
بسم الله الرحمن الرحيم

دانشگاه پیام نور
واحد تهران
دانشکده علوم پایه

پایان نامه برای دریافت درجه کارشناسی ارشد

گرایش :

آمار ریاضی

عنوان :

برآورد ناپارامتری استوار رگرسیون فضایی

استاد راهنما:

دکتر مسعود یار محمدی

استاد مشاور :

دکتر پرویز نصیری

پژوهش گر :

صفیه غریب آبادی

زمستان ۹۱

فهرست مطالب

چکیده :

در این پایان نامه، برآورد ناپارامتری استوار برای رگرسیون فضایی مورد بحث و بررسی قرار می گیرد. به طور دقیق تر برای میدان تصادفی مانای اکید داده شدهٔ $Z_i = (X_i, Y_i)_{i \in N^N, N \geq 1}$ ، یک خانواده از برآوردگرهای ناپارامتری استوار برای یک تابع رگرسیونی بر اساس روش هسته ای در نظر گرفته می شود.

تحت برخی فرضیه های مرکب کلی سازگاری تقریبی و نرمال مجانبی برای این برآوردگرها بدست آورده و در ادامه یک روش استوار برای انتخاب پارامتر هموار جهت تطبیق که به داده های فضایی مورد بحث و بررسی قرار می گیرد.

واژگان کلیدی :

توزیع متقارن- همگرایی تقریباً کامل- میدان تصادفی- رگرسیون ناپارامتری- برآورد هسته- برآورد استوار

پیشگفتار:

در بسیاری از مسائل آماری در زمینه های کاربردی نظیر ، زمین شناسی، اقیانوس شناسی ،اقتصاد سنجی، همه گیرشناسی، علوم محیطی، جنگلداری و غیره ، اثریک بردار کمکی از متغیرها روی برخی از متغیرهای پاسخ در وابستگی فضایی مورد مطالعه قرار گرفته است. هدف از این پایان نامه مطالعه برآورد استوار یک تابع رگرسیون به عنوان یک ابزار تحلیلی برای این نوع از داده است.

به طور کلی برای مدل سازی کردن داده ها از روشهای آماری استفاده می شود، مباحث آماری مربوط به *مدل بندی* داده های فضایی از دیر باز مورد توجه آماردانان بوده است. به عنوان مثال می توان گیون[1] (۱۹۹۵) ، انسیلین و فلورکسی[2] (۱۹۹۵) ،کرسی[3] (۱۹۹۱) و ریپلای[4](۱۹۸۱) اشاره کرد، به طوری که اولین مطالعه توسط ترن[5] (۱۹۹۰)بدست آمده است. او نرمال متقارن بودن برآورد چگالی هسته ای راپایه گذاری کرده است. لو و چن[6](۲۰۰۴) برآورد هسته ای از تابع رگرسیون را مورد مطالعه قرار داده و نرخ همگرایی (همگرایی در احتمال) از این برآورد(برآورد هسته ای) را بدست آورده و بیو و کادر[7] (۲۰۰۴) سازگاری یکنواخت و نرمال متقارن ، برآورد هسته از تابع رگرسیون را ثابت نمود.کاربن [8](۲۰۰۷) مدل اتو رگرسیون ناپارامتری برای پیشگویی میدان های تصادفی را مورد مطالعه قرار دادند. اخیراً لی و ترن (۲۰۰۹)[9] یک روش متقابل ازیک برآورد ناپارامتری از رگرسیون فضایی را بدست آورده اند. این روش بر اساس وزن گذاری روی نزدیک ترین همسایگی پایه گذاری شده است، نشان داده اند که برآوردشان ، مخالف نرمال تقریبی است.

بحث ابتدایی رگرسیون استوار را هوبر[10] در سال (۱۹۶۴) شروع کرده و توسط لی لاینز تحقیقی در مورد رگرسیون استوار بدست آمده است. برای این که تعمیم پیوستگی رگرسیون استوار را مشاهده نمود، می توان به روبینسون[11] (۱۹۸۴) و کلوب و هاردل[12] (۲۰۰۰) ، و همچنین برای

1-Guyon
2-Anselin and Florax
3-Cressi
4-Riply
5-Tran
6-Lu and C
7-Biau and cadre
8-Carbon
9-Li and Tran
10-Huber
11-Robinson
12- Collomb and Hardle

نتایج قبلی ، بونت و فریمن[1] (۲۰۰۹) برای پیشرفت های اخیر و به عنوان مرجع مراجعه نمود. در اغلب این تحقیقات فرض شده است که مشاهدات وابسته از داده های سری زمانی آمده است.

در حالت فضائی، زو و وانگ [2](۲۰۰۸) برآورد خطی مکانی از مدل رگرسیون را مورد مطالعه قرار داده اند. این برآورد شبه پارامتری به وسیلهٔ مینیمم کردن قدر مطلق انحراف آنها بدست می آید. با بررسی های به عمل آمده مسئله برآورد ناپارامتری رگرسیون فضایی بوسیله یک روش استوار موضوع جدیدی بوده و توسط ابدلادرگریباله ، الی لاکساکی، رچیداروان[3] (۲۰۱۰) معرفی شده است .در فصل اول این پایان نامه تعاریف پایه ای ازآمار فضایی و مروری بر رگرسیون کلاسیک و رگرسیون فضایی ارائه می شود. در فصل دوم روشهای استوار در تحلیل داده های فضایی ارائه می شود. در فصل سوم به تجزیه و تحلیل قضایا و لم هایی که مربوط به برآورد ناپارامتری رگرسیون فضایی است پرداخته و در فصل چهارم روشهای شبیه سازی به تفصیل مورد مطالعه قرار می گیرد.

1- Boente and fraiman
2- Xu and Wang
3- Abdelkader Gheriballah, Ali Laksaci, Rachida Rouane´

فصل اول

آمار فضایی

۱ـ ۱: مقدمه

در آمار، وقتی جامعه آماری مورد نمونه برداری قرارمی گیرند، اغلب فرض می شود که مشاهدات حاصل از نمونه از یکدیگر مستقل هستند. با فرض استقلال مشاهدات، استنباط آماری آنها بر پایه قضایای ریاضی و آمار ساده می گردد در عمل با موارد زیادی مواجه می شویم که مشاهدات به نوعی به یکدیگر وابسته هستند به عنوان مثال داده ها در سری های زمانی مثال از داده هایی هستند که در طول زمان همبسته می باشد. داده هایی که در فضای مورد مطالعه وابسته به یکدیگر بوده و این وابستگی نوعاً به محل قرار گرفتن آنها اغلب همبستگی روی زمان را به نمایش می گذارند. به دلیل همبستگی فضایی، روش های کلاسیک آماری برای تجزیه و تحلیل چنین داده هایی قابل استفاده نمی باشند. به همین دلیل شاخه آمارفضایی ایجاد گردیده است. نخستین بار به دنبال روند تکاملی ذخایر معدنی که قبل از سال (۱۹۶۰)آغاز شده بود و به ویژه براساس پژوهشهای افرادی مانند سیشل وکرنگ ، ماترون پژوهشگر فرانسوی با انتشار مقاله ای درسال (۱۹۶۲) پایه های زمین آمار را بنا نهاد و شاخه ای جدید علم آمار را تحت همین عنوان بوجود آورد. زمین آمار که توسط ماترون (۱۹۶۲) پایه گذاری شده است، به شاخه ای از علم آمار گفته می شود به تجزیه و تحلیل داده هایی همبسته که همبستگی شان تابعی از موقعیت قرارگرفتن جهت داده ها می باشد، می پردازد. در واقع بین مقادیر مختلف یک متغیر و یک فاصله و جهت قرارگیری آنها ارتباطی را برقرارمی کند، که این ارتباط فضایی، ساختارفضایی نامیده می شود.

بررسی و مطالعه این ساختار یکی از مسائل مهم آمارفضایی است. در شکل گیری آمار فضایی افراد مختلفی سهیم بوده و به طور موازی با سیشل، کریگ و ماترون فعالیت داشتند. از جمله گایدن (۱۹۶۲) از کشور روسیه که مبحثی به نام تحلیل عینی را با مفاهیم بسیار نزدیک به تعاریف ماترون و دیگران در زمینه هواشناسی پیشنهاد کرده است و یا ماترون سوئدی (۱۹۶۲) روشهایی به نام تجزیه و تحلیل داده های فضایی را ابداع نمود. در این راستا دیوبد (۱۹۷۷) به طور مفصل روش های زمین آماری را با اصطلاحات خاص شرح داده است.

ریپلی (۱۹۸۱) در کتاب آمار فضایی دریچه دیگری گشود و با زبان آماری سخن گفت وی نظرات مربوط به پیشگویی فرایندهای تصادفی که ارتباط نزدیکی با سری های زمانی دارند را به کارگرفت .کرسی (۱۹۹۳) نیز با انتشار کتاب آمار برای داده فضایی موجب پیشرفت گسترده این شاخه از علوم گردید. در سالهای اخیر به مدد تلاش بسیاری از پژوهشگران، آمار فضایی به عنوان وسیله ای کارآمد برای مطالعه داده های فضایی مطرح گردیده است.

اگرچه عمده زمینه های رشد و توسعه آمار فضایی در مسائل مربوط به ذخائر معدن بوده است ولی در زمینه های مختلف از علوم مانند جغرافیا، همه گیرشناسی، اکتشاف نفت، گاز وکنترل کیفیت هوا تحقیقات کاربردی فراوان انجام گرفته است (خالدی۱۳۷۸).

۱-۲ ضرورت استفاده ازآمار فضایی

در آمار کلاسیک مشاهدات مستقل از هم و فارغ از موقعیت مکانی فرض می شوند، در نتیجه نظریهٔ استنباط آماری آنها ساده است. اما در نظر گرفتن موقعیت فضایی و هم چنین وابستگی فضایی بین مشاهدات نمونه، به شناخت بیش تر جامعه مورد بررسی کمک می کند. کرسی (۱۹۹۳) بیان ضرورت در نظر گرفتن همبستگی فضایی در تحلیل داده ها به ذکر مسأله براورد میانگین جامعه ای با واریانس معلوم σ^2 پرداخت. در این مسأله با فرض استقلال مشاهدات و به کمک نمونه تصادفی X_1،...، X_n، از برآوردگر ناریب با کم ترین واریانس $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ با واریانس $\frac{\sigma^2}{n}$استفاده می شود.

اگر توزیع جامعه ایکه از آن نمونه استخراج شده نرمال باشد یا n به اندازه کافی بزرگ باشد، بنابر قضیهٔ حد مرکزی، $\overline{X}$ دارای توزیع نرمال با میانگین μو واریانس $\frac{\sigma^2}{n}$خواهدبود. بنابراین بازه اطمینان ۹۵٪ برای μ به صورت

$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) \qquad (1-1)$$

اما اگر داده ها به طور مستقل نبوده و همبستگی آنها از طریق تابع کواریانس

$$Cov(X_i, X_j) = \sigma^2\rho^{|i-j|}; i, j = 1,...,n \qquad (2-1)$$

قابل بیان باشد،واریانس برآوردگر $\overline{X}$ برابر

$$Var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)$$

$$= \frac{\sigma^2}{n}\left[1 + 2(\frac{\rho}{1-\rho})(1-\frac{1}{n}) - 2(\frac{\rho}{1-\rho})^2\frac{1-\rho^{n-1}}{n}\right] \qquad (3-1)$$

خواهد بود. این مقدار با واریانس برآوردگر در حالت استقلال مشاهدات، یعنی $\frac{\sigma^2}{n}$ تفاوت قابل ملاحظه ای دارد. علاوه بر این، به ازای $n = 10$و $\rho = .26$، $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}[1.608]$و بازهٔ اطمینان ۹۵٪ برای μ به صورت

$$(\overline{X} - 2.485\frac{\sigma}{\sqrt{n}}, \overline{X} + 2.485\frac{\sigma}{\sqrt{n}})$$

خواهد شد. همان طور که ملاحظه می شود، وجود همبستگی داده ها از نوع (۱-۲) موجب افزایش واریانس $\overline{X}$ و نیز افزایش طول برآورد بازه ای ۹۵٪ باشد، بلکه احتمال پوشش آن ۰.۸۷ ٪ است.

فرض کنید X_1،...، X_n و Y_1،...، Y_n دو نمونه مستقل از دو جامعه به ترتیب با توزیع های $N(\mu_x,\sigma^2)$ و $N(\mu_y,\sigma^2)$ باشند در این صورت $\overline{X}$ و $\overline{Y}$ براوردگرهای μ_x و μ_y نیز مستقل و به ترتیب دارای توزیع های $N(\mu_x,\frac{\sigma^2}{n})$ و $N(\mu_y,\frac{\sigma^2}{n})$ خواهند بود. آمارهٔ آزمون برای فرضیه $H_0 : \mu_x = \mu_y$ در مقابل $H_1 : \mu_x \neq \mu_y$ به صورت زیر است

$$Z = \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{2}{n}}} \sim N(0,1)$$

اما اگر مشاهدات هر یک از دو نمونه وابسته بوده دارای کواریانس

$$Cov(X_i X_j) = \sigma^2\rho;\ i \neq j = 1,...,n \qquad (4-1)$$

باشند، آن گاه

$$Var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)$$

$$= \frac{1}{n^2}\left[n\sigma^2 + n(n-1)\sigma^2\rho\right]$$

$$= \frac{\sigma^2}{n}\left[1 + (n-1)\rho\right]$$

و به طور مشابه $Var(\overline{Y}) = \frac{\sigma^2}{n}\left[1 + (n-1)\rho\right]$. در این صورت آمارهٔ آزمون برای فرضیه $H_0 : \mu_x = \mu_y$ به صورت زیر می شود.

$$Z^* = \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{2(1 + \frac{(n-1)\rho}{n}}} \sim N(0,1)$$

بدیهی است مقدار مشاهده شدهٔ Z در نمونه ها همواره بزرگ تر از مقدار Z^* است، یعنی اگر همبستگی داده ها در نظر گرفته نشود، مقدار مشاهده شدهٔ Z خیلی بزرگ تر و به تبع آن p - مقدار کوچک تر و در نتیجه شواهد رد H_0 در داده ها بیش تر از حالتی خواهد بود که همبستگی داده ها منظور می شوند، یعنی آزمون بیش از مواردی که لازم است رد می شود.

نکته قابل توجه دیگر آن که $E(\overline{X}) = \mu_x$ و $\lim_{n\to\infty} Var(\overline{X}) = \sigma^2\rho$،یعنی بدون توجه به همبستگی داده ها، همواره $\overline{X}$ برآوردگری نااریب برای میانگین جامعه است، اما در صورت وابستگی داده ها، $\overline{X}$ برآوردگری ناسازگار است.

اگر رابطه(۱-۳) به صورت $Var(\bar{x}) = \frac{\sigma^2}{n}$ نوشته شود، آن گاه می توان

$$n' = \frac{n}{1 + 2(\frac{\rho}{1-\rho})(1 - \frac{1}{n}) - 2(\frac{\rho}{1-\rho})^2(\frac{1-\rho^{n-1}}{n})} \qquad (5-1)$$

را به عنوان تعداد مشاهدات مستقل معادل n مشاهده همبسته تعبیر کرد. به عنوان مثال اگر $n = 10$ و $\rho = .26$ آن گاه $n' = 6.2$.یعنی ۶ مشاهده مستقل به طور تقریبی نتایجی با همان دقت $n = 10$ مشاهده همبسته ایجاد می کنند. برای مقادیر بزرگ n نیز $n' \approx \frac{n}{\left[\frac{(1+\rho)}{(1-\rho)}\right]}$ بیانگر آن است که در نمونه های بزرگ نیز همبستگی داده ها در دقت استنباط آماری تأثیرگذار است. در نتیجه توجه نکردن به همبستگی فضایی داده ها، دقت نتایج تحلیل آماری آن ها را بسیار تحت تأثیر قرار می دهد و لازم است مراحل مختلف تحلیل آماری داده ها با توجه به همبستگی داده ها انجام شود.

۱-۳: انواع داده های فضایی

در آمار فضایی متغیر مورد اندازه گیری ممکن است گسسته یا پیوسته باشد. فضای موقعیت یا مکان مشاهدات نیز ممکن است پیوسته یا گسسته باشد. فضای موقعیت یا مکان مشاهدات نیز ممکن است پیوسته یا گسسته، نقطه ای یا ناحیه ای، منظم یا نامنظم باشد. وقتی مقدار متغیر در یک ناحیه ثبت می شود، موقعیت ناحیه ای و اگر مقدار متغیر در یک نقطه ثابت یا طول و عرض جغرافیایی معین اندازه گیری شود، موقعیت نقطه ای است. در موقعیت نقطه ای اگر نقاط در فواصل مساوی از هم قرار داشته باشند، موقعیت های منظم و در غیر اینصورت نامنظم اند. در مورد مکان های ناحیه ای اگر ناحیه ها هم شکل و هم اندازه باشند، ناحیه ها منظم و در غیر این صورت نامنظم اند. با توجه به انواع موقعیت ها مشاهدات فضایی به سه گروه عمده، داده های زمین آماری[1]،داده های مشبکه ای[2] و الگوهای نقطه ای تقسیم می شوند.

1-Geostatisticasl data
2-Lattice data

داده های زمین آماری: این نوع داده ها در موقعیت های ثابت و مشخص در ناحیه ای پیوسته مشاهده می شوند. متغیر مورد بررسی ممکن است گسسته یا پیوسته باشد. به عنوان مثال برای حالت پیوسته می توان غلظت مواد معدنی در داخل یک معدن، مقدار ریزش باران در ایستگاه هواشناسی و برای حالت گسسته می توان تعداد نوعی جانور دریایی در یک سری از مکان های نمونه گیری شده در طول یک ساحل را نام برد.

داده های مشبکه ای: این نوع داده ها مربوط به مکان های ناحیه ای هستند، که این مکان ها ممکن است منظم یا نا منظم باشند. تصاویر ماهواره ای از سطح زمین که در آن ها سطح زمین به تعدادی عنصر تصویر[1] کوچک به صورت مشبکهٔ منظم در R^2 تقسیم شده ،مثالی برای داده های مشبکه ای منظم است. تعداد افراد مبتلا به سرطان در تمام مناطق خدمات درمانی کشور مثالی برای داده های مشبکه ای نامنظم است. در این حالت کل افراد مبتلا به سرطان در هر شهرستان در مرکز خدمات درمانی آن شهرستان در نظر گرفته می شوند. تعداد حوادث رانندگی، تعداد اتباع خارجی غیر مجاز ساکن یا نرخ رشد اقتصادی هر استان مثال های دیگری از داده های مشبکه ای هستند. به طور معمول هدف از تحلیل داده های مشبکه ای مدل بندی احتمالاتی مشاهدات است، در صورتی که در زمین آمار پیش گویی مقدار متغیر در یک موقعیت جدید مدنظر است.

الگوی نقطه ای: در این حالت مکان یا موقعیت مشاهده شده خود متغیری تصادفی است. الگوهای نقطه ای شامل تعدادی متناهی از مکان ها در یک ناحیه اند که در آن ها یک صفت خاص اندازه گیری می شود. به طور معمول الگوهای نقطه ای به سه دستهٔ به طور کامل تصادفی فضایی، منظم و خوشه ای تقسیم و به مدل بندی آن ها اقدام می شود. یک مثال برای این نوع مشاهدات موقعیت گونه ای از درختان در یک ناحیه ای جنگلی یا موقعیت مراکز زلزله است (پور طاهری و همکاران ۲۰۰۶)

موقعیت مکانی هر نقطه روی کره زمین به صورت مختصات جغرافیایی بر حسب درجه و دقیقه بیان می شود که بر مبنای هندسه کروی استوار است. در این سیستم کره زمین به ۶۰ نوار عمودی یا منطقه هر یک به طول ۶ درجه تقسیم می شود. محدودهٔ شمالی این سیستم عرض جغرافیایی ۸۴ درجه و محدودهٔ جنوبی آن تا درجه ۸۰ می باشد. برای مناطق بالاتر و پایین تر این عرض ها، از سیتم تصویر صفحه ای تحت عنوان سیستم استرایوگرافیک قطبی جهانی[2] (UPS) استفاده می شود. هر کدام از

3-Pixel
1- Universal polar stereograph

نوارهای ۶۰ گانه با یک عدد مشخص می شود. به نوار بین ۱۷۴ تا ۱۸۰ درجه عدد یک را اختصاص می دهند و هر چه به سمت شرق جلو برویم اعداد افزایش پیدا می کند. بنابراین منطقهٔ ۳۱ دارای طول صفر تا ۶ درجه شرقی است. ایران نیز در منطقه های ۳۸، ۳۹، ۴۰، ۴۱ قرار گرفته است.

نحوه دیگر ارائه موقعیت مکانی هر نقطه روی سطح دو بعدی ، سیستم مختصات دکارتی است. برای تصویر سطح کره زمین بر روی سطح کاغذ از سیستم تصویر مرکاتور جانبی جهانی[1] (UTM) استفاده می شود که متریک است و طول و عرض هر نقطه جغرافیایی را در مختصات دکارتی به صورت xوy نمایش می دهد.

با توجه به این که در آمار فضایی موقعیت مکانی هر نقطه به صورت مختصات دکارتی در تظر گرفته می شود، لازم است تا موقعیت جغرافیایی مشاهدات با استفاده از سیستم UTM به مختصات دکارتی تبدیل شود. لازم به ذکر است که مختصات جدید موقعیت داده های واقع در یک ناحیه در سیتم تبدیل UTM، با تعیین فاصله هر موقعیت از وسط آن ناحیه (محل تقاطع خط استوا با نصف النهار میانی ناحیه) محاسبه شود. بنابراین با استفاده از این سیتم برای داده هایی که در نواحی مختلف قرار دارند، صحیح نیست. در این گونه موارد از سیستم لامبرت[2] استفاده می شود که مبدأ سنجش فاصله موقعیت ها را یک نقطه دلخواه (به طور مثال تقاطع خط استوا با نصف النهار گرینوچ) فرض می کند و فاصله هر موقعیت از آن نقطه را ارائه می دهد.

۱-۴: موقعیتهای فضایی

منظور از موقعیت های فضایی[3] دنباله $\{s_1,...,s_n\}$است، که این نقاط به طور منظم یا بی قاعده در R^d قرار گرفته و داده های $\{Z(s_1),...,Z(s_n)\}$در این موقعیت ها مشاهده می شوند. معمولاً مولفه های $s \in R^d$ به صورت $s = (s[1],...,s[d])$نشان داده می شود. در آمارفضایی همبستگی قوی بین داده ها در موقعیت های نزدیک برقرار میگردد. بسیج (۱۹۷۴) چندین مثال درارتباط با طبقه بندی های مختلف موقعیت ها و متغیرهای مرتبط ارائه می کند. این طبقه بندی بر اساس سه مورد زیرمی باشد:

(۱) سیستم و موقعیت ها (منظم یا نامنظم)

(۲) موقعیت های منفرد (نقطه ها یا ناحیه ها)

2-Universal polar stereographic
3-Lambert
4- Spatial Location or sits

(۳) متغیرهای تصادفی (مرتبط پیوسته یا گسسته)

بر اساس تقسیم بندی فوق داده های فضایی بـه سـه نـوع تقسـیم مـی شـوند. اگـر داده هـادر موقعیت های ثابت مشاهده شوند و متغیر مورد مطالعه پیوسته یا گسسته باشد، داده های زمین آمـاری نامیده می شوند. در صورتی که داده ها مربوط به مکان های ناحیه ای باشند، داده های شبکه ای و در صورتی که سایت یا موقعیت خود یک متغیر تصادفی باشد، الگوهای نقطه ای نامیده می شوند.

۱–۵: تفاوت رگرسیون کلاسیک وفضایی

در رگرسیون کلاسیک مدل آماری $y_i = g(x_i) + \varepsilon_i \quad ;i = ۱,...,n$ با این فرض که خطاها بـه طورمسـتقل دارای توزیـع نرمـال بـا میـانگین صـفر و واریـانس $\sigma^۲$ هسـتند، تجزیـه و تحلیـل می گردد. در عمل پس از انتخاب مدل رگرسیونی بررسی درستی سه فرضیهٔ نرمـال بـودن خطاهـا در اولویت قرارمی گیرد.

درستی فرض های فوق به اعتبار مدل رگرسیونی منجر می گـردد. در رگرسـیون فضـایی بـین متغیرها و درنتیجه خطاها، نوعی همبستگی فضایی وجود دارد. از اینرو در رگرسیون فضـایی بررسـی فـرض خـود ناهمبسـته نبـودن خطاهـا غیـر ضـروری اسـت. همچنـین در مـدل رگرسـیون کلاسیک $\{x_۱,...,x_n\}$، بردار $Z = (Z(s_۱),...,Z(s_n))$ به عنوان نمونـه در ایـن موقعیـت هـا مشـاهده می گردد.

۱–۶:مدل آماری

برای تحلیل داده های فضایی لازم است یک مدل آماری در نظر گرفته شود.در آمار فضایی بـه طور معمول یک میدان تصادفی به عنوان مدل های آماری در نظر گرفتـه مـی شـود. میـدان تصـادفی مجموعه ای از متغیرهای تصادفی مانند $\{Z(s); s \in D\}$است، که در آن مجموعهٔ اندیس گذارِ یک زیـر مجموعه ای از فضای اقلیدیسی ی بعدی، $d \geq ۱$، از R^dاست. در مورد میدان تصـادفی $Z(.)$میـانگین در موقعیت sو کواریانس در موقعیت های $s_۱$و $s_۲$به ترتیب به صورت

$$E(Z(s)) = \mu(s) \;;\; s \in D$$

$$C(s_۱, s_۲) = Cov[Z(s_۱), Z(s_۲)]$$

$$= E[(Z(s_۱) - \mu(s_۱))(Z(s_۲) - \mu(s_۲)), \qquad s_۱, s_۲ \in D]$$

تعریف می شوند. برای $s = s_۱ = s_۲$، واریانس میدان تصادفی $Z(.)$ در مکان sبصورت :

$$Var(Z(s)) = E\left[z(s) - \mu(s)^2\right]$$

حاصل می شود. هر میدان تصادفی $Z(.)$را می توان به صورت

$$Z(s) = \mu(s) + \delta(s),\ s \in D$$

زیر تجزیه کرد، که در آن $\mu(.)$ تغییرات بزرگ مقیاس [1]یا روند[2]و $\delta(.)$، فرایندخطایاتغییرات کوچک مقیاس[3] میدان تصادفی نامیده می شوند. تغییرات کوچک مقیاس ممکن است ناشی از خطای اندازه گیری یا تغییرپذیری در درون موقعیت مشاهده شده باشد وتغییرات بزرگ مقیاس ممکن است ناشی از تغییرات بین موقعیت های مشاهده شده باشند.

به طور معمول تحلیل داده های فضایی براساس مشاهدات نمونه دشوار است، امابرخی ویژگی های میدان تصادفی موجب ساده سازی مسأله خواهند شد، که درادامه معرفی می شوند.

میدان تصادفی

میدان تصادفی[4] برای مدل سازی داده های فضایی به منظور تجزیه و تحلیل آنها به کار می رود. یک میدان تصادفی مجموعه ای از متغیرهای تصادفی مانند $\{Z(s); s \in D\}$است، که در آن D زیر مجموعه ای از فضای اقلیدیسی $d \geq 1$ بعدی R^dاست. برای حالت خاص $d = 1$، $\{Z(s); s \in D\}$ فرایند تصادفی نامیده می شود. مجموعه اندیس گذار D می تواند پیوسته، مانند R^d یا زیر مجموعه ای از R^d، یا گسسته به صورت شبکه ای[5] از موقعیت ها، مانند Z^dیا زیر مجموعه ای از Z^dباشد. میدان تصادفی پیوسته یا شبکه نامیده می شود هر گاه Dبه طور متناظر پیوسته یا گسسته باشد.

شبکه می تواند بصورت اشکال هندسی منظم مانند یا نامنظم باشد. برای داده های زمین آماری Dزیر مجموعه ای از مستطیل های dبعدی با حجم ثابت و پیوسته در R^dاست. برای الگوهای شبکه ای D زیر مجموعه ای ثابت و شمارشی از R^dمی باشد.

تعریف ۱ـ۱

میدان تصادفی $Z(.)$مانای ذاتی[6] نامیده می شود، هرگاه

۱-میانگین میدان تصادفی ثابت یا مستقل از sباشد،یعنی

1-Large scale variation.
2-Trend،
3-Small scale variation،
4- Random Field
5- Lattice

1-Intrinsic stationary

$E(Z(s) = \mu$

۲-واریانس عبارت $(Z(s_۱) - Z(s_۲))$فقط تابع یاز فاصلهٔ موقعیت های $s_۱$و $s_۲$باشد، یعنی

$Var(Z(s_۱) - Z(s_۲)) = ۲\gamma(s_۱ - s_۲),\ s_۱, s_۲ \in R^d$

تعریف ۱ـ۲

میدان تصادفی $Z(.)$، مانای مرتبه دوم۱ (یا مانای ضعیف) نامیده می شود، هرگاه:

۱- میانگین میدان تصادفی $Z(.)$مستقل از sو ثابت باشد. یعنی

$E(Z(s) = \mu,\ s \in D \subset R^d$

۲-کواریانس $Z(s_۱)$و $Z(s_۲)$، فقط تابعی از فاصله موقعیت های $s_۱$و $s_۲$ باشد، یعنی

$Cov(Z(s_۱), Z(s_۲)) = C(s_۱ - s_۲),\ s_۱, s_۲ \in D \subset R^d$

واضح است که تحت مانایی مرتبه دوم

$Var(Z(s)) = Cov(Z(s), Z(s)) = C(۰) = \sigma^۲$

یعنی واریانس میدان تصادفی به موقعیت فضایی بستگی ندارد و تغییرپذیری میدان در همه جا یکسان است.

تعریف۳ـ۱

میدان تصادفی $Z(.)$مانای قوی[۲] نامیده می شود، هر گاه برای همه موقعیت هـای $s_۱$،...، s_nو تـأخیر h، توزیع توأم $Z(s_۱)$،...، $Z(s_n)$باشد. یعنی

$(Z(s_۱),..., Z(s_n)) \underline{D} (Z(s_۱ + h),..., Z(s_n + h))$

بـه بیـان دیگـر، در صـورت انتقـال موقعیـت هـای $s_۱$،...، s_n در راسـتای $h \in R^d$، توزیـع تـوأم $Z(s_۱)$،...، $Z(s_n)$تغییر نمی کند.

تعریف ۱ـ۴

میدان تصادفی $Z(.)$، گاوسی۳ نامیده می شود، هرگاه برای هر $m \geq ۱$ توزیع تـوأم $(Z(s_۱),..., Z(s_m)$، نرمال چند متغیره باشد.

2-Second order stationary
3-Strong stationary
1-Guassian

فصل دوم

روشهای استواردر تحلیل

داده های فضایی

۲-۱: تعریف داده فضایی

داده های فضایی، مشاهداتی هستند که برحسب موقعیت قرار گرفتنشان در فضای مورد مطالعه به یکدیگر وابسته اند. آمار فضایی، شاخه ای از علم آمار است که به تحلیل چنین داده هایی می پردازد. برآورد تغییرنگار، برآورد پارامترهای روند و پیشگویی فضایی از مباحث مهم در آمار فضایی هستند، که معمولاً با پذیرش مفروضاتی صورت می پذیرند. انحراف از این پیش فرضها، ممکن است باعث بروز خطا و کاهش دقت در هر کدام از مباحث مورد اشاره شود. حضور داده های دورافتاده نیز یکی از مهمترین عوامل بروز این انحرافات است. روش های استوار فضایی قادرند تا حد زیادی اثرات، انحراف از پیش فرضها را در تحلیل داده های فضایی کنترل کنند. این فصل از سه بخش تشکیل شده است، در بخش اول تعاریف بررسی شده و در بخش دوم به علت اهمیت داده های دورافتاده شناسایی داده های دورافتاده مورد بررسی قرار گرفته و روش های استوار در حضور داده های دور افتاده در بخش سوم مورد توجه قرار گرفته است.

مشاهدات مبنای استنباط آماری هستند. اما علاوه بر مشاهدات، فرض هایی که درباره موقعیت تحت بررسی صورت می گیرد نقش پایه ای در استنباط آماری دارند. این پیش فرض ها در مورد توزیع، تصادفی بودن، وابستگی پارامترهای نامعلوم و...هستند. انتخاب پیش فرض ها به عنوان مکمل های یک مدل آماری نقش حساس و تعیین کننده ای در استنباط آماری دارند. انتخاب پیش فرض های مناسب، مدل های آماری را آسان وگاهی ممکن می سازند، نیازمند دقت زیاد هستند، ممکن است نتایج نادرستی حاصل شود. به این ترتیب در انتخاب با پیش فرض ها همان قدر که مفید بودن آنها و نقشی که در هموارسازی مسیر تحلیل دارند مورد توجه است باید همان میزان اعتبار آنها به عنوان پایه های استنباط آماری در نظر گرفته شود. هر چقدر دقت در انتخاب پیش فرض ها بیشترباشد، اطمینان از معتبر بودن نتایج بالاتر خواهد بود. اما نباید فراموش کرد که پیش فرض های در نظر گرفته شده هیچ وقت قادر نخواهند بود طبیعت پدیده مورد بررسی را کامل کنند. اختلاف جزئی بین آنچه پیش فرض‌ها می‌گویند وآنچه واقعیت دارد، چیزی طبیعی و قابل اغماض است. ولی اگر تفاوت ها به اندازه ای باشد که صحت پیش فرض ها مورد تردید قرار گیرد نباید از آنها چشم

پوشی کرده در عمل می توان انحرافات زیادی از پیش فرض ها را متصور شد و روش‌هایی را فـراهم آورد که حساسیت کمتری به این انحرافات داشته باشد.

با پیشرفت آمار و توجه به واقعیت‌های فوق تلاش‌هـا بـرای یـافتن روش هـایی کـه بتواننـد حساسیت به تخطی از پیش فرض ها را کنترل کنند، بیشترشد. روش های ناپـارامتری یکـی از گزینـه هایی بود که برای حل این مشکل مورد توجه قرار گرفت. در این روش ها هیچ پیش فرضـی دربـاره توزیع جامعه ای که داده ها در آن آمده اند، در نظر گرفته نمی شود و نتایج آن هر چند با دقت کمتـر اما برای تمامی توزیع جامعه قابل استفاده است. به علت این قابلیت بعضـی از ایـن روش هـا کـاربرد وسیعی در آمار یافتند با این وجود، در مواردی که عملاً استخراج اطلاعات در مورد ساختار جامعه از داده ها مقدور نمی باشد، روش های ناپارامتری ذهن برتری طلـب یـک آمـاردان را اغنـاء نمـی کنـد وهمواره با این پرسش مواجه خواهیم بود که آیا واقعاً نمی توان مدل هایی برای جامعه تحت بررسـی در نظر گرفت؟ آیا اطلاعات زیادی که در مورد ساختار جامعه در دسترس ماست باید فقط بـه خـاطر ترس از ناکامل بودن مدل، کنار گذاشته شود؟ آیا این یک نوع محافظه کاری افراطـی نیسـت؟ آیـا راه هایی برای استفاده از این اطلاعات وجود ندارد؟ این چنین است که نمی توان به سـادگی مـدل هـای پارامتری را با توجه ظرفیتشان در میان اطلاعات موجـود در داده هـا نادیـده گرفـت. بلکـه بـه دنبـال راهکارهایی برای کاهش حساسیت به فاصله پیش فرض ها از واقعیات جامعه گشت. به این صـورت که مدل های پارامتری همچنان محور باشند و روش هایی مورد توجه قرار گیرند که وابسـتگی آنهابـه پیش فرض های مدل کمتر بحران ساز شود به این نیاز به روش های استوار[1] بوجود آمد

۲-۱-۱: روش های آماری استوار

اصلی ترین تعریف روش های آماری استوارتوسط هوبر(۱۹۸۱) ارائه شده است کـه کیفیـت های مطلوب زیر را درنظرگرفت.

۱-باید تحت مدل فرض شده دارای کارایی بهینه یا نزدیک بهینه باشد

۲-عملکرد آنها تحت انحرافات کوچک از فرض های مدل به مقدار کمی دچار آسیب شود.

۳-بعضی از انحرافات بزرگ از مدل سبب فاجعه نشود.

---

1 -roubust