# شناسایی مشاهدات دورافتاده در مدل رگرسیونی خطی-دایره‌ای

سیده صدیقه عظیمی، محمدرضا فریدروحانی
دانشگاه شهید بهشتی، دانشکدهٔ علوم ریاضی

دریافت ۹۷/۰۳/۱۲ پذیرش ۹۷/۰۹/۱۲

## چکیده

یکی از راه‌های شناسایی مشاهدات دورافتاده در مدل‌های رگرسیونی، سنجش دوری مشاهدات از مقدار مورد انتظارشان تحت مدل برازش شده است. در مدل‌های رگرسیونی دایره‌ای-دایره‌ای، این شناسایی با استفاده از فاصلهٔ دایره‌ای امکان‌پذیر است. در این مقاله آمارهٔ اختلاف میانگین‌های خطای دایره‌ای که به‌وسیلهٔ ابوزید و همکاران [۱] برای شناسایی متغیر پاسخ دورافتاده در مدل رگرسیونی دایره‌ای-دایره‌ای ساده معرفی شده است، برای مدل رگرسیونی خطی-دایره‌ای به‌کار رفته و به‌روش شبیه‌سازی مونت‌کارلویی نقاط برینشی این آماره به‌دست آمده است. به‌علاوه با مطالعات شبیه‌سازی عملکرد این آماره بررسی شده است. در نهایت این آماره برای شناسایی پاسخ دورافتاده در دادهٔ سرعت و جهت باد ثبت شده در ایستگاه هواشناسی مهرآباد تهران به روش شبیه‌سازی خودگردان پارامتری به‌کار گرفته شده است.

**واژه‌های کلیدی:** مدل رگرسیونی خطی- دایره‌ای، مشاهدهٔ دورافتاده، آمارهٔ اختلاف میانگین‌های خطای دایره‌ای.

## مقدمه

از مسائلی که آمارشناسان در تحلیل‌های آماری به آن توجه می‌کنند، شناسایی مشاهداتی است که دور از انتظار هستند. مشاهدهٔ دورافتاده را می‌توان به‌عنوان نقطهٔ تنهایی که دور از دیگر مشاهدات است، تعریف کرد. یک مشاهدهٔ دورافتاده می‌تواند در اثر خطای اندازه‌گیری اتفاق بیافتد یا مقداری واقعی باشد که در پژوهش به‌دست آمده است[۲]. در هر دو حالت، شناسایی مشاهدهٔ دورافتاده اهمیت دارد زیرا اگر این مشاهده در اثر خطای اندازه‌گیری به‌دست آمده باشد، می‌توان آن را نادیده گرفت، اما اگر این مشاهده مقداری واقعی باشد، در این صورت شناسایی آن می‌تواند در تحلیل بهتر و پژوهش‌های آینده مفید واقع شود. بلسلی و همکاران [۳] با رویکرد حذف هر مشاهده در هر مرحله و بررسی تأثیر حذف آن مشاهده بر برآورد ضرایب مدل‌های رگرسیونی خطی به شناسایی مشاهدات دورافتاده پرداختند. هم‌چنین پژوهش‌های دیگری در این حوزه به‌وسیلهٔ بکمن و کوک [۴]، برنت و لوییز [۵] و مونت‌گومری و پک [۲] انجام شده است.

برای شناسایی مشاهدات دایره‌ای دورافتاده، با توجه به‌ویژگی تناوبی بودن این داده‌ها باید از معیارهایی استفاده کرد که این ویژگی لحاظ شود. اخیراً، پژوهش‌هایی برای شناسایی مشاهدات دورافتاده در مدل‌های رگرسیونی دایره‌ای-دایره‌ای انجام شده است. ابوزید و همکاران [۶] با تعریف جدیدی از مانده‌های دایره‌ای به شناسایی مشاهدات دورافتاده با استفاده از روش‌های نموداری و عددی در مدل رگرسیونی دایره‌ای-دایره‌ای ساده پرداختند. رامبلی و همکاران [۷] روش ارائه شده به‌وسیلهٔ ابوزید و همکاران [۶] را برای مدل رگرسیونی دایره‌ای- دایره‌ای که داونز و ماردیا [۸] معرفی کرده‌اند، تعمیم داده و به‌روش عددی به شناسایی مشاهدات مؤثر پرداختند. ابوزید و همکاران [۹] آمارهٔ COVRATIO را با رویکرد حذف هر مشاهده در هر مرحله و محاسبهٔ ماتریس کوواریانس ضرایب مدل معرفی کردند.

*نویسنده مسئول sazimi04@gmail.com

و به شناسایی مشاهدات دورافتاده با استفاده از روش عددی در مدل رگرسیونی دایره‌ای-دایره‌ای ساده پرداختند. ابراهیم و همکاران [۱۰] با استفاده از آمارهٔ COVRATIO و به‌روش عددی در مدل رگرسیونی دایره‌ای-دایره‌ای که به‌وسیلهٔ جمالامادکا و سارما [۱۱] معرفی شده است، شناسایی مشاهدات دورافتاده را بررسی کردند. ابوزید و همکاران [۱] با استفاده از آمارهٔ اختلاف میانگین‌های خطای دایره‌ای به شناسایی مشاهدات دورافتاده در مدل رگرسیونی دایره‌ای-دایره‌ای ساده پرداختند و با مطالعهٔ شبیه‌سازی نشان دادند این آماره برای شناسایی نقاط دورافتاده این مدل مناسب است. رامبلی و همکاران [۱۲] با استفاده از آمارهٔ COVRATIO در مدل رگرسیونی دایره‌ای-دایره‌ای که به‌وسیلهٔ داونز و ماردیا [۸] معرفی شده است، شناسایی مشاهدات دورافتاده را بررسی کردند.

با توجه به اهمیت شناسایی مشاهدات دورافتاده در مدل رگرسیونی خطی-دایره‌ای، در این مقاله به شناسایی مشاهدات دورافتاده با استفاده از آمارهٔ اختلاف میانگین‌های خطای دایره‌ای در مدل رگرسیونی خطی- دایره‌ای پرداختیم. بدین منظور در بخش دوم مدل رگرسیونی خطی- دایره‌ای و برآورد بیشینهٔ درست‌نمایی پارامترهای آن که به‌وسیلهٔ فیشر و لی [۱۵] معرفی شده است را مرور می‌کنیم. در بخش سوم با تعریف آمارهٔ اختلاف میانگین‌های خطای دایره‌ای که ابوزید و همکاران [۱] معرفی کرده‌اند، برخی از چندک‌های توزیع این آماره را با استفاده از روش مونت‌کارلویی برای مدل رگرسیونی خطی- دایره‌ای به‌دست آورده و توان این آماره را براساس تغییرات اندازهٔ نمونه و پارامتر تمرکز بررسی می‌کنیم. در نهایت در بخش آخر با استفاده از این آماره به جست‌وجوی مشاهدهٔ دورافتاده در یک مجموعه دادهٔ واقعی هواشناسی با روش خودگردان پارامتری می‌پردازیم.

## مدل رگرسیونی خطی – دایره‌ای

در برخی از حوزه‌های علمی مانند زیست‌شناسی و جغرافیا هدف بررسی رابطهٔ بین متغیرهای خطی و دایره‌ای است. برای مثال در بررسی رابطهٔ جهت حرکت و فاصلهٔ حرکت حیوانات از مبدا تا مقصد معینی و یا در بررسی رابطهٔ جهت باد و سرعت آن در ایستگاه‌های هواشناسی چنین رابطه‌ای وجود دارد. اولین بار مدل رگرسیونی خطی- دایره‌ای به‌وسیلهٔ گولد [۱۳] ارائه شد. جانسون و ورلی [۱۴] با نشان دادن شناساناپذیری مدل گولد، مدل جدیدی با یک متغیر پیشگو برای میانگین سویی پیشنهاد دادند که در آن متغیر پاسخ $\Theta$ به شرط متغیر پیشگوی $x$ دارای توزیع فون‌میزس با میانگین سویی $\mu+2\pi F(x)$ و پارامتر تمرکز $\kappa$ است که در آن $\mu$ و $F(x)$ به‌ترتیب عرض از مبدأ مدل و تابع توزیع تجمعی متغیر خطی $X$ هستند. فیشر و لی [۱۵] تعمیم مدل رگرسیونی جانسون و ورلی [۱۴] را برای بیش از یک متغیر پیشگو ارائه کردند.

فرض کنید $\Theta_i$، $i=1,\ldots,n$، متغیرهای تصادفی مستقل باشند که دارای توزیع فون‌میزس با پارامتر میانگین سویی $\mu$ و پارامتر تمرکز $\kappa$ هستند. فیشر و لی [۱۵] مدل رگرسیونی جانسون و ورلی [۱۴] را به‌صورت (۱) تعمیم دادند:

$$E(\Theta_i \mid x)=\mu_i=\mu+g\left(\beta' \underline{x}_i\right) \qquad (1)$$

که در آن $\underline{x}_i$ بردار $i$ام متغیرهای پیشگو و $\beta$ بردار ضرایب رگرسیونی با $k$ مؤلفه است. تابع $g(.)$ خط حقیقی را بر بازهٔ $[0,2\pi)$ تصویر می‌کند. بنابراین یک انتخاب برای تابع $g(x)$ می‌تواند تابع $\arctan(x)$ باشد. از این‌رو، مدل رگرسیونی خطی- دایره‌ای را می‌توان به‌صورت (۲) در نظر گرفت:

$$\theta_i=\mu+2\arctan\left(\beta' \underline{x}_i\right)+\varepsilon_i \qquad (2)$$

که در آن $\varepsilon_i$ دارای توزیع فون‌میزس با پارامتر میانگین سویی صفر و پارامتر تمرکز $\kappa$ است.

**۱. برآورد بیشینهٔ درست‌نمایی پارامترهای مدل (۱)**

فرض کنید $\theta_i$، $i=1,\ldots,n$، نمونه‌ای مستقل تحت مدل (۱) باشد که در آن $g$ تابعی معلوم است. لگاریتم تابع درست‌نمایی این نمونه متناسب است با

$$l \propto -n\log\left(I_0(\kappa)\right)+\kappa\sum_{i=1}^{n}\cos\left(\theta_i-\mu-g\left(\beta'\underline{x}_i\right)\right) \qquad (3)$$

که در آن $g\left(\beta'\underline{x}_i\right)=g\left(\sum_{j=1}^{k}\beta_j x_{ij}\right)$ و $I_0(\kappa)$ تابع بسل اصلاح شده نوع اول و از مرتبهٔ صفر است. برای $\mu$ و $\kappa$ ثابت از لگاریتم تابع درست‌نمایی (۳) نسبت به بردار $\beta$ مشتق می‌گیریم

$$\frac{\partial l}{\partial \beta}=\left[\frac{\partial g(\beta'\underline{x}_i)}{\partial \beta_j}\right]_{j\times i}\times u$$

که در آن

$$\frac{\partial g(\beta'\underline{x}_i)}{\partial \beta_j}=x_{ij}g'(\beta'\underline{x}_i),\quad i=1,\ldots,n,\quad j=1,\ldots,k$$

و

$$.u'=\left(u_1,\ldots,u_n\right),\quad u_i=\sin\left(\theta_i-\mu-g\left(\beta'\underline{x}_i\right)\right)$$

بنابراین برای یافتن برآورد بیشینهٔ درست‌نمایی پارامتر $\beta$ باید معادلات درست‌نمایی

$$\frac{\partial l}{\partial \beta}=\kappa X'Gu=\underline{0}$$

حل شود که در آن $X$ ماتریس متغیرهای پیشگو، $G_{n\times n}=diag\left(g'(\beta'\underline{x}_1),\ldots,g'(\beta'\underline{x}_n)\right)$ و $g'$ مشتق تابع $g$ است. از آن‌جا که حل این دستگاه به‌صورت عددی انجام می‌شود، می‌توان از روش تقریب تابع امتیاز برای حل آن استفاده کرد. در روش امتیاز فیشر از بسط تیلور $\frac{\partial l}{\partial \beta}$ تا مرتبهٔ دوم استفاده می‌شود. بر این اساس الگوریتم بازگشتی (۴)

$$\beta_{m+1}=\beta_m+\left(X'G^2X\right)^{-1}\left(X'G^2y\right),\quad m=0,1,2,\ldots \qquad (4)$$

برای تقریب زدن برآورد بیشینهٔ درست‌نمایی بردار $\beta$ به‌دست می‌آید که در آن

$$y'=\left(y_1,\ldots,y_n\right),\quad y_i=\frac{u_i}{A(\kappa)g'(\beta'\underline{x}_i)}$$

که $A(\kappa)=\frac{I_1(\kappa)}{I_0(\kappa)}$ و $I_1(\kappa)$ تابع بسل اصلاح شده نوع اول و از مرتبهٔ یک است. مجدداً برای $\kappa$ و $\beta$ ثابت و با مشتق‌گیری از لگاریتم تابع درست‌نمایی (۳) نسبت به $\mu$، برآورد بیشینهٔ درست‌نمایی پارامتر $\mu$ از روابط

$$\overline{R}\sin\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}\sin\theta_i,\quad \overline{R}\cos\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}\cos\theta_i \qquad (5)$$

به‌دست می‌آید یا به‌طور معادل $\hat{\mu}$ برابر میانگین سویی نمونه‌ای است. برآورد بیشینهٔ درست‌نمایی $\kappa$ با فرض ثابت بودن $\mu$ و $\kappa$ نیز عبارت است از:

$$\hat{\kappa}=A^{-1}\left(\overline{R}\right). \qquad (6)$$

بنابراین با آغاز از $\beta_0$ و $\mu_0$ به‌صورت تکراری و به‌ترتیب با استفاده از برابری‌های (۴)، (۵) و (۶) برآورد بیشینهٔ درست‌نمایی پارامترها به‌دست می‌آید.

ابوزید و همکاران [۱] برای شناسایی متغیر پاسخ دورافتاده در مدل رگرسیونی دایره‌ای-دایره‌ای ساده، با رویکرد حذف هر مشاهده آمارۀ میانگین خطای دایره‌ای که به اختصار با نماد MCE[1] نشان داده می‌شود را به‌صورت

$$MCE = ۱ - \frac{۱}{n}\sum_{i=۱}^{n}\cos\left(\theta_i - \hat{\theta}_i\right)$$

معرفی کردند که در آن $n$ اندازۀ نمونه و $\hat{\theta}_i$ برآورد $\theta_i$ از مدل رگرسیونی دایره‌ای-دایره‌ای است. توجه کنید که $MCE \in [۰,۲\pi)$. آمارۀ $MCE$ نوعی میانگین حسابی است که نسبت به‌وجود مشاهدۀ دورافتاده استوار نیست. به‌طوری‌که اگر مشاهدۀ $\theta_i$ دورافتاده باشد، انتظار می‌رود فاصلۀ دایره‌ای بین $\theta_i$ و $\hat{\theta}_i$ بزرگ شود. بنابراین وجود چنین مشاهده‌ای بین داده‌ها باعث افزایش مجموع فاصله‌های دایره‌ای و در نتیجه آمارۀ $MCE$ می‌شود. پس با حذف این مشاهده از مجموعۀ داده‌ها مقدار آمارۀ $MCE$ کاهش می‌یابد. اگر $MCE_{(-i)}$ آمارۀ میانگین خطای دایره‌ای با حذف مشاهدۀ $i$ام باشد، بیشینۀ قدر مطلق اختلاف بین آمارۀ $MCE$ و $MCE_{(-i)}$ به‌صورت

$$DMCE = \max_i \left\{ | MCE - MCE_{(-i)} | \right\}$$

یعنی اختلاف میانگین‌های خطای دایره‌ای تعریف می‌شود. اگر آمارۀ $DMCE$ [2] از نقطۀ برینشی توزیع خود تجاوز کند، آن‌گاه $j$ امین پاسخ دورافتاده است اگر

$$j = \arg\max_i \left\{ | MCE - MCE_{(-i)} | \right\}.$$

آمارۀ DMCE برای شناسایی مشاهدات دورافتاده در مدل‌های رگرسیونی که متغیر پاسخ آن‌ها دایره‌ای است، استفاده می‌شود. در این بخش آمارۀ $DMCE$ را برای شناسایی مشاهدات دورافتاده در مدل رگرسیونی خطی- دایره‌ای (۲) به‌کار می‌بریم. برای این منظور ابتدا باید نقاط برینشی توزیع این آماره محاسبه شود. این نقاط را با استفاده از شبیه‌سازی مونت‌کارلویی به‌دست آورده و سپس توان عملکرد آن در شناسایی مشاهدات دورافتاده بررسی می‌شود. هم‌چنین با استفاده از این آماره الگوریتمی برای شناسایی مشاهدات دورافتاده در داده‌های واقعی با استفاده از روش خودگردان پارامتری ارائه می‌شود.

## ۱. تعیین نقاط برینشی آمارۀ میانگین خطای دایره‌ای

فرض کنید نمونۀ تصادفی $(\theta_۱, x_۱),\ldots,(\theta_n, x_n)$ از متغیرهای تصادفی $(\Theta, X)$ در اختیار باشد که در آن $X$ و $\Theta$ به‌ترتیب متغیرهای تصادفی خطی و دایره‌ای هستند. مدل رگرسیونی خطی- دایره‌ای (۲) را در نظر بگیرید. در ادامه ابتدا نقاط برینشی آمارۀ $DMCE$ را با روش شبیه‌سازی مونت‌کارلویی به‌دست می‌آوریم. این بررسی به‌ازای ۱۰ اندازه نمونۀ $n = 10, 20, 30, 40, 50, 70, \ldots, 150$ و پارامترهای تمرکز $\kappa = 1, 2, 5, 7, 10$ اجرا می‌شود. مدل رگرسیونی خطی- دایره‌ای (۲) یک مدل رگرسیونی چندگانه است. در این پژوهش، شبیه‌سازی را فقط با یک متغیر پیشگو انجام داده‌ایم. الگوریتم شبیه‌سازی بدین‌صورت است:

**الگوریتم ۱ (شبیه‌سازی نقاط برینشی به روش مونت‌کارلویی)**

۱. برای هر اندازه نمونۀ $n$، خطاهای تصادفی دایره‌ای از توزیع فون‌میزس با میانگین صفر و پارامتر تمرکز معین $\kappa_۰$ را تولید کنید،

1. Mean Circular Error
2. Difference of Means Circular Error

۲. از توزیع نرمال با میانگین صفر و انحراف معیار دو برای متغیر تصادفی خطی $X$ نیز نمونه‌ای تصادفی به‌اندازهٔ $n$ تولید کنید،

۳. پارامترهای مدل رگرسیونی خطی- دایره‌ای (۲) را $\mu=\mu_\circ$ و $\beta=\beta_\circ$ در نظر بگیرید که $\mu_\circ$ و $\beta_\circ$ معلوم هستند. مقدار مشاهدات $\theta_i$ را با قرار دادن مشاهدات تولید شدهٔ $(x_i, e_i)$ در گام یک و دو از مدل (۲) به‌دست آورید.

۴. بر اساس $(\theta_i, x_i)$های تولید شده در گام دو و سه، برآوردهای بیشینهٔ درست‌نمایی پارامترها و در نتیجه مقدار برازش داده شدهٔ $\hat{\theta}_i$را به‌دست آورید،

۵. آمارهٔ $MCE$ را از کل مشاهدات به‌دست آورید. سپس با استفاده از مدل رگرسیونی (۲) با حذف $i$امین مشاهده $(i=۱,\ldots,n)$ آمارهٔ $MCE_{(-i)}$ را محاسبه کرده و در نهایت آماره $DMCE$ را به‌دست آورید.

۶. با تکرار گام‌های یک تا پنج به تعداد از پیش تعیین شده M و براساس DMCEهای تولید شده در گام پنجم چندک $(۱-\alpha)$ام را به‌صورت $q=\left[M\left(۱-\alpha\right)\right]$ به‌دست آورید.

اینک الگوریتم ۱ را M=۲۰۰۰ بار برای هر ترکیب معین $n$و $\kappa$ تکرار می‌کنیم. نتیجه این شبیه‌سازی یعنی مقدار چندک‌های ۹۰٪، ۹۵٪ و ۹۹٪ برای ترکیب‌های $n$ و $\kappa$ به‌ازای $\mu_\circ=۰$ و $\beta_\circ=۱$ در جدول ۱ آمده است.

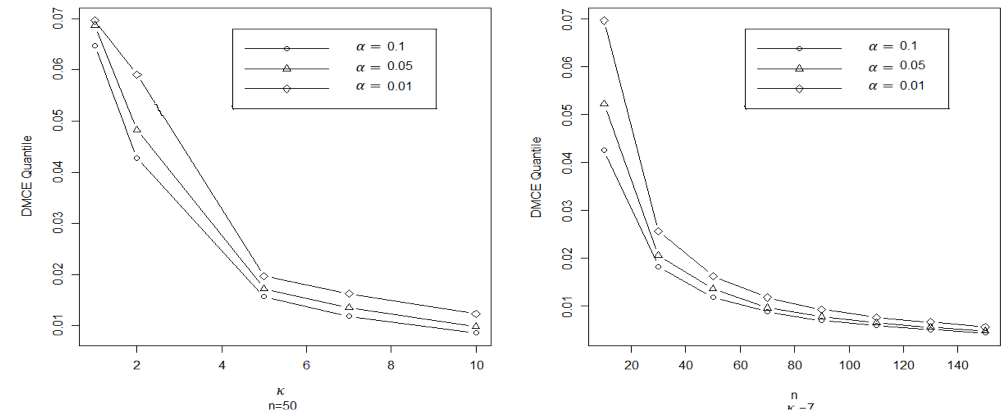
**جدول ۱. چندک‌های شبیه‌سازی شدهٔ توزیع آمارهٔ $DMCE$**

| $\kappa$ | ۱ | | | ۲ | | | ۵ | | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | ٪۹۰ | ٪۹۵ | ٪۹۹ | ٪۹۰ | ٪۹۵ | ٪۹۹ | ٪۹۰ | ٪۹۵ | ٪۹۹ |
| ۱۰ | ۰/۱۲۴۵ | ۰/۱۳۵۸ | ۰/۱۵۷۹ | ۰/۱۱۶۰ | ۰/۱۱۹۰ | ۰/۱۳۴۰ | ۰/۰۶۲۹ | ۰/۰۷۴۴ | ۰/۱۰۳۳ |
| ۲۰ | /۰۸۵۳ | ۰/۰۹۱۸ | ۰/۱۰۲۳ | ۰/۰۵۹۱ | ۰/۰۶۹۹ | ۰/۰۸۹۴ | ۰/۰۳۵۲ | ۰/۰۴۰۷ | ۰/۰۵۰۵ |
| ۳۰ | ۰/۰۷۵۲ | ۰/۰۸۰۹ | ۰/۰۹۱۲ | ۰/۰۵۰۸ | ۰/۰۵۷۸ | ۰/۰۷۲۵ | ۰/۰۲۴۵ | ۰/۰۲۷۱ | ۰/۰۳۲۷ |
| ۴۰ | ۰/۰۶۷۰ | ۰/۰۷۱۳ | ۰/۰۸۲۱ | ۰/۰۴۴۸ | ۰/۰۵۰۱ | ۰/۰۶۳۲ | ۰/۰۱۹۴ | ۰/۰۲۱۹ | ۰/۰۲۵۱ |
| ۵۰ | ۰/۰۶۴۷ | ۰/۰۶۸۷ | ۰/۰۶۹۶ | ۰/۰۴۲۸ | ۰/۰۴۸۲ | ۰/۰۵۹۰ | ۰/۰۱۵۷ | ۰/۰۱۷۲ | ۰/۰۱۹۷ |
| ۷۰ | ۰/۰۵۸۹ | ۰/۰۶۲۸ | ۰/۰۶۸۸ | ۰/۰۳۸۶ | ۰/۰۴۲۸ | ۰/۰۵۱۰ | ۰/۰۱۱۲ | ۰/۰۱۲۶ | ۰/۰۱۴۱ |
| ۹۰ | ۰/۰۵۶۰ | ۰/۰۵۹ | ۰/۰۶۳۷ | ۰/۰۳۶۱ | ۰/۰۴۰۰ | ۰/۰۴۶۲ | ۰/۰۰۸۷ | ۰/۰۰۹۷ | ۰/۰۱۰۸ |
| ۱۱۰ | ۰/۰۵۳۴ | ۰/۰۵۶۰ | ۰/۰۶۰۹ | ۰/۰۳۴۹ | ۰/۰۳۸۰ | ۰/۰۴۲۵ | ۰/۰۰۷۲ | ۰/۰۰۷۷ | ۰/۰۰۸۸ |
| ۱۳۰ | ۰/۰۵۲۵ | ۰/۰۵۴۵ | ۰/۰۵۹۹ | ۰/۰۳۴۰ | ۰/۰۳۶۴ | ۰/۰۴۰۷ | ۰/۰۰۶۰ | ۰/۰۰۶۵ | ۰/۰۰۷۳ |
| ۱۵۰ | ۰/۰۵۱۴ | ۰/۰۵۳۸ | ۰/۰۵۸۵ | ۰/۰۳۲۸ | ۰/۰۳۵۰ | ۰/۰۴۰۴ | ۰/۰۰۵۲ | ۰/۰۰۵۶ | ۰/۰۰۷۰ |

**ادامهٔ جدول ۱**

| $\kappa$ | ۷ | | | ۱۰ | | |
|---|---|---|---|---|---|---|
| $n$ | ٪۹۰ | ٪۹۵ | ٪۹۹ | ٪۹۰ | ٪۹۵ | ٪۹۹ |
| ۱۰ | ۰/۰۴۲۶ | ۰/۰۵۲۱ | ۰/۰۶۹۵ | ۰/۰۳۱۳ | ۰/۰۳۸۲ | ۰/۰۸۳۵ |
| ۲۰ | ۰/۰۲۵۹ | ۰/۰۳۰۳ | ۰/۰۳۸۷ | /۰۱۷۱۰ | ۰/۰۲۰۶ | ۰/۰۲۷۶ |
| ۳۰ | ۰/۰۱۸۲ | ۰/۰۲۰۶ | ۰/۰۲۵۶ | ۰/۰۱۳۰ | ۰/۰۱۵۱ | ۰/۰۲۰۰ |
| ۴۰ | ۰/۰۱۴۰ | ۰/۰۱۶۲ | ۰/۰۲۰۵ | ۰/۰۱۰۳ | ۰/۰۱۱۷ | ۰/۰۱۵۲ |
| ۵۰ | ۰/۰۱۱۸ | ۰/۰۱۳۶ | ۰/۰۱۶۳ | ۰/۰۰۸۶ | ۰/۰۰۹۹ | ۰/۰۱۲۳ |
| ۷۰ | ۰/۰۰۸۷ | ۰/۰۰۹۶ | ۰/۰۱۱۷ | ۰/۰۰۶۵ | ۰/۰۰۷۳ | ۰/۰۰۹۲ |
| ۹۰ | ۰/۰۰۶۹ | ۰/۰۰۷۷ | ۰/۰۰۹۳ | ۰/۰۰۵۲ | ۰/۰۰۶۰ | ۰/۰۰۷۲ |
| ۱۱۰ | ۰/۰۰۵۸ | ۰/۰۰۶۵ | ۰/۰۰۷۶ | ۰/۰۰۴۳ | ۰/۰۰۴۹ | ۰/۰۰۶۳ |
| ۱۳۰ | ۰/۰۰۵۱ | ۰/۰۰۵۶ | ۰/۰۰۶۶ | ۰/۰۰۳۷ | ۰/۰۰۴۲ | ۰/۰۰۵۵ |
| ۱۵۰ | ۰/۰۰۴۴ | ۰/۰۰۴۸ | ۰/۰۰۵۶ | ۰/۰۰۳۳ | ۰/۰۰۳۶ | ۰/۰۰۴۴ |

شکل ۱ چندک‌های ۰/۹۰، ۰/۹۵ و ۰/۹۹ آمارهٔ $DMCE$ را به‌ازای دو حالت $\kappa=۷$ و $n$ های مختلف و $n=۵۰$ و $\kappa$ های مختلف نشان می‌دهد. چنان‌که ملاحظه می‌شود چندک‌های این آماره برای $\kappa(n)$ ثابت تابع کاهشی نسبت $n(\kappa)$ است.




**شکل ۱. نمودار چندک‌های ۰/۹۰، ۰/۹۵ و ۰/۹۹ آمارهٔ $DMCE$ راست: به‌ازای $\kappa=۷$ و $n$ های مختلف چپ: به‌ازای $n=۵۰$ و $\kappa$ های مختلف**

از آن‌جاکه در تحلیل داده‌های واقعی استفاده از الگوریتم ۱ برای یافتن نقاط برینشی منطقی به‌نظر نمی‌رسد، ضروری است پس از برازش مدل رگرسیونی خطی-دایره‌ای به داده‌های واقعی از روش خودگردان پارامتری برای این منظور بهره گرفت. در ادامه چگونگی یافتن نقاط برینشی در تحلیل داده‌های واقعی، بر اساس الگوریتم زیر بیان می‌شود:

**الگوریتم ۲ (شبیه‌سازی نقاط برینشی براساس دادهٔ واقعی به روش خودگردان پارامتری)**

۱. مدل رگرسیونی خطی-دایره‌ای (۲) را به مجموعهٔ داده‌ها برازش دهید و پارامترهای برآورد شده این مدل را $\hat{\mu}$، $\hat{\beta}$ و $\hat{\kappa}$ بنامید،

۲. نمونهٔ به اندازهٔ $n$، از خطاهای تصادفی دایره‌ای از توزیع فون‌میزس با میانگین صفر و پارامتر تمرکز $\hat{\kappa}$ تولید کنید،

۳. براساس نمونه‌های تولید شده از گام دو، $\theta_i$ ها را با استفاده از مدل ۲ و پارامترهای برآورد شده در گام اول تولید کنید،

۴. براساس متغیر پیشگو $x_i$ و $\theta_i$ های تولید شده در گام سه، برآوردهای بیشینهٔ درست‌نمایی پارامترها و در نتیجه مقدار برازش داده شدهٔ $\hat{\theta}_i$ را به‌دست آورید،

۵. آمارهٔ $MCE$ را از کل مشاهدات به‌دست آورید. سپس با استفاده از مدل رگرسیونی (۲) با حذف $i$ امین مشاهده $(i=۱,\dots,n)$ آمارهٔ $MCE_{(-i)}$ را محاسبه کرده و در نهایت آماره $DMCE$ را به‌دست آورید،

۶. با تکرار گام‌های دو تا پنج به تعداد از پیش تعیین شده B و براساس DMCEهای تولید شده در گام ششم چندک $(۱-\alpha)$ام را به‌صورت $q=[B(۱-\alpha)]$ به‌دست آورید.

در بخش پنجم به‌منظور شناسایی مشاهدات دورافتاده در مجموعه دادهٔ واقعی هواشناسی، از این الگوریتم برای یافتن نقاط برینشی آمارهٔ DMCE استفاده شده است.

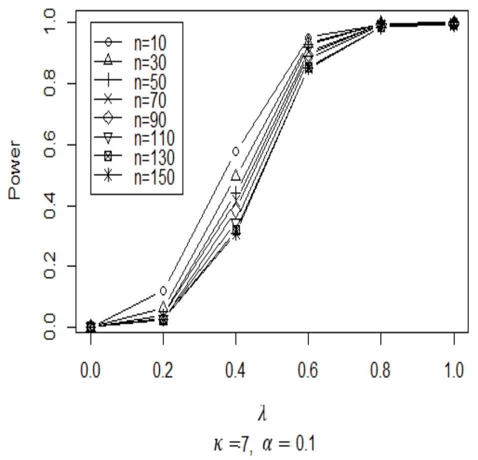
## ۲. بررسی توان آمارهٔ میانگین خطای دایره‌ای

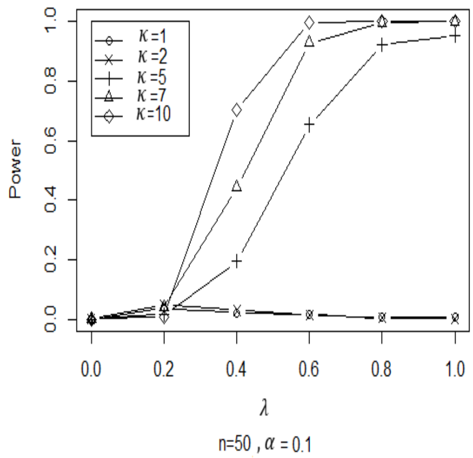
برای بررسی توان آمارهٔ DMCE تحت مدل رگرسیونی خطی- دایره‌ای (۲)، ابتدا $n$ مشاهده از $(\Theta, X)$ تحت مدل (۲) با اجرای گام‌های یک تا چهار الگوریتم ۱ تولید می‌کنیم. سپس از میان آنها، $m$ مشاهده، $۱\le m\le n$، را تحت مدل

$$\theta_d^* = \theta_d + \lambda\pi(\text{mod}\,۲\pi), \quad d = ۱,۲,\ldots,m$$

آلوده می‌کنیم. $\lambda$ میزان آلودگی مشاهدهٔ $d$ام است، به‌طوری‌که $۰ \leq \lambda \leq ۱$. اگر $\lambda = ۰$، در این صورت در موقعیت $d$ام یک مشاهدهٔ غیر دورافتاده تولید می‌شود. در حالی که $\lambda = ۱$، $\theta_d^*$ در مکان پادمد $\theta_d$ قرار می‌گیرد. اینک مقدار آمارهٔ $DMCE$ را با استفاده از الگوریتم ۱، به‌دست می‌آوریم. با ۲۰۰۰ بار تکرار این فرآیند، عملکرد آمارهٔ $DMCE$ را با معیار درصد شناسایی درست مشاهدهٔ دورافتاده در ۲۰۰۰ بار تکرار بررسی می‌شود.

توان عملکرد این آماره را برای $m = ۱$ بررسی کرده‌ایم. نتایج برای $n = ۵۰$ و $\kappa = ۱,۲,۵,۷,۱۰$ در شکل ۲ (راست) نشان داده شده است. توان این آماره برای مقدارهای کوچک $\kappa$، برای مثال $\kappa < ۲$، نزدیک صفر است. اما به‌طور کلی برای مقدارهای بزرگ $\kappa$ توان عملکرد این آماره با افزایش $\lambda$ افزایش می‌یابد که این نتیجه با توجه به افزایش اختلاف مقدار مشاهده شده و مقدار برازش شده (با افزایش $\lambda$) قابل انتظار است. برای $\lambda < ۰/۲$ توان عملکرد پایین است و برای مقدارهای بزرگ‌تر $\lambda$ توان عملکرد آماره افزایش می‌یابد. به‌طورکلی برای مقدار $\lambda > ۰/۶$ توان عملکرد این آماره حداقل به سطح ۶۰٪ (به‌ازای $\kappa = ۵$) و حداکثر به سطح ۹۹/۹٪ (به‌ازای $\kappa = ۱۰$) می‌رسد.




**شکل ۲. نمودار توان چندک $\alpha = ۰/۱$ام آمارهٔ $DMCE$ نسبت به سطح آلودگی $\lambda$ راست: به‌ازای $n = ۵۰$ و $\kappa = ۱,۲,۵,۷,۱۰$ چپ: به‌ازای مقادیر مختلف $n$ و $\kappa = ۷$**

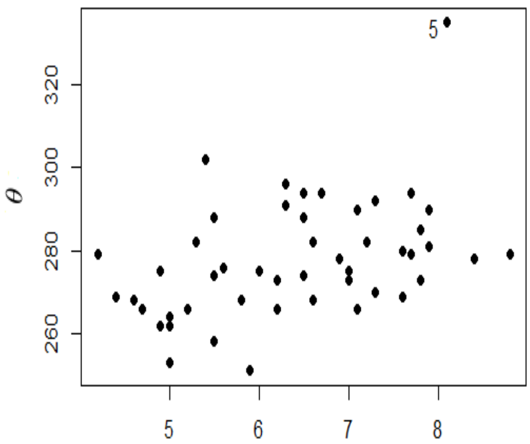
در شکل ۲ (چپ) توان عملکرد آماره به‌ازای $\kappa = ۷$ و مقدارهای مختلف $n$ نشان داده شده است. چنان‌که ملاحظه می‌شود نتایج این شکل نیز مشابه نتایج شکل ۲ (راست) است. برای $\lambda < ۰/۲$ توان عملکرد پایین است (حداکثر ۰/۱)، حال آن‌که مقدارهای بزرگ‌تر $\lambda$ توان عملکرد آماره را افزایش می‌دهد. چنان‌که در این شکل ملاحظه می‌شود با افزایش $n$ به‌ازای یک $\lambda$ ثابت و معلوم و $m = ۱$ توان عملکرد آماره کاهش می‌یابد. این رفتار ناشی از این واقعیت است که با افزایش $n$، سهم تک مشاهدهٔ آلوده از کل مشاهدات کاهش یافته و از این‌رو، تأثیر این مشاهده در کل مشاهدات کم می‌شود.

## جست‌وجوی نقاط دورافتاده در داده‌های هواشناسی

ایستگاه هواشناسی مهرآباد شهر تهران در عرض جغرافیایی ۳۵۴۱ شمالی و طول جغرافیایی ۵۱۱۹ شرقی و در ارتفاع ۱۱۹۰/۸ متر از سطح دریا واقع شده است. این ایستگاه به‌عنوان ایستگاه مبنای شهر تهران شناخته شده است. در بررسی‌های هواشناسی به‌نظر می‌رسد سرعت باد بر جهت باد تأثیر داشته باشد. از این‌رو، میانگین جهت باد و میانگین سرعت باد در ماه مارس طی سال‌های ۱۹۵۱ تا ۲۰۰۰ میلادی (۵۰ مشاهدهٔ مستقل) را به‌ترتیب به واحدهای درجه و گره (هر گره معادل ۱۸۵۲ متر) در این ایستگاه در نظر گرفته‌ایم. این داده‌ها در بانک اطلاعات داده‌های

هواشناسی به آدرس اینترنتی http://www.iranhydrology.net/meteo/meteo.htm در دسترس است. میانگین جهت باد یک متغیر دایره‌ای و میانگین سرعت باد یک متغیر خطی است که به‌ترتیب با نمادهای $\theta \in [۰,۳۶۰^\circ)$ و $x \in \mathbb{R}$ نمایش می‌دهیم. در این بخش هدف شناسایی مشاهدات دورافتاده در این مجموعه داده است.

با رسم نمودار پراکنش $x$ در مقابل $\theta$ (شکل ۳) به‌نظر می‌رسد بین این دو متغیر رابطه وجود دارد. هم‌چنین به نظر می‌رسد مشاهده پنجم از سایر مجموعه مشاهدات میانگین جهت باد دورتر است و احتمالاً مشاهدهٔ دورافتاده است. بنابراین مدل (۲) را به این مجموعه داده برازش می‌دهیم و وجود مشاهدهٔ دورافتاده در این مجموعه داده را با استفاده از آمارهٔ $DMCE$ بررسی می‌کنیم.




**شکل ۳. نمودار پراکنش متغیر پاسخ دایره‌ای $(\theta)$ در مقابل متغیر پیشگوی خطی $(x)$**

ابتدا با استفاده از آمارهٔ واتسون، که واتسون [۱۶] معرفی کرده است، فرض صفر دارای توزیع فون‌میزس بودن متغیر $\theta$ را می‌آزماییم. چون مقدار آمارهٔ این آزمون $W = ۰/۰۸۰۶$ از نقطه بحرانی در سطح پنج درصد (۰/۱۱۷) کوچک‌تر است، فرض این‌که متغیر $\theta$ دارای توزیع فون‌میزس است، معنی‌دار است.

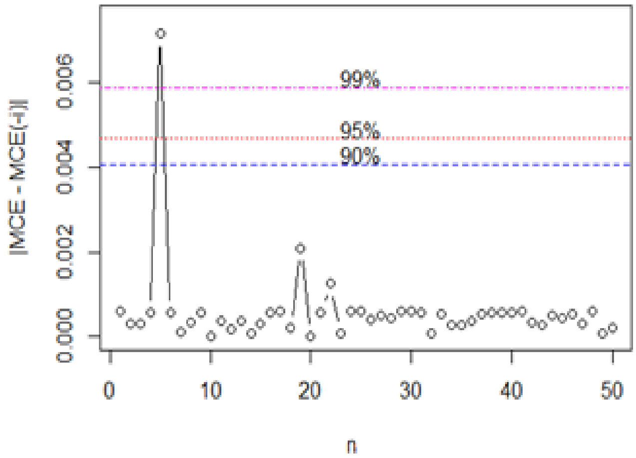
با برازش مدل (۲) به داده‌ها، برآورد بیشینهٔ درست‌نمایی پارامترهای مدل برابر است با:

$$\hat{\mu} = ۲۴۵/۸۶۶۸^\circ, \quad \hat{\beta} = ۰/۰۴۴۲, \quad \hat{\kappa} = ۲۱/۲۵.$$

اینک با استفاده از آمارهٔ $DMCE$ به بررسی وجود مشاهدهٔ دورافتاده در این مجموعه داده می‌پردازیم. در شکل ۴ مقادیر $|MCE - MCE_{(-i)}|$ به‌ازای $i = ۱,\dots,۵۰$ رسم شده است. هم‌چنین با B=۵۰۰ بار تکرار الگوریتم ۲ چندک‌های تجربی ۹۰، ۹۵ و ۹۹ درصد توزیع آمارهٔ $DMCE$ تحت مدل رگرسیونی

$$\theta_i = ۲۴۵/۸۶۶۸ + ۲\arctan\left(۰/۰۴۴۲ x_i\right) + \varepsilon_i$$

که در آن $\varepsilon_i$ دارای توزیع فون‌میزس با پارامتر میانگین سویی صفر و پارامتر تمرکز ۲۱/۲۵ است، به‌ترتیب برابر ۰/۰۰۴۱، ۰/۰۰۴۷ و ۰/۰۰۵۹ به‌دست آمده است. چنان‌که در شکل ۴ ملاحظه می‌شود مقدار آمارهٔ $|MCE - MCE_{(-۵)}|$ از چندک‌های آمارهٔ $DMCE$ بزرگ‌تر است. بنابراین با استفاده از این آماره مشاهدهٔ پنجم یک نقطهٔ دورافتاده است.این نتیجه با ملاحظهٔ نمودار پراکنش $x$ در مقابل $\theta_i$ در شکل ۳ نیز منطقی به‌نظر می‌رسد.




**شکل ۴. آمارۀ $DMCE$ برای متغیر پاسخ $\theta$ و چندک‌های توزیع این آماره برای دادۀ هواشناسی**

## بحث و نتیجه‌گیری

در این مقاله نقاط برینشی آمارۀ $DMCE$ در یک مدل رگرسیونی خطی- دایره‌ای را به‌روش شبیه‌سازی مونت‌کارلویی به‌دست آوردیم. این نقاط با فرض ثابت بودن $\kappa(n)$، نسبت به $n(\kappa)$ کاهشی بودند. هم‌چنین در پژوهش‌های شبیه‌سازی بررسی شد که توان این آماره برای مقدارهای بزرگ $\lambda$ (میزان آلودگی) به‌ازای مقادیر مختلف $n$ و مقادیر بزرگ $\kappa$ نزدیک یک به‌دست آمد. با استفاده از آمارۀ مزبور برای یک مجموعه از دادۀ واقعی، مشاهده‌ای دورافتاده شناسایی شد که با توجه به نمودار پراکنش داده‌ها، تشخیص موجهی به نظر می‌رسد از این‌رو، عملکرد این آماره در مواجهه با داده‌های واقعی نیز مطلوب است. به‌عنوان یک موضوع پژوهشی می‌توان آمارۀ COVRATIO را نیز برای شناسایی نقاط دورافتاده در مدل رگرسیونی خطی- دایره‌ای بررسی کرد و توان آن را با آمارۀ DMCE مقایسه کرد.

## منابع

1. Abuzaid A. H., Hussin A. G., Mohamed I. B., "Detection of outliers in simple circular regression models using the mean circular error statistic", Journal of Statistical Computation and Simulation, 83:2 (2013) 269-277.
2. Montgomery D. C., Peck E. A., "Introduction to linear regression analysis", 2nd Edition, New York: Wiley (1992).
3. Belsley D.A., Kuh E., Welsch R.E., "Regression Diagnostics: Identifying Influential Data and Sources of Collinear it", New York: John Wiley & Sons (1980).
4. Beckman R. J., Cook R. D., "Outliers", Techno metrics, 25 (2) (1983)119-149.
5. Barnett V., Lewis T., "Outliers in statistical data", New York: John Wiley & Sons (1984).
6. Abuzaid A. H., Hussin A. G., Mohamed I. B., "Identifying single outlier in linear circular regression model based on circular distance", Journal of Applied Probability and Statistics, 3 (1) (2008) 107-117.

7. Rambli A. B., Mohamed I., Abuzaid A. H., Hussin, A. G., "Identification of Influential Observations in Circular Regression Model", Proceedings of the Regional Conference on Statistical Sciences (RCSS'10) (2010) 195-203.

8. Downs T. D., Mardia K.V., "Circular regression", Biometrika, 89(3) (2002) 683-697.

9. Abuzaid A. H., Mohamed I. B., Hussin A. G., Rambli A., "Covratio statistic for simple circular regression model", Chiang Mai J. Sci., 38(3) (2011) 321-330.

10. Ibrahim S., Rambli A., Hussin A.G., Mohamed I., "Outlier detection in a circular regression model using COVRATIO statistic", Communications in Statistics-Simulation and Computation, 42 (10) (2013) 2272-2280.

11. Jammalamadaka S. R., Sarma Y. R., "Circular regression. In Statistical Sciences and Data Analysis", edited by Matusita, K. Utrecht: VSP. (1993) 109-128.

12. Rambli A., Yunus R. M., Mohamed I., Hussin A. G., "Outlier Detection in a Circular Regression Model", Sains Malaysiana, 44 (7) (2015) 1027-1032

13. Gould A. L., "A regression technique for angular data", Biometrics,25 (1969) 683-700.

14. Johnson R. A., Wehrly T. E., "Some angular-linear distributions and related regression models", Journal of the American Statistical Association, 73 (1987) 602-606.

15. Fisher N. I., Lee A. J., "Regression models for an angular response", Biometrics, 48 (1992) 665-677.

16. Watson G. S., "Goodness-of-fit tests on the circle, Biometrics", 48 (1961)109-114.