

برازش مدل‌های سری‌زمانی شمارشی روی تعداد مراجعین به مراکز ترک اعتیاد در شهرستان سمنان

امید کریمی*، فاطمه حسینی و سپیده تشریفی

دانشگاه سمنان

چکیده. داده‌های شمارشی برحسب زمان در خیلی از زمینه‌های کاربردی مشاهده می‌شوند. خیلی از پژوهشگران برای تحلیل این داده‌ها از الگوهای سری‌زمانی استفاده می‌کنند. در این مقاله، مدل‌های خطی سری‌زمانی شمارشی پواسن و دوجمله‌ای منفی روی این نوع از داده‌ها با حضور متغیرهای توضیحی مورد مطالعه قرار می‌گیرند. تحلیل درست‌نمایی و ارزیابی مدل سری‌زمانی شمارشی براساس مدل‌های خطی تعمیم‌یافته ارائه و بررسی می‌شود و همچنین داده‌های تعداد معتادین مراجعه کننده به مراکز ترک اعتیاد شهرستان سمنان براساس مدل‌ها و روش‌های بیان شده مورد تحلیل قرار می‌گیرد. با توجه به نتایج حاصل از ملاک‌های انتخاب مدل، مدل سری‌زمانی شمارشی دوجمله‌ای منفی برازش مناسبی برای این داده‌ها است و براساس آن پیش‌بینی ماهانه برای سال ۹۸ صورت گرفت.

واژگان کلیدی: مدل شکنندگی، مدل خطرهای متناسب کاکس، داده بقا، سانسور، مدت زمان بیکاری.

۱- مقدمه

داده‌های سری‌زمانی شمارشی، مانند تعداد ماهانه یک بیماری خاص، تعداد روزهای بارانی در هفته، تعداد معتادین، تعداد مراجعات و غیره در بسیاری از زمینه‌های کاربردی علوم پزشکی، علوم اجتماعی و فیزیکی مشاهده می‌شوند. یک مدل مناسب برای تحلیل داده‌های سری‌زمانی شمارشی، استفاده از مدل رگرسیون پواسن تحت

* نویسنده‌ی عهده‌دار مکاتبات

دریافت: ۱۳۹۹/۵/۸، پذیرش: ۱۴۰۰/۷/۱۴.

مدل‌های خطی تعمیم‌یافته (GLM)^۱ است [۸]. فوکیانوس [۵] مدل رگرسیون پواسون برای داده‌های سری‌زمانی شمارشی را مورد مطالعه قرار داد و روش ماکسیمم درستنمایی برای برآورد پارامترهای مدل به‌کار گرفت.

فرلند و همکاران [۴] مانایی مدل‌های اتورگرسیو شرطی با مقادیر صحیح را مورد تحلیل قرار دادند. کریستو و فوکیانوس [۱] مدل‌های خطی تعمیم‌یافته را با تابع پیوند لگ خطی بررسی کردند و لیبوشیک و همکاران [۷] اثرات مداخله‌ای را در مدل‌های سری‌زمانی دوجمله‌ای منفی در نظر گرفتند. مدل‌های سری‌زمانی شمارشی دارای مشاهدات صحیح نامنفی هستند و باید همبستگی بین آن‌ها به درستی شناسایی شود. یک روش مناسب و منعطف، به‌کارگیری روش GLM نلدر و ودربرن [۹]، برای مدل کردن مشاهدات به صورت شرطی روی اطلاعات گذشته می‌باشد. این روش با انتخاب یک توزیع مناسب برای داده‌های شمارشی و یک تابع پیوند مناسب، پیاده‌سازی می‌شود [۴]، [۲]. در این مقاله به تحلیل مدل‌های سری‌زمانی شمارشی پرداخته می‌شود، سپس این مدل‌ها روی داده‌های ترک اعتیاد شهرستان سمنان پیاده‌سازی می‌شوند.

ساختار مقاله به این صورت است که در بخش ۲ مدل‌های سری‌زمانی شمارشی معرفی و مورد مطالعه قرار می‌گیرند. همچنین جزئیات مدل‌های خطی و لگ خطی سری‌زمانی شمارشی بیان می‌شود. در بخش ۳ داده‌های واقعی، تعداد معتادین مراجعه‌کننده به مراکز ترک اعتیاد شهرستان سمنان براساس مدل‌های ارائه شده مورد بررسی قرار می‌گیرد. در انتها بحث و نتیجه‌گیری ارائه می‌شود.

۲- مدل‌های سری‌زمانی شمارشی

۲-۱- مدل‌سازی رگرسیون پواسون

توزیع پواسون معمولاً برای مدل‌بندی نرخ رویدادهای تصادفی که در برخی از فاصله‌های زمانی ثابت رخ می‌دهند، استفاده می‌شود. اگر فرض شود λ نشان دهنده نرخ ورود است؛ متغیر تصادفی Y نشان دهنده نرخ ورودی‌ها در یک فاصله زمانی ثابت است و دارای توزیع پواسون با تابع چگالی احتمال زیر است:

$$P[Y = y] = \frac{\exp(-\lambda) \lambda^y}{y!}, y = 0, 1, 2$$

میانگین و واریانس Y هر دو برابر λ است. در اکثر مسائل کاربردی، داده‌های شمارشی معمولاً با برخی اطلاعات متغیر کمکی مشاهده می‌شود [۸]. به‌طور کلی فرض کنید که X_1, \dots, X_p متغیر رگرسیونی مشاهده شده همراه با متغیر پاسخ شمارشی Y با توزیع پواسون است. در آن صورت یک مدل رگرسیونی برای ارتباط بین میانگین متغیر پاسخ با متغیرهای کمکی به صورت $\lambda = \beta_0 + \sum_{i=1}^p \beta_i X_i$ است که در آن پارامتر λ به‌خاطر این‌که میانگین توزیع پواسون است، باید مثبت باشد. یک انتخاب ساده برای مدل‌سازی رگرسیون داده‌های شمارشی، مدل لگ خطی به صورت $\log(\lambda) = \beta_0 + \sum_{i=1}^p \beta_i X_i$ است. هر دو مدل متعلق به کلاس مدل‌های خطی تعمیم یافته‌اند که توسط نلدر و ودربرن [۹] معرفی شده است. در تعریف مدل‌های سری زمانی شمارشی از مدل اتورگرسیو مرتبه یک $AR(1)$ به صورت

$$(1) \quad Y_t = \beta_1 Y_{t-1} + \epsilon_t,$$

استفاده می‌شود. که در آن $|\beta_1| < 1$ و $\{\epsilon_t\}$ دنباله متغیرهای تصادفی نرمال با میانگین صفر و واریانس σ^2 (دنباله نوفه سفید) است. این یک مدل استاندارد برای تجزیه و تحلیل سری‌های زمانی است که مقدار فرآیند در زمان t بسته به مقدار فرآیند در زمان $t-1$ به همراه یک خطای تصادفی است [۱۰].

۲-۱-۱- مدل‌های خطی برای سری‌های زمانی شمارشی

فرض کنید که $\{Y_t\}$ یک سری زمانی شمارشی را نشان می‌دهد. مدل اتورگرسیو پواسون برای فرآیند پاسخ $\{Y_t\}$ به صورت

$$(2) \quad Y_t | \mathcal{F}_{t-1}^{y, \lambda} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = \beta_0 + \beta_1 Y_{t-1}, \quad t \geq 1,$$

تعریف می‌شود که در آن $\mathcal{F}_t^{y,\lambda} = \sigma(Y_s, \lambda_s, s \leq t)$ سیگما میدان تولید شده به وسیله $\{Y_0, \dots, Y_t, \lambda_0\}$ است. β_0, β_1 پارامترهای نامنفی و $\{\lambda_t\}$ به عنوان میانگین فرآیند $\{Y_t\}$ با توجه به گذشته آن است. مقادیر مثبت β_0 و β_1 تضمین می‌کند که $\lambda_t > 0$ است، به خاطر این که Y_t یک عدد صحیح نامنفی است. با توجه به این نکات، واضح است که مدل (۲) با مولفه تصادفی توزیع پواسون و مؤلفه سیستماتیک $\lambda_t = \beta_0 + \beta_1 Y_{t-1}$ و تابع ربط همانی در چارچوب مدل‌های خطی تعمیم‌یافته قرار می‌گیرد. مدل (۲) همان پویایی مدل (۱) را نشان می‌دهد، بنابراین داریم:

$$(۳) \quad Y_t = \lambda_t + (Y_t - \lambda_t) = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t, \quad t \geq 1,$$

که در آن $\epsilon_t = Y_t - \lambda_t$ نوفه سفید است، چون ناهمبسته با میانگین صفر و واریانس ثابت است. البته با فرض مانایی Y_t به راحتی می‌توان نشان داد:

$$E(\epsilon_t) = E(Y_t - \lambda_t) = E(E(Y_t - \lambda_t | \mathcal{F}_t)) = 0,$$

$$\begin{aligned} \text{Var}(\epsilon_t) &= \text{Var}(E(\epsilon_t | \mathcal{F}_t)) + E(\text{Var}(\epsilon_t | \mathcal{F}_t)) = E(\lambda_t) \\ &= E(Y_t), \end{aligned}$$

$$\text{Cov}(\epsilon_t, \epsilon_{t+k}) = E(\epsilon_t \epsilon_{t+k}) = E\left(\epsilon_t E(\epsilon_{t+k} | \mathcal{F}_{t+k-1})\right) = 0.$$

با توجه به معادله (۳) و با فرض مانایی داریم:

$$\text{Var}(\epsilon_t) = E(Y_t) = \frac{\beta_0}{1 - \beta_1},$$

بنابراین باید پارامتر $0 < \beta_1 < 1$ باشد.

حالت کلی مدل (۲) با مرتبه‌ی (p, q) به صورت زیر تعریف می‌شود:

$$(۴) \quad \begin{aligned} Y_t | \mathcal{F}_{t-1}^{y,\lambda} &\sim \text{Poisson}(\lambda_t), \\ \lambda_t &= \beta_0 + \sum_{k=1}^p \beta_k Y_{t-k} + \sum_{\ell=1}^q \alpha_\ell \lambda_{t-\ell}, \quad t \geq \max(p, q), \end{aligned}$$

که در آن مانایی مرتبه دوم با شرط $1 < \sum_{k=1}^p \beta_k + \sum_{l=1}^q \alpha_l < 1$ فراهم می‌شود. برای اطلاعات بیشتر به فرلند و همکاران [۴] مراجعه شود. مدل (۲) به مدل واریانس ناهمسانی شرطی اتورگرسیو صحیح مقدار 2 NGARCH معروف است. با در نظر گرفتن تابع پیوند کانونی $\nu_t = \log(\lambda_t)$ خانواده مدل‌های لگ خطی اتورگرسیو به صورت

$$Y_t | \mathcal{F}_{t-1}^{\nu, \nu} \sim \text{Poisson}(\lambda_t),$$

$$(5) \quad \nu_t = \beta_0 + \sum_{k=1}^p \beta_k \log(Y_{t-k} + 1) + \sum_{\ell=1}^q \alpha_\ell \nu_{t-\ell}$$

حاصل می‌شود که در آن $\mathcal{F}_t^{\nu, \nu}$ سیگما میدان تولید شده از $\{Y_0, \dots, Y_t, \nu_0\}$ است. به‌طور کلی، پارامترهای $\beta_0, \beta_k, \alpha_\ell$ می‌توانند مثبت یا منفی باشند. اما باید شرایط خاصی روی پارامترها مشابه مدل قبلی در نظر گرفت تا یک سری زمانی مانا به‌دست آید، برای جزئیات بیشتر در مورد آن به [۱] مراجعه شود. با در نظر گرفتن بردار پارامترهای مدل

$$\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_p, \alpha_1, \dots, \alpha_q)',$$

لگاریتم تابع درست‌نمایی مدل بدون مقدار ثابت را می‌توان به صورت

$$l(\boldsymbol{\theta}) = \sum_{t=1}^n (Y_t \log(\lambda_t(\boldsymbol{\theta})) - \lambda_t(\boldsymbol{\theta})),$$

نوشت که دارای شکل بسته‌ای برای محاسبه برآورد پارامترها نیست و باید از الگوریتم‌های عددی برای به‌دست آوردن برآورد ماکسیمم درست‌نمایی استفاده کرد. در عمل گاهی اوقات با متغیرهای کمکی که روی متغیر پاسخ تأثیرگذارند روبه‌رو می‌شویم. در آن صورت مدل (۵) به صورت زیر بیان می‌شود:

$$\nu_t = \beta_0 + \sum_{k=1}^p \beta_k \log(Y_{t-k} + 1) + \sum_{\ell=1}^q \alpha_\ell \nu_{t-\ell} + \boldsymbol{\eta}^T \mathbf{X}_t,$$

که در آن بردار متغیر کمکی r بعدی $X_t = (X_{t,1}, \dots, X_{t,r})^T$ می باشد و $\eta = (\eta_1, \dots, \eta_r)^T$ بردار پارامتر ضرائب متغیرهای کمکی است. برای نحوه برازش و پیش بینی در این نوع مدل ها به [۴] مراجعه شود. یک حالت دیگر برای پاسخ های شمارشی استفاده از توزیع دو جمله ای منفی است. در توزیع دو جمله ای منفی، واریانس شرطی می تواند از میانگین λ_t بزرگتر باشد که به عنوان یک جایگزین برای وقتی که داده ها بیش پراکنش دارند، معرفی می شود. کریستو و فوکیانوس [۱] مدل دو جمله ای منفی را به صورت

$$Y_t | \mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi)$$

بیان کردند، که بر حسب پارامتر میانگین λ_t و پارامتر پراکنندگی $\phi \in (0, \infty)$ است. تابع جرم احتمال توزیع دو جمله ای منفی به صورت

$$P(Y_t = y | \mathcal{F}_{t-1}) = \frac{\Gamma(\phi + y)}{\Gamma(y + 1) \Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda_t}\right)^\phi \left(\frac{\lambda_t}{\phi + \lambda_t}\right)^y, \\ y = 0, 1, \dots$$

تعریف می شود که در آن $\text{Var}(Y_t | \mathcal{F}) = \lambda_t + \lambda_t^2 / \phi$ است. یعنی، واریانس شرطی با توان دوم λ_t افزایش می یابد. توزیع پواسون یک حالت خاص دو جمله ای منفی است که در آن $\phi \rightarrow \infty$ میل می کند. در توزیع دو جمله ای منفی λ_t مشابه توزیع دو جمله ای مدل بندی می شود. یکی از مسائل مهم در تحلیل داده های واقعی این است که کدام مدل برای این داده ها مناسب است. بنابراین ملاک هایی برای ارزیابی مدل های ارائه شده نیاز است که در ادامه بیان می شود و از این ملاک ها برای تشخیص مدل مناسب در تحلیل داده های ترک اعتیاد شهرستان سمنان استفاده شده است.

۲-۲- ارزیابی مدل

ابزارهایی که برای مدل های خطی تعمیم یافته و نیز برای سری زمانی توسعه یافته اند، می توانند برای ارزیابی برازش مدل و عملکرد پیشگویی آن مورد استفاده قرار گیرند. در کلاس سری های زمانی شمارشی و مدل های خطی تعمیم یافته باید ارزیابی پیشگویی خطی و انتخاب تابع پیوند برای توزیع شرطی، انجام شود. معیارهای ارائه شده در این

بخش، انتخاب مدل مناسب برای یک مجموعه داده را فراهم می‌کند. این معیارها در بخش بعدی روی مجموعه داده واقعی تعداد معتادین مراجعه کننده به مراکز ترک اعتیاد شهرستان سمنان برای انتخاب مدل مناسب استفاده می‌شوند.

با در نظر گرفتن مقادیر برازش برای هر مدل به صورت $\hat{\lambda}_t = \lambda_t(\hat{\theta})$ ، باقیمانده مدل نیز به صورت $r_t = y_t - \hat{\lambda}_t$ به دست می‌آیند. همانند مدل‌های خطی این باقیمانده‌ها برای ارزیابی مدل برازش شده به کار می‌روند.

تابع خودهمبستگی تجربی مانده‌ها برای تشخیص همبستگی‌های سریالی استفاده می‌شوند که توسط مدل برازش شده بیان نشده‌اند. شکل این مانده‌ها در برابر زمان می‌تواند تغییرات فرآیند تولید داده‌ها روی زمان را آشکار سازد. به علاوه، شکل توان

دوم باقیمانده‌ها r_t^2 در برابر مقادیر برازش شده $\hat{\lambda}_t$ رابطه‌ی میانگین و واریانس را نمایش می‌دهد. اگر نقاط، اطراف تابع همانی یا همان خط نیمساز قرار گیرند، توزیع پواسون مناسب است و اگر یک رابطه‌ی درجه دو مشاهده شود، توزیع دوجمله‌ای منفی مناسب است، [۱۱].

فرض کنید $P_t(y) = P(Y_t \leq y | \mathcal{F}_{t-1})$ تابع توزیع تجمعی، $p_t(y) =$

$v_t = \sqrt{\text{Var}(Y_t | \mathcal{F}_{t-1})}$ و $y \in N_{\mathbb{0}}$ تابع احتمال برای $P(Y_t = y | \mathcal{F}_{t-1})$

انحراف استاندارد توزیع پیشگویی باشد که یا دارای توزیع پواسون با میانگین $\hat{\lambda}_t$ و یا دارای توزیع دوجمله‌ای منفی با میانگین $\hat{\lambda}_t$ و ضریب بیش‌پراکنش $\frac{1}{\phi} = \sigma^2$ است.

یک معیار ارزیابی کالبدن احتمالی^۳ برای توزیع پیشگو، تبدیل انتگرال احتمال^۴ (PIT) است که اگر توزیع پیشگو صحیح باشد؛ دارای توزیع یکنواخت است. برای داده‌های شمارشی چادو و همکاران [۲] معیار PIT را برای مقدار مشاهده شد y_t و

توزیع پیشگویی $P_t(y)$ به صورت

$$(۶) \quad F_t(u|y) = \begin{cases} 0 & u \leq P_t(y-1) \\ \frac{u - P_t(y-1)}{P_t(y) - P_t(y-1)} & P_t(y-1) < u < P_t(y) \\ 1 & u \geq P_t(y) \end{cases}$$

تعریف کرده‌اند که میانگین PIT به صورت زیر محاسبه می‌شود:

$$\bar{F}(u) = \frac{1}{n} \sum_{t=1}^n F_t(u|y_t), \quad 0 \leq u \leq 1.$$

برای بررسی این که آیا $\bar{F}(u)$ ، تابع توزیع تجمعی یک توزیع یکنواخت است، چادو و همکاران [۲] ترسیم یک شکل ستونی با H ستون را پیشنهاد کردند که ستون h دارای ارتفاع

$$f_h = \bar{F}(h/H) - \bar{F}\left(\frac{h-1}{H}\right), \quad h = 1, \dots, H$$

است. مقدار H به طور پیش فرض 10 انتخاب می شود. شکل U حاکی از کم پراکنش توزیع پیش گو است در حالی که شکل U وارونه نشان دهنده ی بیش پراکنش است. کالبدن حاشیه ای^۵ به صورت اختلاف میانگین تابع توزیع تجربی مشاهدات به ازای هر $y \in R$ به صورت زیر تعریف می شود:

$$(۷) \quad \frac{1}{n} \sum_{t=1}^n P_t(y) - \frac{1}{n} \sum_{t=1}^n \mathbb{I}(y_t \leq y).$$

در عمل کالبدن حاشیه ای برای مقادیر y در حوزه مقادیر مشاهدات رسم می شود. اگر مدل مناسب باشد، توزیع حاشیه ای پیش بینی ها، شبیه توزیع حاشیه ای مشاهدات است و (۷) باید نزدیک صفر باشد. انحراف عمده از صفر نشان دهنده نامناسب بودن مدل است. نایتینگ و همکاران [۶] نشان دادند که ارزیابی مدل توسط شکل ستونی PIT یا شکل کالبدن حاشیه ای یک شرط لازم و نه کافی برای ایده آل بودن پیش گویی است. آن ها با مدلی موافق هستند که حداکثر شفافیت را در میان تمام مدل های پیشنهادی دارد. شفافیت، تمرکز توزیع پیش بینی است و به وسیله پهنای بازه های پیش بینی اندازه گیری می شود. یک ارزیابی هم زمان کالبدن و شفافیت که در یک امتیاز عددی خلاصه شود، با قوانین امتیازدهی مناسب قابل انجام است، [۶]. تعدادی از قوانین امتیازدهی در جدول ۱ ارائه شده اند. هر یک از قوانین امتیازدهی ویژگی های مختلف توزیع پیش بینی و فاصله آن با داده های مشاهده شده را اندازه گیری می کنند. به جز میانگین توان دوم خطای نرمال شده، مدل با پایین ترین امتیاز ترجیح داده می شود. یعنی مدلی که پایین ترین مقدار را از معیارهای امتیازدهی جدول ۱ اختیار کند مناسب است. امتیاز میانگین توان دوم خطا تنها موردی است که به توزیع بستگی ندارد و با عنوان میانگین توان دوم خطای پیشگویی شناخته می شود. میانگین

توان دوم خطاهای نرمال شده، واریانس مانده‌های پیرسون را اندازه‌گیری می‌کند. اگر مدل مناسب باشد خطای نرمال شده NSE (جدول ۱) نزدیک به یک است. امتیازدهی دیوید-سباستینی نوع دیگری با یک جمله‌ی اضافی برای جبران بیش‌برآورد انحراف استاندارد است.

جدول ۱- قوانین امتیازدهی برای انتخاب مدل مناسب

معیارهای امتیازدهی	
$MSE = \frac{1}{n} \sum_{t=1}^n r_t^2$	میانگین توان دوم خطاها
$NSE = \frac{1}{n} \sum_{t=1}^n \left(\frac{r_t}{v_t}\right)^2$	میانگین توان دوم خطاهای نرمال شده
$DS = \frac{1}{n} \sum_{t=1}^n \left(\left(\frac{r_t}{v_t}\right)^2 + \log(p_t(y_t)) \right)$	امتیازدهی دیوید-سباستینی
$LOG = \frac{\sum_{t=1}^n \log(p_t(y_t))}{n}$	امتیازدهی لگاریتمی
$QUD = -\frac{1}{n} \sum_{t=1}^n \left(\sum_{y=0}^{\infty} p_t(y) - \sum_{y=0}^{\infty} p_t^*(y) \right)$	امتیازدهی درجه دوم
$SPH = -\frac{1}{n} \sum_{t=1}^n \left(p_t(y_t) / \sum_{y=0}^{\infty} p_t^*(y) \right)$	امتیازدهی کروی
$RP = \frac{1}{n} \sum_{t=1}^n \left(\sum_{y=0}^{\infty} (P_t(y) - I(y_t \leq y)) \right)$	امتیازدهی رتبه احتمال

ابزارهای دیگر انتخاب مدل، معیار اطلاع آکائیکه (AIC) و معیار اطلاع بیزی (BIC) هستند. مدل با کمترین مقدار معیار اطلاع مناسب است. معیارهای اطلاع به صورت

$$AIC = -2\tilde{\ell}(\hat{\theta}, \hat{\sigma}^2) + 2df$$

و

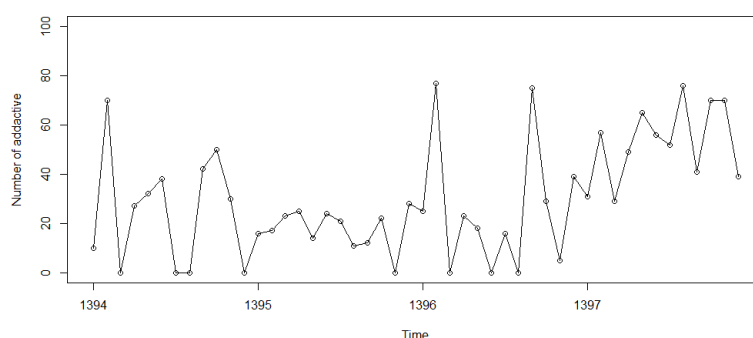
$$BIC = -2\tilde{\ell}(\hat{\theta}, \hat{\sigma}^2) + \log(n_{eff}) df$$

محاسبه می‌شوند که در آن $\tilde{\ell}(\hat{\theta}, \hat{\sigma}^2) = \sum_{t=1}^n \log(p_t(y_t))$ تعداد کل پارامترها (شامل ضریب پراکندگی) و n_{eff} تعداد مشاهدات موثر هستند. به‌طور کلی BIC نسبت به AIC مدل‌های خلاصه‌تری را انتخاب می‌کند. در ادامه روش‌ها و مدل‌های ارائه شده روی داده‌های ترک اعتیاد شهرستان سمنان مورد تحلیل قرار می‌گیرد.

۳- داده‌های ترک اعتیاد شهرستان سمنان

اطلاعات و داده‌های این بخش با همکاری اداره بهزیستی استان سمنان گردآوری شده است. داده‌ها با توجه به مراجعین افراد معتاد به یکی از مواد مخدر به مراکز ترک اعتیاد در شهرستان سمنان طی مدت چهار سال از ماه فروردین سال ۱۳۹۴ تا اسفند ماه سال ۱۳۹۸ به صورت ماهانه جمع‌آوری شده‌اند. مراکز ترک اعتیاد به دو صورت سرپایی و اقامتی می‌باشد که در این مقاله آمار مراکز سرپایی شهرستان سمنان مورد بررسی قرار گرفته است. برای در نظر گرفتن اثر زمان، مدل رگرسیونی روی مشاهدات قبلی لحاظ شده است. اثر فصلی آن با رگرسیون بر λ_{12} یعنی ۱۲ واحد زمانی گذشته (یک سال) تعیین شده‌اند. عوامل بی‌شماری در اعتیاد تاثیرگذار می‌باشد و به همین نسبت برخی عوامل نیز مانع و یا انگیزه‌ای برای ترک اعتیاد است. گروه معتادین مورد مطالعه زنان و مردان می‌باشند. عوامل تاثیرگذار به این شرح هستند: سن، وضعیت شغلی، میزان تحصیلات، وضعیت تأهل، سن شروع مصرف مواد، تعداد دفعاتی که اقدام به ترک کرده است. در مورد وضعیت شغلی، دو گروه شاغل و بیکار در نظر گرفته شده است. میزان تحصیلات گروه مورد بررسی بازه‌ی بی‌سواد تا دکتری را در بر گرفته که آن در دو طبقه‌ی دارای سواد زیر دیپلم و بالای دیپلم گروه‌بندی شده‌اند. همچنین افراد ممکن است، زندگی مشترک را تجربه کرده باشند و بعد از آن به هر دلیلی طلاق گرفته و یا جدا زندگی کنند. در این بررسی تمامی افرادی که مجرد بوده و یا به هر دلیلی جدا از همسرشان زندگی می‌کنند در یک گروه قرار گرفته‌اند و وضعیت تاهل افراد معتاد در دو صورت متاهل و غیرمتاهل لحاظ شده‌اند. داده‌ها با روش‌های ارائه شده مورد تحلیل و بررسی قرار گرفته‌اند و برای انجام این‌کار از نرم‌افزار R بسته‌ی tscount استفاده شده است. تعداد معتادین به مراکز ترک اعتیاد در یک ماه به

عنوان متغیر پاسخ در نظر گرفته شده است. شکل سری زمانی تعداد معتادین مراجعه کننده جهت ترک در یک بازه‌ی چهار ساله در شکل ۱ رسم شده است. ابتدا مدل‌های پواسونی و دوجمله‌ای منفی با توجه به متغیرهای توضیحی برازش شدند، سپس با ملاک‌های معرفی شده در بخش ۲ بهترین مدل برای برازش به داده‌ها انتخاب می‌شود. نتایج حاصل از برازش مدل پواسونی در جدول ۲ خلاصه شده‌اند.



شکل ۱- تعداد معتادین مراجعه‌کننده جهت ترک اعتیاد

جدول ۲- برآورد ضرایب مدل پواسون

فاصله اطمینان ۹۵ درصد	خطای استاندارد	برآورد	پارامترهای مدل
$(۰٫۳-، ۰٫۷۵)$	۰٫۲۶	۰٫۲۳	β (عرض از مبدأ)
$(-۰٫۰۰۵، ۰٫۰۴)$	۰٫۰۲	-۰٫۰۰۷	$(Y_t - 1)\beta_1$
$(-۰٫۰۳، ۰٫۰۴)$	۰٫۰۲	۰٫۰۰۷	$(\lambda_{t-1})\beta_{12}$
$(۰٫۰۴، ۰٫۰۷)$	۰٫۰۰۸	۰٫۰۵	سن (X_1)
$(۰٫۰۱، ۰٫۰۴)$	۰٫۰۰۶	۰٫۰۲	وضعیت تحصیلات (X_2)
$(-۰٫۰۳، -۰٫۰۰۶)$	۰٫۰۰۶	-۰٫۰۲	وضعیت تأهل (X_3)
$(-۰٫۰۵، ۰٫۴۳۳)$	۰٫۱۲	۰٫۱۹	تعداد فرزندان (X_4)
$(-۰٫۰۵، ۰٫۰۱۳)$	۰٫۰۵	۰٫۰۴	بعد خانوار (X_5)
$(۰٫۰۱، ۰٫۰۰۳)$	۰٫۰۰۵	۰٫۰۲	وضعیت شغلی (X_6)
$(-۰٫۰۱، ۰٫۰۰۷)$	۰٫۰۰۴	-۰٫۰۰۱	وضعیت مصرف سیگار (X_7)

بنابراین مدل حاصل برابری است با:

$$M_1: \log(\lambda_t) = 0.23 - 0.007Y_{t-1} + 0.007\lambda_{t-12} + 0.05X_{1t} + 0.02X_{2t} - 0.02X_{3t} + 0.19X_{4t} + 0.04X_{5t} + 0.02X_{6t} - 0.001X_{7t}, \quad t = 1, \dots, 48.$$

که در آن متغیرهای توضیحی X_1, \dots, X_7 در جدول ۲ مشخص شده‌اند. با توجه به فواصل اطمینان ۹۵ درصد جدول ۲ متغیرهای سن، میزان تحصیلات، وضعیت تاهل و وضعیت شغلی تأثیر معنی‌داری روی تعداد مراجعات معتادین به مراکز ترک اعتیاد شهرستان سمنان دارند. نتایج برازش مدل دوجمله‌ای منفی در جدول ۳ خلاصه شده است. معیارهای AIC و BIC در این مدل به ترتیب برابر با ۲۴۴/۰۸۱ و ۲۶۲/۷۹ است که در آن متغیرهای وضعیت شغلی و مصرف سیگار معنی‌دار شده‌اند.

جدول ۳- برآورد ضرایب مدل دوجمله‌ای منفی

پارامترهای مدل	برآورد	خطای استاندارد	فاصله اطمینان ۹۵ درصد
β_0	3.42×10^{-7}	3.93×10^{-4}	$(-7.7 \times 10^{-4}, -7.7 \times 10^{-4})$
β_1	1.01×10^{-11}	8.27×10^{-6}	$(1.62 \times 10^{-5}, 1.62 \times 10^{-5})$
β_{12}	2.88×10^{-10}	1.73×10^{-5}	$(-3.39 \times 10^{-5}, 3.39 \times 10^{-5})$
سن	6.89×10^{-3}	۰/۱۳۷	$(-9.08 \times 10^{-4}, 0.66)$
میزان تحصیلات	۰/۳۳	۰/۱۷	$(-0.844, 0.51)$
وضعیت تاهل	۰/۲۱	۰/۱۵	$(-0.844, 0.51)$
تعداد فرزندان	۰/۴۷	۱/۶۱	$(-2.69, 3.63)$
بعد خانوار	1.28×10^{-6}	۱/۴۱	$(-2.76, 2.76)$
وضعیت شغلی	۰/۴۴	۰/۱۵	$(0.14, 0.73)$
وضعیت مصرف سیگار	۰/۳۱	۰/۱۳	$(0.49, 0.57)$
σ^2	۰/۳۹		

مدل برازش شده به صورت

$$\begin{aligned}
 M_4: \lambda_t = & 3/42 \times 10^{-7} + 1/01 \times 10^{-11} Y_{t-1} \\
 & + 2/88 \times 10^{-10} \lambda_{t-12} + 6/89 \times 10^{-3} X_{1t} \\
 & + 0/33 X_{2t} + 0/21 X_{3t} + 0/46 X_{4t} + 0/13 X_{5t} \\
 & + 0/44 X_{6t} - 0/31 X_{7t}, \quad t = 1, \dots, 48,
 \end{aligned}
 \tag{8}$$

است. از آنجایی که متغیر وضعیت شغلی در هر دو مدل معنی‌دار بود، مدل‌های پواسونی و دوجمله‌ای منفی فقط برای این متغیر هریک به صورت جداگانه به داده‌ها برازش شدند. نتایج برازش مدل پواسونی با متغیر وضعیت شغلی در جدول ۴ خلاصه شده است.

جدول ۴- برآورد ضرایب مدل پواسونی با متغیر وضعیت شغلی

فاصله اطمینان ۹۵ درصد	خطای استاندارد	برآورد	پارامترهای مدل
(۲,۱۴۵, ۲,۵۲۷)	۰/۰۹۷	۲,۳۳	β_0
(-۰/۰۰۵, ۰/۰۶۹)	۰/۰۱۹	۰/۰۳۲	β_1
(-۰/۰۶۴, ۰/۰۲۱)	۰/۰۲۱	-۰/۰۰۲	β_{12}
(۰/۰۳۰, ۰/۰۳۵)	۰/۰۰۱	۰/۰۳۲	وضعیت شغلی

مدل پواسونی با متغیر وضعیت شغلی به صورت

$$\begin{aligned}
 M_5: \log(\lambda_t) = & 2/33 + 0/032 Y_{t-1} - 0/02 \lambda_{t-12} + 0/032 X_{6t}, \\
 & t = 1, \dots, 48,
 \end{aligned}$$

است. جدول ۵ نتایج برازش مدل دوجمله‌ای منفی با متغیر وضعیت شغلی را نشان می‌دهد.

جدول ۵- برآورد ضرایب دوجمله‌ای منفی با متغیر وضعیت شغلی

فاصله اطمینان ۹۵ درصد	خطای استاندارد	برآورد	پارامترهای مدل
(۲,۱۴۵, ۲,۵۲۷)	$2/29 \times 10^{-4}$	7×10^{-8}	β_0
($-1/06 \times 10^{-5}$, $1/06 \times 10^{-5}$)	$5/42 \times 10^{-6}$	59×10^{-11}	β_1
(-۰/۰۱۲, ۰/۰۳۹)	۰/۰۱۲۹	۰/۰۱۳۵	β_{12}
(۱/۱, ۱/۳۸)	۷,۲۳	۱,۲۴	وضعیت شغلی
		۰/۰۷۲	σ^2

همچنین مدل برازش شده به صورت

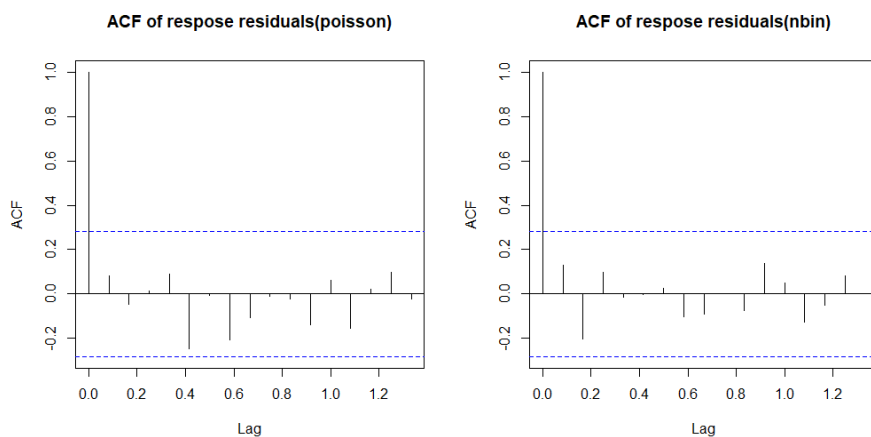
$$M_4: \lambda_t = 7 \times 10^{-8} + 5.9 \times 10^{-11} Y_{t-1} + 0.135 \lambda_{t-12} + 1.24 X_{6t}, \quad t = 1, \dots, 48$$

به دست می‌آید. نتایج معیارهای AIC ، BIC و MSE به دست آمده از برازش مدل‌های ذکر شده در جدول ۶ خلاصه شده است.

جدول ۶- نتایج معیارهای AIC ، BIC و MSE به دست آمده از برازش مدل‌ها

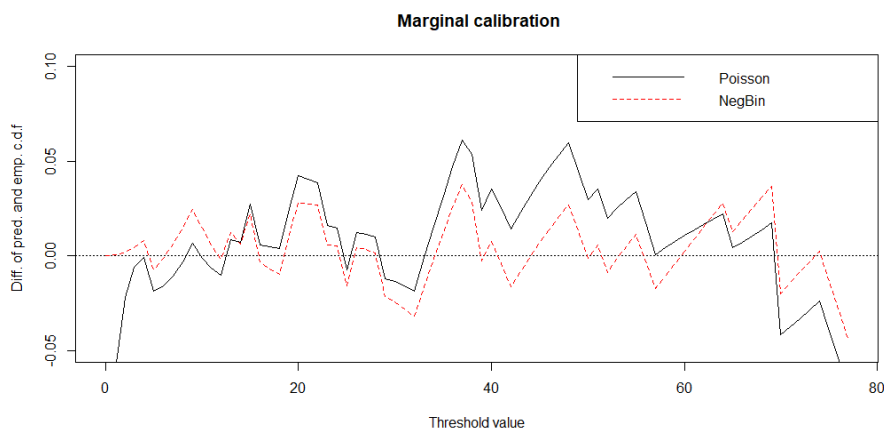
مدل‌ها	AIC	BIC	MSE
M_1	۳۱۱،۴۹	۳۳۰،۲۰	۳۴،۷۸
M_2	۲۴۴،۰۸۱	۲۶۲،۷۹	۱۱،۵۵
M_3	۳۱۱،۴۹	۳۳۰،۲۰	۸۸،۶۴
M_4	۳۰۱،۶۶	۳۱۱،۰۲۵	۳۳،۱۹

با توجه به جدول ۶ همه معیارها مدل دوجمله‌ای منفی M_4 را تأیید می‌کنند، که این مدل برازش بهتری در مقایسه‌ی مدل پواسونی دارد. برای اطمینان بیشتر، از معیارهای دیگر نیز استفاده شده است که در ادامه بررسی می‌شوند. شکل ۲ مانده پاسخ‌ها را برای هر دو مدل پواسونی و دوجمله‌ای منفی نشان می‌دهد. با توجه به تابع خودهمبستگی، هیچ خودهمبستگی یا اثر فصلی مشاهده نمی‌شود.



شکل ۲- شکل خودهمبستگی باقیمانده‌های دو مدل پواسنی (M_1) و دو جمله‌ای منفی (M_2)

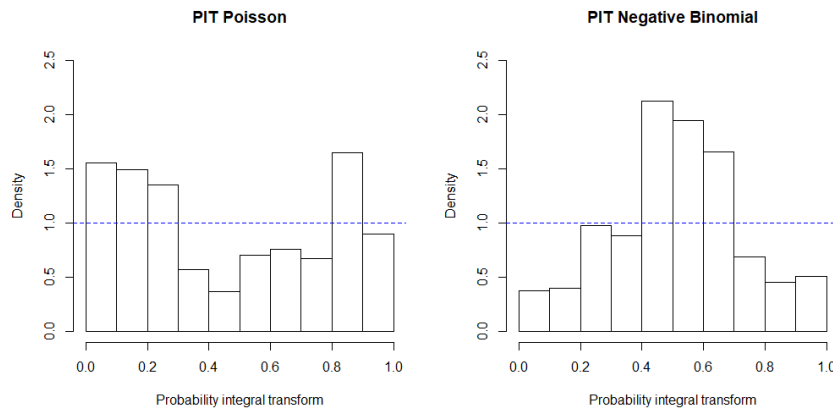
شکل کالبدین حاشیه‌ای در شکل ۳ برای دو مدل رسم شده‌اند، با توجه به فرمول ۶ هرچه این مقادیر به صفر نزدیکتر باشند، برازش بهتر را بیان می‌کنند. در اکثر مواقع شکل دو جمله‌ای منفی به خط صفر نزدیکتر است.



شکل ۳- شکل کالبدین حاشیه‌ای دو مدل پواسنی (M_1) و دو جمله‌ای منفی (M_2)

شکل ۴ مقادیر PIT برای هر دو برازش مدل را نشان می‌دهد. همان‌طور که مشاهده می‌شود، شکل پواسونی U شکل است و پراکنش کمتر را نشان می‌دهد. برخلاف آن

شکل ستونی دوجمله‌ای منفی شکل U برعکس است و بیش‌پراکنش را نشان می‌دهد.



شکل ۴- شکل PIT برای دو مدل پواسونی و دوجمله‌ای منفی

به‌عنوان آخرین ابزار مقایسه و ارزیابی مدل، قوانین امتیازدهی را محاسبه کرده‌ایم که نتایج در جدول ۷ آورده شده است.

جدول ۷- مقایسه مدل پواسونی و دوجمله‌ای منفی با قوانین امتیازدهی

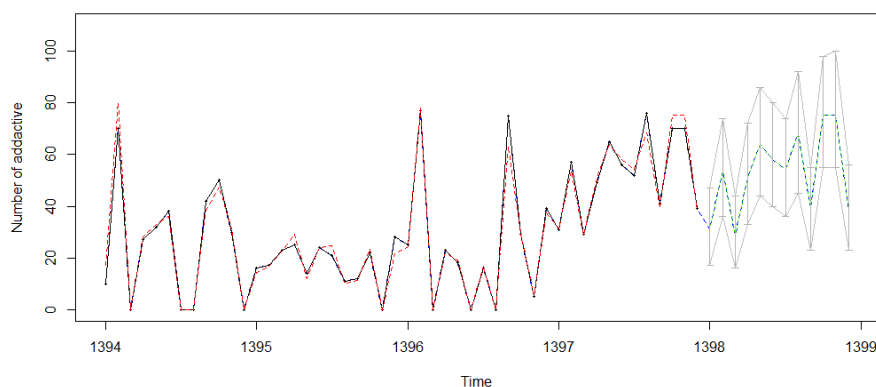
پواسون	دوجمله‌ای منفی	معیارهای
۳۰۳۶	۲۳۳۴	LOG
-۰/۰۸۲	-۰/۲۳۳	QUD
-۰/۲۵۸	-۰/۴۰۲	SPH
۳۲۰۷	۱۷۳۹	RP
۴۴۴۲	۰/۶۴۹	DS
۱/۵۵۴	۰/۲۸۰	NSE
۳۴/۷۸۲	۱۱/۵۵۵	MSE

با توجه به نتایج امتیازدهی چون مقادیر آن برای مدل دوجمله‌ای منفی کمتر است پس برازش بهتری نسبت به مدل پواسونی دارد. تفسیر نتایج ضرائب رگرسیونی از برازش مدل دوجمله‌ای منفی در جدول ۸ به این صورت است که ضریب β_1 مربوط به رگرسیون روی مشاهدات گذشته پاسخ با تاخیر یک واحد زمان و برآورد β_{11} رگرسیون روی مقدار میانگین شرطی ۱۲ واحد گذشته در زمان است. برآورد ضریب

بیش‌پراکنش σ^2 به پارامتر پراکنندگی Φ توزیع دوجمله‌ای منفی مربوط است و این رابطه $\Phi = \frac{1}{\sigma^2}$ است. بر این اساس، مدل برازش شده برای Y_t در طول زمان t به صورت $(Y_t | \mathcal{F}_{t-1} \sim NegBin(\lambda_t, \Phi = 13/07))$ می‌باشد. خطاهای استاندارد برآورد پارامتر رگرسیون و فاصله اطمینان پراکنندگی در جدول ۳ بر اساس تقریب نرمال داده شده در (۶) است. برای ضریب پیش‌پراکنش اضافی σ^2 در توزیع دوجمله‌ای منفی، هیچ تحلیل تقریبی برای خطای استاندارد در دسترس نیست. گاهی مشکل برآورد برای پارامتر پراکنندگی مدل‌ها اتفاق می‌افتد و ممکن است پراکنندگی σ^2 کوچک باشد. مثلاً در این جا مدل برازش شده نزدیک مدل پواسون است و به همین علت مقدار σ^2 کوچک است اما این مسئله عدم قطعیت انتخاب مدل را نشان نمی‌دهد. ضریب برآورد β_1 مربوط به مرتبه‌ی اول اتورگرسیو بسیار کوچک است و حتی پایین‌تر از اندازه‌ی خطای تقریبی آن است که نشان می‌دهد وابستگی قابل توجهی به تعداد رگرسیون‌ها در ماه‌های گذشته وجود ندارد. یک اثر فصلی با ضریب همبستگی مرتبه دوازدهم β_{12} مشاهده می‌شود. با توجه به مدل برازش شده براساس داده‌ها تا اسفند ۱۳۹۷، می‌توان تعداد معتادین مراجعه کننده در سال ۱۳۹۸ پیش‌بینی کرد. نتایج پیش‌بینی در جدول ۸ خلاصه شده‌اند.

جدول ۸- پیش‌بینی تعداد مراجعین در سال ۱۳۹۸ براساس مدل دوجمله‌ای منفی

تیر	خرداد	اردیبهشت	فروردین	
۵۱	۲۹	۵۴	۳۱	۱۳۹۸
آبان	مهر	شهریور	مرداد	
۶۹	۵۴	۵۸	۶۴	۱۳۹۸
اسفند	بهمن	دی	آذر	
۳۹	۷۵	۷۵	۳۹	۱۳۹۸



شکل ۵- مقادیر برازش شده و مقادیر پیش‌بینی شده بر اساس مدل دوجمله‌ای منفی

نمایش گرافیکی پیش‌بینی‌ها در شکل ۵ آمده است. شکل سری زمانی داده‌ها، برازش و پیش‌بینی نیز در شکل ۵ رسم شده است. خطوط شکسته مقادیر برازش شده و خط چین‌ها مقادیر پیش‌بینی همراه با فاصله پیشگویی ۹۵ درصد می‌باشد. در نهایت با توجه به نتایج مشاهده می‌شود که از متغیرهای کمکی در نظر گرفته شده، وضعیت شغلی بر تعداد مراجعین ترک اعتیاد معنی‌دار است. در این میان مدل دوجمله‌ای منفی بهترین برازش را به این داده‌ها دارد.

۴- بحث و نتیجه‌گیری

در این مقاله مدل‌های سری‌زمانی شمارشی بر اساس فرایند پواسن ارائه و مورد مطالعه قرار گرفتند. یک مدل کلی اتورگرسیو تعمیم یافته بیان و معیارهای ارزیابی مدل نیز ارائه شد. مدل‌های سری‌زمانی پواسنی و دوجمله‌ای منفی روی داده‌های ترک اعتیاد شهرستان سمنان به‌کار گرفته شد. همچنین عوامل تأثیرگذار بر تعداد معتادین مراجعه کنند به مراکز ترک اعتیاد مانند سن، وضعیت شغلی و غیره به‌صورت متغیرهای توضیحی در این مدل‌ها مورد تحلیل قرار گرفتند. در میان آن‌ها متغیر وضعیت شغلی که حاکی از سطح رفاه و خطر از دست دادن شغل معتادین می‌باشد، با اطمینان ۹۵ درصد در هر دو مدل پواسنی و دوجمله‌ای تأثیر معنی‌داری بر تعداد مراجعین به مراکز

ترک اعتیاد دارد. همه معیارهای پیشنهادی انتخاب مدل، مدل دوجمله‌ای منفی را برای تحلیل این داده‌ها پیشنهاد کردند. در انتها پیش‌بینی تعداد مراجعات براساس مدل انتخابی صورت گرفت که افزایش تعداد مراجعین را نسبت به سال‌های گذشته نشان می‌داد.

در این مقاله رهیافت درستنمایی برای برآورد پارامترهای مدل به کار گرفته شد، به عنوان پیشنهاد می‌توان از رهیافت بیزی سلسله مراتبی و یا رهیافت بیز تقریبی با وجود پیچیدگی مدل‌ها برای بالا بردن سرعت محاسبات استفاده کرد.

توضیحات

1. Generalized Linear Model
2. Integer-Valued Autoregressive conditional heteroskedasticity
3. Probabilistic calibration
4. Probability integral transform
5. Marginal calibration

مرجع‌ها

- [1] Christou, V., and Fokianos, K. (2014). Quasi-Likelihood Inference for Negative Binomial Time Series Models. *Journal of Time Series Analysis*, **35**, 55–78.
- [2] Czado, C., Gneiting, T., and Held, L. (2009). Predictive Model Assessment for Count Data. *Biometrics*, **65**, 1254–1261.
- [3] Fahrmeir, L., and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.
- [4] Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-Valued GARCH Processes. *J. Time Ser. Anal.*, **27**, 249–329.
- [5] Fokianos, K. (2012). Count time series models. In: *Handbook of Statistics*, **30**, 315–347.

- [6] Gneiting, T., Balabdaui, F., and Raftery, A.E. (2007). Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society B*, **69**, 243–268.
- [7] Liboschik, T., Kerschke, P., Fokianos, K., and Fried, R. (2016). Modelling Interventions in INGARCH Processes. *International Journal of Computer Mathematics*, **93**, 640–657.
- [8] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Second ed. Chapman & Hall, London.
- [9] Nelder, J.A., and Wedderburn, R.W.M. (1972). Generalized Linear Models. *J. R. Stat. Soc. Ser. A*, **135**, 370–384.
- [10] Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press, London.
- [11] Ver, H.J., and Boveng, P. (2007). Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?. *Ecology*, **88**, 2766–2772.

امید کریمی

دکتری آمار

سمنان، دانشگاه سمنان، گروه آمار.

رایانشانی: omid.karimi@semnan.ac.ir

فاطمه حسینی

دکتری آمار

سمنان، دانشگاه سمنان، گروه آمار.

رایانشانی: fatemeh.hoseini@semnan.ac.ir

سپیده تشریفی

فوق لیسانس آمار

سمنان، دانشگاه سمنان، گروه آمار.

رایانشانی: s.tasharofy@gmail.com