مدل‌های رگرسیونی هاردل، پوآسون آماسیده و دوجمله‌ای منفی آماسیده برای تحلیل داده‌های شمارشی با صفر زیاد

احسان بهرامی سامانی[۱]

تاریخ دریافت: ۱۳۹۶/۷/۲۳
تاریخ پذیرش: ۱۳۹۷/۶/۲۶

چکیده:

در این مقاله، مدل رگرسیونی هاردل برای پاسخ‌های شمارشی با تعداد صفر زیاد معرفی می‌شود. یک شیوهٔ برآورد ماکسیمم درست‌نمایی برای برآورد پارامترهای مدل استفاده شده است. کاربردی از مدل معرفی شده در داده‌های بیم‌سنجی ارائه شده است. در این مثال، تعداد زیادی ادعای خسارت برابر صفر وجود دارد که کاربرد مدل با پاسخ شمارشی آماسیده صفر را روشن می‌سازد. مدل‌های رگرسیونی شمارشی مختلفی برای مدل‌بندی چنین پاسخ‌های شمارشی در این مقاله معرفی شده‌اند. از جملهٔ این مدل‌های معرفی شده می‌توان به مدل رگرسیونی پوآسون هاردل و مدل رگرسیونی دوجمله‌ای منفی هاردل اشاره کرد.

واژه‌های کلیدی: توزیع آماسیده صفر، پاسخ‌های سری توانی، مدل رگرسیونی هاردل. تعداد ادعای خسارت.

۱ مقدمه

در بسیاری از مطالعه‌ها ممکن است با مجموعه‌ای از داده‌های شمارشی روبرو شویم. به‌عنوان مثال در مطالعه‌های مربوط به علوم بیم‌سنجی، تعداد ادعای خسارت مربوط به بیمه شخص ثالث در یک سال مشخص به‌عنوان یک پاسخ شمارشی از اهمیت زیادی برخوردار است، به‌طوری‌که این پاسخ ممکن است دارای تعداد صفر زیاد باشد. از سویی دیگر بررسی و تحلیل این پاسخ از اهمیت بسیاری در جهت شناسایی عوامل تشکیل‌دهنده خطر دارا است. همچنین به‌واسطه ساختار پیچیده این داده‌ها، تحلیل و مدل‌سازی آنها کار بسیار پیچیده‌ای است که این موضوع سبب شده است بسیاری از محققان به تحلیل داده‌های گسسته روی آورند که برخی از این تحقیقات در ادامه بیان می‌شود.

مدل‌های شمارشی آماسیده صفر روشی را برای مدل‌بندی جرم احتمالی صفر در توزیع متغیرها فرض می‌کند. مدل‌بندی صفرها برای داده‌های شمارشی در حالت تک متغیره مطالعه شده است [۳، ۱]. همچنین رگرسیون پوآسون آماسیده صفر چندسطحی ارائه شده است [۴]. [۲] روش مطرح شده در [۳] را تعدیل داده و همچنین مدل دوجمله‌ای آماسیده صفر و مدل رگرسیونی آماسیده صفر با پاسخ‌های طولی را مطرح کرد. مدل هاردل برای داده‌های شمارشی آماسیده صفر در [۵] مطالعه شده است. [۸] برای مدل‌سازی پاسخ‌های شمارشی از مدل رگرسیونی پوآسون و مدل رگرسیونی دوجمله‌ای منفی برای پاسخ‌های مقطعی استفاده کردند. [۶] با استفاده از توزیع‌های چندمتغیره طولی پوآسون و دوجمله‌ای منفی به محاسبه حق بیمه پرداختند. همچنین [۷] برای تعیین حق بیمه‌ها در بیمه اتومبیل در طول چند سال از توزیع‌های آماسیده صفر استفاده کردند.

در این مقاله، خانواده‌ای از مدل‌های آماسیده و مدل‌های هاردل تحت عنوان مدل‌های رگرسیونی آماسیده صفر و هاردل برای معرفی می‌شود. از آن‌جایی که توزیع‌های دوجمله ای، پوآسون و دوجمله‌ای منفی جز خانواده توزیع‌های سری توانی هستند، خانواده‌ای از مدل‌هایی مانند مدل‌های آماسیده صفر پوآسون و پوآسون هاردل و همچنین مدل‌های آماسیده صفر دوجمله‌ای منفی و دوجمله‌ای منفی هاردل معرفی می‌شوند و برای تشریح توانمندی آنها یک مطالعه شبیه‌سازی انجام می‌گیرد و در

[۱] عضو هیئت علمی گروه آمار دانشگاه شهید بهشتی تهران، تهران، ایران

نهایت این مدل‌ها روی داده‌های بیم‌سنجی برازش داده شده و نتایج مهمی استخراج می‌گردد.

در بخش دوم به معرفی توزیع سری توانی آماسیده صفر و سری توانی هاردل پرداخته می‌شود. در بخش سوم مدل رگرسیونی مربوط به توزیع‌های بیان شده، معرفی می‌گردد و توابع درست‌نمایی آنها معرفی می‌شود. در بخش چهارم مطالعهٔ شبیه‌سازی به‌منظور بررسی عملکرد مدل‌های رگرسیونی بیان شده، انجام می‌شود. در بخش پنجم، کاربرد مدل‌های بیان شده روی داده‌های بیمه بیان می‌شود و در نهایت در بخش ششم به نتیجه‌گیری پرداخته می‌شود.

۲ برخی توزیع‌های آماری آماسیده صفر در تحلیل داده‌های شمارشی

در این بخش با مدل آماسیده صفر و مدل هاردل آشنا شده سپس توزیع سری‌توانی آماسیده صفر معرفی می‌شود.

۱.۲ توزیع آماسیده صفر

یکی از توزیع‌های قابل استفاده برای متغیرهای گسسته شمارشی با تعداد صفر زیاد، توزیع آماسیده صفر (ZI) است. توزیع آماسیده صفر به دلیل زیاد بودن تعداد صفر در مقادیر متغیر گسسته شمارشی یک پارامتر آمیختگی را برای مقدار صفر در نظر می‌گیرند. فرض می‌شود که Y یک متغیر تصادفی گسسته شمارشی با تعداد صفر زیاد باشد. توزیع آماسیده صفر را با نماد $Y \sim ZIF(\pi_{\circ},\theta)$ نشان داده و تابع جرم احتمال آن به‌صورت زیر است:

$$f_{ZIF}(y)=\pi_{\circ}\, I(y=\circ)+(\text{۱}-\pi_{\circ}) f_F(y|\theta)\, I(y>\circ),$$

که در آن $f_F(y|\theta)$ تابع احتمال توزیع F با پارامتر θ، $f_{ZIF}(y)$ توزیع آماسیده صفر از توزیع F با یک پارامتر آمیختگی $\pi_{\circ}$ و $I(y=\circ)$ تابع نشانگر جرم در صفر است.

۲.۲ توزیع هاردل

تابع جرم احتمال این توزیع به‌صورت زیر است:

$$f_{ZIF}(y)=\pi\, I(y=\circ)+(\text{۱}-\pi)\frac{f(y)}{\text{۱}-f(\circ)}\, I(y>\circ).$$

که در آن پارامتر آمیختگی π است، زمانی که متغیر پاسخ مقدار صفر را اختیار کند. همچنین f یک تابع جرم احتمال سری توانی است، زمانی‌که متغیر پاسخ مقدار صحیح و نامنفی اختیار کند.

۳.۲ توزیع سری‌توانی آماسیده صفر

فرض شود که Y یک متغیر تصادفی گسسته شمارشی با تعداد صفر زیاد باشد. توزیع سری‌توانی آماسیده صفر را با نماد $Y \sim ZIPS(\pi,\theta)$ نشان داده و تابع جرم احتمال آن به‌صورت زیر است:

$$f_{ZIPS}(y)=\pi\, I(y=\circ)+(\text{۱}-\pi) f_{PS}(y|\theta)\, I(y=\circ),$$

که در آن

$$f_{PS}(y\mid\theta)=\frac{a(y)\theta^y}{c(\theta)}, \qquad y=\circ,\text{۱},\text{۲},\ldots.$$

یا به‌عبارت دیگر:

$$p(Y=y)=\begin{cases}\pi+(\text{۱}-\pi)\frac{a(\circ)}{c(\theta)}, & y=\circ,\\ (\text{۱}-\pi)\frac{a(y)\theta^y}{c(\theta)}, & y>\circ\end{cases}$$

$a(.)$که در آن تابعی مثبت، $c(.)$ تابعی مثبت، متناهی و مشتق‌پذیر از θ است.

۴.۲ توزیع سری‌توانی هاردل

فرض شود که Y یک متغیر تصادفی گسسته شمارشی با تعداد صفر زیاد باشد. توزیع سری‌توانی هاردل را با نماد $Y \sim HPS(\pi,\theta)$ نشان داده و تابع جرم احتمال آن به‌صورت زیر است

$$f_{HPS}(y)=\pi\, I(y=\circ)+I(y>\circ)(\text{۱}-\pi)\, f_{PS}(y\mid\theta)/f_{PS}(\circ\mid\theta).$$

یا به‌عبارت دیگر

$$p(Y=y)=\begin{cases}\pi, & y=\circ,\\ (\text{۱}-\pi)\frac{a(y)\theta^y}{a(\circ)}, & y>\circ\end{cases}$$

۳ مدل‌های رگرسیونی با پاسخ آماسیده صفر

در این بخش به معرفی مدل‌های رگرسیونی با پاسخ آماسیده صفر پرداخته می‌شود، مدل‌هایی مانند مدل رگرسیونی سری توانی آماسیده صفر و مدل رگرسیونی هاردل از جملهٔ این مدل‌ها هستند.

۱.۳ مدل رگرسیونی سری‌توانی آماسیده صفر

فرض کنید $Y_1,\ldots,Y_n$ یک نمونهٔ تصادفی از توزیع $ZIPS(\pi_i,\theta_i)$ در نظر گرفته شود، از سویی دیگر با در نظر گرفتن بردار پارامترهای $\theta=(\theta_1,\ldots,\theta_n)'$ و $\pi=(\pi_1,\ldots,\pi_n)'$ مدل رگرسیونی سری‌های توانی آماسیده صفر به‌صورت زیر معرفی می‌شود:

$$logit\,(\pi_i)=B_i^{'}\beta,$$

$$\log\,(\theta_i)=W_i^{'}\gamma,$$

که در آن B_i و W_i سطر i-ام ماتریس‌های طرح (بردارهای متغیرهای کمکی) و همچنین β و γ بردار پارامترهای رگرسیونی مدل هستند که باید برآورد شوند. از سویی دیگر برای برآورد پارامترها تابع درست‌نمایی این مدل به‌صورت زیر است:

$$\begin{aligned} log\ L\,(\beta,\gamma,y)&=\sum_{i=1}^{n}(1-I(y_i>0))log[\pi_i+(1-\pi_i)\frac{a(0)}{c(\theta_i)}]\\ &\quad+\sum_{i=1}^{n}I(y_i>0)log[(1-\pi_i)\frac{a(y_i)\theta_i{}^{y_i}}{c(\theta_i)}],\\ &=\sum_{i=1}^{n}(1-I(y_i>0))log[\frac{e^{B_i^{'}\beta}}{1+e^{B_i^{'}\beta}}+\frac{1}{1+e^{B_i^{'}\beta}}\frac{a(0)}{c(e^{W_i^{'}\gamma})}],\\ &\quad+\sum_{i=1}^{n}I(y_i>0)log[\frac{1}{1+e^{B_i^{'}\beta}}\frac{a(y_i)(e^{W_i^{'}\gamma})^{y_i}}{c(e^{W_i^{'}\gamma})}]. \end{aligned}$$

از رایج‌ترین این مدل‌ها می‌توان به مدل رگرسیونی پوآسون آماسیده صفر و مدل رگرسیونی دوجمله‌ای منفی آماسیده صفر اشاره نمود.

۲.۳ مدل رگرسیونی پوآسون آماسیده صفر

وقتی تعداد صفر زیاد برای توزیع پوآسون رخ دهد، تابع جرم احتمال آن به‌صورت توزیع آمیخته در نظر گرفته می‌شود، این توزیع آمیخته یک پارامتر آمیختگی π را به مقدار صفر متغیر گسسته شمارشی y و یک احتمال $(1-\pi)$ را به تابع جرم احتمال پوآسون با پارامتر μ برای مقادیر مثبت پاسخ متغیر گسسته y نسبت می‌دهد که این تابع جرم احتمال به‌صورت زیر است:

$$Pr(Y=y)=\begin{cases}\pi+(1-\pi)e^{-\mu}, & y=0,\\ (1-\pi)\frac{e^{-\mu}\mu^{y}}{y!}, & y=1,2,\ldots\end{cases}$$

تابع درست‌نمایی برای متغیر پاسخ گسسته شمارشی Y که دارای توزیع ZIP است، به‌صورت زیر محاسبه می‌شود.

فرض کنید $Y_1,\ldots,Y_n$ یک نمونهٔ تصادفی از توزیع $ZIP(\pi,\mu)$ باشد، مدل رگرسیون پوآسون آماسیده صفر به‌صورت است

$$logit\,(\pi_i)=B_i^{'}\beta,$$

$$\log\,(\mu_i)=W_i^{'}\gamma,$$

که در آن B_i و W_i سطر i-ام ماتریس‌های طرح (بردارهای متغیرهای کمکی) و همچنین β و γ بردار پارامترهای رگرسیونی مدل هستند که باید برآورد شوند. از سویی دیگر برای برآورد پارامترها تابع درست‌نمایی این مدل به‌صورت زیر است:

$$\begin{aligned} L\,(\mu,\pi|y)=\prod_{i=1}^{n}[\,&\left(\pi_i+(1-\pi_i)\,e^{-\mu_i}\right)I\,(y_i=0)\\ &((1-\pi_i)\frac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!})I\,(y_i\geq 1)\,]. \end{aligned}$$

۳.۳ مدل رگرسیونی دوجمله‌ای منفی آماسیده صفر

فرض کنید $Y_1,\ldots,Y_n$ یک نمونهٔ تصادفی از توزیع $ZINB(\pi,\mu,\frac{\delta}{1+\delta})$ باشد، لذا تابع جرم احتمال آن به‌صورت زیر در نظر گرفته می‌شود:

$$p\,(Y=y)=\begin{cases}\pi+(1-\pi)(\frac{\delta}{1+\delta})^{\mu}, & y=0\\ (1-\pi)\frac{\Gamma(y+\mu)}{y!\Gamma(\mu)}(\frac{\delta}{1+\delta})^{\mu}(\frac{1}{1+\delta})^{y}, & y=1,2,\ldots\end{cases}$$

که در آن δ پارامتر پراکنش نامیده می‌شود.

مدل رگرسیونی دوجمله‌ای منفی آماسیده صفر به‌صورت زیر است:

$$logit\,(\pi_i)=B_i^{'}\beta,$$

$$\log\,(\mu_i)=W_i^{'}\gamma,$$

که در آن B_i و W_i سطر i-ام ماتریس‌های طرح (بردارهای متغیرهای کمکی) و همچنین β و γ بردار پارامترهای رگرسیونی مدل هستند که باید برآورد شوند. از سویی دیگر برای برآورد پارامترها تابع درست‌نمایی این مدل به‌صورت زیر است:

$$L(\mu,\pi|y)=\prod_{i=1}^{n}\left(\pi_i+(1-\pi_i)\left(\frac{\delta}{1+\delta}\right)^{\mu_i}\right)I(y_i=0)$$
$$\times((1-\pi_i)\frac{(y_i+\mu_i)}{y_i!\,(\mu_i)}\left(\frac{\delta}{1+\delta}\right)^{\mu_i}\left(\frac{1}{1+\delta}\right)^{y_i})I(y_i\geq 1).$$

۴.۳ مدل رگرسیونی سری‌توانی هاردل

فرض کنید $Y_1,\ldots,Y_n$ یک نمونهٔ تصادفی از توزیع $HPS(\pi_i,\theta_i)'$ در نظر گرفته شود، از سویی دیگر با در نظر گرفتن بردار پارامترهای $\theta=(\theta_1,\ldots,\theta_n)'$ و $\pi=(\pi_1,\ldots,\pi_n)$ مدل رگرسیون سری‌های توانی هاردل را به‌صورت زیر است

$$logit(\pi_i)=B_i'\beta,$$
$$\log(\theta_i)=W_i'\gamma,$$

که در آن B_i و W_i سطر i -ام ماتریس‌های طرح (بردارهای متغیرهای کمکی) و همچنین β و γ بردار پارامترهای رگرسیونی مدل هستند که می‌بایست برآورد شوند. از سویی دیگر برای برآورد پارامترها تابع درست‌نمایی این مدل به‌صورت زیر است:

$$log\,L(\beta,\gamma,y)=\sum_{i=1}^{n}I(y_i=0)log[\pi_i]$$
$$+\sum_{i=1}^{n}I(y_i>0)log[(1-\pi_i)\frac{a(y_i)\theta_i^{y_i}}{a(0)}],$$
$$=\sum_{i=1}^{n}I(y_i=0)log[\frac{e^{B_i'\beta}}{1+e^{B_i'\beta}}]$$
$$+\sum_{i=1}^{n}I(y_i>0)log[\frac{1}{1+e^{B_i'\beta}}\frac{a(y_i)(e^{W_i'\gamma})^{y_i}}{a(0)}].$$

۴ مطالعه شبیه‌سازی

در این بخش، طی چند مطالعه شبیه‌سازی به بررسی عملکرد مدل‌های بیان شده در بخش سوم پرداخته می‌شود. برای یک مطالعهٔ شبیه‌سازی از مدل‌های رگرسیونی پوآسون آماسیده صفر، مدل رگرسیونی پوآسون هاردل، مدل‌های رگرسیونی دو جمله‌ای منفی آماسیده صفر و در نهایت مدل رگرسیونی دوجمله‌ای منفی هاردل ابتدا از این توزیع‌ها، داده تولید می‌شود. برای این منظور طی یک بررسی شبیه‌سازی برای اندازه‌های نمونه‌ای ۵۰، ۱۰۰ و ۱۰۰۰ به شیوه زیر، نمونهٔ تصادفی تولید می‌شود و این کار را ۱۰۰۰ بار تکرار خواهد شد.

برای تولید داده، مراحل زیر به ترتیب انجام دهید:

الف) مرحله اول: ابتدا یک عدد تصادفی مانند U از توزیع یکنواخت روی بازهٔ (۰،۱) تولید کنید.

ب) مرحله دوم: متغیر تبیینی X را از توزیع نرمال استاندارد تولید کنید و مقایر واقعی پارامترهای مدل‌ها را به‌صورت $(\beta_0,\beta_1)=$ (۰،۰٫۵) ، $(\gamma_0,\gamma_1)=(0,0.5)$ و $\delta=1$ در نظر بگیرید.

ج) مرحله سوم: متغیر پاسخ را بر اساس یکی از چهار مدل زیر تولید کنید:

ج-۱-مدل اول: مدل رگرسیونی پوآسون آماسیده صفر:

$$Y\sim ZIP(\mu,\pi),$$
$$\mu=\exp(\gamma_0+\gamma_1X),$$
$$logit(\pi)=\beta_0+\beta_1X.$$

ج-۲-مدل دوم: مدل رگرسیونی پوآسون هاردل:

$$Y\sim HurdlePois(\mu,\pi),$$
$$\mu=\exp(\gamma_0+\gamma_1X),$$
$$logit(\pi)=\beta_0+\beta_1X.$$

ج-۳-مدل سوم: مدل رگرسیونی دوجمله‌ای منفی آماسیده صفر:

$$Y\sim ZINB\left(\pi,\mu,\frac{\delta}{1+\delta}\right),$$
$$\mu=\exp(\gamma_0+\gamma_1X),$$
$$logit(\pi)=\beta_0+\beta_1X.$$

ج-۴-مدل چهارم: مدل رگرسیونی دوجمله‌ای منفی هاردل:

$$Y\sim HurdleNB\left(\pi,\mu,\frac{\delta}{1+\delta}\right),$$
$$\mu=\exp(\gamma_0+\gamma_1X),$$
$$logit(\pi)=\beta_0+\beta_1X.$$

جدول ۱. نتایج شبیه‌سازی مربوط به برآورد پارامترهای چهار مدل بیان شده

		$n=۵۰$		$n=۱۰۰$		$n=۱۰۰۰$	
پارامتر	مقدار واقعی	برآورد	انحراف معیار	برآورد	انحراف معیار	برآورد	انحراف خطا
ZIP							
$\beta_۰$	۰/۰۰۰	۰/۰۷۲	۰/۸۷۱	۰/۰۶۴	۰/۷۷۲	۰/۰۰۸	۰/۰۳۴
$\beta_۱$	۰/۵۰۰	۰/۴۳۰	۰/۷۸۷	۰/۵۱۰	۰/۶۲۴	۰/۵۰۳	۰/۰۲۵
$\gamma_۰$	۰/۰۰۰	۰/۰۸۹	۰/۶۲۵	۰/۰۷۶	۰/۴۲۳	۰/۰۶۳	۰/۰۶۳
$\gamma_۱$	۰/۵۰۰	۰/۴۰۰	۰/۶۲۳	۰/۵۳۰	۰/۵۶۶	۰/۴۹۰	۰/۰۴۵
HurdlePois							
$\beta_۰$	۰/۰۰۰	۰/۰۲۵	۰/۲۱۰	۰/۰۱۹	۰/۲۰۰	۰/۰۰۸	۰/۰۴۲
$\beta_۱$	۰/۵۰۰	۰/۴۰۱	۰/۱۹۹	۰/۴۸۵	۰/۱۸۷	۰/۴۹۸	۰/۰۵۶
$\gamma_۰$	۰/۰۰۰	۰/۱۵۰	۰/۲۳۲	۰/۱۲۹	۰/۱۱۲	۰/۰۲۶	۰/۰۷۸
$\gamma_۱$	۰/۵۰۰	۰/۴۸۰	۰/۲۶۵	۰/۴۹۲	۰/۱۰۵	۰/۴۹۵	۰/۰۴۵
ZINB							
$\beta_۰$	۰/۰۰۰	−۰/۳۲۸	۰/۱۷۹	−۰/۰۲۵	۰/۰۹۵	−۰/۰۲۰	۰/۰۲۵
$\beta_۱$	۰/۵۰۰	۰/۴۱۸	۰/۲۳۹	۰/۴۵۹	۰/۱۳۸	−۰/۰۱۷	۰/۰۳۵
$\gamma_۰$	۰/۰۰۰	−۰/۲۵۵	۰/۱۸۵	−۰/۰۲۰	۰/۱۱۲	−۰/۰۰۲	۰/۰۴۵
$\gamma_۱$	۰/۵۰۰	۰/۴۴۲	۰/۲۸۱	۰/۴۴۸	۰/۰۸۸	۰/۴۹۵	۰/۰۵۶
δ	۱/۰۰۰	۱/۲۵۰	۰/۲۱۶	۱/۱۲۵	۰/۱۰۲	۱/۰۲۶	۰/۰۵۵
HurdleNB							
$\beta_۰$	۰/۰۰۰	۰/۱۴۹	۰/۲۱۶	۰/۰۲۶	۰/۱۲۳	۰/۰۲۷	۰/۰۷۹
$\beta_۱$	۰/۵۰۰	۰/۴۶۴	۰/۲۲۱	۰/۴۸۰	۰/۰۸۷	۰/۴۹۹	۰/۰۳۱
$\gamma_۰$	۰/۰۰۰	۰/۱۵۰	۰/۱۳۲	۰/۰۲۹	۰/۱۰۲	۰/۰۱۶	۰/۰۵۸
$\gamma_۱$	۰/۵۰۰	۰/۴۸۲	۰/۱۶۴	۰/۴۹۶	۰/۱۰۵	۰/۴۹۸	۰/۰۴۵
δ	۱/۰۰۰	۰/۸۱۷	۰/۲۱۷	۰/۸۷۶	۰/۱۰۲	۰/۹۲۱	۰/۰۵۲

د) مرحله چهارم: اگر $U < \pi$، آن‌گاه مقدار متغیر پاسخ Y صفر در نظر گرفته می‌شود، در غیر این صورت برای تولید پاسخ از توزیع پوآسون برای مدل اول، پوآسون بریده شده در صفر برای مدل دوم، توزیع دوجمله‌ای منفی برای مدل سوم و درنهایت توزیع دوجمله‌ای منفی بریده شده در صفر برای مدل چهارم استفاده می‌شود.

نتایج شبیه‌سازی مربوط به برآورد پارامترهای چهار مدل بیان شده در جدول ۱ گردآوری شده است. با توجه به نتایج جدول ۱، هر چهار مدل توانایی مناسبی در برآورد پارامترهای مربوط به مدل خود را دارند و با افزایش اندازهٔ نمونه مقدار برآورد پارامترهای مدل در هر چهار مدل به مقدار واقعی پارامتر نزدیک می‌شود. از سوی دیگر برآورد پارامترهای این چهار مدل برآوردهای سازگاری هستند.

۵ تحلیل داده‌های بیم‌سنجی

تعداد ادعای خسارت در داده‌های بیمه شخص ثالث، گروه سن اتومبیل برای سال ۱۳۸۶ در ۲ رده، (۰-۷ و بیشتر از ۷)، نوع اتومبیل در ۳ رده (پژو، پراید، پیکان) و محل رانندگی در ۳ رده (کم‌ترافیک، ترافیک متوسط، پرترافیک) به‌عنوان متغیرهای

مورد علاقه، در این مقاله مورد بررسی قرار گرفته‌اند. تعداد ادعای خسارت به‌عنوان یک متغیر شمارشی درنظر گرفته می‌شود، به‌طوری‌که برای متغیر تعداد ادعای خسارت توزیع‌های بیان شده در بخش دوم، در نظر گرفته می‌شود. در جدول ۲ متغیرهای تبیینی معرفی شده‌اند. در جدول ۳ اطلاعات مربوط به فراوانی تعداد ادعای خسارت برای سال‌های ۱۳۸۶ ارائه گردیده است. همان‌طور که ملاحظه می‌شود در سال ۱۳۸۶ کسانی که هیچ ادعای خسارتی نداشته‌اند، ۷۲ درصد؛ کسانی که یک ادعای خسارت داشته‌اند، ۱۴ درصد؛ کسانی که دو ادعای خسارت داشته‌اند، ۶ درصد؛ کسانی که سه ادعای خسارت داشته‌اند، ۴ درصد؛ کسانی‌که بیش از ۳ ادعای خسارت داشته‌اند، ۴ درصد بوده است.

۱.۵ مدل‌های برازش شده روی داده‌ها

در این بخش مدل‌های بیان شده در بخش سوم روی داده‌های تعداد ادعای خسارت برازش داده می‌شود. این مدل‌ها به شرح زیر هستند:

۱-مدل اول: در این مدل برای تعداد ادعای خسارت در بیمه شخص ثالث (پاسخ گسسته) مدل رگرسیونی پوآسون آماسیده صفر در نظر گرفته می‌شود:

$$Y \sim ZIP\left(\mu,\pi\right),$$

$$\mu= \exp(\gamma_۰+\gamma_۱x_۱+\gamma_۲x_{۱۱}+\gamma_۳x_{۱۲}+\gamma_۴x_{۲۱}+\gamma_۵x_{۲۲}),$$

$$logit(\pi) =\beta_۰+\beta_۱x_۱+\beta_۲x_{۱۱}+\beta_۳x_{۱۲}+\beta_۴x_{۲۱}+\beta_۵x_{۲۲}.$$

۲-مدل دوم: در این مدل برای تعداد ادعای خسارت در بیمه شخص ثالث (پاسخ گسسته) مدل رگرسیونی هاردل پوآسونی در نظر گرفته می‌شود:

$$Y \sim HurdlePois\left(\mu,\pi\right),$$

$$\mu= \exp(\gamma_۰+\gamma_۱x_۱+\gamma_۲x_{۱۱}+\gamma_۳x_{۱۲}+\gamma_۴x_{۲۱}+\gamma_۵x_{۲۲}),$$

$$logit(\pi) =\beta_۰+\beta_۱x_۱+\beta_۲x_{۱۱}+\beta_۳x_{۱۲}+\beta_۴x_{۲۱}+\beta_۵x_{۲۲}.$$

۳-مدل سوم: در این مدل برای تعداد ادعای خسارت در بیمه شخص ثالث (پاسخ گسسته) مدل رگرسیونی دوجمله‌ای منفی آماسیده صفر در نظر گرفته می‌شود:

$$Y \sim ZINB\left(\pi,\mu,\frac{\delta}{۱+\delta}\right),$$

$$\mu= \exp(\gamma_۰+\gamma_۱x_۱+\gamma_۲x_{۱۱}+\gamma_۳x_{۱۲}+\gamma_۴x_{۲۱}+\gamma_۵x_{۲۲}),$$

$$logit(\pi) =\beta_۰+\beta_۱x_۱+\beta_۲x_{۱۱}+\beta_۳x_{۱۲}+\beta_۴x_{۲۱}+\beta_۵x_{۲۲}.$$

۴-مدل چهارم: در این مدل فرض می‌شود برای تعداد ادعای خسارت در بیمه اتومبیل (پاسخ گسسته) مدل رگرسیونی هاردل دوجمله‌ای منفی در نظر گرفته می‌شود:

$$Y \sim HurdleNB\left(\pi,\mu,\frac{\delta}{۱+\delta}\right),$$

$$\mu= \exp(\gamma_۰+\gamma_۱x_۱+\gamma_۲x_{۱۱}+\gamma_۳x_{۱۲}+\gamma_۴x_{۲۱}+\gamma_۵x_{۲۲}),$$

$$logit(\pi) =\beta_۰+\beta_۱x_۱+\beta_۲x_{۱۱}+\beta_۳x_{۱۲}+\beta_۴x_{۲۱}+\beta_۵x_{۲۲}.$$

۲.۵ نتایج

نتایج برآورد پارامترهای مدل‌های ZIP و HurdlePois در جدول ۴ گردآوری شده است. سن ماشین در سطح معنی داری ۰/۰۵ در هر دو مدل در سال ۸۶ روی تعداد ادعای خسارت معنی‌دار است. هرچه‌قدر سن ماشین بیشتر باشد، در این صورت لگاریتم متوسط تعداد ادعای خسارت بیشتر خواهد بود، یعنی سن اتومبیل با متوسط تعداد ادعای خسارت رابطهٔ لگاریتمی دارد و هرچه‌قدر سن اتومبیل بیشتر باشد متوسط تعداد خسارت‌ها بیشتر خواهد بود. همچنین نوع اتومبیل نیز در هیچکدام از مدل‌ها معنی دار نشده است، ولی محل رانندگی در سطح معنی داری ۰/۰۵ در این دو مدل معنی دار است. بر اساس لوجیت نسبت صفرهای زیاد مربوط به تعداد ادعای خسارت، سن اتومبیل و نوع اتومبیل در هیچ سطحی معنی دار نبوده و بنا بر این در هیچ‌کدام از مدل‌ها اثرگذار نیستند. اما محل رانندگی در سطح ۰/۰۵ در این دو مدل معنی دار می‌باشد. هر چه‌قدر محل رانندگی در جاهای کم ترافیک باشد لوجیت نسبت صفرهای زیاد مربوط به تعداد ادعای خسارت کمتر خواهد بود.

جدول ۲. متغیرهای تبیینی مورد استفاده

متغیرهای تبیینی	سن اتومبیل	نوع اتومبیل: پژو	نوع اتومبیل: پراید	محل رانندگی: کم ترافیک	محل رانندگی: ترافیک متوسط
سال۸۶	$x_{۱}$	$x_{۱۱}$	$x_{۱۲}$	$x_{۲۱}$	$x_{۲۲}$

این موضوع به ما نشان می‌دهد که سن اتومبیل و نوع آن در صفرهای زیاد مربوط به تعداد ادعای خسارت که به واسطهٔ وقوع خسارت و گزارش نکردن آن به وجود آمده است، تأثیری ندارد و عامل مؤثر در به وجود آمدن این نوع صفرها محل رانندگی بوده است. نتایج به‌دست‌آمده به نحوی توانایی مدل‌های آماسیده صفر و هاردل را نیز نشان می‌دهد.

نتایج برآورد پارامترهای مدل‌های ZINB و HurdleNB در جدول ۵ گردآوری شده است. نتایج یکسانی نیز در این دو مدل نسبت به دو مدل قبل رخ می‌دهد. در مدل‌های رگرسیونی هردو مدل ZINB و HurdleNB به دلیل این‌که پارامتر پراکنش در سطح معنی داری ۰/۰۵ معنی دار است و همچنین به دلیل وجود صفرهای ساختاری زیاد بر اساس معیارهای AIC این مدل به‌عنوان بهترین مدل انتخاب می‌شود یعنی این‌که برازش بهتری نسبت به مدل‌های دیگر به داده‌های تعداد ادعاهای خسارت دارد.

۶ بحث و نتیجه‌گیری

در این مقاله به معرفی مدل رگرسیونی آماسیده و هاردل سری توانی پرداخته شد و انواع این مدل‌ها مورد بررسی قرار گرفت. این مدل‌ها نقش عمده‌ای در تعیین عوامل مؤثر بر تعداد ادعای خسارت و شدت خسارت ایفا کرده و شرکت‌های بیمه با استفاده از این مدل‌ها می‌توانند عوامل اثرگذار روی تعداد ادعای خسارت را شناسایی کرده و تعداد ادعای خسارت و حق بیمه در چند سال آینده را به‌راحتی پیش‌بینی کنند. در این داده‌ها، عواملی همچون نوع اتومبیل، سن اتومبیل و مکان رانندگی به‌عنوان عوامل اثرگذار روی تعداد ادعای خسارت و شدت خسارت مطرح شده است. مدل‌های مذکور را می‌توان با در نظر گرفتن گمشدگی در داده‌های تعداد ادعای خسارت بررسی کرد.

جدول ۳. درصد فراوانی ادعای خسارت برای سال‌های ۱۳۸۶

تعداد تصادفات	درصد فراوانی ادعاها در سال ۱۳۸۶
۰	۷۲
۱	۱۴
۲	۶
۳	۴
۴	۲
۵	۱
۶	۱
۷	۰
مجموع	۱۰۰

جدول ۴. نتایج مربوط به مدل‌های مورد نظر($HurdlePois,\ ZIP$)

مدل	ZIP		$HurdlePois$	
	برآورد	انحراف معیار	برآورد	انحراف معیار
μ				
عرض از مبدأ	۰/۱۸۱	۰/۴۱۵	۰/۲۰۷	۱/۸۹۰
سن ماشین	**۱.۰۲۹**	**۰/۰۷۷**	**۱.۰۳۰**	**۰/۵۱۷**
نوع ماشین: پژو	۱/۱۲۸	۱/۲۲۳	۱/۱۳۰	۱/۵۱۹
نوع ماشین: پراید	۱/۰۷۳	۱/۱۴۶	۱/۰۷۳	۱/۰۳۲
محل رانندگی: کم ترافیک	**۱.۵۸۸**	**۰.۴۲۳**	**۱.۵۸۸**	**۱.۰۷۶**
محل رانندگی: ترافیک متوسط	**۱.۶۵۷**	**۰.۴۴۶**	**۱.۶۲۹**	**۱.۰۸۸**
π				
عرض از مبدأ	−۰/۳۴۰	۰/۳۳۴	−۰/۴۰۳	۰/۳۰۹
سن ماشین	۰/۲۶۸	۰/۲۴۴	۰/۲۳۴	۰/۲۱۵
نوع ماشین: پژو	۰/۶۴۷	۰/۶۱۲	۰/۶۶۵	۰/۳۳۷
نوع ماشین: پراید	۰/۷۳۴	۰/۵۰۴	۰/۵۰۹	۰/۵۵۱
محل رانندگی: کم ترافیک	−۰/۶۴۳	۰/۰۰۱	−۰/۶۱۱	۰/۰۰۱
محل رانندگی: ترافیک متوسط	−۰/۵۴۳	۰/۱۰۳	−۰/۴۱۱	۰/۰۰۱
معیار AIC	۱۸۶۵/۳۲۳		۱۸۵۷/۱۴۷	

جدول ۵. نتایج مربوط به مدل‌های مورد نظر($HurdleNB,\ ZINB$)

مدل	$ZINB$		$HurdleNB$	
	برآورد	انحراف معیار	برآورد	انحراف معیار
μ				
عرض از مبدأ	−۲/۳۴۸	۱/۲۷۶	−۱/۰۷۹	۱/۸۷۹
سن ماشین	**۲.۹۰۲**	**۰.۶۲۳**	**۱.۲۴۱**	**۰.۵۱۷**
نوع ماشین: پژو	۰/۲۶۶	۰/۱۵۹	۱/۲۳۹	۱/۵۱۹
نوع ماشین: پراید	۱/۸۰۰	۱/۴۰۳	۱/۲۸۹	۱/۰۳۲
محل رانندگی: کم ترافیک	**۱/۴۲۷**	**۰/۰۴۵**	**۱/۸۷۶**	**۰/۰۶۴**
محل رانندگی: ترافیک متوسط	**۱/۲۰۵**	**۰/۰۳۰**	**۲/۲۳۰**	**۰/۰۸۲**
π				
عرض از مبدأ	−۰/۵۴۲	۰/۱۳۵	−۰/۴۵۳	۰/۱۰۸
سن ماشین	۰/۲۷۰	۰/۲۱۳	۰/۲۵۸	۰/۱۴۹
نوع ماشین: پژو	۰/۷۰۹	۰/۶۳۴	۰/۶۷۹۴	۰/۴۳۶۷
نوع ماشین: پراید	−۰/۸۹۰	۰/۰۰۱	−۰/۸۰۴	۰/۰۰۱
محل رانندگی: کم ترافیک	−۰/۸۷۵	۰/۱۰۳	−۰/۷۹۹	۰/۱۰۰
محل رانندگی: ترافیک متوسط	۱/۰۲۰	۰/۰۲۱	۱/۱۳۰	۰/۰۱۶
معیار AIC	۸۱۷/۲۸۹		۸۰۴/۱۱۲	

مراجع

[1] Greene, W. (2005). Functional form and heterogeneity in models for count data, *Foundations and Trends in Econometrics*, **1**(**2**), 113-218.

[2] Hall, D. B. (2000). Zero-Inflated Poisson and binomial regression with random effects: A case study, *Biometrics*, **56**, 1030-1039.

[3] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1-14.

[4] Lee, A., Wang, K., Scott, J., Yan, K., and McLachlan, G. (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros", *Statistical Methods in Medical Research*, **15**, 47-61.

[5] Min, Y. and Agrestic, A. (2005). Random effect models for repeated measures of zero inflated count data, *Statistical Model*, **5**, 5-19.

[6] Boucher, J.,P., Denuit, M. and Guillee' n, M. (2008). Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions, *Journal of the Variance*, **1**(**2**), 135-162.

[7] Boucher, J. P. and Guillee' n, M. (2009). A survey on models for panel count data with applications to insurance, *Journal of the Applied Mathematics*, **3**(**2**), 280-281.

[8] Thomas, H. and Samson, D. (1987). Linear models as aids in insurancedecision making: The estimation of automobile insurance claims, *Journal of the Business*, **15**, 247-256 .