رگرسیونی به داده‌ها برازش می‌دهند، سپس نقاطی که باقیمانده‌های بزرگی در معادله دارند را به عنوان نقاط پرت درنظر می‌گیرند. روش‌های رگرسیون مقاوم نسبت به روش‌های استوار کارایی بالاتری دارند. در این پژوهش یکی از روش‌های رگرسیون مقاوم به نام برآورد حداقل مربعات پیراسته را تشریح می‌کنیم و برتری آن را نسبت به روش برآورد حداقل مربعات مورد بررسی قرار خواهیم داد.

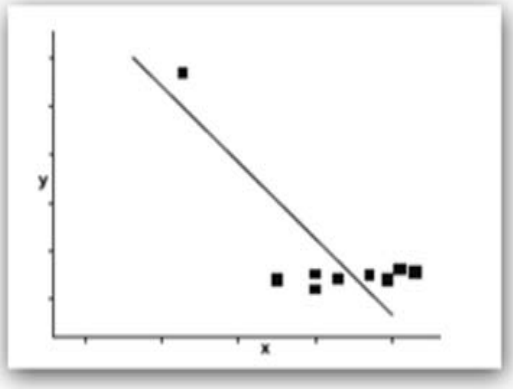
۲ مقایسه نقاط پرت و بانفوذ

نقاط پرت، مشاهداتی هستند که باقیمانده‌های بسیار بزرگی دارند. خطاهای استاندارد شده معمولاً دارای توزیع نرمال می‌باشد. نقاط با مقدار خطای استاندارد بیش از ۲ یا ۳ به عنوان نقاط پرت درنظر گرفته می‌شوند. بنابراین مقدار متغیر وابسته کمیتی است که برای شناسایی نقاط پرت استفاده می‌شود.

نوع دیگر نقاط پرت، نقاط بانفوذ نامیده می‌شوند که در رابطه با متغیرهای مستقل مورد بررسی قرار می‌گیرند. به عبارت دیگر مشاهدات بانفوذ، مشاهداتی هستند که حذف آنها از مدل تغییرات اساسی در مدل به وجود می‌آورد. همان طوری که مشاهده می‌کنید در شکل ۱، داده‌های بانفوذ در مجاورت خط رگرسیونی قرار دارد که شیب خط منفی می‌باشد، اگر این داده حذف شود شیب این خط مثبت خواهد شد و عرض از مبدأ کاهش می‌یابد. بدیهی است این داده نقش مهمی در تعیین برآورد رگرسیون خواهد داشت.

تحقیقات وسیعی در این رابطه انجام گرفته است به طوریکه منجر به مباحث گسترده‌ای چون روش‌های رگرسیون مقاوم و استوار شده است. معمولاً برای روش برآورد حداقل مربعات پیراسته نقطه فروریزش بالا را درنظر می‌گیرند که قادر می‌باشد

از عهده نسبت بزرگی از نقاط پرت برآید. مفهوم نقطه فروریزش بالا را در بخش بعدی شرح می‌دهیم.



شکل ۱: تأثیر یک مشاهده بانفوذ روی معادله رگرسیون

۳ برآوردگرهای با نقطه فروریزش بالا

برآوردگرهای با نقطه فروریزش، برآوردگرهایی هستند که درصدی از نقاط پرت را تحمل می‌نماید. برای روش حداقل مربعات پیراسته، نقطه فروریزش بالا را درنظر می گیرند، زیرا می‌تواند شمار بزرگی از نقاط پرت را تحمل نماید. در اینجا نقطه فروریزش با نمونه متناهی پیشنهاد شده توسط دونوهو و هابر[۱] (۱۹۸۳) را معرفی می‌کنیم. یک نمونه دلخواه با حجم n به صورت $s_n = \{x_1, ..., x_n\}$ را درنظر بگیرید و فرض کنید T_n برآورد رگرسیون باشد، به عبارت دیگر با به کار بردن T_n برای نمونه انتخابی s_n، برآوردی از ضرایب رگرسیون به صورت $T_n(s_n) = \hat{\theta}$ به دست می‌آید. بنابراین نقطه فروریزش برآوردگر T_n در s_n به صورت زیر تعریف می‌شود:

$$\epsilon_n^*(T_n, s_n) = \frac{1}{n} max\{m | maxsup|T_n(z_1, ..., z_n)| < +\infty\} \qquad (1)$$

زمانی که نمونه $(z_1, ..., z_n)$ از نمونه اصلی s_n با جایگذاری از مشاهدات گرفته شده است. نقطه فروریزش معمولاً به s_n بستگی ندارد. به عنوان مثال، با توجه به تعریف نقطه فروریزش میانگین، برابر صفر است در حالی که برای میانه برابر ۰/۵ می‌باشد.

[۱] Donoho and Huber

برای درک این مطلب درنظر بگیرید اگر برای هر داده مقادیر $y_i \to \pm\infty$، بنابراین داریم $\bar{y}_i \to \pm\infty$، در مقابل میانه نمونه با حرکت مقادیر y_i به سمت $\pm\infty$ تأثیر کمی می‌پذیرند. بنابراین میانه نسبت به ناخالص کردن خطاها مقاوم است، در حالیکه برای میانگین اینطور نیست. در حقیقت میانه ۵۰ درصد ناخالصی خطاها را تحمل می‌کند.

۴ روش برآورد حداقل مربعات پیراسته

این روش توسط روسیو[۲] (۱۹۸۳) معرفی شده است. هدف این روش برآورد مینیمم کردن مجموع کوچکترین مربعات باقیمانده‌ها می‌باشد. معمولاً برآورد حداقل مربعات پیراسته را با نقطه فروریزش بالا یعنی ۵۰ درصد درنظر می‌گیرند، به این معنی که نقاط پرت، کمتر از ۵۹۰ درصد داده‌ها را تشکیل می‌دهد.

تعریف: فرض کنید یک مدل رگرسیونی خطی برای یک نمونه (x_i, y_i) به صورت $y_i = \beta^T x_i + \epsilon_i$ $i = (1, ..., n)$ را داریم. برآوردگر حداقل مربعات پیراسته به صورت زیر تعریف می‌شود:

$$\hat{\beta}(LTS) = min \sum_{i=1}^{h} r^2_{[i]}(\beta) \qquad (۲)$$

هنگامی که $r^2_1(\beta) \leq ... \leq r^2_h(\beta)$ مربعات باقیمانده‌های مرتب شده را نشان می‌دهد. ثابت پیراسته h توجیه می‌کند که $\frac{n}{2} < h \leq n$. این ثابت، نقطه فروریزش برآوردگر حداقل مربعات پیراسته را تعیین می‌کند، تعریف (۲) دلالت می‌کند که تعداد $n - h$ مشاهده با بزرگترین باقیمانده، برآوردگر را تحت تأثیر قرار نخواهند داد. برای انتخاب ثابت پیراسته در زیربخش ۳.۴ توضیحات لازم را خواهیم داد.

۱.۴ محاسبه برآورد حداقل مربعات پیراسته

وقتی n به اندازه کافی کوچک باشد یعنی $(n < 50)$، تمام زیرنمونه‌های با حجم h، $q = \binom{n}{h}$ را باید تولید کرده و برای هرکدام از زیرنمونه‌ها مقدار تابع هدف رابطه (۲) را بدست می‌آوریم. در بین این زیرنمونه‌ها هرکدام تابع هدف را مینیمم سازد، برآورد نهایی می‌باشد. متأسفانه وقتی n بزرگ باشد، امکان بررسی کل حالات وجود ندارد، در این حالت باید از تقریب استفاده نمود. یک الگوریتم برای به دست آوردن این تقریب توسط روسیو و درایسن[۳] (۱۹۹۹) پیشنهاد شده است.

[۲] Rousseeuw
[۳] Rousseeuw and Drissen

۲.۴ انتخاب ثابت پیراسته

همان‌طور که در بخش قبا اشاره نمودیم، ثابت پیراسته در بازه $\frac{n}{۲} < h \leq n$ قرار دارد. انتخاب این ثابت تنها به هدف استفاده از رگرسیون حداقل مربعات پیراسته دارد، البته زمانی که نقطه فروریزش بزرگتر انتخاب می‌شود کارایی را بهبود می‌بخشد، زیرا اطلاعات بیشتری از داده‌ها به کار می‌روند. نقطه فروریزش ماکزیمم زمانیکه $h = [\frac{n}{۲}] + \left[\frac{p+۱}{۲}\right]$ به دست می‌آید. غالباً انتخاب ماکزیمم h زمانی به کار می‌رود که روش حداقل مربعات پیراسته برای مقایسه با بعضی از برآوردگرها از جمله حداقل مربعات استفاده می‌شود. از طرف دیگر انتخاب ماکزیمم h نیز ممکن است نسبت به دامنه بزرگی از مقادیر ثابت پیراسته حساس باشد، زیرا این مسئله موجب مقداری ناخالصی و احتمال ساختار مشکوک برای مجموعه داده‌ها می‌شود. این قضیه به طور مثال زمانی اتفاق می‌افتد که مجموعه داده‌ها شامل ترکیبی از دو یا چند جامعه متفاوت باشد، اما این مسئله نقطه ضعف برای برآوردگر مورد بحث محسوب نمی‌شود.

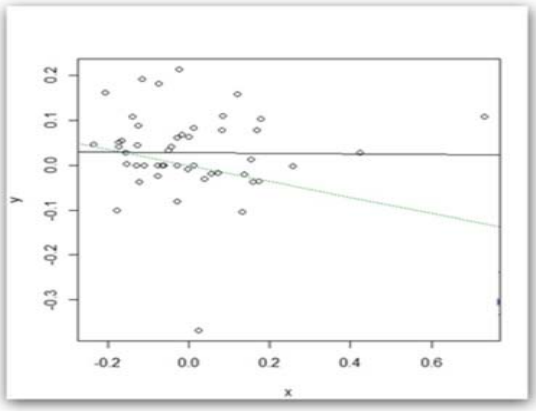
۵ مثال

داده‌های مربوط به یکی از شرکت‌های پذیرفته شده در بورس به نام پتروشیمی خارک با متغیر وابسته (بازده ماهیانه سهم) و متغیر مستقل (شاخص کل بورس به صورت ماهانه) در طول سال‌های ۱۳۸۹-۱۳۸۶ به صورت ماهیانه مورد ارزیابی قرار دادیم. با استفاده از بسته نرم‌افزاری $rrcov$ در محیط R معرفی شده توسط هوتینگ[۴]، رگرسیون حداقل مربعات پیراسته را روی داده‌ها اجرا می‌کنیم. در شکل ۳ خط سبز رنگ رگرسیون حداقل مربعات پیراسته و خط مشکی رنگ رگرسیون حداقل مربعات را نشان می‌دهد. نتایج نشان می‌دهد برآورد حداقل مربعات پیراسته کاراتر از برآورد حداقل مربعات است.

بحث و نتیجه‌گیری

رگرسیون حداقل مربعات، در اکثر تحقیقات اقتصادی استفاده می‌شود که به شدت تحت تأثیر نقاط پرت قرار می‌گیرد. بدین ترتیب برای کنترل نقاط پرت از روش‌های رگرسیون مقاوم از جمله رگرسیون حداقل مربعات پیراسته استفاده می‌شود. مقایسه این دو روش رگرسیون حداقل مربعات و حداقل مربعات پیراسته نشان می‌دهد روش حداقل مربعات پیراسته از کارایی بالاتری برخوردار است.

[۴] Hoeting



شکل ۲:. رگرسیون حداقل مربعات و حداقل مربعات پیراسته

مراجع

Cizek, P., and Visek, J.A (2000), *Least trimmed squares, In Xplore Application Guide*, Hardle, W., Hlavka, Z., Klinke, S. editors, Springer Verlag, 46-64.

Doornik, J.A. (2011), *Robust Estimation Using Least Trimmed Squares*, Institute for Economic ModellingUniversity of Oxford, UK, **2**, 1-17.

Jureckov, J.and Picek, J. (2006), *Roboust Statistical Methods with R*, Chapman Hall/CRC.

Momeni, M., Dehghan Nayeri, M., Faal Ghayoumi, A. and Ghorbani, H. (2010), *Robust Regression and its Application in Financial Data Analysis*, World Academy of Science, Engineering and Technology, 521-526

Raftery, A., Hoeting, J., Volinsky, C., Painter, I and Yee Yeung, K. (2008), *The BMA Package*. http://cran.r-project.org/web/packages/BMA/BMA.pdf.

Rousseeuw, P.J. and Leroy, A.M. (1987), Robust Regression and Outlier Detection, Wiley, New York.

Rousseeuw, P.J. (1984), *Least median of square regression*, Journal of the American Statistical Association, **79**, 871-888.

Ricardo, A. Maronna, R. Douglas, M. and V ctor J. Yohai (2006), *Robust Statistics: Theory and Methods John Wiley Sons*, ISBN: 0-470-01092-4.