آزمون تعیین خطای مشخص‌سازی مدل خطی عام با داده‌های گم‌شده

فیاض بهاری*، صفر پارسی*[۱] و مجتبی گنجعلی**

*گروه آمار و علوم کامپیوتر، دانشگاه محقق اردبیلی

**گروه آمار، دانشگاه شهید بهشتی

تاریخ دریافت: ۱۳۹۶/۱۰/۲۵ تاریخ پذیرش: ۱۳۹۷/۷/۲۳

چکیده: در این مقاله مدل خطی عام را در تحلیل داده‌هایی که در آن‌ها متغیرهای کمکی و متغیر پاسخ گم‌شدگی دارند، در نظر می‌گیریم. برای تعیین نیکویی برازش مدل خطی عام در حضور داده‌های گم‌شده، آزمون جدیدی را بر اساس آزمون رمزی می‌سازیم. نشان می‌دهیم که تحت فرض صفر، آماره‌های آزمون در برخی از حالات از توزیع فیشر تبعیت می‌کند و در برخی از حالات به توزیع شبه-فیشر میل می‌کند. علاوه بر این، عملکرد آماره‌های آزمون‌های مختلف را بر اساس چند مسئله شبیه‌سازی مقایسه می‌کنیم. همچنین از این آماره‌ها برای بررسی کفایت مدل خطی عام در تحلیل داده‌های واقعی استفاده می‌کنیم.

واژه‌های کلیدی : مدل خطی عام، داده‌های گم‌شده، آزمون نیکویی برازش، توزیع شبه-فیشر، آزمون رمزی.

رده‌بندی موضوعی (۲۰۱۰): ۶۲J۰۲، ۶۲F۰۳.

۱-مقدمه

مدل خطی عام[۲] یکی از روش‌های متداول در تعیین رابطه بین پدیده‌ها است. یک مدل خطی عام بیانگر رابطه بین متغیر پاسخ y و بردار متغیرهای کمکی $\boldsymbol{x}$ با بعد m است. مدل خطی عام از رابطه زیر پیروی می‌کند.

$$y = g^T(\mathbf{x})\boldsymbol{\beta} + \varepsilon, \qquad (۱)$$

۱- آدرس الکترونیکی مسئول مقاله: parsi@uma.ac.ir

2- General Linear Model

که در آن $g(\cdot)$ یک تابع برداری معلوم با بعد p، $\boldsymbol{\beta}$ یک پارامتر برداری با بعد p و ε 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؛ به‌طوری‌که 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[1] (RESET) را معرفی کرد که یک ابزار مناسب برای تشخیص عدم‌کفایت مدل است. در برخی از مطالعات ممکن است، برخی از متغیرهای کمکی کنار گذاشته شوند. این امر باعث می‌شود تا یک اریبی حذف متغیر[2] (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[3]، کاربرد فراوانی پیدا کرده است، به‌طوری‌که 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] موردبررسی قرار گرفت. نتایج بررسی آن‌ها نشان می‌دهد که با افزایش عدم‌کفایت 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[4] 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[5] 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 قابل‌چشم‌پوشی[۱] پیروی می‌کنند. پس فرض بر این است که یا الگوی گم‌شدگی داده‌ها، بر مبنای تعریف [۷]، به‌صورت کاملاً تصادفی[۲] (MCAR) و یا به‌صورت تصادفی[۳] (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 جانهی داده‌های گم‌شده موردمطالعه قرار گرفته است و روش‌های برآورد داده ها بحث شده است. اما اخیراً توجه بیشتر محققین بر تصحیح معادلات برآورد است تا بتوان با تصحیح معادلات برآورد، برآوردگرهای نااریب 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 اندازه گیری شده باشد، [۱۳] آزمون نوع امتیاز[۴] را برای بررسی کفایت مدل خطی عام به کار بردد. در حالتی که الگوی گم‌شدگی داده‌ها قابل‌چشم‌پوشی باشد، برخی از محققین راهکارهایی معرفی کردند. از جمله، [۱۴] یک آزمون مناسب دیگر بر اساس تابع امتیاز معرفی کرد. به‌علاوه، [۱۵] روش آزمون مناسبی را بر اساس انتگرال کمترین توان دوم فاصله بین برازش‌های ناپارامتری و پارامتری[۵] ساخت، به‌طوری‌که داده‌های گم‌شده با یک الگوی قابل‌چشم‌پوشی در داده‌ها رخ می دهند.

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] به‌صورت نظری به آن پرداخته است و آن را به مدل‌های خطی تعمیم‌یافته، 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:

$$\sum_{i=1}^{n} \boldsymbol{\psi}_{\boldsymbol{\beta}}(y_i \mid \mathbf{x}_i) = 0, \qquad (۲)$$

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

$$y = g^T(\mathbf{x})\boldsymbol{\beta} + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \cdots + \gamma_q \hat{y}^{q+1} + \eta, \qquad (۳)$$

که در آن $\gamma_1,\cdots,\gamma_q$ پارامترهای جدید مدل خطی عام هستند و η خطای جدید مدل خطی عام است که از توزیع $N(0,\nu^2)$ پیروی می‌کند. آزمون رمزی فرضیات زیر را برای بررسی کفایت مدل خطی عام در نظر می‌گیرد:

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

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 $(g^T(\mathbf{x})\boldsymbol{\beta})^k$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 γ_iها رد شود. در این حالت، مدل رگرسیونی تحت رابطه (۱) ، مدل نامقید[1] و مدل رگرسیونی تحت رابطه (۳)، مدل مقید[2] 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 به‌صورت 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 به‌صورت 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 به‌صورت زیر تعریف می‌شود:

$$F_{Full} = \frac{(SSE_{U,Full} - SSE_{R,Full})/q}{SSE_{R,Full}/(n-p-q)}. \qquad (۴)$$

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 موردبحث 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 قابل‌چشم‌پوشی 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 به‌صورت هم‌زمان رخ دهند. تنها کافی است که یکی از متغیرها به‌صورت کامل مشاهده‌شده 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 به‌صورت ناقص مشاهده‌شده‌اند و استنباط‌ها با داده‌هایی انجام می‌شود که به‌صورت کامل مشاهده‌شده‌اند. این روش، روش حالت کامل[1] (CC) نامیده می‌شود. بنابراین در این حالت، پارامترهای مدل خطی عام از رابطه زیر به دست می‌آید:

$$\sum_{i=1}^{n} \delta_i \boldsymbol{\psi}_{\boldsymbol{\beta}}(y_i \mid \mathbf{x}_i) = 0, \qquad (۵)$$

که در آن، δ_i یک تابع نشانگر است و درصورتی‌که i-امین داده به‌صورت کامل مشاهده‌شده باشد، برابر یک و در غیر این صورت برابر صفر است. در این حالت تابع آزمون رمزی به‌صورت زیر خواهد بود:

$$F_{CC} = \frac{(SSE_{U,CC} - SSE_{R,CC})/q}{SSE_{R,CC}/(\sum_{i=1}^{n}\delta_i - p - q)}, \qquad (۶)$$

که در آن، $SSE_{R,CC}$ و $SSE_{U,CC}$ 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:

$$SSE_{U,CC} = \sum_{i=1}^{n} \delta_i (y_i - \hat{y}_{Ui})^2, \qquad (۷)$$

و بر اساس مدل مقید، داریم:

$$SSE_{R,CC} = \sum_{i=1}^{n} \delta_i (y_i - \hat{y}_{Ri})^2. \qquad (۸)$$

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[2] (IPW) استفاده می‌شود که اولین بار توسط [۱۶] معرفی شد. این روش برآوردهای نااریب 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] موردبررسی قرار گرفته است. درروش، IPW پارامترهای مدل خطی عام از حل معادلات برآورد زیر به دست می‌آید:

$$\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} \boldsymbol{\psi}_{\boldsymbol{\beta}}(y_i \mid \mathbf{x}_i) = 0, \tag{۹}$$

که در آن، π_i احتمال مشاهده i-امین عضو نمونه است. در مطالعات واقعی، مقدار π_i نامعلوم است و می‌توان آن را به‌صورت ناپارامتری زیر برآورد کرد:

$$\hat{\pi}(v_i) = \frac{\sum_{j=1}^{n} \delta_j K_h(v_j - v_i)}{\sum_{j=1}^{n} K_h(v_j - v_i)}, \tag{۱۰}$$

که در آن، v_i، i-امین جزء متغیرهای کاملاً مشاهده‌شده است و $K_h(\cdot)$ یک تابع هسته[1] با پارامتر هموارسازی[2] 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.

درروش وزن معکوس احتمال، آماره آزمون رمزی به‌صورت زیر خواهد بود:

$$F_{IPW} = \frac{(SSE_{U,IPW} - SSE_{R,IPW})/q}{SSE_{R,IPW}/(\sum_{i=1}^{n} \delta_i - p - q)}, \tag{۱۱}$$

که در آن، $SSE_{R,IPW}$ و $SSE_{U,IPW}$ 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:

$$SSE_{U,IPW} = \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} (y_i - \hat{y}_{Ui})^2, \tag{۱۲}$$

و بر اساس مدل مقید، داریم:

$$SSE_{R,IPW} = \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} (y_i - \hat{y}_{Ri})^2. \tag{۱۳}$$

1- Kernel Function
2 -Smoothing Parameter

۲-۳- خواص توزیعی آماره‌های آزمون

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

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

$$F = \frac{\mathbf{y}^T A \mathbf{y} / df_1}{\mathbf{y}^T B \mathbf{y} / df_2}, \qquad (۱۴)$$

آنگاه، می‌گوییم، F دارای توزیع شبه-فیشر $QF(df_1, df_2, A, B)$ است. به‌طوری‌که $\mathbf{y} = (y_1, \cdots, y_n)^T$، 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 بیان‌شده 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. به‌عبارت‌دیگر، توزیع فیشر حالت خاصی از توزیع شبه-فیشر است.

حال، می‌توانیم خواص توزیعی آماره‌ها را به دست آوریم. این ویژگی‌ها در قضیه زیر داده‌شده 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 به‌صورت زیر تعریف می‌کنیم.

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

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

$$Z = \begin{bmatrix} \hat{y}^{2}_{U,1} & \hat{y}^{3}_{U,1} & \cdots & \hat{y}^{q+1}_{U,1} \\ \hat{y}^{2}_{U,2} & \hat{y}^{3}_{U,2} & \cdots & \hat{y}^{q+1}_{U,2} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}^{2}_{U,n} & \hat{y}^{3}_{U,n} & \cdots & \hat{y}^{q+1}_{U,n} \end{bmatrix}.$$

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

$$\begin{aligned} SSE_{R,Full} &= \sum_{i=1}^{n}(y_i - \hat{y}_{Ri})^2 = (\mathbf{y} - \hat{\mathbf{y}}_R)^T(\mathbf{y} - \hat{\mathbf{y}}_R) \\ &= (\mathbf{y} - G(G^TG)^{-1}G^T\mathbf{y})^T(\mathbf{y} - G(G^TG)^{-1}G^T\mathbf{y}) \\ &= \mathbf{y}^T(I - H)^T(I - H)\mathbf{y}, \end{aligned}$$

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

$$SSE_{R,Full} = \mathbf{y}^T(I - H)\mathbf{y}. \quad (۱۵)$$

پس، خواهیم داشت:

$$\frac{SSE_{R,Full}}{\sigma^2} \sim \chi^2(n - p - q). \quad (۱۶)$$

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

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

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

$$\begin{aligned} SSE_{U,Full} - SSE_{R,Full} &= \mathbf{y}^T(G(G^TG)^{-1}G^T - g(X)(g^T(X)g(X))^{-1}g^T(X))\mathbf{y} \\ &:= \mathbf{y}^T H^* \mathbf{y}. \end{aligned} \quad (۱۷)$$

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

$$\frac{SSE_{U,Full} - SSE_{R,Full}}{\sigma^2} \sim \chi^2(q). \qquad (۱۸)$$

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

$$\begin{aligned} df &= trace(G(G^TG)^{-1}G^T - g(X)(g^T(X)g(X))^{-1}g^T(X)) \\ &= p+q-p=q. \end{aligned} \qquad (۱۹)$$

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

درروش وزن معکوس احتمال، می‌توان پارامترهای مدل خطی عام را با استفاده از معادلات برآورد رابطه (۹) به‌صورت زیر استخراج کرد.

$$\hat{\boldsymbol{\beta}} = (X^T\hat{W}X)^{-1}X^T\hat{W}\mathbf{y} \qquad (۲۰)$$

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:

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

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:

$$SSE_{U,IPW}=\mathbf{y}^T(W-Wg(X)(g^T(X)Wg(X))^{-1}g^TW)\mathbf{y}+o_P(1). \quad (۲۱)$$

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

$$SSE_{R,IPW}=\mathbf{y}^T(W-WG(G^TWG)^{-1}G^TW)\mathbf{y}+o_P(1). \quad (۲۲)$$

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

$$\begin{aligned} F_{IPW} &= \frac{(SSE_{U,IPW}-SSE_{R,IPW})/q}{SSE_{R,IPW}/(\sum_{i=1}^{n}\delta_i-p-q)} \\ &= \frac{(\mathbf{y}^T(W[G(G^TWG)^{-1}G^T-g(X)(g^T(X)Wg(X))^{-1}g^T(X)]W)\mathbf{y})/q}{(\mathbf{y}^T(W-WG(G^TWG)^{-1}G^TW)\mathbf{y})/(\sum_{i=1}^{n}\delta_i-p-q)}+o_P(1) \end{aligned} \quad (۲۳)$$

در نهایت می‌توان آماره آزمون روش IPW را به‌صورت زیر نوشت:

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

حال، با استفاده از تعریف ۱ و قضیه مینکوسکی[1] قسمت (ج) قضیه ۱ نتیجه می‌شود. همچنین، درجات آزادی به‌صورت مشابه با اثبات قسمت (ب) قضیه ۱ نتیجه می‌شود. ■

در قضیه ۱، آماره آزمون‌ها درروش داده‌های کامل و روش حالت کامل از توزیع دقیق فیشر پیروی می‌کنند. همچنین، درروش 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 $\lfloor \cdot \rfloor$ 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 مونت‌کارلو 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

1- Minkowski

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:

$$y = g^T(\mathbf{x})\boldsymbol{\beta} + ch(\mathbf{x}) + \varepsilon, \qquad (۲۴)$$

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 $c = ۰$، 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.

$$RESET\,۱:\ y = g^T(\mathbf{x})\boldsymbol{\beta} + \gamma_۱\hat{y}^۲ + \eta,$$
$$RESET\,۲:\ y = g^T(\mathbf{x})\boldsymbol{\beta} + \gamma_۱\hat{y}^۲ + \gamma_۲\hat{y}^۳ + \eta.$$

در هر مرحله، توان تجربی رد فرض صفر با استفاده از ۲۰۰۰ تکرار الگوریتم مونت‌کارلو[1] برآورد شده است.

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

در رابطه (۲۴)، فرض کنید،

$$h(\mathbf{x}) = \sqrt{۱ + sin(۲\pi x_۱) + \frac{(۱ + x_۲^۲)}{۴}},$$
$$g(\mathbf{x}) = (۱ + sin(۲\pi x_۱), \frac{(۱ + x_۲^۲)}{۴})^T,$$

که در آن، $x_۱$ از توزیع یکنواخت $U(۰,۱)$ تولید می‌شود، $x_۲$ از توزیع نرمال $N(۰,۱)$ تولید می شود، ε یک متغیر تصادفی از توزیع نرمال $N(۰,۰/۲۵)$ است و پارامترهای مدل خطی عام برابر با بردار $(۱,۲)$ در نظر گرفته‌شده است.

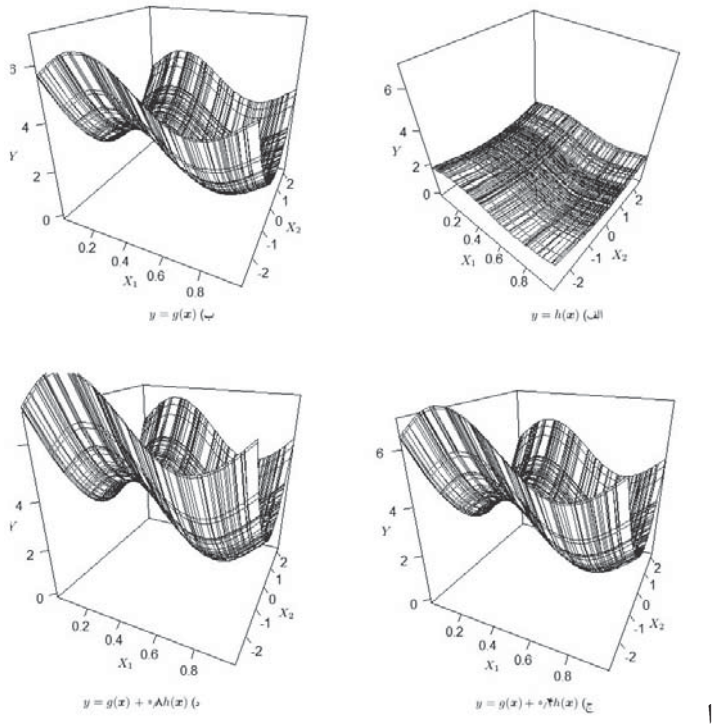
الگوهای گم‌شدگی زیر را برای تولید داده‌های گم‌شده در متغیر پاسخ در نظر می‌گیریم:

حالت اول: $\pi_۱(\mathbf{x}_i) = ۱/(۱ + e^{-(۰/۲۵ + ۰/۵((۱ + sin(۲\pi x_{۱i})) + ۰/۵\frac{(۱ + x_{۲i}^۲)}{۴})})}$،

حالت دوم: $\pi_۲(\mathbf{x}_i) = ۰/۷۳$.

1 -Monte Carlo

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) داده‌شده است.

جدول (۱): مرحله اول: توان تجربی رد فرض صفر بر اساس قید RESET۱ به ازای حجم نمونه ۵۰ و ۱۰۰ (Full: حالت داده‌های کامل؛ CC: حالت کامل؛ IPW: وزن معکوس احتمال).

الگوی گم‌شدگی	C	n = ۵۰			n = ۱۰۰		
		Full	CC	IPW	Full	CC	IPW
$\pi_1: MAR$	۰/۰۰	۰/۰۴۳۵	۰/۰۵۰۰	۰/۰۵۲۵	۰/۰۵۳۰	۰/۵۱۰	۰/۰۴۹۵
	۰/۱۰	۰/۰۶۰۰	۰/۰۵۸۵	۰/۰۶۱۵	۰/۰۷۲۵	۰/۰۶۳۰	۰/۰۶۸۵
	۰/۲۰	۰/۱۰۱۵	۰/۰۸۸۰	۰/۰۸۷۰	۰/۱۳۶۵	۰/۱۱۱۵	۰/۱۱۳۵
	۰/۳۰	۰/۱۵۲۰	۰/۱۲۹۵	۰/۱۲۹۰	۰/۲۸۵۰	۰/۱۹۱۰	۰/۲۰۳۰
	۰/۴۰	۰/۲۲۴۵	۰/۱۷۷۵	۰/۱۸۳۰	۰/۴۲۴۵	۰/۳۰۲۵	۰/۳۲۱۵
	۰/۵۰	۰/۳۲۴۰	۰/۲۳۴۵	۰/۲۴۵۰	۰/۵۷۸۵	۰/۴۳۸۵	۰/۴۵۹۵
	۰/۶۰	۰/۴۳۴۵	۰/۳۱۰۰	۰/۳۳۳۵	۰/۷۳۸۵	۰/۵۸۷۰	۰/۶۱۲۰
	۰/۷۰	۰/۵۴۹۰	۰/۴۱۲۵	۰/۴۲۳۵	۰/۸۵۸۵	۰/۷۲۵۰	۰/۷۴۶۵
	۰/۸۰	۰/۶۶۲۵	۰/۵۰۵۵	۰/۵۱۹۵	۰/۹۳۰۰	۰/۸۲۵۵	۰/۸۴۵۰
	۰/۹۰	۰/۷۶۳۰	۰/۵۸۹۵	۰/۶۰۲۵	۰/۹۶۸۵	۰/۹۰۰۰	۰/۹۱۰۰
	۱/۰۰	۰/۸۳۹۰	۰/۶۸۶۸	۰/۶۹۸۵	۰/۹۸۸۰	۰/۹۴۸۵	۰/۹۵۶۰
	۱/۱۰	۰/۸۹۵۰	۰/۷۶۵۵	۰/۷۷۷۵	۰/۹۹۴۵	۰/۹۷۵۵	۰/۹۷۷۵
	۱/۲۰	۰/۹۳۰۰	۰/۸۳۵۰	۰/۸۴۴۰	۰/۹۹۹۰	۰/۹۸۴۵	۰/۹۸۵۵
$\pi_2: MCAR$	۰/۰۰	۰/۰۴۳۵	۰/۰۵۳۵	۰/۰۵۶۰	۰/۰۵۳۰	۰/۰۵۸۵	۰/۰۵۹۵
	۰/۱۰	۰/۰۶۵۵	۰/۰۶۰۰	۰/۰۶۳۵	۰/۰۷۲۵	۰/۰۶۵۰	۰/۰۶۰۰
	۰/۲۰	۰/۱۰۱۵	۰/۰۸۶۵	۰/۰۸۶۵	۰/۱۳۶۵	۰/۱۱۲۰	۰/۱۱۷۰
	۰/۳۰	۰/۱۵۲۰	۰/۱۲۱۰	۰/۱۲۲۵	۰/۲۵۸۰	۰/۱۹۸۰	۰/۲۰۵۵
	۰/۴۰	۰/۲۲۴۵	۰/۱۷۴۰	۰/۱۸۱۵	۰/۴۱۴۵	۰/۳۰۲۵	۰/۳۱۲۵
	۰/۵۰	۰/۳۲۴۰	۰/۲۴۷۰	۰/۲۵۲۰	۰/۵۷۸۵	۰/۴۴۶۰	۰/۴۵۹۵
	۰/۶۰	۰/۴۳۴۵	۰/۳۳۳۵	۰/۳۳۹۰	۰/۷۳۸۵	۰/۵۸۸۰	۰/۶۰۰۰
	۰/۷۰	۰/۴۴۹۰	۰/۴۱۹۵	۰/۴۲۳۰	۰/۸۵۸۵	۰/۷۲۷۰	۰/۷۳۰۵
	۰/۸۰	۰/۶۶۲۵	۰/۵۱۱۰	۰/۵۲۲۰	۰/۹۳۰۰	۰/۸۲۷۰	۰/۸۳۴۰
	۰/۹۰	۰/۷۶۳۰	۰/۶۰۰۰	۰/۶۱۱۵	۰/۹۶۸۵	۰/۹۰۰۰	۰/۹۰۴۵
	۱/۰۰	۰/۸۳۹۰	۰/۶۹۷۰	۰/۶۹۹۰	۰/۹۸۸۰	۰/۹۴۶۰	۰/۹۴۵۰
	۱/۱۰	۰/۸۹۵۰	۰/۷۷۷۵	۰/۷۸۰۵	۰/۹۹۴۵	۰/۹۶۹۵	۰/۹۶۹۰
	۱/۲۰	۰/۹۳۰۰	۰/۸۳۵۵	۰/۸۳۸۵	۰/۹۹۹۰	۰/۹۸۳۵	۰/۹۸۳۵

جدول (۲): مرحله اول: توان تجربی رد فرض صفر بر اساس قید RESET۲ به ازای حجم نمونه ۵۰ و ۱۰۰ (Full: حالت داده‌های کامل؛ CC: حالت کامل؛ IPW: وزن معکوس احتمال).

الگوی گم‌شدگی	C	n = ۵۰			n = ۱۰۰		
		Full	CC	IPW	Full	CC	IPW
$\pi_1: MAR$	۰/۰۰	۰/۰۵۰۰	۰/۰۵۲۵	۰/۰۴۹۰	۰/۰۵۲۵	۰/۰۴۶۵	۰/۰۴۵۰
	۰/۱۰	۰/۰۶۳۰	۰/۰۶۳۵	۰/۰۵۸۵	۰/۰۷۰۵	۰/۰۶۱۰	۰/۰۶۰۵
	۰/۲۰	۰/۰۸۵۰	۰/۰۷۸۵	۰/۰۸۰۵	۰/۱۲۴۰	۰/۱۰۳۰	۰/۱۰۴۰
	۰/۳۰	۰/۱۳۲۵	۰/۱۰۶۰	۰/۱۱۳۵	۰/۲۳۹۰	۰/۱۷۴۵	۰/۱۷۷۵
	۰/۴۰	۰/۱۹۶۵	۰/۱۵۲۰	۰/۱۶۲۰	۰/۳۷۷۰	۰/۲۷۱۰	۰/۲۹۱۵
	۰/۵۰	۰/۲۸۵۰	۰/۲۰۵۰	۰/۲۱۳۰	۰/۵۳۵۵	۰/۴۰۰۰	۰/۴۱۶۵
	۰/۶۰	۰/۳۷۸۰	۰/۲۷۵۰	۰/۲۸۶۵	۰/۶۸۸۵	۰/۵۳۷۰	۰/۵۵۲۰
	۰/۷۰	۰/۵۰۳۰	۰/۳۵۴۵	۰/۳۶۷۵	۰/۸۲۹۵	۰/۶۷۵۵	۰/۶۹۷۰
	۰/۸۰	۰/۶۰۲۵	۰/۴۵۳۵	۰/۴۵۴۵	۰/۹۰۹۰	۰/۸۰۱۵	۰/۸۱۴۰
	۰/۹۰	۰/۷۰۴۵	۰/۵۴۰۰	۰/۵۴۸۵	۰/۹۵۲۵	۰/۸۷۷۰	۰/۸۸۶۰
	۱/۰۰	۰/۷۹۲۵	۰/۶۲۱۰	۰/۶۲۵۵	۰/۹۸۶۰	۰/۹۳۵۵	۰/۹۳۶۰
	۱/۱۰	۰/۸۶۶۰	۰/۷۱۰۰	۰/۷۰۹۰	۰/۹۹۵۰	۰/۹۶۴۵	۰/۹۶۷۵
	۱/۲۰	۰/۹۱۴۵	۰/۷۹۴۵	۰/۷۸۵۵	۰/۹۹۸۰	۰/۹۸۴۵	۰/۹۸۲۵
$\pi_2: MAR$	۰/۰۰	۰/۰۵۰۰	۰/۰۵۶۰	۰/۰۵۲۰	۰/۰۵۲۵	۰/۰۵۱۵	۰/۰۵۳۰
	۰/۱۰	۰/۰۶۳۰	۰/۰۵۸۰	۰/۰۵۸۰	۰/۰۷۰۵	۰/۰۶۲۵	۰/۰۶۴۵
	۰/۲۰	۰/۰۸۵۰	۰/۰۷۷۵	۰/۰۷۷۰	۰/۱۲۴۰	۰/۰۹۵۵	۰/۱۰۰۵
	۰/۳۰	۰/۱۳۲۵	۰/۱۰۸۰	۰/۱۰۷۵	۰/۲۳۹۰	۰/۱۶۶۵	۰/۱۶۸۰
	۰/۴۰	۰/۱۹۶۵	۰/۱۵۰۵	۰/۱۵۹۵	۰/۳۷۷۰	۰/۲۶۹۵	۰/۲۷۵۵
	۰/۵۰	۰/۲۸۵۰	۰/۲۰۹۰	۰/۲۱۵۰	۰/۵۳۵۵	۰/۴۰۶۵	۰/۴۱۴۷
	۰/۶۰	۰/۳۷۸۰	۰/۲۷۷۵	۰/۲۸۰۰	۰/۶۸۸۵	۰/۵۴۲۰	۰/۵۴۶۰
	۰/۷۰	۰/۵۰۳۰	۰/۳۶۸۰	۰/۳۷۰۵	۰/۸۲۹۸	۰/۶۷۹۵	۰/۶۷۸۵
	۰/۸۰	۰/۶۰۲۵	۰/۴۵۱۵	۰/۴۵۴۰	۰/۹۰۹۰	۰/۸۰۳۰	۰/۸۰۱۰
	۰/۹۰	/۰۷۰۴۵	۰/۵۴۳۰	۰/۵۵۰۵	۰/۹۸۶۰	۰/۹۲۵۰	۰/۹۲۵۵
	۱/۰۰	۰/۷۹۲۵	۰/۶۳۲۰	۰/۶۴۰۵	۰/۹۸۶۰	۰/۹۲۵۰	۰/۹۲۵۵
	۱/۱۰	۰/۸۶۶۰	۰/۷۱۷۵	۰/۷۱۱۵	۰/۹۹۵۰	۰/۹۵۹۵	۰/۹۶۱۰
	۱/۲۰	۰/۹۴۵۰	۰/۷۸۸۰	۰/۷۸۳۰	۰/۹۹۸۰	۰/۹۸۱۵	۰/۹۷۸۵

نمودار (ج) بیانگر مدل خطی عام تحت فرض مقابل با ضریب ثابت $c=۰/۴$ و نمودار (د) بیانگر مدل خطی عام تحت فرض مقابل با ضریب ثابت $c=۰/۸$ 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 ازلحاظ ظاهری تفاوت چندانی نداشته باشند. این ویژگی‌های تابع $h(\mathbf{x})$ باعث می‌شود تا با افزایش مقدار c، توان آزمون به‌صورت آهسته افزایش یابد. درصورتی‌که 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 پیشنهادشده در بدترین حالات نیز مطالعه شود. نتایج مرحله اول با توجه به آماره‌های RESET۱ و RESET۲ به ترتیب در جدول‌های (۱) و (۲) داده‌شده است.

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

برای مدل خطی عام رابطه (۲۴)، فرض کنید، $g(\mathbf{x})=(x_۱,\frac{(۱+x_۲^۲)}{۴},x_۳)^T$ و $h(\mathbf{x})=۳(x_۱+x_۱^۲)$. که در آن، $x_۱$ از توزیع نرمال $N(۰,۱)$، $x_۲$ از توزیع یکنواخت $U(۰,۱)$ و $x_۳$ از توزیع نرمال $N(۰,۱)$ تولید می‌شوند. ε خطای نمونه‌گیری تصادفی از توزیع نرمال $N(۰,۰/۲۵)$ و پارامترهای مدل خطی عام برابر با بردار $(۱,۲,۳)$ است.

الگوهای گم‌شدگی زیر را برای تولید داده‌های گم‌شده در متغیر پاسخ y و متغیر کمکی $x_۱$ در نظر می‌گیریم.

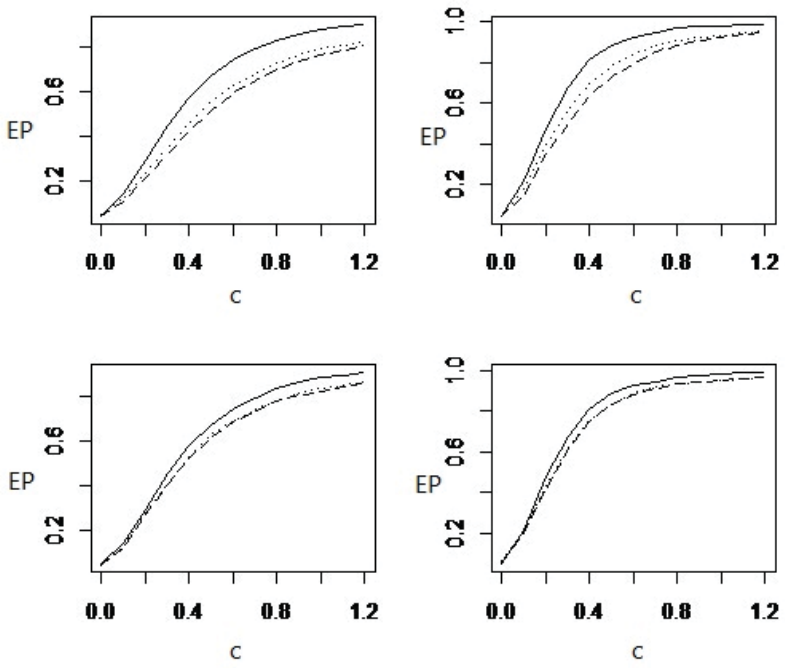
حالت سوم: $\pi_۳(\mathbf{x}_i)=۱/(۱+۰/۵e^{(-۰/۵x_{۳i}^۲)})$،

حالت چهارم: $\pi_۴(\mathbf{x}_i):=\begin{cases}\pi_{۴۱}(\mathbf{x}_i)=۱/(۱+۰/۵e^{(-۰/۵(x_{۱i}^۲+x_{۲i}+x_{۳i}^۲))})\\ \pi_{۴۲}(\mathbf{x}_i)=۱/(۱+۰/۵|۰/۲۵x_{۲i}+۰/۳x_{۳i}^۲|)\end{cases}$.

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 $x_۱$ یا به‌صورت هم‌زمان گم‌شده هستند و یا به‌صورت هم‌زمان مشاهده‌شده‌اند. اما در حالت چهارم، $۱-\pi_{۴۱}$ احتمال آن است که متغیر پاسخ y گم‌شده باشد. از طرف دیگر، $۱-\pi_{۴۲}$ احتمال آن است که متغیر کمکی $x_۱$، گم‌شده باشد، به‌شرط آن‌که متغیر پاسخ y مشاهده‌شده باشد. همچنین با استفاده از الگوی گم‌شدگی $\pi_{۴۱}$، ۱۵/۵٪ داده گم‌شده در متغیر پاسخ y خواهیم داشت و با استفاده از الگوی گم‌شدگی $\pi_{۴۲}$

، ۱۳/۵٪ داده گم‌شده در متغیر کمکی x_1 خواهیم داشت. بنابراین در حالت چهارم نیز به‌طور متوسط ۲۹٪ داده گم‌شده در داده‌ها خواهیم داشت.

نتایج مرحله دوم در جدول‌های (۳) و (۴) داده‌شده است. همچنین، نمودار توان تجربی رد فرض صفر به ازای قید RESET۱ در شکل (۲) داده‌شده است. در شکل (۲)، نمودارهای بالایی، توان تجربی رد فرض صفر را به ازای الگوی گم‌شدگی π_3 نشان می‌دهند. همچنین نمودارهای پایینی، توان تجربی رد فرض صفر را به ازای الگوی گم‌شدگی π_4 نشان می‌دهند. به‌علاوه، نمودارهای سمت چپ، بر اساس نمونه ۵۰ تایی و نمودارهای سمت راست، بر اساس نمونه ۱۰۰ هستند.



شکل (۲): نمودار توان تجربی رد فرض صفر برای مقادیر مختلف c تحت مرحله دوم بر اساس قید RESET۱: نمودارهای بالایی بر اساس الگوی گم‌شدگی π_3 و نمودارهای پایینی بر اساس الگوی گم‌شدگی π_4 هستند. نمودارهای سمت چپ بر اساس نمونه ۵۰ تایی و نمودارهای سمت راست بر اساس نمونه ۱۰۰ تایی هستند. همچنین، نمودارهای خط ممتد، نمودارهای خط تیره و نمودارهای نقطه‌ای به ترتیب توان تجربی آماره آزمون‌های روش داده‌های کامل، روش حالت کامل و روش وزن معکوس احتمال را نشان می‌دهند.

جدول (۳): مرحله دوم: توان تجربی رد فرض صفر بر اساس قید RESET۱ به ازای حجم نمونه ۵۰ و ۱۰۰ (Full: حالت داده‌های کامل؛ CC: حالت کامل؛ IPW: وزن معکوس احتمال).

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

جدول (۴): مرحله دوم: توان تجربی رد فرض صفر بر اساس قید RESET۲ به ازای حجم نمونه ۵۰ و ۱۰۰ (Full: حالت داده‌های کامل؛ CC: حالت کامل؛ IPW: وزن معکوس احتمال).

الگوی گم‌شدگی	C	n = ۵۰			n = ۱۰۰		
		Full	CC	IPW	Full	CC	IPW
$\pi_1: MAR$	۰/۰۰	۰/۰۵۴۰	۰/۰۴۹۰	۰/۰۵۰۰	۰/۰۴۵۵	۰/۰۴۸۵	۰/۰۴۸۵
	۰/۱۰	۰/۱۴۲۵	۰/۱۱۲۵	۰/۱۲۶۰	۰/۲۰۷۰	۰/۱۵۵۰	۰/۱۷۴۰
	۰/۲۰	۰/۲۷۸۰	۰/۲۰۵۵	۰/۲۲۵۰	۰/۴۵۳۵	۰/۳۳۶۵	۰/۳۸۶۵
	۰/۳۰	۰/۴۲۷۰	۰/۳۰۹۰	۰/۳۳۸۰	۰/۶۳۲۰	۰/۴۷۸۵	۰/۵۲۷۰
	۰/۴۰	۰/۵۵۱۵	۰/۴۰۵۵	۰/۴۳۶۰	۰/۷۷۰۰	۰/۵۹۹۰	۰/۶۶۴۵
	۰/۵۰	۰/۶۴۴۵	۰/۴۹۴۵	۰/۵۲۹۰	۰/۸۵۲۵	۰/۶۹۷۰	۰/۷۵۱۰
	۰/۶۰	۰/۷۱۳۵	۰/۵۶۸۵	۰/۶۰۱۵	۰/۹۰۸۰	۰/۷۶۷۰	۰/۸۱۳۵
	۰/۷۰	۰/۷۶۹۵	۰/۶۳۱۵	۰/۶۶۲۵	۰/۹۳۴۰	۰/۸۲۱۵	۰/۸۵۹۵
	۰/۸۰	۰/۸۰۸۰	۰/۶۷۶۰	۰/۷۰۶۰	۰/۹۵۳۰	۰/۸۶۲۰	۰/۸۹۶۵
	۰/۹۰	۰/۸۳۶۵	۰/۷۱۳۵	۰/۷۴۰۰	۰/۹۶۵۵	۰/۸۸۵۵	۰/۹۱۹۰
	۱/۰۰	۰/۸۶۱۵	۰/۷۳۸۰	۰/۷۷۱۰	۰/۹۷۳۰	۰/۹۱۱۰	۰/۹۲۹۰
	۱/۱۰	۰/۸۸۳۵	۰/۷۶۹۵	۰/۷۶۹۵	۰/۹۷۹۰	۰/۹۲۷۵	۰/۹۳۸۵
	۱/۲۰	۰/۸۹۱۵	۰/۷۹۵۵	۰/۸۱۵۰	۰/۹۸۶۵	۰/۹۳۵۵	۰/۹۴۵۰
$\pi_2: MAR$	۰/۰۰	۰/۰۵۴۰	۰/۰۵۵۵	۰/۰۵۵۰	۰/۰۴۵۵	۰/۰۴۸۰	۰/۰۴۹۵
	۰/۱۰	۰/۱۴۲۵	۰/۱۳۴۰	۰/۱۳۲۵	۰/۲۰۷۰	۰/۱۸۸۰	۰/۱۹۵۰
	۰/۲۰	۰/۲۷۸۰	۰/۲۶۳۰	۰/۲۶۸۰	۰/۴۵۳۵	۰/۳۹۹۰	۰/۳۹۸۰
	۰/۳۰	۰/۴۲۷۰	۰/۳۷۴۵	۰/۳۸۱۵	۰/۶۳۲۰	۰/۵۸۴	۰/۵۷۹۰
	۰/۴۰	۰/۵۵۱۵	۰/۴۸۳۰	۰/۴۹۲۵	۰/۷۷۰۰	۰/۷۰۷۵	۰/۷۱۲۰
	۰/۵۰	۰/۶۴۴۵	۰/۵۷۹۵	۰/۵۸۸۰	۰/۸۵۲۵	۰/۸۰۸۰	۰/۸۰۷۵
	۰/۶۰	۰/۷۱۳۵	۰/۶۴۸۵	۰/۶۵۱۵	۰/۹۰۸۰	۰/۸۶۰۰	۰/۸۵۹۰
	۰/۷۰	۰/۷۶۹۵	۰/۷۰۰۰	۰/۷۰۳۵	۰/۹۳۴۰	۰/۸۹۵۵	۰/۸۹۷۵
	۰/۸۰	۰/۸۰۸۰	۰/۷۴۴۰	۰/۷۰۳۳	۰/۹۵۳۰	۰/۹۱۹۰	۰/۹۱۶۵
	۰/۹۰	۰/۸۳۶۵	۰/۷۷۷۰	۰/۷۷۹۵	۰/۹۶۵۵	۰/۹۳۶۰	۰/۹۳۱۰
	۱/۰۰	۰/۸۶۱۵	۰/۸۰۵۰	۰/۸۰۶۵	۰/۹۷۳۰	۰/۹۴۵۵	۰/۹۴۰۰
	۱/۱۰	۰/۸۸۳۵	۰/۸۲۶۰	۰/۸۳۱۰	۰/۹۷۹۰	۰/۹۵۲۰	۰/۹۵۴۵
	۱/۲۰	۰/۸۹۱۵	۰/۸۴۵۰	۰/۸۴۷۰	۰/۹۸۶۵	۰/۹۵۷۵	۰/۹۵۵۰

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 π_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 RESET۲ به دست آورد. با مقایسه نتایج جدول (۳) و جدول (۴)، می‌توان نتیجه گرفت که مطالعه تحت قید RESET۱ از مطالعه تحت قید RESET۲ 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 کفیت 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 RESET۲ 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 RESET۲ 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.

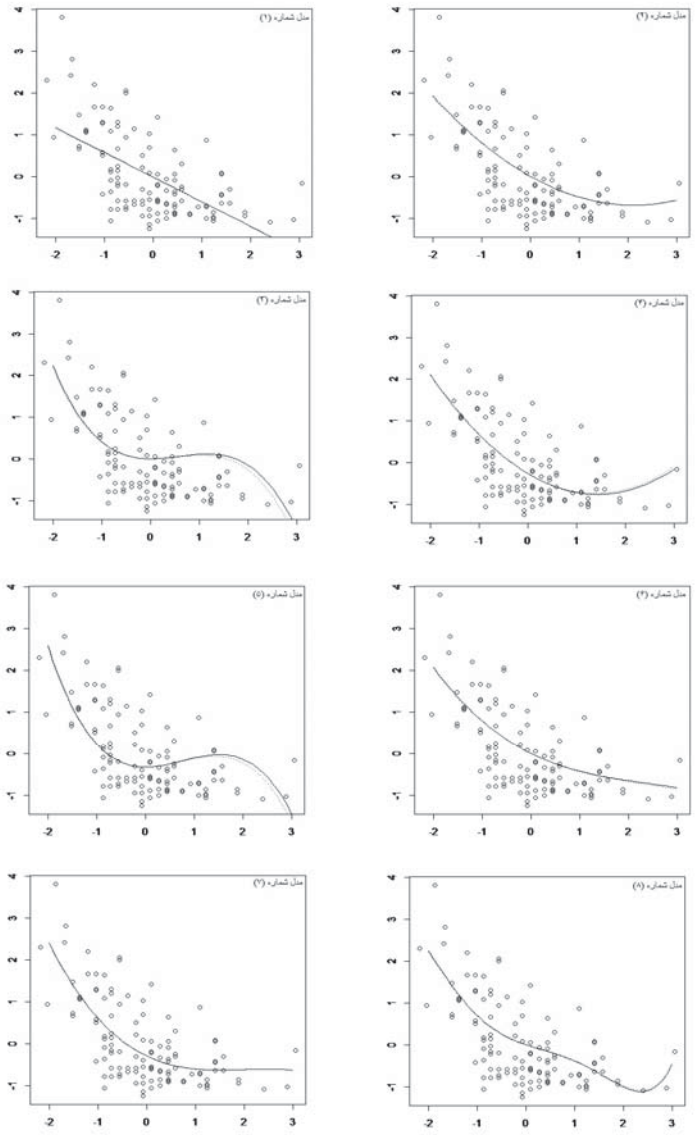
جدول (۵): عملکرد مدل‌های مختلف بر میزان گاز اوزون در مقابل سرعت باد. برآورد هر یک از پارامترها با $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ و $\hat{\beta}_4$ نشان داده‌شده است. برای قید RESET۲ مقادیر آماره آزمون با $F-value$ و مقادیر معنی‌داری آزمون‌ها با $P-value$ داده‌شده است. همچنین، برای هر مدل، 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
		-۰/۰۲۴۳	-۰/۵۸۷۶	-	-	۱۵/۹۳۳۸	۰/۰۰۰۰
۲	$y=\beta_1 x+\beta_2 x^2$	-۰/۶۵۴۴	۰/۱۵۵۵	-	-	۰/۸۰۰۵	۰/۴۵۱۶
		-۰/۶۴۹۱	۰/۱۵۳۳	-	-	۰/۷۳۲۱	۰/۴۶۰۶
۳	$y=\beta_1 x+\beta_2 x^3$	۰/۲۶۰۱	۰/۱۴۹۷	-	-	۱۱/۳۳۲۹	۰/۰۰۰۴
		-۰/۲۳۹۷	-۰/۱۵۳۴	-	-	۱۰/۸۴۵۰	۰/۰۰۰۰
۴	$y=\beta_1+\beta_2 x+\beta_3 x^2$	-۰/۲۶۸۲	-۰/۶۹۶۸	۰/۲۴۴۱	-	۲/۸۳۶۴	۰/۰۶۲۹
		-۰/۲۶۶۱	-۰/۶۸۴۳	۰/۲۴۷۱	-	۲/۶۲۴۱	۰/۰۶۶۹
۵	$y=\beta_1+\beta_2 x^2+\beta_3 x^3$	-۰/۳۲۷۱	۰/۳۸۸۵	-۰/۱۷۱۳	-	۸/۲۸۹۲	۰/۰۰۰۴
		-۰/۳۲۶۸	۰/۳۷۲۵	-۰/۱۷۴۲	-	۸/۷۱۰۴	۰/۰۰۰۰
۶	$y=\beta_1 x+\beta_2 x^2+\beta_3 x^3$	-۰/۵۸۲۷	۰/۱۷۷۷	-۰/۰۲۵۰	-	۲/۰۳۲۲	۰/۱۳۵۹
		-۰/۵۸۳۴	۰/۱۷۲۲	-۰/۰۲۴۵	-	۱/۶۶۷۵	۰/۱۸۱۷
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
		-۰/۲۹۵۳	-۰/۵۵۵۸	۰/۲۹۵۴	۰/۰۴۹۴	۱/۴۹۸۱	۰/۲۱۵۰
۸	$y=\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4 x^5$	-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. به‌طوری‌که پارامتر هموارسازی آن را با استفاده از روش اعتبارسنجی[1] 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 به‌صورت صریح قابل‌محاسبه نبودند. به همین دلیل، از روش شبیه‌سازی برای برآورد این مقادیر استفاده کرده‌ایم. از ۱۰۰۰ تکرار الگوریتم مونت‌کارلو با حجم نمونه ۵۰۰ برای برآورد مقادیر موردنظر استفاده کرده‌ایم.

1- Cross Validation



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

وقتی داده گم‌شده 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-Fisherd distribution.

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