Iranian Journal of Insurance Research JR
Scientific Research Quarterly Journal
Vol. 36/ No.3/ Autumn 2021/pp.43-84
Research Paper

Analysis of Mortality Models with Covariates Missing at Random

Shirin Shoaee[1], Reyhaneh Fathi[2]

Recived: 2021/23/04 **Accepted**: 2021/14/07

Abstract

Objective: Demographic indicators such as mortality rates play a very important role in health, financial and pension policies. Therefore, the accuracy of mathematical models in estimating mortality rates is an important challenge. One of the tasks of actuaries is to construct a suitable mortality model for the available data so that these mortality models can calculate mortality for different ages and longevity, as well as the different information available to individuals on retirement plans. Missing data is a problem that may be faced by actuaries when they are analyzing the real data. Missing data can occur for a variety of reasons, such as unanswered or censored. The presence of missing data can pose a threat to the accuracy of the data analysis results. The purpose of this study is to model the mortality in a retirement plan. In this regard, it is assumed that data are available at the individual level, including date of birth, date of joining the retirement plan, date of completion of the observation, and reason for discontinuation (usually death or right censoring). Information on covariate variables such as gender, benefits or size of pension, demographic geography or health status will also be available. More precisely, this study aims to model the mortality in a retirement plan based on missing data and access to information from various covariate variables, to carefully analyze the structure of different models, to estimate and finally to investigate the financial implications for different mortality experiences containing missing data.

Methodology: In this article, we deal with a pension plan in which each member's future life expectancy is modeled using parametric survival models incorporating covariates which may be missing for some individuals. Likelihood-based techniques estimate parameters, and in this regard, an algorithm is proposed that can perform the estimation task in the best possible way. One of the necessary features to check the adequacy of the statistical model, especially when the data contains missing values, is identifiable. If not identifiable, it can be claimed that the statistical model is not a full rank and is not a suitable model for the data. It is worth noting that the Jacobin matrix needs to be calculated to verify identifiability. As mentioned, in the analysis of

1. Assistant Professor of Department of Actuarial Science, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran. **(Corresponding Author).** sh_shoaee@sbu.ac.ir.
2. MSc in Actuarial Science, Department of Actuarial Science, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran. reyhane.fathi1@gmail.com

DOI: 10.22056/JIR.2021.278996.2872

mortality models with the presence of missing values, the maximum likelihood method can be used. In such cases, an estimation error may often occur when fitting the model, which can be reduced by modeling from a larger population. For this reason, hybrid retirement plans that remain homogeneous are often used. This proposed method can also be useful for calculating financial quantities based on pension factors. In fact, in this proposed method, different data sets with equal or similar death experiences are combined, sample size increases and risk of parameter decreases, which also leads to a reduction in capital requirement. Socio-economic variables such as the level of benefits and geographical characteristics of the population are also considered more if interest rates are low.

Finding: First, complete data are analyzed and modeled for observations of members of a retirement plan, which includes survival time and ancillary variables for each individual. Estimation of parameters is obtained using the maximum likelihood method. however, when the data is missing, it is not easy to estimate the parameters with the maximum likelihood method. In this case, the model parameters are estimated by the maximum likelihood method which are calculated using the proposed algorithm; then, statistical indicators such as identifiability of parameters are calculated to evaluate the performance of the proposed structure and algorithm. Furthermore, the financial effects, in particular the annuity factors, and the mis-estimation risk capital requirements for the mortality experience which includes the maximum covariates variables are calculated and compared with the individual segments when the data are missing. In addition, it can be seen that when the two statistical variables are not observed together, the model is not identifiable according to the data.

Conclusion: It was found that if the data are missing, the statistical model is not always identifiable using the maximum likelihood, and data combination from two or more experiments can avoid identifiable barriers. The methods proposed in this paper can be useful for actuaries when calculating financial committees based on annuity factors. These methods may combine different datasets with equal or similar mortality experiences, increase sample size, and reduce parameter risk, thus, reducing capital requirements. Socio-economic variables such as the level of benefits and geographical characteristics of the population are given more attention if the interest rate is low.

Keywords: Capital Requirement, Parameter Redundant, Full Rank, Likelihood Contribution, Identifiability, Missing At Random, Mortality Model.
JEL-Classification: C13, C24, C51

فصلنامه علمی پژوهشی پژوهشنامه بیمه
مقاله پژوهشی
سال سی‌ششم/شماره ۳/پاییز ۱۴۰۰/ شماره مسلسل۱۴۳/صفحات ۸۴-۴۳

تجزیه‌وتحلیل مدل‌های مرگ‌ومیر با متغیرهای کمکی دارای گمشدگی تصادفی

شیرین شعاعی[۱]، ریحانه فتحی[۲]

تاریخ دریافت: ۱۴۰۰/۰۲/۰۳ **تاریخ پذیرش:** ۱۴۰۰/۰۳/۲۴

چکیده

هدف: این پژوهش با هدف مدل‌بندی مرگ‌ومیر در یک طرح بازنشستگی بر اساس داده‌های گمشده و دسترسی به اطلاعات مختلف از متغیرهای کمکی، تجزیه و تحلیل دقیق ساختار مدل‌های مختلف، برآوردیابی و در نهایت بررسی تأثیرات مالی برای تجربه‌های مختلف مرگ‌ومیر و حاوی داده‌های گمشده انجام شده است.

روش‌شناسی: این مقاله با یک طرح بازنشستگی سروکار دارد که در آن طول عمر آتی هر فرد با مدل‌های بقای پارامتری با ترکیب متغیرهای کمکی که ممکن است برای برخی از افراد گمشده باشند، مدل‌سازی شده است. پارامترها با روش ماکسیمم درستنمایی برآورد شده و الگوریتمی پیشنهاد گردیده که بتواند وظیفه برآوردیابی را به بهترین شکل ممکن انجام دهد.

یافته‌ها: نتایج نشان داد در صورتی‌که داده‌ها گمشده باشند، مدل آماری همیشه با استفاده از ماکسیمم درستنمایی شناسایی‌پذیر نیست و تلفیق داده‌های حاصل از دو یا چند تجربه می‌تواند از موانع شناسایی‌پذیری جلوگیری نماید.

نتیجه‌گیری: روش‌های پیشنهادی این مقاله هنگام محاسبه کمیت‌های مالی مورد علاقه براساس عامل‌های مستمری، برای بیم‌سنجان می‌تواند مفید باشد. این روش‌ها ممکن است مجموعه داده‌های مختلف با تجربه مرگ‌ومیر برابر یا مشابه را با هم ترکیب کنند، اندازه نمونه را افزایش دهند و ریسک پارامتر را کاهش دهند، بنابراین منجر به کاهش الزام سرمایه شوند. متغیرهای اقتصادی-اجتماعی از جمله سطح مزایا و مشخصات جغرافیایی جمعیتی در صورت پایین بودن نرخ بهره بیشتر مورد توجه قرار می‌گیرند.

واژگان‌کلیدی: الزام سرمایه، پارامتر افزونه، رتبه کامل، سهم درستنمایی، شناسایی‌پذیری، گم‌شدگی تصادفی؛ مدل مرگ‌ومیر.

طبقه‌بندی موضوعی: C13، C24، C51

---

۱. استادیار گروه بیم‌سنجی، دانشکده علوم ریاضی، دانشگاه شهید بهشتی، تهران، ایران. (نویسنده مسئول).
sh_shoaee@sbu.ac.ir

۲. کارشناسی ارشد اکچوئری، گروه بیم‌سنجی، دانشکده علوم ریاضی، دانشگاه شهید بهشتی، تهران، ایران.
reyhane.fathi1@gmail.com

مقدمه

شاخص‌های جمعیتی همواره نقش مهمی در تصمیم‌گیری‌های جمعیتی داشته‌اند و نرخ مرگ‌ومیر[1] مهم‌ترین عامل تغییر الگوی جمعیت به‌شمار می‌آید. داشتن اطلاعات درست از آینده جمعیت برای برآورد دقیق جمعیت سالخورده به‌منظور برنامه‌ریزی‌های دقیق مالی جهت تأمین آینده افراد، ساخت جداول آتی عمر مطابق با محتمل‌ترین الگوی مرگ‌ومیر و قیمت‌گذاری صحیح سالانه‌های عمر، تنها بخشی از کاربردهای پیش‌بینی مقادیر آتی نرخ مرگ‌ومیر در جمعیت‌شناسی، بیمه و آمار است.

مدل‌بندی نرخ مرگ‌ومیر پیشینه‌ای طولانی دارد و جداول مرگ‌ومیر برای بیم‌سنجان، یکی از متداول‌ترین و قدیمی‌ترین روش‌های بررسی میزان مرگ‌ومیر و ابزار اصلی توصیف الگوی مرگ‌ومیر تلقی می‌شوند. رویکرد یک جدول مرگ‌ومیر بر این اساس است که آزمایش گذشته بدون هیچ دگرگونی در آینده تکرار شود. لذا تلاش‌های بسیاری برای دستیابی به مدلی که بتواند این جداول را تبیین کند، صورت گرفته است. یافتن مدل ریاضی برای احتمال مرگ‌ومیر از نظر کاربردی مفیدتر است. روشن است که استفاده از تابعی که پارامتر کمی دارد و از نظر هندسی دارای منحنی همواری است به‌مراتب آسان‌تر از کار با جدول عمری است که پارامترهای فراوانی دارد و در موارد بسیاری نیز احتیاج به درون‌یابی، برون‌یابی، هموارسازی و تعدیل‌های متعدد آماری دیگر، در جهت یافتن احتمال‌های فوت دارد که در محاسبه حق‌بیمه فنی استفاده می‌شوند. لذا بیم‌سنج‌ها علاقه‌مند هستند که مدل مرگ‌ومیری را ارائه دهند که بتواند برای سنین و طول عمرهای متفاوت و همچنین اطلاعات متفاوتی که برای افراد در طرح‌های بازنشستگی موجود است، شدت مرگ‌ومیر را محاسبه نماید.

بنابراین، مطالعه حاضر با هدف مدل‌بندی مرگ‌ومیر در یک طرح بازنشستگی انجام شده است. می‌توان فرض نمود که داده‌ها در سطح فردی شامل تاریخ تولد، تاریخ پیوستن به طرح، تاریخ اتمام مشاهده و دلیل قطع مشاهده (معمولاً مرگ یا سانسور از راست) در

1. Mortality Rate

دسترس هستند. همچنین سایر اطلاعات به‌صورت متغیرهای کمکی در دسترس خواهند بود. ازجمله متغیرهای کمکی، جنسیت است که معمولاً شناخته شده است. متغیرهای کمکی دیگری که شامل اطلاعات غنی‌تری هستند، از قبیل مزایا یا اندازه حقوق بازنشستگی، مشخصات جغرافیایی جمعیتی یا وضعیت سلامتی نیز در دسترس هستند.

در طرح‌های بازنشستگی اطلاعات متفاوتی برای هر فرد موجود است و وابستگی مرگ‌ومیر به این اطلاعات به‌خصوص زمانی‌که ریسک روی زیرمجموعه‌هایی از اعضا متمرکز است، بسیار اهمیت دارد. در بررسی داده‌ها در طرح‌های بازنشستگی با مواردی مواجه می‌شویم که برخی از اطلاعات در مورد هر فرد موجود نیست. در تحلیل مدل‌های مرگ‌ومیر با حضور مقادیر گمشده[1] می‌توان از روش ماکسیمم درستنمایی[2] استفاده نمود. در این‌گونه حالت‌ها اغلب هنگام برازش مدل ممکن است خطای برآورد حاصل شود، که می‌توان با مدل‌سازی از یک جمعیت بزرگ‌تر آن را کاهش داد. به‌همین علت معمولاً از طرح‌های بازنشستگی ترکیبی که همگن باقی بمانند، استفاده می‌شود. این امر باعث افزایش اندازه نمونه، کاهش ریسک پارامتر و در نتیجه، کاهش الزام سرمایه[3] می‌گردد.

در این پژوهش بعد از بیان مقدمه، در ادامه، مبانی نظری و پیشینه پژوهش به اختصار مرور شده است. بخش بعدی شامل نمادگذاری مورد نیاز برای توزیع احتمال متغیرهای کمکی، توصیف تابع درستنمایی[4] و توسیع آن برای زمانی‌که داده‌ها گمشده هستند، است. همچنین در ابتدای این بخش، به تحلیل و مدل‌بندی داده‌های کامل برای مشاهدات اعضای یک طرح بازنشستگی که شامل زمان بقا و متغیرهای کمکی برای هر فرد است، پرداخته شده است. همچنین در این بخش براساس مشاهدات کامل از داده‌های صندوق بازنشستگی، ابتدا مجموعه داده‌هایی حاوی داده‌های گمشده با حذف متغیرهای کمکی برای برخی واحدها شبیه‌سازی شده‌اند. سپس مسئله برآورد، شرایط ریاضی برای شناسایی‌پذیری[5] برآورد پارامترها و الگوریتم برازش توصیف شده است. به‌علاوه مشاهده

1. Missing Values
2. Maximum Likelihood
3. Capital Requirement
4. Likelihood Function
5. Identifiability

می‌شود در حالتی‌که دو متغیر آماری تواماً مشاهده نمی‌شوند، مدل با توجه به داده‌ها شناسایی‌پذیر نیست. برای غلبه بر عدم شناسایی‌پذیری، نمونه بسیار کوچکی در نظر گرفته شده که متغیرهای کمکی آن کاملاً مشاهده شده است. در پایان درباره تأثیر مالی ترکیب مجموعه داده‌های مختلف برای برآورد شدت مرگ‌ومیر مطالعه شده و درنهایت نتیجه‌گیری ارائه گردیده است.

۱. مروری بر پیشینه پژوهش

تدوین مدل‌های مرگ‌ومیر از ساختن جداول مرگ‌ومیر آغاز و اولین دستاورد در این زمینه در سال ۱۶۹۳ به‌وسیله ستاره‌شناس مشهور، هالی[1] انجام گرفته است که جدول عمری براساس تعداد فوتی‌های مشاهده شده به‌دست آورد. در سال ۱۷۴۰، اولین جدول عمر به تفکیک مردان و زنان توسط استرویک[2] منتشر شد. اولین مدل تحلیلی مرگ‌ومیر که احتمال بقا را به‌صورت یک تابع افزایشی از سن کنونی فرد در نظر گرفت، توسط دمورآور[3] معرفی شد. سپس گومپرتز[4] در سال ۱۸۲۵ مدل کامل‌تری از مدل دمورآور ارائه نمود. در سال ۱۸۶۰ مکهام[5]، مدل گومپرتز را که برای سنین بالاتر مناسب نبود اصلاح و مدل کامل‌تری ارائه کرد. بعد از او پرکس[6] مدل ریاضی ارائه شده گومپرتز و مکهام را کامل نمود. مدل‌های گومپرتز و مکهام از مدل‌های رایج مرگ‌ومیر باقی ماندند تا این‌که در اوایل قرن بیستم، اقتصاددان و جامعه‌شناس ایتالیایی پارتو[7] نظریات خود که بر مبنای مشکلات جامعه بود را در مدل‌بندی مرگ‌ومیر دخیل نمود. در اواخر قرن بیستم مدل‌های مرگ‌ومیر پیچیده‌ای بررسی شدند که اکثر آن‌ها تعدیل‌شده یا تعمیم‌یافته مدل گومپرتز

1. Halley
2. Struyck
3. De Moivre
4. Gompertz
5. Makeham
6. Perkes
7. Pareto

بودند. از جمله می‌توان به مدل ذکایی و مقصودی (۱۳۸۹) که به بازسازی مدل مرگ‌ومیر بر پایه شکنندگی با استفاده از تعمیم مدل گومپرتز می‌پرداخت، اشاره نمود.

برخلاف مدل‌بندی نرخ مرگ‌ومیر، پیش‌بینی نرخ مرگ‌ومیر پیشینه کوتاه‌تری دارد. تا سه دهه قبل، روش‌های مورد استفاده برای پیش‌بینی مرگ‌ومیر نسبتاً ساده و براساس قضاوت‌های ذهنی انجام می‌شدند. اما در سال‌های گذشته، روش‌های پیشرفته‌تری ارائه شده که کارشناسان بیمه، آمار و جمعیت‌شناسان را به استفاده بیشتر از آنها سوق داده است. از میان این روش‌ها، روش لی-کارتر[1] به‌دلیل ساختار ساده و کاربرد موفق آن در بسیاری از کشورها، مورد استفاده محققین قرار گرفته است. پیدایش روش لی-کارتر به تغییر الگوی امید به زندگی از سال ۱۹۰۰ در آمریکا برمی‌گردد. گرچه این روش بر اساس داده‌های مرگ‌ومیر آمریکا در سال‌های ۱۹۰۰ تا ۱۹۸۷، طراحی شده بود، اما عملکرد بسیار خوبی در مدل‌بندی و پیش‌بینی بلندمدت نرخ مرگ‌ومیر در بسیاری از کشورهای توسعه‌یافته داشته است. به‌طوری‌که کمیجانی و همکاران (۱۳۹۲) به برآورد و پیش‌بینی نرخ مرگ‌ومیر در ایران با استفاده از مدل لی-کارتر پرداخته و عملکرد آن را شرح داده‌اند.

جدیدترین ایده در مطالعه بقا، مرگ‌ومیر تصادفی است که مرگ‌ومیر همانند یک فرایند تصادفی درنظر گرفته می‌شود (یاشین[2]، ۲۰۰۱). اما رویکرد زنجیر مارکوف در مدل‌بندی مرگ‌ومیر به‌طوری که توزیع زمان مرگ مشخص و دارای توزیع فاز- نوع[3] باشد به‌وسیله به‌وسیله لین و لیو[4] (۲۰۰۷) مطرح شده است. همچنین شجاعی آذر و حسن‌زاده (۱۳۹۳) (۱۳۹۳) براساس این ایده به مدل‌بندی مرگ‌ومیر پرداخته‌اند.

در ادامه پژوهش‌های صورت گرفته در این حوزه، مهدوی و همکاران (۱۳۹۰) به کاربرد یک مدل مرگ‌ومیر با چند عامل ریسک پرداختند. آن‌ها به این موضوع اشاره نمودند که مدل‌های مرگ‌ومیر معرفی شده عمدتاً معطوف به سن و جنس افراد است، اما عوامل

1. Lee - Carter
2. Yashin
3. Phase-Type
4. Lin & Liu

دیگری شامل وضعیت اجتماعی – اقتصادی افراد مانند سطح تحصیلات، میزان درآمد، محل زندگی، شغل، شرایط تأهل و عوامل رفتاری می‌تواند در میزان مرگ‌ومیر تأثیر داشته باشند.

۲. مبانی نظری

در بررسی بسیاری از مجموعه داده‌ها، با داده‌های گمشده روبه‌رو می‌شویم. گمشدگی داده‌ها می‌تواند به‌علل مختلفی مانند پاسخ داده‌نشده یا سانسورشده به‌وجود آید. در مجموعه داده‌ها در موارد پاسخ داده‌نشده، برخی از اطلاعات هر فرد موجود نیست. موارد سانسورشده از جایی به بعد تمامی اطلاعات دربارهٔ فرد موجود نیست که می‌تواند به‌دلیل فوت یا هر علت دیگری فرد از مشاهده خارج شود. وجود دادهٔ گمشده می‌تواند منجر به تهدیدی برای صحت نتایج حاصل از تحلیل داده‌ها شود. در نظر نگرفتن دادهٔ گمشده یا استفاده از روش‌های نامناسب برخورد با این داده‌ها، به‌عنوان مثال جایگذاری داده‌های گمشده با میانگین می‌تواند به مشکلاتی همچون آماسیدگی[1] در نقطه‌ای مانند میانگین منجر شود. بسیاری از محققین روش‌هایی را برای مقابله با مشکلات داده گمشده ارائه کرده‌اند. اکثر این روش‌ها وابسته به نوع سازوکار گمشدگی داده‌ها هستند و برحسب آن می‌توان روش تحلیل مناسبی را انتخاب نمود. در این راستا، روبین[2] (۱۹۷۶) سیستم طبقه‌بندی را برای مشاهدات داده گمشده معرفی نمود که به‌طور گسترده مورد استفاده قرار می‌گیرد.

به‌طور معمول سه مکانیسم گمشدگی شامل گمشدگی تصادفی[3]، گمشدگی کاملاً تصادفی[4] و گمشدگی غیرتصادفی[5] وجود دارد. در این راستا، بردار پاسخ برای فرد iام به‌صورت $y_i = (y_{i1}, \dots, y_{ij})$ است. برای هر متغیر پاسخ برداری از متغیرهای نشانگر

1. Inflation
2. Rubin
3. Missing at Random (MAR)
4. Missing Completely at Random (MCAR)
5. Missing not at Random (MNAR)

به‌صورت $m_i = (m_{i1}, \dots, m_{ij})$، که در آن $m_{ij} = ۰$ اگر y_{ij} مشاهده شود و $m_{ij} = ۱$ وقتی y_{ij} مشاهده نشود وجود دارد. بردار مشاهدات برای هر فرد به‌صورت y_i^{obs} (مقادیر مشاهده‌شده) و y_i^{mis} (مقادیر گمشده) تفکیک می‌شود.

در گمشدگی تصادفی داده‌ها به‌طور تصادفی گمشده هستند. بدین مفهوم که احتمال گمشدن داده‌ها در یک متغیر Y به برخی از متغیرهای اندازه‌گیری‌شده دیگر وابسته، اما به مقدار گمشده Y وابسته نیست. به‌عبارت دیگر بین داده‌های گمشده در Y و مقادیر Y پس از جدا کردن سایر متغیرها رابطه‌ای وجود ندارد.

امروزه بسیاری از محققین روش‌هایی را برای مقابله با مشکل داده گمشده ارائه کرده‌اند که اغلب آنها وابسته به نوع سازوکار گمشدگی داده‌ها هستند. از جملهٔ می‌توان به روش‌های ویلکس[1] (۱۹۳۲) و لرد[2] (۱۹۵۵) اشاره کرد که بر توزیع نرمال تمرکز کرده‌اند، درحالی‌که در زمینهٔ داده‌های بقا، اسلاچر و جکسون[3] (۱۹۸۹) یک مدل لوگ خطی برای تجزیه‌وتحلیل داده‌های بقا سانسورشده همراه با عوامل قطعی در نظر گرفتند. هیرینگ و ابراهیم[4] (۲۰۰۱) بر مدل خطرات متناسب کاکس[5] تمرکز کرده‌اند. در این دو مطالعه الگوی کلی داده‌های گمشده مطرح و پارامترها با استفاده از الگوریتم EM[6] برآورد شده‌اند. لیتل و آن[7] (۲۰۰۴) یک روش مبتنی‌بر احتمال قوی MAR را برای داده‌های چند چند متغیره با مقادیر گمشده بر اساس رگرسیون روی اسپلاین‌های نمره‌های تمایل ارائه دادند. تسیاتیس[8] (۲۰۰۷) برای داده‌های گمشده و سانسورشده یک روش با مدل‌های نیمه پارامتری معرفی نمود. چن و همکاران[9] (۲۰۱۴) یک مدل‌بندی برای داده‌های طولی و بقا بر روی داده‌های گمشده و سانسورشده از چپ معرفی نمودند که بر پایهٔ مدل

1. Wilks
2. Lord
3. Schluchter & Jackson
4. Herring & Ibrahim
5. Cox Proportional Hazards Model
6. Expectation Maximization
7. Little & An
8. Tsiatis
9. Chen et al

کاکس، مدل آمیختهٔ خطی و مدل‌های بیزی بود. خو و همکاران[1] (۲۰۱۷) از معادلات برآوردساز مورد انتظار برای تولید یک نظریه کلی استفاده نمودند. شایان ذکر است که رویکرد آن‌ها در شرایطی که متغیرهای کمکی و برآمدها پیوسته یا گسسته هستند قابل استفاده است، اما برای برآورد واریانس از روش بوت استرپ[2] بایستی استفاده نمود. این روش نیمه‌پارامتری، یک طرح با داده‌های مقطعی و متغیرهای مستقل از زمان را در نظر می‌گیرد. در نهایت اونگولو و همکاران[3] (۲۰۱۹) به برآورد پارامترهای مدل برای داده‌های داده‌های بقا بر اساس داده‌های گمشده به‌وسیله اعمال یک توزیع آمیخته برای متغیرهای کمکی و روش بهینه‌سازی مقید پرداختند.

یکی از ویژگی‌های لازم برای کفایت مدل آماری، به‌خصوص زمانی‌که داده‌ها حاوی مقادیر گمشده‌اند، شناسایی‌پذیری است. در صورت عدم شناسایی‌پذیری، می‌توان ادعا نمود که مدل آماری رتبه کامل نبوده و مدل مناسبی برای داده‌ها نیست. یک مدل رتبه کامل مدلی است که پارامتر افزونه ندارد. به‌طور خاص اگر در یک مدل آماری، داده‌ها گمشده وجود داشته باشد، ممکن است پارامتر افزونه حاصل شود و روی شناسایی‌پذیری اثر بگذارد. کچپول و مورگان[4] (۱۹۹۷) در قضیه ۴ مقاله خود نشان دادند دادند اگر یک مدل پارامتر افزونه باشد، به‌صورت محلی شناسایی‌پذیر نیست. مدل‌های آماری که پارامتر افزونه نیستند، به‌صورت رتبه کامل تعریف می‌شوند. اگر یک مدل رتبه کامل باشد، مقادیر ویژهٔ ماتریس اطلاع تجربی ارزیابی‌شده توسط براورد ماکسیمم درستنمایی[5] (MLE) همگی منفی و از صفر دور خواهند بود. برای تشخیص رتبه کامل بودن به ماتریس ژاکوبین نیاز است. اگر رتبهٔ نمادین این ماتریس با تعداد پارامترهای مدل (تعداد ردیف‌های ماتریس ژاکوبین[6]) برابر بود، مدل رتبه کامل[7] است. اما اگر رتبهٔ

1. Xu et al
2. Bootstrap
3. Ungolo et al
4. Catchpole & Morgan
5. Maximum likelihood Estimation
6. Jacobian Matrix
7. Full Rank

رتبهٔ نمادین ماتریس ژاکوبین کمتر از تعداد پارامترهای مدل حاصل شود، مدل پارامتر افزونه[1] و درنتیجه شناسایی‌پذیر نیست. پارامتر افزونه به‌حالتی گفته می‌شود که بتوان مدل آماری مدنظر را با بردار پارامتری کوچک‌تری نیز تعریف نمود. بنابراین مدل دارای تعداد زیادی پارامتر (بیش از تعداد مورد نیاز) است و درنتیجه شناسایی‌پذیر نیست.

در طرح‌های بازنشستگی اطلاعات متفاوتی برای هر فرد موجود است و وابستگی مرگ‌ومیر به این اطلاعات به‌خصوص زمانی‌که ریسک روی زیرمجموعه‌هایی از اعضا متمرکز است، بسیار اهمیت دارد. در مدل‌بندی داده‌ها در طرح‌های بازنشستگی با مواردی مواجه می‌شویم که برخی از اطلاعات در مورد هر فرد موجود نیست. یکی از رویکرده‌های موجود در برآورد پارامترهای مدل‌های مرگ‌ومیر با حضور مقادیر گمشده استفاده از روش ماکسیمم درستنمایی است. اما در اغلب موارد به‌دلیل عدم دسترسی به برخی از داده‌ها در براورد پارامترهای مدل خطای برآورد حاصل شود، که می‌توان آن‌ها را با مدل‌سازی از یک جمعیت بزرگ‌تر کاهش داد. به‌همین علت معمولاً از طرح‌های بازنشستگی ترکیبی که همگن باقی بمانند، استفاده می‌شود. این امر باعث افزایش اندازه نمونه، کاهش ریسک پارامتر و در نتیجه، کاهش الزام سرمایه می‌گردد.

۳. روش‌شناسی پژوهش

۳-۱. مشاهدات و تابع درستنمایی

در این بخش، ساختار تابع درستنمایی توأم برای برآورد بردار پارامترهای مدل در حالت مشاهدات کامل ارائه و سپس برای مواردی که دادهٔ گمشده وجود دارد، توسیع شده است.

1. Parameter Redundant

۳-۱-۱. مشاهدات کامل

در تجزیه و تحلیل رگرسیونی، متغیرهای کمکی اغلب قطعی درنظر گرفته می‌شوند، در این بخش آن‌ها را به‌عنوان متغیر تصادفی مدل‌بندی می‌کنیم. مجموعهٔ متغیرهای کمکی تصادفی یک فرد با بردار تصادفی p بعدی Z نشان می‌دهیم. مقدار تحقق‌یافته z، شدت مرگ‌ومیر فرد $\mu_x(z;\tau)$ را تعیین می‌کند. تابع توزیع و تابع چگالی یا احتمال Z، به‌ترتیب با $F_z(z;\zeta)$ و $f_z(z;\zeta)$ نشان داده می‌شود که ζ یک بردار پارامتری ناشناخته است. فرض‌کنید n تعداد افراد مشاهده‌شده در یک دورهٔ زمانی محدود است. نمادهای زیر را تعریف می‌کنیم:

- ✓ x_i سن فرد iام در آغاز دورهٔ مشاهده است،
- ✓ t_i مقدار تحقق‌یافته متغیر تصادفی T_i، بیانگر زمان زندگی‌شده فرد iام در طول دوره است،
- ✓ d_i شمارنده‌ای است که مقدار تحقق‌یافته متغیر تصادفی D_i را نشان می‌دهد؛ اگر فرد iام فوت کرده باشد مقدار آن یک و اگر بر اثر سانسور از مشاهده خارج شده باشد مقدار آن صفر است،
- ✓ z_i مقدار تحقق‌یافته متغیر تصادفی Z_i، که متغیرهای کمکی فرد iام را مشخص می‌کند.

در طول یک دوره مشاهدات متناهی برای آن دسته از افرادی که فوت می‌کنند، T_i بیانگر طول‌عمر باقی‌مانده برای سن x_i است. برای افراد زنده (سانسورشده از راست)، تمام آن چیزی که می‌دانیم این است که طول‌عمر باقی‌مانده آن‌ها بیشتر از T_i خواهد بود.

فرض می‌کنیم متغیرهای کمکی $Z_n, \dots, Z_1$ مستقل و هم‌توزیع[1] $(i.i.d.)$ و روی مقادیر تحقق‌یافتهٔ خود شرطی هستند. به‌علاوه سنین مشاهده‌شدهٔ $x_n, \dots, x_1$ و طول‌عمرهای مشاهده‌شدهٔ $T_n, \dots, T_1$ به‌صورت $i.i.d.$ هستند. درنهایت فرض می‌شود که سانسورها

1. Independent and Identically Distributed (i.i.d)

به‌صورت ناآگاهی‌بخش هستند. یعنی می‌توان طول‌عمرها را به‌عنوان متغیر تصادفی در نظر و توزیع آن‌را مستقل از زمان سانسور، برآورد نمود، (مک‌دونالد[1] و همکاران، ۲۰۱۸). در ادامه فرض می‌کنیم مشاهدات به‌صورت کامل در دسترس هستند، به‌طوری‌که هر بردار متغیر کمکی z_i به‌طور کامل مشاهده شده است. اگر سهم فرد iام در تابع درستنمایی را با L_i نمایش دهیم، بنابراین تابع درستنمایی کل به‌صورت $L = \prod L_i$ خواهد بود.

برای افرادی که فوت می‌شوند ($d_i = ۱$) سهم تابع درستنمایی به‌وسیلهٔ ضرب چگالی شرطی T_i روی x_i و z_i مربوطه در تابع چگالی یا احتمال z_i به‌دست می‌آید که به‌صورت زیر محاسبه می‌گردد:

$$L_i(\tau, \zeta) = f_{x_i}(t_i|z_i;\tau) f_Z(z_i;\zeta), \qquad (۱)$$

که در آن (τ, ζ) بردار پارامترهای نامعلوم هستند و در رابطهٔ فوق چگالی توام (T, Z) به‌عنوان مدل کامل[2] (برای سن x) تعریف می‌شود، و همچنین به چگالی T شرطی شده روی Z مدل جزئی[3] گفته می‌شود. عبارت $f_{x_i}(t_i|z_i;\tau)$ به این معنی است که فرد x ساله بعد از سن $x + t$، دیگر زنده نیست.

برای افرادی که طول‌عمرهای مشاهده‌شدهٔ آن‌ها سانسورشده از راست است ($d_i = ۰$)، زمان دقیق فوت مشاهده نشده است. لذا سهم تابع درستنمایی برای این افراد به‌صورت زیر محاسبه می‌شود:

$$L_i(\tau, \zeta) = S_{x_i}(t_i, z_i;\tau) f_Z(z_i;\zeta), \qquad (۲)$$

که $S_{x_i}(t_i, z_i;\tau)$ تابع بقا و بیانگر این است که فرد x ساله حداقل t سال زنده مانده است و داریم:

$$S_{x_i}(t_i, z_i;\tau) = exp\left\{-\int_0^{t_i} \mu_{x_i+s}(z_i;\tau) ds\right\}, \qquad (۳)$$

1. Macdonald
2. Full Model
3. Partial Model

که μ_{x_i+s} بیانگر شدت مرگ‌ومیر فرد iام دارای سن $x+s$ است. با تلفیق این دو رابطه و ارتباط شدت مرگ‌ومیر، تابع احتمال و بقا، سهم مشارکت در تابع درستنمایی فرد iام به‌صورت زیر است:

$$L_i(\tau,\zeta) = exp\left\{-\int_0^{t_i}\mu_{x_i+s}(z_i;\tau)ds\right\}\mu_{x_i+t_i}(z_i;\tau)^{d_i}f_Z(z_i;\zeta). \quad (۴)$$

اگر عناصر Z برای تمام افراد مشاهده شود، تابع درستنمایی به یک تابع از τ و یک تابع از ζ تبدیل می‌شود:

$$L(\tau,\zeta) = \underbrace{\prod_i exp\left\{-\int_0^{t_i}\mu_{x_i+s}(z_i;\tau)ds\right\}\mu_{x_i+t_i}(z_i;\tau)^{d_i}}_{L^{\tau}(\tau)}\underbrace{\prod_i f_Z(z_i;\zeta).}_{L^{\zeta}(\zeta)} \quad (۵)$$

در این حالت τ می‌تواند به‌صورت مستقل از ζ برآورد و توزیع متغیرهای کمکی در تحلیل رگرسیونی نادیده گرفته می‌شود. در بخش بعد مشاهده می‌شود که اگر بعضی از مشاهدات Z_i گمشده باشند، این کار امکان‌پذیر نیست و پارامترهای τ و ζ باید توأماً برآورد شوند.

۳-۱-۲. مشاهدات گمشده

در بررسی بسیاری از مجموعه داده‌ها، با داده‌های گمشده روبه‌رو می‌شویم. گمشدگی داده‌ها می‌تواند به‌علل مختلفی مانند پاسخ داده‌نشده یا سانسورشده به‌وجود آید. حال فرض کنید بعضی از مولفه‌های متغیرهای کمکی Z برای برخی از افراد تحت مطالعه گمشده هستند. لذا بردار p بعدی $Z_i = (Z_{i,p},\dots,Z_{i,1})$ به دو بردار Z_i^{obs} و Z_i^{mis} تقسیم می‌شود. برای هر فرد تابع چگالی بردار متغیرهای کمکی $Z_i = \left(Z_i^{obs}, Z_i^{mis}\right)$ به‌صورت زیر تجزیه می‌شود:

$$f_{Z_i}(z_i) = f_{Z_i^{obs}}\left(z_i^{obs};\zeta\right)f_{Z_i^{mis}|Z_i^{obs}}\left(z_i^{mis}|z_i^{obs};\zeta\right). \quad (۶)$$

متغیرهای کمکی دارای گمشدگی تصادفی هستند، لذا عبارت $f_{Z_i^{mis}|Z_i^{obs}}\left(z_i^{mis}|z_i^{obs};\zeta\right)$ را می‌توان نادیده گرفت. مشابه رابطهٔ (۴)، میزان سهم

درستنمایی برای هر فرد، بر پایهٔ طول‌عمر مشاهده شدهٔ T، سن x و مجموعهٔ متغیرهای کمکی مشاهده‌شده Z^{obs} به‌صورت زیر است:

$$L_i(\tau,\zeta) = \int_{S_{z_i^{mis}}} exp\left\{-\int_0^{t_i} \mu_{x_i+s}(z_i^{obs},y;\tau)ds\right\} \mu_{x_i+t_i}^{d_i}(z_i^{obs},y;\tau) f_Z\left(z_i^{obs},y;\zeta\right)dy, \quad (۷)$$

که روی فضای وضعیت[1] $S_{z_i^{mis}}$ از Z_i^{mis} انتگرال گرفته می‌شود. اگر مولفه‌های Z_i^{mis} دارای فضای وضعیت گسسته باشند، انتگرال روی $S_{z_i^{mis}}$ تبدیل به مجموع خواهد شد.

در حالت داده‌های گمشده، تابع درستنمایی کل را نمی‌توان مشابه رابطه (۵) به دو تابع جداگانه از τ و ζ عامل‌بندی نمود. بنابراین نمی‌توان τ را به‌طور مستقل از ζ برآورد کرد و بایستی آن‌ها را به‌صورت توأم برپایهٔ طول عمر مشاهده‌شده و متغیرهای کمکی برآورد نماییم.

همچنین زمانی‌که داده‌ها گمشده و پارامترهای τ و ζ هر دو نامعلوم هستند، شناسایی‌پذیری مدل کامل را نمی‌توان تصدیق نمود.

٤. تجزیه و تحلیل داده‌ها

٤-١. تحلیل و مدل‌بندی بر اساس داده‌های کامل

هدف این بخش مدل‌بندی داده‌ها برای مشاهدات اعضای یک صندوق بازنشستگی است. برای هر عضو صندوق، دو متغیر کمکی وجود دارد. این مجموعه دادهٔ کامل نامیده می‌شوند.

٤-١-١. مجموعه داده‌های کامل

مجموعه داده‌های کامل دارای مشخصه‌های زیر است:

- ✓ ۱۸۷۴۱ مستمری در حال پرداخت است،

1. State Space

- ✓ مجموع سال‌های در معرض ریسک افراد برابر ۴/۱۷۲۶۰۱ سال است،
- ✓ تعداد افراد فوت شده ۴۹۵۶ نفر است،
- ✓ دورهٔ مشاهده از ۱۰ نوامبر ۱۹۹۲ تا ۳۱ دسامبر ۲۰۰۹ است،
- ✓ مقدار مزایای سالانه برای هر فرد مشاهده‌شده را با B نشان می‌دهیم که یک متغیر کمکی با دو سطح مزایای بالا و مزایای پایین است،
- ✓ مشخصات جغرافیایی جمعیتی برای هر فرد برپایهٔ سیستم طبقه‌بندی موزاییک[1]، را با C نمایش می‌دهیم و شامل سه سطح ۰ و ۱ و ۲ است. سطح ۰ نمایندهٔ محروم‌ترین مناطق و سطح ۲ مناطقی با کم‌ترین محرومیت است.

دو ایتم آخر، بردار متغیرهای کمکی z_i را تشکیل می‌دهند. در این مجموعه داده مزایا بر حسب پوند است و می‌توان با آن به‌عنوان متغیر پیوسته برخورد نمود، اما به چند دلیل آن‌را به‌صورت متغیر گسسته لحاظ می‌نماییم. اول: مکدونالد و همکاران (۲۰۱۸) نشان دادند که استفادهٔ مستقیم آن در مدل مناسب نیست. دوم: اعمال متغیر پیوسته به‌عنوان یک متغیر طبقه‌بندی شده، از لحاظ امساک در رویکرد مدل‌بندی راحت‌تر است. سوم: مادریگال[2] و همکاران (۲۰۱۱) مشاهده کردند که مقدار پایین حقوق بازنشستگی می‌تواند یک شاخص گمراه‌کننده برای ثروتمندی فرد بوده که می‌تواند به این دلیل رخ دهد که یک کارمند (یا مستخدم، عضو) برای خدمات طولانی مدت حقوق پایینی داشته یا برای خدمات کوتاه مدت حقوق بالایی داشته است.

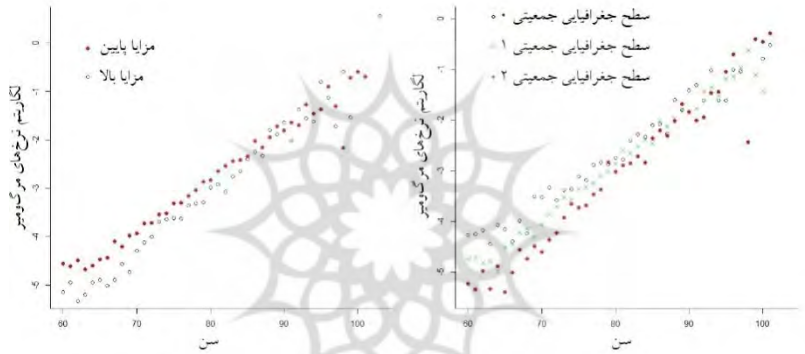
٤-١-٢. تجزیه و تحلیل مقدماتی داده‌ها

در این بخش ویژگی‌های مجموعه داده کامل را توصیف می‌کنیم. ابتدا نرخ خام مرگ‌ومیر D_x/E_x در هر سال واحد از سن محاسبه و در شکل ۱ نمایش داده‌شده است. نرخ‌های مرگ‌ومیر به‌طور جداگانه برای هر دو سطح مزایا (نمودار سمت چپ) و سه گروه مشخصات جغرافیایی جمعیتی (نمودار سمت راست) رسم شده است. در این نمودار D_x

1. Mosaic
2. Madrigal

بیانگر تعداد افرادی است که بین سن‌های دقیق x و $x+1$ فوت شده‌اند و E_x (در معرض ریسک[1]) بیانگر مجموع تمام زمان‌هایی است که تمام افراد بین این دو سن زنده هستند. براساس شکل (۱) سه نتیجه مهم حاصل می‌شود:

- ✓ تقریباً یک رابطهٔ لگ-خطی بین سن و نرخ‌های مرگ‌ومیر وجود دارد.
- ✓ بین مزایای بالاتر و مرگ‌ومیر پایین‌تر ارتباط وجود دارد.
- ✓ بین مشخصات جغرافیایی جمعیتی دارای کم‌ترین محرومیت، با مرگ‌ومیر پایین‌تر ارتباط وجود دارد.



شکل ۱. نمودار خام نرخ مرگ‌ومیر D_x/E_x بر اساس سن

منبع: یافته‌های تحقیق

٤-۱-۳. انتخاب مدل پارامتری

برای مدل‌بندی و به‌منظور ساده نگه داشتن مدل، از اختلاف کوچک بین نرخ مرگ‌ومیر برای گروه‌های مختلف در سنین بالا چشم‌پوشی و از مدل گومپرتز (گومپرتز، (۱۸۲۵)) با جداسازی گروه‌های خاص استفاده می‌کنیم که در آن نیروی مرگ‌ومیر برای فرد iام از رابطهٔ زیر حاصل می‌شود:

$$\mu(b_i, c_i; \tau) = exp\{\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i=High]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\}, \quad (۸)$$

1. Exposed to Risk

که $\tau = (\alpha, \beta, \gamma, \delta_1, \delta_2)$ بردار پارامترهای نامعلوم، پارامتر β تأثیر سن، γ تأثیر مزایای بالا و δ_1 و δ_2 تأثیر مشخصات جغرافیایی جمعیتی مختلف است. تابع شدت مرگ‌ومیر ارائه شده در رابطه ۸ متناظر با مدل بقای $S_{x_i}(t_i|b_i, c_i; \tau)$ است که از رابطه زیر محاسبه می‌گردد:

$$\begin{aligned} S_{x_i}(t_i|b_i, c_i; \tau) &= exp\left\{-\int_0^t \mu_{x+s}\, ds\right\} \\ &= exp\left\{-\left(\frac{\exp(\beta t_i) - 1}{\beta}\right) exp\big(\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i = High]} \right. \\ &\left. + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\big)\right\}. \end{aligned} \quad (۹)$$

تابع چگالی برای طول‌عمر باقی‌مانده به‌صورت زیر حاصل می‌شود:

$$f_{x_i}(t_i|b_i, c_i; \tau) = S_{x_i}(t_i|b_i, c_i; \tau)\mu_{x_i+t_i}(b_i, c_i; \tau). \quad (۱۰)$$

طبق رابطهٔ (۸) پارامترهای γ، δ_1 و δ_2 بیانگر تفاوت نرخ مرگ‌ومیر بین گروه‌های متفاوت است. مرگ‌ومیر پایهٔ $\exp(\alpha + \beta x)$ در مدل، بیانگر نرخ مرگ‌ومیر برای فرد x ساله که دارای مزایا پایین و در مناطق جغرافیایی جمعیتی با محرومیت بیشتر زندگی می‌کند، است.

٤-١-٤. نتایج تجربی برای مجموعه داده‌های کامل

در ادامه به بررسی بهترین برازش مدل در رابطه ۸ بر اساس مدل‌های آشیانه‌ای با ترکیب‌های مناسب از پارامترهای γ، δ_1 و δ_2، می‌پردازیم. در این راستا ۴ مدل مختلف حاصل می‌شود. ساختار این مدل‌ها بر اساس نیروی مرگ‌ومیر در جدول زیر ارائه شده است:

جدول۱. ساختار ٤ مدل بر اساس نیروی مرگ‌ومیر

مدل	ساختار مدل
$\mathbf{M_0}$	$\exp\{\alpha + \beta x_i\};$
$\mathbf{M_1}$	$\exp\{\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i=High]}\};$
$\mathbf{M_2}$	$\exp\{\alpha + \beta x_i + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\};$
$\mathbf{M_3}$	$\exp\{\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i=High]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\}.$

منبع: یافته‌های تحقیق

در جدول ۱، مدل M_0 یک مدل گومپرتز بدون متغیر کمکی، مدل‌های M_1 و M_2 تنها دارای یک متغیر کمکی (به‌ترتیب مقدار مزایا و مشخصات جغرافیایی جمعیتی) و مدل M_3 هر دو متغیر کمکی را دارد. برای مجموعه دادهٔ کامل، تابع درستنمایی به‌صورت رابطهٔ (۵) است. لذا تنها برآورد τ نیاز است، پس کافی است فقط عامل $L^{\tau}(\tau)$ را درنظر گرفته و $L^{\zeta}(\zeta)$ را نادیده بگیریم. به‌عنوان مثال رابطهٔ $L^{\tau}(\tau)$ برای مدل M_3 به‌صورت زیر است:

$$L^{\tau}(\tau) = \prod_i exp\left\{-\int_{.}^{t_i} exp\{\alpha + \beta(x_i + s) + \gamma \mathbb{1}_{[b_i=High]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\} ds\right\} \times exp\{\alpha + \beta(x_i + t_i) + \gamma \mathbb{1}_{[b_i=High]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\}^{d_i}, \quad (۱۱)$$

لذا لگاریتم درستنمایی به‌صورت زیر حاصل می‌شود:

$$l(\tau) = -\sum_i exp\left\{-\left(\frac{\exp(\beta t_i) - 1}{\beta}\right)\exp\left(\alpha + \beta x_i + \gamma \mathbb{1}_{[b_i=High]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\right)\right\} + \sum_{d_i=1}\left(\alpha + \beta(x_i + t_i) + \gamma \mathbb{1}_{[b_i=High]} + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\right). \quad (۱۲)$$

تابع درستنمایی برای سایر مدل‌ها به‌طریق مشابه حاصل می‌گردد. برای بررسی نیکویی برازش مدل‌ها، معیار اطلاع بیزی[1] (BIC) را محاسبه و آن‌ها با یکدیگر مقایسه می‌کنیم. این نتایج برای مجموعه داده‌های کامل در جدول ۲ ارائه شده‌است.

جدول ۲. برآورد پارامترها، لگاریتم درستنمایی، BIC و برآورد خطای استاندارد برآورد پارامترها.

مدل	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\delta}_1$	$\hat{\delta}_2$	$L^r(\tau)$	BIC
M_0	−۱۱/۵۸	۰/۱۱	–	–	–	−۲۰۵۹۲/۶۱	۴۱۲۰۴/۹۰
$\hat{\sigma}(\hat{\theta}) \times ۱۰^{۳}$	۹۹/۱۳۰	۶۷/۱	–	–	–	–	–
M_1	−۱۱/۵۴	۰/۱۱	−۰/۲۸	–	–	−۲۰۵۶۱/۸۸	۴۱۱۵۳/۲۷
$\hat{\sigma}(\hat{\theta}) \times ۱۰^{۳}$	۵۱/۱۳۱	۶۸/۱	۶۳/۱۳	–	–	–	–
M_2	−۱۱/۴۲	۰/۱۱	–	−۰/۲۴	−۰/۴۸	−۲۰۵۲۸/۹۳	۴۱۰۹۷/۲۱
$\hat{\sigma}(\hat{\theta}) \times ۱۰^{۳}$	۹۷/۱۳۲	۶۸/۱	–	۵۷/۳۳	۰۴/۴۳	–	–
M_3	−۱۱/۴۱	۰/۱۱	−۰/۲۱	−۰/۲۲	−۰/۴۳	−۲۰۵۱۲/۶۹	**٤١٠٧٤/٥٧**
$\hat{\sigma}(\hat{\theta}) \times ۱۰^{۳}$	۲۵/۱۳۳	۶۹/۱	۸۵/۳۶	۷۵/۳۳	۸۹/۴۳	–	–

منبع: یافته‌های تحقیق

براساس نتایج جدول ۲، مدل‌های حاوی متغیرهای کمکی B و C (به‌صورت جداگانه یا توأم)، برازش مدل را بهبود می‌بخشند. کمترین مقدار BIC برای مدل M_3 است که هر دو متغیر کمکی را دارد. بنابراین M_3 مدل بهتری برای داده‌های موجود است. برای بررسی کیفیت برازش مدل‌ها، باقی‌مانده انحراف پواسون به‌صورت رابطه زیر محاسبه می‌شود:

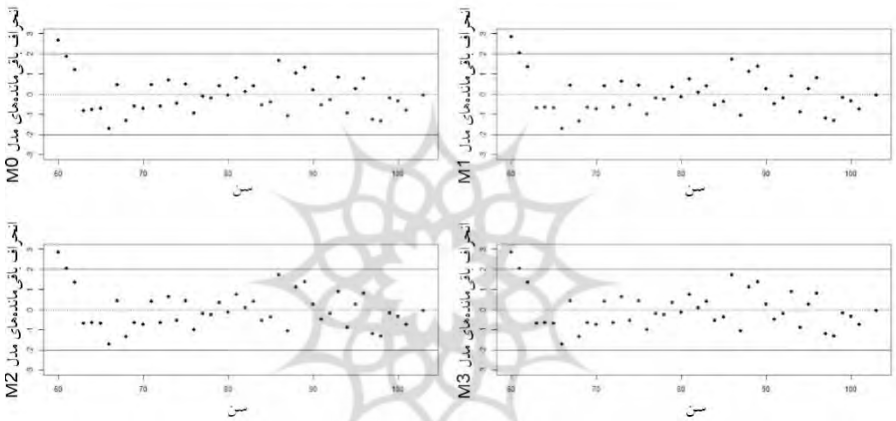
$$r_i = sign(d_i - \Lambda_i)\sqrt{2\left[d_i log\frac{d_i}{\Lambda_i} - (d_i - \Lambda_i)\right]}, \qquad (۱۳)$$

که d_i مجموع تعداد مرگ‌ومیرها در سن iام و $\Lambda_i = \sum_j \lambda_{i,j}$ است. پارامتر توزیع پواسون λ برای مدل M_1 و $i = ۱, \dots, ۴۴$ و $j = ۱, \dots, ۱۸۷۴۱$ از رابطه زیر محاسبه می‌شود:

1. Bayesian Information Criterion

$$\lambda_{i,j} = \left(\frac{\exp\left(\beta \min\left((60 + i - x_{i,j}), (la_{i,j} - x_{i,j})\right)\right) - 1}{\beta} \right) \times \exp\left(\alpha + \beta x_{i,j} + \gamma \mathbb{1}_{[b_{i,j}=High]}\right), \quad (۱۴)$$

که $la_{i,j} = x_{i,j} + t_{i,j}$ و $\Lambda_i = \sum_{j=۱}^{۱۸۷۴۱} \lambda_{i,j}$ است. باقی‌مانده انحراف پواسون برای همه مدل‌ها در شکل ۲ رسم شده است. مشاهده می‌شود که باقی‌مانده‌ها در اطراف صفر متمرکز و از مدل خاصی تبعیت نمی‌کنند.



شکل ۲. نمودار باقی‌مانده انحراف پواسون در برابر سن برای هر ٤ مدل

منبع: یافته‌های تحقیق

با توجه به مقادیر ویژهٔ ماتریس هسین در برآورد پارامترها که همگی منفی و از صفر دور هستند، لذا تمامی مدل‌ها شناسایی‌پذیر خواهند بود. همچنین باتوجه به مجموعه داده‌های کامل، توزیع توام متغیرهای کمکی B و C به‌وسیلهٔ فراوانی نسبی آن‌ها که در جدول ۳ گزارش شده، برآورد شده است. این مقادیر همان مقادیر بردار پارامتری ζ هستند.

جدول ۳. فراوانی نسبی سطح مزایا و مشخصات جغرافیایی جمعیتی برای طرح بازنشستگی

مشخصات جغرافیایی جمعیتی (C) / سطح مزایا (B)	مشخصات جغرافیایی جمعیتی (C)			
	۰	۱	۲	
پایین	۰/۱۹	۰/۴۳	۰/۱۳	۰/۷۵
بالا	۰/۰۳	۰/۱۳	۰/۰۹	۰/۲۵
	۰/۲۲	۰/۵۶	۰/۲۲	۱

منبع: یافته‌های تحقیق

٤-۲. تحلیل و مدل‌بندی بر اساس مشاهدات گمشده

در این بخش با استفاده از حذف برخی از اطلاعات افراد، دادهٔ گمشده را شبیه‌سازی می‌کنیم. برای این منظور مجموعه داده کامل را به‌طور تصادفی به چند بخش تقسیم و یکی از متغیرهای کمکی یا هر دو آن‌ها را در هر زیر مجموعه دادهٔ کامل حذف می‌کنیم.

٤-۲-۱. تولید مجموعه داده حاوی مشاهدات گمشده

برای تولید مجموعه داده‌های حاوی مشاهدات گمشده، مجموعه دادهٔ کامل را که شامل n فرد است، به‌صورت تصادفی به دو زیرمجموعه تقسیم و آن‌ها را با P_1 و P_2 نمایش می‌دهیم که هرکدام به‌ترتیب شامل $n_1 = ۹۳۷۰$ و $n_2 = ۹۳۷۱$ عضوهستند ($n_1 + n_2 = n$). فرض می‌کنیم برای هر فرد در P_1 تنها متغیر کمکی B را مشاهده کردیم (C را حذف کردیم) و برای هر فرد در P_2 تنها متغیر کمکی C را مشاهده کردیم (B را حذف کردیم). این دو زیرمجموعهٔ را به‌عنوان دو طرح بازنشستگی جداگانه درنظر می‌گیریم که علاقه‌مند به مدل‌بندی آن‌ها به‌صورت توأم هستیم.

از آنجائی که مجموعه دادهٔ ترکیبی توسط یک منبع تولید شده (مجموعه داده کامل)، لذا هیچ ناهمگنی بین داده‌ها وجود ندارد و می‌توان یک قانون مرگ‌ومیر یکسان برای اعضای P_1 و P_2 درنظرگرفت.

به‌دلیل این‌که متغیرهای گمشده رسته‌ای هستند، انتگرال در رابطه ۷ به یک مجموع متناهی روی فضای نمونهٔ مشاهدات گمشده تبدیل می‌شود. برای هر فرد در P_1 که مشخصات جغرافیایی جمعیتی مشاهده نشده است، سهم تابع درستنمایی به‌صورت زیر است:

$$L_i(\tau,\zeta) \propto \sum_{c\in\{0,1,2\}} S_{x_i}(t_i|b_i,c;\tau)\mu^{d_i}_{x_i+t_i}(b_i,c;\tau)f_{B,C}(b_i,c;\zeta). \qquad (۱۵)$$

و برای اعضای P_2 که متغیر مزایا مشاهده نشده است، سهم تابع درستنمایی به‌فرم زیر است:

$$L_i(\tau,\zeta) \propto \sum_{b\in\{Low,High\}} S_{x_i}(t_i|b,c_i;\tau)\mu^{d_i}_{x_i+t_i}(b,c_i;\tau)f_{B,C}(b,c_i;\zeta). \qquad (۱۶)$$

بنابراین تابع درستنمایی به‌صورت زیر است:

$$L(\tau,\zeta|t,x,b_{obs},c_{obs},d) \propto \prod_{\{i\in P1\}} L_i \prod_{\{i\in P2\}} L_i. \qquad (۱۷)$$

هدف براورد توأم پارامتر τ از تابع مرگ‌ومیر، پارامتر ζ از تابع احتمال توأم برای سطوح مزایا و مشخصات جغرافیایی جمعیتی است. یک روش منطقی ساختن یک توزیع توأم مناسب برای B و C است.

٤-۲-۲. شناسایی‌پذیری

حال دو شرط ریاضی که برای شناسایی‌پذیری بردار پارامتری $\theta = (\tau,\zeta)$ لازم است را ارائه می‌کنیم. فرض می‌کنیم که هیچ‌کدام از متغیرهای کمکی B و C برای یک فرد مشاهده نشده‌اند. سهم درستنمایی توسط توزیع آمیختهٔ متناهی که اجزای آن تمام نتایج ممکن (B,C) است، ارائه می‌شود. براساس رابطهٔ (۷) تابع چگالی احتمال آمیخته با ۶ مولفه، را می‌توان به‌صورت زیر نوشت:

$$f_x(t;\tau,\zeta) = \zeta_1 f_x(t|b=Low, c=0;\tau) + \zeta_2 f_x(t|b=High, c=0;\tau) + \zeta_3 f_x(t|b=Low, c=1;\tau) + \zeta_4 f_x(t|b=High, c=1;\tau) + \zeta_5 f_x(t|b=Low, c=2;\tau) + \zeta_6 f_x(t|b=High, c=2;\tau). \quad (۱۸)$$

شروط زیر برای شناسایی‌پذیری مدل آمیختهٔ متناهی لازم است (مک‌لاکن و پیل[1]، ۲۰۰۰):

۱. برای سنین مختلف با طول‌عمرهای متفاوت، مقادیر متفاوت (B,C) باید تابع چگالی احتمال متفاوتی را نتیجه دهند. یعنی برای $h \neq k$ باید داشته باشیم:

$$f_x(t|(b,c)=h;\tau) \neq f_x(t|(b,c)=k;\tau). \quad (۱۹)$$

که این شرط در مورد تابع بقا، به این معنا است که مقادیر متفاوت (B,C) تابع مرگ‌ومیر متفاوتی را بایستی نتیجه دهند.

۲. مقادیر ζ_j برای $j = ۱, \dots, ۶$ بایستی به‌صورت $۰ < \zeta_j < ۱$ باشد.

شایان ذکر است که شرط‌ها لازم هستند، اما کافی نیستند، (تیترینگتون و همکاران[2]، ۱۹۸۵). اگر دو شرط بالا برقرار شد، مدل آمیخته در رابطهٔ ۱۸ شناسایی‌پذیر است.

زمانی‌که داده گمشده وجود دارد، شناسایی‌پذیری مدل را از روش‌های دیگری نیز می‌توان بررسی نمود. هنگام برآورد پارامترهای مدل ممکن است برآوردهایی به‌دست آیند که دارای واریانس‌های بسیار بزرگ و یا بسیار کوچک و نزدیک به صفر هستند و یا ممکن است برآوردها در فضای پارامتری موردنظر، تعریف نشوند، که می‌توان نتیجه گرفت مدل شناسایی‌پذیر نیست. همچنین شناسایی‌پذیری را می‌توان از طریق مقادیر ویژهٔ ماتریس هسین[3] (حاصل‌شده در مرحله برآوردیابی) نیز بررسی نمود. در صورتی که مقادیر ویژه[4]، همگی منفی و از صفر دور به‌دست آیند، نتیجه می‌شود مدل آماری شناسایی‌پذیر است.

1. McLachlan & Peel
2. Titterington et al
3. Hessian Matrix
4. Eigenvalues

٤-٢-٣. الگوریتم برازش

زمانی‌که داده‌ها گمشده هستند، به‌حداکثر رساندن تابع درستنمایی ممکن است به‌آسانی صورت نگیرد. مسئلهٔ کلی محاسباتی این است که چگونه محدودیت‌ها روی بردار پارامتری ζ را با توجه به توزیع توام B و C، کنترل نماییم (عناصر ζ بین صفر و یک محدود شده‌اند و مجموع آن‌ها برابر یک است). به‌همین دلایل هر بهینه‌سازی مقید معمولی ممکن است شامل عدم پایداری و استواری شود.

الگوریتم EM که ابتدا توسط دمپستر و همکاران[1] (۱۹۹۷) ارائه شد، برآورد را برای برخی مسائل آسان می‌نمود. درادامه یک اصلاح جدید برای تبدیل مسئله بهینه‌سازی مقید به بهینه‌سازی بدون قید که منجر به همگرایی سریع‌تر نسبت به الگوریتم EM و نیز استواری برآوردها نسبت به مقادیر اولیهٔ متفاوت می‌شود، ارائه می‌شود. برای تبدیل مسئلهٔ بهینه‌سازی مقید به یک مسئلهٔ بدون قید، از تبدیل نسبت لگاریتمی ایزومتریک[2] (ILR) که توسط اگزکو و همکاران[3] (۲۰۰۳) ارائه شد، استفاده می‌شود. این تبدیل حاوی نگاشت حفظ فاصله از k بعدی ساده به فضای $\mathbb{R}^{k-1}$ بعدی است و بیشتر در تجزیه‌وتحلیل داده‌های ترکیبی برای پارامترسازی مجدد بردار$\zeta = (\zeta_1, \dots, \zeta_k)$ استفاده می‌شود.

$$\pi = ILR(\zeta) = \Psi^T \log(\zeta) \qquad (۲۰)$$

که π یک بردار $k-1$ بعدی و Ψ یک ماتریس با ابعاد $\left(k \times (k-1)\right)$ است که ستون‌های $(\psi_1, \dots, \psi_{k-1})$ پایه‌های یکامتعامد را برای ابرصفحه از $\mathbb{R}^k$ متعامد به بردار k واحدی ۱ نشان را می‌دهند که به‌صورت زیر تعریف می‌شود:

$$\psi_i = \sqrt{\frac{i}{i+1}}\left[\underbrace{\frac{1}{i}, \dots \frac{1}{i}}_{\text{عنصر } i}, -1, \underbrace{0, \dots, 0}_{\text{عنصر } k-i-1}\right]. \qquad (۲۱)$$

1. Dempster et al
2. Isometric Log-Ratio Transform
3. Egozcue et al

برای به‌دست آوردن ζ برای هر π موجود، معکوس تبدیل ILR در رابطه ۲۰ به‌صورت زیر است:

$$\zeta = ILR^{-1}(\pi) = \left[\frac{\exp(\psi_1\pi)}{\sum_{i=1}^{k}\exp(\psi_i\pi)}, \dots, \frac{\exp(\psi_k\pi)}{\sum_{i=1}^{k}\exp(\psi_i\pi)}\right]. \quad (۲۲)$$

حال از تبدیل ILR برای تغییر مسئلهٔ بهینه‌سازی مقید به یک مسئلهٔ بدون قید، استفاده می‌کنیم. لذا الگوریتم بهینه‌سازی پیشنهادی شامل گام‌های زیر می‌شود:

۱. مقادیر اولیه‌ای برای بردارهای پارامتری $\hat{\tau}^{(0)}$ و $\hat{\zeta}^{(0)}$ در نظر بگیرید.

۲. براساس تبدیل ILR مقدار $\hat{\pi}^{(0)}$ را از روی $\hat{\zeta}^{(0)}$ محاسبه نمایید.

$$\hat{\pi}^{(0)} = ILR\left(\hat{\zeta}^{(0)}\right). \quad (۲۳)$$

۳. در تابع لگاریتم درستنمایی ζ_j را با مولفه‌های فردی زیر جایگذاری نمایید.

$$\frac{\exp(\psi_j\pi)}{\sum_{i=1}^{k}\exp(\psi_i\pi)}, \quad j = 1, \dots, k \quad (۲۴)$$

۴. مقادیر $\hat{\tau}$ و $\hat{\pi}$ را به‌وسیلهٔ

$$(\hat{\tau}, \hat{\pi}) = \arg\max_{(\tau,\pi)} \log L\left(\tau, ILR^{-1}(\pi)\right). \quad (۲۵)$$

۵. با استفاده از طرح تکراری نیوتن رافسون با مقادیر اولیهٔ $\hat{\tau}^{(0)}$ و $\hat{\pi}^{(0)}$ محاسبه نمایید.

۶. تبدیل ILR معکوس را برای محاسبهٔ $\hat{\zeta}$ از روی $\hat{\pi}$ استفاده نمایید.

$$\hat{\zeta} = ILR^{-1}(\hat{\pi}). \quad (۲۶)$$

از آنجائی که جواب رابطهٔ درستنمایی می‌تواند از انتخاب مقدارهای اولیه، تأثیر بگیرد. در پیاده‌سازی، چندین مقدار اولیه درنظر گرفته شد و MLE ای انتخاب شد که بیش‌ترین مقدار را برای تابع لگاریتم درستنمایی نتیجه می‌داد.

٤-٢-٤. نتایج تجربی برای مجموعه مشاهدات گمشده

برای تحلیل خصوصیات برآوردگر $(\hat{\tau},\hat{\zeta})$، ابتدا مجموعه دادهٔ کامل را به‌صورت تصادفی به دو بخش مطابق ساختار بخش ۲-۴ تقسیم می‌کنیم. بدین ترتیب دو مجموعه دادهٔ مصنوعی P_1 و P_2 ایجاد می‌گردد و سپس آن‌ها را به‌صورت مصنوعی ترکیب می‌نماییم که برای برآورد $\theta = (\tau,\zeta)$ استفاده می‌شوند. برآوردگر بر پایهٔ P_1 و P_2 با $\hat{\theta}_1$ نمایش داده می‌شود. با توجه به تابع چگالی رابطه ۱۸، لگاریتم درستنمایی برای مجموعه P_1 به‌صورت زیر است:

$$
\begin{aligned}
l_i(\tau,\zeta) \propto\ & (\zeta_1+\zeta_2)exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp\big(\alpha+\beta x_i \right. \\
& \left. +\gamma \mathbb{1}_{[b_i=High]}\big)\right\} \\
& \times exp\{\alpha+\beta(x_i+t_i)+\gamma\mathbb{1}_{[b_i=High]}\}^{d_i} \\
& +(\zeta_3+\zeta_4)exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp\big(\alpha+\beta x_i+\gamma\mathbb{1}_{[b_i=High]} \right. \\
& \left. +\delta_1\big)\right\} \\
& \times exp\{\alpha+\beta(x_i+t_i)+\gamma\mathbb{1}_{[b_i=High]}+\delta_1\}^{d_i} \\
& +(\zeta_5+\zeta_6)exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp\big(\alpha+\beta x_i+\gamma\mathbb{1}_{[b_i=High]} \right. \\
& \left. +\delta_2\big)\right\} \\
& \times exp\{\alpha+\beta(x_i+t_i)+\gamma\mathbb{1}_{[b_i=High]}+\delta_2\}^{d_i}.
\end{aligned}
\tag{۲۷}
$$

لگاریتم درستنمایی برای مجموعه P_2 به‌صورت زیر است:

$$
\begin{aligned}
& l_i(\tau,\zeta) \propto (\zeta_1+\zeta_3+\zeta_5) \\
& \times exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp\big(\alpha+\beta x_i+\delta_1\mathbb{1}_{[c_i=1]}+\delta_2\mathbb{1}_{[c_i=2]}\big)\right\} \\
& \times exp\{\alpha+\beta(x_i+t_i)+\delta_1\mathbb{1}_{[c_i=1]}+\delta_2\mathbb{1}_{[c_i=2]}\}^{d_i} \\
& + \\
& \times exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp\big(\alpha+\beta x_i+\gamma+\delta_1\mathbb{1}_{[c_i=1]}+\delta_2\mathbb{1}_{[c_i=2]}\big)\right\}
\end{aligned}
\tag{۲۸}
$$

$$\times exp\{\alpha + \beta(x_i + t_i) + \gamma + \delta_1 \mathbb{1}_{[c_i=1]} + \delta_2 \mathbb{1}_{[c_i=2]}\}^{d_i}.$$

ابتدا تنها یک‌بار سناریو تقسیم‌بندی تصادفی به P_1 و P_2 را انجام می‌دهیم. برآورد پارامترها در این سناریو را با $(\hat{\theta}_1)$ نمایش می‌دهیم. سناریوی دیگر ۱۰۰۰ بار تقسیم کردن مجموعه دادهٔ کامل به‌صورت تصادفی است و میانگین برآورد پارامترها $(\bar{\hat{\theta}} = \frac{1}{1000}\sum_k \hat{\theta}_k)$ و همچنین انحراف استاندارد پارامترها $(\sigma_{PAR} \times 10^3)$ $(\sigma_{PAR} = \sqrt{\frac{1}{999}\sum_k(\hat{\theta}_k - \bar{\hat{\theta}})^2})$ محاسبه می‌نماییم. برآورد پارامترها برای مجموعه دادهٔ کامل را با $(\hat{\theta})$ نمایش می‌دهیم. کلیه نتایج در جدول ۴ ارائه شده است.

جدول ٤. برآورد و انحراف استاندارد پارامترها در داده کامل و گمشده با سناریو تقسیم‌بندی مختلف

برآورد پارامتر	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\zeta}_1$	$\hat{\zeta}_2$	$\hat{\zeta}_3$	$\hat{\zeta}_4$	$\hat{\zeta}_5$
$\hat{\theta}$	-۴۱/۱۱	۱۱/۰	-۲۱/۰	-۲۲/۰	-۴۳/۰	۱۹/۰	۰۳/۰	۴۳/۰	۱۳/۰	۱۳/۰
$\hat{\theta}_1$	-۴۱/۱۱	۱۱/۰	-۰۸/۰	-۲۷/۰	-۳۷/۰	۲۱/۰	۰۰/۰	۵۴/۰	۰۲/۰	۰۰/۰
$\bar{\hat{\theta}}$	-۴۲/۱۱	۱۱/۰	-۱۳/۰	-۲۳/۰	-۳۴/۰	۲۲/۰	۰۰/۰	۵۲/۰	۰۴/۰	۰۱/۰
$\sigma_{PAR} \times 10^3$	۴۳/۱۷	۲۱/۰	۲۴/۱۰۱	۲۲/۳۳	۹۴/۱۱۴	۰۹/۳	۸۷/۰	۷۴/۱۷	۴۹/۱۷	۰۸/۱۷

منبع: یافته‌های تحقیق

برای بررسی دلایل برآورد ضعیف به‌دست آمده از ترکیب P_1 و P_2، به بررسی شناسایی‌پذیری پارامترها می‌پردازیم. برای این منظور، براساس تابع مرگ‌ومیر موجود تابع ϕ را برای مقادیر مختلف (B, C) و با توجه به سطح پایه و سن x به صورت زیر تعریف می‌کنیم:

$$\phi = exp\left[\gamma \mathbb{1}_{[b=High]} + \delta_1 \mathbb{1}_{[c=c_1]} + \delta_2 \mathbb{1}_{[c=c_2]}\right].$$

جدول ۵ مقادیر $\hat{\phi}$ را برای مقادیر مختلف (B, C) برای مجموعه دادهٔ کامل، نشان می‌دهد.

جدول ۵. مقادیر برآورد ϕ ($\widehat{\Phi}$) برای مجموعه دادهٔ کامل

مشخصات جغرافیایی جمعیتی (C) / سطح مزایا (B)	۰	۱	۲
پایین	۱	۰/۸۰	۰/۶۵
بالا	۰/۸۱	۰/۶۵	۰/۵۳

منبع: یافته‌های تحقیق

در جدول ۵ مشاهده می‌شود که فردی با مشخصات جغرافیایی جمعیتی «صفر» و سطح مزایای بالا، با فردی که سطح مزایا پایین دارد و مشخصات جغرافیایی جمعیتی او «یک» است، مرگ‌ومیر بسیار مشابهی (در سن‌های یکسان) دارند. این موضوع برای افرادی که مزایای بالا و مشخصات جغرافیایی جمعیتی آن‌ها یک است، با افرادی که دارای مزایای پایین و مشخصات جغرافیایی جمعیتی «دو» هستند، نیز برقرار است. بنابراین شرایط شناسایی‌پذیری، برقرار نمی‌باشد.

همچنین شناسایی‌پذیری پارامترها در مدل مطرح‌شده با داده‌های موجود، می‌تواند از طریق بررسی مقادیر ویژهٔ ماتریس هسین تحلیل شود. مقادیر ویژه برای تمام ۱۰۰۰ مجموعه دادهٔ شبیه‌سازی شده محاسبه شده است. مشاهده می‌شود که دو مقدار بزرگ‌تر مقادیر ویژهٔ ماتریس هسین بسیار نزدیک به صفر هستند که نشان دهندهٔ این است که ماتریس اطلاع برآورد شده، رتبه کامل نیست و مدل برای داده‌های موجود، دارای پارامتر افزونه است. علاوه‌بر این ممکن است یک مسئله در ریسک مدل موجود باشد. ما یک مدل گومپرتز برای داده‌ها استفاده کرده‌ایم که ممکن است نتواند تابع مرگ‌ومیر درستی را برای این تجربهٔ مرگ‌ومیر منعکس کند و یا ممکن است که اختلافات مرگ‌ومیر به‌دلیل ویژگی‌های مختلف اقتصادی و اجتماعی به‌طور متناسب بر روی لگاریتم خطر تأثیر نگذارد.

٤-٣. دسترسی به اطلاعات بیشتر

مسئلهٔ شناسایی‌پذیری زمانی رخ می‌دهد که دو متغیر کمکی که هیچ‌گاه به‌صورت مشترک مشاهده نشده‌اند، مانع از استنباط بیشتر درمورد تابع مرگ‌ومیر می‌شود. اکنون برای بررسی تأثیر دسترسی به مجموعه دادهٔ بیشتر حالت‌های مختلفی را مدنظر قرار می‌دهیم. مجموعه داده P_3 اندازهٔ نمونهٔ کوچکتری به تعداد n_3 دارد که نمایندهٔ طرح بازنشستگی است که هیچ اطلاعاتی از دو متغیر کمکی، گمشده نیستند. برای نشان دادن استواری نتایج، مجموعه دادهٔ P_4 با اندازه نمونهٔ بزرگتر به تعداد n_4 در نظر گرفته می‌شود که نمایندهٔ طرح بازنشستگی است که در آن B و C هر دو گمشده هستند.

حال چهار طرح بازنشستگی P_1، P_2، P_3 و P_4 داریم که از طریق تقسیم‌بندی مجموعه داده‌های کامل به‌صورت تصادفی و حذف عوامل مقتضی به‌دست آورده می‌شوند. در این رویکرد، ۵ درصد به‌طور کامل مشاهده شده‌اند (P_3)، ۲۰ درصد فقط B مشاهده شده (P_1)، ۲۰ درصد فقط C مشاهده شده (P_2) و در ۵۵ درصد باقی‌مانده هر دو B و C مشاهده نشده‌اند (P_4). از این‌رو $n_1 = n_2 =$ ۳۷۴۸، $n_3 =$ ۹۳۷ و $n_4 =$ ۱۰۳۰۸ است. این نحوه تقسیم‌بندی تصادفی ۱۰۰۰ بار تکرار می‌شود. بنابراین سهم تابع درستنمایی برای هر فرد در P_3 (بدون مشاهده گمشده) به‌صورت زیر است:

$$L_i(\tau,\zeta) \propto S_{x_i}(t_i|b_i,c_i;\tau)\mu^{d_i}_{x_i+t_i}(b_i,c_i;\tau)f_{B,C}(b_i,c_i;\zeta). \quad (۲۹)$$

لذا:

$$\begin{aligned} L_i(\tau,\zeta) &\propto (\zeta_1+\zeta_2+\zeta_3+\zeta_4+\zeta_5+\zeta_6) \\ &\times exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp\big(\alpha+\beta x_i+\gamma\mathbb{1}_{[b_i=High]}+\delta_1\mathbb{1}_{[c_i=1]}\right. \\ &\left.+\delta_2\mathbb{1}_{[c_i=2]}\big)\right\} \\ &\times exp\big\{\alpha+\beta(x_i+t_i)+\gamma\mathbb{1}_{[b_i=High]}+\delta_1\mathbb{1}_{[c_i=1]}+\delta_2\mathbb{1}_{[c_i=2]}\big\}^{d_i}. \end{aligned} \quad (۳۰)$$

و سهم تابع درستنمایی برای افراد در P_4 (هر دو B و C گمشده) به‌صورت زیر است:

$$L_i(\tau,\zeta) \propto \sum_{b\in\{Low,High\}}\sum_{c\in\{0,1,2\}} S_{x_i}(t_i|b,c;\tau)\mu^{d_i}_{x_i+t_i}(b,c;\tau)f_{B,C}(b,c;\zeta). \quad (۳۱)$$

در نتیجه داریم:

$$
\begin{aligned}
L_i(\tau,\zeta) \propto\ & \zeta_1 exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp(\alpha+\beta x_i)\right\}exp\{\alpha+\beta(x_i \\
& +t_i)\}^{d_i} \\
& +\zeta_2 exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp(\alpha+\beta x_i+\gamma)\right\}exp\{\alpha+\beta(x_i+t_i) \\
& +\gamma\}^{d_i} \\
& +\zeta_3 exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp(\alpha+\beta x_i+\delta_1)\right\}exp\{\alpha+\beta(x_i+t_i) \\
& +\delta_1\}^{d_i} \\
& +\zeta_4 exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp(\alpha+\beta x_i+\gamma+\delta_1)\right\}exp\{\alpha+\beta(x_i \\
& +t_i)+\gamma+\delta_1\}^{d_i} \\
& +\zeta_5 exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp(\alpha+\beta x_i+\delta_2)\right\}exp\{\alpha+\beta(x_i+t_i) \\
& +\delta_2\}^{d_i} \\
& +\zeta_6 exp\left\{-\left(\frac{\exp(\beta t_i)-1}{\beta}\right)exp(\alpha+\beta x_i+\gamma+\delta_2)\right\}exp\{\alpha+\beta(x_i \\
& +t_i)+\gamma+\delta_2\}^{d_i}.
\end{aligned}
\tag{۳۲}
$$

در ادامه، مقادیر برآورد پارامترها و خطاهای استاندارد به‌دست آمده از مجموعه داده‌های مختلف را با هم مقایسه می‌کنیم. برآوردها بر اساس ترکیب چهار مجموعه داده (با P_1-P_4 مشخص شده)، براساس ترکیب P_1، P_2 و P_3 (با P_1-P_3 مشخص شده) یعنی به غیر از مجموعه داده بزرگتر که هر دو B و C مشاهده نشده‌اند؛ و تنها برای P_3 (بدون داده گمشده ولی اندازه نمونهٔ خیلی کوچک) را در نظر می‌گیریم.

برآورد پارامترها در مجموعه دادهٔ کامل ($\hat{\theta}$)، برآورد پارامترها در یک بار تقسیم مجموعه داده کامل به‌صورت تصادفی ($\hat{\theta}_1$)، برآورد خطاهای استاندارد آن‌ها ($\bar{\hat{\sigma}}(\hat{\theta}_1)\times ۱۰^{۳}$)، میانگین برآورد پارامترها به‌وسیله ۱۰۰۰ بار تقسیم کردن مجموعه داده‌های کامل به‌صورت تصادفی ($\bar{\hat{\theta}}=\frac{۱}{۱۰۰۰}\sum_k \hat{\theta}_k$)، میانگین برآورد خطای استاندارد و انحراف استاندارد پارامترها به ترتیب ($\bar{\hat{\sigma}}(\hat{\theta})=\frac{۱}{۱۰۰۰}\sum_k \hat{\sigma}(\hat{\theta}_k)\times ۱۰۰۰$) و $\sigma_{PAR}=\sqrt{\frac{۱}{۹۹۹}\sum_k(\hat{\theta}_k-\bar{\hat{\theta}})^{۲}}\times ۱۰۰۰$ محاسبه و در جدول ۶ نمایش داده شده است.

این برآوردها را با نتایج حاصل از مجموعه دادهٔ کامل (خط اول جدول ۶) مقایسه می‌کنیم. خطاهای استاندارد با استفاده از منفی معکوس ماتریس اطلاع تجربی برآورد شده‌اند. پروسهٔ برآورد در بخش ۵، شامل (τ, ζ) است. از این رو برآورد خطای استاندارد برای (τ, ζ) نیازمند این است که از روش دلتا برای توابع برداری استفاده شود. این روش تنها برای برآورد خطاهای استاندارد ζ استفاده می‌شود. تابع مورد استفاده، همان رابطه ۲۲ است. بنابراین مشتق تابع مد نظر نسبت به π، یک برداری k بعدی به‌صورت زیر است:

$$\zeta'_j = \frac{\psi_j \exp(\psi_j \pi) \sum_{i=1}^{k} \exp(\psi_i \pi) - \exp(\psi_j \pi) \sum_{i=1}^{k} \psi_i \exp(\psi_i \pi)}{\left(\sum_{i=1}^{k} \exp(\psi_i \pi)\right)^2}. \quad (۳۳)$$

که $j = 1, \dots, k$ است. بنابراین، واریانس آن به‌طور تقریبی از رابطه $var(\zeta) = \zeta' \Sigma \zeta'^T$ حاصل می‌شود، که در آن Σ ماتریس واریانس کوواریانسی است که از طریق منفی معکوس ماتریس هسین به‌دست آمده در برآورد بردار پارامتری π حاصل می‌شود. در نهایت با جذر گرفتن از رابطهٔ فوق، خطای استاندارد مربوط به بردار پارامتری ζ حاصل می‌گردد.

همان‌طور که از مقادیر $\hat{\theta}_k$ و $\bar{\hat{\theta}}$ مشاهده می‌شود برآورد پارامترها با برآوردهای به‌دست آمده براساس مجموعه دادهٔ کامل، مطابقت دارند. اما، چون بخش‌هایی از مجموعه داده کامل از استنباط خارج می‌شوند، خطای استاندارد برآوردشدهٔ پارامترها $\hat{\sigma}(\hat{\theta})$ به‌دلیل کوچک‌تر بودن اندازه نمونه افزایش می‌یابد.

افزایش خطای استاندارد برآورد شده، برای پارامترهای γ، δ_1 و δ_2 به این دلیل است که تفاوت‌های مرگ‌ومیر بر پایهٔ مزایا و مشخصات جغرافیایی جمعیتی، وابسته به متغیرهایی است که گاهی اوقات گمشده هستند. برای پارامترهای α و β این امر تأثیر کم‌تری دارد. به این دلیل که متغیر کمکی سن برای تمام افراد مشاهده‌شده است و خط مبنا برای همه افراد در چهار مجموعه داده طرح‌های بازنشستگی شبیه‌سازی‌شده، مشترک است. این تفاوت‌ها بیشتر هنگام مقایسهٔ خطای استاندارد برآورد شده $\hat{\gamma}$، $\widehat{\delta_1}$ و $\widehat{\delta_2}$ وقتی‌که از هر چهار طرح بازنشستگی $(P_1 - P_4)$ با برآورد فقط بر اساس طرح‌های بازنشستگی $P_1 - P_3$ استفاده می‌شود، تأیید می‌شوند.

برای بررسی شناسایی‌پذیری مدل کامل با ۴ مجموعه داده $P_1 - P_4$، مشابه قبل می‌توان مقادیر ویژه ماتریس هسین حاصل از برآورد هر چهار مجموعه داده، را محاسبه نمود.

مشاهده می‌شود که در تمام ۱۰۰۰ طرح تصادفی داده گمشده، مقادیر ویژه همگی منفی و دور از صفر هستند. همچنین نزدیک‌ترین مقدار ویژه به صفر در تمام شبیه‌سازی‌ها برابر با ۹/۹۳- است. بنابراین مدل بررسی شده، شناسایی‌پذیر است.

جدول ٦. برآورد پارامترها و برآورد خطای استاندارد مربوطه در سناریوهای مختلف

مجموعه داده	برآورد	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\zeta}_1$	$\hat{\zeta}_2$	$\hat{\zeta}_3$	$\hat{\zeta}_4$	$\hat{\zeta}_5$
کامل	$\hat{\theta}$	-۱۱/۴۱	۰/۱۱	-۰/۲۱	-۰/۲۲	-۰/۴۳	۰/۱۹	۰/۰۳	۰/۴۳	۰/۱۳	۰/۱۳
$P_1 - P_4$	$\hat{\theta}_1$	-۱۱/۴۶	۰/۱۱	-۰/۰۸	-۰/۱۶	-۰/۳۹	۰/۱۹	۰/۰۲	۰/۴۳	۰/۱۴	۰/۱۳
	$\bar{\hat{\sigma}}(\hat{\theta}_1)$ $\times 10^3$	۱۳۶/۲۶	۱/۷۱	۶۷/۹۶	۶۳/۱۶	۸۴/۷۱	۶/۶۹	۴/۰۸	۹/۴۴	۷/۹۴	۷/۹۰
	$\bar{\hat{\theta}}$	-۱۱/۴۵	۰/۱۱	-۰/۱۸	-۰/۱۹	-۰/۳۶	۰/۱۹	۰/۰۳	۰/۴۳	۰/۱۳	۰/۱۳
	$\bar{\hat{\sigma}}(\hat{\theta})$ $\times 10^3$	۱۳۶/۴۲	۱/۷۲	۶۸/۴۱	۶۱/۶۵	۸۲/۱۶	۶/۹۴	۴/۶۸	۹/۴۴	۷/۹۴	۷/۶۶
	σ_{PAR} $\times 10^3$	۲۵/۴۹	۰/۲۹	۵۰/۹۷	۴۸/۵۸	۶۳/۲۹	۶/۱۴	۴/۴۸	۸/۵۶	۷/۳۰	۷/۴۷
$P_1 - P_3$	$\hat{\theta}_1$	-۱۱/۵۲	۰/۱۱	-۰/۱۰	-۰/۱۹	-۰/۴۷	۰/۱۹	۰/۰۲	۰/۴۳	۰/۱۴	۰/۱۳
	$\bar{\hat{\sigma}}(\hat{\theta}_1)$ $\times 10^3$	۱۹۹/۲۱	۲/۵۲	۷۷/۰۹	۶۹/۰۱	۹۴/۲۹	۶/۶۹	۴/۰۹	۹/۴۴	۷/۹۵	۷/۹۱
	$\bar{\hat{\theta}}$	-۱۱/۴۳	۰/۱۱	-۰/۲۰	-۰/۲۱	-۰/۴۱	۰/۱۹	۰/۰۳	۰/۴۳	۰/۱۳	۰/۱۳
	$\bar{\hat{\sigma}}(\hat{\theta})$ $\times 10^3$	۲۰۰/۶۱	۲/۵۴	۷۴/۰۵	۶۶/۴۶	۸۸/۵۵	۶/۹۵	۴/۶۹	۹/۴۵	۷/۹۵	۷/۶۷
	σ_{PAR} $\times 10^3$	۱۵۴/۳۲	۱/۹۷	۵۹/۷۳	۵۴/۹۷	۷۱/۳۹	۶/۴۲	۴/۵۷	۸/۴۵	۷/۳۸	۷/۳۷
P_3	$\hat{\theta}_1$	-۱۱/۵۴	۰/۱۱	-۰/۱۶	-۰/۱۳	-۰/۶۶	۰/۱۹	۰/۰۲	۰/۴۵	۰/۱۴	۰/۱۱
	$\bar{\hat{\sigma}}(\hat{\theta}_1)$ $\times 10^3$	۶۱۱/۵۲	۷/۷۴	۱۷۳/۳۷	۱۵۰/۹۲	۲۲۷/۵۲	۱۲/۹۰	۴/۳۶	۱۶/۲۵	۱۱/۴۴	۱۰/۳۵
	$\bar{\hat{\theta}}$	-۱۱/۵۱	۰/۱۱	-۰/۲۱	-۰/۲۲	-۰/۴۴	۰/۱۹	۰/۰۳	۰/۴۳	۰/۱۳	۰/۱۳
	$\bar{\hat{\sigma}}(\hat{\theta})$ $\times 10^3$	۶۰۴/۹۷	۷/۶۶	۱۶۶/۵۷	۱۵۲/۲۵	۱۹۹/۰۵	۱۲/۸۰	۵/۱۴	۱۶/۱۵	۱۱/۱۰	۱۰/۹۸
	σ_{PAR} $\times 10^3$	۶۳۴/۰۷	۸/۱۳	۱۵۲/۹۵	۱۴۶/۰۲	۱۹۰/۲۱	۱۲/۶۹	۵/۰۱	۱۵/۹۴	۱۰/۸۶	۱۰/۷۰

منبع: یافته‌های تحقیق

٤-٤. بررسی تأثیر مالی در استنباط آمار براساس مشاهدات گمشده

٤-٤-١. بررسی نرخ مرگ‌ومیر و عامل مستمری

فرض کنید متغیر تصادفی Y_x نشان دهندهٔ ارزش فعلی جریان‌های نقدی در یک صندوق بازنشستگی برای عضو x ساله است که در طول‌عمر باقی‌مانده خود آن را دریافت می‌کند. اگر این جریان نقدی به میزان یک واحد هر ساله به‌طور پیوسته پرداخت شود، مقدار مورد انتظار برای Y_x، عامل مستمری نامیده شده که با $\overline{a}_x$ نمایش و به‌صورت زیر محاسبه می‌شود (دیکسون و همکاران[1]، ۲۰۱۳):

$$\overline{a}_x = \mathbb{E}(Y_x) = \int_{.}^{۱۲۰-x} \exp(-rt)\, S_x(t)dt. \qquad (۳۴)$$

به‌دلیل گمشدگی داده‌ها، $\hat{\tau}$ به $\hat{\zeta}$ وابسته است، توزیع نمونه‌گیری $\hat{\theta} = (\hat{\tau}, \hat{\zeta})$ به‌صورت مجانبی دارای توزیع نرمال چندمتغیره[2] با میانگین θ و ماتریس واریانس کوواریانس $\boldsymbol{\Sigma}$ است، که $\boldsymbol{\Sigma}$ از طریق منفی معکوس ماتریس اطلاع فیشر[3] برآورد می‌شود.

برای بررسی نرخ مرگ‌ومیر و عامل مستمری، مقادیر عامل مستمری و درصد تغییر آن در مقایسه با عامل مستمری در مدل M_0 محاسبه و نتایج در جدول ۷ ارائه‌شده است. برآورد پارامترها با مشاهدات هر چهار طرح $P_1 - P_4$ در نظر گرفته می‌شوند. تجزیه و تحلیل عددی برای دو نرخ بهره ۱ و ۳ درصد و دو فرد به سن‌های ۶۵ و ۷۱ سال انجام شده است. اولین سن، همان سن معمول بازنشستگی و دومین سن به‌طور تقریبی، میانگین سن افراد زنده در مجموعه داده کامل در زمان اتمام دوره مشاهده است.

نتایج جدول نشان می‌دهد که برای سن‌های انتخابی و دو نرخ بهره محاسبه‌شده، بین عامل‌های مستمری به‌دست آمده از مدل مبتنی‌بر متغیرهای کمکی[4] اقتصادی-اجتماعی و مقادیر به‌دست آمده از مدل M_0 تفاوت‌های قابل توجهی وجود دارد. این اختلافات برای نرخ بهره پایین‌تر، بیشتر است که بیانگر این است که وقتی نرخ بهره پایین است،

1. Dickson et al
2. Multivariate Normal Distribution
3. Fisher Information Matrix

مرگ‌ومیر عامل خطر مهم‌تری برای مستمری است. برای بررسی تأثیر مالی استفاده از مدل مرگ‌ومیر معرفی شده، می‌توان الزام سرمایه را بررسی نمود که نتیجهٔ آن وابسته به عامل مستمری و در نتیجه مدل مرگ‌ومیر است.

جدول ۷. مقادیرمستمری، درصد تغییرات بر اساس مدل M_0 و نرخ‌های بهره مختلف

سن	مجموعه داده	تقسیم‌بندی	$r = ۱٪$		$r = ۳٪$	
			$\overline{a}_x$	درصد تغییر		درصد تغییر
۶۵	$P_1 - P_4$	M_0	۷۷/۱۶	–	۷۰/۱۳	–
		$b = L, c = ۰$	۶۱/۱۵	-۹۳/۶	۹۰/۱۲	-۸۱/۵
		$b = H, c = ۰$	۱۱/۱۶	-۹۳/۳	۲۵/۱۳	-۲۶/۳
		$b = L, c = ۱$	۵۷/۱۶	-۲۱/۱	۵۶/۱۳	-۹۸/۰
		$b = H, c = ۱$	۰۸/۱۷	۸۴/۱	۹۱/۱۳	۵۴/۱
		$b = L, c = ۲$	۹۵/۱۷	۰۱/۷	۴۹/۱۴	۷۷/۵
		$b = H, c = ۲$	۴۷/۱۸	۱۰/۱۰	۶۱/۱۴	۶۶/۶
۷۱	$P_1 - P_4$	M_0	۹۸/۱۲	–	۰۲/۱۱	–
		$b = L, c = ۰$	۹۰/۱۱	-۳۳/۸	۲۲/۱۰	-۲۸/۷
		$b = H, c = ۰$	۳۶/۱۲	-۷۸/۴	۵۷/۱۰	-۱۵/۴
		$b = L, c = ۱$	۷۸/۱۲	-۵۴/۱	۸۸/۱۰	-۳۲/۱
		$b = H, c = ۱$	۲۶/۱۳	۱۰/۲	۲۲/۱۱	۸۳/۱
		$b = L, c = ۲$	۰۷/۱۴	۳۴/۸	۸۱/۱۱	۱۶/۷
		$b = H, c = ۲$	۵۶/۱۴	۱۱/۱۲	۹۴/۱۱	۲۹/۸

منبع: یافته‌های تحقیق

٤-٤-١. ریسک برآورد نادرست الزام سرمایه

هدف این بخش نحوه کاهش الزام سرمایه براساس طرح بازنشستگی موجود، و ترکیب طرح‌های بازنشستگی با دیگر با هدف تخمین پارامترهای نرخ مرگ‌ومیر است.

با استفاده از رویکرد ریچارد[1] (۲۰۱۶) برای محاسبه SCR[2]، بر اساس مجموعهٔ بزرگی از مقادیر پارامتر جایگزین، ارزیابی‌های مکرری از کل پرتفوی مستمری انجام می‌دهیم. مقدار پرتفوی به‌عنوان تابعی از پارامتر θ با $P(\theta)$ نشان داده می‌شود. الزام سرمایه ناشی از ریسک پارامترها با برآورد نادرست، به شرح زیر محاسبه می‌شود:

$$\left(\frac{\text{چندک ۹۹/۵ تابع P}(\theta)}{\text{میانگین P}(\theta)} - ۱\right) \times ۱۰۰. \qquad (۳۵)$$

برای یافتن برآورد صدک و میانگین $P(\theta)$ ، ۱۰۰۰۰ مقدار شبیه‌سازی شده برای θ محاسبه می‌شود. مقادیر شبیه‌سازی شده را با θ' نمایش داده و kامین مقدار شبیه‌سازی شده برای پرتفوی ($P(\theta')$) به‌صورت زیر محاسبه می‌شود:

$$\mathrm{P}(\theta'_k) = \sum_{i=۱}^{n} w_i \overline{a}_{x_i}(\theta'_k). \qquad (۳۶)$$

که در این رابطه، w_i نشان‌دهنده مبلغ مستمری شخص iام است. بنابراین ریسک برآورد نادرست الزام سرمایه را به‌عنوان مثال برای طرح بازنشستگی P_1 به سه روش محاسبه می‌کنیم:

- ✓ پارامتر تابع مرگ‌ومیر τ فقط با استفاده از داده‌های موجود در P_1 برآورد می‌شود. به‌منظور ساده نگه داشتن کارها و جلوگیری از مسئلهٔ شناسایی‌پذیری به‌دلیل نبود C، تابع مرگ‌ومیر مانند مدل M_1 است که پارامترهای آن‌را می‌توان با حداکثر درستنمایی تخمین زد.
- ✓ تمام پارامترهای تابع مرگ‌ومیر با استفاده از ترکیب طرح‌های $P_1 - P_3$، از طریق رویکرد داده‌های گمشده، برآورد می‌شوند.
- ✓ تمام پارامترهای تابع مرگ‌ومیر با استفاده از ترکیب طرح‌های $P_1 - P_4$ برآورد می‌شوند.

1. Richards
2. The Solvency Capital Requirement

این کار را برای برای طرح بازنشستگی P_3 تکرار می‌کنیم که در این طرح هر دو متغیر کمکی موجود است. الزامات سرمایه برای ریسک برآورد نادرست براساس این دو رویکرد و دو نرخ بهره، در جدول ۸ نشان داده شده‌است.

جدول ۸. الزام سرمایه برای طرح بازنشستگی P_1 و P_3 بر اساس دو نرخ بهرهٔ فرضی و طرح‌های مورد استفاده برای برآورد نرخ مرگ‌ومیر

نرخ بهره و الزام سرمایه	نمونه‌ها					
طرح بازنشستگی P_1	$P_1 - P_4$		$P_1 - P_3$		P_1	
نرخ بهره	٪۱	٪۳	٪۱	٪۳	٪۱	٪۳
الزام سرمایه	٪۳۰/۱	٪۰۹/۱	٪۹۵/۱	٪۶۲/۱	٪۸۲/۲	٪۲۷/۲
طرح بازنشستگی P_3	$P_1 - P_4$		$P_1 - P_3$		P_3	
نرخ بهره	٪۱	٪۳	٪۱	٪۳	٪۱	٪۳
الزام سرمایه	٪۲۶/۱	٪۰۳/۱	٪۸۸/۱	٪۵۴/۱	٪۷۹/۵	٪۶۱/۴

منبع: یافته‌های تحقیق

از نتایج این جدول مشاهده می‌شود که در نظر گرفتن سایر مجموعه داده‌ها برای برآورد نرخ مرگ‌ومیر، می‌تواند الزام سرمایه را کاهش دهد. درنظر گرفتن مشاهدات برای وقتی که در آن‌ها بعضی از متغیرهای کمکی یا تمامی آن‌ها گمشده هستند، اندازه نمونه را افزایش داده و سبب کاهش عدم حتمیت پارامتر می‌شود. این نتایج نشان می‌دهند که در نظر گرفتن نمونه‌هایی با مشاهدات گمشده، می‌تواند تغییرپذیری برآورد نمونه‌گیری را کاهش دهد و به نوبه خود سبب کاهش الزام سرمایه شود. و یکی از راه‌های به‌دست آوردن چنین مشاهدات اضافی، ترکیب داده‌های تجربی از طرح‌های مختلف بازنشستگی است با این فرض که اعضای آن طرح‌ها از قانون مرگ‌ومیر یکسانی پیروی کنند.

۵. جمع‌بندی و پیشنهادها

در این پژوهش به ترکیب داده‌های حاصل از طرح‌های بازنشستگی مختلف، بافرض اینکه قانون احتمال یکسانی برای طول‌عمر آتی دارند، اما اطلاعات متغیر کمکی متفاوت است، پرداخته شد. اگر متغیرهای کمکی حتی در یک تجربه مرگ‌ومیر گمشده باشد، می‌توان از همین روش‌ها استفاده کرد.

همچنین به‌دست آوردن داده‌های کامل برای زیرمجموعهٔ کوچکی از اعضا به ما این امکان را می‌دهد داده‌های حاصل از دو یا چند تجربه را با هم تلفیق و از موانع شناسایی‌پذیری جلوگیری کنیم. این نتیجه از لحاظ عملی اهمیت دارد، زیرا به‌دست آوردن اطلاعات کامل برای همه اعضای صندوق‌های بازنشستگی مختلف ممکن است دشوار یا گران باشد.

روش‌های مطرح‌شده می‌تواند برای محاسبه کمیت‌های مالی مورد علاقه براساس عامل‌های مستمری، برای بیم‌سنج‌ها مفید باشد. این روش مجموعه داده‌های مختلف با تجربه مرگ‌ومیر برابر یا مشابه را با هم ترکیب، اندازه نمونه را افزایش و ریسک پارامتر را کاهش می‌دهد، بنابراین منجر به کاهش الزام سرمایه می‌شوند. متغیرهای اقتصادی–اجتماعی از جمله سطح مزایا و مشخصات جغرافیایی جمعیتی در صورت پایین بودن نرخ بهره بیشتر مورد توجه قرار می‌گیرند.

بررسی این‌که آیا می‌توان بدون کنار گذاشتن داده‌های مربوطه در وضعیت پارامتر افزونه، استنباط آماری کرد، مساله مورد بررسی قرار نگرفته‌است و این مسئله می‌تواند در چارچوب استنباط بیزی بهتر حل شود، زیرا هنگامی‌که ماتریس اطلاع فیشر معین مثبت اکید نیست، برآوردگر ماکسیمم درستنمایی، توزیع نرمال مجانبی ندارد (وانتانابه[1]، ۲۰۱۰). همچنین یکی از مفروضات مدل‌بندی این بود که اختلافات مرگ‌ومیر بین گروه‌های اقتصادی–اجتماعی با افزایش سن تغییر نمی‌کند. لذا تأثیر و تجزیه و تحلیل یک عبارت اثر متقابل می‌تواند در بررسی‌های آتی مدنظر قرار گیرد.

1. Watanabe

ملاحظات اخلاقی

حامی مالی

این مقاله حامی مالی ندارد.

مشارکت نویسندگان

تمام نویسندگان در آماده سازی این مقاله مشارکت کرده‌اند.

تعارض منافع

بنا به اظهار نویسندگان، در این مقاله هیچ‌گونه تعارض منافعی وجود ندارد.

تعهد کپی‌رایت

منابع

ذکایی، محمد و مقصودی، مسطوره. (۱۳۸۹). بازسازی مدل‌های مرگ‌ومیر بر پایه شکنندگی با استفاده از تعمیم توزیع گومپرتز. *فصلنامه صنعت بیمه*، ۲۵(۴): ۸۵-۵۹.

شجاعی‌آذر، زهرا و حسن‌زاده، امین. (۱۳۹۳). کاربرد مدل‌های فاز-نوع در مدل‌بندی مرگ‌ومیر. *پژوهشنامه بیمه*، ۲۹(۱): ۱۰۵-۱۲۶.

کمیجانی، اکبر.، کوششی، مجید و نیاکان، لیلی. (۱۳۹۲). برآورد و پیش‌بینی نرخ مرگ‌ومیر در ایران با استفاده از مدل لی-کارتر. *پژوهشنامه بیمه*، ۲۸(۴): ۲۵-۱.

مهدوی، غدیر.، دقیقی اصل، علیرضا و لطفی، نیر. (۱۳۹۰). کاربرد یک مدل مرگ‌ومیر با چند عامل ریسک در فسخ قراردادهای بیمه عمر (مورد مطالعه: یک شرکت بیمه). *پژوهشنامه بیمه*، ۲۶(۳): ۲۸-۱.

Catchpole, E. A. & Morgan, B. J. T. (1997). Detecting parameter redundancy. *Biometrika*, 84(1): 187–196.

Chen, Q., May, R. C., Ibrahim, J. G., Chu, H. & Cole, S. R. (2014). Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Statistics in Medicine*, 33(26): 4560–4576.

Dempster, A. P., Laird, N. M. & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1): 1–38.

Dickson, D., Hardy, M. & Waters, H. (2013). Actuarial mathematics for life contingent risks. international series on actuarial science. Cambridge University Press.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the

value of life contingencies. *Philosophical Transactions of the Royal Society of London*, 115: 513–583.

Herring, A. H. & Ibrahim, J. G. (2001). Likelihood-Based methods for missing covariates in the cox proportional hazards model. *Journal of the American Statistical Association*, 96(453): 292–302.

Lin, X. S. & Liu, X. (2007). Markov aging process and Phase-Type law of mortality. *North American Actuarial Journal*. 11(4): 92–109.

Little, R. & An, H. (2004). Robust Likelihood-Based analysis of multivariate data with missing values. *Statistica Sinica*, 14(3): 949–968.

Lord, F. M. (1955). Estimation of parameters from incomplete data. *Journal of the American Statistical Association*, 50(271): 870–876.

Madrigal, A. M., Matthews, F. E., Patel, D., Gaches, A. & Baxter, S. (2011). What longevity predictors should be allowed for when valuing pension scheme liabilities? *British Actuarial Journal*, 16(1): 1–38.

Macdonald, A. S., Richards, S. J. & Currie, I. D. (2018). Modelling mortality with actuarial applications. International Series on Actuarial Science. Cambridge University Press.

McLachlan, G. & Peel, D. (2000). Finite mixture models. Wiley Series in Probability and Statistics, New York.

Richards, S. J. (2016). Mis-Estimation risk: Measurement and impact. *British Actuarial Journal*, 21(3): 429–457.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3): 581–592.

Schluchter, M. D. & Jackson, K. L. (1989). Log-Linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, 84(405): 42–52.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985). Statistical analysis of finite mixture distributions. New York, Wiley.

Tsiatis, A. (2007). Semiparametric theory and missing data. Springer Science & Business Media.

Ungolo, F., Christiansen, M. C., Kleinow, T. & MacDonald, A. S. (2019). Survival analysis of pension scheme mortality when data are missing. *Scandinavian Actuarial Journal*, 2019 (6): 523–547.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(116): 3571–3594.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3): 163–195.

Xu, Y., Kim, J. K. & Li, Y. (2017). Semiparametric estimation for measurement error models with validation data. *Canadian Journal of Statistics*, 45(2): 185–201.

Yashin, A. (2001). Mortality models incorporating theoretical concepts of ageing. *In Forecasting Mortality in Developed Countries*, 261–280.