

## آزمون تعیین خطای مشخص سازی مدل خطی عام با داده های گم شده

فیاض بهاری\*، صفر پارسی\*<sup>۱</sup> و مجتبی گنجعلی\*\*

\*گروه آمار و علوم کامپیوتر، دانشگاه محقق اردبیلی

\*\*گروه آمار، دانشگاه شهید بهشتی

تاریخ دریافت: تاریخ پذیرش:

**چکیده:** در این مقاله مدل خطی عام را در تحلیل داده هایی که در آن ها متغیرهای کمکی و متغیر پاسخ گم شدهگی دارند، در نظر می گیریم. برای تعیین نیکویی برازش مدل خطی عام در حضور داده های گم شده، آزمون جدیدی را بر اساس آزمون رمزی می سازیم. نشان می دهیم که تحت فرض صفر، آماره های آزمون در برخی از حالات از توزیع فیشر تبعیت می کند و در برخی از حالات به توزیع شبه-فیشر میل می کند. علاوه بر این، عملکرد آماره های آزمون های مختلف را بر اساس چند مسئله شبیه سازی مقایسه می کنیم. همچنین از این آماره ها برای بررسی کفایت مدل خطی عام در تحلیل داده های واقعی استفاده می کنیم.

**واژه های کلیدی:** مدل خطی عام، داده های گم شده، آزمون نیکویی برازش، توزیع شبه-فیشر، آزمون رمزی.

رده بندی موضوعی (۲۰۱۰): ۶۲J۰۲، ۶۲F۰۳.

### ۱-مقدمه

مدل خطی عام<sup>۲</sup> یکی از روش های متداول در تعیین رابطه بین پدیده ها است. یک مدل خطی عام بیانگر رابطه بین متغیر پاسخ  $y$  و بردار متغیرهای کمکی  $x$  با بعد  $m$  است. مدل خطی عام از رابطه زیر پیروی می کند.

$$y = g^T(\mathbf{x})\boldsymbol{\beta} + \varepsilon, \quad (1)$$

۱- آدرس الکترونیکی مسئول مقاله: [parsi@uma.ac.ir](mailto:parsi@uma.ac.ir)

که در آن  $g(\cdot)$  یک تابع برداری معلوم با بعد  $p$ ،  $\beta$  یک پارامتر برداری با بعد  $p$  و  $\varepsilon$  خطای نمونه‌گیری با واریانس ثابت است. همچنین فرض بر این است که  $E(\varepsilon | \mathbf{x}) = 0$  و  $E(\varepsilon^2 | \mathbf{x}) < \infty$ . مدل خطی ساده حالت خاصی از مدل خطی عام است؛ به طوری که وقتی  $g(\mathbf{x}) = x$ ، مدل خطی ساده نتیجه می‌شود. به همین دلیل مدل خطی عام از مدل خطی ساده، انعطاف‌پذیرتر است و به ما اجازه استنباط در مورد مدل‌های پیچیده‌تر را می‌دهد. برای ملاحظه جزئیات بیشتر در این زمینه به [۱] مراجعه شود.

یک مشکل اساسی در مطالعه مدل‌های خطی، کفایت مدل برازش‌شده است. بر همین اساس [۲] آزمون تعیین خطای مشخص‌سازی مدل رگرسیونی و یا به عبارت دقیق‌تر آزمون تعیین خطای معادله رگرسیونی<sup>۱</sup> (RESET) را معرفی کرد که یک ابزار مناسب برای تشخیص عدم کفایت مدل است. در برخی از مطالعات ممکن است، برخی از متغیرهای کمکی کنار گذاشته شوند. این امر باعث می‌شود تا یک اریبی حذف متغیر<sup>۲</sup> (OVB) به وجود آید که باعث می‌شود متغیر پاسخ به‌درستی پیش‌بینی نشود. آزمون رمزی کمک می‌کند تا اثر متغیرهای حذف‌شده در مدل را تشخیص دهیم. به همین دلیل، آزمون رمزی، به‌ویژه در مباحث اقتصادسنجی برای تحلیل داده‌های فضایی<sup>۳</sup>، کاربرد فراوانی پیدا کرده است، به طوری که در متون اقتصادسنجی آزمون رمزی به آزمون تعیین خطای مشخص‌سازی مدل معروف است. به‌عنوان نمونه، اخیراً [۳] از آزمون رمزی برای تشخیص چگونگی تأثیر متغیرهای فضایی بر OVB استفاده کردند. عملکرد خوب آزمون رمزی توسط بسیاری از محققین از جمله [۴] مورد بررسی قرار گرفت. نتایج بررسی آن‌ها نشان می‌دهد که با افزایش عدم کفایت مدل، احتمال رد مدل برازش‌شده افزایش می‌یابد. در واقع، کاربرد واقعی آزمون رمزی در تشخیص شکل تابعی مدل خطی برازش‌شده است. این آزمون کفایت مدل برازش‌شده را می‌سنجد. در همین راستا، [۵] آزمون رمزی را به مدل‌های خطی تعمیم‌یافته<sup>۴</sup> بسط داد. نتایج خوب آزمون رمزی برای مدل خطی تعمیم‌یافته با استفاده از شبیه‌سازی در [۵] به‌خوبی نمایان است.

در این مقاله هدف ما بررسی کفایت مدل‌های خطی عام با استفاده از راهکار رمزی است. در این راستا، مشکل اساسی دیگر مطالعات آماری، یعنی حضور داده‌های گم‌شده<sup>۵</sup> در مطالعات را در نظر می‌گیریم. برازش مدل خطی عام بدون در نظر گرفتن اثر داده‌های گم‌شده، ممکن است منجر به برآوردهای اریب شود که باعث می‌شود از کفایت مدل خطی عام کاسته شود. با استفاده از

1- Regression Equation Specification Error Test

2- Omission-Variable Bias

3-Spatial Data

4- Generalized Linear Models

5 -Missing Data

روش‌های پیشرفته آماری و ابزار پیشرفته محاسباتی، روش‌هایی برای مواجهه با این مشکل مطرح شد. با دسته‌بندی الگوی داده‌های گم‌شده به سه گروه توسط [۶] کار با داده‌های گم‌شده شکل ساده‌تری به خود گرفت. در این مقاله فرض بر این است که داده‌ها از الگوی قابل چشم‌پوشی<sup>۱</sup> پیروی می‌کنند. پس فرض بر این است که یا الگوی گم‌شدگی داده‌ها، بر مبنای تعریف [۷]، به صورت کاملاً تصادفی<sup>۲</sup> (MCAR) و یا به صورت تصادفی<sup>۳</sup> (MAR) است. همچنین، فرض بر این است که داده‌های گم‌شده می‌توانند هم در متغیرهای کمکی و هم در متغیر پاسخ رخ دهند. در [۸] روش‌های جهانی داده‌های گم‌شده مورد مطالعه قرار گرفته است و روش‌های برآورد داده‌ها بحث شده است. اما اخیراً توجه بیشتر محققین بر تصحیح معادلات برآورد است تا بتوان با تصحیح معادلات برآورد، برآوردگرهای ناریب و سازگار را به دست آورد. برای مطالعه بیشتر در این زمینه به [۹] و [۱۰] مراجعه شود.

در مطالعات بدون داده گم‌شده [۱۱] و [۱۲] آزمون نیکویی برازش مدل خطی عام را به ترتیب با استفاده از روش‌های پارامتری و ناپارامتری انجام دادند. در حالتی که متغیر پاسخ با خطا اندازه‌گیری شده باشد، [۱۳] آزمون نوع امتیاز<sup>۴</sup> را برای بررسی کفایت مدل خطی عام به کار برد. در حالتی که الگوی گم‌شدگی داده‌ها قابل چشم‌پوشی باشد، برخی از محققین راهکارهایی معرفی کردند. از جمله، [۱۴] یک آزمون مناسب دیگر بر اساس تابع امتیاز معرفی کرد. به علاوه، [۱۵] روش آزمون مناسبی را بر اساس انتگرال کمترین توان دوم فاصله بین برازش‌های ناپارامتری و پارامتری<sup>۵</sup> ساخت، به طوری که داده‌های گم‌شده با یک الگوی قابل چشم‌پوشی در داده‌ها رخ می‌دهند.

هدف ما در این مقاله، تعمیم آزمون رمزی برای مدل خطی عام و به‌ویژه تعمیم این آزمون برای تشخیص کفایت مدل خطی عام در حضور داده‌های گم‌شده است.

در بخش بعد، آماره‌های آزمون مناسب را بر اساس روش‌های مناسب برآورد، در حالت داده‌های کامل و داده‌های ناکامل معرفی خواهیم کرد. در ادامه ویژگی‌های توزیعی آماره‌های ارائه‌شده را به دست می‌آوریم. در بخش ۳، آماره‌های ارائه‌شده را بر اساس شبیه‌سازی چند مسئله آماری، مقایسه خواهیم کرد. در بخش ۴، بر اساس آماره‌های ارائه‌شده، یک مسئله واقعی را در اقتصاد تحلیل خواهیم کرد.

- 
- 1- Ignorable Mechanism
  - 2- Missing Completely at Random
  - 3- Missing at Random
  - 4- Score-Type Test
  - 5- Minimum Integrated Square Distance Between the Nonparametric and Parametric Fits

## ۲-آزمون‌های مشخص‌سازی مدل خطی عام و ویژگی‌های آن‌ها

از آزمون رمزی تنها برای بررسی کفایت مدل خطی ساده و مدل خطی چندگانه استفاده می‌شود. این آزمون بیشتر در کارهای عملی موردتوجه محققین قرار گرفته است و تنها [۵] به صورت نظری به آن پرداخته است و آن را به مدل‌های خطی تعمیم یافته، بسط داده است.

در این بخش، ابتدا آزمون رمزی را در حالت داده‌های کامل، برای تشخیص کفایت مدل خطی عام بسط می‌دهیم. سپس در بخش ۲-۲ تحت داده‌های گم‌شده، آماره آزمون جدیدی را ارائه خواهیم داد که از داده‌های گم‌شده نیز در مشخص‌سازی مدل خطی عام استفاده می‌کند. در هر مرحله پارامترهای مدل خطی عام را در حالت داده‌های کامل و حالت داده‌های ناکامل با استفاده از روش کمترین توان‌های دوم خطا، برآورد می‌کنیم و برای مشخص‌سازی مدل خطی عام بر اساس روش برآورد اشاره شده، آماره آزمون مناسب را خواهیم ساخت.

### ۲-۱- حالت داده‌های کامل

برای داده‌های کامل، (Full) معادله برآورد زیر را برای برآورد پارامترهای مدل خطی عام رابطه (۱) در نظر می‌گیریم:

$$\sum_{i=1}^n \Psi_{\beta}(y_i | \mathbf{x}_i) = 0, \quad (2)$$

که در آن برای،  $i = 1, \dots, n$ ،  $\Psi_{\beta}(y_i | \mathbf{x}_i) = g^T(\mathbf{x}_i)(y_i - g^T(\mathbf{x}_i)\beta)$  تابع امتیاز حاصل از روش برآورد کمترین توان‌های دوم خطا است. فرض می‌کنیم  $\hat{y}_i$  مقادیر برازش شده متغیر پاسخ از مدل خطی رابطه (۱) برای  $i$ -امین مشاهده باشد. مدل خطی عام جدید زیر را در نظر می‌گیریم:

$$y = g^T(\mathbf{x})\beta + \gamma_1 \hat{y}^1 + \gamma_2 \hat{y}^2 + \dots + \gamma_q \hat{y}^{q+1} + \eta, \quad (3)$$

که در آن پارامترهای جدید مدل خطی عام هستند و  $\eta$  خطای جدید مدل خطی عام است که از توزیع  $N(0, V^2)$  پیروی می‌کند. آزمون رمزی فرضیات زیر را برای بررسی کفایت مدل خطی عام در نظر می‌گیرد:

$$H_0: \gamma_1 = \dots = \gamma_q = 0 \quad vs \quad H_1: \exists i \ni \gamma_i \neq 0.$$

فرض صفر بیان می‌کند که مدل برازش شده مناسب است اما فرض مقابل بیان می‌کند که مدل برازش شده کفایت و مشخص‌سازی لازم برای بیان رابطه بین متغیر پاسخ و کمکی را ندارد. در واقع آزمون رمزی میزان تأثیر  $(g^T(\mathbf{x})\beta)^k$  ها را بر متغیر پاسخ را با استفاده از مقادیر برآورد شده آن یعنی  $\hat{y}_i$  ها بیان می‌کند و توابع چندجمله‌ای از آن را به فرض مقابل برای بررسی

کفایت مدل اضافه می‌کند. در آزمون رمزی مدل کفایت لازم را ندارد اگر و تنها اگر فرض صفر بودن هر یک از  $\gamma_i$  ها رد شود. در این حالت، مدل رگرسیونی تحت رابطه (۱)، مدل نامقید<sup>۱</sup> و مدل رگرسیونی تحت رابطه (۳)، مدل مقید<sup>۲</sup> نامیده می‌شود. بر اساس تابع آزمون رمزی در حالت مدل خطی ساده، توابع زیر را برای ساخت آماره آزمون رمزی برای مدل خطی عام تعریف می‌کنیم. مجموع توان‌های دوم خطا برای مدل نامقید در رابطه (۱) در حالت داده‌های کامل به صورت زیر تعریف می‌شود:

$$SSE_{U,Full} = \sum_{i=1}^n (y_i - \hat{y}_{Ui})^2,$$

که در آن  $\hat{y}_{Ui}$  مقادیر برازش شده از مدل نامقید هستند. مجموع توان‌های دوم خطا برای مدل مقید رابطه (۳) در حالت داده‌های کامل به صورت زیر تعریف می‌شود:

$$SSE_{R,Full} = \sum_{i=1}^n (y_i - \hat{y}_{Ri})^2,$$

که در آن  $\hat{y}_{Ri}$  مقادیر برازش شده از مدل مقید هستند. بنابراین، آماره آزمون رمزی در حالت داده‌های کامل به صورت زیر تعریف می‌شود:

$$F_{Full} = \frac{(SSE_{U,Full} - SSE_{R,Full}) / q}{SSE_{R,Full} / (n - p - q)}. \quad (4)$$

وقتی داده گم شده در مجموعه داده‌ها وجود نداشته باشد، از این آماره آزمون برای بررسی کفایت مدل خطی عام استفاده می‌کنیم که ویژگی‌های توزیعی آن در قسمت پایانی این بخش مورد بحث قرار می‌گیرد.

## ۲-۲- حالت داده‌های گم شده

بر اساس دسته‌بندی [۶]، گوییم الگوی گم‌شدگی داده‌ها MCAR است، اگر و تنها اگر گم‌شدگی داده‌ها نه به مشاهدات و نه به داده‌های گم‌شده وابسته باشند. همچنین گوییم الگوی گم‌شدگی داده‌ها MAR است، اگر و تنها اگر گم‌شدگی داده‌ها تنها به مشاهدات وابسته باشد. این الگوها که گم‌شدگی داده‌ها در آن‌ها به داده‌های گم‌شده وابسته نیست، الگوهای قابل چشم‌پوشی نامیده می‌شوند.

1- Unrestricted Model

2- Restricted Model

در این بخش حالتی را در نظر می‌گیریم که داده‌های گم‌شده در متغیر پاسخ و یا در متغیر کمکی رخ می‌دهند. حتی فرض بر این است که داده‌های گم‌شده می‌توانند در متغیر کمکی و متغیر پاسخ به صورت هم‌زمان رخ دهند. تنها کافی است که یکی از متغیرها به صورت کامل مشاهده شده باشد تا بتوان توابع مجهول از جمله احتمال گم‌شدگی داده‌ها را برآورد کرد. در این حالت، ساده‌ترین روش برای استنباط در مورد داده‌ها، کنار گذاشتن داده‌هایی است که به صورت ناقص مشاهده شده‌اند و استنباط‌ها با داده‌هایی انجام می‌شود که به صورت کامل مشاهده شده‌اند. این روش، روش حالت کامل<sup>۱</sup> (CC) نامیده می‌شود. بنابراین در این حالت، پارامترهای مدل خطی عام از رابطه زیر به دست می‌آید:

$$\sum_{i=1}^n \delta_i \psi_{\beta}(y_i | \mathbf{x}_i) = 0, \quad (5)$$

که در آن،  $\delta_i$  یک تابع نشانگر است و در صورتی که  $i$ -امین داده به صورت کامل مشاهده شده باشد، برابر یک و در غیر این صورت برابر صفر است. در این حالت تابع آزمون رمزی به صورت زیر خواهد بود:

$$F_{CC} = \frac{(SSE_{U,CC} - SSE_{R,CC}) / q}{SSE_{R,CC} / (\sum_{i=1}^n \delta_i - p - q)}, \quad (6)$$

که در آن،  $SSE_{U,CC}$  و  $SSE_{R,CC}$  به ترتیب، مجموع توان‌های دوم خطای مدل‌های مقید و نامقید بر اساس روش برآورد حالت کامل هستند. در رابطه (۶)، بر اساس مدل نامقید، داریم:

$$SSE_{U,CC} = \sum_{i=1}^n \delta_i (y_i - \hat{y}_{Ui})^2, \quad (7)$$

و بر اساس مدل مقید، داریم:

$$SSE_{R,CC} = \sum_{i=1}^n \delta_i (y_i - \hat{y}_{Ri})^2. \quad (8)$$

وقتی الگوی گم‌شدگی داده‌ها MAR است، ممکن است، روش حالت کامل برآوردهای اریب را نتیجه دهد. بنابراین، از یک روش دیگر برای برآورد پارامترهای مدل خطی عام استفاده می‌کنیم. در این روش از وزن‌های معکوس احتمال<sup>۲</sup> (IPW) استفاده می‌شود که اولین بار توسط [۱۶] معرفی شد. این روش برآوردهای ناریب را نتیجه می‌دهد. عملکرد خوب روش IPW در مقایسه

1- Complete Case

2-Inverse Probability Weights

با روش CC توسط بسیاری از محققین از جمله [۹] و [۱۷] مورد بررسی قرار گرفته است. در روش، IPW پارامترهای مدل خطی عام از حل معادلات برآورد زیر به دست می‌آید:

$$\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} \psi_{\beta}(y_i | \mathbf{x}_i) = 0, \quad (9)$$

که در آن،  $\pi_i$  احتمال مشاهده  $i$ -امین عضو نمونه است. در مطالعات واقعی، مقدار  $\pi_i$  نامعلوم است و می‌توان آن را به صورت ناپارامتری زیر برآورد کرد:

$$\hat{\pi}(v_i) = \frac{\sum_{j=1}^n \delta_j K_h(v_j - v_i)}{\sum_{j=1}^n K_h(v_j - v_i)}, \quad (10)$$

که در آن،  $v_i$ ،  $i$ -امین جزء متغیرهای کاملاً مشاهده شده است و  $K_h(\cdot)$  یک تابع هسته<sup>۱</sup> با پارامتر هموارسازی<sup>۲</sup>  $h$  است. برای مشاهده جزئیات بیشتر در مورد روش وزن معکوس احتمال، رجوع شود به [۱۶] و برای مشاهده جزئیات بیشتر در مورد وزن‌های مناسب به [۱۰] مراجعه شود.

در روش وزن معکوس احتمال، آماره آزمون رمزی به صورت زیر خواهد بود:

$$F_{IPW} = \frac{(SSE_{U,IPW} - SSE_{R,IPW}) / q}{SSE_{R,IPW} / (\sum_{i=1}^n \delta_i - p - q)}, \quad (11)$$

که در آن،  $SSE_{U,IPW}$  و  $SSE_{R,IPW}$  به ترتیب مجموع توان‌های دوم خطای مقید و نامقید برای مدل برازش شده بر اساس روش IPW هستند. در رابطه (۱۱)، بر اساس مدل نامقید، داریم:

$$SSE_{U,IPW} = \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} (y_i - \hat{y}_{U_i})^2, \quad (12)$$

و بر اساس مدل مقید، داریم:

$$SSE_{R,IPW} = \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} (y_i - \hat{y}_{R_i})^2. \quad (13)$$

1- Kernel Function

2 -Smoothing Parameter

### ۲-۳- خواص توزیعی آماره‌های آزمون

در این بخش، برخی از خواص آماره‌های آزمون ارائه شده در بخش ۲ را مورد بررسی قرار می‌دهیم. برای رسیدن به این هدف، نیاز داریم تا تعریف زیر را انجام دهیم.

**تعریف ۱:** اگر  $y_1, y_2, \dots, y_n \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$  فرض کنید:

$$F = \frac{\mathbf{y}^T A \mathbf{y} / df_1}{\mathbf{y}^T B \mathbf{y} / df_2} \quad (14)$$

آنگاه، می‌گوییم،  $F$  دارای توزیع شبه-فیشر  $QF(df_1, df_2, A, B)$  است. به طوری که  $A$  و  $B$  ماتریس‌های  $n \times n$  و  $df_1$  و  $df_2$  درجات آزادی توزیع شبه-فیشر هستند که به ترتیب رتبه ماتریس‌های  $A$  و  $B$  هستند.

این توزیع دارای فرم بسته نیست و تنها تحت شرایطی که در تذکر زیر بیان شده است، به توزیع فیشر تبدیل می‌شود.

**تذکر ۱:** اگر  $A$  و  $B$  ماتریس‌های متقارن و خودتوان باشند و  $AB = 0$ . آنگاه،  $F$  از توزیع فیشر  $F(df_1, df_2)$  پیروی می‌کند. به عبارت دیگر، توزیع فیشر حالت خاصی از توزیع شبه-فیشر است. حال، می‌توانیم خواص توزیعی آماره‌ها را به دست آوریم. این ویژگی‌ها در قضیه زیر داده شده است.

**قضیه ۱:** تحت فرض صفر داریم:

الف)  $F_{Full} \sim F(q, n - p - q)$

ب)  $F_{CC} \sim F(q, \sum_{i=1}^n \delta_i - p - q)$

همچنین، به ازای حجم نمونه‌های بزرگ داریم:

ج)  $F_{IPW} \xrightarrow{D} QF(q, \sum_{i=1}^n \delta_i - p - q, A, B)$

که در آن اگر  $G = (g(X), Z)$  و  $W$  یک ماتریس قطری تعریف شود که درایه‌های قطر اصلی آن برابر با  $\frac{\delta_i}{\pi_i}$  است و  $X$  ماتریس مشاهدات متغیرهای کمکی باشد. آنگاه،  $A$ ،  $B$  و  $Z$  را به صورت زیر تعریف می‌کنیم.

$$A = (W[G(G^T W G)^{-1} G^T - g(X)(g^T(X) W g(X))^{-1} g^T(X)] W)$$



$$B = (W - WG(G^T W G)^{-1} G^T W),$$

$$Z = \begin{bmatrix} \hat{y}_{U,1}^r & \hat{y}_{U,1}^r & \cdots & \hat{y}_{U,1}^{q+1} \\ \hat{y}_{U,r}^r & \hat{y}_{U,r}^r & \cdots & \hat{y}_{U,r}^{q+1} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{U,n}^r & \hat{y}_{U,n}^r & \cdots & \hat{y}_{U,n}^{q+1} \end{bmatrix}.$$

برهان: برای اثبات قسمت (الف) قضیه ۱، می‌توانیم  $SSE_{R,Full}$  را به صورت زیر بنویسیم:

$$\begin{aligned} SSE_{R,Full} &= \sum_{i=1}^n (y_i - \hat{y}_{Ri})^r = (\mathbf{y} - \hat{\mathbf{y}}_R)^T (\mathbf{y} - \hat{\mathbf{y}}_R) \\ &= (\mathbf{y} - G(G^T G)^{-1} G^T \mathbf{y})^T (\mathbf{y} - G(G^T G)^{-1} G^T \mathbf{y}) \\ &= \mathbf{y}^T (I - H)^T (I - H) \mathbf{y}, \end{aligned}$$

که در آن،  $I$  یک ماتریس همبندی  $n \times n$  و  $H = G(G^T G)^{-1} G^T$ . به راحتی می‌توان ثابت کرد که  $H$  و در نتیجه  $I - H$  ماتریس‌های متقارن خودتوان هستند. بنابراین می‌توان نوشت:

$$SSE_{R,Full} = \mathbf{y}^T (I - H) \mathbf{y}. \quad (15)$$

پس، خواهیم داشت:

$$\frac{SSE_{R,Full}}{\sigma^r} \sim \chi^r(n - p - q). \quad (16)$$

که در آن، درجه آزادی از رتبه ماتریس  $I - H$  حاصل شده است. چون  $I - H$  یک ماتریس متقارن خودتوان است، پس رتبه آن برابر با اثر آن خواهد بود. پس،

$$\begin{aligned} df &= \text{trace}(I - H) = \text{trace}(I) - \text{trace}(G(G^T G)^{-1} G^T) \\ &= n - \text{trace}((G^T G)^{-1} G^T G) = n - \text{trace}(I_{(p+q)(p+q)}) = n - p - q, \end{aligned}$$

که در آن،  $I_{(p+q)(p+q)}$  یک ماتریس همبندی  $(p+q) \times (p+q)$  است. از طرف دیگر، به صورت مشابه با عبارت بالا، می‌توانیم  $SSE_{U,Full} - SSE_{R,Full}$  را به صورت زیر بازنویسی کنیم:

$$\begin{aligned} SSE_{U,Full} - SSE_{R,Full} &= \mathbf{y}^T (G(G^T G)^{-1} G^T - g(X)(g^T(X)g(X))^{-1} g^T(X)) \mathbf{y} \\ &:= \mathbf{y}^T H^* \mathbf{y}. \end{aligned} \quad (17)$$

با توجه به رابطه  $G(G^T G)^{-1} g(X) = g(X)$  و فرمول معکوس ماتریس افراز شده، به راحتی می‌توان نتیجه گرفت که  $H^*$  یک ماتریس متقارن خودتوان است. بنابراین، می‌توان نوشت:

$$\frac{SSE_{U,Full} - SSE_{R,Full}}{\sigma^2} \sim \chi^2(q). \quad (18)$$

درجه آزادی از اثر ماتریس خودتوان  $H^*$  به دست می‌آید که برابر است با:

$$\begin{aligned} df &= \text{trace}(G(G^T G)^{-1} G^T - g(X)(g^T(X)g(X))^{-1} g^T(X)) \\ &= p + q - p = q. \end{aligned} \quad (19)$$

حال، با استفاده از رابطه‌های (۱۶) و (۱۸)، قسمت (الف) قضیه ۱ حاصل می‌شود.

خواص توزیعی آماره آزمون روش حالت کامل از این حقیقت که ما از داده‌های ناقص چشم‌پوشی می‌کنیم و از داده‌های کامل برای استنباط در مورد مدل استفاده می‌کنیم، حاصل می‌شود. در

این حالت، ما از  $\sum_{i=1}^n \delta_i$  تعداد داده در استنباطها استفاده می‌کنیم. بنابراین به صورت مشابه با روش داده‌های کامل، آماره آزمون  $F_{CC}$  نیز از توزیع فیشر با درجات آزادی کاهش یافته پیروی می‌کند  $(F(q, \sum_{i=1}^n \delta_i - p - q))$ .

در روش وزن معکوس احتمال، می‌توان پارامترهای مدل خطی عام را با استفاده از معادلات برآورد رابطه (۹) به صورت زیر استخراج کرد.

$$\hat{\beta} = (X^T \hat{W} X)^{-1} X^T \hat{W} y \quad (20)$$

که در آن،  $\hat{W}$  یک ماتریس قطری است و درایه  $i$ -ام قطر اصلی آن برابر است با  $\frac{\delta_i}{\hat{\pi}_i}$ . از طرف دیگر، می‌توان نوشت:

$$\begin{aligned} SSE_{U,IPW} &= \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})^T \hat{W} (\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - g(X)(g^T(X) \hat{W} g(X))^{-1} g^T(X) \hat{W} \mathbf{y})^T \hat{W} \\ &\quad \times (\mathbf{y} - g(X)(g^T(X) \hat{W} g(X))^{-1} g^T(X) \hat{W} \mathbf{y}) \\ &= \mathbf{y}^T (I - g(X)(g^T(X) \hat{W} g(X))^{-1} g^T(X) \hat{W})^T \hat{W} \\ &\quad \times (I - g(X)(g^T(X) \hat{W} g(X))^{-1} g^T(X) \hat{W}) \mathbf{y} \\ &= \mathbf{y}^T (\hat{W} - \hat{W} g(X)(g^T(X) \hat{W} g(X))^{-1} g^T(X) \hat{W}) \mathbf{y}. \end{aligned}$$

با توجه به این‌که درایه‌های قطر اصلی ماتریس  $\hat{W}$  شامل  $\hat{\pi}(v_i)$  ها است. می‌توان ثابت کرد که به ازای حجم نمونه بزرگ و پارامتر هموارسازی کوچک،  $E(\hat{\pi}(\cdot)) = \pi(\cdot) + o_p(1)$  و  $Var(\hat{\pi}(\cdot)) = o_p(1)$  که در آن،  $o_p(1)$  نماد اوی کوچک در احتمال است. بنابراین با استفاده از نابرابری چبیشف نتیجه می‌شود که  $\hat{\pi}(\cdot) = \pi(\cdot) + o_p(1)$ . پس درایه  $i$ -ام قطر اصلی ماتریس  $\hat{W}$  معادل خواهد بود با  $\frac{\delta_i}{\pi_i} + o_p(1)$ . حال با توجه به پیوستگی تابع  $SSE_{U,IPW}$  در  $\hat{W}$  می‌توان نتیجه گرفت:

$$SSE_{U,IPW} = \mathbf{y}^T (W - Wg(X)(g^T(X)Wg(X))^{-1}g^T(X)W) \mathbf{y} + o_p(1). \quad (21)$$

که در آن،  $W$  یک ماتریس قطری است و درایه  $i$ -ام قطر اصلی آن برابر است با  $\frac{\delta_i}{\pi_i}$ . به صورت مشابه می‌توان نتیجه گرفت:

$$SSE_{R,IPW} = \mathbf{y}^T (W - WG(G^T WG)^{-1}G^T W) \mathbf{y} + o_p(1). \quad (22)$$

بنابراین، با به کار بردن چند عملیات ریاضی ساده، خواهیم داشت:

$$\begin{aligned} F_{IPW} &= \frac{(SSE_{U,IPW} - SSE_{R,IPW}) / q}{SSE_{R,IPW} / (\sum_{i=1}^n \delta_i - p - q)} \\ &= \frac{(\mathbf{y}^T (W [G(G^T WG)^{-1}G^T - g(X)(g^T(X)Wg(X))^{-1}g^T(X)W] \mathbf{y}) / q + o_p(1))}{(\mathbf{y}^T (W - WG(G^T WG)^{-1}G^T W) \mathbf{y}) / (\sum_{i=1}^n \delta_i - p - q)} \end{aligned} \quad (23)$$

در نهایت می‌توان آماره آزمون روش IPW را به صورت زیر نوشت:

$$\begin{aligned} F_{IPW} &= \frac{(SSE_{U,IPW} - SSE_{R,IPW}) / q}{SSE_{R,IPW} / (\sum_{i=1}^n \delta_i - p - q)} \\ &= \frac{(\mathbf{y}^T (W [G(G^T WG)^{-1}G^T - g(X)(g^T(X)Wg(X))^{-1}g^T(X)W] \mathbf{y}) / q + o_p(1))}{(\mathbf{y}^T (W - WG(G^T WG)^{-1}G^T W) \mathbf{y}) / (\sum_{i=1}^n \delta_i - p - q)} \end{aligned}$$

حال، با استفاده از تعریف ۱ و قضیه مینکوسکی<sup>۱</sup> قسمت (ج) قضیه ۱ نتیجه می‌شود. همچنین، درجات آزادی به صورت مشابه با اثبات قسمت (ب) قضیه ۱ نتیجه می‌شود. ■

در قضیه ۱، آماره آزمون‌ها در روش داده‌های کامل و روش حالت کامل از توزیع دقیق فیشر پیروی می‌کنند. همچنین، در روش وزن معکوس احتمال، نیاز به برآورد مقادیر نامعلوم  $\pi(\cdot)$  داریم که باعث می‌شود تا آماره آزمون این روش به توزیع شبه-فیشر همگرا شود.

**تذکر ۲:** در مسئله شبیه‌سازی و مطالعه داده‌های واقعی برای محاسبه  $p$ -value نیاز داریم که مقادیر چندک‌های توزیع شبه-فیشر را به دست آوریم. برای این منظور چندک‌های  $p$ -ام توزیع شبه-فیشر را از الگوریتم زیر با استفاده از شبیه‌سازی تقریب می‌زنیم:

گام ۱. تابع آزمون  $F_{IPW}$  را  $N$  بار تولید کنید.

گام ۲. در این گام  $F_{IPW}$ ‌ها را از گام ۱ مرتب کنید و مقدار  $\lfloor pN + 1 \rfloor$ -ام آن را ذخیره کنید (نماد  $[ \cdot ]$  بیانگر جزء صحیح است).

گام ۳. گام‌های ۱ و ۲ را  $t$  بار تکرار کنید.

گام ۴. برای تقریب چندک  $p$ -ام، از مقادیر ذخیره‌شده  $F_{IPW}$ ‌ها در گام ۲، میانگین بگیرید.

$F_{IPW}$  با توجه به رابطه (۱۱) تولید می‌شود. همچنین  $t$  تعداد تکرارهای الگوریتم مونت کارلو را نشان می‌دهد. بنابراین هر چه مقدار  $t$  بزرگ‌تر باشد، انتظار داریم مقادیر برآورد شده از توزیع شبه-فیشر به مقدار واقعی نزدیک‌تر باشند. مطالعات اولیه ما نشان می‌دهد که انتخاب مقادیر بزرگ‌تر از ۵۰۰ به برآوردهای مناسب منجر می‌شود. در این مقاله مقدار  $t$ ، ۱۰۰۰ اختیار شده است.

### ۳- شبیه‌سازی چند مسئله

در این بخش، عملکرد تابع آزمون‌های معرفی‌شده را بحث خواهیم کرد. مسئله شبیه‌سازی را به دو مرحله تقسیم می‌کنیم. در مرحله اول، مدل خطی عام با دو متغیر کمکی را در نظر می‌گیریم و در مرحله دوم، مدل خطی عام با سه متغیر کمکی را در نظر می‌گیریم. برای مشاهده اثر حجم نمونه در مطالعات، در هر مرحله از شبیه‌سازی از حجم نمونه‌های ۵۰ و ۱۰۰ استفاده می‌کنیم. همچنین، برای مشاهده اثر الگوی گم‌شدگی بر تحلیل‌ها، از الگوهای مختلف استفاده می‌کنیم.

در آزمون رمزی فرض مقابل باعث می‌شود تا مدل خطی عام تحت این فرض به مقادیر برآورد شده متغیر پاسخ از فرض صفر وابسته باشد. این امر باعث می‌شود تا نتوان از فرض مقابل برای

تولید داده‌ها استفاده کرد. برای حل این مشکل همانند آزمون‌های ارائه‌شده به روش تابع امتیاز عمل می‌کنیم که بر صحت مسئله تأثیرگذار نیست. داده‌ها را از مدل خطی عام زیر تولید می‌کنیم:

$$y = g^T(\mathbf{x})\boldsymbol{\beta} + ch(\mathbf{x}) + \varepsilon, \quad (24)$$

و مدل خطی عام رابطه (۱) را بر داده‌ها برازش می‌دهیم. در هر مرحله برای مقادیر مختلف  $c$ ، توان تجربی (EP) رد فرض صفر را حساب می‌کنیم که با افزایش مقدار  $c$  از فرض صفر دورتر می‌شویم. وقتی  $c = 0$ ، مدل خطی عام رابطه (۲۴) با مدل خطی عام رابطه (۱) معادل است. همچنین، دو نوع متفاوت از مدل خطی عام مقید زیر را در نظر می‌گیریم تا اثر قیدهای مختلف را نیز مشاهده کنیم.

$$RESET \ 1: y = g^T(\mathbf{x})\boldsymbol{\beta} + \gamma_1 \hat{y}^1 + \eta,$$

$$RESET \ 2: y = g^T(\mathbf{x})\boldsymbol{\beta} + \gamma_1 \hat{y}^1 + \gamma_2 \hat{y}^2 + \eta.$$

در هر مرحله، توان تجربی رد فرض صفر با استفاده از ۲۰۰۰ تکرار الگوریتم مونت کارلو<sup>۱</sup> برآورد شده است.

### ۳-۱- مرحله اول: مدل خطی عام با دو متغیر کمکی

در رابطه (۲۴)، فرض کنید،

$$h(\mathbf{x}) = \sqrt{1 + \sin(2\pi x_1) + \frac{(1+x_2^2)}{4}},$$

$$g(\mathbf{x}) = (1 + \sin(2\pi x_1), \frac{(1+x_2^2)}{4})^T,$$

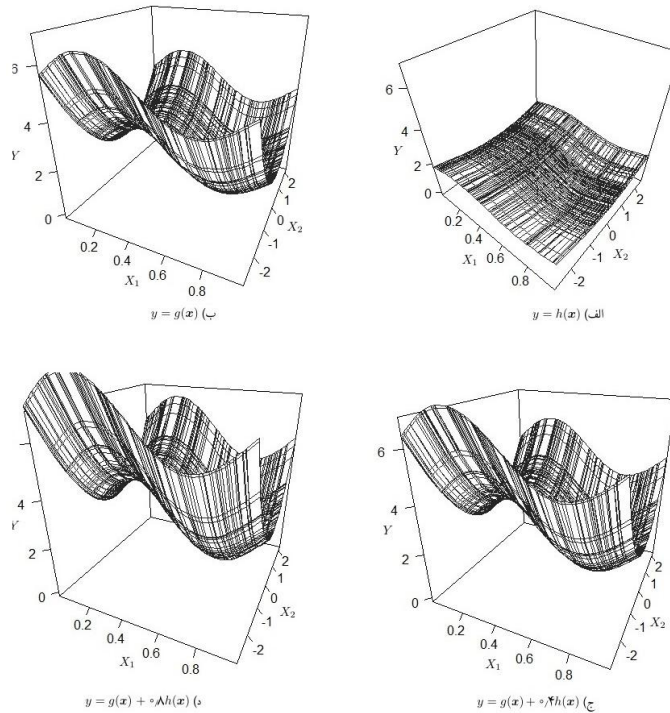
که در آن،  $x_1$  از توزیع یکنواخت  $U(0,1)$  تولید می‌شود،  $x_2$  از توزیع نرمال  $N(0,1)$  تولید می‌شود،  $\varepsilon$  یک متغیر تصادفی از توزیع نرمال  $N(0,0.25)$  است و پارامترهای مدل خطی عام برابر با بردار  $(1,2)$  در نظر گرفته شده است.

الگوهای گم‌شدگی زیر را برای تولید داده‌های گم‌شده در متغیر پاسخ در نظر می‌گیریم:

$$\pi_1(\mathbf{x}_i) = 1 / (1 + e^{-\left(\frac{0.25 + 0.5((1 + \sin(2\pi x_{1i})) + \frac{(1+x_{2i}^2)}{4})}{4}\right)})$$

$$\pi_2(\mathbf{x}_i) = 0.73$$

با استفاده از هر دو الگوی گم‌شدگی، تقریباً ۲۷٪ داده گم‌شده در متغیر پاسخ رخ می‌دهد. درحالی‌که در اولین تابع گم‌شدگی از الگوی گم‌شدگی MAR پیروی می‌کند و تابع بعدی از الگوی MCAR پیروی می‌کند.



شکل (۱): نمودارهای مختلف مربوط به اجزای مدل‌های خطی عام تحت فرض صفر و فرض مقابل. نمودار (الف) جزء  $h(\mathbf{x})$  از فرض مقابل را نشان می‌دهد که در هر مرحله از شبیه‌سازی به مدل خطی عام تحت فرض صفر (نمودار (ب)) با ضریب  $c$  اضافه می‌شود تا فرض مقابل ساخته شود. نمودارهای (ج) و (د) نیز به ترتیب مدل خطی عام را تحت فرض مقابل به ازای مقادیر  $c = 0.4$  و  $c = 0.8$  نشان می‌دهند.

در شکل (۱) نمودارهای مختلف رسم شده است تا دید بهتری از شکل مدل خطی عام تحت فرض صفر در مقایسه با مدل خطی عام تحت فرض مقابل داشته باشیم. در این شکل، نمودار (ب) مدل خطی عام را تحت فرض صفر نشان می‌دهد که با اضافه شدن ضریبی از تابع  $h(\mathbf{x})$ ، مدل‌های مفروض تحت رابطه (۲۴) به دست می‌آید. دو نمونه از نمودارهای مفروض تحت رابطه (۲۴) در شکل (۱) در قسمت (ج) و (د) داده شده است.

جدول (۱): مرحله اول: توان تجربی رد فرض صفر بر اساس قید RESET<sup>۱</sup> به ازای حجم نمونه ۵۰ و ۱۰۰ (Full): حالت داده‌های کامل؛ CC: حالت کامل؛ IPW: وزن معکوس احتمال).

الگوی گم‌شدگی	$n$						
	۱۰۰			۵۰			
	IPW	CC	Full	IPW	CC	Full	
$\pi_r: MAR$	۰/۰۴۹۵	۰/۵۱۰	۰/۰۵۳۰	۰/۰۵۲۵	۰/۰۵۰۰	۰/۰۴۳۵	۰/۰۰
	۰/۰۶۸۵	۰/۰۶۳۰	۰/۰۷۲۵	۰/۰۶۱۵	۰/۰۵۸۵	۰/۰۶۰۰	۰/۱۰
	۰/۱۱۳۵	۰/۱۱۱۵	۰/۱۳۶۵	۰/۰۸۷۰	۰/۰۸۸۰	۰/۱۰۱۵	۰/۲۰
	۰/۲۰۳۰	۰/۱۹۱۰	۰/۲۸۵۰	۰/۱۲۹۰	۰/۱۲۹۵	۰/۱۵۲۰	۰/۳۰
	۰/۳۲۱۵	۰/۳۰۲۵	۰/۴۲۴۵	۰/۱۸۳۰	۰/۱۷۷۵	۰/۲۲۴۵	۰/۴۰
	۰/۴۵۹۵	۰/۴۳۸۵	۰/۵۷۸۵	۰/۲۴۵۰	۰/۲۳۴۵	۰/۳۲۴۰	۰/۵۰
	۰/۶۱۲۰	۰/۵۸۷۰	۰/۷۳۸۵	۰/۳۳۳۵	۰/۳۱۰۰	۰/۴۳۴۵	۰/۶۰
	۰/۷۴۶۵	۰/۷۲۵۰	۰/۸۵۸۵	۰/۴۲۳۵	۰/۴۱۲۵	۰/۵۴۹۰	۰/۷۰
	۰/۸۴۵۰	۰/۸۲۵۵	۰/۹۳۰۰	۰/۵۱۹۵	۰/۵۰۵۵	۰/۶۶۲۵	۰/۸۰
	۰/۹۱۰۰	۰/۹۰۰۰	۰/۹۶۸۵	۰/۶۰۲۵	۰/۵۸۹۵	۰/۷۶۳۰	۰/۹۰
	۰/۹۵۶۰	۰/۹۴۸۵	۰/۹۸۸۰	۰/۶۹۸۵	۰/۶۸۶۸	۰/۸۳۹۰	۱/۰۰
	۰/۹۷۷۵	۰/۹۷۵۵	۰/۹۹۴۵	۰/۷۷۷۵	۰/۷۶۵۵	۰/۸۹۵۰	۱/۱۰
۰/۹۸۵۵	۰/۹۸۴۵	۰/۹۹۹۰	۰/۸۴۴۰	۰/۸۳۵۰	۰/۹۳۰۰	۱/۲۰	
$\pi_r: MCAR$	۰/۰۵۹۵	۰/۰۵۸۵	۰/۰۵۳۰	۰/۰۵۶۰	۰/۰۵۳۵	۰/۰۴۳۵	۰/۰۰
	۰/۰۶۰۰	۰/۰۶۵۰	۰/۰۷۲۵	۰/۰۶۳۵	۰/۰۶۰۰	۰/۰۶۵۵	۰/۱۰
	۰/۱۱۷۰	۰/۱۱۱۲۰	۰/۱۳۶۵	۰/۰۸۶۵	۰/۰۸۶۵	۰/۱۰۱۵	۰/۲۰
	۰/۲۰۵۵	۰/۱۹۸۰	۰/۲۵۸۰	۰/۱۲۲۵	۰/۱۲۱۰	۰/۱۵۲۰	۰/۳۰
	۰/۳۱۲۵	۰/۳۰۲۵	۰/۴۱۴۵	۰/۱۸۱۵	۰/۱۷۴۰	۰/۲۲۴۵	۰/۴۰
	۰/۴۵۹۵	۰/۴۴۶۰	۰/۵۷۸۵	۰/۲۵۲۰	۰/۲۴۷۰	۰/۳۲۴۰	۰/۵۰
	۰/۶۰۰۰	۰/۵۸۸۰	۰/۷۳۸۵	۰/۳۳۹۰	۰/۳۳۳۵	۰/۴۳۴۵	۰/۶۰
	۰/۷۳۰۵	۰/۷۲۷۰	۰/۸۵۸۵	۰/۴۲۳۰	۰/۴۱۹۵	۰/۵۴۹۰	۰/۷۰
	۰/۸۳۴۰	۰/۸۲۷۰	۰/۹۳۰۰	۰/۵۲۲۰	۰/۵۱۱۰	۰/۶۶۲۵	۰/۸۰
	۰/۹۰۴۵	۰/۹۰۰۰	۰/۹۶۸۵	۰/۶۱۱۵	۰/۶۰۰۰	۰/۷۶۳۰	۰/۹۰
	۰/۹۴۵۰	۰/۹۴۶۰	۰/۹۸۸۰	۰/۶۹۹۰	۰/۶۹۷۰	۰/۸۳۹۰	۱/۰۰
	۰/۹۶۹۰	۰/۹۶۹۵	۰/۹۹۴۵	۰/۷۸۰۵	۰/۷۷۷۵	۰/۸۹۵۰	۱/۱۰
۰/۹۸۳۵	۰/۹۸۳۵	۰/۹۹۹۰	۰/۸۳۸۵	۰/۸۳۵۵	۰/۹۳۰۰	۱/۲۰	

جدول (۲): مرحله اول: توان تجربی رد فرض صفر بر اساس قید RESET<sub>۲</sub> به ازای حجم نمونه ۵۰ و ۱۰۰ (Full): حالت داده‌های کامل؛ CC: حالت کامل؛ IPW: وزن معکوس احتمال).

الگوی گم‌شدگی	<i>n</i>						
	۱۰۰			۵۰			
	IPW	CC	Full	IPW	CC	Full	
$\pi_1: MAR$	۰/۰۴۵۰	۰/۰۴۶۵	۰/۰۵۲۵	۰/۰۴۹۰	۰/۰۵۲۵	۰/۰۵۰۰	۰/۰۰
	۰/۰۶۰۵	۰/۰۶۱۰	۰/۰۷۰۵	۰/۰۵۸۵	۰/۰۶۳۵	۰/۰۶۳۰	۰/۱۰
	۰/۱۰۴۰	۰/۱۰۳۰	۰/۱۲۴۰	۰/۰۸۰۵	۰/۰۷۸۵	۰/۰۸۵۰	۰/۲۰
	۰/۱۷۷۵	۰/۱۷۴۵	۰/۲۳۹۰	۰/۱۱۳۵	۰/۱۰۶۰	۰/۱۳۲۵	۰/۳۰
	۰/۲۹۱۵	۰/۲۷۱۰	۰/۳۷۷۰	۰/۱۶۲۰	۰/۱۵۲۰	۰/۱۹۶۵	۰/۴۰
	۰/۴۱۶۵	۰/۴۰۰۰	۰/۵۳۵۵	۰/۲۱۳۰	۰/۲۰۵۰	۰/۲۸۵۰	۰/۵۰
	۰/۵۵۲۰	۰/۵۳۷۰	۰/۶۸۸۵	۰/۲۸۶۵	۰/۲۷۵۰	۰/۳۷۸۰	۰/۶۰
	۰/۶۹۷۰	۰/۶۷۵۵	۰/۸۲۹۵	۰/۳۶۷۵	۰/۳۵۴۵	۰/۵۰۳۰	۰/۷۰
	۰/۸۱۴۰	۰/۸۰۱۵	۰/۹۰۹۰	۰/۴۵۴۵	۰/۴۵۳۵	۰/۶۰۲۵	۰/۸۰
	۰/۸۸۶۰	۰/۸۷۷۰	۰/۹۵۲۵	۰/۵۴۸۵	۰/۵۴۰۰	۰/۷۰۴۵	۰/۹۰
	۰/۹۳۶۰	۰/۹۳۵۵	۰/۹۸۶۰	۰/۶۲۵۵	۰/۶۲۱۰	۰/۷۹۲۵	۱/۰۰
	۰/۹۶۷۵	۰/۹۶۴۵	۰/۹۹۵۰	۰/۷۰۹۰	۰/۷۱۰۰	۰/۸۶۶۰	۱/۱۰
۰/۹۸۲۵	۰/۹۸۴۵	۰/۹۹۸۰	۰/۷۸۵۵	۰/۷۹۴۵	۰/۹۱۴۵	۱/۲۰	
$\pi_2: MAR$	۰/۰۵۳۰	۰/۰۵۱۵	۰/۰۵۲۵	۰/۰۵۲۰	۰/۰۵۶۰	۰/۰۵۰۰	۰/۰۰
	۰/۰۶۴۵	۰/۰۶۲۵	۰/۰۷۰۵	۰/۰۵۸۰	۰/۰۵۸۰	۰/۰۶۳۰	۰/۱۰
	۰/۱۰۰۵	۰/۰۹۵۵	۰/۱۲۴۰	۰/۰۷۷۰	۰/۰۷۷۵	۰/۰۸۵۰	۰/۲۰
	۰/۱۶۸۰	۰/۱۶۶۵	۰/۲۳۹۰	۰/۱۰۷۵	۰/۱۰۸۰	۰/۱۳۲۵	۰/۳۰
	۰/۲۷۵۵	۰/۲۶۹۵	۰/۳۷۷۰	۰/۱۵۹۵	۰/۱۵۰۵	۰/۱۹۶۵	۰/۴۰
	۰/۴۱۴۷	۰/۴۰۶۵	۰/۵۳۵۵	۰/۲۱۵۰	۰/۲۰۹۰	۰/۲۸۵۰	۰/۵۰
	۰/۵۴۶۰	۰/۵۴۲۰	۰/۶۸۸۵	۰/۲۸۰۰	۰/۲۷۷۵	۰/۳۷۸۰	۰/۶۰
	۰/۶۷۸۵	۰/۶۷۹۵	۰/۸۲۹۸	۰/۳۷۰۵	۰/۳۶۸۰	۰/۵۰۳۰	۰/۷۰
	۰/۸۰۱۰	۰/۸۰۳۰	۰/۹۰۹۰	۰/۴۵۴۰	۰/۴۵۱۵	۰/۶۰۲۵	۰/۸۰
	۰/۹۲۵۵	۰/۹۲۵۰	۰/۹۸۶۰	۰/۵۵۰۵	۰/۵۴۳۰	۰/۷۰۴۵	۰/۹۰
	۰/۹۲۵۵	۰/۹۲۵۰	۰/۹۸۶۰	۰/۶۴۰۵	۰/۶۳۲۰	۰/۷۹۲۵	۱/۰۰
	۰/۹۶۱۰	۰/۹۵۹۵	۰/۹۹۵۰	۰/۷۱۱۵	۰/۷۱۷۵	۰/۸۶۶۰	۱/۱۰
۰/۹۷۸۵	۰/۹۸۱۵	۰/۹۹۸۰	۰/۷۸۳۰	۰/۷۸۸۰	۰/۹۴۵۰	۱/۲۰	



نمودار (ج) بیانگر مدل خطی عام تحت فرض مقابل با ضریب ثابت  $c = 0/4$  و نمودار (د) بیانگر مدل خطی عام تحت فرض مقابل با ضریب ثابت  $c = 0/8$  است. تابع  $h(\mathbf{x})$  یک تابع مثبت است، بنابراین با افزایش مقدار  $c$ ، مدل‌های جدید اندکی به سمت بالا انتقال می‌یابند. همچنین شکل تقریباً مسطح تابع  $h(\mathbf{x})$  باعث می‌شود تا نمودارهای تحت فرض مقابل در مقایسه با نمودار تحت فرض صفر از لحاظ ظاهری تفاوت چندانی نداشته باشند. این ویژگی‌های تابع  $h(\mathbf{x})$  باعث می‌شود تا با افزایش مقدار  $c$ ، توان آزمون به‌صورت آهسته افزایش یابد. در صورتی که تابع  $h(\mathbf{x})$  را طوری انتخاب می‌کردیم که شکل نمودارها تحت فرض مقابل تغییرات بیشتری می‌داشت، با افزایش مقدار  $c$  احتمال رد فرض صفر از حالت قبلی خیلی بیشتر می‌شد. به همین دلیل  $h(\mathbf{x})$  را تقریباً مسطح اختیار کردیم تا عملکرد آماره‌های پیشنهاد شده در بدترین حالات نیز مطالعه شود. نتایج مرحله اول با توجه به آماره‌های  $RESET^1$  و  $RESET^2$  به ترتیب در جدول‌های (۱) و (۲) داده شده است.

### ۳-۲- مرحله دوم: مدل خطی عام با یک متغیر کمکی

برای مدل خطی عام رابطه (۲۴)، فرض کنید،  $g(\mathbf{x}) = (x_1, \frac{(1+x_2^2)}{4}, x_2)^T$  و  $h(\mathbf{x}) = 3(x_1 + x_2^2)$  که در آن،  $x_1$  از توزیع نرمال  $N(0, 1)$ ،  $x_2$  از توزیع یکنواخت  $U(0, 1)$  و  $x_3$  از توزیع نرمال  $N(0, 1)$  تولید می‌شوند.  $\varepsilon$  خطای نمونه‌گیری تصادفی از توزیع نرمال  $N(0, 0/25)$  و پارامترهای مدل خطی عام برابر با بردار  $(1, 2, 3)$  است.

الگوهای گم‌شدگی زیر را برای تولید داده‌های گم‌شده در متغیر پاسخ  $y$  و متغیر کمکی  $x_1$  در نظر می‌گیریم.

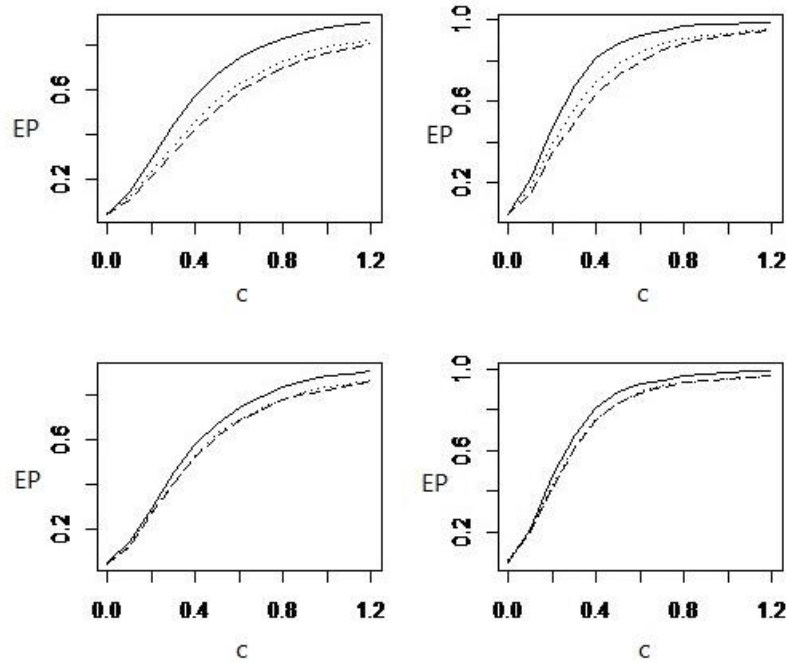
$$\text{حالت سوم: } \pi_{x_2}(\mathbf{x}_i) = 1 / (1 + 0/5 e^{(-0/5 x_{2i}^2)})$$

$$\text{حالت چهارم: } \pi_{x_3}(\mathbf{x}_i) := \begin{cases} \pi_{x_1}(\mathbf{x}_i) = 1 / (1 + 0/5 e^{(-0/5 (x_{1i}^2 + x_{2i} + x_{3i}^2)}) \\ \pi_{x_2}(\mathbf{x}_i) = 1 / (1 + 0/5 | 0/25 x_{2i} + 0/3 x_{3i}^2 |) \end{cases}$$

با استفاده از الگوهای فوق، در هر دو حالت، به‌طور متوسط ۲۹٪ داده گم‌شده در متغیرها ایجاد می‌شود. در حالت سوم، متغیر پاسخ  $y$  و متغیر کمکی  $x_1$  یا به‌صورت هم‌زمان گم‌شده هستند و یا به‌صورت هم‌زمان مشاهده شده‌اند. اما در حالت چهارم،  $1 - \pi_{x_1}$  احتمال آن است که متغیر پاسخ  $y$  گم‌شده باشد. از طرف دیگر،  $1 - \pi_{x_2}$  احتمال آن است که متغیر کمکی  $x_1$  گم‌شده باشد، به شرط آن که متغیر پاسخ  $y$  مشاهده شده باشد. همچنین با استفاده از الگوی گم‌شدگی  $\pi_{x_3}$ ، ۱۵/۵٪ داده گم‌شده در متغیر پاسخ  $y$  خواهیم داشت و با استفاده از الگوی گم‌شدگی  $\pi_{x_2}$

،  $13/5\%$  داده گم شده در متغیر کمکی  $x_1$  خواهیم داشت. بنابراین در حالت چهارم نیز به طور متوسط  $29\%$  داده گم شده در داده ها خواهیم داشت.

نتایج مرحله دوم در جدول های (۳) و (۴) داده شده است. همچنین، نمودار توان تجربی رد فرض صفر به ازای قید  $1$  RESET در شکل (۲) داده شده است. در شکل (۲)، نمودارهای بالایی، توان تجربی رد فرض صفر را به ازای الگوی گم شدگی  $\pi_p$  نشان می دهند. همچنین نمودارهای پایینی، توان تجربی رد فرض صفر را به ازای الگوی گم شدگی  $\pi_f$  نشان می دهند. به علاوه، نمودارهای سمت چپ، بر اساس نمونه  $5^\circ$  تایی و نمودارهای سمت راست، بر اساس نمونه  $10^\circ$  هستند.



**شکل (۲):** نمودار توان تجربی رد فرض صفر برای مقادیر مختلف  $c$  تحت مرحله دوم بر اساس قید  $1$  RESET: نمودارهای بالایی بر اساس الگوی گم شدگی  $\pi_p$  و نمودارهای پایینی بر اساس الگوی گم شدگی  $\pi_f$  هستند. نمودارهای سمت چپ بر اساس نمونه  $5^\circ$  تایی و نمودارهای سمت راست بر اساس نمونه  $10^\circ$  تایی هستند. همچنین، نمودارهای خط ممتد، نمودارهای خط تیره و نمودارهای نقطه ای به ترتیب توان تجربی آماره آزمون های روش داده های کامل، روش حالت کامل و روش وزن معکوس احتمال را نشان می دهند.

جدول (۳): مرحله دوم: توان تجربی رد فرض صفر بر اساس قید RESET<sup>۱</sup> به ازای حجم نمونه ۵۰ و ۱۰۰ (Full): حالت داده‌های کامل؛ CC: حالت کامل؛ IPW: وزن معکوس احتمال).

گم‌شدگی	<i>n</i>						C
	۱۰۰			۵۰			
	IPW	CC	Full	IPW	CC	Full	
$\pi_1: MAR$	۰/۰۴۸۵	۰/۰۴۸۵	۰/۰۴۵۵	۰/۰۵۴۰	۰/۰۴۷۰	۰/۰۴۵۵	۰/۰۰
	۰/۱۷۰۰	۰/۱۴۵۰	۰/۲۱۳۵	۰/۱۲۶۵	۰/۱۱۰۵	۰/۱۴۸۰	۰/۱۰
	۰/۴۰۱۰	۰/۳۵۰۵	۰/۴۷۶۵	۰/۲۳۵۰	۰/۲۱۳۵	۰/۲۹۵۵	۰/۲۰
	۰/۵۶۸۵	۰/۵۰۱۰	۰/۶۷۶۰	۰/۳۵۳۰	۰/۳۲۳۰	۰/۴۵۵۰	۰/۳۰
	۰/۶۹۸۰	۰/۶۳۲۰	۰/۸۱۰۰	۰/۴۶۱۰	۰/۴۲۱۰	۰/۵۷۶۵	۰/۴۰
	۰/۷۸۳۰	۰/۷۲۹۰	۰/۸۸۳۵	۰/۵۵۷۵	۰/۵۱۷۰	۰/۶۷۰۰	۰/۵۰
	۰/۸۴۳۵	۰/۷۹۷۵	۰/۹۲۴۰	۰/۶۳۰۵	۰/۵۹۲۵	۰/۷۴۲۰	۰/۶۰
	۰/۸۸۱۵	۰/۸۵۱۵	۰/۹۴۵۵	۰/۶۸۲۵	۰/۶۵۳۰	۰/۷۹۳۵	۰/۷۰
	۰/۹۰۹۰	۰/۸۸۴۰	۰/۹۶۶۰	۰/۷۳۰۰	۰/۵۹۹۰	۰/۸۳۱۰	۰/۸۰
	۰/۹۲۵۰	۰/۹۰۸۵	۰/۹۷۳۰	۰/۷۶۵۰	۰/۷۳۴۵	۰/۸۶۲۰	۰/۹۰
	۰/۹۳۳۵	۰/۹۲۳۰	۰/۹۷۹۵	۰/۷۹۲۵	۰/۷۶۶۰	۰/۸۸۲۰	۱/۰۰
	۰/۹۴۷۵	۰/۹۳۶۵	۰/۹۸۵۵	۰/۸۱۲۰	۰/۷۹۱۰	۰/۸۹۴۰	۱/۱۰
۰/۹۵۴۰	۰/۹۴۴۵	۰/۹۸۷۵	۰/۸۲۹۵	۰/۸۰۹۵	۰/۹۰۴۵	۱/۲۰	
$\pi_2: MAR$	۰/۰۶۴۵	۰/۰۵۷۵	۰/۰۴۵۵	۰/۰۵۰۰	۰/۰۴۶۰	۰/۰۵۰۰	۰/۰۰
	۰/۲۱۴۵	۰/۲۰۰۰	۰/۲۱۳۵	۰/۱۲۹۰	۰/۱۲۶۵	۰/۱۴۸۰	۰/۱۰
	۰/۴۳۱۵	۰/۴۱۹۰	۰/۴۷۶۵	۰/۲۸۳۵	۰/۲۷۴۵	۰/۲۹۵۵	۰/۲۰
	۰/۶۱۸۵	۰/۶۱۳۰	۰/۶۷۶۰	۰/۴۰۲۷	۰/۴۰۸۰	۰/۴۵۵۰	۰/۳۰
	۰/۷۵۷۰	۰/۷۴۸۵	۰/۸۱۰۰	۰/۵۲۹۵	۰/۵۲۱۵	۰/۵۷۶۵	۰/۴۰
	۰/۸۲۹۵	۰/۸۳۰۰	۰/۸۸۳۵	۰/۶۲۶۰	۰/۶۱۷۵	۰/۶۷۰۰	۰/۵۰
	۰/۸۸۶۰	۰/۸۷۹۵	۰/۹۲۴۰	۰/۶۹۳۰	۰/۶۸۲۵	۰/۷۴۲۰	۰/۶۰
	۰/۹۱۴۵	۰/۹۱۱۵	۰/۹۴۵۵	۰/۷۳۹۰	۰/۷۳۳۰	۰/۷۹۳۵	۰/۷۰
	۰/۹۳۲۰	۰/۹۳۳۵	۰/۹۶۶۰	۰/۷۸۰۰	۰/۷۷۴۵	۰/۸۳۱۰	۰/۸۰
	۰/۹۴۳۵	۰/۹۴۲۵	۰/۹۷۳۰	۰/۸۱۳۵	۰/۸۰۷۵	۰/۸۶۲۰	۰/۹۰
	۰/۹۵۰۵	۰/۹۵۳۰	۰/۹۷۹۵	۰/۸۳۳۵	۰/۸۲۲۵	۰/۸۸۲۰	۱/۰۰
	۰/۹۵۵۰	۰/۹۶۱۰	۰/۹۸۵۵	۰/۸۴۹۵	۰/۸۴۵۰	۰/۸۹۴۰	۱/۱۰
۰/۹۶۳۰	۰/۹۶۵۰	۰/۹۸۷۵	۰/۸۵۹۵	۰/۸۵۷۰	۰/۹۰۴۵	۱/۲۰	

جدول (۴): مرحله دوم: توان تجربی رد فرض صفر بر اساس قید RESET<sub>۲</sub> به ازای حجم نمونه ۵۰ و ۱۰۰ (Full): حالت داده‌های کامل؛ CC: حالت کامل؛ IPW: وزن معکوس احتمال).

الگوی گم‌شدگی	$n$						
	۱۰۰			۵۰			
	IPW	CC	Full	IPW	CC	Full	
$\pi_1: MAR$	۰/۰۴۸۵	۰/۰۴۸۵	۰/۰۴۵۵	۰/۰۵۰۰	۰/۰۴۹۰	۰/۰۵۴۰	۰/۰۰
	۰/۱۷۴۰	۰/۱۵۵۰	۰/۲۰۷۰	۰/۱۲۶۰	۰/۱۱۲۵	۰/۱۴۲۵	۰/۱۰
	۰/۳۸۶۵	۰/۳۳۶۵	۰/۴۵۳۵	۰/۲۲۵۰	۰/۲۰۵۵	۰/۲۷۸۰	۰/۲۰
	۰/۵۲۷۰	۰/۴۷۸۵	۰/۶۳۲۰	۰/۳۳۸۰	۰/۳۰۹۰	۰/۴۲۷۰	۰/۳۰
	۰/۶۶۴۵	۰/۵۹۹۰	۰/۷۷۰۰	۰/۴۳۶۰	۰/۴۰۵۵	۰/۵۵۱۵	۰/۴۰
	۰/۷۵۱۰	۰/۶۹۷۰	۰/۸۵۲۵	۰/۵۲۹۰	۰/۴۹۴۵	۰/۶۴۴۵	۰/۵۰
	۰/۸۱۳۵	۰/۷۶۷۰	۰/۹۰۸۰	۰/۶۰۱۵	۰/۵۶۸۵	۰/۷۱۳۵	۰/۶۰
	۰/۸۵۹۵	۰/۸۲۱۵	۰/۹۳۴۰	۰/۶۶۲۵	۰/۶۳۱۵	۰/۷۶۹۵	۰/۷۰
	۰/۸۹۶۵	۰/۸۶۲۰	۰/۹۵۳۰	۰/۷۰۶۰	۰/۶۷۶۰	۰/۸۰۸۰	۰/۸۰
	۰/۹۱۹۰	۰/۸۸۵۵	۰/۹۶۵۵	۰/۷۴۰۰	۰/۷۱۳۵	۰/۸۳۶۵	۰/۹۰
	۰/۹۲۹۰	۰/۹۱۱۰	۰/۹۷۳۰	۰/۷۷۱۰	۰/۷۳۸۰	۰/۸۶۱۵	۱/۰۰
	۰/۹۳۸۵	۰/۹۲۷۵	۰/۹۷۹۰	۰/۷۶۹۵	۰/۷۶۹۵	۰/۸۸۳۵	۱/۱۰
۰/۹۴۵۰	۰/۹۳۵۵	۰/۹۸۶۵	۰/۸۱۵۰	۰/۷۹۵۵	۰/۸۹۱۵	۱/۲۰	
$\pi_2: MAR$	۰/۰۴۹۵	۰/۰۴۸۰	۰/۰۴۵۵	۰/۰۵۵۰	۰/۰۵۵۵	۰/۰۵۴۰	۰/۰۰
	۰/۱۹۵۰	۰/۱۸۸۰	۰/۲۰۷۰	۰/۱۳۲۵	۰/۱۳۴۰	۰/۱۴۲۵	۰/۱۰
	۰/۳۹۸۰	۰/۳۹۹۰	۰/۴۵۳۵	۰/۲۶۸۰	۰/۲۶۳۰	۰/۲۷۸۰	۰/۲۰
	۰/۵۷۹۰	۰/۵۸۴۰	۰/۶۳۲۰	۰/۳۸۱۵	۰/۳۷۴۵	۰/۴۲۷۰	۰/۳۰
	۰/۷۱۲۰	۰/۷۰۷۵	۰/۷۷۰۰	۰/۴۹۲۵	۰/۴۸۳۰	۰/۵۵۱۵	۰/۴۰
	۰/۸۰۷۵	۰/۸۰۸۰	۰/۸۵۲۵	۰/۵۸۸۰	۰/۵۷۹۵	۰/۶۴۴۵	۰/۵۰
	۰/۸۵۹۰	۰/۸۶۰۰	۰/۹۰۸۰	۰/۶۵۱۵	۰/۶۴۸۵	۰/۷۱۳۵	۰/۶۰
	۰/۸۹۷۵	۰/۸۹۵۵	۰/۹۳۴۰	۰/۷۰۳۵	۰/۷۰۰۰	۰/۷۶۹۵	۰/۷۰
	۰/۹۱۶۵	۰/۹۱۹۰	۰/۹۵۳۰	۰/۷۰۳۳	۰/۷۴۴۰	۰/۸۰۸۰	۰/۸۰
	۰/۹۳۱۰	۰/۹۳۶۰	۰/۹۶۵۵	۰/۷۷۹۵	۰/۷۷۷۰	۰/۸۳۶۵	۰/۹۰
	۰/۹۴۰۰	۰/۹۴۵۵	۰/۹۷۳۰	۰/۸۰۶۵	۰/۸۰۵۰	۰/۸۶۱۵	۱/۰۰
	۰/۹۵۴۵	۰/۹۵۲۰	۰/۹۷۹۰	۰/۸۳۱۰	۰/۸۲۶۰	۰/۸۸۳۵	۱/۱۰
۰/۹۵۵۰	۰/۹۵۷۵	۰/۹۸۶۵	۰/۸۴۷۰	۰/۸۴۵۰	۰/۸۹۱۵	۱/۲۰	

عملکرد بسیار خوب روش IPW در مقایسه با روش CC از شکل (۲) به‌خوبی نمایان است. به‌طوری‌که به ازای الگوی اول این عملکرد نمود بیشتری دارد. وقتی الگوی گم‌شدگی  $\pi_4$  است، توان تجربی رد فرض صفر روش CC به توان تجربی رد فرض صفر روش IPW نزدیک‌تر می‌شود. این نتایج را می‌توان از جدول (۴) نیز تحت قید RESET<sub>2</sub> به دست آورد. با مقایسه نتایج جدول (۳) و جدول (۴)، می‌توان نتیجه گرفت که مطالعه تحت قید RESET<sub>1</sub> از مطالعه تحت قید RESET<sub>2</sub> پرتوان‌تر است. این نتیجه‌گیری می‌تواند از انتخاب تابع  $h(\mathbf{x})$  نشأت بگیرد. زیرا ما  $h(\mathbf{x})$  را طوری انتخاب کرده‌ایم که تابعی از درایه‌های تابع برداری  $g(\mathbf{x})$  است. انتخاب دیگری از  $h(\mathbf{x})$ ، می‌توانست نتایج دیگری را به دست دهد.

#### ۴- داده‌های واقعی

در این بخش، از آزمون‌های مشخص‌سازی مدل پیشنهادشده، برای بررسی کفایت مدل برازش شده بر داده‌های واقعی استفاده می‌شود. به همین منظور، یک مجموعه داده واقعی در نظر گرفته می‌شود که برخی از اطلاعات آن گم‌شده است. داده‌های مربوط به وضع کیفیت هوای شهر نیویورک در سال ۱۹۷۳ به ازای ۱۵۳ روز مختلف در نظر گرفته شد. این داده‌ها در بسته محاسباتی "speff2trial" با نام داده "airquality" در نرم‌افزار "R" گنجانده شده است. جزئیات این داده‌ها در [۱۸] داده‌شده است. ما متغیر مربوط به سطح گاز اوزون ( $Y$ ) برحسب تعداد ذرات در میلیارد (ppb) را به‌عنوان متغیر پاسخ در نظر می‌گیریم که نزدیک به ۲۴ درصد از اطلاعات آن گم‌شده است. همچنین سرعت باد ( $X$ ) برحسب مایل بر ساعت (mph) را به‌عنوان متغیر کمکی در نظر می‌گیریم که به‌صورت کامل مشاهده‌شده است.

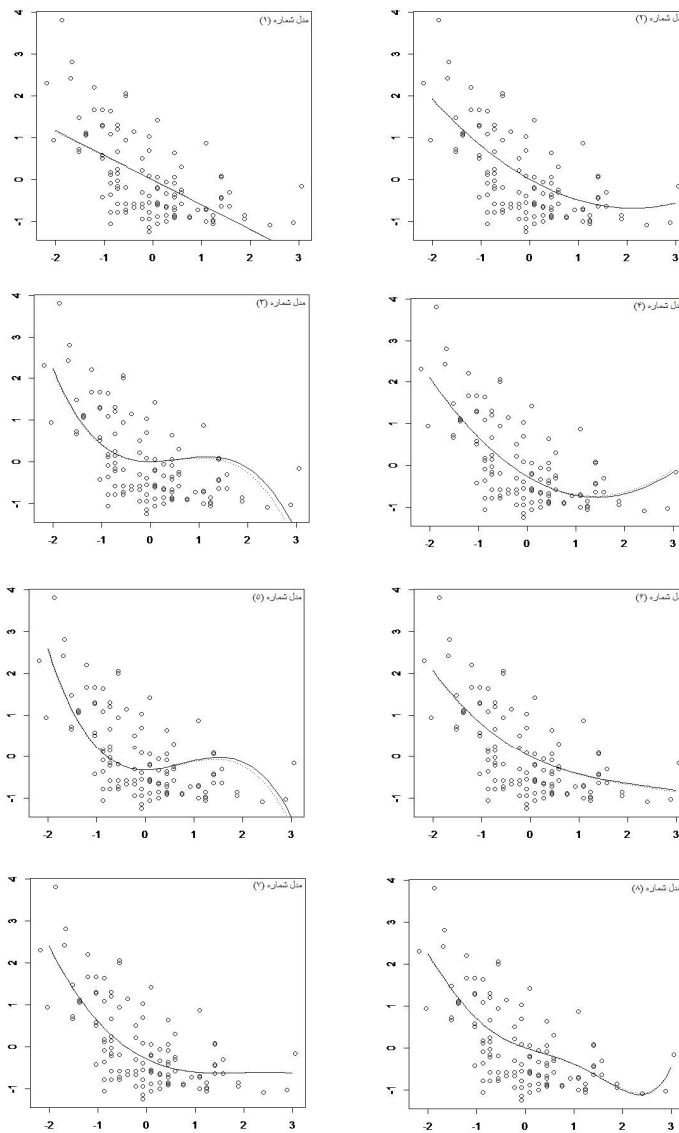
برای سهولت محاسبات، داده‌ها بر اساس داده‌های مشاهده‌شده، استاندارد شده‌اند. مدل‌های مختلف برای بحث در مورد رابطه بین دو متغیر برازش داده‌شده است و صحت آن‌ها با استفاده از آزمون مشخص‌سازی مدل RESET<sub>2</sub> بررسی شده است. نتایج این بررسی در جدول (۵) درج شده است. همچنین، شکل (۳) برخی از مدل‌های برازش شده بر داده‌ها را نشان می‌دهد. از جدول (۵) می‌توان نتیجه گرفت که برازش مدل خطی ساده برای این داده‌ها مناسب نیست. این نتیجه‌گیری از شکل (۳) نیز قابل استنتاج است. علاوه بر مدل شماره ۱، مشخص‌سازی مدل‌های شماره ۳ و ۵ نیز در سطح ۰/۰۵ رد می‌شود. این مدل‌ها کفایت لازم برای بررسی میزان ذرات گاز اوزون در مقابل سرعت وزش باد را ندارند. از بین مدل‌هایی که مشخص‌سازی آن‌ها با استفاده از آزمون مشخص‌سازی RESET<sub>2</sub> در سطح ۰/۰۵ رد نمی‌شوند. مدل‌های ۲ و ۸ از بالاترین سطح معنی‌داری برخوردار هستند و با توجه به شکل (۳) به‌خوبی بر داده‌ها برازش شده‌اند. همچنین با توجه به این‌که مدل شماره ۲ در مقایسه با مدل شماره ۸ از سطح معنی‌داری بالاتر و فرمول ساده‌تری برخوردار است، به‌عنوان مدل خطی عام مناسب بر این داده‌ها انتخاب می‌شود.

**جدول (۵):** عملکرد مدل‌های مختلف بر میزان گاز اوزون در مقابل سرعت باد. برآورد هر یک از پارامترها با  $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$  نشان داده شده است. برای قید RESET مقادیر آماره آزمون با  $F$ -value و مقادیر معنی‌داری آزمون‌ها با  $P$ -value داده شده است. همچنین، برای هر مدل، ردیف بالا بیانگر نتایج روش CC و ردیف پایین بیانگر نتایج روش IPW است.

مدل	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$F$ -value	$p$ -value
۱ $y = \beta_1 + \beta_2 x$	-۰/۰۱۶۱	-۰/۵۹۲۸	-	-	۱۵/۱۴۶۶	۰/۰۰۰۰
۲ $y = \beta_1 x + \beta_2 x^2$	-۰/۰۲۴۳	-۰/۵۸۷۶	-	-	۱۵/۹۳۳۸	۰/۰۰۰۰
۳ $y = \beta_1 x + \beta_2 x^2$	۰/۲۶۰۱	۰/۱۴۹۷	-	-	۱۱/۳۳۲۹	۰/۰۰۰۴
۴ $y = \beta_1 + \beta_2 x + \beta_3 x^2$	-۰/۲۶۸۲	-۰/۶۹۶۸	۰/۲۴۴۱	-	۲/۸۳۶۴	۰/۰۶۲۹
۵ $y = \beta_1 + \beta_2 x^2 + \beta_3 x^3$	-۰/۳۲۷۱	۰/۳۸۸۵	-۰/۱۷۱۳	-	۸/۲۸۹۲	۰/۰۰۰۴
۶ $y = \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	-۰/۵۸۲۷	۰/۱۷۷۷	-۰/۰۲۵۰	-	۲/۰۳۲۲	۰/۱۳۵۹
۷ $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3$	-۰/۳۰۰۰	-۰/۵۵۸۵	۰/۲۹۸۸	-۰/۰۵۰۰	۱/۷۴۶۲	۰/۱۷۹۲
۸ $y = \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$	-۰/۳۸۹۲	۰/۱۵۵۱	-۰/۱۷۱۶	۰/۰۱۶۳	۱/۴۱۴۳	۰/۲۴۷۵
	-۰/۴۰۶۴	۰/۱۵۴۵	-۰/۱۵۷۸	۰/۰۱۴۹	۱/۳۰۴۶	۰/۲۶۰۶

### ۵- بحث و نتیجه‌گیری

در مبحث شبیه‌سازی برای برآورد مقادیر مجهول از تابع هسته نرمال استاندارد استفاده کردیم. به طوری که پارامتر هموارسازی آن را با استفاده از روش اعتبارسنجی<sup>۱</sup> متقابل برآورد کردیم. همچنین، برای به دست آوردن توان تجربی و سطح معناداری آزمون‌ها به ترتیب به چندک‌ها و تابع توزیع شبه-فیشر نیاز داشتیم که به صورت صریح قابل محاسبه نبودند. به همین دلیل، از روش شبیه‌سازی برای برآورد این مقادیر استفاده کرده‌ایم. از ۱۰۰۰ تکرار الگوریتم مونت کارلو با حجم نمونه ۵۰۰ برای برآورد مقادیر موردنظر استفاده کرده‌ایم.



شکل (۳): نمودار مدل‌های برازش شده: نمودار با خط ممتد بیانگر روش CC و نمودار نقطه‌چین بیانگر روش IPW است.

وقتی داده گم شده در متغیرها وجود دارد، از روش‌هایی استفاده کرده‌ایم که مشاهدات ناقص را کنار می‌گذارد که باعث کاهش درجه آزادی توزیع آماره آزمون‌ها می‌شود. اگرچه، روش‌هایی وجود دارند که از مشاهدات ناقص در استنباط مدل استفاده می‌کنند. کار بعدی ما می‌تواند، استفاده از مدل‌های برآورد و ساخت تابع آزمون‌هایی باشد که مانع از کاهش درجه آزادی تابع آزمون‌ها می‌شود.

از روش‌های ناپارامتری برای برآورد مقدار مجهول  $\pi(\cdot)$  استفاده کردیم. هرچند، روش‌های دیگری از جمله روش برآورد رگرسیون لوژستیک وجود دارند که ممکن است در برخی از حالات عملکرد بهتری نسبت به روش برآورد ناپارامتری داشته باشند. از طرفی، چون برآورد ناپارامتری در این حالت همواره عملکرد قابل قبولی دارد، با استفاده از این روش، مقادیر نامعلوم برآورد شده است.

در مبحث شبیه‌سازی مشاهده کردیم که استفاده از روش‌های نوین می‌تواند در استنباط‌های آماری با داده‌های ناقص مفید باشد. در کل آزمون مشخص‌سازی رمزی با استفاده از روش برآورد IPW در مقایسه با روش برآورد CC عملکرد قابل قبول‌تری دارد. هر چند به ازای الگوی گم‌شدگی MCAR، عملکرد آزمون مشخص‌سازی رمزی با استفاده از روش برآورد CC بهبود می‌یابد اما همچنان با به‌کارگیری روش برآورد IPW به ازای هر دو الگوی گم‌شدگی نتایج بهتری نسبت به روش CC حاصل می‌شود. همچنین، با توجه به این‌که تشخیص الگوی گم‌شدگی داده‌ها کار بسیار پیچیده‌ای است، استفاده از روش IPW توصیه می‌شود. بنابراین آزمون مشخص‌سازی رمزی با استفاده از روش برآورد وزن معکوس احتمال می‌تواند یک گزینه مناسب در تعیین صحت مدل باشد.

#### منابع

- [1] Madsen, H. and Thyregod, P. 2010. *Introduction to general and generalized linear models*. CRC Press.
- [2] Ramsey, G. B. (1969). Test for specification error in classical linear least square regression analysis. *Journal of the Royal Statistical Society*, **31**, 350-71.
- [3] Griffith, D. A. and Chun, Y. (2016). Evaluating eigenvector spatial filter corrections for omitted georeference variables. *Econometrics*, **21**, 1-12.
- [4] Shukur, G. and Mantalos, P. (2004). Size and power of the RESET test as applied to systems of equations. A Bootstrap Approach, *Journal of Modern Applied Statistical Methods*, **3**, 370-385.
- [5] Sapra, S. (2005). A regression error specification test (RESET) for generalized linear model. *Economics Bulletin*, **3**, 1-6.
- [6] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.



- [7] Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data. Second Edition*, Wiley-Interscience, New York.
- [8] Basilevsky, A., Sabourin, D., Hum, D. and Anderson, A. (1985). Missing data estimators in the general linear model: an evaluation of simulated data as an experimental design. *Communications in Statistics-Simulation and Computation*, **14**(2), 371-394.
- [9] Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, **87**, 1227-1237.
- [10] Wang, S. and Wang, C.Y. (2001). A note on kernel assisted estimators in missing covariate regression. *Statistics and Probability letters*, **55**, 439-449.
- [11] Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, **21**, 1921-1947.
- [12] Hardle, W., Mammen, E. and Muller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, **93**(444), 1461-1474.
- [13] Zhu, L.X. and Cui, H.J. (2005). Testing lack-of-fit for general linear errors in variables models. *Statistica Sinica*, **15**, 1049--1068.
- [14] Guo, X. and Xu, W. (2012). Goodness-of-fit tests for general linear models with covariates missed at random. *Journal of Statistical Planning and Inference*, **142**, 2047-2058.
- [15] Li, X. (2012). Lack-of-fit testing of a regression model with response missing at random. *Journal of Statistical Planning and Inference*, **142**(1), 155-170.
- [16] Zhao, L. P. and Lipsitz, S. (1992). Design and analysis of two-stage studies. *Statistics in Medicine*, **11**, 769-782.
- [17] Carpenter, J. R. and Kenward, M. G. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal statistical Society*, **169**, 571-584.
- [18] Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical methods for data analysis*. Belmont, CA: Wadsworth.

## **General Linear Model Specification Error Test with Missing Data**

Fayyaz Bahari\*, Safar Parsi\*, and Mojtaba Ganjali\*\*.

\*Department of Statistics and Computer Sciences, University of Mohaghegh Ardabili, Ardabil, Iran.

\*\*Department of Statistics, Shahid Beheshti University, Tehran, Iran.

### **Abstract**

In this paper, we consider a general linear model where missing data may occur in response and covariate variables. We propose a new test based on Ramsey's test to identify goodness of fit for general linear model with missing data. We show that under the null hypothesis, our test functions for complete case analysis follow a Fisher distribution and the other test function used for analysis with available data converges in distribution to Quasi-Fisher distribution. Furthermore, we compare proposed test functions by using some simulation studies. Also, we apply our methods in analyzing a real data set.

**Keywords:** General linear model, Missing data, Goodness of fit, Ramsey's test, Quasi-Fisher distribution.

**Mathematics Subject Classification (2010):** 62J02, 62F03.

