مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۵، شماره‌ی ۲، پاییز و زمستان ۱۳۹۳، صص ۱۶۱ – ۱۷۴

تحلیل قیمت مسکن شهر تهران با استفاده از مدل‌های جمعی تعمیم‌یافته

علی محمدیان مصمم* و ملیحه عباسی

دانشگاه زنجان

چکیده: مدل‌های جمعی تعمیم‌یافته در سال‌های اخیر به‌طور گسترده‌ای مورد استفاده قرار گرفته‌اند. در حالی که روش‌های برازش مدل‌های جمعی تعمیم‌یافته بر روی داده‌های مستقل به‌خوبی گسترش یافته‌اند، اما بر روی داده‌هایی که وابستگی فضایی دارند مطالعات زیادی انجام نگرفته است. در این مقاله کاربرد این مدل‌ها را بر روی داده‌های قیمت مسکن شهر تهران که وابستگی فضایی دارند، مورد تحلیل قرار خواهیم داد.

واژگان کلیدی: مدل‌های جمعی تعمیم‌یافته؛ مدل‌های خطی تعمیم‌یافته؛ وابستگی فضایی؛ توابع هموار.

۱- مقدمه

مدل‌های جمعی تعمیم‌یافته (GAM)[1] اولین بار توسط هستی و تیبشیرانی [۷] و [۸] مطرح شدند. در این مدل‌ها فرض می‌شود که میانگین متغیر پاسخ از طریق یک تابع پیوند به متغیر پیشگو وابسته است. همانند مدل‌های خطی تعمیم‌یافته، در مدل‌های جمعی تعمیم‌یافته نیز متغیر پاسخ می‌تواند هر توزیعی از خانواده‌ی نمایی را دارا باشد. تنها تفاوت بین مدل‌های خطی تعمیم‌یافته و مدل‌های جمعی تعمیم‌یافته این است که در مدل‌های جمعی تعمیم‌یافته، توابع هموار نامعلوم می‌توانند به‌عنوان پیشگو در مدل حضور داشته باشند. قدرت مدل‌های جمعی تعمیم‌یافته، توانایی آن‌ها برای کار با روابط ناخطی و غیریکنوا بین متغیر پاسخ و مجموعه‌ی متغیرهای تبیینی می‌باشد. به‌دلیل انعطاف بالای

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت: ۱۳۹۳/۸/۲۶، پذیرش: ۱۳۹۴/۶/۲۴.

مدل‌های جمعی تعمیم‌یافته، این مدل‌ها در سال‌های اخیر به‌طور گسترده‌ای مورد استفاده قرار گرفته‌اند. برای مثال به [۱]، [۴]، [۶]، [۹]، [۱۰]، [۱۵] و [۱۶] مراجعه شود.
با وجود این که مطالعات فراوانی در ارتباط با برازش مدل‌های جمعی تعمیم‌یافته و مدل‌های رگرسیون ناپارامتری با داده‌های مستقل صورت گرفته است [۵]، [۱۳]، [۱۷]، اما در مورد داده‌های همبسته بررسی‌هایی زیادی انجام نشده است که به اندک مقاله‌های موجود در [۱۴] می‌توان اشاره کرد.
ساختار کلی مقاله به‌صورت زیر می‌باشد: در بخش دوم مروری بر مفاهیم مقدماتی آمار فضایی انجام می‌گیرد. در بخش۳ مدل‌های پارامتری و در بخش ۴ مدل‌های ناپارامتری معرفی می‌شوند. در بخش ۵ سعی داریم با یک مثال کاربردی به بررسی عملکرد مدل‌های جمعی تعمیم‌یافته با وابستگی فضایی و بدون وابستگی فضایی و نیز مقایسه‌ی عملکرد این مدل‌ها در مقایسه با مدل‌های خطی تعمیم‌یافته بپردازیم. مقاله با یک بحث و نتیجه‌گیری در بخش ۶ به پایان می‌رسد.

۲- آمار فضایی

در اغلب روش‌های آماری معمولاً فرض بر این است که مشاهدات تحت شرایط یکسان و به‌صورت مستقل از هم جمع‌اوری شده‌اند. فرض استقلال کمک شایانی به تسهیل مبانی نظری می‌نماید. اما در عمل ممکن است این فرض ما را از واقعیت دور کرده و موجب از بین رفتن اطلاعات زیادی شود. در مطالعات میدانی اغلب با داده‌هایی سروکار داریم که مستقل از یکدیگر نیستند و به‌طور خاص وابستگی آن‌ها ناشی از موقعیت و مکان قرار گرفتن داده‌ها در فضای مورد مطالعه است. به‌دلیل وجود وابستگی فضایی بین آن‌ها، روش‌های معمول آماری برای تحلیل چنین داده‌هایی که داده‌های فضایی نامیده می‌شود، قابل استفاده نیستند.
برای تحلیل و مدل‌سازی چنین داده‌هایی لازم است به‌نحوی ساختار وابستگی داده‌ها در تحلیل آن‌ها لحاظ شود. بدیهی است که این وابستگی باید تابعی از فاصله‌ی بین موقعیت مشاهدات باشد، به‌طوری که مشاهدات نزدیک به هم وابستگی بیش‌تر و مشاهدات دورتر از هم، وابستگی کم‌تری داشته باشند. بدیهی است که تحلیل آماری داده‌های فضایی با روش‌های آماری معمول، مقدور نیست، زیرا شرط اساسی که همان استقلال داده‌ها است، محقق نمی‌باشد. لذا شاخه‌ی آمار فضایی برای تحلیل این‌گونه مشاهدات شکل گرفته و در

حال توسعه یافتن و فراهم آوردن فن‌های مختلف می‌باشد. در واقع آمار فضایی یکی از شاخه‌های علم آمار است که در آن به بررسی متغیرهایی پرداخته می‌شود که از خود ساختار وابستگی فضایی نشان می‌دهند و تلاش می‌شود این ساختار، که همان ارتباط بین مقادیر متغیر و فاصله و جهت قرارگیری آن‌ها است، تعیین و برای افزایش دقت در تحلیل آماری آن‌ها مورد استفاده قرار گیرد. برای توضیحات بیش‌تر به [۳] و [۱۲] مراجعه شود.

۳- مدل‌های پارامتری

۳-۱- مدل‌های خطی

مدل‌های خطی مدل‌های آماری هستند که در آن یک متغیر پاسخ به‌صورت ترکیب خطی از یک پیشگو و جمله‌ی خطای تصادفی با میانگین صفر نمایش داده می‌شود. همان‌گونه که از اسم آن مشخص است در مدل‌های خطی متغیر پیشگو به‌صورت خطی به متغیرها وابسته است. استنباط آماری از چنین مدل‌هایی معمولاً بر این فرض استوار است که متغیر پاسخ توزیع نرمال دارد. مدل‌های خطی به‌طور گسترده‌ای در بیش‌تر شاخه‌های علوم، هم در طرح آزمایش‌ها و هم برای سایر کارهای مدل‌سازی مانند رگرسیون چندگانه به کار برده می‌شود.

تعداد n مشاهده‌ی x_i و y_i را در نظر بگیرید که y_i مشاهده‌ای از متغیر تصادفی Y_i است. فرض کنید یک مدل مناسب برای ارتباط بین ماترس طرح X و بردار پاسخ Y به‌صورت زیر باشد:

$$Y_i = \mu_i + \varepsilon_i', \qquad i = ۱, \dots, n,$$

به‌طوری که $\mu_i = X_i\beta$ که در آن β یک پارامتر مجهول است و O_i'ها متغیرهای تصادفی به‌طور توام مستقل با میانگین صفر و واریانس $\sigma^۲$ هستند. بنابراین Y_i دارای میانگین $\mu_i \equiv E(Y_i)$ می‌باشد.

۳-۲- مدل‌های خطی تعمیم‌یافته

مدل‌های خطی تعمیم‌یافته [۱۱] به متغیر پاسخ اجازه می‌دهند که توزیعی غیر از توزیع نرمال را نیز داشته باشد و درجه‌ای از ناخطی بودن را در ساختار مدل می‌پذیرند. در

مدل‌های خطی تعمیم‌یافته تا حدی فرض خطی بودن اکید مدل‌های خطی قابل اغماض است، به این صورت که مقدار مورد انتظار متغیر پاسخ به یک تابع یکنوای هموار از پیشگو خطی وابسته می‌باشد. همچنین فرض این که متغیر پاسخ توزیع نرمال دارد را می‌توان حذف کرد. بنا بر این متغیر پاسخ می‌تواند هر توزیعی از خانواده‌ی نمایی (مثل توزیع نرمال، پواسون، دوجمله‌ای، گاما) را داشته باشد. مدل خطی تعمیم‌یافته ساختار اصلی زیر را دارد:

$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i' \boldsymbol{\beta}$$

که در آن $\boldsymbol{\mu}_i \equiv E(\boldsymbol{Y}_i)$، g یک تابع پیوند یکنوای هموار، بردار $\boldsymbol{X}_i'$، iامین سطر از ماتریس طرح $\boldsymbol{X}$ و $\boldsymbol{\beta}$ بردار پارامترهای نامعلوم است. چون مدل‌های خطی تعمیم‌یافته از طریق پیشگوی خطی $\boldsymbol{X}'\boldsymbol{\beta}$ تعیین می‌شود، بیش‌تر مفاهیم و نظریه‌های کلی مدل‌سازی خطی با مقداری تغییر، به مدل‌سازی خطی تعمیم‌یافته انتقال می‌یابد. تحلیل مدل خطی تعمیم‌یافته همانند مدل‌های خطی است، به جز این که باید یک توزیع و تابع پیوند انتخاب شوند. البته اگر تابع همانی به‌عنوان پیوند انتخاب شود، همراه با توزیع نرمال، در این‌صورت مدل‌های خطی معمولی به‌عنوان یک حالت خاص مدل‌های خطی تعمیم‌یافته می‌توانند در نظر گرفته شوند.

برای تعمیم مدل‌های خطی به مدل‌های خطی تعمیم‌یافته ملزم به پرداخت هزینه هستیم. در مدل‌های خطی تعمیم‌یافته برازش مدل باید با استفاده از آلگوریتم‌های تکراری انجام شود و نتایج توزیعی که برای استنباط به‌کار می‌روند تقریبی هستند.

۴- مدل‌های ناپارامتری

مدل ناپارامتری عمومی به شکل زیر می‌باشد:

$$Y - f\left(X_1, X_2, \dots, X_k\right) + \varepsilon'$$

مدل ناپارامتری چندمتغیره می‌تواند همه‌ی اثرات متقابل بین متغیرهای مستقل روی Y را در نظر بگیرد. این مدل در شرایط زیر ایده‌آل است:

الف) بیش‌تر از دو متغیر پیشگو نداشته باشیم،

ب) الگوی ناخطی بودن پیچیده باشد و بنا بر این به‌راحتی نتوان با یک تبدیل ساده مثل رگرسیون (خطی) چندگانه آن را مدل‌سازی کرد، و
ج) اندازه‌ی نمونه به قدر کافی بزرگ باشد.
همچنین تفسیر مدل ناپارامتری عمومی وقتی که به آن متغیرهای تبیینی بیش‌تری اضافه می‌کنیم غیر ممکن و ناپایدار می‌شود. این محدودیت‌ها ما را به سمت مدل‌های جمعی و مدل‌های جمعی تعمیم‌یافته سوق می‌دهند.

۱-۴- مدل‌های جمعی تعمیم‌یافته

مدل جمعی تعمیم‌یافته یک مدل خطی تعمیم‌یافته با یک پیشگوی خطی بوده که شامل مجموعی از توابع هموار برای متغیرهاست. مدل جمعی تعمیم‌یافته در حالت کلی ساختاری به شکل زیر دارد:

$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i'\boldsymbol{\beta} + \sum_{i=1}^{m} \boldsymbol{f}_j(\boldsymbol{x}_{ij}) \qquad (1)$$

که در آن iامین متغیر پاسخ دارای توزیعی از خانواده‌ی نمایی، $\boldsymbol{\mu}_i \equiv E(\boldsymbol{Y}_i)$، $\boldsymbol{X}_i'$، iامین سطر ماتریس طرح برای مؤلفه‌های اکیداً پارامتری، $\boldsymbol{f}_j$ها توابعی هموار از متغیرهای کمکی $\boldsymbol{x}_{ij}$ و g یک تابع پیوند معلوم و یکنوا می‌باشد که دارای خاصیت دوبار مشتق‌پذیری است. به‌عبارت دیگر هدف مدل‌های جمعی تعمیم‌یافته بیشینه کردن کیفیت پیش‌بینی متغیر پاسخ Y با انواع توزیع‌های نمایی از طریق براورد توابع ناپارامتری از متغیرهای پیشگو می‌باشد که از طریق یک تابع پیوند با متغیر مستقل در ارتباط هستند. این مدل وابستگی تقریبی متغیر پاسخ را روی متغیرهای کمکی، به‌جای استفاده از روابط پارامتری، از طریق نمایش توابع هموار تعیین می‌کند.
مسلماً برای مدلی با این انعطاف‌پذیری بالا باید هزینه‌ای پرداخت شود. برای برازش چنین مدلی علاوه بر نمایش توابع هموار، باید به روشی مناسب آن‌ها را هموار ساخت. روش‌های مختلفی برای هموارسازی توابع تک‌متغیره و چندمتغیره وجود دارد [۸]. یک هموارگر ابزاری برای توصیف متغیر پاسخ Y به‌عنوان تابعی از متغیرهای کمکی $X_1, \ldots, X_p$ می‌باشد. این ابزار براوردی از روند موجود برای متغیر پاسخ را ظاهر می‌سازد که کم‌تر از خودY تغییرپذیری دارد و در نتیجه هموارگر نامیده می‌شود. ویژگی مهم هموارگر، طبیعت ناپارامتری آن است؛ زیرا هموارگر الگوی پارامتری برای وابستگی

Y روی $X_1, \ldots, X_p$ را در نظر نمی‌گیرد. از جمله هموارگرهای مهم می‌توان هموارگر اسپلاین تجربی درجه‌ی سه برای توابع تک‌متغیره و هموارگر صفحه باریک برای توابع چندمتغیره را نام برد. در این مقاله از هموارگر اسپلاین تجربی درجه‌ی سه برای هموارسازی استفاده شده است.

۲-۴- ملاک‌های انتخاب مدل

۱-۲-۴- ملاک انحراف

انحراف (Deviance) یا آماره‌ی نسبت درستنمایی، برای مدل برازش شده‌ی $\hat{\mu}$ به‌صورت

$$D(y, \hat{\mu}) = ۲\{L(\mu_{\max}; y) - L(\hat{\mu}; y)\},$$

تعریف می‌شود که در آن $\mu_{\max}$ مقدار پارامتری است که $L(\mu; y)$ را روی تمام μها بیشینه می‌کند. به چنین مدلی مدل اشباع‌شده[۲] می‌گویند. در واقع معیار انحراف، همانند معیار مجموع توان‌های دوم مانده برای مدل‌های عمومی است و می‌توان از آن برای دست‌یابی به نیکویی برازش و مقایسه‌ی مدل‌ها استفاده کرد.

۲-۲-۴- معیار اطلاع آکائیکه

در حالت کلی معیار اطلاع آکائیکه (AIC)[۳] برابر است با:

$$\text{AIC} = ۲\{k - \ln(L)\}$$

که در آن k تعداد پارامترهای مدل و L بیشینه‌ی مقدار تابع درستنمایی برای مدل براوردشده است. از میان مدل‌های به کار رفته برای برازش به داده‌ها، مدلی ترجیح داده می‌شود که کم‌ترین مقدار AIC را دارا باشد [۲].

۵- تحلیل داده‌های قیمت مسکن

قیمت زمین و مسکن از چند سال پیش تا کنون به طور برگشت‌ناپذیری رو به فزونی گذاشته است؛ به‌طوری که سودآوری سرمایه‌های انباشته در زمین و مسکن، قابل مقایسه

با هیچ یک از بخش‌های دیگر اقتصاد کشور نیست و زمین و مسکن، به جای کالاهای مصرفی بادوام، به کالاهای سرمایه‌ای پربازده تبدیل شده‌اند.
در سال گذشته نیز بار دیگر قیمت زمین و مسکن افزایش فزاینده‌ای یافت. این معضل در شهر تهران چشمگیرتر بود. از این رو در این بخش سعی داریم عوامل مختلف تأثیرگذار در قیمت مسکن را در شهر تهران مورد بررسی قرار داده و کاربرد مدل‌های جمعی تعمیم‌یافته را با این داده‌ها شرح دهیم.
داده‌های مورد مطالعه شامل قیمت فروش واحدهای مسکونی مناطق ۲۲ گانه‌ی شهر تهران در مرداد ماه سال ۱۳۹۲ می‌باشد. پس از اصلاح و خذف برخی داده‌های دورافتاده بررسی را با ۳۱۸۲ داده انجام می‌دهیم. در این مطالعه قیمت کل یک واحد مسکونی به عنوان متغیر پاسخ در نظر گرفته شده و متغیرهای تبیینی عبارت‌اند از: تعداد اتاق خواب، متراژ، طبقه، سن بنا و طول و عرض جغرافیایی مناطق ۲۲ گانه‌ی شهر تهران.
در این مطالعه دو مدل جمعی تعمیم‌یافته و یک مدل خطی تعمیم‌یافته را به داده‌ها برازش می‌دهیم.

۱-۵- برازش مدل‌های جمعی تعمیم‌یافته با ۴ متغیر تبیینی

در این قسمت مدل‌های جمعی تعمیم‌یافته (۱) را با در نظر گرفتن قیمت کل یک واحد مسکونی به‌عنوان متغیر پاسخ به متغیرهای تبیینی تعداد اتاق خواب، متراژ، طبقه و سن بنا برازش می‌دهیم.
مدل‌های مورد نظر به‌صورت زیر می‌باشد:

$$Y_1 = f_1(Bedrooms) + f_2(Total\ Room) + f_3(Floor) + f_4(Age) + f_5(x, y) + \varepsilon$$

$$Y_2 = f_1(Bedrooms) + f_2(Total\ Room) + f_3(Floor) + f_4(Age) + \varepsilon$$

همچنین وابستگی فضایی را به‌صورت $f(x, y)$ وارد مدل می‌کنیم که x و y طول و عرض جغرافیایی هر منطقه می‌باشد. این تابع دومتغیره یک تابع هموار بوده که به‌صورت ناپارامتری با استفاده از روش‌های اسپلاین براورد می‌شود. یک روش استفاده از اسپلاین نوار باریک[۴] می‌باشد و روش دیگر استفاده از فرض تفکیک‌پذیری این تابع و در نتیجه

استفاده از حاصل ضرب کرونیکر است که هر دو روش در تابع gam در بسته‌ی آماری mgcv در نرم افزار R قابل اجرا می‌باشد.
علاوه بر دو مدل بالا مدل خطی تعمیم‌یافته را از طریق تابع lm به داده‌ها برازش می‌دهیم. در جدول ۱ ملاک انحراف و AIC و همچنین احتمال پوشش بازه‌های اطمینان ۹۵% (CI) نشان داده شده است. برای محاسبه‌ی احتمال پوشش بازه‌های اطمینان ۹۵%، که در واقع نسبت نقاط داده‌هایی است که توسط بازه‌های اطمینان ۹۵% پوشش داده شده‌اند، به این صورت عمل می‌کنیم: ابتدا با استفاده از رسم نمودار Q-Q از نرمال بودن مانده‌های هر مدل اطمینان حاصل می‌کنیم. سپس از طریق تابع predict میانگین و انحراف معیار هر یک از مدل‌ها را تولید کرده و بازه‌های اطمینان ۹۵% را برای آن‌ها می‌سازیم. حال مجموع تعداد نقاطی را که در آن متغیر پاسخ مورد نظر((log(Price) درون این بازه‌های اطمینان قرار می‌گیرند محاسبه کرده و بر تعداد کل داده‌ها تقسیم می‌کنیم و از این طریق نسبت پوشش بازه‌های اطمینان ۹۵% حاصل می‌شوند. ملاحظه می‌شود که بازه‌های اطمینان تولیدشده از طریق مدل اول (با وابستگی فضایی) بیش‌ترین میزان پوشش را دارد یعنی متغیر پاسخ مورد مطالعه توسط این مدل بهتر از سایر مدل‌ها توصیف می‌شود. کم‌ترین میزان پوشش را مدل خطی دارد که نشان می‌دهد در این مثال مدل خطی بدتر عمل می‌کند. از طرفی همان‌طور که ملاحظه می‌شود به طور کلی میزان پوشش بازه‌های اطمینان برای تمامی مدل‌ها، مقدار کمی است و از میان این مدل‌ها مدل اول بهتر عمل کرده است. این نشان می‌دهد که نیاز داریم متغیر تبیینی دیگری نیز به مدل‌ها اضافه کنیم تا درصد پوشش مطلوبی به‌دست آوریم. همچنین ملاحظه می‌شود که مدل اول کم‌ترین مقدار AIC را دارد، در نتیجه بهترین مدل در بین مدل‌های مذکور می‌باشد. از طرفی با توجه به مقدار انحراف‌ها ملاحظه می‌شود که میزان انحراف مدل خطی نسبت به مدل‌های دیگر بسیار چشمگیرتر است. نتایج نشان می‌دهد مدل خطی برای این داده‌ها بدتر از سایر مدل‌ها عمل می‌کند.

جدول ۱- درصد پوشش متغیر پاسخ توسط بازه‌های اطمینان و ملاک‌های انتخاب مدل

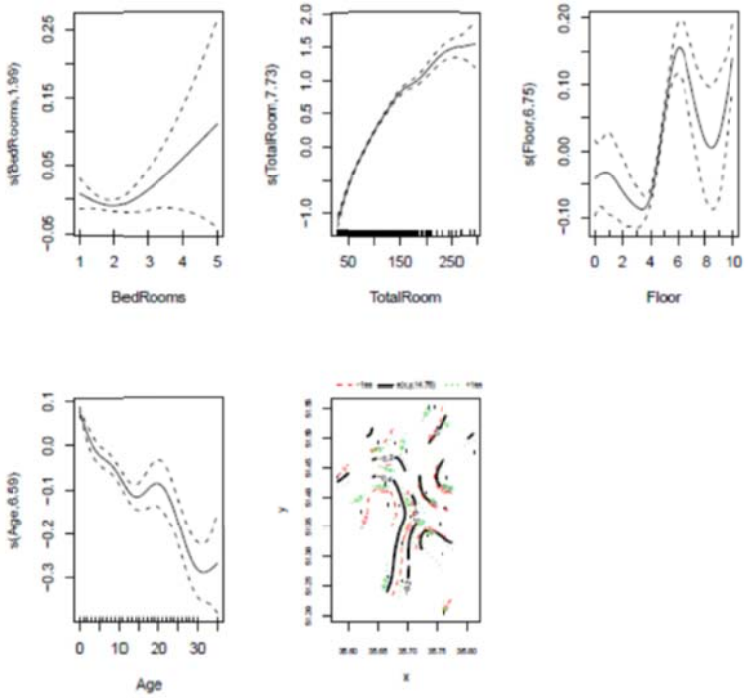
مدل	۹۵% پوشش CI	انحراف	AIC
مدل جمعی تعمیم‌یافته با وابستگی فضایی	۱۵/۳۱	۱۸۴/۱۶	۳۷/۷۲
مدل جمعی تعمیم‌یافته بدون وابستگی فضایی	۱۱/۴۲	۲۲۷/۵۹	۶۷۹/۷۶
مدل خطی تعمیم‌یافته	۵/۹۹	۲۷۳/۹۱	۱۲۳۵/۳۲

برای این که رفتار متغیرهای تبیینی مختلف را در برابر متغیر پاسخ نشان دهیم، توابع هموار هر کدام از متغیرها را برای مدل اول رسم می‌کنیم. در شکل ۱ خط توپر، مقدار پیش‌بینی‌شده‌ی توابع هموار را به‌صورت تابعی از محور افقی نشان می‌دهد. خطوط نقطه‌چین دو برابر انحرافات استاندارد هستند. محور عمودی نشان‌دهنده‌ی اثر ناخطی متغییر تبیینی روی متغیر پاسخ بوده وقتی که مدل طوری استاندارد شده است که میانگین پاسخ صفر شود.

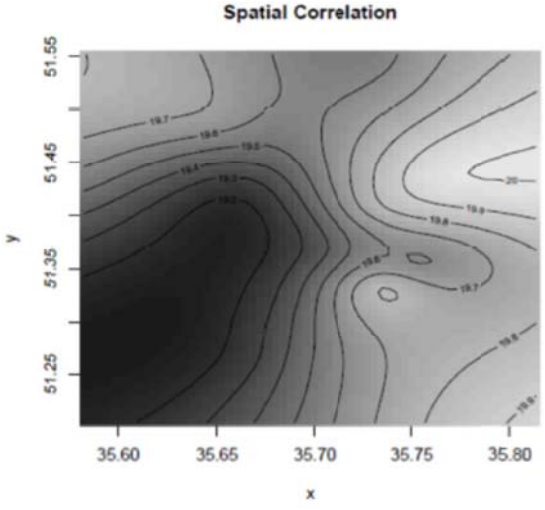
طبق نمودار تقریباً تمامی توابع هموار معنی‌دار هستند. با افزایش متغیر تعداد اتاق خواب متغیر پاسخ افزایش می‌یابد، اما برای تعداد اتاق خواب بیش‌تر از ۳ پراکندگی متغیر پاسخ بیش‌تر است، به‌عبارتی واریانس متغیر پاسخ برای تعداد اتاق خواب‌های زیاد، بیش‌تر است. با افزایش متغیر متراژ، متغیر پاسخ یعنی قیمت کل افزایش می‌یابد اما پراکندگی در متراژهای بالاتر، بیش‌تر می‌شود. متغیر پاسخ با افزایش متغیر طبقه تقریباً افزایش می‌یابد و طبقات ۵ تا ۷ بیش‌ترین قیمت را دارا بوده‌اند. قیمت کل واحد مسکونی با افزایش سن بنا کاهش پیدا می‌کند. نمودار آخر نمودار طول و عرض جغرافیایی یعنی وابستگی فضایی را نشان می‌دهد. برای ملاحظه‌ی هرچه بهتر وابستگی فضایی، نقشه‌ی برجسته‌ی اثر فضایی روی متغیر پاسخ با ثابت فرض کردن اثر سایر متغیرها در شکل ۲ رسم شده است.

نواحی پررنگ در این نمودار تراکم کم‌تر داده‌ها را نشان می‌دهد و در نواحی کم‌رنگ‌تر تراکم داده‌ها بیش‌تر است. این نشان می‌دهد که قیمت مسکن بر حسب مناطق مختلف، متفاوت است.

برای این که انعطاف مدل‌های جمعی تعمیم‌یافته را در مقایسه با مدل‌های خطی نشان دهیم، مدل‌های ذکرشده در بخش ۱-۵ را با افزودن یک متغیر تبیینی دیگر به داده‌ها برازش داده و نتایج حاصل را بررسی خواهیم کرد. اکنون متغیر تبیینی «ارزش هر متر مربع زمین مسکونی» را نیز به مدل‌ها اضافه می‌کنیم. مدل‌های بخش قبل به شکل زیر تغییر می‌یابند:



شکل ۱- توابع هموار براوردشده برای مدل با وابستگی فضایی



شکل ۲- نمودار طول و عرض جغرافیایی

$$Z_1 = f_1(Bedrooms) + f_2(Total\ Room) + f_3(Floor) + f_4(Age) + f_5(Value) + f_6(x, y) + \varepsilon$$

$$Z_2 = f_1(Bedrooms) + f_2(Total\ Room) + f_3(Floor) + f_4(Age) + f_5(Value) + +$$

$$L = Bedrooms + Total\ Room + Floor + Age + Value + \varepsilon$$

میزان پوشش بازه‌های اطمینان ۹۵% و معیارهای انتخاب مدل برای مدل‌های مذکور محاسبه شده و در جدول ۲ خلاصه شده است. ملاحظه می‌شود که در صد پوشش بازه‌های اطمینان بعد از افزودن متغیر تبیینی «ارزش یک متر مربع زمین مسکونی» به مدل‌ها، به‌طور چشمگیری افزایش یافته، در حالی که در مورد مدل خطی تغییر چندانی صورت نگرفته است. زیرا یکی از ویژگی‌های مدل‌های جمعی تعمیم‌یافته، انعطاف‌پذیری آن‌ها می‌باشد. این مدل‌ها نسبت به افزایش متغیرها حساسیت زیادی از خود نشان می‌دهند. همچنین افزایش چشمگیر نسبت‌های پوشش برای متغیر پاسخ بعد از اضافه کردن متغیر تبیینی «ارزش یک متر مربع زمین مسکونی» به مدل، نشان می‌دهد که این متغیر تبیینی تأثیر زیادی روی متغیر پاسخ دارد. تمامی معیارهای سنجش مدل نیز بهبود یافته‌اند. برای مثال مقدار معیار انحراف در مدل اول از ۱۸۴/۱۶ در بخش قبل به ۱/۳۶ در این بخش کاهش یافته است که این نشان‌دهنده‌ی این است که در این حالت مدل بسیار خوب عمل می‌کند.

جدول ۲- درصد پوشش متغیر پاسخ با ۵ متغیر تبیینی توسط بازه‌های اطمینان و ملاک‌های انتخاب مدل

مدل	۹۵% پوشش CI	انحراف	AIC
مدل جمعی تعمیم‌یافته با وابستگی فضایی	۷۹٫۲۹	۱٫۳۶	-۱۵۴۴۹٫۲۳
مدل جمعی تعمیم‌یافته بدون وابستگی فضایی	۷۵٫۶۱	۱٫۴۱	-۱۵۳۷۰٫۴۳
مدل خطی تعمیم‌یافته	۶٫۱۴	۵۹٫۶۸	-۳۶۱۸٫۸۲

۶- بحث و نتیجه‌گیری

در این مقاله کاربرد مدل‌های جمعی تعمیم‌یافته در داده‌های فضایی با یک مثال کاربردی در مورد قیمت مسکن بررسی شده است. با بررسی داده‌های قیمت مسکن در مناطق ۲۲گانه‌ی شهر تهران توان مدل‌های جمعی تعمیم‌یافته در مقایسه با مدل‌های خطی

تعمیم‌یافته نشان داده شد. با بررسی مدل جمعی تعمیم‌یافته و نمایش توابع هموار اثر متغیرهای تبیینی بر روی متغیر پاسخ به‌خوبی قابل مشاهده است. همچنین در این مطالعه انعطاف‌پذیری مدل‌های جمعی تعمیم‌یافته با اضافه کردن متغیر تبیینی به مدل نشان داده شد. این نتایج حاکی از این است که مدل‌های جمعی تعمیم‌یافته عملکرد مناسبی در تحلیل داده‌های قیمت مسکن داشته است.

توضیحات

1. Generalized Additive Models
2. Saturated Model
3. Akaike-Information Criterion
4. Thin plate

مرجع‌ها

[1] Abe, M. (1999). A Generalized Additive Model for Discrete-Choice Data. *Journal of Business and Economic Statistics*, **17**, 271-284.

[2] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B.N. Petrov and F. Csake (eds.), Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 267-281.

[3] Cressie, N. (1993). *Statistics for Spatial Data*, Wiley, New York.

[4] Frescino, T.S., Edwards, T.C. and Moisen, G.G. (2001). Modeling Spatially Explicit Forest Structural Attributes Using Generalized Additive Models. *Journal of Vegetation Science*, **12**, 15-26.

[5] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

[6] Guisan, A., Edwards, T.C. and Hastie, T.J. (2002). Generalized Linear and Generalized Additive Models in Studies of Species Distributions: Setting the Scene. *Ecological Modeling*, **157**, 89-100.

[7] Hastie, T., and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1**, 297-318.

[8] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC, New York.

[9] Hastie, T.J. and Tibshirani, R.J. (1995). Generalized Additive Models for Medical Research. *Statistical Methods in Medical Research*, **4**, 187-196.

[10] Lehmann, A. (1998). GIS Modeling of Submerged Macrophyte Distribution Using Generalized Additive Models. *Plant Ecology*, **139**, 113-124.

[11] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, Series A **135**, 370–384.

[12] Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Verlag, New York.

[13] Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.

[14] Kneib, T., Hothorn, T. and Tutz, G. (2009). Variable Selection and Model Choice in Geoadditive Regression Models. *Biometrics*, **65**, 626-634.

[15] Wang, S., Morishima, G., Sharma, R. and Gilbertson, L. (2009). The Use of Generalized Additive Models for Forecasting the Abundance of Queets River Coho Salmon. *North American Journal of Fisheries Management*, **29**, 423-433.

[16] Webster, T., Vieira, V., Weinberg, J. and Aschengrau, A. (2006). Method for Mapping Population-based Case-Control Studies: An Application Using Generalized Additive Models. *International Journal of Health Geographics*, **5**, 26.

[17] Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R. Boca Raton*, Chapman and Hall/CRC, Florida.

علی محمدیان مصمم
دکتری آمار
زنجان، دانشگاه زنجان، گروه آمار.
رایانشانی: a.m.mosammam@znu.ac.ir

ملیحه عباسی
فوق لیسانس آمار
زنجان، دانشگاه زنجان، گروه آمار.
رایانشانی: malihe_abbasi@znu.ac.ir