معرفی بسته $CircOutlier$ برای شناسایی داده‌های پرت در رگرسیون دایره‌ای-دایره‌ای

آزاده غضنفری حصاری [١]، مجید سرمد [٢]

چکیده:

یکی از مسائل مهم در هر تحلیل آماری، وجود مشاهدات غیرمنتظره است. بعضی از مشاهدات بخشی از مسائل مورد مطالعه نیستند و به عنوان داده پرت شناخته شده‌اند. بررسی‌ها نشان داده است که داده‌های پرت بر عملکرد روش‌های استاندارد آماری در مدل‌ها و پیش‌بینی‌ها تأثیر می‌گذارد. هدف این مقاله ارائه بسته‌ی موجود در نرم‌افزار R برای شناسایی داده پرت در رگرسیون دایره‌ای-دایره‌ای است که توسط نگارنده این مقاله نوشته شده است. ابتدا توضیح مختصری در مورد داده دایره‌ای و رگرسیون دایره‌ای داده می‌شود، سپس بسته‌های موجود در نرم‌افزار R برای انجام رگرسیون دایره‌ای معرفی شده، توابع موجود در بسته $CircOutlier$ شرح داده می‌شود و برای هر کدام از توابع مثالی ارائه خواهد شد.

واژه‌های کلیدی: شناسایی داده پرت در رگرسیون دایره‌ای-دایره‌ای، نرم‌افزارR، بسته $CircOutlier$.

١ مقدمه

در تجزیه و تحلیل داده‌های آماری، اغلب با مقادیری که مشکوک یا تعجب‌آور به نظر می‌رسند، مواجه می‌شویم. چنین مقادیری ممکن است نقاط دورافتاده باشند، که از اصطلاح "داده پرت" برای توصیف این مقادیر که بر اساس برخی از معیارهای آماری، مغایر با بقیه نمونه باشد، استفاده می‌کنیم. (کالِت [٣]). داده‌های جهت‌دار نیز از آلودگی به داده‌های پرت مستثنی نیستند. علاقه به توسعه روش‌های تجزیه و تحلیل داده‌های جهت‌دار یک موضوع قدیمی در آمار ریاضی است. اولین مطالعات مربوط به رگرسیون دایره‌ای به زمان گولد [٢] برمی‌گردد. رگرسیون دایره‌ای، یک روش تجزیه و تحلیل است برای زمانی که متغیر وابسته، یک نقطه روی محیط یک دایره و سطح یک کره می‌باشد. برای مثال، ممکن است علاقه‌مند به ثبت جهت باد و یا جهت حرکت ابرها باشیم. در چنین مواردی، ممکن است علاقه‌مند به مسائل مربوط به همبستگی و یا ارتباط بین این متغیرها و همچنین رگرسیون با هدف پیش‌بینی یک متغیر با توجه به متغیر دیگر باشیم.

رگرسیون دایره‌ای خود شامل سه بخش است:

رگرسیون دایره‌ای-دایره‌ای: زمانی که هر دو متغیر پاسخ و توضیحی دایره‌ای هستند.

رگرسیون دایره‌ای-خطی: زمانی که متغیر توضیحی روی خط حقیقی تعریف می‌شود و متغیر پاسخ دایره‌ای است.

رگرسیون خطی-دایره‌ای: زمانی که متغیر پاسخ خطی و متغیر توضیحی دایره‌ای است.

ما در این مقاله به بررسی داده پرت در رگرسیون دایره‌ای-دایره‌ای می‌پردازیم. قسمت‌های مختلف این مقاله به شرح زیر است. در بخش دوم به معرفی رگرسیون دایره‌ای-دایره‌ای، توزیع فون میزس و بسته‌های موجود در نرم‌افزار R برای رگرسیون دایره‌ای-دایره‌ای می‌پردازیم. در بخش سه روشی را برای شناسایی داده پرت در رگرسیون دایره‌ای-دایره‌ای معرفی می‌کنیم و در بخش چهار نرم‌افزارهای مورد استفاده برای شناسایی داده پرت، بسته $CircOutlier$ و توابع موجود در آن را مورد بررسی قرار می‌دهیم. در بخش پنج مثال‌هایی برای کاربرد این بسته ارائه می‌دهیم. در نهایت در بخش شش یک نتیجه‌گیری کوتاه از کارهای انجام شده در این مقاله بیان می‌کنیم.

[١] دانشجوی کارشناسی ارشد آمار ریاضی دانشگاه فردوسی مشهد، $azade.ghazanfarihesari@stu-mail.um.ac.ir$

[٢] استادیار دانشگاه فردوسی مشهد، $sarmad@um.ac.ir$

۲ رگرسیون دایره‌ای-دایره‌ای

در رگرسیون دایره‌ای ساده که یک رابطه خطی بین متغیرهای دایره‌ای X و Y فرض شده است، برای مشاهدات دایره‌ای $(x_۱, y_۱), ..., (x_n, y_n)$ مدل زیر را داریم:

$$y_i = \alpha + \beta x_i + \varepsilon_i \ \ (mod ۲\pi), \quad i = ۱, ۲, ..., n, \qquad (۱)$$

که ε_i خطا تصادفی دایره‌ای از توزیع فون‌میزس [۳] با میانگین صفر و پارامتر مرکزی k است. کاربرد این مدل در تحلیل و مقایسه جهت باد و موج به دست آمده از دو روش متفاوت، به منظور مقایسه یک ابزار جدید اندازه‌گیری جهت باد با یک ابزار قدیمی اندازه‌گیری جهت باد است.

۱.۲ توزیع فون‌میزس

متغیر تصادفی X دارای توزیع فون‌میزس یا توزیع نرمال دایره‌ای [۴] است، اگر دارای تابع چگالی زیر باشد:

$$f(x; \mu, k) = \frac{۱}{۲\pi I_۰(k)} e^{k \cos(x-\mu)} \quad , ۰ \le x < ۲\pi,$$

که در آن $۰ \le \mu < ۲\pi$ و $k \ge ۰$ پارامترهای توزیع هستند و $I_۰(k)$ در ثابت نرمال‌ساز، تابع بسل [۵] اصلاح شده نوع اول و مرتبه صفر است که به صورت زیر بدست می‌آید:

$$I_۰(k) = \frac{۱}{۲\pi} \int_۰^{۲\pi} e^{k \cos(x)} dx.$$

این توزیع به عنوان یک مدل آماری توسط فون‌میزس [۷] معرفی شد و قبل از آن توسط لانگوین[۸] در زمینه فیزیک مورد بحث قرار گرفت. به دلیل اهمیت و شباهت آن به توزیع نرمال روی خط حقیقی، آن را یک توزیع نرمال دایره‌ای می‌نامند. تابع چگالی فون میزس دارای خواص زیر است:

الف) تقارن

ب) مُد در μ

پ) پادمُد [۶] در $(\mu \pm \pi)$

۲.۲ بسته‌های موجود در نرم‌افزار Rبرای رگرسیون دایره‌ای-دایره‌ای

دو بسته‌ی $circular$ و $CircStats$ در نرم‌افزار R موجود هستند که توابع موجود در آن‌ها تحلیل‌های مربوط به داده‌های دایره‌ای از جمله رگرسیون دایره‌ای-دایره‌ای، محاسبه میانگین و واریانس داده‌های دایره‌ای و ... را انجام می‌دهند. برای دریافت این بسته‌ها به آدرس زیر می‌توان مراجعه کرد:

`http://cran.um.ac.ir/web/packages`

۳ شناسایی داده پرت در رگرسیون دایره‌ای-دایره‌ای

رائو [۵] فاصله دایره‌ای بین دو مشاهده دایره‌ای θ_i و θ_j را به صورت زیر تعریف کرد:

$$d_{ij} = ۱ - \cos(\theta_i - \theta_j).$$

مشاهده می‌کنیم که d_{ij} یک تابع صعودی یکنواخت از $(\theta_i - \theta_j)$ است و $d_{ij} \in [۰, ۲]$.

ماردیا [۳] یک انحراف زاویه‌ای از مشاهدات برای مقادیر برازش داده شده آن‌ها در مدل دایره‌ای تعریف کرد که از این آماره برای شناسایی داده‌های پرت موجود در مدل (۱) استفاده می‌کنیم. ماردیا [۳] میانگین خطای دایره‌ای را به صورت زیر تعریف می‌کند:

$$MCE = ۱ - \frac{\sum_{i=۱}^{n} \cos(y_i - \widehat{y}_i)}{n},$$

که در آن n حجم نمونه و $\widehat{y}_i$ مقدار برآورد شده از y_i تحت مدل (۱) است. توجه کنید که $MCE \in [۰, ۲]$. اگر یک مشاهده داده‌ای پرت باشد، انتظار می‌رود که فاصله دایره‌ای بین y_i و مقدار $\widehat{y}_i$ مرتبط با آن، بزرگ باشد. بنابراین، وجود چنین مشاهداتی در یک مجموعه داده، مجموع همه فواصل دایره‌ای و هم‌چنین آماره میانگین خطای دایره‌ای را افزایش خواهد داد. به این ترتیب با

[۳] Von Mises

[۴] Circular Normal

[۵] Bessel Function

[۶] Antimode

حذف i -امین مشاهده از مجموعه مشاهدات، مقدار آماره کاهش خواهد یافت. این مقدار کاهش یافته با $MCE_{(-i)}$ نشان داده می‌شود. ماکسیمم قدرمطلق اختلاف بین مقادیر آماره‌ی MCE و $MCE_{(-i)}$ به صورت زیر تعریف شده است:

$$DMCE = \max_i\{|MCE - MCE_{(-i)}|\}, \quad i = ۱, ۲, ..., n.$$

اگر قدرمطلق اختلاف آماره‌های MCE و $MCE_{(-i)}$ از مقدار $DMCE$ که بر اساس حجم نمونه و برآورد پارامتر مرکزی k از جدول شبیه‌سازی شده از توزیع $DMCE$ تعیین می‌شود، بیشتر باشد، i-امین مشاهده یک داده پرت است.

٤ معرفی بسته $CircOutlier$

موضوع داده‌های پرت در مدل رگرسیون خطی بسیار مورد علاقه محققانی چون بارنت و لِویس [۲]، مونتگومری و پِک[۷] و ... بوده است. بنابراین در نرم‌افزارهای $Minitab$، $S-plus$ و R، بسته‌هایی برای آزمون شناسایی داده پرت در مدل رگرسیون خطی فراهم شده است. بسته $CircOutlier$ توسط نگارنده این مقاله در سال ۲۰۱۵ میلادی نوشته شده است. با استفاده از این بسته، شناسایی داده پرت در رگرسیون دایره‌ای-دایره‌ای انجام می‌شود. توابعی برای اصلاح داده پرت و برآورد ضرایب در رگرسیون دایره‌ای-دایره‌ای نیز در این بسته موجود است.

۱.٤ توابع موجود در $CircOutlier$

با استفاده از توابع موجود در بسته $CircOutlier$ می‌توان به شناسایی داده پرت در رگرسیون دایره‌ای-دایره‌ای پرداخت. از جمله این توابع عبارتند از:

`MCE`: این تابع میانگین خطای دایره‌ای (MCE) را در داده‌های دایره‌ای محاسبه می‌کند.

`MCe`: این تابع با حذف i-امین مشاهده از مجموعه مشاهدات، مقدار آماره $MCE_{(-i)}$ را محاسبه می‌کند.

`DMCEE`: این تابع قدرمطلق اختلاف بین مقادیر آماره‌های MCE و $MCE_{(-i)}$ را محاسبه کرده، نمودار پراکنش این مقادیر را رسم می‌کند و سپس برآورد پارامتر مرکزی، k، را به دست می‌آورد. در نهایت با توجه به حجم نمونه و برآورد پارامتر مرکزی، خط مربوط به مقدار $DMCE$ را به نمودار پراکنش برازش می‌دهد. هر داده‌ای که بالای این خط قرار بگیرد، داده پرت محسوب می‌شود. مقادیر $DMCE$ در جدولی با همین نام در بسته $CircOutlier$ موجود است.

`Predict`: برآورد درستنمایی ماکسیمم پارامترها را در مدل (۱) محاسبه می‌کند. توجه داشته باشید که در بسته $circular$ برآورد پارامترهای مدل (۱) به روش مستقیم محاسبه نمی‌شود یعنی با به کار بردن دستور `lm.circular` ضرایبی به دست می‌آید که با استفاده از این ضرایب می‌توان دو معادله نوشت و از طریق آنها برآورد پارامترها را محاسبه نمود.

`Huberized`: این تابع به شناسایی داده پرت در داده‌های دایره‌ای می‌پردازد و پس از شناسایی آن‌ها را اصلاح می‌کند.

در ادامه با استفاده از دو دسته از داده‌ها با نام داده‌های $wind$۲ و $wind$ هر یک از توابع در قالب یک مثال شرح داده می‌شود.

۲.٤ داده‌های $wind$۲

این داده‌ها مربوط به اندازه‌گیری جهت باد در ساحل هامبرساید[۷] از دریای شمال می‌باشند که طی دوره ۲۲/۷ روز (تقریباً ۲۳ روز) ثبت شده‌اند. لازم به ذکر است این داده‌ها توسط ابوزید و همکاران [۱] مورد بررسی قرار گرفته‌اند. داده‌ها با عنوان $wind$۲ در بسته $CircOutlie$ موجود می‌باشند. این داده‌ها با استفاده از دو ابزار مختلف اندازه‌گیری جهت باد، یک سیستم رادار فرکانس بالا $(Radar)$ و یک لنگر موج شناور $(Anchored)$ اندازه‌گیری شده‌اند.

۳.٤ داده‌های $wind$

این داده‌ها قدرمطلق اختلاف آماره‌های MCE و $MCE_{(-i)}$ در داده‌های $wind$۲ هستند که بعد از انجام محاسبات ریاضی روی داده‌های $wind$۲ با عنوان $wind$ در بسته $CircOutlier$ قرار داده شده‌اند.

[۷] Humberside

۵ کاربرد بسته $CircOutlier$

مثال ۱.۵. در این مثال ابتدا داده‌های $wind2$ فراخوانی می‌گردد. در این داده‌ها مقادیر $Radar$ همان x و مقادیر $Anchored$ همان y در رگرسیون دایره‌ای–دایره‌ای هستند. سپس دستور MCE نوشته می‌گردد.

```
data(wind2)
MCE=MCE(y,Y,n)
```

که در آن n حجم نمونه و Y مقادیر برآورد شده از متغیر پاسخ دایره‌ای y تحت مدل (۱) است. در خروجی دستور MCE مقدار میانگین خطای دایره‌ای مشاهده می‌شود.

```
0.06712991
```

مقدار Y را می‌توان با استفاده از دستور `circ.reg` در بسته $CircStats$ به شکل زیر محاسبه نمود.

```
library(CircStats)
Y=circ.reg(Radar,Anchored)$fi
```

در خروجی این دستور مقادیر برآورد شده y یعنی Y محاسبه می‌شود.

```
Circular Data:
Type = angles
Units = radians
Template = geographics
Modulo = asis
Zero = 1.570796
Rotation = clock

       1           2           3
0.93457989  0.86234415  1.10409492...
   128         129
 0.71939203  0.57470826
```

حال دستور `MCe` را برای داده‌ها به کار می‌بریم.

```
MCe=MCe(u)
```

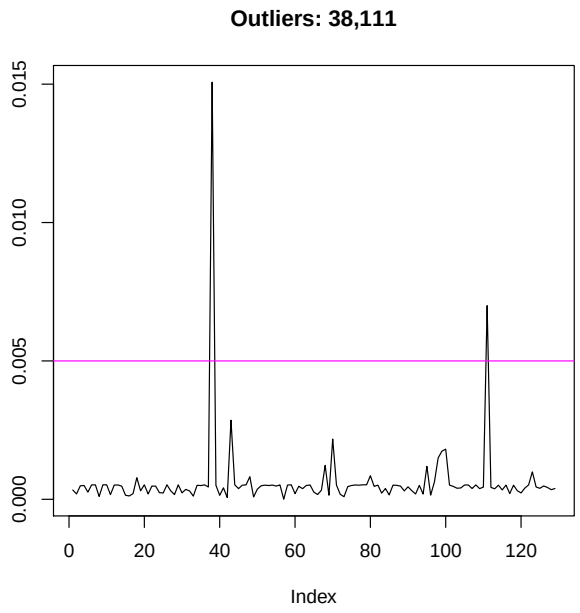
در این تابع u کسینوس اختلاف بین مقادیر مشاهده شده از متغیر پاسخ y و مقادیر برازش داده شده Y تحت مدل (۱) است. این تابع با حذف i–امین داده، به محاسبه آماره میانگین خطای دایره‌ای برای مجوعه داده‌های کاهش یافته $(MCE_{(-i)})$ می‌پردازد. در خروجی این دستور به تعداد داده‌ها مقدار میانگین خطای دایره‌ای مشاهده می‌شود.

```
[1] 0.06746705 0.06732443 0.06761756
    0.06763040 0.06738977 0.06765047
[7] 0.06765286 0.06723266 0.06765436
    0.06765431 0.06729894  ...
[127] 0.06755854 0.06747845 0.06751919
```

حال دستور `DMCEE` را برای داده‌ها به کار می‌بریم.

```
DMCEE=DMCEE(x, y, b)
```

که در آن x متغیر توضیح دایره‌ای، y متغیر پاسخ دایره‌ای و b سطح معنی‌داری ۰/۱ یا ۰/۰۵است. در خروجی، نمودار پراکنش مربوط به قدرمطلق اختلاف بین مقادیر آماره‌های MCE و $MCE_{(-i)}$ مشاهده می‌شود. هم‌چنین خط مربوط به $DMCE$ نیز به نمودار برازش داده شده است. هر داده‌ای که بالای این خط قرار می‌گیرد، داده پرت است.



شکل ۱. شناسایی داده پرت به روش $DMCE$

همانطور که از نمودار مشخص است دو مشاهده ۳۸ و ۱۱۱ داده پرت هستند.

در نهایت برای این داده‌ها دستور *Predict* را به کار می‌بریم.

```
Predict=Predict(x,y)
```

در این دستور x متغیر توضیح دایره‌ای، y متغیر پاسخ دایره‌ای است. خروجی این تابع برآورد ضرایب را در مدل (۱) به روش درستنمایی ماکسیمم محاسبه می‌کند.

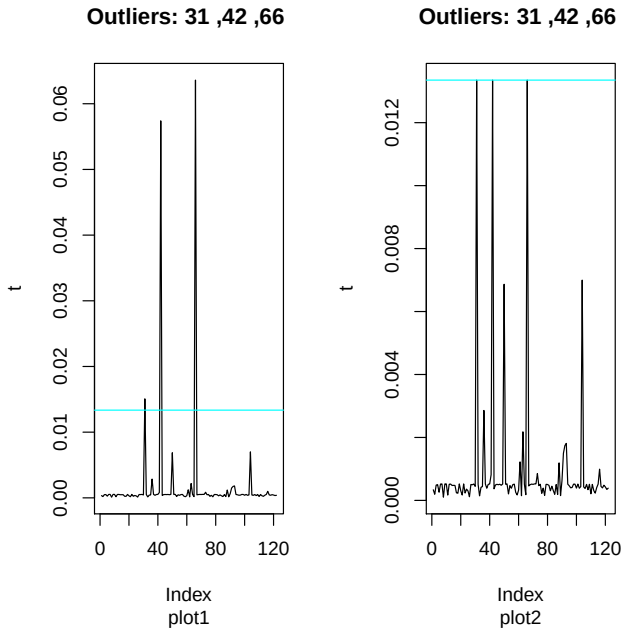
```
$output
      alpha1     beta1
[1,] 0.153068 0.9757867
```

در اینجا $alpha$۱ برآورد ضریب α و $beta$۱ برآورد ضریب β در مدل (۱) هستند.

مثال ۲.۵. در این مثال ابتدا داده‌های $wind$ از بسته $CircOutlier$ فراخوانی، سپس دستور *Huberized* نوشته می‌شود. در خروجی این تابع دو نمودار که مربوط به شناسایی داده‌پرت و اصلاح آن به روش هیوبر[۸]، میانگین و انحراف استاندارد داده‌ها در حضور داده پرت و بعد از اصلاح آن مشاهده می‌شود.

```
data(wind)
Huberized=Huberized(wind)
$output
          m           s
[1,] 0.001709521 0.007761697
        m1          s1
0.0009231165 0.00245804
```

نمودار اول مربوط به شناسایی داده پرت در داده‌ها می‌باشد و نمودار دوم با استفاده از روش هیوبر به اصلاح داده‌های پرت شناسایی شده پرداخته است. در روش هیوبر بیشتر از اطلاعات به دست آمده از داده‌ها استفاده شده است. این روش به جای حذف داده پرت به اصلاح آن می‌پردازد. فرض کنید دو برآورد اولیه $\widehat{\mu}_.$ (میانگین) و $\widehat{s}_.$ (انحراف معیار) را داریم. اگر یک مقدار x_i بالای خط $\widehat{\mu}_. + (۱/۵ * \widehat{s}_.)$ قرار بگیرد آن را با مقدار $\widehat{\mu}_. + (۱/۵ * \widehat{s}_.)$ جایگزین می‌کنیم. به‌طور مشابه اگر یک مقدار x_i پایین خط $\widehat{\mu}_. - (۱/۵ * \widehat{s}_.)$ قرار بگیرد آن را با مقدار $\widehat{\mu}_. - (۱/۵ * \widehat{s}_.)$ جایگزین می‌کنیم. سپس میانگین $\widehat{\mu}_۱ = mean(x_i)$ و انحراف معیار داده‌ها $\widehat{s}_۱ = ۱/۱۳۴ * sd(x_i)$ را به دست می‌آوریم. (عامل ۱/۱۳٤ از توزیع نرمال به دست آمده، و مقدار ۱/۵ اغلب در فرایند وینزورایزیشن[۹] استفاده می‌شود.) هم‌چنین m و s مربوط به میانگین و انحراف استاندارد دایره‌ای داده‌ها در حضور داده پرت، $m۱$ و $s۱$ مربوط به میانگین و انحراف استاندارد دایره‌ای داده‌ها بعد از اصلاح داده پرت می‌باشد.



شکل۲. شناسایی و اصلاح داده پرت

٦ نتیجه‌گیری

با کمک بسته $CircOutlier$ می‌توان داده‌های پرت در رگرسیون دایره‌ای-دایره‌ای را تشخیص داد و هم‌چنین با استفاده از روش هیوبر، مدل رگرسیون را اصلاح کرد.

[۸] Huber

[۹] Winsorization

مراجع

[1] Abuzaid, A. H., Hussin, A. G. and Mohamed, I. B., (2013), Detection of outliers in simple circular regression models using the mean circular error statistic, *Comm. Statist. Simulation Comput*, **83**, 269-277.

[2] Barnett, V., and Lewis, T., (1984), *Outliers in Statistical Data*, Second Edition, John Wiley and Sons, New York.

[3] Collet, D. (1980), Outliers in circular data. *Journal of Applied Statistics* , **29**, 50-57.

[4] Gould, A. L., (1969), A regression technique for angular response, *Biometrics*, **299**, 50-57.

[5] Langvin, P., (1905), Magnetime et theorie des electons, *Ann. Chim. Phys.*, **5**, 71-127.

[6] Mardia, K. V., (1972), *Statistics of Directional Data*, Academic Peress, London.

[7] Montgomery, D. C., and Peck, E. A., (1984), *Intrroduction to Linear Regression Analysis*, Second Edition, Wiley, New York.

[8] Rao, J. S. (1969), Some contributions to the analysis of circular data. Ph.D. thesis, Indian Statistical Institute, Calcutte, India.

[9] Von Mises, R., (1918), *Über die "Ganzzahligkeit" der Atmogewichte und Vervandte Fragen*, Phiysikal. Z., **19**, 490-500.