سیزدهمین کنفرانس آمار ایران
۴-۲ شهریور ۱۳۹۵

مدل رگرسیون بتا فضایی و مدل‌بندی میزان طلاق استان‌ها

لیدا کلهری ندرآبادی [۱]، محسن محمدزاده

گروه آمار دانشگاه تربیت مدرس، lida.kalhori@modares.ac.ir ، mohsen_m@modares.ac.ir

چکیده مولا مقادیر نسبت‌ها و نرخ‌ها در بازه (۱و۰) قرار دارند و در بسیاری از زمینه‌های کاربردی به عنوان تحقق‌های متغیر پاسخ در نظر گرفته می‌شوند. در سال‌های اخیر مدل رگرسیون بتا برای مدلبندی اینگونه مشاهدات معرفی شده است. این مدل شامل یک پارامتر دقت نیز هست که می‌تواند ثابت باشد یا آن هم از طریق یک تابع پیوند مدل‌بندی شود. توزیع بتا با تغییر مقادیر پارامترهایش به فرم‌های متفاوتی ظاهر می‌شود که در بردارنده فرم‌های نامتقارن نیز می‌باشد. بنابراین، مدل رگرسیون بتا در مقایسه با مدل‌های رگرسیون خطی از انعطاف‌پذیری بیشتری برخوردار است. از طرفی فرض نابستگی مشاهدات متغیرپاسخ یکی از فرض‌های اساسی مدل‌های رگرسیونی است، در حالی که برخی از داده‌ها مانند داده‌های فضایی بر حسب موقعیتی که مشاهده می‌شوند وابسته‌اند. در این مقاله مدل رگرسیون بتا فضایی و نحوه برآورد بیزی پارامترهای آن معرفی می‌شود و داده‌های نسبت طلاق به ازدواج به دست آمده از سازمان ثبت احوال کشور به عنوان کاربردی از این مدل ارائه و مدل‌بندی می‌شوند.

واژه‌های کلیدی: برآورد بیزی، مدل رگرسیونی بتا فضایی، نسبت طلاق به ازدواج

کد موضوع‌بندی ریاضی (۲۰۱۰): ۹۱D۲۵. محور تخصصی: آمار رسمی (کد ۵۰۹)

۱ مقدمه

برای بررسی رابطه متغیرهای پاسخ و متغیرهای تبیینی معمولاً با فرض آن که متغیر پاسخ یا تبدیلی از آن دارای توزیع نرمال است، واریانس آن ثابت و مؤلفه‌های خطا ناهمبسته هستند، از مدل‌های رگرسیونی یا رگرسیونی تعمیم‌یافته استفاده می‌شود. اما گاهی در عمل این فرض‌ها غیرواقع‌گرایانه هستند. یکی از اهداف استفاده از مدل‌های رگرسیونی ساختن الگوهایی مفید برای پیشگویی مناسب است. وقتی حوزه مقادیر متغیر پاسخ بازه (۱و۰) باشد، استفاده از این مدل‌ها ممکن است منجر به پیشگویی مقادیری خارج از این بازه شود. علاوه بر آن ممکن است پراکندگی نسبت‌هایی که در بازه (۱و۰) قرار دارند متقارن نباشند و برازش مدل با در نظر گرفتن توزیع نرمال

[۱] لیدا کلهری ندرآبادی : lida.kalhori@modares.ac.ir

برای متغیر پاسخ به نتایج گمراه کننده‌ای منجر شود. یک رهیافت برای حل این مسأله استفاده از تبدیل‌هایی مناسب بر متغیر پاسخ است، به گونه‌ای که مقادیر آن در مجموعه اعداد حقیقی واقع شوند و سپس مدل‌ها بر داده‌های تبدیل یافته برازش شوند. در این‌صورت پارامترهای مدل برازانده شده به داده‌های تبدیل یافته در واقع مفسر داده‌های تبدیل یافته هستند و بیان روشنی برای توضیح رابطه بین متغیرهای تبیینی و میانگین داده‌های اولیه به دست نمی‌دهند. 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. این توزیع شامل دو پارامتر میانگین و دقت است که می‌توانند ثابت یا متغیر باشند. وقتی پارامتر دقت ثابت باشد، شرط همگنی واریانس برقرار است و می‌توان مانند مدل‌های تعمیم‌یافته عمل نمود. اما در حالاتی که پارامتر دقت متغیر باشد، شرط همگنی واریانس برقرار نیست و لازم است علاوه بر میانگین، با استفاده از توابع پیوند مناسب مدلی برای پارامتر دقت نیز در نظر گرفت. این مدل می‌تواند تابعی از متغیرهای تبیینی باشد که لزوما با متغیرهای تبیینی مدل میانگین یکسان نیستند. فراری و کریباری(۲۰۰۴) الگویی خطی برای متغیرهای تبیینی در نظر گرفتند و بدون استفاده از تبدیل بر متغیر پاسخ، به برازش مدل پرداخته و پارامترهای آن را به روش ماکسیمم درستنمایی برآورد کردند. در این مدل‌ها از اصول مدل‌های خطی تعمیم‌یافته نلدر و مک‌کالا (۱۹۸۹) استفاده شده است که در آن‌ها متغیرهای تبیینی و میانگین متغیر پاسخ از طریق یک تابع پیوند مناسب به هم مرتبط می‌شوند. کپدا و گامرمن (۲۰۰۵) و اسمیتسون و ورکولین (۲۰۰۶)، مدل رگرسیون بتا را با مدل‌بندی همزمان میانگین و پارامتر دقت توسعه دادند. در این مطالعات لوجیت میانگین و لگاریتم پارامتر دقت با در نظر گرفتن الگویی خطی به متغیرهای تبیینی ربط داده شده‌اند. مدل آمیخته خطی تعمیم‌یافته بتا اولین بار توسط زیمپریچ (۲۰۱۰) در مطالعه داده‌های طولی از طریق وارد کردن اثر تصادفی در مدل میانگین برای بررسی تغییر سرعت واکنش افراد نسبت به محرک‌ها با افزایش سن، مطرح شد. فیگارو و همکاران (۲۰۱۳) برآوردهای بیزی پارامترهای مدل آمیخته خطی تعمیم‌یافته بتا را در دو وضعیت ثابت و متغیر بودن پارامتر دقت، با رهیافت بیزی ارایه کردند. برای مطالعه اثرات فضایی در مدل رگرسیون بتا، کپدا و همکاران (۲۰۱۲) مدل‌بندی همزمان پارامترهای میانگین و دقت را با در نظر گرفتن یک الگوی همبستگی برای لوجیت میانگین و لگاریتم پارامتر دقت پیشنهاد دادند. در مطالعه آن‌ها مداخله اثر فضایی با استفاده از حاصل ضرب ماتریس وزن در متغیر پاسخ به عنوان یک متغیر تبیینی مدل انجام شد. کپدا و نونز (۲۰۱۳) این مدل را برای مطالعه کیفیت تحصیلات در کلمبیا مورد استفاده قرار دادند.

آمار ارائه شده توسط سازمان ثبت احوال ایران نشان می‌دهد که نسبت طلاق به ازدواج در استان تهران نسبت به سایر استان‌ها بیشتر است. پایتخت به عنوان مرکز سیاسي کشور ، بیشترین امکانات عمرانی و تولیدي و تسهیلات اجتماعي و رفاهي را در خود جاي داده است و همچنین نقش این استان در امور آموزش عالي، بهداشت و درمان و صنعت، وجود فرصت‌هاي شغلي با سایر استان‌ها متفاوت است. با توجه به اهمیت پیامدهای ناشی از طلاق، هدف از این مقاله بررسی برخی از عوامل اجتماعی و اقتصادی مؤثر بر نسبت طلاق به ازدواج است. با توجه به این که فرهنگ نواحی مختلف می‌تواند بر طلاق اثر داشته باشد، اثر فاصله جغرافیایی را به عنوان یک متغیر موثر بر طلاق مورد بررسی قرار می‌دهیم. برای این منظور نسبت طلاق به ازدواج در هر استان به نسبت طلاق به ازدواج در تهران را به عنوان متغیر پاسخ در نظر می‌گیریم و از این پس آن را میزان طلاق می‌نامیم. در بخش‌های بعد عوامل موثر بر میزان طلاق را با استفاده از مدل رگرسیون بتا فضایی مورد بررسی قرار می‌دهیم. در بخش دو مدل رگرسیون بتا و تعمیم آن به مدل رگرسیون بتا فضایی معرفی می‌شوند. در بخش سه عوامل موثر بر میزان طلاق به عنوان کاربردی از مدل بررسی شده و برآورد پارامترهای مدل با استفاده رهیافت بیزی به دست می‌آیند. بخش پایانی به بحث و نتیجه‌گیری اختصاص یافته است.

۲ مدل رگرسیون بتا فضایی

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 $Beta(a,b)$ را با فرض $\mu = \frac{a}{a+b}$ و $\phi = a+b$ به صورت

$$\pi(y,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \qquad 0 < y < 1$$

در نظر گرفتند، که در آن $0 < \mu < 1$ و $\phi > 0$ در نتیجه $E(Y) = \mu$ ، $V(Y) = \frac{\mu(1-\mu)}{1+\phi}$ ، که در آن ϕ پارامتر دقت نامیده می‌شود. در مواردی که مقادیر متغیر پاسخ در بازه (α,β) قرار گرفته باشند می‌توان $\frac{y-\alpha}{\beta-\alpha}$ را به جای متغیر پاسخ مدل‌بندی کرد.
توزیع بتای بازپارامتریده با توجه به هدف فراری و کریباری (۲۰۰۴) برای مدل‌بندی میانگین توزیع بتا تعریف شده است. اگر $y_1,...,y_n$ متغیرهای تصادفی مستقل و هر یک از توزیع $Beta(\mu_i\phi,(1-\mu_i)\phi)$ ، $i = 1,...,n$ باشند، آنگاه مدل رگرسیون بتا به‌صورت $g(\mu_i) = \sum_{j=0}^{k} x_{ij}\beta_j$ معرفی می‌شود که در آن $x_{i0},...,x_{ik}$ متغیرهای تبیینی و $\beta = (\beta_0,...,\beta_k)^T$ بردارضرایب رگرسیونی هستند. در این حالت فرض شده است که ϕ یک پارامتر نامعلوم با مقدار ثابت است. اما در حالاتی که پارامتر دقت متغیر باشد، شرط همگنی واریانس برقرار نیست و لازم است علاوه بر میانگین، با استفاده از توابع پیوند مناسب مدلی برای پارامتر دقت نیز در نظر گرفت. مدل‌بندی همزمان پارامترهای میانگین و دقت توسط کپدا و گامرمن (۲۰۰۵) و اسمیتسون و ورکولین (۲۰۰۶) انجام شد. اگر $y_1,...,y_n$ متغیرهای تصادفی مستقل از توزیع $Beta(\mu_i\phi_i,(1-\mu_i)\phi_i)$ ، $i = 1,...,n$ باشند، آنگاه مدل رگرسیون بتا را می‌توان به‌صورت $h(\phi_i) = \sum_{l=0}^{m} z_{il}\delta_l$ در نظر گرفت، که در آن همانند قبل $z_{i0},...,z_{il}$ متغیرهای تبیینی و $\delta = (\delta_0,...,\delta_m)^T$ بردار ضرایب رگرسیونی هستند. این مدل می‌تواند تابعی از متغیرهای تبیینی باشد که لزوما با متغیرهای تبیینی مدل میانگین یکسان نیستند.

از آنجاییکه در بسیاري از زمینه‌های کاربردی مانند مطالعات محیطی با داده‌های فضایی مواجه می‌شویم که برحسب موقعیت قرارگرفتنشان به یکدیگر وابسته‌اند، کپدا و همکاران (۲۰۱۲) مدل‌بندی همزمان پارامترهای میانگین و دقت را با در نظر گرفتن یک الگوی همبستگی برای لوجیت میانگین و لگاریتم پارامتر دقت پیشنهاد دادند. در مطالعه آن‌ها مداخله اثر فضایی در مدل رگرسیون بتا با استفاده از حاصل ضرب ماتریس وزن در متغیر پاسخ به عنوان یک متغیر تبیینی مدل انجام شد. مدل آن‌ها به صورت

$$logit(\mu) = X\beta + \rho WY$$

$$log(\Phi) = Z\delta + \lambda WY$$

ارائه شد که در آن ρ و λ ضرایب رگرسیونی و W ماتریس وزن فضایی و Y بردار متغیر پاسخ است. برای برآورد بیزی پارامترها با در نظر گرفتن توزیع‌های پیشین مستقل، توزیع پسین همزمان برای پارامترهای مدل رگرسیون بتا فضایی با فرض ثابت بودن یا ثابت نبودن پارامتر دقت به‌ترتیب به‌صورت

$$f(\beta,\rho,\phi \mid y) \propto \prod_{i=1}^{n} f(y_i \mid \beta,\rho,\phi) f(\rho) f(\beta) f(\phi)$$

$$f(\beta,\delta,\rho,\lambda \mid y) \propto \prod_{i=1}^{n} f(y_i \mid \beta,\delta,\rho,\lambda) f(\beta) f(\delta) f(\rho) f(\lambda)$$

و توزیع‌های شرطی کامل نیز به ترتیب به‌صورت

$$f(\beta \mid \rho,\phi,y), f(\rho \mid \beta,\phi,y), f(\phi \mid \beta,\rho,y)$$

$$f(\beta \mid \delta,\rho,\lambda,y), f(\delta \mid \beta,\rho,\lambda,y), f(\rho \mid \beta,\delta,\lambda,y), f(\lambda \mid \beta,\delta,\rho,y)$$

ارائه می‌شوند. برآورد پارامترهای مدل با نمونه‌گیری گیبز از توزیع‌های پسین شرطی به دست می‌آیند.

۳ بررسی علل موثر بر میزان طلاق

در این بخش میزان طلاق با استفاده از مدل رگرسیون بتای فضایی مدل‌بندی و تحلیل می‌شود. با استفاده از اطلاعات به دست آمده از سرشماری نفوس و مسکن ۱۳۹۰، نرخ بیکاری استان، نسبت افراد دارای تحصیلات عالی از کل افراد تحصیل کرده استان، ارزش افزوده استان، نسبت افراد ازدواج کرده به جمعیت در سن ازدواج استان و نسبت افراد در سن ازدواج به جمعیت استان به عنوان متغیرهای تبیینی کاندید در نظر گرفته می‌شوند. از آنجایی‌که متغیر پاسخ از تقسیم کردن نسبت طلاق به ازدواج در استان بر نسبت طلاق به ازدواج در تهران به دست می‌آید، متغیرهای تبیینی از روی متغیرهای فوق‌الذکر به‌گونه‌ای ساخته می‌شوند که اطلاعات هر دو استان در مدل وارد شود. بنابراین متغیر تبیینی i ام برای استان k ام، از تقسیم مقدار متغیر i ام در استان k ام به مقدار متغیر i ام در استان تهران به دست می‌آید. فاصله اقلیدسی بین مراکز استان‌ها و تهران به عنوان متغیر تبیین کننده اثر فضایی در نظر گرفته می‌شود. در واقع اثر فضایی بیانگر تاثیر موقعیت جغرافیایی بر میزان طلاق است. با فرض همگن بودن واریانس پاسخ، میزان طلاق به صورت

$$logit(\mu) = X\beta + \rho WY$$

مدل‌بندی می‌شود. برای ضرایب رگرسیونی مدل، توزیع پیشین $N(۰, ۱۰۰)$ و برای ϕ توزیع پیشین $Gamma(۰/۰۱, ۰/۰۱)$ در نظر گرفته می‌شود. پس از برازش مدل‌هایی با متغیرهای تبیینی مختلف، مدل نهایی در جدول ۱ ارائه شده است:

جدول ۱: برآوردهای بیزی پارامترهای مدل رگرسیون بتا فضایی برای میزان طلاق

ضرایب رگرسیونی	برآورد	انحراف استاندارد	بازه اطمینان ۹۵ درصد	
			کران پایین	کران بالا
عرض از مبدا	۱/۹۵	۰/۸۶	۰/۴۳	۳/۷۹
نسبت افراد ازدواج کرده	-۱/۲۷	۰/۵۶	- ۲/۳۹	-۰/۲۴
فاصله از تهران	-۰/۴۵	۰/۲۱	-۰/۸۷	-۰/۰۳

همان‌طور که ملاحظه می‌شود نسبت افراد ازدواج کرده و فاصله از تهران با اطمینان ۹۵٪ بر میزان طلاق اثر معنی‌دار دارند. بنابراین مدل رگرسیون میزان طلاق به صورت

$$logit(\mu) = ۱/۹۵ - ۱/۲۷(\text{نسبت افراد ازدواج کرده}) - ۰/۴۵(\text{فاصله})$$

حاصل می‌شود. ملاحظه می‌شود که هر چه نسبت افراد ازدواج کرده در استان نسبت به افراد ازدواج کرده در استان تهران افزایش یابد، نسبت طلاق به ازدواج در مقایسه با نسبت طلاق به ازدواج در استان کاهش می‌یابد. همچنین با افزایش فاصله از تهران میزان طلاق کاهش می‌یابد. نتیجه به دست آمده می‌تواند نشان دهنده این مطلب باشد که با افزایش فاصله از تهران به دلیل تغییر شرایط فرهنگی و میزان توسعه یافتگی استان‌ها نسبت طلاق به ازدواج کاهش می‌یابد. در ادامه مدل‌بندی همزمان میانگین و پارامتر دقت انجام می‌شود.

$$logit(\mu) = X\beta + \rho WY$$

$$log(\Phi) = Z\delta + \lambda WY$$

برای ضرایب رگرسیونی مدل، توزیع پیشین $N(۰, ۱۰۰)$ در نظر گرفته می‌شود. پس از برازش مدل‌هایی با متغیرهای تبیینی مختلف، مدل نهایی در جدول ۲ ارائه شده است.

جدول ۲: برآوردهای بیزی پارامترهای مدل رگرسیون بتا فضایی برای میزان طلاق با مدل‌بندی همزمان میانگین و پارامتر دقت

				بازه اطمینان ۹۵ درصد	
	ضرایب رگرسیونی	برآورد	انحراف استاندارد	کران پایین	کران بالا
مدل میانگین	عرض از مبدا	۱/۸۴	۰/۷۵	۰/۳۷	۳/۴۵
	نسبت افراد ازدواج کرده	-۱/۲۵	۰/۴۶	- ۲/۲۱	-۰/۳۳
	فاصله از تهران	-۰/۸۲	۰/۲۲	-۱/۲۶	-۰/۳۸
مدل پارامتر دقت	عرض از مبدا	۲/۸۵	۰/۲۶	۲/۳۱	۳/۳۲
	فاصله از تهران	-۱/۴۶	۰/۵۰	-۲/۳۸	-۰/۴۰۱

همان‌طور که ملاحظه می‌شود نسبت افراد ازدواج کرده و فاصله از تهران با اطمینان ۹۵٪ بر میزان طلاق اثر معنی‌دار دارند. بنابراین مدل رگرسیون میزان طلاق به صورت

$$logit(\mu) = ۱/۸۴ - ۱/۲۵(\text{نسبت افراد ازدواج کرده}) - ۰/۸۲(\text{فاصله})$$

$$log(\Phi) = ۲/۸۵ - ۱/۴۶(\text{فاصله})$$

به دست می‌آید. ملاحظه می‌شود که با افزایش فاصله از تهران پارامتر دقت کاهش می‌یابد، به عبارت دیگر واریانس میزان طلاق در کشور ثابت نیست و با افزایش فاصله از تهران افزایش می‌یابد. نسبت طلاق به ازدواج در استان‌های نزدیک به تهران در مقایسه با مقدار این نسبت در سایر استان‌ها اختلاف کمتری دارند. ملاک انحراف اطلاع (DIC) برای مدل‌های اول و دوم به ترتیب برابر با ۳۴/۹۳- و ۴۲/۲۵- حاصل شده است که نشان می‌دهد مدل فضایی دوم در مقایسه با مدل اول از کارایی بیشتری برخوردار است.

بحث و نتیجه‌گیری

طلاق یکی از پدیده‌های اجتماعی است که می‌تواند از دیدگاه‌های مختلف مورد بررسی قرار گیرد. چون امکانات فرهنگی، سیاسی و اقتصادی در استان تهران تمرکز یافته است و این استان از بعد توسعه در مقایسه با سایر استان‌ها پیشرو است در این مقاله سعی شده بر

اساس نتایج سرشماری نفوس و مسکن ۱۳۹۰، نسبت طلاق به ازدواج در استان‌ها با نسبت طلاق به ازدواج تهران مقایسه شود. نرخ بیکاری استان، نسبت افراد دارای تحصیلات عالی از کل افراد تحصیل کرده استان، ارزش افزوده استان، نسبت افراد ازدواج کرده استان به جمعیت در سن ازدواج استان و نسبت افراد در سن ازدواج به جمعیت استان به عنوان متغیرهای تبیینی کاندید در نظر گرفته شدند. با توجه به ماهیت متغیر پاسخ که از نوع نرخ است و دامنه تغییرات آن به بازه (۱و۰) محدود می‌شود و علاوه بر آن بین مشاهدات متغیر پاسخ وابستگی فضایی وجود دارد، مدل رگرسیون بتا فضایی برای مدل بندی مشاهدات معرفی و مورد استفاده قرار گرفت. نتایج به دست آمده حاکی از آن است که واریانس پاسخ ثابت نیست و به فاصله استان‌ها با استان تهران رابطه مستقیم دارد. علاوه بر آن میانگین میزان طلاق با نسبت افراد ازدواج کرده استان به نسبت افراد ازدواج کرده تهران رابطه معکوس دارد، این نتیجه می‌تواند نشان دهنده این باشد که شانس وقوع طلاق در سایر استان‌ها در مقایسه با تهران کمتر است.

مراجع

نتایج تفصیلی سرشماری عمومی نفوس و مسکن ۱۳۹۰ (۱۳۹۱) تهران، مرکز آمار ایران .

نسبت طلاق به ازدواج ۱۳۹۰، سازمان ثبت احوال کشور.

Cepeda, E. .Dand Gamerman, D. (2005), Bayesian Methodology for ModelingP arameters in the Two ParameterE xponentialF amily, *Revista Estadística,***57**, 168-169.

Cepeda, E. .D and Núñez, .A V. (2013), Spatial Double Generalized Beta Regression Models Extensions and Application to Study Quality of Education in Colombia, *Journal of Educational and Behavioral Statistics,* **38**, 604-628.

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

Ferrari, S. and Cribari-Nieto, F. (2004), Beta Regression for ModellingR ates 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 Data Analysis,* **61**, 137– 147.

Nelder, J. A. and McCullagh, P. (1989), *Generalized Linear Models,* Second Edition, Chapman and Hall, London.

Smithson, M. and Verkuilen, J. (2006), A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables, *Psychological Methods,* **11**, 54-71.

Zimprich, D. (2010), Modeling Change in SkewedV ariables using Mixed Beta Regression Models, *Research in Human Development,* 7, 9.