مدل‌های آماری برای داده‌های شمارشی وابسته به زمان

حسین باغیشنی [۱] ، سید محمد مهدی طباطبایی [۱]

چکیده:

در سال‌های اخیر پیشرفت‌های قابل ملاحظه‌ای در مدل‌بندی داده‌های وابسته به زمان صورت پذیرفته است. در این مقاله دو رده از مدل‌هایی که برای مدل‌بندی داده‌های وابسته به زمان به کار می‌روند، با تأکید بر شمارشی بودن آنها، معرفی شده‌اند. این دو رده اولین بار توسط کاکس در سال (۱۹۸۱) معرفی شدند و شامل مدل‌های مشاهده مبنا و پارامتر مبنا می‌شوند. این دو رده و برخی از خواص آنها معرفی شده و بطور خلاصه روش‌های برآورد پارامترهای مدل عنوان شده‌اند. در پایان به منظور نمایش کاربرد وسیع آنها، بکارگیری یکی از اعضای مدل‌های مشاهده مبنا با عنوان مدل‌های GLARMA، برای مدل‌بندی تعداد تصادفات رانندگی درون‌شهری مشهد نشان داده شده است.

واژه‌های کلیدی: مدل‌های مشاهده مبنا، مدل‌های پارامتر مبنا، فرآیند پنهان، معادلات برآورد، مدل‌های خطی تعمیم‌یافته خودبرگشت میانگین متحرک (GLARMA).

۱ مقدمه

اولین بار کاکس [۵] دو رده از مدل‌های آماری را برای مدل‌بندی داده‌های وابسته به زمان دسته‌بندی کرد: مدل‌های مشاهده مبنا[۲] و مدل‌های پارامتر مبنا[۳]. این دسته‌بندی مبتنی بر نظریه مدل‌های فضای حالت[۴] است که ساختار آنها بر اساس دو معادله مشاهده و حالت شکل می‌گیرد. برای هر دو مدل پارامتر مبنا و مشاهده مبنا، معادله‌های مشاهده یکسان هستند و اختلاف بین این دو رده از مدل‌ها به مشخص کردن معادله حالت و مدل‌بندی وابستگی بین مشاهدات در هر کدام برمی‌گردد. مدل‌های پارامتر مبنا همبستگی را توسط یک فرآیند پنهان[۵] که مستقل از مشاهدات گذشته سری مشاهده شده است، وارد مدل می‌کنند. در حالیکه مدل‌های مشاهده مبنا، همبستگی را به صورت تابعی از مشاهدات گذشته و متغیرهای تبیینی، مدل‌بندی می‌کنند.

این دو رده، از جمله مدل‌های قدرتمندی هستند که برای مدل‌بندی داده‌های شمارشی همبسته بکار گرفته می‌شوند. با فرض مدل‌بندی داده‌های شمارشی با توزیع پواسون، این دو رده از مدل‌ها برای داده‌های شمارشی، تعمیمی از یک مدل رگرسیون پواسون معمولی هستند. اگر فرض

[۱] گروه آمار – دانشگاه فردوسی مشهد
[۲] Observation Driven Models
[۳] Parameter Driven Models
[۴] State-Space Models
[۵] Latent Process

کنیم دنباله مشاهدات $\{y_t\}_{t=0}^{T}$ شمارشی هستند، به این معنی که برای هر $t \in \{0, 1, 2, \ldots\}$ ، $Y_t \in \mathbb{N} \bigcup \{0\}$، آنگاه در یک مدل رگرسیون پواسون معمولی فرض می‌شود که این مشاهدات با شرط متغیرهای تبیینی، از یک توزیع پواسون با میانگین μ_t استخراج شده‌اند، بطوریکه:

$$P(Y_t = y_t) = \mu_t^{y_t} \frac{e^{-\mu_t}}{y_t!} \qquad (1)$$

که μ_t تابعی از بردار k بعدی متغیرهای تبیینی $\boldsymbol{x}_t$ و پارامترهای نامعلوم $\boldsymbol{\beta}$ می‌باشد. این تابع، تابع پیوند[6] نامیده می‌شود زیرا رابطه بین μ_t و $\boldsymbol{x}_t'\boldsymbol{\beta}$ را مشخص می‌کند. یک تابع پیوند مناسب برای توزیع پواسون، $\mu_t = \exp\{\boldsymbol{x}_t'\boldsymbol{\beta}\}$ است. با انتخاب این تابع پیوند، نتیجه می‌شود $E[Y_t|\boldsymbol{x}_t] = Var[Y_t|\boldsymbol{x}_t] = \exp\{\boldsymbol{x}_t'\boldsymbol{\beta}\}$.

دو رده مدل‌های پارامتر مبنا و مشاهده مبنا، رگرسیون پواسون معمولی را با ضرب میانگین μ_t توسط یک متغیر تصادفی مثبت $\exp\{Z_t\}$ برای لحاظ کردن همبستگی مشاهدات، تصحیح می‌کنند. گاهی مواقع از نماد زیر استفاده می‌شود:

$$\begin{aligned} \lambda_t &= \mu_t \exp\{Z_t\} = \exp\{\boldsymbol{x}_t'\boldsymbol{\beta} + Z_t\} \\ &= \exp\{W_t\} \end{aligned} \qquad (2)$$

بطوریکه تشخیص یکی از دو رده مدل‌ها به نحوه ساخت Z_t برمی‌گردد.

کاکس [5] دسته‌بندی فوق را پیشنهاد کرد. زگر [17] از مدل‌های پارامتر مبنا برای مدل‌بندی اطلاعات تعداد بیماران عصبی استفاده کرد. وی از یک رهیافت معادلات برآورد برای برآورد پارامترهای مدل بهره برد. وست و دیگران [16] مدل‌های پارامتر مبنا را از دیدگاه بیزی مورد بررسی قرار دادند. آزالینی [2]، استراتلی و دیگران [14]، و اندرسون و ایتکین [1] نیز مطالعات مختلفی بر روی این مدل‌ها انجام داده‌اند. کمپل [3] از مدل‌های پارامتر مبنا برای بررسی رابطه بین دمای محیط و مرگ ناگهانی نوزادان، استفاده کرد. چان و لدولتر [4] از روش‌های نمونه‌گیری مونت کارلو برای محاسبه برآوردگرهای درستنمایی ماکسیمم پارامترهای یک مدل پارامتر مبنا استفاده کردند. دیویس و همکاران [8] روشی برای تشخیص وجود فرآیند پنهان در مدل، پیشنهاد دادند. علاوه بر این توزیع مجانبی برآوردگرهای درستنمایی ماکسیمم یک مدل خطی تعمیم‌یافته را در حضور فرآیند پنهان محاسبه کرده و تصحیحی برای اریبی شدید برآوردگرهای پیشنهادی زگر [17] ارایه دادند.

در مورد مدل‌های مشاهده مبنا نیز مطالعات زیادی صورت گرفته و اعضای مختلفی از آنها معرفی شده‌اند. دیویس و همکاران [7] مدل‌های مشاهده مبنا را برای داده‌های شمارشی پواسونی بطور کامل مورد بررسی قرار دادند و خواص بزرگ نمونه‌ای برآوردگرهای پیشنهادی خود را معرفی کردند. همچنین دیویس و همکاران [9] بیشتر جزییات مربوط به این مدل‌ها را برای داده‌های شمارشی وابسته به زمان بررسی کردند. زگر و کواکیش [18] نیز یک رهیافت شبه‌درستنمایی[7] را در مدل‌های رگرسیونی مارکوفی که حالت خاصی از مدل‌های مشاهده مبنا هستند، بکار گرفتند. بخش‌های ۲ و ۳، به ترتیب به

[6] Link Function
[7] Quasi-Likelihood

معرفی مدل‌های مشاهده مبنا و پارامتر مبنا و خواص آنها و همچنین روش‌های برآورد پارامترهای مدل متناظر با هر کدام، پرداخته‌اند. بخش ۴ نیز کاربردی از یک مدل GLARMA[۸] را، که عضوی از رده مدل‌های مشاهده مبنا می‌باشد، بر روی تعداد تصادفات رانندگی درون‌شهری مشهد نشان می‌دهد. در نهایت در بخش ۵، خلاصه‌ای از نتایج عملی بکارگیری دو خانواده مدل‌های پارامتر مبنا و مشاهده مبنا، مطرح شده و مورد بحث قرار گرفته‌اند.

۲ مدل‌های مشاهده مبنا

مدل‌های مشاهده مبنا در برخی موارد با نام مدل‌های انتقالی[۹] در متون مربوط به داده‌های طولی یاد می‌شوند (دیگل و دیگران [۱۱]). برای معرفی مدل ابتدا فرض کنید $\Im_{t-1} = \sigma(\boldsymbol{x}_t, Y_s, s \le t-1)$ سیگما میدان تولید شده توسط مشاهدات گذشته Y و مقادیر گذشته و حال متغیرهای تبیینی $\boldsymbol{x}_t'$ باشد. اکنون فرض می‌شود که $Y_t|\Im_{t-1}$ دنباله‌ای از متغیرهای مستقل از توزیع پواسون با میانگین λ_t هستند. علاوه بر این فرض می‌شود فرآیند حالت $\log(\lambda_t)$ از یک مدل خطی به شکل $\boldsymbol{x}_t'\boldsymbol{\beta} + Z_t$ پیروی می‌کند که در آن Z_t یک فرآیند میانگین متحرک نامتناهی مبتنی بر اغتشاشاتی است که توسط خود داده‌ها تولید می‌شوند و یک دنباله اختلاف مارتینگل[۱۰] تشکیل می‌دهند:

$$Z_t = \sum_{i=1}^{\infty} \pi_i \epsilon_{t-i}. \qquad (۳)$$

تعریف ۱. دنباله $\{\epsilon_t\}$ یک دنباله اختلاف مارتینگل متناظر با $\Im_t$ نامیده می‌شود، هرگاه برای تمام tها، $E(\epsilon_t|\Im_{t-1}) = 0$.

استفاده از این مدل در کارهای عملی، منوط به مشخص کردن شکل تابعی برای محاسبه ϵ_tها می‌باشد. دیویس و همکاران [۶و۹] از باقی‌مانده‌های پیرسون استفاده کردند. یعنی:

$$\epsilon_t = \frac{Y_t - \lambda_t}{\lambda_t^{1/2}}$$

چون $E[Y_t|\Im_{t-1}] = \lambda_t$، براحتی می‌توان نشان داد که $\{\epsilon_t\}$ها تشکیل یک دنباله اختلاف مارتینگل با واریانس واحد می‌دهند. در واقع در این مدل، پارامتری برای واریانس ϵ_t وجود ندارد. به عنوان یک نتیجه، واریانس و خودهمبستگی $\{Z_t\}$ بطور کامل توسط پارامترهای π_i قابل بیان هستند:

$$\begin{aligned} \sigma_Z^2 &= \sum_{i=1}^{\infty} \pi_i^2 \\ \rho_z(h) &= \frac{\sum_{i=1}^{\infty} \pi_i \pi_{i+h}}{\sum_{i=1}^{\infty} \pi_i^2} \end{aligned}$$

در این مدل، امید ریاضی شرطی مشاهدات با شرط متغیرهای تبیینی اریب است. یعنی:

$$E[Y_t|\boldsymbol{x}_t] = \mu_t . E[\exp\{Z_t\}] \qquad (۴)$$

از آنجایی که ویژگی‌های توزیعی Z_t یا $\exp\{Z_t\}$ نامعلوم است، این معادله بیشتر از این ساده نمی‌شود. دیویس و همکاران [۹] محاسبه تقریبی نااریب از μ_t را مبتنی بر

[۸] Generalized Linear Autoregressive Moving Average
[۹] Transitional Models
[۱۰] Martingale Difference Sequence

مقادیر مشاهده شده x_t بصورت زیر پیشنهاد کردند:

$$\hat{\mu_t} = \exp\left\{\hat{W}_t - ۰/۵\hat{\sigma}_Z^۲\right\}$$

برای برازش مدل به یک سری زمانی مشاهده شده، باید پارامترهای مدل برآورد شوند. اما برآورد کردن پارامترهای یک مجموع نامتناهی بطور مستقیم، غیرممکن است. بنابراین لازم است تا فرآیند میانگین متحرک نامتناهی، برحسب تعدادی متناهی از پارامترها مشخص شود. دیویس و همکاران [۹] پیشنهاد کردند که وزن‌های میانگین متحرک π_i به عنوان ضرایب یک صافی خودبرگشت میانگین متحرک، پارامتری شوند:

$$\pi(B) = \sum_{i=۱}^{\infty} \pi_i B^i = \frac{\theta(B)}{\phi(B)} - ۱ \qquad (۵)$$

بطوریکه:

$$\begin{aligned} \theta(B) &= ۱ + \sum_{i=۱}^{q} \theta_i B^i \\ \phi(B) &= ۱ - \sum_{i=۱}^{p} \phi_i B^i \end{aligned} \qquad (۶)$$

چندجمله‌ای‌هایی، با نماد عملگر پس‌رو B، هستند که تمام ریشه‌هایشان خارج دایره به شعاع واحد قرار می‌گیرند. بنابراین Z_t را می‌توان با یک فرآیند ARMA نمایش داد:

$$Z_t = \sum_{i=۱}^{p} \phi_i \left(Z_{t-i} + \epsilon_{t-i}\right) + \sum_{i=۱}^{q} \theta_i \epsilon_{t-i}. \qquad (۷)$$

این دسته از مدل‌ها با عنوان مدل‌های خطی تعمیم‌یافته خودبرگشت میانگین متحرک (GLARMA) از مرتبه (p,q) یا $GLARMA(p,q)$ نامگذاری شده‌اند.

۱.۲ برآورد و استنباط

پارامترهای مدل معرفی شده در بالا، با بردار $\boldsymbol{\delta} = (\beta_۱, \ldots, \beta_k, \phi_۱, \ldots, \phi_p, \theta_۱, \ldots, \theta_q)$ که از بعد $(k+p+q)$ است، نشان داده می‌شوند. معمولاً برآورد پارامترهای $\boldsymbol{\delta}$ با استفاده از روش‌های ماکسیمم درستنمایی انجام می‌شود. بطور کلی برآورد پارامترها در مدل‌های مشاهده مبنا با استفاده از روش درستنمایی شرطی کردن بر روی مقادیر اولیه براحتی صورت می‌پذیرد. در این مدل‌ها، تابع درستنمایی شرطی و مشتقات اول و دوم آن براحتی و در یک الگوریتم تکراری با استفاده از روش عددی نیوتون—رافسون قابل محاسبه هستند. انحراف معیار برآوردگرها نیز با در نظر گرفتن همبستگی موجود در داده‌ها محاسبه می‌شوند.

روش محاسبه برآوردگرهای ماکسیمم درستنمایی یک مدل مشاهده مبنا بر اساس توزیع پواسون و با استفاده از باقیمانده‌های پیرسون به عنوان دنباله اختلاف مارتینگل حاصل از مدل، توسط دیویس و همکاران [۹] بطور کامل تشریح شده است. لازم به ذکر است که کدم و فوکیانوس [۱۲] برآورد پارامترها را به روش درستنمایی جزیی[۱۱] پایه‌ریزی کرده‌اند. البته نتایج استنباط در این مدل‌ها به دو روش درستنمایی جزیی و حاشیه‌ای، با هم تطابق دارند[۱۲].

چنانچه تابع لگاریتم درستنمایی بصورت $l(\boldsymbol{\delta}) = \sum_{t=۱}^{n} \log f\left(y_t | \Im_{t-۱}\right)$ تعریف شود که در آن f تابع احتمال شرطی Y_t باشد، آنگاه با فرض توزیع

[۱۱] Partial Likelihood

پواسون، تابع زیر بایستی برحسب $\boldsymbol{\delta}$ ماکسیمم شود:

$$l(\boldsymbol{\delta}) = \sum_{t=1}^{n} \left(y_t W_t(\boldsymbol{\delta}) - e^{W_t(\boldsymbol{\delta})} \right) \qquad (۸)$$

برآوردگرماکسیمم درستنمایی $\boldsymbol{\delta}$ با $\hat{\boldsymbol{\delta}}$ نشان داده می‌شود. استنباط در مدل‌های مشاهده مبنا با استفاده از نتیجه زیر قابل انجام است:

$$\hat{\boldsymbol{\delta}} \approx N\left(\boldsymbol{\delta}_0, \frac{1}{T}\Omega\right) \qquad (۹)$$

که $\boldsymbol{\delta}_0$ مقدار درست پارامتر است و برآوردی از Ω با رابطه زیر قابل دستیابی است:

$$\hat{\Omega} = -\left(\frac{1}{T} \frac{\partial^2 l(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}'} |_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}} \right) \qquad (۱۰)$$

در مرحله برازش مدل نیز، Z_t با استفاده از تکرارهای ARMA محاسبه می‌شود. برای $t \leq 0$، $e_t = 0$ و $Z_t = 0$ و برای $t > 0$، مراحل سه گانه زیر تکرار می‌شوند:

$$\begin{aligned}
\hat{Z}_t &= \phi_1(\hat{Z}_{t-1} + e_{t-1}) + \ldots \\
&+ \phi_p(\hat{Z}_{t-p} + e_{t-p}) + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} \\
W_t &= \boldsymbol{x}_t'\boldsymbol{\beta} + \hat{Z}_t \\
e_t &= (Y_t - e^{W_t}) e^{-\frac{W_t}{2}}
\end{aligned}$$

۳ مدل‌های پارامتر مبنا

همانطور که در مقدمه گفته شد، یک مدل پارامتر مبنا برای داده‌های شمارشی معمولاً مبتنی بر فرض رگرسیون پواسون برای مدل‌بندی اثرات متغیرهای تبیینی و یک فرآیند پنهان برای محسوب کردن همبستگی و فراپراکنش[۱۲] مشاهدات در مدل است. با در نظر گرفتن این ساختار، سه رهیافت متفاوت را می‌توان معرفی کرد. در اولین رهیافت برای $\exp\{Z_t\}$ توزیع گاما فرض می‌شود و در واقع برای داده‌های شمارشی، توزیع پواسون مخلوط شده با اثرات تصادفی گاما در نظر گرفته می‌شود. با این فرض‌ها، چگالی‌های حاشیه‌ای دوجمله‌ای منفی حاصل می‌شوند که بر اساس این چگالی‌های حاشیه‌ای، برآورد پارامترها انجام می‌گیرد.

رهیافت دوم، برای Z_t فرض توزیع نرمال را در نظر می‌گیرد [۴]. بکارگیری این رهیافت، برآوردگرهای کارا نتیجه می‌دهد اما نیاز به محاسبات سنگین عددی و الگوریتم‌های پیچیده آماری دارد. شکل توزیع‌های حاشیه‌ای و گشتاورهای آنها نیز به صورت بسته وجود ندارند. استفاده از الگوریتم‌های پیچیده آماری مانند الگوریتم MCEM[۱۳] در این رهیافت معمول‌اند. رهیافت سوم، مبتنی بر معادلات برآورد تعمیم‌یافته[۱۴] با مشخص کردن گشتاورهای مرتبه اول و دوم Y_t است و برآورد پارامترها از طریق آماره‌های مناسب نمونه‌ای انجام می‌گیرد (زگر[۱۷]).

با توجه به تعریف مدل‌های پارامتر مبنا، مشاهدات شمارشی y_t با شرط فرآیند پنهان Z_t بطور مستقل از یک توزیع پواسون با میانگین λ_t تولید شده‌اند:

$$Y_t | Z_{t-1} \sim Po(\lambda_t).$$

در این مدل، امید ریاضی شرطی Y_t با شرط متغیرهای

[۱۲] Overdispersion
[۱۳] Monte Carlo EM
[۱۴] Generalized Estimating Equations

تبیینی مدل، برابر است با $E[Y_t|\boldsymbol{x}_t] = \mu_t.E[\exp\{Z_t\}]$. توزیع Z_t یا $\exp\{Z_t\}$ را می‌توان طوری تعیین کرد که $E[\exp\{Z_t\}] = ۱$ [۱۷]. بنابراین برخلاف مدل‌های مشاهده مبنا، در این رده از مدل‌ها می‌توان ضرایب متغیرهای تبیینی را بطور مفیدی تعبیر کرد. در واقع می‌توان ضریب رگرسیونی یک متغیر تبیینی را به عنوان تغییر نسبی در میانگین حاشیه‌ای Y_t در مقیاس لگاریتمی به ازای یک واحد تغییر در متغیر مورد نظر تعبیر کرد.

[۴] فرض نرمال بودن فرآیند پنهان Z_t را در نظر گرفتند. بویژه فرض کردند که Z_t یک فرآیند مانای گاوسی $AR(۱)$ است، یعنی $Z_t = \phi Z_{t-۱} + \nu_t$ که در آن $\{\nu_t\}$ یک دنباله مستقل و هم‌توزیع با توزیع $N(۰, \sigma_\nu^۲)$ می‌باشد. چون $\{Z_t\}$ یک فرآیند گاوسی با $E[Z_t] = ۰$ است، متغیر تصادفی $\exp\{Z_t\}$ دارای توزیع لگ–نرمال با $E[\exp\{Z_t\}] = \exp\{۰/۵Var[Z_t]\}$ خواهد بود. به عنوان یک نتیجه می‌توان با کاستن $۰/۵\sigma_Z^۲$ از برآورد ضریب ثابت رگرسیون، مقداری نارایب برای $E[Y_t|\boldsymbol{x}_t]$ در این حالت بدست آورد.

۱.۳ برآورد و استنباط

مشکل اساسی مدل‌های پارامتر مبنا، محاسبه تابع درستنمایی آنها و در نتیجه برآورد پارامترها و استنباط در مورد آنهاست. پیش‌بینی مشاهدات آینده سری نیز نسبت به مدل‌های مشاهده مبنا خیلی مشکل‌تر انجام می‌شود. با توجه به این مشکل اساسی، چندین روش و رهیافت برای برآورد و استنباط در این رده از مدل‌ها، توسط افراد مختلف پیشنهاد شده‌اند که این قسمت بطور خلاصه به معرفی برخی از آنها پرداخته است.

۱.۱.۳ رهیافت معادلات برآورد

چنانچه فرض کنیم $\exp\{Z_t\} = \epsilon_t$، آنگاه $E[Y_t|\epsilon_t] = \exp\left(\boldsymbol{x}_t'\boldsymbol{\beta}\right)\epsilon_t$. در این حالت زگر [۱۷] رهیافت معادلات برآورد را برای برآورد ضرایب رگرسیون β، پیشنهاد کرد که مشابه روش شبه‌درستنمایی عمل می‌کند. تمام آنچه که برای برآورد پارامترهای رگرسیونی در این رهیافت لازم است، دو گشتاور حاشیه‌ای اول و دوم Y_t است.

با شرط معلوم بودن فرآیند پنهان ϵ_t، فرض کنید:

$$\begin{aligned} u_t &= E(Y_t|\epsilon_t) = e^{\boldsymbol{x}_t'\boldsymbol{\beta}}\epsilon_t, \\ w_t &= Var(Y_t|\epsilon_t) = u_t \end{aligned} \qquad (۱۱)$$

همچنین فرض کنید ϵ_t یک فرآیند ایستای نامنفی غیرقابل مشاهده با میانگین $E(\epsilon_t) = ۱$ و تابع کوواریانس $Cov(\epsilon_t, \epsilon_{t+h}) = \sigma^۲\rho_\epsilon(h)$ باشد. بنابراین گشتاورهای حاشیه‌ای Y_t عبارتند از:

$$\begin{aligned} \mu_t &= E(Y_t) = \exp(\boldsymbol{x}_t'\boldsymbol{\beta}), \\ \delta_t &= Var(Y_t) = \mu_t + \sigma^۲\mu_t^۲ \end{aligned} \qquad (۱۲)$$

اعمال محدودیت گشتاور اول، $E(\epsilon_t) = ۱$، به دلیل قابل تفکیک شدن میانگین فرآیند پنهان از ضریب ثابت رگرسیون است. زیرا چنانچه این شرط را قایل نشویم، میانگین فرآیند پنهان و ضریب ثابت با هم مخلوط می‌شوند و قابل تفکیک و شناسایی نخواهند بود. در واقع با این محدودیت، میانگین حاشیه‌ای μ_t تنها به $\boldsymbol{x}_t'\beta$ وابسته خواهد شد و به گشتاورهای فرآیند پنهان وابسته

نخواهد بود. از طرفی

$$\rho_Y(t,h)=\frac{\rho_\epsilon(h)}{\left[\{1+(\sigma^2\mu_t)^{-1}\}\{1+(\sigma^2\mu_{t+h})^{-1}\}\right]^{1/2}}$$

چون مخرج کسر برابری آخر بیشتر از یک است، خودهمبستگی در Y_t کمتر یا مساوی خودهمبستگی در ϵ_t است. بنابراین از روی خود مشاهدات نمی‌توان ساختار همبستگی موجود در داده‌ها را درست تشخیص داد و با استفاده از مشاهدات، بطور کلی همبستگی موجود در آنها کم برآورد می‌شود.

۲.۱.۳ رهیافت مدل‌های خطی تعمیم‌یافته

مدل پیشنهادی چان ولدولتر [۴] نیز مشکل مدل زگر [۱۷] را داراست. یعنی خودهمبستگی موجود در داده‌های شمارشی نمی‌تواند خودهمبستگی مدل را که توسط فرآیند پنهان تولید می‌شود، با دقت برآورد کند. حتی در این حالت چنانچه $Z_t\sim N\left(-\frac{\sigma_Z^2}{2},\sigma_Z^2\right)$، همان مدل زگر [۱۷] حاصل می‌شود. در عمل در چنین مواقعی، بایستی قبل از برآورد تابع خودهمبستگی مدل، رگرسیون پواسون بر روی متغیرهای تبیینی x_t' برازش داده شود و سپس ساختار خودهمبستگی از روی باقی‌مانده‌های حاصل از مدل خطی تعمیم‌یافته برازش داده شده، برآورد شود. بنابراین به منظور آزمون وجود فرآیند پنهان و در ادامه آن تشخیص ساختار همبستگی مدل، لازم است یک روش برآورد سازگار برای بردار ضرایب رگرسیون فراهم آید. یک روش ساده و معمول برای محاسبه برآوردگرهای سازگار β، استفاده از مدل خطی تعمیم‌یافته استاندارد است، زیرا برآوردگرهایی که از این روش بدست می‌آیند، سازگار و دارای توزیع مجانبی نرمال می‌باشند.

۳.۱.۳ رهیافت الگوریتم MCEM

در ابتدای این بخش اشاره شد که مشکل اصلی مدل‌های پارامتر مبنا، محاسبه تابع درستنمایی آنهاست. بنابراین می‌توان از روش‌های مختلفی برای تقریب این تابع استفاده کرد. الگوریتم EM (دمپسترو همکاران [۱۰]) یکی از روش‌های متداول برای محاسبه برآوردگرهای درستنمایی ماکسیمم در اینگونه مسایل پیچیده است. اما در مورد داده‌های شمارشی مشکلی که وجود دارد، غیرقابل اجرا بودن مرحله E، محاسبه امید ریاضی شرطی تابع درستنمایی توأم مشاهدات و فرآیند پنهان به شرط مشاهدات، الگوریتم است. زیرا توزیع شرطی فرآیند پنهان با شرط مشاهدات، خیلی پیچیده است. تقریب این امید ریاضی توسط روش‌های تربیع عددی[۱۵] نیز به دلیل بالا بودن بعد انتگرال‌ها، ناکاراست (شان و مک‌کالاک[۱۳]). روشی کاراتر و در عین حال ساده برای اجرای مرحله E، استفاده از روش‌های نمونه‌گیری زنجیر مارکوف مانند نمونه‌گیری گیبز یا الگوریتم متروپولیس است. در نتیجه بجای مرحله E در الگوریتم EM، مرحله MCE اجرا می‌شود. الگوریتم حاصل را الگوریتم MCEM گویند که اولین بار در [۱۵] پیشنهاد شد.

یک خاصیت مهم الگوریتم EM این است که درستنمایی داده‌های مشاهده شده همیشه همراه با دنباله EM صعود می‌کند. برای الگوریتم MCEM این خاصیت برقرار نیست، اما چان ولدولتر [۴] نشان دادند که تحت شرایط

[۱۵] Numerical Quadrature

مناسب نظم، یک دنباله MCEM با احتمال بالا به ماکسیمم کننده درستنمایی داده‌های مشاهده شده نزدیک می‌شود.

۴ تعداد تصادفات رانندگی درون‌شهری مشهد

این قسمت به تشریح کاربرد یکی از اعضای مدل‌های مشاهده مبنا معروف به مدل GLARMA پرداخته است. مجموعه داده‌های شمارشی که برای تحلیل در نظر گرفته شده‌اند، تعداد کل تصادفات ماهیانه رانندگی درون‌شهری مشهد برای دو سال ۱۳۸۲ و ۸۳ می‌باشند. متأسفانه مجموعه اطلاعات دیگر از سایر عواملی که می‌توانند بر این تعداد تصادفات موثر باشند، در دسترس نبودند. لذا با توجه به وابسته بودن مشاهدات به زمان، و رفتار فصلی سری زمانی مشاهدات که نمودار آن در شکل ۱ نمایش داده شده است، اقدام به برازش یک مدل GLARMA شد. روش برآورد پارامترها و برازش مدل، همان روشی است که برای مدل‌های مشاهده مبنا توضیح داده شد. برای مدل‌بندی اثر فصلی موجود در داده‌ها از مولفه‌های هارمونیک ماهیانه و فصلی استفاده شده است. علاوه بر این با توجه به رشد افزایشی سری زمانی مشاهدات، یک مولفه روند خطی t/n نیز در مدل در نظر گرفته شده است. دلیل استفاده از t/n بجای t برای مولفه روند، بدست آوردن برآوردی سازگار برای ضریب رگرسیونی این مولفه است.

با توجه به مدل‌های مختلفی که بر روی داده‌ها برازش داده شدند، برای مولفه ARMA مدل GLARMA دو تأخیر ۲ و ۳ مولفه AR معنی‌دار تشخیص داده شدند. دو تأخیر y_{t-2} و y_{t-3}، ساختار همبستگی موجود در داده‌ها پس از حذف اثر متغیرهای تبیینی معرفی شده در بالا را مدل‌بندی می‌کنند. نتایج خلاصه شده برآورد مدل در جدول ۱ گزارش شده‌اند. شکل ۲ که مقادیر برازش داده شده مشاهدات، مبتنی بر مدل، را در مقابل مقادیر واقعی آنها نمودار کرده است، بیانگر مناسب بودن مدل برازش داده شده بر داده‌هاست. در واقع می‌توان گفت این مدل، الگوی تغییرات داده‌ها را بخوبی مدل‌بندی کرده است.

۵ بحث و نتیجه‌گیری

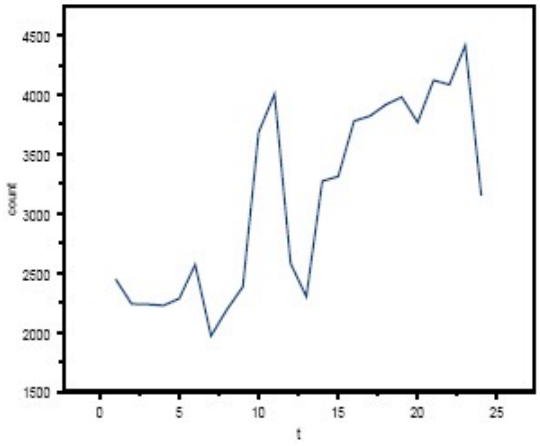
در بسیاری از مسایل عملی، هدف اصلی یافتن رابطه بین متغیرهای تبیینی اندازه‌گیری شده با داده‌های شمارشی مشاهده شده وابسته به زمان است. در اغلب موارد، متغیرهای تبیینی که لزوم آنها در مدل احساس می‌شوند، وجود دارند. در چنین مواقعی، برازش مدل‌های مختلف سری زمانی شمارشی قابل مقایسه بر روی داده‌ها به عنوان قسمتی از یک تحقیق، لازم می‌شود. به عنوان مثال در مطالعه تأثیر سیاست‌های اجرایی دولت در تعداد حوادث جاده‌ای منجر به مرگ، باید همه نواحی و استان‌های کشور بطور جداگانه بررسی شوند. زیرا زمان‌بندی و طبیعت متغیرهای مرتبط با این اطلاعات در نواحی مختلف، متغیر هستند. در این موارد استفاده از روش‌هایی که به سادگی قابل اجرا باشند و تأثیر متغیرها را بر روی سری زمانی شمارشی مشاهده شده به سرعت محاسبه کنند و علاوه بر این همبستگی موجود در داده‌ها را نیز لحاظ کنند، ارجحیت دارند.

در حال حاضر برازش مدل‌های پارامتر مبنا نیاز به محاسبات کامپیوتری پیچیده و سنگین دارد. اما تعبیر پارامترهای مربوط به متغیرهای تبیینی در این مدل‌ها، ساده و مفید می‌باشد و می‌توان ضریب هر متغیر را به عنوان تغییر نسبی در میانگین حاشیه‌ای Y_t در مقیاس لگاریتمی به ازای یک واحد تغییر در متغیر تبیینی مورد نظر، تفسیر کرد. از طرف دیگر مدل‌های مشاهده مبنا براحتی و با سرعت، قابل اجرا و برازش هستند. علاوه بر این، این مدل‌ها استنباط اثرات ثابت مدل را در حضور همبستگی داده‌ها تصحیح می‌کنند. اما تعابیر پارامترها بسته به ساختار همبستگی در نظر گرفته شده در مدل، پیچیده می‌شوند.

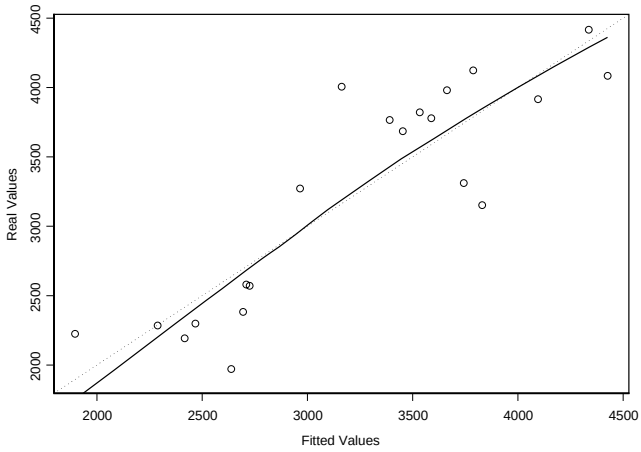
قدردانی: در پایان لازم می‌دانیم از مسئول واحد آمار سازمان آمار و فن‌آوری اطلاعات شهرداری مشهد، سرکار خانم حاجی‌زاده، برای ارایه اطلاعات مربوط به تصادفات رانندگی شهر مشهد، قدردانی کنیم. همچنین از داوران محترم این مقاله بخاطر نظرات سازنده‌شان در بهبود این اثر، تشکر و قدردانی می‌کنیم.

جدول ۱: برآورد پارامترهای مدل مشاهده مبنا

پارامتر	برآورد	انحراف معیار	t—مقدار
ثابت	۱۰/۴۲	۰/۰۱۳	۷۹/۷۱
t/n	−۰/۱۰۷	۰/۰۰۸	−۱۳/۰۱
$y_{t-۲}$	−۰/۰۰۰۱۶	۰/۰۰۰۰۱	−۱۵/۶۴
$y_{t-۳}$	−۰/۰۰۰۱۵	۰/۰۰۰۰۱	−۱۵/۱
$\sin(\frac{۲t\pi}{۱۲})$	−۰/۰۹۵	۰/۰۰۶۶	−۱۴/۴۴
$\cos(\frac{۲t\pi}{۱۲})$	۰/۱۱۲	./۰۰۹	۱۲/۴۶
$\sin(\frac{۲t\pi}{۴})$	۰/۷۷۵	۰/۰۴۵	۱۷/۲
$\cos(\frac{۲t\pi}{۴})$	−۱/۱۶	۰/۰۵۴	−۲۱/۳۶



شکل ۱: نمودار سری زمانی تعداد تصادفات رانندگی شهر مشهد



شکل ۲: مقادیر برازش شده در مقابل مقادیر واقعی سری

مراجع

[1] Anderson, D.A. and Aitkin, M. ,1985, Variance Component Models with Binary Response: Interviewer Variability, *Journal of the Royal Statistical Society, B,* **47**, 203-210.

[2] Azzalini, A. ,1982, Approximate Filtering of Parameter-Driven Processes, *Journal of Time Series Analysis,* **3**, 219-223.

[3] Campbell, M.J. ,1994, Time Series Regression for Counts: An Investigation into the Rela-

tionship between Sudden Death Syndrome and Environmental Temperature, *Journal of the Royal Statistical Society, A,* **157** 191-208.

[4] Chan, K.S. and Ledolter, J. ,1995, Monte Carlo EM Estimation for Time Series Models Involving Counts, *Journal of the American Statistical Association,* **90**, 242-252.

[5] Cox, D.R. ,1981, Statistical Analysis of Time Series: Some Recent Developments, *Scandinavian Journal of Statistics,* **8**, 93-115.

[6] Davis, R.A., Dunsmuir, W.T.M. and Streett, S.B. ,2003, Observation-Driven Models for Poisson Counts, *Biometrika,* **90**, 777-790.

[7] Davis, R.A., Dunsmuir, W.T.M. and Streett, S.B. ,2001, Observation-Driven Models for Poisson Counts, *Technical Report.*

[8] Davis, R.A., Dunsmuir, W.T.M. and Wang, Y. ,2000, On Autocorrelation in a Poisson Regression Model, *Biometrika,* **87**, 491-506.

[9] Davis, R.A., Dunsmuir, W.T.M. and Wang, Y. ,1999, Modelling Time Series of Count Data, *In Asymptotics, Nonparametrics, and Time Series,* Ed. S. Ghosh, pp. 63-114, New York, Marcel Dekker.

[10] Dempster, A.P., Laird, N.M. and Rubin, D. ,1977, Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion), *Journal of the Royal Statistical Society, B,* **39**, 1-38.

[11] Diggle, P.J., Liang, K.Y. and Zeger, S.L. ,1994,*Analysis of Longitudinal Data*, Oxford University Press, Oxford.

[12] Kedem, B. and Fokianos, K. ,2002,*Regression Models for Time Series Analysis,* Wiley, New York.

[13] Shun, Z. and McCullogh, C.E. ,1995, Laplace Approximation of High-Dimensional Integrals, *Journal of the Royal Statistical Society, B,* **57**, 749-760.

[14] Stiratelli, R., Laird, N. and Ware, J.H. ,1985, Random Effects Models for Serial Observations with Binary Response, *Biometrics,* **40**, 961-971.

[15] Wei, C.G. and Tanner, M.A. ,1990, A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms, *Journal of the American Statistical Association,* **85**, 699-704.

[16] West, M., Harrison, P.J. and Migon, H.S. ,1985, Dynamic Generalized Linear Models and Bayesian Forecasting, *Journal of the American Statistical Association,* **80**, 73-96.

[17] Zeger, S.L. ,1988, A Regression Model for Time Series of Counts, *Biometrika,* **75**, 621-629.

[18] Zeger, S.L. and Qaqish, B. ,1988, Markov Regression Models for Time Series: A QuasiLikelihood Approach, *Biometrics,* **44**, 1019-1031.