چهارمین سمینار آمار فضایی و کاربردهای آن
4th Seminar on Spatial Statistics and Its Applications
۵–۶ آبان ماه ۱۴۰۰
27-28 Oct. 2021

مدل رگرسیون آمیخته خطی تعمیم‌یافته بتا فضایی‌ـ زمانی و کاربرد آن در آمارگیری هزینه و درآمد

لیدا کلهری
پژوهشکده آمار

چکیده: در سال‌های اخیر مدل‌بندی داده‌های محدود در بازه $(۰, ۱)$ مورد توجه قرار گرفته است و مدل رگرسیون بتا برای مدل‌بندی این‌گونه مشاهدات معرفی شده است. در این مقاله مدل‌بندی داده‌ها در بازه $(۰, ۱)$ با در نظر گرفتن ساختار همبستگی فضایی ـ زمانی تفکیک‌پذیر بررسی شده است و برآورد پارامترها با رهیافت بیزی به‌دست آمده‌اند. کاربست این مدل برای تحلیل داده‌های آمارگیری هزینه و درآمد در شهر تهران ارائه شده است.

واژه‌های کلیدی: رگرسیون بتا، همبستگی فضایی‌ـ زمانی، رهیافت بیزی، آمارگیری هزینه و درآمد
کد موضوع‌بندی ریاضی (۲۰۱۰): 62F15، 62M30.

۱ مقدمه

در سال‌های اخیر تمرکز زیادی بر مدل‌بندی داده‌هایی مانند نرخ‌ها و نسبت‌ها که در بازه $(۰, ۱)$ تغییر می‌کنند، شده است. اهمیت این موضوع به دلیل محدودیت‌های به‌کارگیری مدل‌های موجود برای تحلیل این‌گونه داده‌ها است. به عنوان مثال، استفاده از مدل‌های رگرسیونی یا رگرسیون تعمیم‌یافته برای مدل‌بندی داده‌های محدود دربازه $(۰, ۱)$ می‌تواند به پیشگویی خارج از دامنه منجر شود. فراری و کریباری (۲۰۰۴) برای اجتناب از اعمال تبدیل به داده‌های اولیه و با در نظر گرفتن خواص انعطاف‌پذیری توزیع بتا، استفاده از مدل رگرسیونی بتا را پیشنهاد کردند که در آن متغیر پاسخ دارای توزیع بتای بازپارامتریده است. فیگورا و همکاران (۲۰۱۳) پارامترهای میانگین و دقت توزیع بتای بازپارامتریده را با در نظر گرفتن اثرات تصادفی مدل‌بندی نموده و براورد پارامترهای مدل با رهیافت بیزی را مورد بررسی قرار داده‌اند.

کپدا وهمکاران (۲۰۱۲)، لاگوس الوارز و همکاران (۲۰۱۷) و کلهری و محمدزاده (۲۰۱۸) مدل‌بندی داده‌های فضایی با استفاده از مدل رگرسیون بتای تعمیم‌یافته را مورد مطالعه قرار داده‌اند. بسیاری از انواع داده‌ها در حوزه آمار رسمی

* نام وآدرس پست الکترونیکی ارائه‌دهنده مقاله: لیدا کلهری، kalhori@srtc.ac.ir

علاوه بر همبستگی فضایی در طول زمان نیز وابسته هستند، که به عنوان داده‌های فضایی_زمانی شناخته می‌شوند. تاکنون مطالعات متعددی برای مدل‌بندی این‌گونه داده‌ها صورت گرفته است. با این وجود مدل‌بندی داده‌های فضایی_زمانی با دامنه تغییرات محدود در یک بازه، کمتر مورد توجه قرار گرفته است. در این مقاله مدل‌بندی داده‌های همبسته فضایی محدود در بازه $(۰, ۱)$، در طول زمان مورد مطالعه قرار خواهد گرفت و براورد پارامترها با رهیافت بیزی انجام می‌شود. در بخش ۲ مدل آماری معرفی می‌شود. بخش ۳ به مطالعه شبیه‌سازی اختصاص یافته است و در بخش ۴ کاربست مدل بر داده‌های آمارگیری هزینه و درآمد ارائه شده است.

۲ مدل آماری

در این بخش ابتدا مدل رگرسیون بتا به اختصار معرفی می‌شود و پس از آن مدل مورد استفاده برای تحلیل داده‌های فضایی_زمانی در بازه $(۰, ۱)$ بیان خواهد شد. تاکنون طیف گسترده‌ای از مدل‌های آماری برای مدل‌بندی داده‌های فضایی_ زمانی معرفی شده و قابل استفاده هستند. از آن‌جا که هبستگی فضایی و زمانی به‌صورت جداگانه قابل مدل‌بندی هستند، می‌توان از همه فنون موجود در سری‌های زمانی و آمار فضایی برای مدل‌بندی ساختار همبستگی فضایی_زمانی تفکیک‌پذیر بهره گرفت (محمدزاده، ۱۳۹۴). هدف این مطالعه تحلیل داده‌های فضایی_زمانی با استفاده از یک مدل تفکیک‌پذیر است و اثرات متقابل فضایی_ زمانی در دامنه این تحقیق نیست.

توزیع بتا از انعطاف پذیری بالایی برخوردار است و با تغییر مقادیر پارامترهایش به فرم‌های متمایزی ظاهر می‌شود. با توجه به این که در مدل‌های رگرسیونی، الگوی رفتار میانگین متغیر پاسخ مشروط بر متغیرهای تبیینی مورد بررسی قرار می‌گیرد، فراری و کریباری (۲۰۰۴) توزیع بتای بازپازامتریده را برای مطالعه نسبت‌ها پیشنهاد کردند. آن‌ها پارامترهای توزیع بتا را به‌گونه‌ای بازنویسی کردند که مدل رگرسیونی بر اساس میانگین متغیر پاسخ معین شود. بنابراین توزیع بتا با پارامترهای (a, b) را با فرض $\mu = \frac{a}{a+b}$ و $\phi = a + b$ به صورت

(۱.۲)

$$\pi(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((۱-\mu)\phi)} y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱} \qquad ۰ < y < ۱,$$

در نظر گرفتند که در آن $۰ < \mu < ۱$ و $\phi > ۰$، در نتیجه $E(Y) = \mu$، $V(Y) = \frac{\mu(۱-\mu)}{۱+\phi}$، که در آن ϕ پارامتر دقت نامیده می‌شود. با فرض ثابت بودن پارامتر دقت، مدل رگرسیون بتا توسط فراری و کریباری (۲۰۰۴) به‌صورت زیر

$$g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta} \qquad i = ۱, \ldots, n, \tag{۲.۲}$$

ارائه شد که در آن $\mathbf{x}_i = (x_{i۰}, \ldots, x_{ip})$ بردار متغیرهای تبیینی، $\boldsymbol{\beta} = (\beta_۰, \ldots, \beta_p)^\top$ بردار ضرایب رگرسیونی، و $g(\cdot)$ تابع پیوند مناسب است. در این مقاله، برای مطالعه اثرات فضایی ـزمانی یک مدل آمیخته خطی تعمیم یافته از رگرسیون بتا با اثرات تفکیک‌پذیر معرفی می‌شود.

فرض کنید $\mathbf{y}(s,t) = (y(s_۱,t_۱), \ldots, y(s_n,t_T))^\top$ مشاهدات متغیر $Y(s,t) = (Y(s_۱,t_۱), \ldots, Y(s_n,t_T))^\top$ در n موقعیت مکانی $s_۱, \ldots, s_n$ و T زمان $t_۱, \ldots, t_T$ باشند. به‌منظور ساده‌سازی در نوشتار از این پس فرض می‌شود y_{ij} و Y_{ij} برای $i = ۱, \ldots, n$ و $j = ۱, \ldots, T$ به ترتیب نمایشگر $y(s_i,t_j)$ و $Y(s_i,t_j)$ باشند. هدف مدل‌بندی متغیرهای تصادفی همبسته فضایی_زمانی $Y_{۱۱}, \ldots, Y_{nT}$ از توزیع بتا است که همبستگی فضایی آن‌ها از طریق یک میدان تصادفی گاوسی[1] (GRF) و همبستگی زمانی آن‌ها از طریق یک مدل سری زمانی خودهمبسته در نظر گرفته شود.

[1] Gaussian Random Field

فرض کنید $\boldsymbol{\tau} = (\tau(s_1), \ldots, \tau(s_n))^\top \equiv (\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_n)^\top$ نمایانگر یک میدان تصادفی مانای مرتبه دوم گاوسی با میانگین صفر باشد، یعنی $E(\boldsymbol{\tau}) = 0$. همچنین فرض کنید $\nu = (\nu(t_1), \ldots, \nu(t_T))^\top \equiv (\nu_1, \ldots, \nu_T)^\top$ نمایانگر مؤلفه‌های یک سری زمانی خودهمبسته باشند. متغیرهای تصادفی Y به شرط مؤلفه تصادفی $\boldsymbol{\tau}$ و ν از یکدیگر مستقل هستند و از توزیع بتا پیروی می‌کنند. به عبارت دیگر

$$Y \mid \boldsymbol{\tau}, \nu \sim \mathrm{Beta}(\mu(\mathrm{s,t})\phi(\mathrm{s,t}), (1-\mu(\mathrm{s,t}))\phi(\mathrm{s,t})),$$

که توزیع آن از جایگذاری $\mu(s,t)$ و $\phi(s,t)$ در (۱.۲) به‌دست می‌آید. بنابراین مدل آمیخته خطی تعمیم‌یافته بتا با اثرات تفکیک‌پذیر فضایی ـ زمانی به‌صورت

$$\mathrm{g}(E(Y_{ij} \mid \tau_i, \nu_j)) = \mathbf{x}_{ij}\boldsymbol{\beta} + \tau_i + \nu_j \quad i = s_1, \ldots, s_n, \;\; j = t_1, \ldots, t_T, \tag{۳.۲}$$

تعریف می‌شود که یک مدل خطی در اثرات ثابت و تصادفی است. در این مدل $\mathrm{g}(\cdot)$ تابع پیوند یک به یک و مشتق‌پذیر با دامنه اعداد حقیقی، $\mathbf{x}_{ij} = (x_{ij.}, \ldots, x_{ijm})$ و $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)^\top$ به ترتیب بردار متغیرهای تبیینی و ضرایب رگرسیونی، τ_i نمایشگر اثر تصادفی است که ساختار همبستگی فضایی داده‌ها را در مدل وارد می‌کند و ν_j نمایانگر روند داده‌ها در طول زمان بر اساس یک مدل سری زمانی خودهمبسته مرتبه p به صورت $\nu_j = \mathrm{c} + \sum_{k=1}^{p} \varphi_k \nu_{j-k}$ است، که در آن c ثابت مدل و $\varphi_k, \ldots, \varphi_1$ پارامترهای مدل هستند. فرایند تصادفی $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^\top$ از توزیع نرمال چند متغیره به‌صورت $\boldsymbol{\tau} \sim \mathrm{N_n}(0, \Sigma_\tau)$ پیروی می‌کند، که در آن

$$(\Sigma_\tau)_{i\ell} = Cov(\tau_i, \tau_\ell) = \sigma^2_. Corr(\tau_i, \tau_\ell), \qquad i, \ell = s_1, \ldots, s_n$$

مؤلفه‌های Σ_τ تابعی از فاصله بین نقاط هستند به‌گونه‌ای که تشریح کننده ساختار کوواریانس فضایی باشند.

بر اساس مطالعات گذشته در مدل‌های رگرسیون بتا، تابع پیوند لوجیت برای مدل‌بندی میانگین در نظر گرفته می‌شود. بنابراین (۳.۲) به‌صورت

$$\mathrm{logij}(\mu_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta} + \tau_i + \nu_j, \quad i = s_1, \ldots, s_n, \quad j = t_1, \ldots, t_T, \tag{۴.۲}$$

در نظر گرفته می‌شود. در نتیجه با فرض ثابت بودن پارامتر دقت توزیع بتا، مدل سلسه مراتبی به‌صورت زیر خواهد بود:

$$Y_{ij} \mid \boldsymbol{\beta}, \phi, \tau_i, \nu_j \sim \mathrm{Beta}(\mu(\mathrm{s_i, t_j})\phi, (1-\mu(\mathrm{s_i, t_j}))\phi),$$

$$\boldsymbol{\tau} \mid \eta \sim \mathrm{N_n}(0, \Sigma_\tau),$$

$$\nu_j \mid \varphi \sim \mathrm{N}\left(\nu_{\mathrm{j}-1}, \sigma^2_\nu\right), \mathrm{j} = 2, \ldots, \mathrm{t_T},$$

که در آن $\eta = (\sigma^2_., \psi^{-1})$ و φ به ترتیب پارامترهای مربوط به ساختار کوواریانس فضایی و مدل سری زمانی هستند. با در نظر گرفتن توزیع‌های پیشین دلخواه $\pi(\boldsymbol{\beta})$ ، $\pi(\phi)$، $\pi(\sigma^2_.)$،$\pi(\psi^{-1})$و $\pi(\sigma^2_\nu)$ توزیع پسین توام پارامترها به‌صورت

$$f\left(\boldsymbol{\beta}, \sigma^2_., \psi^{-1}, \sigma^2_\nu, \phi, \tau, \nu \mid y\right) \propto \prod_{i=1}^{s_n}\prod_{j=1}^{t_T} f\left(y_{ij} \mid \tau_i, \nu_j, \phi\right) \prod_{i=1}^{s_n} f\left(\tau_i \mid \beta, \sigma^2_., \psi^{-1}\right) \prod_{j=1}^{t_T} f\left(\nu_j \mid \sigma^2_\nu\right) \tag{۵.۲}$$

$$\times\, \pi(\beta)\, \pi\left(\sigma^2_.\right)\, \pi\left(\psi^{-1}\right)\, \pi\left(\sigma^2_\nu\right)\, \pi(\phi),$$

حاصل می‌شود که داری فرم بسته نیست و بنابراین برآورد پاارمترها با استفاده از نمونه‌گیری گیبز از توزیع‌های شرطی کامل حاصل می‌شود.

۳ مطالعه شبیه‌سازی

در این بخش با استفاده مطالعه شبیه‌سازی براورد پارامترهای مدل با رهیافت بیزی به‌دست می آیند. تولید داده‌ها با در نظر گرفتن مقادیر $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top = (-1, 1/5)^\top$، $\varphi = (\varphi_1, \varphi_2, \varphi_3) = (0/45, -0/2, 0/15)$ و با به کار بردن تابع کوواریانس همسانگرد نمایی $(\Sigma_\tau)_{i\ell} = \sigma^2_\tau \exp(-|s_i - s_\ell|\psi^{-1})$ انجام شده است. متغیرهای کمکی از توزیع یکنواخت تولید شده‌اند. پاسخ‌ها به شرط متغیر پنهان و مؤلفه زمان از توزیع بتا با پارامترهای $(\mu_{ij}\phi, (1-\mu_{ij})\phi)$ تولید شده‌اند که در آن، $\phi = 50$ و $\mu_{ij} = (\exp(\mathbf{x}_{ij}\boldsymbol{\beta}+\tau_i+\nu_j))/(1+\exp(\mathbf{x}_{ij}\boldsymbol{\beta}+\tau_i+\nu_j))$. برای برآورد بیزی پارامترهای مدل، توزیع‌های پیشین‌های متداول به‌صورت $\beta_i \sim \mathrm{N}(0, 100)$، $\sigma^2_\tau \sim \mathrm{IG}(0/01, 0/01)$، $\psi^{-1} \sim \mathrm{U}(0/01, 20)$، $\nu_1 \sim \mathrm{N}(0, \sigma^2_{\nu_1})$، $\nu_j \sim \mathrm{N}(\nu_j, \sigma^2_\nu)$، به ازای $j > 1$، $\sigma^2_\nu \sim \mathrm{IG}(0/01, 0/01)$ و $\phi \sim \mathrm{G}(0/01, 0/01)$ در نظر گرفته می‌شوند. مقادیر پارامترهای توزیع‌های پیشین به‌گونه‌ای انتخاب شده‌اند که توزیع‌های پیشین ناآگاهی بخش[2] باشند. نتایج شبیه‌سازی بر روی یک مشبکه 10×10 با $h = 1000$ تکرار برای مدل‌های $AR(1)$، $AR(2)$ و $AR(3)$ در جدول ۱ آمده است. اریبی نسبی RelBias و مجذور میانگین مربعات خطا (RMSE) برای هر پارامتر θ از روابط

جدول ۱: براورد پارامترهای مدل رگرسیون خطی تعمیم‌یافته بتا فضایی‌ـ زمانی

		پارامترها							
		β_0	β_1	φ_1	φ_2	φ_3	ψ^{-1}	σ^2_τ	ϕ
مدل	مقدار واقعی	-1	1/5	0/45	$-0/2$	0/15	10	0/5	50
$AR(1)$	برآورد	$-1/17$	1/68	0/52	*	*	11/01	0/65	45/2
	اریبی نسبی	0/17	0/13	0/16	*	*	0/10	0/16	$-0/10$
	مجذور میانگین مربعات خطا	0/30	0/39	0/27	*	*	1/88	0/17	5/57
$AR(2)$	برآورد	$-1/14$	1/53	0/44	$-0/24$	*	9/21	0/59	45/44
	اریبی نسبی	0/14	0/02	$-0/03$	0/23	*	$-0/08$	0/18	$-0/09$
	مجذور میانگین مربعات خطا	0/31	0/22	0/21	0/15	*	1/72	0/18	5/61
$AR(3)$	برآورد	$-1/04$	1/38	0/43	$-0/17$	0/12	10/30	0/45	47/36
	اریبی نسبی	0/04	$-0/08$	$-0/04$	$-0/17$	$-0/18$	0/03	$-0/10$	$-0/05$
	مجذور میانگین مربعات خطا	0/37	0/29	0/19	0/07	0/16	1/08	0/11	7/26

$$\mathrm{RelBias}(\theta) = \frac{1}{h}\sum_{a=1}^{h}\left(\widehat{\theta}_a/\theta - 1\right), \qquad \mathrm{RMSE}(\theta) = \left\{\frac{1}{h}\sum_{a=1}^{h}\left(\widehat{\theta}_a - \theta\right)^2\right\}^{1/2} \tag{۱.۳}$$

محاسبه شده است.

۴ تحلیل سهم هزینه خوراک از کل هزینه‌های خانوار در شهر تهران

سهم هزینه خوراک خانوار از کل هزینه‌ها می‌تواند به‌عنوان معیاری برای تشخیص وضعیت رفاه خانوارها به کار گرفته شود. از آن‌جا که مقدار این سهم در بازه $(0, 1)$ تغییر می‌کند، از مدل معرفی شده در بخش قبل برای تحلیل سهم هزینه خوراک خانوارها از کل هزینه‌های خانوارهای شهر تهران در فصل‌های متوالی یک سال استفاده می‌شود. بعد خانوار، درآمد، و سطح زیر بنای محل سکونت متغیرهای کمکی هستند که در مدل‌بندی داده‌ها مورد بررسی قرار گرفته‌اند. برآورد پارامترها پس

[2] Noninformative
[3] Deviance Information Criterion

از داغیدن ۵۰۰۰۰۰ نمونه از ۷۵۰۰۰۰ نمونه توزیع پسین و انتخاب یک نمونه از هر ۵۰ نمونه باقی‌مانده انجام می‌شود. مقایسه مقدار معیار انحراف اطلاع [۳] مدل با اثرات فضایی‌ـ زمانی (۱۲۹/۳۰−) ، مدل با اثر فضایی (۷۲/۹۱−)، مدل با مؤلفه اثر زمان (۱۲۴/۰۹−) و مدل رگرسیون بتا (۶۷/۰۵−) نشان می‌دهد که به‌کارگیری اثرات فضایی‌ـ زمانی باعث بهبود برازش مدل شده است. نتایج برازش مدل با اثرات فضایی‌ـ زمانی در جدول ۲ آمده است.

جدول ۲: برآورد پارامترها برای مدل سهم هزینه خوراک خانوار از کل هزینه‌ها در شهر تهران

منبع	پارامتر	برآورد	خطای زنجیره مارکوف	بازه باور ٪۹۵
عرض از مبدأ	β_0	−۱/۷۸	۰/۰۳	(−۳/۳۲, −۰/۵۱)
درآمد	β_1	−۰/۲۱	۰/۰۱	(−۰/۴۱, −۰/۰۰۵)
بعد خانوار	β_2	۰/۱۹	۰/۰۱	(۰/۰۰۵, ۰/۳۹)
پارامتر دقت	ϕ	۳۴/۸۷	۰/۱۴	(۲۰/۲۸, ۵۵/۵۲)
واریانس فضایی	σ^2	۳/۴۶	۰/۷۵	(۰/۹۱, ۹/۶۹)
دامنه فضایی	ψ^{-1}	۷/۳۴	۱/۴۷	(۱/۷۵, ۹/۸۳)
پارامتر سری زمانی	φ	۰/۶۶	۰/۰۸	(۰/۱۸, ۰/۹۳)

همان‌طور که در جدول ۲ ملاحظه می‌شود درآمد و بعد خانوار بر سهم هزینه خوراک از کل هزینه‌های خانوار مؤثر بوده‌اند و با افزایش درآمد سهم هزینه خوراک کاهش می‌یابد و با افزایش بعد خانوار این سهم افزایش می‌یابد. پارامترهای مربوط به اثر فضایی و سری زمانی نیز معنی دارند.

بحث و نتیجه‌گیری

مدل‌بندی داده‌های محدود در بازه (۰, ۱) از اهمیت ویژه‌ای برخوردار است. به عنوان مثال در حوزه آمار رسمی نرخ بیکاری، سهم هزینه خوراک از هزینه‌های خانوار، میزان خانوارهای بهره‌مند از خدمات سلامت و بسیاری موارد دیگر در این حوزه می‌گنجد. ارائه یک مدل مناسب با لحاظ کردن همبستگی فضایی‌ـ زمانی آن‌ها از اهمیت ویژه‌ای برخوردار است. در این مقاله یک مدل رگرسیون بتا با در نظر گرفتن اثرات تفکیک‌پذیر فضایی‌ـ زمانی معرفی شد. تحلیل سهم هزینه خوراک خانوارها از کل هزینه‌ها در شهر تهران نشان داد که علاوه بر درآمد و بعد خانوار، اثرات فضایی و زمانی نیز معنی‌دار است و لحاظ کردن آن‌ها در مدل می‌تواند در بهبود برازش مدل بر داده‌ها مؤثر باشد. در حوزه آمار رسمي، توسعه این مدل با الگوهای پیچیده‌تر همبستگی فضایی‌ـ زمانی، می‌تواند کمک مؤثری در راستای تحلیل داده‌ها و پیشگویی مناسب در جهت برنامه‌ریزی و سیاست‌گزاری‌های آتی داشته باشد.

مراجع

محمدزاده، م.، (۱۳۹۸)، *آمار فضایی و کاربردهای آن*، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

Cepeda-Cuervo, E., Urdinola, B. P., and Rodríguez, D. (2012), Double Generalized Spatial Econometric Models, *Communications in Statistics-Simulation and Computation*, **41**, 671-685.

Ferrari, S., Cribari-Nieto, F. (2004), Beta Regression for Modelling Rates and Proportions, *Journal*

of Applied Statistics, **31**, 799-815.

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B., and Ferrari, S. L. (2013), Mixed Beta Regression: A Bayesian Perspective, *Computational Statistics and Data Analysis* , **61**, 137– 147.

Kalhori Nadrabadi, L., Mohhamadzadeh, M. (2018), Bayesian Inference for Spatial Beta Generalized Linear Mixed Models,*Journal of Sciences, Islamic Republic of Iran*, **29**, 173-185.

Lagos-Alvarez, B. M., Fustos-Toribio, R., Figueroa-Zuniga, J., and Mateu, J. (2017), Geostatistical Mixed Beta regression: a Bayesian Approach,*Stochastic Environmental Research and Risk Assessment*,**31**, 571-584.