Research Article

The Iranian Journal of Health Insurance | Volume 5, Issue 2, Summer 2022 | Pages: 96-107

journal.ihio.gov.ir

A Credibility Approach to Predicting Healthcare Costs: A Case Study of Dental Costs

Ehsan Jahanbani[1] MSc, **Amirteymor Payandeh Najafabadi**[1*] PhD, **Khaled Masoumifard**[1] PhD

[1] Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran

Correspondence to: Amirteymor Payandeh Najafabadi, Email: amirtpayandeh@gmail.com

Received: March 16, 2022 **Revised:** June 25, 2022 **Accepted:** July 24, 2022 **Online Published:** August 16, 2022

Abstract

Introduction: The increase in the cost of medical services and the aging of Iran's population have created problems for the medical insurance industry, which has turned this field into one of the riskiest fields. In this situation, one of the most basic needs of this field is to fit a suitable model to the treatment losses, so that the insurer can check the trend of the treatment costs, have an accurate estimate of the next year's losses, and calculate the corresponding insurance premium.

Methods: Based on an empirical fundamental study, the loss data related to group treatment insurance policies issued from the middle of 2017 to the middle of 2019 was collected by Day insurance company and using mixed distributions, dental claims were modeled and the loss of the next year was predicted using the credibility method.

Results: Due to the heterogeneity in the data, mixed distribution will be more suitable due to its high flexibility. The insured people were divided into 2 high-risk and low-risk categories according to their characteristics, and a distribution was given to each of these categories. The general distribution of damages is considered as a mixed distribution of these distributions, and the mixed weights of this distribution were estimated using logistic regression.

Conclusion: One of the most basic issues in all types of insurances is determining their insurance premiums. In this research, using the theory of credibility, the next year's insurance premium is calculated for each of the insured according to their risk parameter. In fact, credibility theory combines the existing insurance premium with past losses and provides the adjusted premium.

Keywords: Finite Mixed Distribution, Complementary Treatment, Logistic Regression, Dentistry

Highlights

1. By using credibility, the existing insurance premium can be combined with the past damages and the adjusted insurance premium can be considered.
2. In case of a more accurate estimate of the treatment costs, a fair insurance premium can be calculated for each insured according to its risk parameters, and in addition to increasing the attractiveness for the policyholder, the insurance company will suffer less loss.

Citation:
Jahanbani E, Payandeh Najafabadi A, Masoumifard K. A credibility approach to predicting healthcare costs: a case study of dental costs. Iran J Health Insur. 2022;5(2):96-107.

مقاله پژوهشی

فصلنامه علمی- پژوهشی بیمه سلامت ایران | دوره ۵، شماره ۲، تابستان ۱۴۰۱ | صفحات ۹۶ تا ۱۰۷

journal.ihio.gov.ir

رویکرد باورمندی به پیشگویی هزینه‌های درمان: مطالعه موردی بررسی هزینه‌های دندانپزشکی

احسان جهان‌بانی[1] MSc، امیرتیمور پاینده نجف‌آبادی[1*] PhD، خالد معصومی‌فرد[1] PhD

[1] گروه بیمسنجی، دانشکده علوم ریاضی، دانشگاه شهید بهشتی، تهران، ایران

[*] **نویسنده مسئول:** امیرتیمور پاینده نجف‌آبادی، پست الکترونیک: amirtpayandeh@gmail.com

دریافت: ۱۴۰۰/۱۲/۲۵ تصحیح: ۱۴۰۱/۰۴/۰۴ پذیرش: ۱۴۰۱/۰۵/۰۲ انتشار آنلاین: ۱۴۰۱/۰۵/۲۵

چکیده

مقدمه: افزایش هزینه‌های خدمات پزشکی و سالمند شدن جمعیت ایران، مشکلاتی را برای صنعت بیمه درمان به وجود آورده که این حوزه را به یکی از پرمخاطره‌ترین حوزه‌ها تبدیل کرده است. در این شرایط یکی از اساسی‌ترین نیازهای این حوزه، برازش مدلی مناسب به خسارت‌های درمان است، به‌طوری‌که بیمه‌گر بتواند روند هزینه‌های درمان را بررسی کند، برآورد دقیقی از خسارت‌های سال آینده داشته باشد و حق بیمه متناسب با آن را محاسبه کند.

روش بررسی: با مطالعه بنیادی تجربی، داده‌های خسارت مربوط به بیمه‌نامه‌های درمان گروهی صادره از اواسط سال ۱۳۹۷ تا اواسط سال ۱۳۹۹ شرک بیمه دی جمع‌آوری شد و با استفاده از توزیع‌های آمیخته، خسارت‌های دندانپزشکی مدل‌بندی شد و با استفاده از روش باورمندی خسارت سال آینده پیش‌بینی شد.

یافته‌ها: باتوجه به ناهمگنی موجود در داده‌ها، توزیع‌های آمیخته به دلیل انعطاف‌پذیری بالا برازش مناسب‌تری خواهد داشت. افراد بیمه شده باتوجه به مشخصات آنها در ۲ دسته پرریسک و کم‌ریسک تقسیم شدند و به هر یک از این دسته‌ها یک توزیع برازش داده شده است. توزیع کلی خسارت‌ها را یک توزیع آمیخته از این توزیع‌ها در نظر گرفته که وزن‌های آمیختگی این توزیع با استفاده از رگرسیون لجستیک برآورد شد.

نتیجه‌گیری: یکی از اساسی‌ترین مسائل در انواع بیمه‌ها، تعیین حق بیمه آنهاست. در این پژوهش با استفاده از نظریه باورمندی، برای هر یک از بیمه شده‌ها باتوجه به پارامتر ریسک آنها حق بیمه سال آینده محاسبه می‌شود. درواقع باورمندی، حق بیمه موجود را با خسارت‌های گذشته ترکیب کرده و حق بیمه تعدیل شده را ارائه می‌دهد.

واژگان کلیدی: توزیع آمیخته متناهی، درمان تکمیلی، رگرسیون لجستیک، دندانپزشکی

نکات ویژه

۱. با استفاده از باورمندی، می‌توان حق بیمه موجود را با خسارت‌های گذشته ترکیب و حق بیمه تعدیل شده را درنظر گرفت.

۲. درصورت برآورد دقیق‌تر از هزینه‌های درمان، می‌توان حق بیمه منصفانه‌ای برای هر بیمه شده باتوجه به پارامترهای ریسک آن محاسبه و علاوه‌بر افزایش جذابیت برای بیمه‌گذار، شرکت بیمه نیز متحمل زیان کمتری خواهد شد.

مقدمه

طی پنجاه سال اخیر توسعه اقتصادی اجتماعی، کاهش باروری و به دنبال آن کاهش رشد جمعیت و افزایش امید به زندگی، منجر به ایجاد تغییرات هنگفتی در ساختار جمعیت جهان شده است، به‌نحوی که طی این مدت تعداد سالمندان به‌طور کلانی افزایش یافته است. این امر هزینه‌های زیادی را، به‌خصوص در بخش بهداشت و درمان، به همراه داشته است. افزایش سالمندان، واقعیت انکارناپذیر جوامع کنونی است. این پدیده هشداری به جوامع برای توجه بیشتر به مسائل این جمعیت رو به رشد است. مسائلی چون نبود حمایت‌های اجتماعی، نداشتن شغل و نقش اجتماعی، مخارج زندگی و به‌ویژه هزینه‌های بهداشتی درمانی سرسام‌آور و غیره ازجمله مواردی است که لزوم توجه بیشتر به این قشر از جامعه

را نمایان می‌سازد. در سال ۲۰۲۰ تخمین زده شد یک میلیارد نفر (۱۳درصد) از جمعیت جهان، افراد سالخورده ۶۵ سال و بالاتر هستند. برآورد شده که تا سال ۲۰۳۰، از هر ۶ نفر در جهان، یک نفر ۶۰ ساله یا بیشتر خواهد بود. انتظار می‌رود این رقم در سال ۲۰۵۰ تقریباً سه برابر شده و به حدود یک و نیم میلیارد نفر (۱۶ درصد از جمعیت جهان) برسد. اگرچه در حال حاضر توسعه یافته‌ترین کشورها، سالخورده‌ترین نمودار جمعیت را دارند، اما سریع‌ترین سالخوردگی جمعیت در کشورهای کمتر توسعه یافته اتفاق می‌افتد. این پدیده با کاهش نرخ باروری و عمر طولانی شکل‌گرفته است. با کاهش تولد و افزایش طول عمر، سهم افراد سالخورده از کل جمعیت در حال افزایش است [۱].

جمعیت ایران نیز در حال حاضر در مرحله انتقال ساختار سنی از جوانی به سالخوردگی است. هم‌اکنون درصد کمتری از جمعیت کشور در سن سالخوردگی هستند، اما باتوجه به کاهش سریع باروری در چند دهه اخیر افزایش تعداد و درصد سالمندان در سال‌های آینده، برنامه‌ریزی آینده‌محور برای حل مشکلات این قشر از جمعیت ضروری است [۲]. موضوع درمان در تمام کشورها، به‌خصوص کشورهای در حال توسعه، یکی از مسائل مهم است و دولت‌ها می‌کوشند شیوه‌های خدمات‌رسانی و ایجاد امکانات لازم در زمینه‌های درمانی را بهبود بخشند [۳]. یکی از راه‌های مناسب برای فراهم ساختن پشتوانه بسیار مطلوب، به‌منظور حمایت از وضعیت اقتصادی خانواده و گذر از موقعیت‌های اضطراری ناشی از بیماری‌ها و حوادث، تهیه بیمه‌نامه درمان است [۴]. در حقیقت، بیمه سازوکاری است که به افراد ریسک‌پذیر اجازه می‌دهد از ریسک و فقدان اطمینانی که با آن مواجه هستند، رها شوند. مردم با پرداخت حق بیمه تعیین شده، زیان و مخارج ناشی از بیماری را کاهش داده و به این ترتیب رفاه خود را افزایش می‌دهند [۳ و ۴]. سالمند شدن جمعیت ایران و افزایش هزینه‌های بازنشستگی، تقاضا برای خدمات درمانی را بسیار افزایش داده است. واضح است که تمام این موارد، متضمن هزینه برای صنعت بیمه کشور خواهد بود. بنابراین مدل‌بندی هزینه‌های درمان، نیاز اصلی صنعت بیمه کشور بوده و بیمه‌گر اجتماعی باید با رویکرد مالی و بر اساس خصوصیات هر جامعه، روند هزینه‌های درمان را به‌دقت بررسی کرده و بتواند مدل مناسبی را برای پیش‌بینی هزینه‌های درمان در آینده ارائه دهد[۵ و ۶].

از مهم‌ترین وظایف شرکت‌های بیمه، پیش‌بینی میزان و تعداد خسارت‌ها در سال آتی است. روش‌های مختلفی مانند سری‌های زمانی، روش بیزی و نظریه باورمندی برای انجام چنین پیش‌بینی‌های وجود دارد. نظریه باورمندی روشی غیرپارامتریک است درحالی‌که روش بیزی و سری‌های زمانی پارامتریک هستند. همه این روش‌ها برای پیش‌بینی تعداد خسارت‌ها در سال آتی قابل اطمینان هستند اما باتوجه به آسان بودن اجرای نظریه باورمندی، استفاده از این روش پیشنهاد می‌شود. باورمندی به‌عنوان ابزاری عملی، عمدتاً حاصل کار اکچوئرهای آمریکای شمالی است، با این وجود ریشه در نظریه ریسک و برآوردهای آماری دارد . نظریه باورمندی تلاش در راستای ایجاد مدلی است که توسط آن بتوان درباره محیطی که در آن دستگاه‌های بیمه عمل می‌کنند، اطلاع حاصل کرد. نظریه باورمندی سابقه طولانی در علم بیم‌سنجی دارد که اساس آن به موبرای (۱۹۱۴) برمی‌گردد [۷]. ویتنی (۱۹۱۸) [۸] شکل جالبی از میانگین وزنی متوسط خسارت‌های وارده بر یک دسته و سایر دسته‌ها را به‌منظور پیش‌بینی خسارت‌های وارده در سال‌های آتی ارائه کرد. کِفِر (۱۹۲۹) [۹] برای اولین بار استفاده از روش بیزی برای نرخ‌گذاری در بیمه‌های عمر گروهی را پیشنهاد داد. به دنبال آن بیلی (۱۹۵۰) [۱۰] رابطه نظریه باورمندی خطی و روش بیزی را نشان داد. در مسائل پیشرفته نظریه باورمندی، برای برآورد حق بیمه، روش‌های بیزی به‌کار گرفته می‌شود که به روش‌های باورمندی بیزی موسوم هستند. روش‌های بیزی با ارائه بولمن و بولمن-استراب (۱۹۶۰) [۱۱] وارد علم بیم‌سنجی شد. میلر و هیکمن در سال ۱۹۵۷ از نظریه باورمندی در توزیع تجمعی خسارت‌ها استفاده کردند، به‌علاوه پین کویت (۱۹۹۷) استفاده از نظریه باورمندی را در بیمه اتومبیل مورد مطالعه قرار داد [۱۲]. چون باورمندی به الگوهایی می‌پردازد که پاسخ‌هایی مطابق با محیط و پیرامون خود ارائه می‌دهد، بیشتر در علم بیم‌سنجی نفوذ کرده است. نظریه باورمندی تلاش در راستای ایجاد مدلی است که توسط آن بتوان درباره محیطی که در آن دستگاه‌های بیمه عمل می‌کنند، اطلاع حاصل کرد. نظریه باورمندی مجموعه‌ای از ابزار کمّی است که به بیمه‌گر اجازه می‌دهد حوادث آتی را برای یک ریسک یا گروهی از ریسک‌ها نرخ‌بندی کند، به این معنا که حق بیمه‌های آتی را بر اساس حوادث گذشته تعدیل می‌کند.

توزیع‌های آمیخته

امروزه تحلیلگران داده، به لطف روش‌های محاسباتی قوی و همچنین طیف گسترده‌تری از توزیعهای مدل‌سازی، قادر به توصیف، تخمین، پیش‌بینی و استنباط در مورد دستگاه‌های پیچیده موردعلاقه هستند. مدل‌های آمیخته تصویری جذاب از این جنبه‌ها فراهم می‌کند. اگرچه مدل‌های آمیخته درون یک خانواده از توزیع‌های پارامتری تقریب‌های سازگاری ارائه می‌کند، اما دارای چالش‌های محاسباتی بسیار پیچیده‌ای است. توزیع‌های آمیخته شامل یک تعداد متناهی یا نامتناهی از عضوهاست و ممکن است نوع توزیع‌ها از همدیگر متفاوت باشند که این خاصیت ویژگی‌های متفاوت داده‌ها را توصیف می‌کند. بنابراین، آنها توصیفهای بسیار دقیق‌تری از دستگاه‌های پیچیده مانند نجوم، بوم‌شناسی، بیوانفورماتیک، علوم رایانه، اقتصاد، مهندسی، روباتیک و زیست‌شناسی ارائه می‌کنند. توزیع‌های آمیخته اهمیت زیادی در مدل‌بندی پدیده‌های تصادفی دارد. در دهه گذشته، مدل‌های آمیخته متناهی بسیار مورد توجه قرار گرفته است، زیرا انعطاف‌پذیری مدل‌های آمیخته بسیار بالاست و راه‌های ساده‌ای برای مدل‌بندی داده‌های که شکل توزیع آنها نامشخص است ارائه می‌کند. باتوجه به گفته باکس (۱۹۷۶) [۱۳] که «همه مدل‌ها اشتباه هستند، اما برخی مفید هستند»، مدلهای آمیخته نیز بسیار مفید هستند، درصورتی‌که عمدتاً درست نیستند، اما می‌توانند به‌عنوان جایگزینهای مفیدی برای برآورد چگالی، خوشه‌بندی، ناهمگنی غیرقابل مشاهده و غیره باشند. مدل‌سازی آمیخته را می‌توان برای طیف گسترده‌ای از داده‌ها مانند: تک‌متغیره یا چندمتغیره، پیوسته یا طبقه‌ای، مقطعی، سری زمانی و شبکه‌ها استفاده کرد. معمولاً در مواردی که جامعه آماری ناهمگن و ترکیبی از چند زیرجامعه باشد توزیع‌های آمیخته بسیار کاربرد دارد. مدل‌های آمیخته بیش از ۱۵۰ سال است که در بسیاری از شاخه‌های مدل‌سازی آماری به‌عنوان ابزاری همه‌کاره و چندوجهی برای توانمند ساختن مجموعه توزیع‌های احتمال برای مدل‌بندی داده‌ها مورد استفاده قرار گرفته است. از نظر تاریخی، مفهوم توزیع‌های آمیخته برای اولین بار توسط نیوکام (۱۸۸۶) به‌عنوان الگویی برای محیط‌های خارجی یافت شد و بعداً توسط کارل پیرسون و والتر ولدون مورد مطالعه قرار گرفت. توزیع‌های آمیخته به دلیل انعطاف‌پذیر بودن در توصیف بسیاری از پدیده‌های تصادفی مفید هستند. کاربرد توزیع‌های آمیخته معمولاً در مواردی است که جامعه آماری ناهمگن و ترکیبی از چند زیرجامعه باشد. در چنین حالتی مشاهداتی که از این جامعه به دست می‌آیند، می‌توانند با احتمال معینی به هر یک از این زیرجامعه‌ها تعلق داشته باشند. برای مدل‌بندی داده‌های بیمه‌ای، باتوجه به پارامتر ریسک آنها که در کلاس‌های مختلفی قرار می‌گیرند، استفاده از توزیع‌های پارامتری شناخته شده کار مشکل و پیچیده‌ای است زیرا با یک جامعه یکسان و همگن سروکار نداریم و پارامتر ریسک برای زیرجامعه‌های مختلف متفاوت است. به همین دلیل جامعه را خوشه‌بندی کرده و برای هریک از این خوشه‌ها می‌توان توزیع‌های متفاوتی در نظر گرفت. در حوزه بیمه توزیع‌های آماری متفاوتی برای بررسی زیان مورد استفاده قرار می‌گیرد که یکی از آنها توزیع‌های آمیخته است. توزیع‌های آمیخته به دلیل انعطاف‌پذیری بالای آنها در مدل‌بندی داده‌ها در حوزه بیمه مورد توجه بیمه‌سنجها قرار گرفته و به‌طور گسترده‌ای برای مدل‌بندی داده‌های به‌کار می‌رود که مشاهدات از گروه‌های مختلف آمده باشد.

تعریف: فرض کنید که F خانواده از توابع توزیع باشد. می‌گوییم Y یک متغیر تصادفی با یک توزیع آمیخته متناهی است. اگر تابع توزیع آن به فرم زیر باشد:

$$F(y) = \sum_{i=1}^{k} w_j \, F_j(y|\theta_j) \qquad (1)$$

که $F_j \epsilon F$ تابع توزیع از k جامعه قابل‌تشخیص است و $\boldsymbol{w} = (w_1, w_2, \ldots, w_k)$، $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ و $0 \leq w_j \leq 1$ وزن‌های آمیختگی به‌طوری که $\sum_{j=1}^{k} w_j = 1$.

رگرسیون لجستیک

رگرسیون یک رابطه آماری بین دو یا بیش از دو متغیر است که هر تغییری در مقدار متغیر مستقل این معادله، متغیر وابسته را تحت تأثیر قرار می‌دهد. اما وقتی‌که متغیر پاسخ (وابسته) کمّی نباشد و یک متغیر گسسته باشد مثلاً بله یا خیر، درست یا غلط، رأی‌داده یا نداده، کالای خریداری شده یا نه؟ از رگرسیون لجستیک استفاده می‌شود. یکی از روش‌های دسته‌بندی در مبحث یادگیری ماشین رگرسیون لجستیک است. در این روش رگرسیونی، از مفهوم و شیوه محاسبه نسبت بخت استفاده می‌شود. فرض کنید فرم مدل

رگرسیون خطی ساده به‌صورت $Z=\beta_0+\beta_1 X+\epsilon$ باشد همان‌طور که دیده می‌شود این رابطه، معادله یک خط است که جمله خطا یا همان ϵ به آن اضافه شده است. در این حالت اگر(Z) مقدار برآورد برای متغیر وابسته باشد، می‌توان آن را میانگین مشاهدات برای متغیر وابسته به ازای مقدار ثابت متغیر مستقل در نظر گرفت. اگر مقدار متغیر وابسته (Z)، باینری (دو وضعیتی) و شامل صفر و ۱ باشد مشخص است که دارای توزیع برنولی است و امید ریاضی آن به‌صورت $\hat{z}=E(Z|X=x)=P(Z=1|X=x)=p(x)$ محاسبه می‌شود. برای مشخص کردن مدل رابطه بین متغیر وابسته و مستقل به‌جای رابطه خطی، به تابعی احتیاج داریم که در حدود ۰ تا ۱ تغییر کند:

$$p(x)=\hat{z}=E(Z=1|X=x)=\frac{e^{b_0+b_1x_1}}{1+e^{b_0+b_1x_1}} \quad (۲)$$

به‌منظور برآورد پارامترهای این مدل، می‌توان از تبدیل لوجیت استفاده کرد.

$$g(x)=\ln\left(\frac{p(x)}{1-p(x)}\right)=\ln\left(e^{b_0+b_1x_1}\right)=b_0+b_1x_1 \quad (۳)$$

با استفاده از تابع درست‌نمایی و حداکثرسازی آن می‌توان پارامترها را برآورد کرد. با این کار به یک دستگاه معادلات می‌رسیم که متأسفانه برای حل آن روش تحلیلی وجود ندارد و باید به کمک روش‌های عددی برآورد را انجام داد. در این پژوهش هدف این است که بیمه شده‌ها را به دو خوشه پرریسک و کم‌ریسک تقسیم کرده و به هر یک از این خوشه‌ها یک توزیع مناسب برازش داده شود، یکی از روش‌های خوشه‌بندی K-means است. مسئله مهمی که در این رابطه به وجود می‌آید، نرمال‌سازی داده‌ها در خوشه‌بندی است زیرا باید ویژگی‌ها در محاسبه فاصله بدون مقیاس باشند تا بزرگی واحد اندازه‌گیری هر بُعد باعث اریبی مقدار تابع فاصله به سمت یک ویژگی نشود. برای اندازه‌گیری شباهت یا فاصله بین مشاهدات از فاصله اقلیدسی یا منهتن استفاده کرد که در زیر نحوه محاسبه بین دونقطه p بعدی x و y ارائه شده است:

$$x=(x_1,\ldots,x_p) \qquad y=(y_1,\ldots,y_p)$$

$$D_{Euc}(\mathbf{x},\mathbf{y})=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2} \quad (۴)$$

$$D_{Man}(\mathbf{x},\mathbf{y})=\sum_{i=1}^{p}|x_i-y_i| \quad (۵)$$

نکته حائز اهمیت در اینجا این است که داده‌ها را با بایستی بر اساس مقیاسی یکسان در معیارهای اندازه‌گیری فاصله به‌کار برد (داده‌ها را نرمال‌سازی کرد). محاسبه فاصله باید براساس داده‌های بدون مقیاس صورت بگیرد تا ویژگی یا متغیرهای خاص با واحد اندازه‌گیری بزرگ، باعث اریبی در مقدار فاصله بین نقطه‌ها نشود، در غیر این صورت ممکن است نتایج خوشه‌بندی بستگی زیادی به متغیری داشته باشد که دارای مقدارهای بزرگ‌تر و درنتیجه واریانس بزرگ‌تری خواهد داشت.

روش بررسی

برای مدل‌بندی داده‌های بیمه‌ای، باتوجه به پارامتر ریسک بیمه شده که در کلاس‌های مختلفی قرار می‌گیرند، استفاده از توزیع‌های پارامتری تک عضوی مناسب نیست زیرا پارامتر ریسک برای کلاس‌های مختلف متفاوت است و با جامعه همگن و یکسانی سروکار نداریم. توزیع‌های آمیخته به دلیل انعطاف‌پذیری بالای آنها، در توصیف بسیاری از پدیده‌های تصادفی کاربرد دارند. معمولاً درمواردی که جامعه ناهمگن است برای مدل‌بندی داده‌ها، توزیع‌های آمیخته بسیار مفید هستند. در این پژوهش، باتوجه به ناهمگنی موجود در خسارت‌های دندانپزشکی که به دلیل عواملی چون سن، جنسیت، شغل و غیره است، نمی‌توان با توزیع‌های پارامتری تک‌عضوی مدل‌بندی کرد، به همین خاطر با استفاده از توزیع‌های آمیخته متناهی داده‌ها را مدل‌بندی کرده و باتوجه به مدل برازش داده شده، حق بیمه باورمندی برای سال آینده محاسبه می‌شود.

باورمندی

یکی از اساسی‌ترین مسائل در انواع بیمه‌ها، تعیین حق بیمه آنهاست. نظریه باورمندی در بیمه، به‌عنوان ابزاری برای تعیین و تعدیل حق بیمه به‌کار می‌رود. درواقع باورمندی حق بیمه موجود را با خسارت‌های گذشته ترکیب کرده و حق بیمه تعدیل شده را ارائه می‌دهد. اما در مدل‌های بیمه هرچقدر حجم نمونه بیشتر باشد، باورمندی افزایش می‌یابد. چون باورمندی به الگوهایی می‌پردازد که پاسخ‌هایی مطابق با محیط و پیرامون خود ارائه می‌دهد، بیشتر در علوم اکچوئری نفوذ کرده است. نظریه باورمندی مجموعه‌ای از ابزار کمّی است که به بیمه‌گر اجازه می‌دهد حوادث آتی را برای یک ریسک یا گروهی از ریسک‌ها نرخ‌بندی کند. به این معنا که

حق بیمه‌های آتی را بر اســاس حوادث گذشته تعدیل می‌کند. اگر حوادث یک بیمه‌گذار یا گروهی از بیمه‌گذاران در دســترس باشــد، دانش آماری ما ممکن اســت ما را به استفاده از میانگین نمونه‌ای یا برخی برآوردگرهای ناریب دیگر متقاعد کند. اما نظریه باورمندی به ما می‌گوید بهینه آن است که تنها قسمتی از وزن را به این حوادث بدهیم و بقیه وزن را به برآوردگرهایی که از اطلاعات دیگری به دست می‌آیند داده شود. فرض کنید که $Y_1,Y_2,\ldots,Y_n$ شدت خسارت‌های یک فرد بیمه شده در n سال گذشته باشد. شرکت بیمه معمولاً یک نرخ بیمه‌ای μ برای کل داشــتمان برآورد کرده است. باورمندی یک دستورالعمل برای ترکیب کردن داده‌های گذشته و μ است. بر اساس دیدگاه باورمندی، یک برآوردگر سازگار برای حق‌بیمه به‌صورت زیر قابل محاسبه است:

$$\pi_c=\xi\bar{Y}+(1-\xi)\mu \qquad (۶)$$

که $0<\xi<1$ عامل باورمندی، $\bar{Y}$ میانگین مشــاهدات گذشته و μ میانگین توزیع پیشین است. از دیدگاه بیمه‌گر، اگر گذشته بیمه‌گذار پایدار باشد، یعنی واریانس خسارت‌ها کم باشد، به نظر می‌رسد که از $\bar{Y}$ برای پیش‌بینی نتایج دوره بعد اســتفاده شــود. همچنین اگر گذشــته افراد دارای تغییرپذیری بیشتری باشد، بهتر است که برای پیش‌بینی نتایج دوره بعد، کمتر از $\bar{Y}$ اســتفاده شود و وزن میانگین توزیع پیشین μ بیشتر باشد. فرض کنید که شرکت بیمه‌ای باتوجه به مشــاهدات مربوط به خسارت‌های n سال گذشته یک بیمه شده با پارامتر ریســک Θ می‌خواهد حق بیمه انفرادی را برای سال n+1 آم به دست آورد، در این نوع حق بیمه باتوجه به پارامتر ریسک هر فرد Θ، برای هر فرد یک حق بیمه خاص به دست می‌آید. حق بیمه انفرادی (خالص) با پارامتر ریسک Θ برابر با $\mu(\theta=E[Y_{n+1}\mid\Theta])$ است. این نوع حق بیمــه را حق بیمه عادلانه نیز می‌نامند و بهترین حق بیمه ممکن اســت، اما مشکل آنجاســت که θ و به دنبال آن $\mu(\theta)$ نامعلوم اســت. اگر $\mu(\theta)$ یک برآوردگر دلخواه برای $\mu(\theta)$ و $E[\mu(\theta)]=E[Y]$ برآوردگــر بیــز برای $\mu(\theta)$(حق‌بیمه انفرادی) باشــد می‌توان نشــان داد که برآوردگر بیز $\mu(\theta)$ تحت تابع زیان درجه دوم نســبت به هر برآوردگر دلخواه $\mu(\theta)$ بهتر اســت. اما در بسیاری از موارد محاســبه برآوردگر بیز کاری بسیار مشــکل و دشوار است به همین خاطر از برآوردگر باورمندی استفاده می‌شود. برای محاسبه برآوردگر باورمندی، ابتدا باید تابع درســت‌نمایی را محاسبه کرد که بتوان توزیع پســین را به دست آورد و درنهایت برآوردگر باورمندی محاسبه شود. بنابراین تابع درست‌نمایی برای توزیع آمیخته که در رابطه ۱ ارائه شد به‌صورت زیر خواهد بود:

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{p}|\mathbf{y})=\prod_{i=1}^{n}\sum_{j=1}^{k}w_j\,f(y_i|\,\theta_j) \qquad (۷)$$

همان‌طور که ملاحظه می‌شود بسط تابع درست‌نمایی فوق،k^n جمله خواهد داشت، به همین دلیل محاسبه تابع درست‌نمایی، تابع توزیع پسین و برآوردگرهای بیز، حتی برای حالتی که توزیع پیشین مزدوج برای (θ_i,p_i) در نظر گرفته شــود، کار بسیار مشکلی است. برای حل این مشــکل یک‌ســری نمادهای جدید در جدول ۱ ارائه شده که با استفاده از این نمادهای جدید فرم ساده‌تری برای تابع درست‌نمایی فوق پیدا کــرده و باتوجه به آن می‌توان برآوردگر باورمندی را برای توزیع‌های آمیخته محاسبه کرد.

قضیه۱) فرض کنید که نمونه تصادفی $Y_1,Y_2,\ldots,Y_n$ از توزیع آمیخته k عضوی رابطه ۱ آمده اســت. تابع درست‌نمایی برای توزیع آمیخته رابطه ۱ یک فرمول بازگشتی به‌صورت زیر خواهد بود:

$$L_k(\boldsymbol{\psi}|\boldsymbol{Y})=\sum_{i=0}^{n}w_k^i\,(1-w_k)^{n-i}\sum_{r=1}^{\binom{n}{i}}L_{k-1}(\boldsymbol{\psi}(-k)|\widetilde{Y_{B_{ir}^c}})\prod_{k\in B_{ir}}f_k(y|\theta_k) \qquad (۸)$$

که $L_{k-1}\left(\boldsymbol{\psi}(-k)\middle|\widetilde{Y_{B_{ir}^c}}\right)$ تابع درستنمایی برای تابع چگالی

$$g^*\left(y|\boldsymbol{\psi}(-k)\right)=\frac{w_1}{1-w_k}f_1(y|\theta_1)+\frac{w_2}{1-w_k}f_2(y|\theta_2)+\cdots+\frac{w_{k-1}}{1-w_k}f_{k-1}(y|\theta_{k-1})$$

خواهد بود.

هدف ما در این پژوهش این اســت که بیمه شــده‌ها را در دو دسته پرریســک و کم‌ریسک تقســیم کنیم. برای این منظور فرض بر این اســت که متغیر پاسخ Z_i دو مقدار ۰ و ۱ را اختیار می‌کند به این معنی که $Z_i=1$ یعنی بیمه شــده i ام متعلق به جامعه پرریســک و $Z_i=0$ یعنی بیمه شــده i ام متعلق به جامعه کم‌ریســک خواهد بود. در تحلیل عوامل مؤثر بر پرریســک یا کم‌ریسک بودن بیمه شده‌ها، متغیرهای همچون جنسیت، شغل، سن و غیره ممکن است تأثیرگذار باشــد، برای این منظور اگر $X_i=(X_{i1},X_{i2},\ldots,X_{ik})$ نشان‌دهنده بردار متغیرهای مســتقل باشــد. آنگاه احتمال اینکه بیمه شده به جامعه پرریسک متعلق باشد با استفاده از رگرسیون لجستیک به‌صورت زیر قابل محاسبه است:

$$w_1=w_1(\mathbf{x},\boldsymbol{\alpha})=P(Z_i=1|\mathbf{X}=\mathbf{x})=\frac{\exp\{\beta_{.1}+\boldsymbol{\beta}_1'\mathbf{x}\}}{\{1+\exp\{\beta_{.1}+\boldsymbol{\beta}_1'\mathbf{x}\}\}} \qquad (۹)$$

بنابراین اگر توزیع برازش داده شده به جامعه کم‌ریسک $f(x|\theta_1)$ و به جامعه پرریسک $f(x|\theta_2)$ باشد، آنگاه توزیع آمیخته برازش داده شده به‌صورت زیر خواهد بود:

$$f(x|\theta_1,\theta_2)=\left(1-\frac{\exp\{\beta_{1}+\boldsymbol{\beta}'\mathbf{x}\}}{\{1+\exp\{\beta_{1}+\boldsymbol{\beta}'\mathbf{x}\}\}}\right)f(x|\theta_1)+\frac{\exp\{\beta_{1}+\boldsymbol{\beta}'\mathbf{x}\}}{\{1+\exp\{\beta_{1}+\boldsymbol{\beta}'\mathbf{x}\}\}}\,f(x|\theta_2) \qquad (10)$$

قضیه ۲) فرض کنید که نمونه تصادفی $Y_1,Y_2,\ldots,Y_n$ از توزیع آمیخته k عضوی رابطه ۱ آمده است به‌طوری‌که وزن‌های آمیختگی در این توزیع با استفاده از رگرسیون لجستیک به‌صورت $w_k=w_k(x,\boldsymbol{\alpha})=\frac{\exp\{\beta_{.k}+\boldsymbol{\beta}_k'x\}}{\{1+\sum_{j=1}^{k-1}\exp\{\beta_{.j}+\boldsymbol{\beta}_j'x\}\}}$ به‌دست‌آمده باشد. در این حالت برآوردگر باورمندی به‌صورت زیر قابل‌محاسبه است:

$$\begin{aligned}&E_k(Y_{n+1}|Y_1,Y_2,\ldots,Y_n)\\&=\sum_{i=0}^{n}\sum_{r=1}^{\binom{n}{i}}w_k^i(1-w_k)^{n-i}\left[w_kE_1(Y_{n+1}|\tilde{Y}_{B_{ir}}\in PoP_k)\right.\\&\left.+(1-w_k)E_{k-1}\left(Y_{n+1}\middle|\tilde{Y}_{B_{ir}^c}\notin PoP_k\right)\right]\end{aligned} \qquad (11)$$

اثبات) ابتدا باید تابع توزیع متغیر تصادفی Y_1 محاسبه شود

$$G_{Y_1|\boldsymbol{\psi}}(t)=P(Y_1\le t|\boldsymbol{\psi})=\sum_{j=1}^{k}P(Y_1\le t|Y_1\in PoP_j,\boldsymbol{\psi})P(Y_1\in PoP_j|\boldsymbol{\psi})=\sum_{j=1}^{k}w_j\,F_j(t) \qquad (12)$$

باتوجه به تابع توزیع فوق، تابع چگالی به‌صورت $g(y|\boldsymbol{\psi})=w_1f_1(y|\theta_1)+w_2f_2(y|\theta_2)+\cdots+w_kf_k(y|\theta_k)$ است. می‌توان تابع چگالی فوق را به‌صورت زیر بازنویسی کرد:

$$g(y|\boldsymbol{\psi})=w_kf_k(x|\theta_k)+(1-w_k)f^*(y|\boldsymbol{\psi}(-k)) \qquad (13)$$

که

$$f^*(y|\boldsymbol{\psi}(-k))=\frac{w_1}{1-w_k}f_1(y|\theta_1)+\frac{w_2}{1-w_k}f_2(y|\theta_2)+\cdots+\frac{w_{k-1}}{1-w_k}f_{k-1}(y|\theta_{k-1})$$

باتوجه به تابع چگالی فوق برآوردگر باورمندی به صورت زیر قابل محاسبه است:

$$\begin{aligned}&E_k(Y_{n+1}|Y_1,Y_2,\ldots,Y_n)=E\left(E\left(Y_{n+1}|\theta_k,\boldsymbol{\psi}(-k)\right)|\tilde{Y}\in\cup_{l=1}^{k}PoP_l\right)=\sum_{i=0}^{n}\sum_{r=1}^{n}E(Y_{n+1}|\tilde{Y}_{B_{ir}}\in PoP_k,\tilde{Y}_{B_{ir}^c}\\&\notin PoP_k)P(\tilde{Y}_{B_{ir}}\in PoP_k,\tilde{Y}_{B_{ir}^c}\notin PoP_k\left|\tilde{Y}\in\left(\bigcup_{l=1}^{k}PoP_l\right)\right.\\&=\sum_{i=0}^{n}\sum_{r=1}^{n}w_k^i(1-w_k)^{n-i}\,E\left(E\left(Y_{n+1}|\theta_k,\boldsymbol{\psi}(-k)\right)|\tilde{Y}_{B_{ir}}\in PoP_k,\tilde{Y}_{B_{ir}^c}\notin PoP_k\right)\\&=\sum_{i=0}^{n}\sum_{r=1}^{n}w_k^i(1-w_k)^{n-i}\left[w_kE\left(\mu(\theta_k)|\tilde{Y}_{B_{ir}}\in PoP_k\right)+(1-w_k)E\left(\mu(\boldsymbol{\psi}(-k))\right|\tilde{Y}_{B_{ir}^c}\notin PoP_k)\right]=\\&=\sum_{i=0}^{n}\sum_{r=1}^{n}w_k^i(1-w_k)^{n-i}\left[w_kE(Y_{n+1}|\tilde{Y}_{B_{ir}}\in PoP_k)\right.\\&\left.+(1-w_k)E_{k-1}\left(Y_{n+1}\middle|\tilde{Y}_{B_{ir}^c}\notin PoP_k\right)\right]\end{aligned} \qquad (14)$$

مراحل محاسبه برآوردگر باورمندی برای توزیع‌های آمیخته در

یافته‌ها

داده‌های مورد استفاده در این مطالعه شامل خسارت‌های مربوط به مجموعه بیمه‌نامه‌های درمان گروهی صادره طی دو سال از اواسط سال ۱۳۹۷ تا اواسط ۱۳۹۹ است. به‌طوری که در هر قرارداد بنا بر نوع قرارداد که شامل متغیرهای سقف تعهد و فرانشیز است، توسط بیمه‌شدگان که اطلاعات سن، جنسیت (زن و مرد)، وضعیت (اصلی و تحت تکفل)، استان محل زندگی و گروه شغلی (اتحادیه/اصناف/خانه کارگر/تعاونی، دانشگاه‌های علوم پزشکی و مراکز درمانی، وزارت کشور، استانداری‌ها، شهرداری‌ها، بخشداری و نظایر آن، محیط‌های اداری (به‌غیراز طبقه ۲ و ۳) و محیط‌های تولیدی و صنعتی) آنها در اختیار است، خسارت‌هایی درخصوص خدمات مختلف ثبت شده، اما هدف در این پژوهش مدل‌بندی و پیش‌بینی خسارت‌های مربوط به دندانپزشکی است. بااینکه ۴۴درصد از خسارت‌های دندانپزشکی مربوط به آقایان و ۵۶درصد از آنها مربوط به خانم‌هاست، برای بررسی تفاوت بین خسارت‌های دندانپزشکی آقایان و خانم‌ها با استفاده از روش تحلیل واریانس (ANOVA)، پی- مقدار برابر با $2*10^{-16}$ شده که در سطح α=۰،۰۵ فرض صفر رد می‌شود. بدین‌معنا که بین خسارت‌های دندانپزشکی آقایان و خانم‌ها به‌طور معناداری تفاوت وجود دارد. به‌منظور تجزیه و تحلیل متغیرها، ابتدا آمار توصیفی از داده‌ها ارائه می‌شود که نتایج آن در جدول ۳ آمده است. در این جدول مقادیر حداکثر، حداقل، میانگین، میانه و انحراف معیار هر یک از متغیرهای جنسیت، شغل و نسبت بیمه شده برای خسارت‌های مربوط به دندانپزشکی است.

جدول ۳. متغیرهای توصیفی پژوهش

با نگاهی به شاخص سلامت دندان گروه‌های سنی مختلف در ایران،

جدول ۱| لیست علائم و نمادها

نماد	تعریف
K	تعداد عضوهای مدل آمیخته
N	تعداد مشاهدات
$\tilde{Y}$	$(Y_1,Y_2,\ldots,Y_n)$
S^n	$\{1,2,\ldots,n\}$
B_{ir}	rامین زیرمجموعه i عضوی از S^n
B_{ir}^c	متمم مجموعه بالا
S_i^n	$=\{B_{ir}:r=1,2,\ldots,\binom{n}{i}\}$
$\tilde{Y_{B_{ir}}}$	$=(Y_{k_1},Y_{k_2},\ldots,Y_{ki}),\quad k_1,k_2,\ldots,k_n\in B_{ir}$
$\tilde{Y_{B_{ir}^c}}$	$=(Y,Y_{k_2},\ldots,Y_{k_{n-i}}),\quad k_1,k_2,\ldots,k_{n-i}\in B_{ir}^c$
$\boldsymbol{\Psi}$	$=(\theta_1,\theta_2,\ldots,\theta_K)$
$\boldsymbol{\Psi}(-\mathbf{l})$	$=(\theta_1,\theta_2,\ldots,\theta_{l-1},\theta_{l+1},\ldots,\theta_K)$
$E_K(.)$	امید ریاضی برای توزیع آمیخته با k عضو

می‌توان به این نتیجه رسید که شاخص سلامت دندان‌های شیری برابر با ۱۶/۵ درصد، دندان‌های دائمی در کودکان با رده سنی ۱۲ سال برابر با ۹/۲ درصد، برای گروه سنی ۳۵-۴۴ سال برابر با ۹۸/۱۳ و گروه سنی ۶۵-۷۴ سال برابر با ۲۵/۲۷ درصد است. در گروه سنی ۶۵-۷۴ سال تعداد قابل‌توجهی از دندان‌های پوسیده درمان شده وجود دارد. باتوجه به موارد گفته شده می‌توان فهمید که هزینه‌های دندانپزشکی یکی از پرهزینه‌ترین خدمات درمانی بوده و همه افراد حداقل چند بار در طول زندگی به دندانپزشکی مراجعه کرده‌اند. این هزینه در کشور ما، بعد از هزینه بستری شدن، در رتبه دوم قرار دارد. به‌گونه‌ای که حدود ۱۵-۲۰ درصد از کل هزینه‌های خدمات درمانی-دارویی را دربر می‌گیرد. همین موضوع اهمیت بیمه تکمیلی دندانپزشکی را نشان می‌دهد. خدمات درمانی تعهدات دندانپزشکی سازمان‌های بیمه‌گر شامل خدمات ویزیت (معاینه و تشخیص، تجویز دارو، اعزام به رادیولوژی و درخواست آزمایش)، رادیوگرافی تک‌دندان (گرافی پری اپیکال یا بایت وینگ)، کشیدن دندان قدامی، کشیدن دندان خلفی، کشیدن دندان عقل، جراحی دندان نهفته در نسج نرم، جراحی دندان نهفته در نسج سخت، جرم‌گیری بالای لثه، جرم‌گیری زیر لثه، بروساژ هر فک (نوعی جرم‌گیری)، پر کردن تک‌سطحی و ترمیم دندان است. بیمه درمان تکمیلی گروهی بین نهاد متقاضی و بیمه به‌صورت قراردادی منعقد می‌شود. شرایط و پوشش‌های این بیمه با توافق طرفین مشخص می‌شود. همچنین سقف تعهدات و پوشش‌های بیمه تکمیلی انفرادی به‌صورت ثابت انجام می‌شود. هدف از این مطالعه این است که مدل‌بندی خسارت‌های مربوط به پوشش دندانپزشکی افراد بیمه شده و برآوردگر باورمندی ارائه شود. در نمودار ۱ چگالی مربوط به خسارت‌های دندانپزشکی که بر عدد ۱۰۰۰۰ تقسیم شده‌اند، رسم شده است.

باتوجه به نمودار چگالی مشخص است که یک ناهمگنی در خسارت‌ها وجود دارد که این ناهمگنی به دلیل مشخصه‌های مانند سن، جنسیت، شغل، موقعیت جغرافیایی، نسبت با بیمه شده و... است. بنابراین هدف این است که بیمه‌شده‌ها را باتوجه به مشخصه‌هایی که دارند به دو زیرجامعه پرریسک و کم‌ریسک تقسیم کرد. برای این منظور با استفاده از روش K-means مرکز هر یک از خوشه‌ها به‌دست آمده است که یک نمایش گرافیکی دومتغیره از خوشه‌بندی داده‌ها در شکل ۱ است. در این شکل، داده‌ها به‌صورت نقاط در یک نمودار دومتغیره و خوشه‌ها به‌صورت بیضی‌هایی با اندازه‌ها و اشکال مختلف نمایش داده شده است.

حال بعد از اینکه داده‌ها در خوشه‌های پرریسک و کم‌ریسک تقسیم شدند باید به هر یک از این زیرجامعه‌ها توزیع مناسب برازش داده شود. برای این منظور ۴ توزیع نرمال، نمایی، گاما و لگ‌نرمال را به هر یک زیرجامعه‌های پرریسک و کم‌ریسک برازش داده و مقدار

جدول۲| به‌اختصار توضیح داده شده است

الگوریتم محاسبه برآوردگر باورمندی برای توزیع‌های آمیخته متناهی

ورودی: Y مجموعه داده‌های خسارت n سال گذشته یک بیمه شده خاص، مرکز خوشه‌ها

خروجی: محاسبه وزن‌های آمیختگی، برازش توزیع آمیخته، محاسبه برآوردگر باورمندی

۱) انتخاب $\boldsymbol{\mu_1, \mu_2, \dots, \mu_k}$ به‌عنوان مرکز خوشه‌ها

۲) فاصله تک‌تک مشاهدات را از مرکز خوشه‌ها $\boldsymbol{(y_i - \mu_l)^2 \quad i = 1, 2, \dots, n \quad l = 1, 2, \dots, k}$

۳) مشخص کردن خوشه هر یک از مشاهدات $\boldsymbol{Z_i = \arg\min(y_i - \mu_l)^2 \quad l = 1, \dots, k \quad i = 1, 2, \dots, n}$

۴) محاسبه مرکز خوشه‌ها با توجه به مشاهداتی که در هر خوشه قرار می‌گیرد $\boldsymbol{\widehat{\mu_l} = \frac{1}{n_l}\sum_{i=1}^{n} y_i \, I_{\{z_i = l\}}}$

۵) تکرار مراحل ۲ تا ۴ تا زمانی که مرکز خوشه‌ها تغییر نکند

۶) برازش توزیع مناسب به هر یک از خوشه‌ها

۷) محاسبه وزن‌های آمیختگی $\boldsymbol{w_i}$ با استفاده از تابع لجستیک $\boldsymbol{\mathbf{w_i} = w_i(\mathbf{x}, \alpha) = \frac{\exp\{\beta_{0i} + \beta_i' \mathbf{x}\}}{\{1 + \sum_{j=1}^{k-1} \exp\{\beta_{0j} + \beta_j' \mathbf{x}\}\}}}$

برازش توزیع نهایی به‌صورت آمیخته‌ای از توزیع‌های برازش داده‌شده در گام پنجم با وزن‌های آمیختگی محاسبه‌شده در گام ششم

۸) محاسبه برآوردگر باورمندی با استفاده از توزیع آمیخته برازش داده‌شده

آماره‌های آزمون کولموگروف اسمیرنوف، اندرسون دارلینگ و معیارهای AIC و BIC با همدیگر مقایسه شده‌اند که درنهایت توزیع برازش داده شده به زیرجامعه کم‌ریسک دارای توزیع لگ‌نرمال با پارامترهای μ= ۴۱۹۵/۱۵ و σ=۸۱۳۱۹/۰ و توزیع برازش داده شده به زیر جامعه پرریسک دارای توزیع لگ‌نرمال با پارامترهای μ= ۱۵,۵۹۴۶ و σ= ۰,۹۶۸۴۷۶۲ بوده است. بنابراین توزیع برازش داده شده به خسارت‌های دندانپزشکی به‌صورت زیر خواهد بود:

$$f(x|\theta_1,\theta_2)=\left(1-\frac{\exp\{\beta_{.1}+\boldsymbol{\beta}'_1\mathbf{x}\}}{\{1+\exp\{\beta_{.1}+\boldsymbol{\beta}'_1\mathbf{x}\}\}}\right)LN_{(15/4,0/8)}+\frac{\exp\{\beta_{.1}+\boldsymbol{\beta}'_1\mathbf{x}\}}{\{1+\exp\{\beta_{.1}+\boldsymbol{\beta}'_1\mathbf{x}\}\}}LN_{(15/6,0/96)} \quad (15)$$

یا می‌توان گفت که لگاریتم خسارت‌ها دارای توزیع آمیخته متناهی از نرمال خواهد بود:

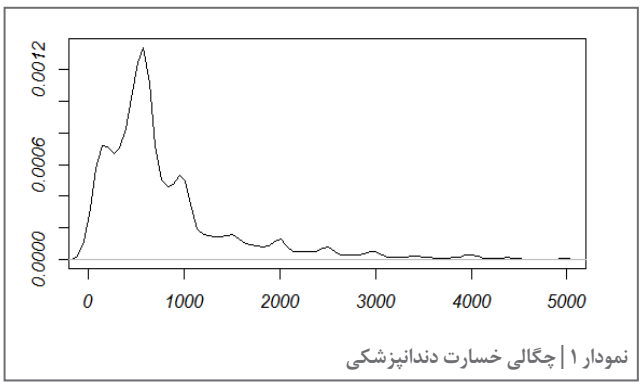
$$f(x|\theta_1,\theta_2)=\left(1-\frac{\exp\{\beta_{.1}+\boldsymbol{\beta}'_1\mathbf{x}\}}{\{1+\exp\{\beta_{.1}+\boldsymbol{\beta}'_1\mathbf{x}\}\}}\right)N_{(15/4,0/8)}+\frac{\exp\{\beta_{.1}+\boldsymbol{\beta}'_1\mathbf{x}\}}{\{1+\exp\{\beta_{.1}+\boldsymbol{\beta}'_1\mathbf{x}\}\}}N_{(15/6,0/96)} \quad (16)$$

حال بایستی با استفاده از رگرسیون لجستیک احتمال تعلق به هریک از زیر جامعه‌ها (وزن‌های آمیختگی) را برآورد کرد. همان‌طور که قبلاً اشاره شد، متغیرهای کمکی جنسیت، شغل، سن، وضعیت و موقعیت جغرافیایی ازجمله عواملی بود که ممکن است در ناهمگنی داده‌ها تاثیرگذار باشد. براساس نتایج مربوط به آزمون معنادار هر یک از این متغیرها در جدول ۴ نشان داده شده که پی- مقدار برای هر یک از این متغیرها در سطح α=0.05 معنادار است. بدین‌معنا که هر یک از این عوامل بر خسارت‌های دندانپزشکی تأثیرگذار خواهد بود. نتایج مربوط برآورد پارامترها و پی-مقدار در جدول ۴ آمده است. همان‌طور که در جدول زیر نشان داده شده است مقدار sig برای هریک از این متغیرها گزارش شده است. باتوجه به اینکه مقدار sig (پی-مقدار) برای هر یک از این متغیرها کمتر از ۰/۰۵ است می‌توان گفت که هر یک از این متغیرها در میزان خسارت یا اینکه بیمه شده متعلق به جامعه پرریسک یا کم‌ریسک است تأثیرگذار خواهد بود.

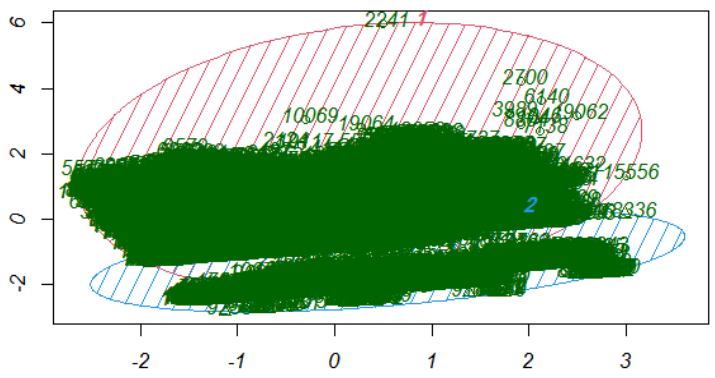
باتوجه به اینکه متغیر پاسخ Y دو مقدار صفر و یک را اختیار می‌کند (Y=1 یعنی بیمه شده متعلق به زیرجامعه پرریسک و Y=0 یعنی بیمه شده متعلق به زیرجامعه کم ریسک) احتمال اینکه یک بیمه شده با مشخصات $X=(X_1,X_2,X_3,X_4,X_5)$ متعلق به جامعه پرریسک باشد به‌صورت زیر قابل‌محاسبه است:

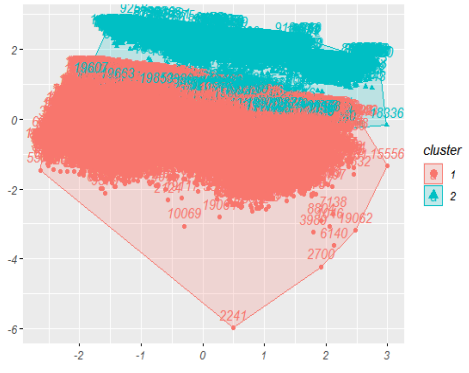
$$P(Y=1|X=x)=\frac{\exp\{۳۰/۸۱۰۶+۴/۷۹۴۹X_1-۵/۱۶X_2+۰/۰۶۷۸X_3-۱۸/۹۶X_4+۱/۱۸۷۱X_5\}}{\{1+\exp\{۳۰/۸۱۰۶+۴/۷۹۴۹X_1-۵/۱۶X_2+۰/۰۶۷۸X_3-۱۸/۹۶X_4+۱/۱۸۷۱X_5\}\}} \quad (17)$$

به‌عنوان‌مثال، فرض کنید که خسارت‌های $Y_1,Y_2,\ldots,Y_n$ مربوط به یک بیمه شده با مشخصه‌های جنسیت خانم (X_1=1)، بیمه شده اصلی (X_2=0)، سن ۶۰ سال (X_3=60)، طبقه شغلی ۳ (X_4=3) در استان سمنان



نمودار ۱ | چگالی خسارت دندانپزشکی





شکل ۱ | خوشه‌بندی داده‌ها

(X_5=16)باشد که از چگالی برازش داده شده در رابطه (۴) با پیشین‌های $N(\mu_k,b_k^2)$ k=1,2 پیروی می‌کند. باتوجه به این مفروضات این بیمه شده با احتمال $P(Y=1 \mid X=x)$=۸۵۶۷۵۴۳/۰ متعلق به جامعه پرریسک و با احتمال $P(Y=0 \mid X=x)$=۱۴۳۲۴۵۷/۰ متعلق به جامعه کم‌ریسک خواهد بود. بنابراین لگاریتم خسارت‌های این بیمه شده توزیع آمیخته‌ای به‌صورت:

$$f(y)=۱۴۳۲۴۵۷N(۱۵/۴۱۹۵,۰/۸۱۳۱۹)+۰/۸۵۶۷۴۳\ N۱۵$$

خواهد بود. حال برای محاسبه برآوردگر باورمندی فرض کنید که ۱۶/۱۹۵۰۲، ۱۳/۹۲۸۲۳، ۱۵/۶۹۷۶۰، ۱۵/۰۰۵۱۵، ۱۵/۳۰۲۹۳، ۱۶/۵۴۰۰۵، ۱۶/۰۳۶۲۶، ۱۶/۸۴۸۲۳، ۱۴/۴۹۷۱۶، ۱۴/۷۵۲۵۸ لگاریتم خسارت‌های ده سال گذشته این بیمه شده باشد و پارامترهای توزیع پیشین به این صورت 5/b_2=0,5/μ_1=8,μ_2=9,b_1=0 باشد. باتوجه به این مفروضات برآوردگر باورمندی به‌صورت زیر قابل‌محاسبه است:

$$\theta_۲|\tilde{Y}_{B_{ir}} \in PoP_۲ \sim N\left(\frac{b_۲^۲\sum_{k\in B_{ir}} x_k+\mu_۲\sigma_۲^۲}{ib_۲^۲+\sigma_۲^۲},\frac{\sigma_۲^۲b_۲^۲}{ib_۲^۲+\sigma_۲^۲}\right) \quad (۱۸)$$

$$\theta_۱|\tilde{Y}_{B_{ir}{}^c} \in PoP_۱ \sim N\left(\frac{b_۱^۲\sum_{k\in B_{ir}^c} x_k+\mu_۱\sigma_۱^۲}{(n-i)b_۱^۲+\sigma_۱^۲},\frac{\sigma_۱^۲b_۱^۲}{(n-i)b_۱^۲+\sigma_۱^۲}\right) \quad (۱۹)$$

$$\begin{aligned}E_۲(Y_{n+۱}|Y_۱,Y_۲,\dots,Y_n) &= \\ &= \sum_{i=۰}^{n}\sum_{r=۱}^{\binom{n}{i}} ۰/۸۵۶^i\ ۰/۱۴۳^{n-i}\left[۰/۸۵۶E_۱\left(Y_{n+۱}|\tilde{Y}_{B_{ir}} \in PoP_۲\right)\right. \\ &= ۱۵/۳۹۳۲۴ \quad (۲۰) \\ &\left. + ۰/۱۴۳E_۱\left(Y_{n+۱}\middle|\tilde{Y}_{B_{ir}^c} \notin PoP_۲\right)\right]\end{aligned}$$

همان‌طور که در بالا اشاره شد پیش‌بینی مربوط به خسارت سال n+۱ ام برای شخصی با مشخصات جنسیت خانم که به‌عنوان بیمه شده اصلی بودند در ۶۰ سالگی که در طبقه شغلی ۳ و استان سمنان قرار داشتند برابر ۱۵/۳۹۳۲۴ است. درنتیجه عواملی مانند سن، جنسیت، موقعیت جغرافیایی، طبقه شغلی و وضعیت بیمه شده در پیش‌بینی خسارت در نظر گرفته شدند.

بحث

ردنر و والکر [۱۴] به برخی از مشکلات برآوردگر ماکسیمم درست‌نمایی (MLE) برای توزیع‌های آمیخته متناهی اشاره کردند. اولاً نمی‌توان فرض کرد که تابع درست‌نمایی به‌طورکلی دارای کران بالاست، ممکن است واگرا باشد. ثانیاً اغلب کمینه‌های محلی بهینه زیادی برای تابع درست‌نمایی وجود دارد. مک لاچلان و باسفورد [۱۵] این مشکل را اظهار داشتند که عملکرد الگوریتم EM نسبت به مقدار اولیه پارامترهای مدل بسیار حساس است. اگر مقادیر اولیه نزدیک به مقادیر تکین در تابع درست‌نمایی باشد، الگوریتم EM ممکن است واگرا شود و سرعت همگرایی بسیار کند شود. روش دیگر برای برآورد پارامترهای مدل آمیخته، برآورد بیزی است که تابع درست‌نمایی داده‌ها با توزیع پیشین پارامترها، برای استخراج یک استنباط، ترکیب می‌شوند. یکی از مزایای روش بیزی نسبت به روش MLE، استفاده از اطلاعات قبلی است، این مزیت باعث می‌شود که با تعداد کمی از داده‌ها استنباط همچنان قابل انجام باشد. روش MLE، به‌ویژه برای مدل‌هایی که پارامترهای زیادی دارند، وقتی مجموعه داده کوچک است، برآورد مناسب و خوبی نیست. با این حال، در موقعیت‌های خاص، برآورد بیزی پارامترهای یک توزیع آمیخته از محاسبات طولانی برای انجام استنتاج برخوردار است، زیرا انتگرال‌ها باید بر روی یک فضای چندبعدی غیرمستقیم بالقوه انجام شوند. در الگوی بیزی برگر [۱۶]، بساگ و همکاران[۱۷] بیانیه‌های احتمال در مورد پارامترهای نامشخص، توزیع پیشین یا نظر نخبگان در تحلیل‌ها و توصیف‌های سلسله مراتبی از مدل‌های محلی- موضعی و سراسر بیان شده است. همچنین درزمینه مدل‌های آمیخته به‌صورت محلی-موضعی نیز می‌توان به مطالعه پاینده و همکاران [۱۸] اشاره کرد که ساختار پیچیده از یک مدل آمیخته را با استفاده از متغیرهای پنهان به یک ساختار ساده‌تر تجزیه کرد. همان‌طور که گفته شد، از زمان پیدایش الگوریتم EM، ML تاکنون متداول‌ترین روش برای برازش توزیع‌های آمیخته بوده است. کاربرد الگوریتم EM برای محاسبه برآوردهای ML مدلهای آمیخته پارامتری زمانی کاربرد دارد

جدول۳| برآورد پارامترهای رگرسیون لجستیک (یافته‌های پژوهش)

متغیرها	عرض از مبدأ	جنسیت	وضعیت	سن	شغل	موقعیت جغرافیایی
برآورد پارامتر	۳۰/۸۱	۴/۷۹	-۵/۱۶	۰/۰۶۷۸	-۱۸/۹۶	۱/۱۸
پی مقدار	$۲*۱۰^{-۱۶}$	$۲*۱۰^{-۱۶}$	$۱/۲۵*۱۰^{-۱۴}$	$۲.۹۱*۱۰^{-۷}$	$۲*۱۰^{-۱۶}$	$۲*۱۰^{-۱۶}$
Odd ratio		$۱/۲*۱۰^{۲}$	$۵/۸۳*۱۰^{-۳}$	$۱/۰۷*۱۰^{۰}$	$۵/۷۱*۱۰^{-۹}$	$۳/۲۸*۱۰^{۰}$

که داده‌ها ناکامل باشند. استفاده از الگوریتم MM توسط هانتر و لانگ [۱۹] و لانگ [۲۰] و همچنین در مطالعه نگوین و مک‌لاچان [۲۱] برای محاسبه برآوردهای ML در توزیع‌های آمیخته، در حالتی که محاسبه گام E در الگوریتم EM مشکل باشد، بسیار مورد توجه قرار گرفته است. همانند الگوریتم EM، برآورد بیزی با استفاده از روش‌های MCMC مبتنی بر کار دمپستر [۲۲] است که اظهار داشت یک مدل آمیخته متناهی را چگونه می‌توان بر اساس داده‌های ناکامل یا داده‌های گم شده ارزیابی کرد. همان‌طور که قبلاً بحث شد محاسبه تابع درست‌نمایی برای توزیع‌های آمیخته برحسب k^n جمله خواهد شد که محاسبه آن کار وقت‌گیر و بسیار مشکلی است. پاینده و ساکی زاده [۲۳] برای تابع درست‌نمایی توزیع‌های آمیخته یک تقریب ارائه کردند که برای حالت‌هایی که تعداد آمیخته‌ها زیاد باشد مناسب نیست. تاکنون برای تابع درست‌نمایی در مدل‌های آمیخته فرم بسته و دقیقی ارائه نشده بود. در این مطالعه با اعمال شرایطی جدید فرم ساده‌تر و بسته‌ای برای تابع درست‌نمایی ارائه شد که این فرم بسته باعث می‌شود به میزان قابل توجهی پیچیدگی محاسبات کاهش پیدا کند و همچنین برای محاسبه برآوردگر باورمندی عوامل مهمی که در پیش‌بینی خسارت تأثیرگذار بودند در نظر گرفته شد. لازم به ذکر است با استفاده از رویکرد به کار گرفته شده در این مطالعه می‌توان هر عامل دیگری را که در محاسبه برآوردگر باورمندی تأثیرگذار باشد، بدون پیچیدگی خاصی به مدل اضافه کرد و اثر آن را کنترل کرد.

نتیجه‌گیری

از مهم‌ترین وظایف شرکت‌های بیمه، پیش‌بینی میزان و تعداد خسارت‌ها در سال آتی است. روش‌های مختلفی مانند سری‌های زمانی، روش بیزی و نظریه باورمندی برای انجام چنین پیش‌بینی‌های وجود دارد. یکی از مزیت‌های روش باورمندی نسبت به بقیه روش‌ها، استفاده از اطلاعات قبلی است. این مزیت باعث می‌شود که با تعداد کمی از داده‌ها استنباط همچنان قابل انجام باشد. اما مسئله‌ای که در این زمینه وجود دارد این است که محاسبه تابع درست‌نمایی و برآوردگر باورمندی برای توزیع‌های آمیخته کاری مشکل و دشوار بود که با اعمال شرایطی جدید فرم ساده و بسته‌ای برای تابع درست‌نمایی ارائه و بنابراین مفروضات برآوردگر باورمندی برای توزیع‌های آمیخته محاسبه شد. سپس باتوجه به روش ارائه شده خسارت‌های مربوط به مجموعه بیمه‌نامه‌های درمان گروهی صادره طی دو سال از اواسط سال ۱۳۹۷ تا اواسط ۱۳۹۹ شرک بیمه دی مدل‌بندی شد که برای این منظور باتوجه به ناهمگنی موجود در داده‌های خسارت مربوط به بیمه درمان تکمیلی، بیمه‌شده‌ها را در دو دسته پرریسک و کم‌ریسک تقسیم کرده و به هر یک از این زیرجامعه‌ها یک توزیع مناسب برازش داده شد و توزیع کلی خسارت‌ها یک توزیع آمیخته از این توزیع‌ها در نظر گرفته شد که وزن‌های آمیختگی این توزیع را یک تابع لجستیک از متغیرهای مربوط به هر بیمه شده در نظر گرفته و درنهایت باتوجه به مدل برازش داده شده برآوردگر، باورمندی برای خسارت‌های دندانپزشکی محاسبه شد.

با استفاده از نتایج این مطالعه می‌توان عوامل مهم و تأثیرگذار مانند سن، جنسیت، طبقه شغلی و... را به مدل اضافه کرد و پیش‌بینی خسارت سال آینده را بر اساس این فاکتورها انجام داد. این پژوهش با ارائه مدل‌های ریاضی، مشکل مربوط به مدل‌بندی خسارت‌ها را مرتفع ساخته و با استفاده از رویکرد باورمندی خسارت‌های سال آینده مربوط به دندانپزشکی پیش‌بینی شد. بنابراین نتایج این مطالعه می‌تواند به بهبود بیمه درمان کمک کرده و مشکلات این حوزه را تا حد زیادی برطرف سازد. به‌علاوه باتوجه به این مدل، برآورد دقیق‌تری از هزینه‌های درمان داشت و حق بیمه منصفانه‌ای برای هر بیمه شده باتوجه به پارامترهای ریسک آن محاسبه کرد. به‌این‌ترتیب علاوه بر اینکه محصول طراحی شده، جذابیت بیشتری برای بیمه‌گذار خواهد داشت و با قیمت مناسب ارائه خواهد شد، شرکت بیمه نیز متحمل زیان کمتری خواهد شد.

تشکر و قدردانی: مقاله حاضر از پایان نامه در مقطع دکترای تخصصی استخراج شده است.

تأییدیه اخلاق: مقاله حاضر با کد IR.SBU.REC.1400.262 مورد تأیید کمیته اخلاق دانشگاه شهید بهشتی قرار گرفته است.

تضاد منافع: کلیه نویسندگان هیچ‌گونه تضاد منافعی را اعلام نکردند.

سهم نویسندگان: سهم نویسندگان در این مقاله یکسان است.

منابع مالی: هیچ‌گونه حمایت مالی از پژوهش حاضر صورت نپذیرفته است.

References

1. World Health Organization. World report on ageing and health. World Health Organization; 2015 Oct 22.
2. Sadeghu R. Population and development in Iran: Dimensions

and challenges. Tehran: National Population Studies & Comprehensive Management Institute. 2009. [Perisan]

3. Rezaeian M, Asadpour M, Hadavi M. Research barriers from the perspective of academic members and strategies for confronting with these barriers in Rafsanjan University of Medical Sciences, Iran. Health System Research. 2013;9(3):269-76. [Persian]
4. Hakimzadeh SM, Hosseini Shokouh SM, Bahadori M, Tahernezhad K. Research needs assessment and priority setting for health economics: a mixed method study in Iran. J Mil Med. 2014;16(1):23-8. [Persian]
5. Akbarzadeh BAR, M, Esmaeili M, Kimiafar Kh. Medical information management and assessment of direct costs of treatment of lung cancer. Health Information Management. 2009;5(210):151-8. [Persian]
6. Pearson K. Contributions to the mathematical theory of evolution. Philos Trans R Soc Lond A. 1894;185:71-110. doi: 10.1098/rsta.1894.0003.
7. Feddern HA. Hybridization between the western Atlantic angelfishes, Holacanthus isabelita and H. ciliaris. Bulletin of Marine Science. 1968;18(2):351-82.
8. Whitney AW. Theory of experience rating. 1918.
9. Keffer R. An experience rating formula. Transactions of the Actuarial Society of America. 1929;30:130-9.
10. Bailey AL. Credibility Procedures: Laplace's generalization of Bayes' Rule and the combination of collateral knowledge with observed data. New York State Insurance Department; 1950.
11. Bühlmann H, Straub E. Glaubwürdigkeit für schadensätze. Bulletin of the Swiss Association of Actuaries. 1970;70(1):111-33.
12. Norberg R. Credibility theory. Encyclopedia of Actuarial Science. 2004;1:398-406. doi: 10.1002/9780470012505.tac068.
13. Box GE. Science and statistics. J Am Stat Assoc. 1976;71(356):791-9. doi: 10.1080/01621459.1976.10480949.
14. Redner RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review. 1984;26(2):195-239. doi: 10.1137/1026034.
15. McLachlan GJ, Basford KE. Mixture models: Inference and applications to clustering. New York: M. Dekker; 1988.
16. Berger JO. Prior information and subjective probability. InStatistical Decision Theory and Bayesian Analysis 1985 (pp. 74-117). Springer, New York, NY. doi: 10.1007/978-1-4757-4286-2_3.
17. Besag J, Green P, Higdon D, Mengersen K. Bayesian computation and stochastic systems. Statistical Science. 1995:3-41. doi: 10.1214/ss/1177010123.
18. Najafabadi AT, Barmalzan G, Aghaei S. A weighted model confidence set: applications to local and mixture model confidence sets. arXiv preprint arXiv:1701.05455. 2017 Jan 19. doi: 10.1504/IJMMNO.2017.086797.
19. Hunter DR, Lange K. A tutorial on MM algorithms. Am Stat. 2004;58(1):30-7. doi: 10.1198/0003130042836.
20. Lange K. The MM algorithm. In Optimization 2013 (pp. 185-219). Springer, New York, NY. 10.1007/978-1-4614-5838-8_8.
21. Nguyen HD, McLachlan GJ. Maximum likelihood estimation of triangular and polygonal distributions. Comput Stat Data Anal. 2016;102:23-36. doi: 10.1016/j.csda.2016.04.003.
22. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EMalgorithm. J R Stat Soc. 1977;39(1):1-38. doi: 10.1111/j.2517-6161.1977.tb01600.x.
23. Najafabadi AT, Sakizadeh M. Designing an Optimal Bonus--Malus System Using the Number of Reported Claims, Steady-State Distribution, and Mixture Claim Size Distribution. arXiv preprint arXiv:1701.05441. 2017 Jan 19.