Iran J. Insur. Res., 3(3): 361-371, Summer 2014, Serial #9

Iranian Journal of Insurance Research
(IJIR)

Homepage: https://ijir.irc.ac.ir/?lang=en

Insurance Research Center (IRC)

ORIGINAL RESEARCH PAPER

Analysis of longitudinal count responses for the number of claims with a high number of zeros in the third party insurance portfolio of Iran

F. Salavati, E. Bahrami Samani*

Department of Insurance Statistics, Shahid Beheshti University, Tehran, Iran

ARTICLE INFO

Article History
Received: 07 January 2013
Revised: 19 February 2013
Accepted: 02 June 2014

Keywords
EM-algorithm; Third Party Liability Insurance; No. of Claims; Zero- Inflated Distribution; Longitudinal Responses.

ABSTRACT

In third-party insurance, due to the reward-penalty system and the use of the end-of-year bonus system, the insured does not report his small losses to the insurance company. This creates many zeros in the number of claims for the insured. On the other hand, it is important for insurance companies to analyze the number of claims and the risk factors on this answer. For this purpose, some models with count responses using power series distribution such as Poisson regression model and negative binomial regression model and zero-accumulated power series distribution such as zero-accumulated Poisson regression model and zero-accumulated negative binomial regression for third-party insurance data analysis It is used with many zeros. In this paper, these models can be generalized to longitudinal third-party insurance data with a large number of zeros. A likelihood-based approach is used to obtain model parameter estimates. In this method, the EM algorithm is also used to estimate the parameters for models with zero accumulated response. Finally, to explain the usefulness of the proposed models, real longitudinal data of third party insurance has been analyzed.

***Corresponding Author:**
Email: *ehsan_bahrami_samani@yahoo.com*
DOI: 10.22056/ijir.2014.03.08

نشریه علمی پژوهشنامه بیمه

سایت نشریه: https://ijir.irc.ac.ir/?lang=fa

مقاله علمی

تحلیل پاسخ‌های شمارشی طولی برای تعداد ادعاهای خسارت با تعداد صفر زیاد در پرتفوی بیمه شخص ثالث کشور ایران

فرید صلواتی، احسان بهرامی سامانی*

گروه آمار بیمه، دانشگاه شهید بهشتی، تهران، ایران

اطلاعات مقاله

تاریخ دریافت: ۱۸ دی ۱۳۹۱

تاریخ داوری: ۰۱ اسفند ۱۳۹۱

تاریخ پذیرش: ۱۲ خرداد ۱۳۹۳

کلمات کلیدی

الگوریتم EM

بیمه شخص ثالث

تعداد ادعای خسارت

توزیع آماسیده-صفر

داده‌های طولی

چکیده:

در بیمه شخص ثالث به دلیل وجود سیستم پاداش- جریمه و استفاده از سیستم پاداش آخر سال، بیمه‌گذار خسارت‌های کوچک خود را به شرکت بیمه گزارش نمی‌دهد. این کار باعث ایجاد صفرهای زیاد در تعداد ادعای خسارت بیمه‌گذار می‌شود. ازسوی‌دیگر تحلیل تعداد ادعای خسارت و عوامل تشکیل‌دهنده خطر روی این پاسخ، برای شرکت‌های بیمه حایز اهمیت است. برای این منظور، برخی از مدل‌ها با پاسخ‌های شمارشی با استفاده از توزیع سری‌های توانی مانند مدل رگرسیون پواسون و مدل رگرسیون دوجمله‌ای منفی و توزیع سری‌های توانی آماسیده صفر مانند مدل رگرسیون پواسون آماسیده صفر و رگرسیون دوجمله‌ای منفی آماسیده صفر برای تحلیل داده‌های بیمه شخص ثالث با تعداد صفر زیاد استفاده می‌شود. در این مقاله می‌توان این مدل‌ها را برای داده‌های طولی بیمه شخص ثالث با تعداد صفر زیاد تعمیم داد. یک شیوه درستنمایی مبنا برای به‌دست‌آوردن برآورد پارامترهای مدل استفاده شده است. در این روش از الگوریتم EM نیز در برآورد پارامترها برای مدل‌هایی با پاسخ آماسیده صفر استفاده شده است. در نهایت برای تشریح سودمندی مدل‌های پیشنهادشده، داده‌های واقعی طولی بیمه شخص ثالث، مورد تجزیه و تحلیل قرار گرفته است.

***نویسنده مسئول:**

ایمیل: ehsan_bahrami_samani@yahoo.com

DOI: 10.22056/ijir.2014.03.08

مقدمه

از موضوع‌های مهم در بیمه شخص ثالث، تحلیل و انتخاب مدل مناسب برای برازش روی داده‌های تعداد ادعای خسارت است. تحلیل رگرسیون با پاسخ‌های شمارشی، اجازه شناسایی عامل‌های ریسک و پیشگویی فراوانی مورد انتظار ادعاها با توجه به ویژگی‌های قرارداد را می‌دهد. در بیمه منظور از پاسخ‌ها، تعداد ادعاهای خسارتی است که شخص بیمه‌گذار به شرکت‌های بیمه گزارش می‌دهد. شرکت‌های بیمه برای به‌دست‌آوردن حق‌بیمه از مدل‌های شمارشی شناخته‌شده مانند مدل پواسون[1] برای تعداد ادعای خسارت استفاده می‌کنند که این مدل‌ها در بیمه اتومبیل به دلیل آنکه این نوع داده‌ها دارای صفرهای زیادی می‌باشند، کارایی پایین‌تری نسبت به مدل‌های آماسیده صفر[2] دارد. بیشتر این صفرها به دلیل گزارش دروغین بیمه‌گذارها برای استفاده از سیستم تخفیف آخر سال است. مدل‌های آماسیده صفر (که در ادامه معرفی می‌شود) به دلیل اینکه یک پارامتر احتمالی را برای این صفرها درنظر می‌گیرد، توانایی برآورد احتمال این ادعاهای دروغین را بر اساس سابقه بیمه‌گذار- که همان متغیرهای تبیینی و تعداد ادعای خسارت سال‌های گذشته بیمه‌گذار است- دارد.

این تحقیق بر آن است که با مقایسه مدل‌های پواسون و دوجمله‌ای منفی[3] با مدل‌های پواسون آماسیده صفر[4] و دو جمله‌ای منفی آماسیده صفر[5]، بهترین مدل را بر اساس معیارهای مورد نظر به‌دست‌آورد. برای این منظور علاوه بر تعداد ادعای خسارت، متغیرهای تبیینی مانند نوع اتومبیل، سن اتومبیل و محل رانندگی نیز درنظرگرفته‌می‌شود، سپس با واردکردن این متغیرها در مدل‌های مورد بررسی، تأثیر آنها را در مدل مورد ارزیابی قرار می‌دهد. در نهایت بهترین مدل از بین مدل‌های شمارشی رایج و مدل‌های آماسیده صفر را برای برازش به داده‌های طولی تعداد ادعای خسارت با صفر زیاد معرفی می‌نماید.

مبانی نظری پژوهش

بیان مسئله و اهمیت تحقیق

برای تحلیل تعداد ادعای خسارت، پاسخ Y_{it} ، $(i=1\ldots,k, \quad t=1,\ldots,T)$ به عنوان تعداد ادعای خسارت بیمه‌گذار iام در زمان tام درنظرگرفته می‌شود. به‌طوری‌که این پاسخ‌ها یک پاسخ طولی است که به شرکت بیمه گزارش داده شده است. همچنین یک سری از عوامل تشکیل‌دهنده خطر وجود دارند که روی این پاسخ اثرگذار می‌باشند، به‌عنوان‌مثال در بیمه شخص ثالث به صورت نمونه برای متغیرهای تبیینی، می‌توان اطلاعات مربوط به‌راننده، وسیله نقلیه بیمه‌شده و محل رانندگی را نام برد. همچنین در این حالت، به دلیل وجود عوامل تشکیل‌دهنده خطری که قابل مشاهده و اندازه‌گیری نمی‌باشند اما بر تعداد تصادف‌ها و همچنین تعداد ادعاهای خسارت تأثیر دارند، نیاز به درنظرگرفتن اثرهای تصادفی در مدل‌های پیشنهادی با پاسخ‌های طولی است.

تاکنون هیچ تحقیقی در زمینه برازش و مقایسه مدل‌های رگرسیون آماسیده صفر برای تعداد ادعای خسارت در طول چند سال صورت نگرفته است. در این مقاله با معرفی مدل‌های آماسیده صفر طولی برای تعداد ادعای خسارت طولی به تشخیص عوامل تشکیل‌دهنده خطر مؤثر روی این پاسخ‌ها می‌پردازیم، همچنین با مقایسه مدل‌های شمارشی در این زمینه، بهترین مدل ممکن روی داده‌های بیمه شخص ثالث معرفی می‌شود.

مروری بر پیشینه پژوهش

برای تحلیل و مدل‌سازی روی تعداد ادعای خسارت، مدل‌های شمارشی همچون مدل‌های رگرسیون پواسون[6] و رگرسیون دوجمله‌ای منفی[7] توسط توماس و سامسون[8] با پاسخ‌های مقطعی مورد استفاده قرار گرفت. پژوهش‌های مختلفی نیز در زمینه محاسبه حق‌بیمه با استفاده از

[1]. Poisson
[2]. Zero-Inflated
[3]. Negative Binomial
[4]. Zero-Inflated Poisson
[5]. Zero-Inflated Negative Binomial
[6]. Poisson Regression
[7]. Negative Binomial Regression
[8]. Thomas and Samson, 1987

توزیع‌های چندمتغیره طولی پواسون و دوجمله‌ای منفی توسط بوچر و همکاران[1] صورت گرفته است. ازآنجایی‌که در تعداد ادعای خسارت ممکن است با صفر زیاد روبه‌رو شویم، نیاز به استفاده از مدل‌های رگرسیون آماسیده صفر برای برازش و تحلیل روی این داده‌هاست.

مدل رگرسیون پواسون آماسیده صفر نیز برای برازش به داده‌های شمارشی با صفر زیاد توسط لمبرت[2] مورد استفاده قرار گرفت. همچنین با درنظرگرفتن توزیع دوجمله‌ای منفی به جای پواسون در مدل‌های آماسیده صفر، هیلبرن[3] مدل دوجمله‌ای منفی آماسیده صفر را معرفی کرد. هال[4] اظهار کرد که این مدل‌ها قابل تعمیم به پاسخ‌های طولی نیز می‌باشند و تحت عنوان مدل رگرسیون آماسیده صفر با پاسخ‌های طولی مطرح شد. این دو مدل کاربردهای بسیاری را روی تحلیل تعداد ادعای خسارت با صفر زیاد ایفا می‌کنند. ییپ و یاو[5] نیز به مقایسه مدل‌های آماسیده صفر و مدل‌های شمارشی رایج در حالت مقطعی پرداخته‌اند. همچنین برای تعیین حق‌بیمه‌ها در بیمه اتومبیل برای تعداد ادعای خسارت در طول چند سال بوچر و گیلن[6] از توزیع‌های آماسیده صفر استفاده کرده‌اند.

چند مفهوم مهم در این تحقیق

در این بخش با مفهوم پاسخ‌های طولی و توزیع‌های آماسیده صفر آشنا شده و سپس به بررسی مدل‌های رگرسیون آماسیده صفر مقطعی و طولی مورد استفاده برای تعداد ادعای خسارت با تعداد صفر زیاد در داده‌های بیمه شخص ثالث می‌پردازیم. از مهم‌ترین این مدل‌ها می‌توان به مدل رگرسیون پواسون آماسیده صفر، مدل رگرسیون دوجمله‌ای منفی آماسیده صفر در حالت‌های مقطعی از زمان و در طول چند سال اشاره نمود.

پاسخ‌های طولی

پاسخ‌هایی هستند که در طول زمان و فضای مشخصی جمع‌آوری می‌شوند که هدف اولیه آن آشکارسازی تغییرات پاسخ در طول زمان و همچنین عامل‌های تأثیرگذار روی این تغییرات است. این پاسخ‌ها می‌تواند تعداد ادعای خسارت‌ها در طول چند سال برای هر بیمه‌گذار باشد. اگر تعداد ادعای خسارت برای یک سال مشخص درنظرگرفته‌شود این پاسخ‌ها به صورت مقطعی خواهند بود اما اگر تعداد ادعای خسارت گزارش‌داده‌شده برای هر بیمه‌گذار در طول چند سال مورد نظر باشد، تعداد ادعای خسارت به صورت طولی درنظرگرفته‌می‌شود.

توزیع‌های سری توانی

X دارای توزیع سری‌های توانی است، هرگاه تابع جرم احتمال آن به صورت زیر باشد:

$$P(X=x)=\frac{a(x)\theta^{x}}{c(\theta)} \qquad x=0,1,\ldots$$

که در آن $a(x)>0$ و $c(\theta)$ تابعی مثبت، متناهی و مشتق‌پذیر از θ است. توزیع‌های پواسون و دوجمله‌ای منفی، نمونه‌ای از توزیع‌های سری توانی می‌باشند.

مدل آماسیده صفر

مدل‌های آماسیده صفر یک پارامتر احتمالی اضافی را برای مقدار صفر که نمی‌تواند به‌طورکامل توسط فرض مدل برآورد شود، معرفی می‌کند و برای مدل‌هایی با بیش پراکنش و صفر زیاد به‌کارمی‌رود که تابع احتمالی آن به صورت زیر است:

$$f_{ZID}(y)=\phi_0 I(y=0)+(1-\phi_0)f_D(y|\theta)$$

- $f_D(y|\theta)$: تابع احتمال از توزیع D با پارامتر θ؛
- $f_{ZID}(y)$: مدل آماسیده صفر از توزیع D با یک پارامتر اضافی ϕ_0 برای احتمال آماسیده صفر.

---

[1]. Boucher et al., 2008
[2]. Lambert, 1992
[3]. Heilbron, 1994
[4]. Hall, 2000
[5]. Yip and Yau, 2005
[6]. Boucher and Guille´n, 2009

مدل رگرسیونی سری‌های توانی طولی

مدل رگرسیون سری‌های توانی آماسیده صفر زمانی به‌کارمی‌رود که مدل سری‌های توانی برازش‌شده، مقدار صفر را کم برآورد کند یا به‌عبارت‌دیگر تعداد صفر زیاد رخ دهد. دراین‌حالت یک پارامتر اضافه مانند ϕ_{it} به مدل اضافه می‌شود و توزیع آمیخته حاصل به‌این‌صورت است که احتمال ϕ_{it} را به رخداد صفر و احتمال $1-\phi_{it}$ را به رخداد توزیع سری‌های توانی می‌دهد. چون در مثال مورد بررسی تعداد صفر زیاد رخ می‌دهد در این بخش به بررسی مدل رگرسیون سری‌های توانی آماسیده صفر طولی پرداخته می‌شود. تابع جرم احتمال سری‌های توانی آماسیده صفر به شرط اثر تصادفی b_i به صورت زیر درنظرگرفته می‌شود:

$$\mathrm{P}(Y_{it}=y_{it})=\begin{cases}\phi_{it}+(1-\phi_{it})\dfrac{a(0)}{c(\lambda_{it})} & y_{it}=0\\[2ex] \phi_{it}(1-\phi_{it})\dfrac{a(y_{it})\lambda_{it}^{y_{it}}}{c(y_{it})} & y_{it}=1,2,\ldots\end{cases}$$

فرض کنید بردار پاسخ $\mathbf{Y}$ شامل داده‌های K گروه مستقل $\mathbf{Y}=(\mathbf{Y}_1^T,\ldots,\mathbf{Y}_K^T)^T$ باشد که $\mathbf{Y}_i=(Y_{i1},\ldots,Y_{iT_i})^T$ است. در این مدل داریم:

$$\boldsymbol{\lambda}_i=(\lambda_{i1},\ldots,\lambda_{iT_i})^T$$
$$\phi=(\phi_{i1},\ldots,\phi_{iT_i})^T$$

در نتیجه مدل رگرسیون سری‌های توانی با پاسخ‌های طولی آماسیده صفر به‌این‌صورت درنظرگرفته می‌شود:

$$Y_{it}\sim ZIPS(\lambda_{it},\phi_{it})$$
$$Log(\boldsymbol{\lambda}_{it})=\mathbf{B}_{it}^{'}\beta_t+\sigma b_i$$
$$Logit(\phi_{it})=\mathbf{G}_{it}^{'}\gamma_t\quad i=1,\ldots,K,\quad t=1,\ldots,T$$

که در آن $\mathbf{B}_{it}$ و $\mathbf{G}_{it}$ بردار مربوط به متغیرهای تبیینی شامل عوامل تشکیل‌دهنده خطر و اثرگذار روی تعداد ادعای خسارت می‌باشند. ازسوی‌دیگر b_i به عنوان اثر تصادفی مربوط به فرد iام به صورت توزیع گاما برای توزیع شرطی پواسون به شرط اثر تصادفی و توزیع بتا برای توزیع دوجمله‌ای منفی به شرط اثر تصادفی درنظرگرفته‌می‌شود (Hausman et al., 1984). اگر $\psi=(\gamma^T,\beta^T,\sigma)^T$ برداری مرکب از پارامترها باشد آنگاه لگاریتم تابع درستنمایی برای مدل رگرسیونی سری‌های توانی آماسیده صفر به‌این‌صورت به‌دست‌می‌آید:

$$LogL_{ZIPS}(\psi;y)=\sum_{i=1}^{K}Log\int_{-\infty}^{+\infty}[\prod_{t=1}^{T_i}Pr(Y_{it}=y_{it}\,|\,b_i)]\mathrm{g}(\mathrm{b}_i)\ \mathrm{b}_i$$

که در آن:

$$\Pr(\mathrm{Y}_{it}=\mathrm{y}_{it}\,|\,\mathrm{b}_i)=[(\phi_{it}+(1-\phi_{it})\frac{a(0)}{c(\lambda_{it})})^{1-u_{it}}\ ((1-\phi_{it})\frac{a(y_{it})\lambda_{it}^{y_{it}}}{c(\lambda_{it})})^{u_{it}}]$$
$$=[(\frac{exp(G_{it}\gamma_t)}{1+exp(G_{it}\gamma_t)}+\frac{1}{1+exp(G_{it}\gamma_t)}\frac{a(0)}{c(exp(B_{it}\beta_t+\sigma\beta_i))})^{1-u_{it}}\times$$
$$(\frac{1}{1+exp(G_{it}\gamma_t)}\frac{a(y_{it})exp(B_{it}\beta_t+\sigma\beta_i)^{y_{it}}}{c(exp(B_{it}\beta_t+\sigma\beta_i))})^{u_{it}}]$$

همچنین u_{it} به‌این‌صورت تعریف می‌شود:

$$u_{it} = \begin{cases} 0 & y_{it} = 0 \\ 1 & y_{it} > 0 \end{cases}$$

به دو دلیل به‌دست‌آوردن برآورد ماکسیمم درستنمایی برای لگاریتم تابع نمایی بالا سخت و پیچیده است و نمی‌توان لگاریتم تابع نمایی ماکسیمم شود.

- دلیل اول: وجود جمله مجموع توابع نمایی (جمله اول در فرمول بالا) که ماکسیمم‌کردن آن بسیار سخت و پیچیده است؛

- دلیل دوم: ازآنجایی‌که در توزیع سری توانی آماسیده صفر، مقادیر صفری که پاسخ Y_{it} اختیار می‌کند از دو منبع می‌باشند، به‌طوری‌که یکی از آنها از توزیع سری توانی تولید شده و دیگری از هیچ توزیع خاصی تولید نمی‌شود بلکه همیشه مقدار آن صفر(صفر مطلق) بوده است، این موضوع سبب می‌شود که نتوان در به‌دست‌آوردن ماکسیمم لگاریتم تابع نمایی، این صفرها را تشخیص داد و این موضوع باعث پیچیدگی و سختی در ماکسیمم‌کردن لگاریتم تابع درستنمایی می‌شود. برای رفع این مشکل از الگوریتم EM مشابه حالت مقطعی استفاده می‌شود. ابتدا متغیر تصادفی Z_{it} به‌این‌صورت تعریف می‌شود: $Z_{it} = 1$ وقتی که Y_{it} صفر (صفر مطلق از توزیع آماسیده صفر) است و $Z_{it} = 0$ وقتی‌که Y_{it} از توزیع سری توانی تولیدشده باشد. لازم به ذکر است که بردار مربوط به متغیر تصادفی $(Z_{1t}, Z_{2t}, \cdots, Z_{nt})'$، به‌عنوان متغیر پنهان و گم‌شده درنظرگرفته می‌شود و باید توسط الگوریتم EM ابتدا این بردار به دست آید و سپس با مشخص شدن منبع صفرهای موجود در تابع درستنمایی، لگاریتم تابع درستنمایی به راحتی ماکسیمم می‌شود. بنابراین تابع درستنمایی برای داده‌های کامل $(\mathbf{y}, \mathbf{z}, \mathbf{b})$ به‌این‌صورت خواهد بود:

$$LogL_c(\psi; \mathbf{y}, \mathbf{z}, \mathbf{b}) = log f(\mathbf{b}; \psi) + log(\mathbf{y}, \mathbf{z} \mid \mathbf{b}; \psi)$$
$$= \sum_{i=1}^{k} log\phi(b_i) + \sum_{i=1}^{k}\sum_{t=1}^{T_i} \{[z_{it}\mathbf{G}_{it}\gamma - log(1 + e^{\mathbf{G}_{it}\gamma})]$$
$$+ (1 - z_{it}) \times [y_{it}(\mathbf{B}_{it}\beta + \sigma b_i) - logc(exp(\mathbf{B}_{it}\beta + \sigma b_i)) + log(a(y_{it})]\}$$

که لگاریتم تابع درستنمایی به دو قسمت مجزا تقسیم می‌شود که قسمت اول تابعی از δ و قسمت دوم تابعی از β می‌باشند. لگاریتم تابع درستنمایی در این حالت ساده‌تر است.

مدل‌های شمارشی برای تعداد ادعای خسارت طولی در بیمه شخص ثالث

در این بخش مدل‌های مطرح‌شده، روی داده‌های مربوط به تعداد ادعای خسارت در بیمه شخص ثالث مورد بررسی قرار می‌گیرد. در این مدل‌ها تعداد ادعای خسارت بیمه‌گذار iام در زمان tام در داده‌های صنعت بیمه کشور ایران برای سال‌های ۱۳۸۸، ۱۳۸۹ و ۱۳۹۰ می‌باشد. همچنین متغیرهای تبیینی مورد علاقه عبارت‌اند از:

سن اتومبیل (x_{i1t})، نوع اتومبیل (x_{i2}) و محل رانندگی (x_{i3}) که در این متغیرها مدل‌های برازش‌شده عبارت‌اند از:

- مدل اول: مدل رگرسیون پواسون

$$Y_{it} \mid b_i \sim Poisson(\lambda_{it})$$
$$log\lambda_{it} = \beta_0 + \beta_{it}x_{i1t} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sigma b_i$$

- مدل دوم: مدل رگرسیون دوجمله‌ای منفی

$$Y_{it} \mid b_i \sim NB(k, p_{it}) \quad p_{it} = \frac{k}{k + \lambda_{it}}$$
$$log\lambda_{it} = \beta_0 + \beta_{it}x_{i1t} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sigma b_i$$

- مدل سوم: مدل رگرسیون پواسون آماسیده صفر

$$Y_{it} \mid b_i \sim ZIP(\phi_{it}, \lambda_{it})$$
$$log\lambda_{it} = \beta_0 + \beta_{it}x_{i1t} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sigma b_i$$

$$\log it\frac{\phi_{it}}{1-\phi_{it}}=\gamma_0+\gamma_{1t}x_{i1t}+\gamma_2x_{i2}+\gamma_3x_{i3}$$

- مدل چهارم: مدل رگرسیون دوجمله‌ای منفی آماسیده صفر

$$Y_i\,|\,b_i \sim ZINB(k,p_{it},\phi_{it}) \quad p_{it}=\frac{k}{k+\lambda_{it}}$$

$$log\lambda_{it}=\beta_0+\beta_{it}x_{i1t}+\beta_2x_{i2}+\beta_3x_{i3}+\sigma b_i$$

$$\log it\frac{\phi_{it}}{1-\phi_{it}}=\gamma_0+\gamma_{1t}x_{i1t}+\gamma_2x_{i2}+\gamma_3x_{i3}$$

در این مدل‌ها b_i اثر تصادفی با توزیع گاما برای توزیع شرطی پواسون و بتا برای دوجمله‌ای منفی درنظرگرفته‌شده است که در آن پاسخ‌های Y_{it} به شرط اثر تصادفی b_i، از هم مستقل درنظرگرفته‌می‌شوند و σ پارامتر مربوط به این اثر تصادفی است. همچنین عوامل تشکیل‌دهنده خطر که قابل مشاهده و اندازه‌گیری نمی‌باشند در b_i قرار می‌گیرند. نحوه برآورد پارامترها با استفاده از نرم‌افزار R و بسته کامپیوتری VGAM است.

روش شناسی پژوهش

فرضیه‌های تحقیق

تعداد ادعای خسارت در داده‌های بیمه شخص ثالث، گروه سن اتومبیل برای سال ۱۳۸۸ در ۱۱رده، (۰-۱، ۱-۲،... ،۱۰-۹ و بیشتر از ۱۰) ، برای سال ۱۳۸۹ در ۱۰ رده(۱-۲، ۲-۳،...،۱۰-۹ و بیشتر از ۱۰) و برای سال ۱۳۹۰ در ۹ رده (۲-۳، ۳-۴،...، ۱۰-۹ و بیشتر از ۱۰)، نوع اتومبیل در ۳ رده(پژو، پراید، پیکان) و محل رانندگی در ۳ رده (کم ترافیک، ترافیک متوسط، پر ترافیک) به عنوان متغیرهای مورد علاقه، در این مقاله مورد بررسی قرار گرفته‌اند. همچنین تعداد ادعای خسارت به عنوان یک متغیر هدف شمارشی درنظرگرفته‌می‌شود، به‌طوری‌که دارای توزیع احتمالی پواسون است. بر اساس این توزیع احتمالی می‌توان مدل‌های رگرسیونی ارائه نمود که ارتباط بین تعداد ادعای خسارت با متغیرهای گروه سن اتومبیل، نوع اتومبیل و محل رانندگی مورد بررسی قرار می‌گیرد.

سؤالات تحقیق

سؤالات تحقیق را می‌توان به این شرح خلاصه نمود:

-آیا می‌توان تعداد ادعاهای خسارت در طول چند سال آینده را بااستفاده از مدل‌های آماسیده صفر پیش‌بینی نمود؟

- مهم‌ترین عوامل مؤثر روی تعداد متوسط ادعای خسارت در داده‌های بیمه شخص ثالث، چه عواملی است؟

- مسئله وجود تعداد صفر زیاد در داده‌های مربوط به تعداد ادعای خسارت بیمه شخص ثالث چیست ؟ چه مشکلاتی در تجزیه‌وتحلیل این داده‌ها ایجاد می‌کند؟ تحلیل این داده‌ها به چه صورت انجام می‌شود؟ همچنین مهم‌ترین عوامل مؤثر روی تعداد متوسط ادعای خسارت با تعداد صفر زیاد در داده‌های بیمه شخص ثالث در طول سال‌های مختلف، چه عواملی است؟

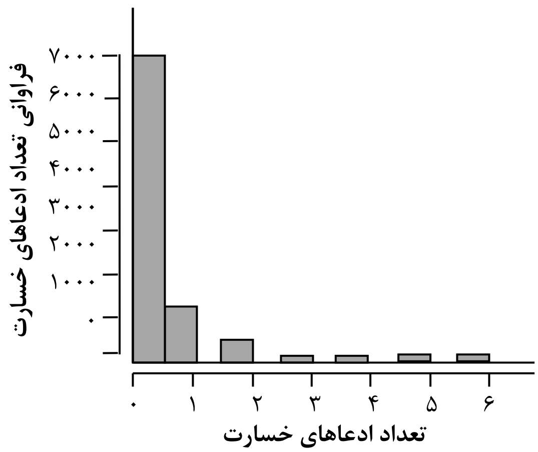
- بررسی مسئله بیش پراکنش روی تعداد ادعای خسارت چیست؟ چه اثری روی تعداد ادعای خسارت دارد؟ تحلیل این داده‌ها به چه صورت انجام می‌شود؟

- چه مدل‌های آماری مناسبی برای متوسط تعداد ادعای خسارت با مسئله بیش پراکنش و تعداد صفر زیاد در داده‌های بیمه شخص ثالث در طول سال‌های مختلف، می‌توان ارائه نمود؟

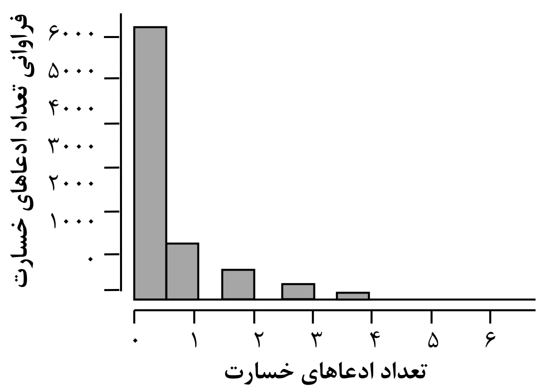
جمع‌آوری داده‌ها و اطلاعات

داده‌های مربوط به بیمه اتومبیل بیمه مرکزی ج.ا.ا در سال ۱۳۸۸، ۱۳۸۹ و ۱۳۹۰ درنظرگرفته‌می‌شود که ۸۵۶۶ فرد را پوشش می‌دهد. متغیر مورد علاقه تعداد مراجعه‌های یک فرد به شرکت بیمه برای گزارش خسارت در طی سه سال و متغیرهای تبیینی نیز مانند: گروه سن اتومبیل برای سال ۱۳۸۸ در ۱۱ رده، (۰-۱، ۱-۲،... ،۱۰-۹ و بیشتر از ۱۰)، برای سال ۱۳۸۹ در ۱۰ رده (۱-۲، ۲-۳،...،۱۰-۹ و بیشتر از ۱۰) و برای سال ۱۳۹۰ در ۹ رده (۲-۳، ۳-۴،...، ۱۰-۹ و بیشتر از ۱۰)، نوع اتومبیل در ۳ رده (پژو، پراید و پیکان) و محل رانندگی در ۳ رده (کم‌ترافیک، ترافیک متوسط و

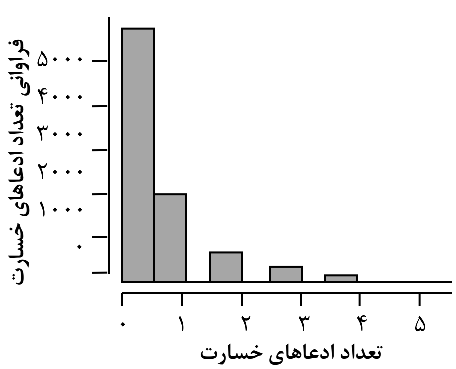
پرترافیک) مورد بررسی قرار گرفته‌اند. در نمودارهای ۱، ۲ و ۳، نمودار میله‌ای فراوانی ادعاهای خسارت برای سال‌های ۱۳۸۸، ۱۳۸۹ و ۱۳۹۰ نشان داده شده است.



نمودار ۱: نمودار میله‌ای فراوانی ادعای خسارت برای سال ۱۳۸۸



نمودار ۲: نمودار میله‌ای فراوانی ادعای خسارت برای سال ۱۳۸۹



نمودار ۳: نمودار میله‌ای فراوانی ادعای خسارت برای سال ۱۳۹۰

در نمودارهای ۱، ۲ و ۳ ملاحظه می‌شود که فراوانی تعداد ادعای خسارت دارای صفر زیاد می‌باشند، همچنین با درنظرگرفتن جدول ۱ نیز اطلاعات مربوط به فراوانی تعداد ادعای خسارت برای سال‌های ۱۳۸۸، ۱۳۸۹ و ۱۳۹۰ ارائه گردیده است.

جدول ۱: فراوانی ادعای خسارت برای سال‌های ۱۳۸۸، ۱۳۸۹ و ۱۳۹۰

تعداد تصادفات	فراوانی ادعاها در سال ۱۳۸۸	فراوانی ادعاها در سال ۱۳۸۹	فراوانی ادعاها در سال ۱۳۹۰
۰	۷۰۲۰	۶۹۱۳	۶۸۴۳
۱	۱۲۹۹	۱۱۷۹	۱۱۶۳
۲	۱۹۵	۲۹۳	۲۲۸
۳	۲۸	۱۰۹	۱۸۸
۴	۱۶	۶۲	۴۰
۵	۵	۷	۳
۶	۲	۳	۲
مجموع	۸۵۶۵	۸۵۶۶	۸۴۶۷

همانطور که ملاحظه می‌شود در سال ۱۳۸۸ درصد کسانی که هیچ ادعای خسارتی نداشته‌اند ۸۲٪، کسانی که یک ادعای خسارت داشته‌اند ۱۵٪، کسانی که دو ادعای خسارت داشته‌اند ۲/۳٪، کسانی که سه ادعای خسارت داشته‌اند ۰/۴٪ و کسانی که چهار و بیشتر از چهار ادعای خسارت داشته‌اند ۰/۳٪ است. در سال ۱۳۸۹ درصد کسانی که هیچ ادعای خسارتی نداشته‌اند ۷۹/۷٪، کسانی که یک ادعا داشته‌اند ۱۳/۷٪، کسانی که دو ادعای خسارت داشته‌اند ۳/۴٪، کسانی که سه ادعای خسارت داشته‌اند ۲/۴٪ و کسانی که چهار و بیشتر از چهار ادعای خسارت داشته‌اند ۰/۸٪ است. در سال ۱۳۹۰ نیز درصد کسانی که هیچ ادعای خسارتی نداشته‌اند ۷۸/۵٪، کسانی که یک ادعا داشته‌اند ۱۵/۵٪، کسانی که دو ادعای خسارت داشته‌اند ۳/۳٪، کسانی که سه ادعای خسارت داشته‌اند ۲/۲٪ و کسانی که چهار و بیشتر از چهار ادعای خسارت داشته‌اند ۰/۵٪ است.

تجزیه‌وتحلیل اطلاعات

در جدول ۲ نیز بر اساس لگاریتم متوسط ادعای خسارت سن اتومبیل برای سال‌های ۱۳۸۸، ۱۳۸۹ و ۱۳۹۰ در سطح معنی‌داری ۰/۱ در مدل‌های رگرسیونی پواسون آماسیده صفر و دوجمله‌ای منفی آماسیده صفر معنی‌دار است. هر چقدر سن اتومبیل بیشتر باشد، لگاریتم متوسط خسارت بیشتر خواهد بود. یعنی سن اتومبیل با متوسط خسارت، رابطه لگاریتمی دارد و هرچقدر سن اتومبیل بیشتر باشد متوسط تعداد خسارت‌ها بیشتر خواهد بود. همچنین نوع اتومبیل نیز در هیچ‌کدام از مدل‌ها معنی‌دار نشده است. ولی محل رانندگی در سطح معنی‌داری ۰/۱ در مدل رگرسیونی پواسون آماسیده صفر و دوجمله‌ای منفی آماسیده صفر معنی‌دار است. هرچقدر محل رانندگی، در جاهای کم ترافیک باشد لگاریتم متوسط خسارت کمتر خواهد بود. اثر تصادفی نیز در سطح معنی‌داری ۰/۰۵ در مدل رگرسیونی پواسون آماسیده صفر و دوجمله‌ای منفی آماسیده صفر معنی‌دار می‌باشد که نشان‌دهنده این است که عواملی در مدل وجود داشته که تحت کنترل و اندازه‌گیری ما نبوده است. بر اساس لوجیت نسبت صفرهای ساختاری نیز سن اتومبیل در سال‌های ۱۳۸۸، ۱۳۸۹ و ۱۳۹۰ و نوع اتومبیل در هیچ سطحی معنی‌دار نبوده و بنابراین در هیچ‌کدام از مدل‌ها اثرگذار نمی‌باشند. اما محل رانندگی در سطح ۰/۰۵ در مدل رگرسیونی پواسون آماسیده صفر و دوجمله‌ای منفی آماسیده صفر معنی‌دار است. هر چقدر محل رانندگی در جاهای کم ترافیک باشد لوجیت نسبت صفرهای ساختاری کمتر خواهد بود. این موضوع به ما نشان می‌دهد که سن اتومبیل و نوع آن در صفرهای ساختاری که به واسطه وقوع خسارت و گزارش‌نکردن آن به‌وجودآمده‌است، تأثیری ندارد و عامل مؤثر در به‌وجودآمدن این نوع صفرها محل رانندگی بوده است. البته ممکن است متغیرهای تبیینی دیگری نیز در به‌وجودآمدن نسبت صفرهای ساختاری نقش داشته باشند ولی در این مقاله فقط اثر این سه متغیر مورد بررسی قرار گرفته است. نتایج به‌دست‌آمده به نحوی توانایی مدل‌های آماسیده صفر را نیز نشان می‌دهد. در مدل رگرسیونی دوجمله‌ای و دوجمله‌ای منفی آماسیده صفر به دلیل اینکه پارامتر پراکنش در سطح

معنی‌داری ۰/۰۵ معنی‌دار است و همچنین به دلیل وجود صفرهای ساختاری زیاد براساس معیارهای AIC و BIC این مدل به عنوان بهترین مدل انتخاب می‌شود؛ یعنی اینکه برازش بهتری نسبت به مدل‌های دیگر به داده‌های تعداد ادعاهای خسارت دارد.

جدول ۲: نتایج برازش مدل‌های رگرسیونی به داده‌ها با پاسخ‌های طولی

مدل	پارامتر	ZINB	ZIP	NB	Poisson
عرض از مبدأ	β_0	−۱/۵۷۰۱ (۰/۱۰۰۶)***	−۰/۷۸۶۲ (۰/۰۸۲۶)***	−۱/۵۶۳۱ (۰/۰۵۲۳)***	−۱/۵۶۲۹ (۰/۰۸۹۱)***
۱۳۸۸ سن اتومبیل (بیشتر از ۱۰سال)	β_{11}	۰/۰۱۱۴ (۰/۰۰۶۸)*	۰/۰۱۴۳ (۰/۰۰۷۳)*	۰/۰۰۵۷ (۰/۰۰۴۸)	۰/۰۰۴۵ (۰/۰۰۳۵)
۱۳۸۹ سن اتومبیل (بیشتر از ۱۰سال)	β_{12}	۰/۰۱۲۷ (۰/۰۰۶۶)*	۰/۰۱۸۳ (۰/۰۰۹۴)*	۰/۰۰۶۹ (۰/۰۰۴۶)	۰/۰۰۵۸ (۰/۰۰۴۱)
۱۳۹۰ سن اتومبیل (بیشتر از ۱۰سال)	β_{13}	۰/۰۱۳۳ (۰/۰۰۷۰)*	۰/۰۱۷۱ (۰/۰۰۸۷)*	۰/۰۰۷۲ (۰/۰۰۵۱)	۰/۰۰۵۱ (۰/۰۰۴۹)
نوع اتومبیل (پژو)	β_2	−۰/۰۰۵۶ (۰/۰۴۲۵)	−۰/۰۲۱۱ (۰/۰۶۲۸)	−۰/۰۳۸۰ (۰/۰۳۱۱)	−۰/۰۳۷۷ (۰/۰۲۸۱)
محل رانندگی (کم ترافیک)	β_3	−۰/۰۱۱۲ (۰/۰۰۵۹)*	−۰/۱۱۲۷ (۰/۰۶۰۴)*	−۰/۰۳۶۸ (۰/۰۳۱۴)	−۰/۰۳۶۸ (۰/۰۲۸۴)
اثر تصادفی	δ	۰/۰۱۴۲ (۰/۰۰۶۶)**	۰/۰۰۳۲ (۰/۰۰۱۵)**	۰/۰۰۲۴ (۰/۰۰۱۹)	۰/۰۰۳۳ (۰/۰۲۶)
عرض از مبدأ	γ_0	−۰/۵۴۲۳ (۰/۱۳۲۵)***	−۱//۲۳۲ (۰/۳۰۵۵)***	−	−
۱۳۸۸ سن اتومبیل (بیشتر از ۱۰سال)	γ_{11}	۰/۱۷۰۶ (۰/۲۱۳۳)	۰/۰۱۷۳ (۰/۰۱۶۴)	−	−
۱۳۸۹ سن اتومبیل (بیشتر از ۱۰سال)	γ_{12}	۰/۱۶۸۶ (۰/۱۰۱۳)	۰/۰۱۳۹ (۰/۰۱۰۶)	−	−
۱۳۹۰ سن اتومبیل (بیشتر از ۱۰سال)	γ_{13}	۰/۱۲۱۶ (۰/۱۱۶۲)	۰/۰۱۲۲ (۰/۰۱۷۳)	−	−
نوع اتومبیل (پژو)	γ_2	۰/۷۹۹۴ (۰/۶۳۴۳)	۰/۰۳۴ (۰/۱۳۰۲)	−	−
محل رانندگی (کم ترافیک)	γ_3	−۰/۹۸۳ (۰/۴۰۱۳)**	−۰/۳۳۸۲ (۰/۱۲۹۹)**	−	−
پارامتر پراکنش		۱/۰۲۵۴ (۰/۰۱۱۴)**	−	۱/۱۳۳۴ (۰/۰۵۶۲)**	−
منفی لگاریتم نمایی		۴۹۱۱/۵۸	۴۹۳۵/۶۱	۴۹۱۷/۲۵	۵۰۰۱/۷۶
معیار AIC		۹۸۴۷/۱۶	۹۸۹۳/۱۲	۹۸۴۸/۵۶	۱۰۰۱۵/۵۲
معیار BIC		۹۸۹۲/۴۱	۹۹۴۱/۱۱	۹۸۸۴/۳۲	۱۰۰۶۳/۲۱

* معنی‌داری در سطح ۰/۱، ** معنی‌داری در سطح ۰/۰۵، *** معنی‌داری در سطح ۰/۰۰۱

عبارت داخل پرانتز انحراف معیار برآورد پارامترهاست.

نتایج و بحث

در این تحقیق به بررسی و مقایسه مدل‌های شمارشی طولی آماسیده صفر برای پاسخ‌های مربوط به تعداد ادعای خسارت در بیمه شخص ثالث پرداخته شد. این مدل‌ها نقش بسیار زیادی در تعیین عوامل تشکیل‌دهنده خطر روی تعداد ادعای خسارت ایفا می‌کنند و بر اساس این مدل‌ها، به شرکت‌های بیمه و آگاهی کافی در مورد عوامل تشکیل‌دهنده خطر می‌دهند به‌طوری‌که برای چند سال آینده تعداد ادعای خسارت مربوط به بیمه

شخص ثالث قابل پیش‌بینی است. در این داده‌ها، سن اتومبیل در طول سال‌های ۱۳۸۸، ۱۳۸۹ و ۱۳۹۰ و محل رانندگی به عنوان عوامل مؤثر و تشکیل‌دهنده خطر روی تعداد ادعای خسارت در این سه سال مطرح شده است.

از این مدل‌ها می توان برای داده‌های بیمه بدنه اتومبیل نیز استفاده کرد.

منابع و ماخذ

Boucher, J.P.; Denuit, M.; Guille´n, M., (2008). Models of insurance claim counts with time dependence based on generalisation of poisson and negative binomial distributions. Journal of the Variance, 2(1), pp. 135–162.

Boucher, J.P.; Guille´n, M., (2009). A survey on models for panel count data with applications to insurance. Journal of the Applied Mathematics, 3(2), pp. 280-281.

Dempster, A.P.; Larid, N.M.; Rubin, D.B., (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Statist, 39, pp. 1-38.

Hall, D.B., (2000). Zero-Inflated poisson and binomial regression with random effects: A case study. Journal of the Biometrics, 56, pp. 1030-1039.

Hausman, J.; Hall, B.; Griliches, Z., (1984). Econometric Models for Count Data with Application to the Patents-R and D Relationship. Econometrica, 52, pp. 909–938.

Heilbron, D., (1994). Zero-Altered and other regression models for count data with added zeros. Journal of the Biometrical , 36, pp. 531–547.

Lambert, D., (1992). Zero-Inflated poisson regression, with an application to defects in manufacturing. Journal of the Technometrics, 34, pp. 1-14.

Thomas, H.; Samson, D., (1987). Linear models as aids in insurance decision making: The estimation of automobile insurance claims. Journal of the Business, 15, pp. 247–256.

Yip, K.C.H.; Yau, K.K.W., (2005). On modeling claim frequency data in general insurance with extra zeros. Journal of the Mathematics and Economics, 36, pp. 153-163.