
بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

رشته آمار، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

مدل‌بندی پاسخ‌های وابسته با استفاده از مفصل گاوسی

نگارنده: خدیجه سلطانی

استادان راهنما

دکتر حسین باغیشنی
دکتر محمد آرشی

شهریور ۱۳۹۶

شکر و سپاس خدا را که بزرگترین امید و یاور در لحظه لحظه زندگیست. ماحصل آموخته هایم را تقدیم می کنم به آنان که مهر آسمانی شان آرام بخش آلام زمینی ام است...

به پدرم، او که صبوری اش، دل ناامیدم را سپیده دم امیدواری است، همین که هست دلگرمی ام، کافیست.

به مادرم، آرامش جاودانه ام که مهربانی اش بی انتهاست، مادرم، تو باشی و خدا، دنیا برایم بس است

بوسه بر دستان پر مهرتان؛ که با هم آغاز کردیم و در کنار هم آموختیم و به امید هم به آینده چشم می دوزیم. خوشبختی تان منتهای آرزوی من است.

و تقدیم به استادان بزرگوارم، جناب آقای دکتر حسین باغیشنی و محمد آرشی، که این مسیر را برایم هموار و کاستی هایم را نادیده گرفتند...

سپاس‌گزاری...

نهال را باران باید و آفتاب، تا سیراب شود از آب حیات و استوار در مسیر کمال...

بسی شایسته است از زحمات بی‌دریغ استادان بزرگوار و فرهیخته‌ام جناب آقای دکتر باغیشنی و جناب آقای دکتر آرشی، آموزگارانی که برایم توزیع‌های نرمال مسیر زندگی، نمونه گیری خوب از جامعه و رسیدن به تعالی با کمترین واریانس را معنا کردند تقدیر و تشکر نمایم. همچنین از استادان گرانقدرم آقای دکتر ربیعی و خانم دکتر اقبال که با پذیرش داوری این رساله صمیمانه مرا مورد عنایت خویش قرار دادند سپاسگزارم. در پایان از زحمات و الطاف بی‌دریغ همه اساتید دلسوز و خستگی ناپذیرم که در راه علم و آموختن همواره مهربانانه دست‌هایم را رها نکردند کمال تشکر را دارم.

خدیجه سلطانی
شهریور ۱۳۹۶

تعهد نامه

اینجانب **خدیجه سلطانی** دانشجوی کارشناسی ارشد رشته **آمار علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **مدل‌بندی پاسخ‌های وابسته با استفاده از مفصل گاوسی** ، تحت راهنمایی **حسین باغیشنی** و **محمد آرشی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام‘‘ دانشگاه صنعتی شاهرود ’’ یا ‘‘ Shahrood University of Technology ‘‘ به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

خدیجه سلطانی
شهریور ۱۳۹۶

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

مدل‌های رگرسیونی از جمله ابزار پرکاربرد در علوم مختلف از جمله پزشکی، علوم طبیعی، علوم اجتماعی، اقتصاد، و محیط‌شناسی هستند. در مواردی که متغیر پاسخ غیرنرمال باشد، معمولا به‌جای مدل‌های خطی از مدل‌های خطی تعمیم‌یافته استفاده می‌شود. اما پذیره اساسی در این مدل‌ها استقلال بین مشاهدات است. در موارد متعددی با پاسخ‌هایی مواجه می‌شویم که نوعی ساختار وابستگی بر آن‌ها حاکم است. به‌عنوان چند نمونه می‌توان به داده‌های طولی، سری زمانی و داده‌های فضایی اشاره کرد.

در این موارد باید وابستگی پاسخ‌ها را به‌شکلی به مدل وارد کرد. برای این کار رهیافت‌های مختلفی معرفی شده‌اند. رهیافت معمول استفاده از مدل‌های آمیخته خطی تعمیم‌یافته است. استنباط در این رده از مدل‌ها (در هر دو دیدگاه کلاسیک و بیزی) با مشکلات محاسباتی جدی روبرو است. یک راهکار جانشین جدید برای لحاظ کردن وابستگی پاسخ‌ها، استفاده از مدل‌های رگرسیون حاشیه‌ای مبتنی بر تابع مفصل است.

در این پایان‌نامه، ابتدا مدل رگرسیون حاشیه‌ای مفصل گاوسی را معرفی می‌کنیم. سپس استنباط آماری مدل و معیارهای برازش را معرفی و با استفاده از آزمون مشخص‌سازی هاسمن درستی انتخاب مدل را بررسی می‌کنیم. رهیافت استنباط مورد نظر ما مبتنی بر تابع درستنمایی است. در ادامه به ارزیابی مدل با مثال‌های شبیه‌سازی و واقعی می‌پردازیم.

کلمات کلیدی: آزمون مشخص‌سازی هاسمن، استنباط مبتنی بر درستنمایی، تابع مفصل، داده‌های طولی، داده‌های فضایی، رگرسیون حاشیه‌ای، سری زمانی، مفصل گاوسی.

لیست مقالات مستخرج از پایان‌نامه

۱. سلطانی، خ. باغیشنی، ح. آرشی، م. (۱۳۹۵) . کاربرد رگرسیون حاشیه‌ای مبتنی بر مفصل گاوسی در تحلیل داده‌های بیماری صرع. سیزدهمین کنفرانس آمار ایران. دانشگاه شهید باهنر کرمان.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

تعاریف و مفاهیم مورد نیاز

۱.۱ مقدمه

مدل‌های رگرسیونی از جمله مدل‌های پرکاربرد در علوم مختلف از جمله پزشکی، علوم طبیعی، علوم اجتماعی، اقتصاد، محیط‌شناسی و غیره است. در حالتی که متغیر پاسخ غیرنرمال باشد، به‌جای مدل‌های خطی از مدل‌های خطی تعمیم‌یافته استفاده می‌شود. اما پذیره اساسی در این مدل‌ها استقلال بین مشاهدات است. در موارد متعددی با پاسخ‌هایی مواجه می‌شویم که نوعی وابستگی بین آن‌ها برقرار است که می‌توان به‌عنوان نمونه به‌داده‌های طولی، سری‌زمانی و داده‌های‌فضایی اشاره کرد. مثلا تعداد افراد مبتلا به سرطان خون در نواحی مختلف پاسخی شمارشی از نوع فضایی است. در این موارد بایستی وابستگی پاسخ‌ها را به‌شکلی در مدل لحاظ کرد، برای این کار روش‌هایی وجود دارد که روش معمول آن استفاده از مدل‌های آمیخته خطی تعمیم‌یافته است. یک روش جانشین جدید برای لحاظ کردن وابستگی پاسخ‌های غیرنرمال، استفاده از مدل‌های رگرسیون حاشیه‌ای مبتنی بر تابع مفصل[۱] است. این رده، چارچوب کلی منعطفی برای مدل‌بندی پاسخ‌های وابسته (پیوسته و گسسته) را فراهم می‌کند.

مفصل‌ها نقش مهمی در مدل‌بندی ساختار وابستگی و همچنین ساخت توزیع‌های چندمتغیره دارند. در حالت کلی مفصل‌ها به دو دسته پارامتری و ناپارامتری تقسیم می‌شوند که در عمل از آن‌ها استفاده می‌گردد. مفصل‌های ناپارامتری به دلیل اینکه به هیچ‌گونه پیش فرضی نیاز ندارند، بیشتر مورد توجه

[۱] Copula

قرار گرفته‌اند. در این فصل تعاریف، مفاهیم و قضایایی که در فصل‌های بعد مورد استفاده قرار گرفته‌اند را بیان می‌کنیم.

۲.۱ الگوی خطی تعمیم‌یافته

تحلیل رگرسیون مجموعه‌ای از تکنیک‌های آماری برای الگوسازی و بررسی رابطه بین یک متغیر پاسخ مورد علاقه و مجموعه‌ای از متغیرهای پیشگو یا تبیین است. می‌توان گفت تحلیل رگرسیونی، پرکاربردترین روش در بین تکنیک‌های آماری است. یک الگوی رگرسیون بسیار مهم، الگوی رگرسیون خطی

$$y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (۱.۱)$$

است، که در آن y متغیر پاسخ، $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)^T$ متغیرهای پیشگو، $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ پارامترهای نامعلوم و ϵ جمله خطای تصادفی است. الگویی که نسبت به پارامترهای مجهول خطی نیست یک الگوی رگرسیون غیرخطی نامیده می‌شود. برای مثال، الگوی

$$y = \beta_1 e^{\beta_2 x} + \epsilon$$

نسبت به پارامترهای مجهول β_1 و β_2 خطی نیست. به‌طور کلی یک الگوی رگرسیون را به‌صورت

$$y = f(\mathbf{x}, \boldsymbol{\beta}) + \epsilon \qquad (۲.۱)$$

می‌توان نوشت، که در آن $\boldsymbol{\beta}$ بردار p بعدی پارامترهای مجهول و ϵ جمله خطای تصادفی ناهم‌بسته با $E(\epsilon) = 0$ و $Var(\epsilon) = \sigma^2$ است. اغلب توزیع خطاها را در مدل رگرسیون نرمال در نظر می‌گیرند. بر اساس (۲.۱)، داریم

$$\begin{aligned} E(y|\mathbf{x}) &= E[f(\mathbf{x}, \boldsymbol{\beta}) + \epsilon] \\ &= f(\mathbf{x}, \boldsymbol{\beta}) \end{aligned}$$

که در آن $f(\mathbf{x}, \boldsymbol{\beta})$ تابع انتظار الگوی رگرسیون نامیده می‌شود. در مدل رگرسیون خطی این تابع انتظار، تابعی خطی از $\boldsymbol{\beta}$ است، اما ضابطه $f(\mathbf{x}, \boldsymbol{\beta})$ می‌تواند غیرخطی نیز باشد. در الگوی رگرسیون غیرخطی دست‌کم یکی از مشتقات جزئی تابع انتظار نسبت به پارامترها حداقل به یکی از پارامترها وابسته است. برای تشریح این مطلب الگوی رگرسیون خطی (۲.۱) را با تابع انتظار رگرسیونی زیر در نظر می‌گیریم:

$$f(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j.$$

مشتقات جزئی تابع انتظار عبارتند از

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\beta})}{\partial \beta_j} = x_j, \quad j = 0, 1, \ldots, k$$

که در آن $x_0 = 1$ یک متغیر ظاهری است که عرض از مبدأ را نشان می‌دهد. توجه کنید در این حالت مشتقات جزئی، توابعی از پارامترهای مجهول نیستند. اکنون الگوی رگرسیون غیرخطی زیر را در نظر می‌گیریم:

$$\begin{aligned} y &= f(x, \boldsymbol{\beta}) + \epsilon \\ &= \beta_1 e^{\beta_2 x} + \epsilon. \end{aligned}$$

مشتقات جزئی تابع انتظار رگرسیونی نسبت به β_1 و β_2 عبارتند از

$$\frac{\partial f(x, \boldsymbol{\beta})}{\partial \beta_1} = e^{\beta_2 x}$$

$$\frac{\partial f(x, \boldsymbol{\beta})}{\partial \beta_2} = \beta_1 e^{\beta_2 x}.$$

چون مشتقات جزئی توابعی از پارامترهای مجهول β_1 و β_2 هستند، لذا این الگو غیرخطی می‌باشد. واضح است که وقتی با الگوهای رگرسیون خطی و غیرخطی سروکار داریم، توزیع نرمال نقش محوری را ایفا می‌کند. در حقیقت در روش‌های استنباطی مربوط به الگوهای رگرسیون خطی و غیرخطی فرض بر این است که متغیر پاسخ y از توزیع نرمال تبعیت می‌کند.

برای مثال، فرض کنید متغیر پاسخ گسسته، از نوع شمارشی باشد. به‌عنوان مثال اغلب با شمارش عیب‌ها یا پیشامدهای نادری چون آسیب‌ها، بیمارانی با امراض خاص و حتی با وقوع پدیده‌های طبیعی از قبیل زمین لرزه‌ها و طوفان‌های وابسته به آن مواجه می‌شویم. امکان دیگر یک متغیر پاسخ دودویی[۲] است. مطالعاتی که در آن‌ها متغیر پاسخ موفقیت یا شکست (یعنی صفر یا یک) است، که تقریباً در تمام زمینه‌های علوم و مهندسی نسبتاً متداول هستند. وضعیت‌های زیادی نیز وجود دارند که متغیر پاسخ در آن‌ها پیوسته است، اما فرض نرمال بودن کاملاً غیرواقعی است. مثال‌هایی در این مورد عبارتند از توزیع فشارها در اجزاء مکانیکی و زمان زوال اجزاء الکترونیکی یا سیستم‌ها.

این نوع پاسخ‌ها نامنفی بوده و معمولاً یک رفتار چوله به راست جدی را نشان می‌دهند. رده مدل‌های خطی تعمیم‌یافته[۳](GLM) برای برازش الگوهای رگرسیون به داده‌های پاسخ یک‌متغیری توسعه داده شده‌اند (نلدرو ودبرن، ۱۹۷۴) که در آن متغیر پاسخ دارای توزیعی از خانواده توزیع‌های نمایی است. این خانواده بسیاری از توزیع‌های مهم مانند نرمال، دوجمله‌ای، پواسن، هندسی، دوجمله‌ای منفی، نمایی و گاما را شامل می‌شود. اگر y_i، $i = 1, \ldots, n$ مقادیر پاسخ را نشان دهد، آن‌گاه GLM با

$$g[E(y_i|x_i)] = x_i^T \beta$$

نشان داده می‌شود که $\mathbf{x}_i$ بردار متغیرهای تبیینی برای مشاهده iام بوده و $\boldsymbol{\beta}$ بردار پارامترها یا ضرایب رگرسیون است. هر الگوی خطی تعمیم‌یافته سه جزء دارد: توزیع متغیر پاسخ (که گاهی اوقات ساختار خطا نامیده می‌شود)، یک پیشگوی خطی که متغیرهای تبیینی را شامل می‌شود و یک تابع پیوند $g(\cdot)$ که پیشگوی خطی را به میانگین طبیعی متغیر پاسخ مربوط می‌کند. برای مثال، الگوی رگرسیون خطی

[۲] Binary
[۳] Generalized linear model

در معادله (۱.۱) را در نظر می‌گیریم. با توجه به اینکه توزیع پاسخ نرمال بوده، پیشگوی خطی عبارت است از

$$x^T\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

و تابع پیوند یک پیوند همانی، $g(\alpha) = \alpha$، است یعنی

$$g[E(y|x)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

بنابراین، الگوی رگرسیون خطی استاندارد در معادله (۱.۱) عضوی از رده GLM است. بسته به انتخاب تابع پیوند $g(\cdot)$، یک GLM می‌تواند یک الگوی غیرخطی را شامل شود. برای مثال، اگر از تابع پیوند لگاریتمی $g(\alpha) = \ln(\alpha)$ استفاده کنیم، آن‌گاه

$$E(y|x) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}.$$

الگوی خطی تعمیم‌یافته را می‌توان به‌عنوان یکسان‌سازی الگوهای خطی و غیرخطی تلقی کرد که یک خانواده‌ای از توزیع‌های پاسخ نرمال و غیرنرمال را با هم متحد می‌کند و برازش الگو و استنباط را در سایه یک چارچوب مشترک انجام می‌دهد. علاوه بر این، نرم‌افزارهای رایانه‌ای که از این رویکرد یکسان‌سازی حمایت می‌کنند، به‌طور وسیعی در دسترس بوده و استفاده از آن‌ها آسان است. به این ترتیب، در حالی که کاربرد اولیه الگوهای GLM به علوم زیستی و صنایع داروسازی منحصر است ولی کاربردهای آن در سایر زمینه‌های علوم و مهندسی به سرعت گسترش یافته‌اند. با توجه به مفهوم الگوهای خطی تعمیم‌یافته‌ای که ارائه کردیم دو مورد مهم نمایان می‌شود: توزیع پاسخ و الگویی که میانگین پاسخ را به متغیرهای رگرسیون مربوط می‌کند. این دو مورد مستقل از یکدیگر نیستند، زیرا انواع الگوهای معین برای بعضی از توزیع‌ها نسبت به سایر توزیع‌ها مناسب‌تر می‌باشند. برای مثال، اگر پاسخ دودویی باشد توزیع برنولی برای آن استفاده می‌شود که امید ریاضی این توزیع مقداری بین صفر و یک است. پس مدلی که در آن احتمال منفی یا بزرگ‌تر از یک باشد، مطلوب نیست. علاوه بر این، اگر پاسخ شمارشی باشد، از توزیع پواسن استفاده می‌شود که پارامتر توزیع پواسن باید مقداری نامنفی باشد، پس نمی‌توان از مدلی استفاده کرد که برای این پارامتر مقادیر منفی ایجاد می‌کند. اکنون ساختار زیر را به عنوان تعمیمی از مدل‌های دوجمله‌ای و پواسن در نظر می‌گیریم. درحقیقت خانواده توزیع‌های نمایی می‌تواند به‌عنوان مدل متغیر پاسخ y در نظر گرفته شود. در این حالت، توزیع احتمال به‌صورت زیر مدل می‌شود:

$$F(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{\alpha(\phi)} + c(y,\phi)\right\}.$$

با استفاده از مدل فوق فرض می‌کنیم مشاهدات پاسخ $y_1, y_2, \ldots, y_n$ به ترتیب دارای میانگین‌های $\mu_1, \mu_2, \ldots, \mu_n$ می‌باشند به‌طوری‌که مشاهده y_i دارای توزیعی است که عضوی از خانواده توزیع‌های نمایی است. بخش منظمی از الگو پیشگوهای $x_1, x_2, \ldots, x_k$ را شامل می‌شود و الگو حول پیشگوی خطی $\eta = \mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \sum_{i=1}^{k}\beta_i x_i$ ساخته می‌شود. ارتباط این پیشگوی خطی η و متغیر پاسخ y از طریق تابع پیوند[۴] زیر برقرار می‌شود

$$\eta_i = g(\mu_i), \quad i = 1, 2, \ldots, n.$$

[۴] Link function

اصطلاح پیوند از این حقیقت ناشی می‌شود که این رابطه، بین میانگین پاسخ و پیشگوی خطی پیوند ایجاد می‌کند. دراین حالت پاسخ مورد انتظار عبارت است از

$$E(y_i) = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}).$$

در حقیقت، در رگرسیون خطی چندگانه الگوی

$$\mu_i = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}, \quad i = 1, 2, \ldots, n$$

حالت خاصی را پیشنهاد می‌کند که در آن $g(\mu_i) = \mu_i$ است. بنابراین، تابع پیوندی که مورد استفاده قرار می‌گیرد یک پیوند همانی است؛ همچنین فرض می‌کنیم واریانس $(i = 1, 2, \ldots, n), \sigma_i^2$ تابعی از میانگین μ_i است. تابع پیوند را می‌توان به‌صورت‌های مختلف انتخاب کرد. اگر

$$\eta_i = \theta_i$$

را انتخاب کنیم، در آن صورت η_i یک پیوند متعارف است. جدول ۱.۱ پیوندهای متعارفی را برای متداول‌ترین انتخاب‌های توزیع‌هایی که با الگوی خطی تعمیم‌یافته انتخاب می‌شوند، نشان می‌دهد. توابع پیوندی دیگری وجود دارند که از آن‌ها می‌توان با یک الگوی خطی تعمیم‌یافته استفاده کرد. چند مورد از این توابع عبارتند از

۱. پیوند پرابیت[۵]

$$\eta_i = \Phi^{-1}[E(y_i)]$$

که در آن $\Phi(\cdot)$ تابع توزیع تجمعی نرمال استاندارد را نشان می‌دهد؛

۲. پیوند لگ‌ـلگ مکمل[۶]

$$\eta_i = \ln[\ln[1 - \mu_i]]$$

۳. پیوند خانواده توانی[۷]

$$\eta_i = \begin{cases} \mu_i^{\lambda} & \lambda \neq 0 \\ \ln[\mu_i] & \lambda = 0 \end{cases}$$

یک نظر بسیار بنیادی این است که در الگوی خطی تعمیم‌یافته دو جزء وجود دارد؛ توزیع پاسخ (توزیع خطا نیز نامیده می‌شود) و تابع پیوند. انتخاب تابع پیوند را به نوعی مشابه انتخاب تبدیلی از پاسخ می‌توان تلقی کرد، در عین حال، درک این مطلب که تابع پیوند یک تبدیل میانگین جامعه است نه داده‌ها، با اهمیت است. برخلاف یک تبدیل، تابع پیوند از ویژگی توزیع طبیعی پاسخ استفاده می‌کند. به‌طور عینی همان‌طور که استفاده نکردن از یک تبدیل درست می‌تواند مشکلاتی را در برازش الگوی خطی به بار آورد، انتخاب‌های نادرست تابع پیوند نیز می‌تواند به مشکلاتی در مورد یک الگوی خطی تعمیم‌یافته منجر شود.

---

[۵] Probit link
[۶] C log-log link
[۷] Power family link

جدول ۱.۱: پیوندهای متعارف برای الگوهای خطی تعمیم‌یافته

	پیوند متعارف	توزیع
(پیوند همانی)	$\eta_i = \mu_i$	نرمال
(پیوند لجستیک)	$\eta_i = \ln(\frac{p}{1-p})$	دوجمله‌ای
(پیوند لگاریتمی)	$\eta_i = \ln(\mu_i)$	پواسن
(پیوند وارون)	$\eta_i = \frac{1}{\mu_i}$	نمایی
(پیوند وارون)	$\eta_i = \frac{1}{\mu_i}$	گاما

صورت‌های بی‌شماری از رگرسیون‌ها وجود دارد که می‌توان در مدل‌سازی متغیر پاسخ از آن‌ها استفاده کرد. هر صورت دارای اهمیت و شرایط خاصی است که در آن مناسب‌ترین کارایی خود را داراست. در این بخش، چند مدل معمول رگرسیون را به‌طور مختصر شرح می‌دهیم.

تعریف ۱.۲.۱. (رگرسیون خطی ساده[۸]) رگرسیون خطی ساده یک ابزار آماری است که در آن به‌بررسی رابطه یک متغیر مستقل (پیش‌بین) و یک متغیر وابسته پرداخته می‌شود. از طرفی، مقیاس هر دو متغیر (وابسته و مستقل) در سطح سنجش کمترین فاصله را دارد. بنابراین، در رگرسیون خطی ساده، مقادیر متغیر وابسته (پاسخ) را روی مقادیر متغیر دیگر (متغیر پیش‌گو) به کمک یک معادله خطی (خط مستقیم) برآورد می‌شود.

تعریف ۲.۲.۱. (رگرسیون خطی چندگانه[۹]) حال اگر تعداد متغیرهای پیش‌گو بیش‌از یک عدد شود، مدل رگرسیون، خطی چندگانه نامیده می‌شود.

معادله رگرسیون خطی چندگانه به‌صورت زیر است

$$E(y\,|X) = X\beta,$$

که در آن $\mathbf{y} = (y_1, ..., y_n)^T$، $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$، $\mathbf{x}_i \in \mathbb{R}^{\mathbf{k}}$ و $\boldsymbol{\beta} = (\beta_1, ..., \beta_i)^T$.

تعریف ۳.۲.۱. (رگرسیون لجستیک[۱۰]) رگرسیون لجستیک برای یافتن احتمال رویداد مبنی بر موفقیت و شکست به کار می‌رود. در طبیعت زمانی از رگرسیون لجستیک استفاده می‌کنیم که متغیر وابسته دودویی باشد. این مدل جز رده GLM است که تابع آن به‌صورت زیر تعریف می‌شود

$$\text{logit}(p) \;=\; \ln\left(\frac{p}{1-p}\right) \;= \mathbf{x}_i^T\boldsymbol{\beta} = \sum_{j=1}^{k} x_i\beta_j, \quad i = 1, 2, ..., n$$

که در آن

$$p = \Pr(y_i = 1\,|X) = \frac{e^{\beta_1 x_{1,i}+...+\beta_k x_{k,i}}}{1 + e^{\beta_1 x_{1,i}+...+\beta_k x_{k,i}}}$$

و $X = (x_1, ..., x_n)^T$.

[۸] Simple linear regression
[۹] Multiple linear regression
[۱۰] Logistic regression

تعریف ۱.۲.۱. (رگرسیون لجستیک چندگانه[۱۱]) رگرسیون لجستیک چندگانه به وضعیتی گفته می‌شود که متغیر وابسته بیش‌از یک سطح دارد. برای نمونه اگر بخواهیم ببینیم که چه عواملی موجب می‌شود که مخاطبان کدام یک از شبکه‌های یک یا دو یا سه را برای دیدن انتخاب کنند، از رگرسیون لجستیک چندگانه سود می‌بریم. در این مثال، عوامل مختلف به‌عنوان متغیر مستقل و انتخاب شبکه به‌عنوان متغیر وابسته شناخته می‌شود. این متغیر وابسته سه سطح دارد که پاسخ آن آری یا خیر است (استفاده از شبکه ۱، استفاده از مطالب شبکه ۲، استفاده از محصولات رسانه‌ای شبکه ۳).

تعریف ۱.۲.۵. (رگرسیون چندجمله ای[۱۲]) یک معادله رگرسیون را یک معادله رگرسیون چندجمله‌ای گوییم اگر توان متغیر مستقل بیش‌تر از ۱ باشد. معادله زیر یک معادله چند جمله‌ای را نشان می‌دهد:

$$y = a + bx^2.$$

در این روش رگرسیون، مناسب‌ترین خط، خط مستقیم نیست بلکه یک منحنی است که متناسب با نقاط رسم می‌شود.

تعریف ۱.۲.۶. (رگرسیون پواسن[۱۳]) رگرسیون پوآسون یك مدل خطی عمومی برای تجزیه و تحلیل رگرسیونی است که برای مدل داده‌های قابل شمارش و جداول احتمالات استفاده شده است. این مدل متعلق به رده GLM است، رگرسیون پواسون فرض می‌کند متغیر پاسخ Y دارای توزیع پواسون است. مدل رگرسیونی پواسون گاه به عنوان مدل خطی لگاریتمی[۱۴] شناخته می‌شود، به ویژه هنگامی که مدل‌های جدول احتمالاتی مورد استفاده قرار گیرند. به‌بیان دیگر، هنگامی که فراوانی متغیر وابسته برای فرآیند مدل‌سازی کم باشد و به اصطلاح داده‌های مشاهده شده دارای فراوانی صفر در خود باشند در این حالت مدل‌های لگاریتم طبیعی توانایی پیش‌بینی مناسب فراوانی متغیر وابسته را ندارند. دلیل اصلی این امر فرض توزیع نرمال داده‌هاست در این شرایط مدل رگرسیون پواسون می‌تواند مورد استفاده قرار گیرد. در رگرسیون پواسن متغیر پاسخ به‌صورت زیر مدل می‌شود:

$$\log(E(y\,|x)) = \mathbf{x}^T\boldsymbol{\beta},$$

با داشتن بردار پارامتر رگرسیون پواسن β و بردار متغیرهای پیش‌بین $\mathbf{x}$، می‌توان پیش‌بینی را به‌صورت زیر به‌دست آورد:

$$E(y\,|x) = e^{x^T\beta}.$$

تعریف ۱.۲.۷. (رگرسیون رتبه‌ای[۱۵]) در رگرسیون رتبه‌ای اثر متغیر و یا متغیرهای مستقل برروی متغیر وابسته‌ای که سطوح مختلف و رتبه‌ای دارد، مدل‌بندی می‌شود.

---

[۱۱] Multiple logistic regression
[۱۲] Ranking regression
[۱۳] Poisson regression
[۱۴] Log-linear model
[۱۵] Ranking regression

تعریف ۸.۲.۱. (رگرسیون پرابیت[۱۶]) هنگامی که خروجی (متغیر وابسته) دو حالت داشته باشد (آری یا خیر) و هدف بررسی شدت متغیر پیش‌بین در پیش‌بینی این دو حالت باشد، از مدل رگرسیون پروبیت استفاده می‌شود. برای اطلاعات بیشتر درخصوص رده مدل‌های GLM و کاربردهای آن مک‌کالاک و نلدر (۱۹۸۹) را ببینید.

۳.۱ نظریه مفصل

مطالعه مفصل‌ها و کاربرد آن‌ها در آمار، یک پدیده نسبتاً مدرن است. از یک نقطه نظر مفصل‌ها توابعی هستند که توابع توزیع چند متغیره را به توابع توزیع حاشیه‌ای آن‌ها متصل می‌کنند و از دیدگاه دیگر مفصل‌ها را می‌توان به‌صورت توابع توزیع چند متغیره‌ای تعریف کرد که توابع توزیع حاشیه‌ای آن‌ها به‌صورت یکنواخت روی فاصله $(0,1)$ توزیع شده است.

معمولاً در نوشته‌های آماری، مفصل را با واژه کاپولا [۱۷] می‌شناسیم، واژه کاپولا یک اسم لاتین است که به معنای ''یک پیوند، گره و یا زنجیر(به دیکشنری لاتین کاسل مراجعه کنید)''[۱۸] می‌باشد.

این واژه اولین بار در سال ۱۹۵۹ توسط اسکلار در یک قضیه به‌کار رفت. در این قضیه از کاپولا برای توصیف توابعی که توابع توزیع تک بعدی آن‌ها برای ساخت تابع توزیع چندمتغیره به هم متصل می‌شوند، استفاده شد.

طبق مطالعات ''فیشر''مفصل‌ها به دو دلیل مورد علاقه آماردانان می‌باشند:

۱. چون روشی برای مطالعه انواعی از اندازه‌های وابستگی هستند که به مقیاس بستگی ندارند.[۱۹]

۲. چون بعضی مواقع با استفاده از شبیه‌سازی‌ها مشاهده می‌شود که مفصل‌ها یک نقطه شروع برای ساختن خانواده توزیع‌های دو متغیره هستند.''

طبق مطالعه هافدینگ (۱۹۴۱-۱۹۴۰) می‌توان توزیع‌های دو متغیره‌ای را یافت که تکیه‌گاه آن‌ها در مربع $[-\frac{1}{2},\frac{1}{2}]^2$ واقع شده و توابع توزیع حاشیه‌ای آن‌ها به‌صورت یکنواخت در فاصله $[-\frac{1}{2},\frac{1}{2}]$ پخش شده‌اند، طبق نظر شوایزر در سال ۱۹۹۱: ''اگر هافدینگ برای نرمال‌سازی‌اش از مربع $[0,1]^2$ به جای $[-\frac{1}{2},\frac{1}{2}]^2$ استفاده می‌کرد، اینک او کاشف مفصل‌ها بود.''

هافدینگ بهترین کران‌های ممکن را نیز برای این توابع به‌دست آورد، توزیع‌های متناظر با این کران‌ها را تعیین نموده و اندازه‌های وابستگی را که پایای مقیاسی (پایا تحت تبدیلات اکیداً صعودی) هستند را مورد مطالعه قرار داد. پس از هافدینگ، فرشه و اسکلار، توابعی که آن‌ها را امروزه تحت عنوان تابع مفصل می‌شناسیم، توسط دانشمندان دیگر مانند، کیملدورف و سمپسون (۱۹۷۵)، (تحت عنوان نمایش‌های یکنواخت) و گالامبوس (۱۹۷۸) و دی‌هیولز (۱۹۷۸)، (تحت عنوان توابع وابستگی) نیز به‌دست آمدند.

---

[۱۶] Probit regression
[۱۷] Copula
[۱۸] Cassell's Latin Dictionary
[۱۹] Scale-free

۱.۳.۱ تابع مفصل

وابستگی بین متغیرهای تصادفی حقیقی $X_1, X_2, \ldots, X_n$ به‌طور کامل به‌وسیله‌ی تابع توزیع توأم آن‌ها، $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ توصیف می‌شود. شرایط لازم و کافی برای این‌که تابع از راست پیوسته‌ی $F : \mathbb{R}^n \to [0,1]$ یک تابع توزیع باشد، عبارت‌اند از:

$$
\begin{aligned}
&i)\ \lim_{x_j \to -\infty} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = 0, \quad j = 1, 2, \ldots, n \\
&ii)\ \lim_{x_j \to +\infty} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = 1, \quad j = 1, 2, \ldots, n \\
&iii)\ \forall (a_1, a_2, \ldots, a_n), (b_1, b_2, \ldots, b_n), \quad a_i < b_i \\
&\sum_{i_1=1}^{2} \sum_{i_2=1}^{2} \cdots \sum_{i_n=1}^{2} (-1)^{i_1+i_2\cdots+i_n} F_{X_1,X_2,\ldots,X_n}(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \geq 0 \\
&x_{i_1} = a_i, \quad x_{i_2} = b_i, \quad i = 1, 2, \ldots, n
\end{aligned}
$$

اگر تابع F دارای مشتق‌های مرتبه 1 تا n باشد، آن‌گاه ویژگی سوم معادل است با

$$\frac{\partial^n F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)}{\partial_{x_1}, \partial_{x_2}, \ldots, \partial_{x_n}} \geq 0.$$

در بسیاری از مسائل کاربردی، توزیع‌های حاشیه‌ای معلوم یا قابل برآورد هستند اما توزیع توأم نامعلوم یا برآورد آن دشوار است. مسأله تعیین توزیع توأم یا برآورد آن با استفاده از توزیع‌های حاشیه‌ای از موارد مهم در بسیاری از زمینه‌های کاربردی است. ایده تفکیک تابع توزیع توأم به دو قسمت، که یکی توصیف‌کننده ساختار وابستگی و دیگری حاشیه‌ای‌هاست، منجر به شکل‌گیری مفهوم مفصل شد که نخستین بار توسط اسکالار (۱۹۵۹) مطرح گردید. اگر متغیرهای تصادفی $X_1, X_2, \ldots, X_n$ با تابع توزیع توأم $F_{X_1,X_2,\ldots,X_n}(\cdot, \ldots, \cdot)$ و توابع توزیع حاشیه‌ای $F_{X_1}(\cdot), F_{X_2}(\cdot), \ldots, F_{X_n}(\cdot)$ باشند، آن‌گاه یک مفصل تابعی مانند $C(\cdot, \ldots, \cdot)$ است که

$$C : F_{X_1}, F_{X_2}, \ldots, F_{X_n} \ \to F_{X_1,X_2,\ldots,X_n}.$$

تعریف ۱.۳.۱. تابع $C(u_1, u_2, \cdots, u_n) : [0,1]^n \longrightarrow [0,1]$ را یک تابع مفصل (در حالت n بعدی) گوییم، اگر و فقط اگر

۱. مفصل C زمین‌خورده [۲۰] باشد:

یعنی برای هر $(u_1, \cdots, u_n) \in [0,1]^n$ حداقل یک $j \in \{1, \cdots, n\}$ وجود داشته باشد به‌طوری‌که اگر $u_j = 0$، آن‌گاه،

$$C(u_1, \cdots, u_n) = 0.$$

[۲۰] Grounded

۲. مفصل C، n صعودی باشد:

یعنی برای هر یک از دو بردار n بعدی u و v که

$$u = (u_1, u_2, \cdots, u_n) \in [0,1]^n$$
$$v = (v_1, v_2, \cdots, v_n) \in [0,1]^n$$

و

$$v_j \leq u_j, \quad j \in \{1, \cdots, n\}$$

رابطهٔ زیر برقرار باشد

$$V_c(B) = D_{u_n}^{v_n} \cdots D_{u_1}^{v_1} C(t_1, \cdots, t_n) \geq 0,$$

که در آن

$$B = ([u_1, v_1] \times \cdots \times [u_n, v_n]), \quad t \in [0,1]^n,$$

و

$$D_{u_j}^{v_j} C(t_1, \cdots, t_n) = C(t_1, \cdots, t_{j-1}, v_j, t_{j+1}, \cdots, t_n) - CC(t_1, \cdots, t_{j-1}, u_j, t_{j+1}, \cdots, t_n).$$

۳. مفصل C دارای توابع مفصل حاشیه‌ای C_k، $k \in \{1, \cdots, n\}$ می‌باشد که،

$$C_k(u) = u, \quad u \in [0,1]$$

قضیه ۲.۳.۱ (قضیه اسکلار)**.** اگر $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ یک تابع توزیع n متغیره با تابع توزیع‌های حاشیه‌ای $F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_n}(x_n)$ باشد، آن‌گاه یک مفصل n بعدی C وجود دارد به‌طوری که

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = C(F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_n}(x_n)) \qquad (۳.۱)$$

و بر عکس اگر C یک مفصل و $F_{X_1}(x_1), F_{X_2}(x_2), \ldots, F_{X_n}(x_n)$ توابع توزیع دلخواه باشند، تابع n متغیره F تعریف‌شده در (۳.۱)، یک تابع توزیع توأم n متغیره با حاشیه‌ای‌های $F_{X_1}, F_{X_2}, \ldots, F_{X_n}$ است. به‌ویژه اگر توزیع‌های حاشیه‌ای همگی پیوسته باشند، آن‌گاه تابع C یکتاست.

در صورتی که تابع توزیع توأم دارای حاشیه‌ای‌های پیوسته باشد، بنابه قضیه اسکلار، تابع مفصل آن‌ها به آسانی به‌صورت زیر به‌دست می‌آید

$$C(u_1, u_2, \ldots, u_n) = F_{X_1,X_2,\ldots,X_n}(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2), \ldots, F_{X_n}^{-1}(u_n)).$$

بنابر مطالب بیان‌شده، اگر $X_1, X_2, \ldots, X_n$ متغیرهای تصادفی با توابع توزیع پیوسته باشند، آن‌گاه بنابه تبدیل انتگرال احتمال، $U_j = F_j(X_j)$ دارای توزیع یکنواخت استاندارد است، و در نتیجه C تابع توزیع توأم متغیرهای تصادفی یکنواخت بر بازه $[0,1]$ است.

مثال ۱.۳.۳. توزیع نمایی دومتغیره گامبل با تابع توزیع

$$F_\theta(x,y)=\begin{cases}1-e^{-x}-e^{-y}+e^{-(x+y+\theta xy)} & x\geq 0,\ y\geq 0,\ \theta\in[0,1]\\ 0 & \text{در غیر این صورت}\end{cases}$$

دارای توزیع‌های حاشیه‌ای نمایی

$$F_Y(y)=\begin{cases}1-e^{-y} & y\geq 0\\ 0 & \text{سایر جاها}\end{cases}\quad,\quad F_X(x)=\begin{cases}1-e^{-x} & x\geq 0\\ 0 & \text{سایر جاها}\end{cases}$$

و شبه‌وارون‌های زیر به‌ازای $u,v\in(0,1)$ است.

$$F_Y^{-1}(v)=-\ln(1-v)\quad,\quad F_X^{-1}(u)=-\ln(1-u)$$

بنابراین تابع مفصل آن به‌صورت زیر است:

$$C_\theta(u,v)=u+v-1+(1-u)(1-v)e^{-\theta\ln(1-u)\ln(1-v)}$$

اگر توزیع‌های حاشیه‌ای پیوسته باشند، با مشتق‌گیری از طرفین (۳.۱) به‌ازای $n=2$، تابع چگالی توأم براساس تابع مفصل، به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned}f_{X,Y}(x,y)&=\frac{\partial^2F_{X,Y}(x,y)}{\partial x\partial y}=f_X(x)f_Y(y)\frac{\partial^2C(u,v)}{\partial u\partial v}\\&=f_X(x)f_Y(y)c(F_X(x),F_Y(y))\end{aligned}$$

و $c(u,v)=\frac{\partial^2C(u,v)}{\partial u\partial v}$ تابع چگالی مفصل نامیده می‌شود.

در ادامه به‌معرفی چند خانواده از مفصل می‌پردازیم.

تعریف ۱.۳.۴. (مفصل گاوسی[۲۱]) این نوع مفصل در امور مالی کاربردهای زیادی دارد و از توزیع نرمال دومتغیره استاندارد با ضریب همبستگی ρ، با استفاده از قضیه اسکالار، ساخته می‌شود. اگر $\Phi_\rho(\cdot)$ و $\Phi(\cdot)$ به‌ترتیب توابع توزیع نرمال دومتغیره استاندارد با ضریب همبستگی ρ و توزیع نرمال استاندارد باشند، آن‌گاه تابع مفصل گاوسی به‌صورت زیر تعریف می‌شود

$$C_\rho(u,v)=\Phi_\rho\left(\Phi^{-1}(u),\Phi^{-1}(v)\right)\quad u,v\in[0,1].$$

تابع چگالی مفصل گاوسی نیز با مشتق‌گیری از تابع مفصل به‌دست می‌آید. یعنی

$$c_\rho(u,v)=\frac{\phi_{X,Y,\rho}\left(\Phi^{-1}(u),\Phi^{-1}(v)\right)}{\phi\left(\Phi^{-1}(u)\right)\phi\left(\Phi^{-1}(v)\right)}$$

که

$$\phi_{X,Y,\rho}(x,y)=\frac{1}{2\pi(1-\rho^2)}exp\left\{-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)\right\}$$

و $\phi(\cdot)$ تابع چگالی نرمال استاندارد است.

[۲۱] Gaussian

تعریف ۵.۳.۱. (مفصل ارشمیدسی[۲۲]) یکی از مهمترین مفصل‌های خانواده مفصل‌ها است که دارای ویژگی‌هایی مانند شرکت‌پذیری و ساختارهای متنوع وابستگی است. برخلاف مفصل‌های بیضوی مانند گاوسی، اغلب مفصل‌های این خانواده دارای شکل بسته می‌باشند که از قضیه اسکالار نتیجه نمی‌شود. یک مفصل ساده از این خانواده

$$C(u,v) = \Psi^{-1}(\Psi(u) + \Psi(v)) \qquad (۴.۱)$$

می‌باشد که مفصل ارشمیدسی نامیده می‌شود و تابع توزیع توأم براساس این خانواده از مفصل

$$F_{X,Y}(x,y) = \Psi^{-1}(\Psi(F_X(x)) + \Psi(F_Y(y)))$$

است. Ψ تابع مولد نامیده می‌شود و دارای ویژگی‌های زیر است:

$$\Psi(1) = 0,\ \lim_{x\to 0}\Psi(x) = \infty\,,\ \Psi'(x) < 0\,,\ \Psi''(x) > 0$$

به‌چند مورد از مفصل‌های این خانواده اشاره می‌کنیم:

الف. مفصل حاصل‌ضربی در این نوع مفصل، وابستگی بین متغیرها وجود ندارد و با قرار دادن $\Psi(x) = -\ln(x)$ در رابطه (۴.۱) به‌دست می‌آید. پس در این نوع مفصل

$$C(u,v) = uv$$

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

ب. مفصل کلایتون[۲۳] این مفصل به‌عنوان عضوی از خانواده مفصل‌های ارشمیدسی، توسط کلایتون معرفی گردید، و با جایگزینی $\Psi(x) = x^\theta - 1$ به‌ازای $\theta \le 0$ در رابطه (۴.۱) به‌دست می‌آید. بنابراین

$$C(u,v) = (u^\theta + v^\theta - 1)^{\frac{1}{\theta}}$$

$$F_{X,Y}(x,y) = \left(F_X(x)^\theta + F_Y(y)^\theta - 1\right)^{\frac{1}{\theta}}$$

این مفصل دارای وابستگی دمی پایینی می‌باشد و وابستگی دمی بالایی آن صفر است.

ج. مفصل گامبل[۲۴] این مفصل عضو دیگری از خانواده مفصل‌های ارشمیدسی است، و با جایگزینی $\Psi(x) = (-\ln(x))^\alpha$ در (۴.۱) به‌دست می‌آید. این مفصل دارای وابستگی دمی بالایی است و وابستگی دمی پایینی آن صفر است. این مفصل نخستین بار توسط گامبل (۱۹۶۰) معرفی گردید.

د. مفصل فرانک [۲۵] عضو دیگری از خانواده مفصل‌های ارشمیدسی است، این مفصل با جایگزینی $\Psi(x) = \left[\frac{e^{\alpha x}-1}{e^{\alpha}-1}\right]$ در (۴.۱) حاصل می‌شود. تابع مفصل فرانک نیز دارای ناهمبستگی در دم‌ها می‌باشد.

[۲۲] Archimedean
[۲۳] Clayton
[۲۴] Gumbel
[۲۵] Frank

تعریف ۱.۳.۶. مفصل فارلی ـ گامبل ـ مورگنشترن [۲۶] این نوع مفصل به‌صورت زیر تعریف می‌شود:

$$C_\alpha(u,v) = uv[1+\alpha(1-u)(1-v)]\ ,\ |\alpha| \leq 1$$

تعریف ۱.۳.۷. مفصل بقاء اگر C تابع مفصل متغیرهای تصادفی X, Y باشد آن‌گاه

$$\begin{aligned}\bar{F}_{X,Y}(x,y) &= P(X>x, Y>y)\\ &= 1 - F_X(x) - F_Y(y) + F_{X,Y}(x,y)\\ &= \bar{F}_X(x) + \bar{F}_Y(y) - 1 + C(F_X(x), F_Y(y))\\ &= \bar{F}_X(x) + \bar{F}_Y(y) - 1 + C(1-\bar{F}_X(x), 1-\bar{F}_Y(y))\end{aligned}$$

بنابراین اگر تابع $\hat{C} : [0,1]^2 \to [0,1]$ را به‌صورت زیر تعریف کنیم

$$\hat{C}(u,v) = u + v - 1 + C(1-u, 1-v)$$

آن‌گاه

$$\bar{F}_{X,Y}(x,y) = \hat{C}(\bar{F}_X(x)\ ,\ \bar{F}_Y(y))$$

تابع $\hat{C}$ یک تابع مفصل است که مفصل بقای X, Y نامیده می‌شود که C مفصل (X,Y) است.

روش های پارامتری به سه دلیل عمده کمتر مورد توجه هستند. اول این که همان‌طور که گوی [۲۷] (۲۰۰۹) مطرح کرده است روش پارامتری نیازمند محاسبات زیاد و گاهی اوقات یافتن تابع مفصل بهینه بسیار دشوار است. دوم این‌که همان‌طور که از بدفورد [۲۸] و همکاران (۲۰۱۴) نقل شده است، یافتن تابع مفصل بهینه تبدیل یک مسئله دشوار به مسئله دشوار دیگر است، یعنی به‌جای یافتن توزیع توأم باید به‌دنبال تابع مفصل بهینه باشیم. همچنین با محدودکردن خود به کلاس مفصل‌های پارامتری از همه مزایای مفصل‌ها استفاده نشده است. سوم، معلوم نیست کلاس خانواده پارامتری به‌اندازه کافی برای بیان ساختار وابستگی متغیرها در حالت کلی مناسب باشند.

با توجه به دلایل مطرح شده، مفصلهای ناپارامتری که هیچ‌گونه پیش‌فرضی در مورد ساختار وابستگی بین متغیرها نیاز ندارند، بیش‌تر مورد توجه هستند. تعدادی از این مفصل‌های ناپارامتری در ادامه معرفی می‌شوند:

تعریف ۱.۳.۸. مفصل تجربی مفصل تجربی به عنوان اولین مفصل ناپارامتری توسط دهیوولز [۲۹] (۱۹۷۹) معرفی گردید. تابع مفصل تجربی یک براوردگر ناپارامتری از مفصل معمولی است که می‌توان از ان برای بسیاری از اهداف از جمله آزمون استقلال استفاده کرد. همچنین این مفصل از تابع توزیع تجربی به‌دست می‌آید، که به‌صورت زیر است:

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} I(\frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v)$$

[۲۶] Farlie-Gumbel-Morgenstern
[۲۷] Gui
[۲۸] Bedford
[۲۹] Deheuvels

روش برآورد:

$$F_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x, Y_i \le y)$$

$$C(x,y) = F_{X,Y}\left(F_X^{-1}(x), F_Y^{-1}(y))\right)$$

$$\begin{aligned} C_n(x,y) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i \le F_{n,X^{-1}}(x), Y_i \le F_{n,Y^{-1}}(y)) \\ &= \frac{1}{n}\sum_{i=1}^{n} I(F_{n,X}(X_i) \le x, F_{n,Y}(Y_i) \le y) \end{aligned}$$

این مفصل برای بسیاری از اهداف از جمله آزمون‌های استقلال مورد استفاده قرار می‌گیرد. دهیوولز همگرایی تقریبا همه‌جا $(a.s.)$ و همگرایی یکنواخت مفصل تجربی را به مفصل اصلی نشان داد.

۴.۱ انواع ساختارهای همبستگی

مدل‌های خطی تعمیم‌یافته، کلاس جامعی از مدل‌ها، برای تحلیل رگرسیونی داده‌های پیش‌بین گسسته و پیوسته که دارای توزیع نمایی باشند، فراهم آورده است؛ به این ترتیب که با اعمال یک ترکیب غیرخطی مناسب روی میانگین متغیر پاسخ ترکیبی خطی از متغیرهای توضیحی می‌سازد. تعمیم مدل‌های خطی تعمیم‌یافته به داده‌های هم‌بسته و چندمتغیره، توانایی این مدل‌ها را در تحلیل داده‌های هم‌بسته ایجاد نموده است. نمونه‌ای از این تعمیم، مدل‌های حاشیه‌ای می‌باشند که مشخصات کلی آن به‌صورت زیر است:

آ) میانگین حاشیه‌ای یا احتمال پاسخ برای نمونه iام به‌صورت زیر تعریف می‌شود:

$$g[E(y_{ij}|\mathbf{x}_{ij})] = g(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}, \quad i = 1,2,\ldots,n \quad j = 1,2,\ldots,t_i$$

که در آن $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T, \mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$، بردارهای p بعدی از متغیرهای توضیحی و پارامترها و g یک تابع پیوند مناسب و معلوم می‌باشند.

ب) واریانس حاشیه‌ای که تابعی از میانگین حاشیه‌ای است، به‌صورت زیر تعیین می‌شود:

$$var(y_{ij}|x_{ij}) = v(\mu_{ij})\phi$$

که در آن v یک تابع معلوم و ϕ، پارامتر مقیاس است.

پ) همبستگی بین y_{ij} و $y_{ij'}$،که تابعی از میانگین‌های حاشیه‌ای و پارامتر اضافی α است، به‌صورت زیر تعیین می‌شود:

$$corr(y_{ij}, y_{ij'}) = \rho(\mu_{ij}, \mu_{ijT}, \boldsymbol{\alpha})$$

که در آنρ،تابعی معلوم و $\boldsymbol{\alpha}$، برداری از پارامترهای نامعلوم می‌باشند.

معادلات برآوردگر تعمیم‌یافته[۳۰] (GEE) که توسط لیانگ و زیگر (۱۹۸۶) پیشنهاد شده است، یک روش برآورد در مدل‌های خطی تعمیم‌یافته به‌شمار می‌آید.

این روش که تعمیمی از روش شبه‌درستنمایی[۳۱] (ودربرن،۱۹۷۴) به داده‌های چندمتغیره و هم‌بسته می‌باشد، دسته‌ای از معادلات برآوردگر برای پارامترهای مدل رگرسیونی ارائه می‌دهد.در تعیین این معادلات، به گشتاورهای مرتبه اول و دوم پاسخ‌ها و تعیین ماتریس همبستگی عملی برای واحدها نیاز است؛ و درباره توزیع توأم پاسخ‌ها هیچ فرضی اختیار نمی‌شود.

با فرض این‌که، $\mu_i(\boldsymbol{\beta})$ و $\mathbf{V}_i(\boldsymbol{\alpha})$ به‌ترتیب میانگین (گشتاور مرتبه اول) و ماتریس واریانس‌ـ کوواریانس (گشتاور مرتبه دوم) برای Y_i باشند، معادلات برآوردگر تعمیم‌یافته به‌صورت زیر تعریف می‌شوند:

$$\mathbf{S}(\boldsymbol{\beta},\boldsymbol{\alpha})=\sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\beta},\boldsymbol{\alpha})=\sum_{i=1}^{n}\mathbf{D}_i^T\mathbf{V}_i^{-1}(\boldsymbol{\alpha})(\mathbf{Y}_i-\mu_i(\boldsymbol{\beta}))=0$$

به‌طوری‌که:

$$\mathbf{D}_i=\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}$$

و ماتریس واریانس‌ـ کوواریانس $\mathbf{V}_i(\boldsymbol{\alpha})$، به‌صورت زیر نمایش داده می‌شود:

$$\mathbf{V}_i(\boldsymbol{\alpha})=\phi\mathbf{A}_i^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{\frac{1}{2}}$$

که در آن، $\mathbf{A}_i$ یک ماتریس قطری با عناصر قطری $v(\mu_{ij})$ می‌باشد. همچنین برای $\mathbf{R}_i(\boldsymbol{\alpha})$ که به‌عنوان ماتریس همبستگی عملی شناخته می‌شود؛ ساختارهای متعددی در نظر گرفته شده است. نظر به‌این‌که برآورد پارامترهای رگرسیونی به انتخاب این ساختارها وابسته است، آشنایی با انواع ساختارها امری ضروری است که در ادامه به‌معرفی آن‌ها می پردازیم.

۱.۴.۱ ساختار استقلال

ساختار استقلال[۳۲]، ساده‌ترین نوع ساختار و در عین حال نامعقول‌ترین نوع به‌دلیل صفر فرض شدن (نادیده گرفتن) همبستگی بین مشاهدات می‌باشد. با این وجود برای مقایسه با سایر ساختارها به‌عنوان ساختار پیش‌فرض در بسیاری از نرم‌افزارهای آماری آورده شده است. نحوه نمایش و شکل برداری آن به‌ترتیب به‌صورت زیر می‌باشد:

$$corr(y_{ij},y_{ij'})=\begin{cases}1 & j=j'\\ 0 & j\neq j'\end{cases}$$

[۳۰] Generalized estimating equations
[۳۱] Quasi Likelihood
[۳۲] Independent

ماتریس حاصل از این ساختار وابستگی برای یک نمونه چهارتایی به‌صورت زیر است:

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

۲.۴.۱ همبستگی تبادل‌پذیر

همبستگی تبادل‌پذیر[۳۳]، ساده‌ترین نوع همبستگی است که همبستگی بین تمام مشاهدات را بدون در نظر گرفتن فاصله (لگ) آن‌ها یکسان در نظر می‌گیرد و به‌صورت زیر نمایش داده می‌شود:

$$corr(y_{ij}, y_{ij'}) = \begin{cases} 1 & j = j' \\ \alpha & j \neq j' \end{cases}$$

و شکل ماتریسی آن، به‌عنوان مثال برای واحدی با چهار مشاهده به‌صورت زیر است:

$$\mathbf{R} = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{pmatrix}.$$

۳.۴.۱ همبستگی غیرساختاری

ویژگی همبستگی غیرساختاری[۳۴] فرض منحصر به‌فرد بودن همبستگی بین هر دو جفت از مشاهدات می‌باشد که اگرچه واقع‌گرایانه است، اما به‌دلیل عدم وجود ساختاری خاص و زیاد بودن پارامترها، پیچیده‌ترین نوع همبستگی به‌شمار می‌آید. لذا این ساختار برای مجموعه داده‌ها با حجم بالای مشاهدات تکرارشده مناسب نمی‌باشد، زیرا در چنین مواقعی با انتخاب این ساختار در بیش‌تر نرم‌افزارهای آماری با پیغام خطا روبرو خواهیم شد.

$$corr(y_{ij}, y_{ij'}) = \begin{cases} 1 & j = j' \\ \alpha_{ij'} & j \neq j' \end{cases}$$

مشابه ساختار تبادل‌پذیر، صورت ماتریسی این نوع ساختار برای چهار مشاهده به شکل زیر است:

$$\mathbf{R} = \begin{pmatrix} 1 & \alpha_{21} & \alpha_{31} & \alpha_{41} \\ \alpha_{21} & 1 & \alpha_{32} & \alpha_{42} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1 \end{pmatrix}.$$

[۳۳] Exchangeable, Compound Symmetry
[۳۴] Unstructed General

۴.۴.۱ همبستگی خودبرگشت مرتبه اول

همبستگی خودبرگشت مرتبه اول [۳۵] رابطه‌ای معکوس بین همبستگی و فاصله‌ی مشاهدات اعمال می‌کند یعنی اعمال بالاترین همبستگی به مشاهدات مجاور و کاهش همبستگی با فاصله گرفتن مشاهدات از یکدیگر. این همبستگی که به‌صورت زیر نمایش داده می‌شود برای داده‌هایی با فواصل برابر مناسب بوده و به‌جهت اعمال همبستگی به تمام مشاهدات و عدم پیچیدگی به‌دلیل کم بودن تعداد پارامترها جزو ساختارهای پرکاربرد می‌باشد.

$$corr(y_{ij}, y_{ij'}) = \alpha^{|j-j'|}$$

$$\mathbf{R} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}.$$

۵.۴.۱ ساختار m ـ وابستگی

ساختار همبستگی بعدی، ساختار m ـ وابستگی [۳۶] است که در آن فرض می‌شود، تنها مشاهداتی که اختلاف فاصله آن‌ها کمتر از m است دارای همبستگی می‌باشند. نمایش این همبستگی به‌صورت زیر می‌باشد:

$$corr(y_{ij}, y_{ij'}) = \begin{cases} \alpha_{|j-j'|} & |j-j'| \leq m \\ 0 & \text{در غیر این‌صورت} \end{cases}$$

و شکل ماتریسی آن، به‌عنوان مثال برای واحدی با چهار مشاهده و $m = 2$ به‌صورت زیر است:

$$\mathbf{R} = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & 0 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ 0 & \alpha_1 & \alpha_2 & 1 \end{pmatrix}.$$

بنابراین تمام جفت مشاهداتی که اختلاف فاصله آن‌ها برابر باشد دارای همبستگی مشابه می‌باشند. این ساختار به دلیل زیاد بودن تعداد پارامترها نسبت‌به ساختار خودبرگشت پیچیده‌تر است و هنگامی که تعداد تکرارها زیاد و m بزرگ باشد، این پیچیدگی افزایش می‌یابد.

نکته ۱. تاکنون تمامی ساختارهای معرفی‌شده برای مجموعه داده‌هایی با فواصل برابر مناسب می‌باشند. وجود داده‌هایی با فواصل نابرابر موجب شده است که ساختارهای همبستگی مناسب این داده‌ها نیز روی کار آیند. اساس کار به گونه‌ای است که ابتدا فواصل واقعی را شناسایی و سپس فاصله اقلیدسی بین فواصل را محاسبه کرده و از این فاصله جدید به‌عنوان فواصل داده‌ها استفاده می‌شود.

[۳۵] AR(1) Auto-regressive
[۳۶] m-dependent, Teoplitz

شناسایی فواصل واقعی نیازمند اطلاعات اولیه می‌باشد و این یکی از مشکلات این نوع داده‌ها است. در ادامه نمونه‌هایی از این ساختارها را معرفی خواهیم نمود.

۶.۴.۱ همبستگی توانی فضایی

همان‌طور که گفته شد، ساختار خودبرگشت مرتبه اول برای داده‌هایی با فواصل برابر مناسب می‌باشد. یکی از ساختارهای جایگزین برای داده‌ها با فواصل نابرابر همبستگی توانی فضایی [۳۷] است. این ساختار با محاسبه اندازه اقلیدسی بین فواصل واقعی، رابطه‌ای معکوس را بین فواصل و همبستگی اعمال می‌کند. اگر اندازه اقلیدسی بین فواصل واقعی را با $\widehat{d_{jj'}}$ نشان دهیم آن‌گاه ساختار همبستگی و فرم ماتریسی توانی فضایی به‌صورت زیر خواهد بود:

$$corr(y_{ij}, y_{ij'}) = \alpha^{\widehat{d_{jj'}}}$$

با توجه به فرم‌های بالا مشخص است که در صورت برابری فواصل، ساختارهای خودبرگشت مرتبه اول و توانی‌فضایی هم‌ارز خواهند بود.

۷.۴.۱ همبستگی نمایی فضایی

هنگامی که کاهش همبستگی با زیاد شدن فواصل به‌اندازه‌ای باشد که ساختار توانی‌فضایی نیز قادر به‌نمایش آن نباشد از ساختار نمایی‌فضایی [۳۸] استفاده می‌شود. در این ساختار با افزایش فاصله اقلیدسی فواصل، همبستگی با نرخ نمایی کاهش می‌یابد. این ساختار که نسبت به ساختار توانی فضایی پیچیده‌تر و در عین حال منعطف‌تر است به‌صورت زیر نمایش داده می‌شود:

$$corr(y_{ij}, y_{ij'}) = exp\{-\frac{d_{jj'}}{\alpha}\}$$

۸.۴.۱ همبستگی گاوسی فضایی

ساختار گاوسی فضایی [۳۹] زمانی به‌کار می‌رود که حتی ساختار نمایی فضایی نیز در نمایش کاهش شدید همبستگی با افزایش فاصله ناتوان باشد. این ساختار معمولا به‌یکی از دو صورت زیر نمایش داده می‌شود:

$$corr(y_{ij}, y_{ij'}) = exp\{-\frac{d_{jj'}}{\alpha^2}\}$$

$$corr(y_{ij}, y_{ij'}) = exp\{-\frac{d_{jj'}}{\alpha}\}^2$$

---

[۳۷] Spatial Power
[۳۸] Spatial Eponential
[۳۹] Spatial Gaussian

۹.۴.۱ سایر همبستگی‌های فضایی

از جمله سایر ساختارهای وابستگی فضایی می‌توان به ساختارهای خطی[۴۰] درجه دوم[۴۱] کروی[۴۲] ، اشاره کرد. توابع این ساختارهای وابستگی به‌ترتیب عبارتند از

$$corr(y_{ij}, y_{ij'}) = 1 - (\frac{d_{jj'}}{\alpha})I(d_{jj'<\alpha})$$

$$corr(y_{ij}, y_{ij'}) = \frac{(\frac{d_{jj'}}{\alpha})^2}{1 + (\frac{d_{jj'}}{\alpha})^2}$$

$$corr(y_{ij}, y_{ij'}) = [1 - 1.5(\frac{d_{jj'}}{\alpha}) + 0.5(\frac{d_{jj'}}{\alpha})^2]I(d_{jj'<\alpha})$$

۱۰.۴.۱ همبستگی باقی‌مانده ترکیبی

ممکن است مواقعی پیش آید که به‌دلایل زیر هیچ‌کدام از ساختارهای معرفی شده برای داده‌ها مناسب نباشند: ۱) مدنظر بودن ساختاری خاص توسط تحلیل‌گر به‌علت تجربه ۲) اطلاعات اولیه ۳) اظهار نظر یک متخصص در این حالت با ترکیب ساختارهای معرفی شده، ساختارهای جدیدی به‌نام باقی‌مانده ترکیبی[۴۳] را معرفی می‌کنند. نمایش این ساختار به‌صورت زیر است:

$$corr(y_{ij}, y_{ij'}) = R_1 + R_2$$

به‌طوری‌که R_1 و R_2 می‌توانند هر یک از ساختارهای معرفی شده باشند. در پایان لازم به ذکر است که در این نرم افزارها،از شاخص‌های GIC و CIC به‌عنوان معیارهایی برای تشخیص ساختار همبستگی مناسب، تشخیص کارایی برآوردگرها و نیکویی برازش مدل استفاده می‌گردد. این شاخص‌ها تعمیمی از معیار AIC برای داده‌های هم‌بسته می‌باشد.

۵.۱ مدل‌های اتورگرسیو میانگین متحرک

در علم آمار و پردازش سیگنال مدل اتورگرسیو میانگین متحرک[۴۴](ARMA) که گاهی به آن مدل باکس–جنکینز[۴۵] نیز می‌گویند، مدلی است که معمولاً برای سنجش داده‌های سری زمانی مورد استفاده قرار می‌گیرد.

برای داده‌های سری زمانی به‌صورت X_t، مدل ARMA ابزاری برای مطالعه و پیش‌بینی مقادیر آتی چنین سری‌هایی است. این مدل شامل دو بخش اتورگرسیو[۴۶] به‌اختصار (AR) و میانگین متحرک[۴۷]

[۴۰] Spatial Linear
[۴۱] Spatial Rational Quadratic
[۴۲] Spatial Spherical
[۴۳] Hybrid Residual
[۴۴] Autoregressive moving average model
[۴۵] Box-Jenkins
[۴۶] Autoregressive
[۴۷] Moving Average

به‌اختصار (MA) است. بنابراین مدل ARMA را در ادبیات علمی به‌صورت $ARMA\,(p,q)$ نمایش می‌دهند. که در آن p مرتبه مدل AR و q مرتبه مدل MA است.

۱.۵.۱ مدل اتورگرسیو

مدل AR با مرتبه p به‌صورت زیر است:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t. \tag{۵.۱}$$

۲.۵.۱ مدل میانگین متحرک

مدل MA با مرتبه q به‌صورت زیر است:

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \tag{۶.۱}$$

در این مدل $\theta_1, \theta_2, ..., \theta_q$ پارامترهای مدل هستند. μ امید ریاضی X_t (اغلب برابر صفر در نظر گرفته می‌شود). ... $\varepsilon_t, \varepsilon_{t-i}$ خطای مدل هستند.

۳.۵.۱ مدل اتورگرسیو میانگین متحرک

مدل $ARMA$ مدلی با مرتبه p اتورگرسیو و q میانگین متحرک، که شامل دو مدل ذکر شده به‌صورت زیر است:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \tag{۷.۱}$$

مدل ARMA زمانی مناسب است که سیستم تابعی از شوک‌های مشاهده‌ناپذیر باشد. برای مثال قیمت سهام که علاوه بر شوک‌های اطلاعاتی در بازار تحت تاثیر شوک‌های رفتاری آحاد نیز هست.

۶.۱ آمار فضایی

اغلب در مطالعات محیطی با داده‌هایی سر و کار داریم که مستقل از یکدیگر نیستند و وابستگی خاصی بین آن‌ها وجود دارد که ناشی از مکان (موقعیت جغرافیایی) آن در فضای تحت مطالعه است به‌این‌گونه داده‌ها، داده‌های فضایی[۴۸] می‌گویند. به‌عنوان مثال‌هایی از این نوع داده‌ها می‌توان میزان شیوع یک بیماری در استان‌های مختلف، درصد خلوص طلا در نقاط مختلف یک معدن و عوامل موثر بر کیفیت آب‌های زیرزمینی در نقاط مختلف یک استان را نام برد.

[۴۸] Spatial data

دو ویژگی عمده داده‌های‌فضایی، نمایش هر داده با موقعیت آن در فضای مورد مطالعه و وابستگی فضایی آن‌هاست. معمولاً داده‌های‌فضایی جمع‌آوری شده از موقعیت‌های مجاور، وابستگی بیش‌تری دارند و با افزایش فاصله بین موقعیت داده‌ها، وابستگی کاهش می‌یابد.

اگرچه گسترش آمار فضایی عمدتاً توسط آماردانان صورت می‌گیرد، اما پیدایش و کاربرد عملی این شاخه از آمار به‌وسیله زمین‌شناسان و تحت عنوان زمین‌آمار[۴۹] صورت گرفت. نخستین بار به دنبال روند تکاملی ذخایر معدنی که قبل از سال ۱۹۶۰ آغاز شده بود، ماترون با انتشار مقاله‌ای در سال ۱۹۶۲ پایه‌های زمین‌آمار را بنا نهاد. زمین‌آمار به شاخه‌ای از علم آمار گفته می‌شود که مبتنی بر تحلیل متغیرهایی است که دارای ساختار فضایی بوده و با داده‌ها یا متغیرهای فضایی سر و کار دارد. در این شاخه، ناحیه جغرافیایی یک ناحیه چگال[۵۰] است، به این معنی که بین هر دو موقعیت مکانی می‌توان یک مکان دیگر نیز در نظر گرفت.

داده‌های‌فضایی در مطالعات متعدد علمی مانند هواشناسی، محیط زیست، همه‌گیرشناسی، زمین‌شناسی و اقیانوس‌شناسی در طول زمان نیز به‌یکدیگر وابسته‌اند. مشاهداتی که هم از نظر موقعیت فضایی و هم از نظر موقعیت زمانی وابسته باشند، داده‌های‌فضایی ـ زمانی[۵۱] نامیده می‌شوند و معمولاً از طریق تابع کوواریانس مدل‌بندی می‌شود. تحلیل این‌گونه داده‌ها مستلزم تعیین دو نوع ساختار وابستگی هستند: وابستگی فضایی و وابستگی زمانی.

در سال‌های اخیر همراه با توسعه زمینه‌های مختلف آمار از جمله آمار بیزی، آمار فضایی نیز شاهد رشد و توسعه چشم‌گیری بوده است و به‌مدد این پیشرفت‌ها آمار فضایی به‌عنوان وسیله‌ای کارآمد برای مطالعه داده‌های‌فضایی و فضایی ـ زمانی مطرح گردیده است. نخستین بار کیتانیدیس (۱۹۸۶) با اشاره به مسائل و دشواری‌های موجود در روش‌های کلاسیک تحلیل داده‌های‌فضایی، به‌کارگیری روش‌های بیزی را پیشنهاد نمود. اما از آن‌جاکه استفاده از رهیافت بیزی برای تحلیل داده‌های‌فضایی دشوار و نیازمند محاسبات پیچیده می‌باشد، در ابتدا چندان مورد توجه قرار نگرفت. اما با توسعه و گسترش آمار بیزی و ابداع روش‌های محاسبات بیزی پیشرفته نظیر روش‌های مونت‌کارلوی زنجیر مارکوفی[۵۲](MCMC)، تحلیل بیزی داده‌های‌فضایی نیز رشد و توسعه چشم‌گیری یافته است. مطالب این بخش با استفاده از مجموعه مقالات دومین کارگاه آموزشی نظریه مفصل و انواع وابستگی‌ها آبان ماه ۹۱ جمع‌آوری شده است.

در ادامه، به‌معرفی برخی از مفاهیم کلیدی آمار فضایی می‌پردازیم.

۱.۶.۱ داده‌های فضایی

داده‌های‌فضایی به داده‌های وابسته‌ای گفته می‌شود که وابستگی آن‌ها ناشی از موقعیت و مکان قرار گرفتن آن‌ها در فضای (جغرافیایی) مورد مطالعه است. داده‌های‌فضایی بر اساس این‌که موقعیت‌های

[۴۹] Geostatistics
[۵۰] Dense
[۵۱] Spatio-temporal data
[۵۲] Markov Chain Monte Carlo

فضایی و متغیرهای تصادفی مرتبط به‌چه صورت باشند، به سه گروه داده‌های زمین‌آماری[۵۳]، داده‌های مشبکه‌ای[۵۴] و الگوهای نقطه‌ای[۵۵] تقسیم می‌شوند. مطالب این بخش از محمدزاده (۱۳۹۲) استخراج شده‌اند.

داده‌های زمین‌آماری: این نوع داده‌ها در موقعیت‌های ثابت و مشخص در ناحیه‌ای چگال مشاهده می‌شوند. یعنی بین دو موقعیت مفروض، امکان وجود موقعیت‌های دیگر هم هست و متغیر مورد مطالعه ممکن است پیوسته یا گسسته باشد.

داده‌های مشبکه‌ای: این نوع داده‌ها مربوط به نواحی غیرچگال هستند، یعنی بین موقعیت‌های مشاهده شده امکان یافتن یک موقعیت جدید وجود ندارد. این مکان‌ها ممکن است منظم یا نامنظم باشند.

الگوی نقطه‌ای: در این نوع داده‌ها موقعیت مشاهده‌شده، خود یک متغیر تصادفی است. این داده‌ها به سه دسته به‌طور کامل تصادفی فضایی[۵۶] (CSR)، منظم[۵۷] و خوشه‌ای[۵۸] تقسیم شده و اقدام به مدل‌بندی آن‌ها می‌شود.

۲.۶.۱ مدل‌های آماری فضایی

تحلیل داده‌های‌فضایی نیازمند در نظر گرفتن یک مدل آماری است. معمولاً مدل‌بندی داده‌های‌فضایی با تعریف یک میدان تصادفی[۵۹] صورت می‌گیرد.

تعریف ۱.۶.۱ (میدان تصادفی). میدان تصادفی مجموعه‌ای از متغیرهای تصادفی مانند

$$Y(\cdot) = \{Y(\mathbf{s});\ \mathbf{s} \in D\}$$

است که درآن مجموعه اندیس‌گذار D یک زیرمجموعه از فضای اقلیدسی d بعدی، $d \geqslant 1$، از R^d است.

تعریف ۲.۶.۱ (توابع میانگین، واریانس و کوواریانس میدان تصادفی). برای میدان تصادفی $Y(\cdot)$، میانگین در موقعیت $\mathbf{s}$ و کوواریانس در موقعیت‌های $\mathbf{s}_1$ و $\mathbf{s}_2$ به‌ترتیب به‌صورت

$$\begin{aligned} \mu(\mathbf{s}) &= E(Y(\mathbf{s})), \quad \mathbf{s} \in D \\ C(\mathbf{s}_1, \mathbf{s}_2) &= Cov(Y(\mathbf{s}_1), Y(\mathbf{s}_2)) = E\left[(Y(\mathbf{s}_1) - \mu(\mathbf{s}_1))(Y(\mathbf{s}_2) - \mu(\mathbf{s}_2))\right],\ \mathbf{s}_1, \mathbf{s}_1 \in D \end{aligned}$$

تعریف می‌شوند. برای $\mathbf{s}_1 = \mathbf{s}_2 = \mathbf{s}$، واریانس میدان تصادفی $Y(\cdot)$ در مکان $\mathbf{s}$ به‌صورت

$$Var\left[Y(\mathbf{s})\right] = E\left[(Y(\mathbf{s}_1) - \mu(\mathbf{s}_1)\right]^2 = C(\mathbf{s}, \mathbf{s}) = C(\mathbf{s})$$

حاصل می‌شود.

---

[۵۳] Geostatistical data
[۵۴] Lattice data
[۵۵] Point patterns
[۵۶] Complete spatial randomness
[۵۷] Regular
[۵۸] Cluster
[۵۹] Random field

هر میدان تصادفی را می‌توان به‌صورت

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \mathbf{s} \in D$$

تجزیه کرد، که درآن $\mu(\cdot)$ تغییرات بزرگ‌مقیاس[۶۰] یا روند[۶۱] و $\delta(\cdot)$ تغییرات کوچک‌مقیاس[۶۲] یا فرآیند خطا نامیده می‌شود. عبارت روند ممکن است ناشی از تغییرپذیری بین موقعیت‌های مشاهده‌شده باشد، اما فرآیند خطا ممکن است از خطای اندازه‌گیری یا تغییرپذیری در درون موقعیت مشاهده‌شده ناشی شود. به‌طور معمول، تحلیل داده‌های‌فضایی بر اساس مشاهدات نمونه دشوار است، اما افزودن ویژگی‌هایی مانند مانایی[۶۳] و همسانگردی[۶۴] به ساختار همبستگی فضایی، مسأله را ساده‌تر می‌کند که در ادامه به‌معرفی آن‌ها می‌پردازیم. (امیدی و محمدزاده، ۱۳۹۱)

۳.۶.۱ توابع مفصل و آمار فضایی

برای پیشگویی در آمار فضایی میدان‌تصادفی $Z = \{Z(s); s \in D \subset \mathbb{R}^d\}$ در موقعیت مشخص s_0 بر اساس تحقق‌های میدانی تصادفی یعنی $z = (z(s_1), \ldots, z(s_n))$ در موقعیت‌های $s_n, \ldots, s_n$ است. پیشگویی فضایی معمولا با فرض گاوسی بودن میدان تصادفی انجام می‌شود. گاهی این فرض واقع‌گرایانه نیست و لازم است توزیع $Z = (Z(s_1), \ldots, Z(s_n))$ تعیین شود. با استفاده از توابع مفصل می‌توان توزیع توأم $Z(s)$ و $Z(s+h)$ را به‌صورت

$$K_h(F_Z(z_1), F_Z(z_2)) = P(Z(s) \leqslant z_1, Z(s+h) \leqslant z_2)$$

به‌دست آورد. با توجه به خواص وابستگی داده‌های‌فضایی تابع مفصل $K_h(\cdot,\cdot)$ معتبر است هرگاه وابستگی مثبت میان متغیرها را نتیجه دهد، به‌علاوه وقتی $\|h\|$ به بی‌نهایت میل کند تابع مفصل حاصل‌ضرب و موقعی که $\|h\|$ به صفر میل کند باید کران بالای فرشه‌ـ هافدینگ را نتیجه دهد، یعنی

$$\lim_{\|h\|\to\infty} K_h(u,v) = uv \quad , \quad \lim_{\|h\|\to 0} K_h(u,v) = M(u,v)$$

بدیهی است که تمام توابع مفصل ارائه شده در منابع لزوماً واجد شرایط یک مفصل فضای معتبر نیستند. در ادامه توابع مفصل فضایی معتبر معرفی می‌شوند. برخی از مفصل‌های فضایی معتبر موجود در منابع بر اساس تابع مفصل گاوسی و تبدیل‌های از توزیع نرمال چندمتغیره به‌دست آمده‌اند. مفصل گاوسی به‌صورت

$$K(u_1, \ldots, u_n) = \Phi_{0,\Lambda}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)\right) \qquad (۸.۱)$$

تعریف می‌شود. که در آن $\Phi_{0,\Lambda}$ نشان‌دهنده توزیع نرمال n متغیره با بردار میانگین $\circ$ و ماتریس واریانس Λ است. از آنجاکه این مفصل برای همبستگی صفر تابع مفصل استقلال و برای همبستگی یک کران

[۶۰] Large scale variation
[۶۱] Trend
[۶۲] Small scale variation
[۶۳] Stationary
[۶۴] Isotropic

بالای فرشه‌ـ هافدینگ را نتیجه می‌دهد، یک تابع مفصل فضای معتبر است. اکنون تبدیل‌هایی از توزیع نرمال چندمتغیره که منجربه ساخت مفصل‌های فضایی معتبر می‌شوند، مورد بررسی قرار می‌گیرند.

مفصل خی ـ دو غیرمرکزی: باردوسی (۲۰۰۶) با استفاده از تبدیل $g(Y)=Y^2$ ، که در آن Y دارای توزیع نرمال چندمتغیره است، مفصل خی‌ـ دوی غیرمرکزی را با تابع مفصل چگال به‌صورت

$$k(u,v)=K_2\frac{e^{\frac{-2\sqrt{xy}r+(x+y)r^2}{2(r^2-1)}}\left(1+e^{\frac{2r\sqrt{xy}}{r^2-1}}\right)}{\sqrt{4\pi xy}}$$

به‌دست آورد. این مفصل برای همبستگی صفر استقلال و برای همبستگی کران بالای فرشه‌ـ هافدینگ را نتیجه می‌دهد.

توزیع V ـ تبدیل یافته تابع مفصل نرمال چندمتغیره: این مفصل توسط باردوسی و لای (۲۰۰۸) و با استفاده از تبدیل

$$X_j=g(Y_j)=\begin{cases}k(Y_j-m) & Y_j\geqslant m\\ m-Y_j & Y_j<m\end{cases}$$

معرفی شد، که در آن توزیع‌های یک و چندمتغیره X به‌ترتیب به‌صورت

$$H_1(x)=P(X\leqslant x)=\Phi\left(\frac{k}{m}+1\right)-\Phi(-x-m)$$

و

$$\begin{aligned}H_n(x_1,\ldots,x_n)&=P(X_1\leqslant x_1,\ldots,X_n\leqslant x_n)\\&=\Phi\left(\frac{k}{m}+1\right)-\Phi(-x-m)\\&=\sum_{i=0}^{2^n-1}(-1)^i\Phi(\xi_i+m)\end{aligned}$$

به‌دست می‌آیند،که در آن$\left\{b(-1)^{i1}x_1,\ldots,b(-1)^{in}x_n\right\}$ $\xi_i^T=$ و

$$b(x)=\begin{cases}-1 & x<0\\ \frac{1}{k} & Y_j<m\end{cases}$$

تابع مفصل فضایی معتبر متناظر با این تبدیل به‌صورت

$$K(u_1,\ldots,u_n)=\sum_{i=0}^{2^n-1}(-1)^i\Phi\left(\xi_i^T+m\right)$$

حاصل می‌شود،که در آن $\left\{b(-1)^{i_1}H_1^{-1}(u_1),\ldots,b(-1)^{i_n}H_n^{-1}(u_n)\right\}$ $\xi_i^T=$.

مفصل ترکیبی:گرایلر و پیسما (2011) با استفاده از ترکیب خطی از توابع مفصل، تابع مفصل فضایی

معتبر را براساس فواصل بین موقعیت‌ها به‌صورت زیر تعریف کردند.

$$K(u,v)=\begin{cases}\lambda_1 M(u,v)+(1-\lambda_1)K_{1,h}(u,v) & 0\leqslant h\leqslant h_1\\ \vdots & \\ \lambda_i M(u,v)+(1-\lambda_i)K_{1,h}(u,v) & h_{i-1}\leqslant h\leqslant h_i\\ \vdots & \\ \lambda_l M(u,v)+(1-\lambda_l)K_{1,h}(u,v) & h_l-1\leqslant h\leqslant h_l\end{cases}$$

این تابع ویژگی تابع مفصل فضایی معتبر را براساس فاصله‌ها حفظ می‌کند، اما انتخاب خانواده مفصل مناسب در هربازه بایستی به‌گونه‌ای باشد که همواره همبستگی مثبت را نتیجه دهد و با افزایش فاصلهٔ موقعیت‌ها همبستگی کاهش یابد. به‌عنوان مثال تابع مفصل گامبل ـ بارنت که به‌ازای $\theta\in(0,1]$ به‌صورت $K(u,v)=uv\exp(-\theta\log u\log v)$ است، همواره وابستگی منفی را نتیجه می‌دهد و نمی‌تواند یک تابع مفصل فضایی معتبر باشد.

۷.۱ نمونه‌گیری نقاط مهم

در آمار ، نمونه‌گیری نقاط مهم یک روش کلی برای برآورد خواص یک توزیع خاص است، در حالی‌که تنها نمونه‌هایی از توزیع مختلف از توزیع مورد نظر تولید می‌شود. فرض کنید که مسئله ما یافتن $\mu=E(f(X))=\int_D f(x)p(x)dx$ باشد که در آن p یک تابع چگالی احتمال بر روی $D\subseteq\mathbb{R}^d$ می‌باشد و f انتگرال‌ده می‌باشد. $p(x)=0$ را به‌ازای $x\notin D$ در نظر می‌گیریم. اگر qیک تابع چگالی احتمال مثبت بر روی $\mathbb{R}^d$ باشد، آن‌گاه داریم

$$\mu=\int_D f(x)p(x)dx=\int_D\frac{f(x)p(x)}{q(x)}q(x)dx=\mathbb{E}_q\left(\frac{f(X)p(X)}{q(X)}\right). \qquad (۹.۱)$$

که در آن $E_q(\cdot)$ نشان‌دهنده امید ریاضی برای $X\sim q$ می‌باشد. زمانی که $X\sim q$ باشد برای واریانس، کوواریانس و همبستگی از $Var_q(\cdot)$ و $Cov_q(\cdot,\cdot)$ و $Corr_q(\cdot,\cdot)$ استفاده می‌کنیم. سپس هدف اصلی ما یافتن $E_p(f(X))$ می‌باشد. با اعمال یک تعدیل ضربی بر f ، نمونه‌گیری از q به جای p را تنظیم می‌کنیم. تابع تعدیل $p(x)/q(x)$ نسبت درستنمایی نامیده می‌شود. توزیع q توزیع نقاط مهم و p توزیع اسمی است.

ضرورتی ندارد که توزیع نقاط مهم q همه جا مثبت باشد. کافی است $q(x)>0$ هرگاه

$$f(x)p(x)\neq 0$$

باشد. بنابراین، برای $Q=\{x\,|q(x)>0\}$ داریم $x\in Q$ هرگاه داشته باشیم $f(x)p(x)\neq 0$ بنابراین اگر $x\in D\cap Q^c$ باشد، با توجه به $f(x)=0$، در حالی که اگر $x\in Q\cap D^c$ باشد خواهیم

داشت $p(x)=0$. حال داریم

$$\begin{aligned} E_q\left(\frac{f(X)p(X)}{q(X)}\right) &= \int_Q \frac{f(x)p(x)}{q(x)}q(x)\,dx = \int_Q f(x)p(x)\,dx \\ &= \int_Q f(x)p(x)\,dx + \int_{Q\cap D^c} f(x)p(x)\,dx - \int_{D\cap Q^c} f(x)p(x)\,dx \\ &= \int_D f(x)p(x)\,dx = \mu. \end{aligned} \tag{۱۰.۱}$$

جالب خواهد بود که بدانیم برای x با $q(x)=0$ در مخرج کسر چه اتفاقی می‌افتد. پاسخ این است که چنین نقاطی $x\in Q$ وجود ندارد و ما در حین نمونه‌گیری $X\in q$ با چنین نقاطی روبرو نخواهیم شد. در ادامه مثال‌هایی مشاهده خواهیم کرد که در آن $q(x)$ های نزدیک به 0 منجر به مشکلات بزرگی می‌شوند، اما $p(x)=0$ مشکل‌ساز نیست اگر $f(x)p(x)=0$ نیز برقرار باشد.

زمانی که می‌خواهیم q برای توابع مختلف f_j عمل کند آن‌گاه نیاز داریم که $q(x)>0$ هرگاه $p(x)>0$ باشد کافی خواهد بود، و به ما اجازه افزودن توابع جدید f_j را به لیست بعد از انجام نمونه‌گیری خواهد داد.

برآورد نمونه‌گیری نقاط مهم $\mu - E_p(f(X))$ برابر است با

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)p(X_i)}{q(X_i)}, \quad X_i\sim q. \tag{۱۱.۱}$$

برای استفاده از (۱۱.۱) ما باید بتوانیم fp/q را محاسبه کنیم. با فرض اینکه می‌توانیم f را محاسبه کنیم، این برآورد نیازمند این است که ما بتوانیم $p(x)/q(x)$ را در هر x ای که می‌خواهیم نمونه‌گیری کنیم محاسبه کنیم. زمانی که p یا q دارای یک ثابت نرمال‌سازی باشد، آن‌گاه ما به یک برآورد نسبت متوصل خواهیم شد (۱۰.۱). در حال حاضر، فرض می‌کنیم که p/q قابل‌محاسبه باشد و واریانس $\hat{\mu}_q$ را مطالعه می‌کنیم.

قضیه ۱.۷.۱. $\hat{\mu}_q$ را مطابق (۱۱.۱) در نظر بگیرید که در آن $\mu=\int_D f(x)p(x)\,dx$ و $q(x)>0$. اگر $f(x)p(x)\neq 0$ ، آن‌گاه $E_q(\hat{\mu}_q)$ و $Var_q(\hat{\mu}_q)=\sigma_q^2/n$ که در آن اگر

$$\begin{aligned} \sigma_q^2 &= \int_D \frac{(f(x)p(x))^2}{q(x)}dx - \mu^2 \\ &= \int_D \frac{(f(x)p(x)-\mu q(x))^2}{q(x)}dx. \end{aligned} \tag{۱۲.۱}$$

اثبات. معادله $E_q(\hat{\mu}_q)=\mu$ به‌طور مستقیم از (۱۰.۱) نتیجه‌گیری می‌شود. با استفاده از $Q=\{x\,|q(x)>0\}$ ، می‌توان نوشت

$$\begin{aligned} \mathrm{Var}_q(\hat{\mu}_q) &= \frac{1}{n}\left[\int_Q\left(\frac{f(x)p(x)}{q(x)}\right)^2 q(x)\,dx - \mu^2\right] \\ &= \frac{1}{n}\left[\int_D\left(\frac{f(x)p(x)}{q(x)}\right)^2 q(x)\,dx - \mu^2\right] \end{aligned} \tag{۱۳.۱}$$

زیرا سهم x در $D \cap Q^c$ و $Q \cap D^c$ در انتگرال صفر است. تغییرات ساده دو شکل از معادله (۱۲.۱) را ارائه می دهد.

برای ایجاد یک بازه اطمینان برای μ ، باید σ_q^2 را برآورد کنیم. از عبارت دوم در معادله (۱۲.۱) نتیجه می‌گیریم که

$$\sigma_q^2 = E_q\left(\left(f\left(X\right)p\left(X\right) - \mu q\left(X\right)\right)^2 / q(X)^2\right) \tag{۱۴.۱}$$

به این دلیل که x_i از q نمونه‌گیری شده است، برآورد واریانس طبیعی برابر است با

$$\hat{\sigma}_q^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f\left(x_i\right)p\left(x_i\right)}{q\left(x_i\right)} - \hat{\mu}_q\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\omega_i f\left(x_i\right) - \hat{\mu}_q\right)^2 \tag{۱۵.۱}$$

که در آن $\omega_i = p\left(x_i\right)/q\left(x_i\right)$ می‌باشد.

دومین عبارت انتگرالی در (۱۲.۱) نشان می‌دهد که نمونه‌گیری نقاط مهم چگونه می‌تواند موفق یا ناموفق باشد. صورت کسر در انتگرالده در سمت راست (۱۲.۱) کوچک است زمانی که $f\left(x\right)p\left(x\right) - \mu q\left(x\right)$ نزدیک به صفر باشد، یعنی زمانی که $q\left(x\right)$ تقریباً متناسب با $f\left(x\right)p\left(x\right)$ باشد. از مخرج کسر می‌بینیم که ناحیه‌هایی با مقادیر کوچک $q\left(x\right)$ ،هر عدم تناسبی را که در صورت کسر ظاهر می‌شود شدیداً بزرگ نمایی می‌کند.

فرض کنید که $f\left(x\right) \geqslant 0$ برای تمام x ها برقرار باشد، و برای رد کردن یک حالت بدیهی، فرض کنید که $\mu > 0$ نیز برقرار باشد. آن‌گاه $q_{opt}\left(x\right) \equiv f\left(x\right)p\left(x\right)/\mu$ یک چگالی احتمال است و داریم $\sigma_{q_{opt}}^2 = 0$.

این بهینه است اما قابل استفاده نیست زیرا $\hat{\mu}_{q_{opt}}$ برابر می‌شود با یک میانگین از $f\left(x_i\right)p\left(x_i\right)/q\left(x_i\right) =$ μ و این بدین معنی است که ما می‌توانیم μ را مستقیماً از f،pو qو بدون نمونه‌گیری محاسبه کنیم. به طور مشابه برای $f\left(x\right) \leqslant 0$ با $\mu < 0$ یک واریانس صفر برای $q = -fp/\mu$ به دست می آید.

اگرچه چگالی‌های نمونه‌گیری نقاط مهم با واریانس صفر قابل استفاده نیستند، اما یک دید مناسب برای طراحی یک طرح و رویه مناسب برای نمونه‌گیری نقاط مهم فراهم می‌کند. فرض کنید که $f\left(x\right) \geqslant 0$ باشد. شاید برای q مناسب باشد که اسپایک[۶۵] در آن نقاطی مشابه با f و یا p رخ بدهد اما بهتر است که در نقاطی مشابه با fp رخ بدهد. علاوه بر این، بهترین q،q ای است که متناسب با fp باشد ونه $\sqrt{fp}$ و یا f^2p یا ترکیبات دیگری از آن.

انتخاب یک توزیع خوب برای نمونه‌گیری از نقاط مهم نیازمند حدس زدن علمی و احتمالاً پژوهش عددی می‌باشد. در بسیاری از کاربردها یک دانش محیطی درباره محل اسپایک‌ها وجود دارد. در یک زمینه اقتصادی ممکن است بدانیم که کدام نوسان سهام منجر به یک گزینه برای رسیدن به ارزش بیشینه می‌شود. برای یک سیستم صف‌بندی دانستن این نکته آسان است که چه ترکیبی از ورودی‌ها منجر به اضافه‌بار سیستم خواهد شد.

به‌طورکلی، چگالی q^* که σ_q^2 را به حداقل می‌رساند، متناسب است با $|f\left(x\right)|p\left(x\right)$ (کان و مارشال، ۱۹۵۳)، خارج از حالت‌های بدیهی که در آن $\int |f\left(x\right)|p\left(x\right)dx = 0$ می‌باشد. در این چگالی بهینه در

[۶۵] Spike

مسائلی که fp می‌تواند برای برخی x ها مثبت و برای برخی x های دیگر منفی باشد $\sigma^2_{q^*} = 0$ برقرار نیست. زمانی که fp هم مقادیر مثبت و هم مقادیر منفی را اختیار می‌کند، همچنان یک روش واریانس صفر وجود دارد، اما نیازمند نمونه‌گیری در دو نقطه است.

برای اثبات اینکه $q^*(x) = |f(x)|p(x)/E_p(|f(X)|)$ بهینه است، q را هر چگالی‌ای قرار دهید که مثبت باشد هرگاه داشته باشیم $fp \neq 0$. آن‌گاه داریم

$$\begin{aligned}\mu^2 + \sigma^2_{q^*} &= \int \frac{f(x)^2 p(x)^2}{q^*(x)} dx = \int \frac{f(x)^2 p(x)^2}{|f(x)|p(x)/E_p(|f(X)|)} dx \\ &= E_p(|f(X)|)^2 = E_q(|f(x)|p(x)/q(X))^2 \\ &\leqslant E_q(f(x)^2 p(x)^2/q(X)^2) = \mu^2 + \sigma^2_q.\end{aligned}$$

ما می‌توانیم از نسبت درستنمایی $\omega(x) = p(x)/q(x)$ به عنوان راهی برای فهمیدن این‌که کدام چگالی‌های نمونه‌گیری نقاط مهم خوب و کدام‌یک بد هستند استفاده کنیم. عبارت اول در σ_{q^2} برابر است با $\int f(x)^2 p(x)^2/q(x) dx$.

دقت کنید که می‌توان این عبارت را به‌صورت $E_p(f(X)^2\omega(X)) = E_q(f(X)^2\omega(X)^2)$ بنویسیم. ظهور q در مخرج ω بدین معنی است که چگالی‌های نقاط مهم دم‌سبک[۶۶] q خطرناک هستند. اگر باهوش یا خوش‌شانس باشیم، آن‌گاه f می‌تواند در محلی که موردنیاز است کوچک باشد تا مخرج کوچک را جبران کند. اما ما اغلب نیاز داریم که نمونه مشابه با انتگرال‌ده‌های چندگانه f را استفاده کنیم، و بنا بر قاعده q باید دم‌هایی حداقل به سنگینی p داشته باشد.

۸.۱ برخی مفاهیم و تعاریف مورد نیاز

این تعاریف در بخش‌های مختلف پایان نامه مورد استفاده می‌باشند.

تعریف ۱.۸.۱. (داده‌های طولی[۶۷]) مطالعه طولی کاربرد فراوانی در علوم پزشکی و اجتماعی دارد. با توجه به انجام اندازه‌گیری‌های مکرر در این‌گونه مطالعات، شرط استقلال بین مشاهدات برقرار نبوده و لذا باید مدل‌های مناسب انتخاب شوند. داده‌های طولی به داده‌هایی از مطالعات مشاهده‌ای گفته می‌شود که درآن گروه مورد بررسی، در طول زمان معینی، مورد مطالعه قرار می‌گیرد.

تعریف ۲.۸.۱. (داده‌های مشاهده‌ای[۶۸]) در آمار و همه‌گیرشناسی، داده‌های مشاهده‌ای داده‌های حاصل از مشاهدهٔ پدیده‌ها به‌منظور یافتن روابط علّی میان متغیرها هستند. در این نوع مطالعه، پژوهشگر در مجموعهٔ تحت بررسی مداخله‌ای نمی‌کند و صرفاً به مشاهدهٔ پدیده‌ها می‌پردازد. مثالی از این نوع داده‌ها، داده‌های حاصل از بررسی تاثیر سیگار کشیدن بر سرطان ریه در مردان است.

تعریف ۳.۸.۱. (داده‌های رسته‌ای[۶۹]) داده‌های رسته‌ای در آمار به داده‌هایی گفته می‌شود که مقادیر مشاهده شده به‌یک دسته تعلق داشته باشد. مثال‌هایی از این داده‌ها گروه خونی افراد، استان محل

[۶۶] light-tailed
[۶۷] Longitudinal data
[۶۸] Observational data
[۶۹] Categorical data

زندگی افراد یا انواع سنگ (آذرین، رسوبی، یا دگرگونی) است. یکی از روش‌های تحلیل این نوع داده، تبدیل یک متغیر با k دسته به k متغیر دودویی به‌صورتی است که هر متغیر بودن در هر دسته را نشان دهد.

تعریف ۴.۸.۱. (سانسور[۷۰]) زمان بقای یك عنصر (فرد) را سانسور شده گویند هرگاه زمان بقای آن عنصر (فرد) تا پایان دوره‌ی مطالعه مشاهده نشود. به‌عبارت دیگر زمان بقای یك عنصر را سانسور شده گویند هرگاه نقطه پایان پیشامد مورد نظر مشاهده نشود. بنابراین داده سانسور شده یک داده‌ی ناقص است.

معمولاً دلایل عمده سانسور عبارت از پایان مطالعه، عدم امکان پیگیری و خارج شدن عنصر (فرد) از مطالعه می‌باشند. سانسور را می‌توان به‌صورت سانسور راست، سانسور چپ، سانسور فاصله‌ای و غیره تقسیم‌بندی نمود.

تعریف ۵.۸.۱. (توزیع نرمال چندمتغیره[۷۱]) گوییم بردار تصادفی p بعدی x دارای توزیع نرمال p متغیره با بردار میانگین μ و ماتریس واریانس ـ کوواریانس معین مثبت Σ است.

می‌نویسیم $x \sim MVN(\mu, \Sigma)$، اگر تابع چگالی احتمال آن به‌صورت زیر باشد

$$f(x) = (2\pi)^{-p/2} \left|\sum\right|^{-\frac{1}{2}} exp\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\}. \qquad (۱۶.۱)$$

اگر $x \sim MVN(\mu, \Sigma)$، آن‌گاه تابع مشخصه x عبارت‌است از

$$\phi_x(t) = exp(it^T\mu - \frac{1}{2}t^T\Sigma t)$$

همچنین $E(x) = \mu$ و $cov(x) = \Sigma$.

تعریف ۶.۸.۱. (فرآیند تصادفی[۷۲]) مجموعه متغیرهای تصادفی $\{X_t : t \in T\}$ را فرآیند تصادفی با مجموعه اندیس‌گذار T و فضای حالت E می‌گوییم. اگر $x \in E$ و $X_t = x$، می‌گوییم فرآیند در زمان (یا در مرحله) t در مجموعه x قرار دارد. اگر T مجموعه‌ای شمارا باشد، فرآیند را زمان گسسته و اگر مجموعه‌هایی به‌صورت $(0, \infty)$ یا $(-\infty, \infty)$ باشد آن را زمان پیوسته می‌گوییم. برای هر w از فضای نمونه‌ای مجموعه $\{X_t(w) : t \in T\}$ را که زیر مجموعه‌ای از E است، تحقق یا مسیر نمونه‌ای فرآیند می‌گوییم.

تعریف ۷.۸.۱. (معیار اطلاع آکائیک[۷۳](AIC)) معیاری برای سنجش نیکویی برازش است. این معیار بر اساس مفهوم انتروپی بنا شده است و نشان می‌دهد که استفاده از یک مدل آماری به چه میزان باعث از دست رفتن اطلاعات می‌شود. به‌عبارت دیگر، این معیار تعادلی میان دقت مدل و پیچیدگی آن برقرار می‌کند. این معیار توسط هیروتسوگو آکائیکه برای انتخاب بهترین مدل‌آماری پیشنهاد شد.

---

[۷۰] Censor
[۷۱] Multivariate Normal Distribution
[۷۲] accidental process
[۷۳] Akaike information criterion

با توجه به داده‌ها، چند مدل رقیب ممکن است با توجه به مقدار AIC رتبه بندی شوند و مدل دارای کمترین AIC بهترین است. از مقدار AIC می‌توان استنباط نمود که به‌عنوان مثال سه مدل بهتر وضعیت نسبتاً یکسانی دارند و بقیه مدل‌ها به‌مراتب بدتر هستند، اما معیاری برای انتخاب مقدار آستانه‌ای برای AIC که بتوان مدلی را به‌واسطه داشتن AIC بزرگتر از این مقدار رد کرد وجود ندارد.

تعریف ۸.۸.۱. (زنجیره مارکوف[۷۴]) زنجیره مارکف که به‌افتخار آندری مارکوف ریاضی‌دان اهل روسیه این‌گونه نام‌گذاری شده یک سیستم ریاضی است که در آن انتقال از یک حالت به حالت دیگر صورت می‌گیرد که البته تعداد این حالات قابل شمارش است. زنجیره مارکف یک فرایند تصادفی بدون حافظه است بدین معنی که توزیع احتمال شرطی حالت بعد تنها به‌حالت فعلی بستگی دارد و به‌وقایع قبل از آن وابسته نیست. این نوع بدون حافظه بودن خاصیت مارکف نام دارد. زنجیره مارکف در مدل‌سازی دنیای واقعی کاربردهای زیادی دارد.

یک زنجیره‌مارکف دنباله‌ای از متغیرهای تصادفی $X_1, X_2, \ldots$ است که دارای خاصیت مارکف هستند یعنی:

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n). \quad (۱۷.۱)$$

مقادیر ممکن برای X_i مجموعه قابل شمارشی را می‌سازند که فضای حالت نام دارد.

تعریف ۹.۸.۱. (تابع بسل[۷۵]) توابع بسل، اولین بار توسط دانیل برنولی تعریف شدند و سپس فردریش بسل فرم عمومی آن را بررسی نمود. توابع بسل جواب‌های معادله دیفرانسیل زیر می‌باشند

$$x^2 \frac{d^2y}{dx^2} + x \frac{dy}{dx} + (x^2 - \alpha^2)y = 0$$

معادله بسل از معادلات قابل حل با سری‌هاست، و دارای نقطه تکین منظم است. نقطه $x = 0$ تنها نقطه غیرعادی معادله فوق است. جواب‌های معادله به توابع بسل معروفند. در معادلهٔ بالا α یک عدد حقیقی یا یک عدد مختلط

تعریف ۱۰.۸.۱. (تابع نرم[۷۶]) در ریاضی و رشته‌های مربوط به‌آن نرم در مواردی استفاده دارد که عناصر به مقادیر مثبت محدود باشند. تابع حقیقی $\|\cdot\|$ تعریف شده بر فضای برداری X را نُرم نامیم اگر در سه خاصیت زیر صدق کند:

۱. به‌ازای هر $x \in X$، $\|x\| \geqslant 0$، $\|x\| = 0$ اگر و فقط اگر $x = 0$

۲. به‌ازای هر x و $\alpha \in \mathbb{R}$ ، $\|\alpha x\| = |\alpha| \, \|x\|$

۳. به‌ازای هر x و $y \in X$، $\|x + y\| \leqslant \|x\| + \|y\|$

تعریف ۱۱.۸.۱. (گرادیان[۷۷]) در حسابان بردار گرادیان یک میدان نرده‌ای، میدانی برداری است که مؤلفه‌های آن نرخ تغییر میدان نخستین را در جهت‌های مختلف نشان می‌دهد. جهت خود میدان

[۷۴] Markovian process
[۷۵] Bessel function
[۷۶] Soft Function
[۷۷] Gradient

برداری گرادیان جهت بیشینهٔ تغییرات است. به تعبیر دیگر برداری که اندازه و جهت حداکثر نرخ فضائی تغییر یک کمیت عددی را نمایش می دهد، گرادیان آن کمیت عددی تعریف می کنیم.

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right)$$

در حالت خاص برای اسکالر $f(x, y, z)$، گرادیان f در دستگاه کارتزین به‌صورت زیر نوشته می‌شود:

$$\operatorname{grad} f = \frac{\partial f}{\partial x} i + \frac{\partial f}{\partial y} j + \frac{\partial f}{\partial z} k = \nabla f$$

فصل ۲

معرفی مدل رگرسیون حاشیه‌ای مفصل گاوسی

۱.۲ مقدمه

اکثر مدل‌بندی‌های آماری بر اساس استقلال بین داده‌ها انجام می‌پذیرد. با این وجود در بسیاری از موارد متغیرهای مورد مطالعه به نوعی به یکدیگر وابسته‌اند. ساختارهای متنوعی برای توصیف و تحلیل چنین وابستگی‌هایی معرفی شده‌اند. مدل‌های ساخته شده برمبنای مفاصل از مهم‌ترین و پرکاربردترین این مدل‌ها می‌باشد. مفاصل که اولین بار توسط اسکالر (۱۹۵۹) معرفی شده است، ساختارهای وابستگی مورد مطالعه را از توزیع‌های حاشیه‌ای منفک می‌سازد. از یک نقطه نظر مفصل‌ها توابعی هستند که توابع توزیع چندمتغیره را به توابع توزیع حاشیه‌ای آن‌ها پیوند می‌دهند و از دیدگاه دیگر، مفصل‌ها را می‌توان به‌صورت توابع توزیع چندمتغیره‌ای تعریف کرد که توابع توزیع حاشیه‌ای آن‌ها به‌صورت یکنواخت روی فاصله (۱و۰) توزیع شده است. چنانچه توزیع‌های حاشیه‌ای پیوسته باشند تابع مفصل یکتاست. به علاوه تابع مفصل، تحت تبدیلات اکیداً صعودی پایا می‌باشد. در اقتصاد این توابع به‌خوبی توانسته‌اند وابستگی متغیرهای ریسک را مدل‌سازی نمایند و از این منظر مورد توجه اقتصاددانان می‌باشند. جو (۱۹۷۷) ماری و کوتز (۲۰۰۰) و نلسن (۲۰۰۶) شرح جامعی از توابع مفصل و خصوصیات آن‌ها ارائه داده‌اند. به‌دلیل ویژگی‌های قابل توجه، توابع مفصل در مسائل مالی مورد استفاده قرار می‌گیرند تا ساختار وابستگی بین سرمایه‌ها را مدل‌بندی کنند و به‌عنوان ابزاری توانا در تحلیل سری‌های چندمتغیره به‌کار می‌روند. این توابع توسط امبریچ (۱۹۹۹) برای اولین بار در کاربردهای مالی مورد استفاده قرار گرفته‌اند. چون توابع

مفصل، توزیع‌های حاشیه‌ای را از ساختار وابستگی جدا می‌سازند، تابع مفصل مناسب برای یک کاربرد خاص تابعی است که به بهترین نحو ممکن همبستگی بین داده‌ها را نشان می‌دهد.

هدف از به‌کارگیری تابع مفصل برطرف کردن محدودیت‌های قابل توجه در استفاده از توزیع‌های چند متغیره کلاسیک (مانند نرمال، لگ‌نرمال، گامبل و ...) در مدل‌سازی همبستگی متغیرهای وابسته بوده که مهم‌ترین این محدودیت‌ها لزوم مشخص بودن توابع توزیع حاشیه‌ای و یکسان بودن آن‌ها است. در به‌خدمت گرفتن تابع مفصل الزامی در مشخص کردن توابع توزیع حاشیه‌ای و برآورد پارامتر آن‌ها وجود ندارد، به‌گونه‌ای که برای برآورد پارامتر توابع مفصل می‌توان از توزیع‌های تجربی متغیرهای حاشیه‌ای استفاده کرد.

یک رده مهم از مفاصل به‌نام مفصل‌های گاوسی به‌دلیل انعطاف‌پذیری و سادگی محاسبات مورد توجه می‌باشند. به‌نظر می‌رسد این انتخاب مزیت‌های زیادی داشته باشد زیرا به‌طور طبیعی چند ویژگی مشهور توزیع نرمال چندمتغیره را به‌ارث برده است به‌عنوان مثال می‌توانید به نیکولولوپلوس و همکاران (۲۰۱۱) مراجعه کنید. دراین فصل به‌بررسی و طبقه‌بندی مدل‌های مفصل گاوسی برای تحلیل رگرسیون حاشیه‌ای مشاهدات وابسته غیرنرمال می‌پردازیم. این کلاس یک مدل طبیعی رگرسیون‌خطی سنتی با خطاهای همبستگی نرمال را فراهم می‌کند. هر نوع از پاسخ‌های پیوسته، گسسته و قطعی مجاز است. وابستگی به‌راحتی با توجه به خطاهای نرمال چندمتغیره مدل‌سازی می‌شود. نتیجه‌گیری از طریق رویکرد احتمالی انجام می‌شود. در حالی که عملکرد احتمالی در یک شکل بسته برای پاسخ‌های پیوسته در دسترس است، در مقادیر عددی غیرپیوسته استفاده می‌شود. در بخش دوم و سوم ضمن توصیف مسئله، مدل‌های رگرسیون حاشیه‌ای مفصل‌های گاوسی را معرفی و مثال‌هایی برای بیان کاربردهای مفصل ارائه شده است. در بخش چهارم مشخصه‌های مدل را بررسی و ویژگی‌های وابستگی را تجزیه و تحلیل می‌کنیم.

مدل‌های رگرسیون حاشیه‌ای برای پاسخ‌های غیرنرمال هم‌بسته، معمولاً توسط رویکرد محبوب معادلات برآورد کلی، لیانگ و زیگر (۱۹۸۶) متناسب می‌شوند.

علی‌رغم مزیت‌های مختلف نظری و عملی، تجزیه و تحلیل درستنمایی، مدل‌های غیرنرمال رگرسیون حاشیه‌ای، گستردگی کمتری دارند برای این مورد به‌بررسی دیگل و همکاران (۲۰۰۲) مراجعه کنید. علت اصلی، دشواری در شناسایی طبقات کلی، توزیع‌های چندمتغیره برای پاسخ‌های رسته‌ای و گسسته است. با این وجود، تجزیه و تحلیل درستنمایی مدل‌های حاشیه‌ای مدنظر نویسندگان بسیاری بوده است که دراین زمنیه می‌توانید به‌بررسی مولنبرگز و وربک (۲۰۰۵) مراجعه کنید.

سونگ (۲۰۰۰) در مفصل‌های گاوسی چهارچوب کلی انعطاف‌پذیری برای مدل‌سازی پاسخ‌های وابسته به‌هر نوع را فراهم کرد. مفصل گاوسی، سادگی تفسیر در مدل‌سازی حاشیه‌ای را با انعطاف‌پذیری در مشخصه ساختار وابسته ترکیب نمود. با وجود انجام این کار، رگرسیون مفصل گاوسی برای پاسخ‌های وابسته غیرمستمر کاربرد محدودی دارد زیرا تابع درستنمایی به تقریب انتگرال‌هایی با ابعاد بزرگ نیاز دارد.

به‌طور خاص، نشان می‌دهیم با پارامترسازی مناسب ماتریس همبستگی مفصل گاوسی، ایده‌های سونگ (۲۰۰۰) را می‌توان بر تحلیل سری‌های زمانی و مشاهدات با همبستگی فضایی اعمال نمود. ما از

طریق حداکثر درستنمایی مدل برازش را پیشنهاد کردیم. در حالت پیوسته، تابع احتمال به‌شکل بسته خواهد بود. در غیراین‌صورت، در موارد رسته‌ای، تقریب عددی درستنمایی مورد نیاز خواهد بود.

روش‌های بحث شده در این پایان‌نامه در R بسته $gcmr$ (رگرسیون حاشیه‌ای مفصل گاوسی[۱]) پیاده‌سازی شده که در وب سایت $http://cran.r-project.org/web/packages/gcmr$ نیز موجود است.

یکی از اهداف این پایان‌نامه این است که نشان دهد مدل‌های رگرسیون مفصلی گاوسی را می‌توان برای تحلیل فرایندهایی با ابعاد بزرگتر نیز در سری‌های‌زمانی و آمارفضایی مورد استفاده قرار داد.

۲.۲ توصیف مسئله

فرض کنید $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^T$ بردارِ متغیرهای پیوسته، گسسته، یا رسته‌ای تصادفی وابسته باشد که مقدار مشاهده‌شده آن‌را با $\mathbf{y} = (y_1, \ldots, y_n)^T$ نشان می‌دهیم. ممکن است وابستگی در فرم‌های مختلفی دیده شود، به‌عنوان مثال اندازه‌های تکراری در یک موضوع مشابه و یا مشاهداتی که به لحاظ زمانی بصورت متوالی جمع‌آوری شده می‌توان نام برد. در این بخش موقعیت‌هایی را در نظر گرفتیم که هدف علمی اولیه آن‌ها چگونگی توزیع Y_i براساس تغییرات در یک بردار متغیر کمکی $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$، p تغییر می‌کند. وابستگی، به‌عنوان اولویت دوم اهمیت در نظر گرفته می‌شود.

تابع چگالی Y_i با $p_i(y_i; \boldsymbol{\lambda}) = p(y_i | \mathbf{x}_i; \boldsymbol{\lambda})$ نشان داده می‌شود و $\mathbf{x}_i$ را به‌صورت مفروض دارد، از این رو متغیرهای کمکی اجازه دارند هم بر میانگین Y_i و هم کل توزیع حاشیه‌ای تک‌متغیره نیز تاثیر داشته باشند. دراین چهارچوب، چگالی $p_i(y_i; \boldsymbol{\lambda})$ ، مدل رگرسیون را شناسایی می‌کند. احتمال تقریبی شبیه‌سازی‌شده بدون پیش‌فرض‌های بیش‌تر در مورد ماهیت وابستگی در میان پاسخ‌ها، می‌تواند پارامترهای حاشیه $\boldsymbol{\lambda}$ با فرضیه‌های استقلال زیر نتیجه‌گیری کند،

$$L_{ind}(\lambda; y) = \prod_{i=1}^{n} p_i(y_i; \lambda). \qquad (۱.۲)$$

اگر توزیع‌های حاشیه‌ای $p_i(y_i; \boldsymbol{\lambda})$ به‌درستی مشخص شده باشند براوردگر درستنمایی ماکزیمم وابسته $\hat{\boldsymbol{\lambda}}_{ind} = \text{argmax}_{\boldsymbol{\lambda}} \varphi_{ind}(\boldsymbol{\lambda}; \mathbf{y})$ ، می‌تواند $\boldsymbol{\lambda}$ را با هیچ مشخصه‌ای از توزیع مشترک $\mathbf{Y}$ بر مدل ماتریس مفروض $\mathbf{X} = (X_1, \ldots, X_n)^T$ برآورد کند. علی‌رغم این مشخصه مهم، دلایل دیگری نیز در مورد براوردگر فوق می‌توان در نظر گرفت. اول، ممکن است به‌دلیل از دست دادن بخش قابل توجهی از کارایی خود، زمانی که وابستگی قابل توجه باشد دچار مشکل شود. دوم، تئوری استاندارد درستنمایی اعمال نشده و تصحیح خطاهای استاندارد باید براساس فرمول‌های ساندویچی[۲] صورت گرفته باشد، که محاسبات آن‌ها زمانی که بردار پاسخ $\mathbf{Y}$ به زیربردارهای وابسته تجزیه نشده باشد دشوار است. در نهایت، پیش‌بینی‌هایی که وابستگی را مورد توجه قرار نداده باشند کیفیت پایینی دارند. (ویت (۱۹۹۴))

با توجه به دلایل فوق، تکمیل کردن مدل رگرسیون $p_i(y_i; \boldsymbol{\lambda})$ به همراه ساختار وابسته برای به‌دست آوردن نتایج استنتاجی و پیش‌بینی دقیق اهمیت دارد. مدل ایده‌آل، دارای کلیه توزیع‌های مشترک $\mathbf{Y}$

[۱] GCMR

[۲] Sandwich-type formulas

با تک‌متغیرهای حاشیه‌ای $p_i(y_i;\boldsymbol{\lambda}),\ i=1,\ldots,n$ خواهد بود. با این‌حال، این مدل نیمه‌پارامتری گسترده‌تر برای کاربردهای عملی بسیار کلی به‌نظر می‌رسد. از این‌رو، در ادامه ما مدل پارامتری دقیق‌تری ارائه می‌دهیم که برای بسیاری از کاربردها به‌اندازه کافی انعطاف‌پذیر است.

۳.۲ مدل‌های رگرسیون حاشیه‌ای مفصل‌های گاوسی

متفاوت از صورت ارائه‌شده در فصل اول یک مدل رگرسیون را به‌صورت زیر نیز می‌توان تعریف کرد

$$y_i = g(x_i, \epsilon_i; \boldsymbol{\lambda}), \quad i=1,\ldots,n \tag{۲.۲}$$

در اینجا $g(\cdot)$ تابع مناسب رگرسیون $\mathbf{x}_i$ و متغیر تصادفی مشاهده نشده ϵ_i است که به‌عنوان عبارت خطا نشان داده می‌شود. فرض بر این است که مدل رگرسیون ۲.۲ به‌عنوان یک بردار پارامترهای $\boldsymbol{\lambda}$ شناخته می‌شود. رابطه زیر در میان مشخصه‌های احتمالی برای تابع $g(\cdot)$ انتخاب مناسبی است

$$y_i = F_i^{-1}\left\{\Phi(\epsilon_i);\boldsymbol{\lambda}\right\},\ i=1,\ldots,n, \tag{۳.۲}$$

در اینجا ϵ_i متغیر نرمال استاندارد است، $F_i(\cdot;\boldsymbol{\lambda}) = F(\cdot\,|x_i;\boldsymbol{\lambda})$ و $\Phi(\cdot)$ توزیع تجمعی توابع Y_i و X_i که متغیر استاندارد نرمال نیز می‌باشند. با توجه به‌برهان تبدیل یکپارچه مدل رگرسیون ۳.۲ توزیع حاشیه‌ای مطلوب را برای پاسخ‌های Y_i ایجاد می‌کند و ϵ_i را در عبارات خطای نرمال مشخص می‌سازد. مشخصه ۳.۲ شامل کلیه مدل‌های رگرسیون پارامتری ممکن برای پاسخ‌های پیوسته و غیرپیوسته می‌باشد. به‌عنوان مثال، مدل رگرسیون خطی گاوسی $Y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \sigma\epsilon_i$ متناظر با مجموعه

$$F_i(Y_i;\boldsymbol{\lambda}) = \phi\{(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma\}$$

در ۳.۲ ، به‌همراه $\boldsymbol{\lambda} = (\boldsymbol{\beta}^T,\sigma)^T$ می‌باشد درحالی‌که مدل لگاریتم‌خطی پواسون از صورت زیر به‌دست می‌آید:

$$F_i(y_i;\boldsymbol{\lambda}) = \sum_{j=0}^{Y_i} \frac{e^{-\mu_i},\mu_i^3}{j!}$$

که در اینجا $\mu_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$ و $\boldsymbol{\lambda} \equiv \boldsymbol{\beta}$ است.

برای توسعه بیش‌تر، مهم است توجه داشته باشید که نگاشت بین پاسخ Y_i و عبارت خطای ϵ_i تنها در موارد پیوسته یک‌به‌یک است، در غیراین‌صورت نگاشت یک‌به‌چند است و رابطه ۳.۲ نمی‌تواند بین Y_i و ϵ_i وارونه شود.

در ادامه به تحلیل مدل رگرسیون زمانی که وابستگی وجود دارد می‌پردازیم. سپس مشخصات مدل با فرض اینکه $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^T$ بردار خطاها با توزیع نرمال چندمتغیره باشد را بررسی می‌کنیم. برای این منظور فرض کنید

$$\boldsymbol{\epsilon} \sim MVN(0,\Omega) \tag{۴.۲}$$

که در آن Ω یک ماتریس همبستگی است که شکل خاصی از مشاهدات پیش‌بین متناظر با $\Omega = I_n$ ، ماتریس شناسایی $n \times n$ است می‌باشد. شناسایی مدل نیازمند این است‌که ϵ بردار میانگین صفر و واریانس‌های واحد داشته باشد زیرا مشخصه‌های تک‌متغیره به‌صورت جداگانه در حاشیه‌های $p_i(y_i; \boldsymbol{\lambda})$ مدل‌سازی می‌شوند.

مشخصه‌های مدل به‌خوبی می‌تواند مولفه حاشیه‌ای (۳.۲) را از مولفه وابسته (۴.۲) جدا کند، مولفه وابسته برحسب یک فرایند نرمال چندمتغیره توصیف شده است. شکل‌های مختلف وابستگی در داده‌ها می‌تواند با تنظیم مناسب ماتریس همبستگی Ω به‌عنوان تابع یک بردار پارامتر $\boldsymbol{\tau}$ مدل‌سازی شود، برای مشاهده برخی مثال‌های رایج در این زمینه به‌بخش ۴.۲ مراجعه کنید. کل بردار پارامترها توسط $\boldsymbol{\theta} = (\boldsymbol{\lambda}^T, \boldsymbol{\tau}^T)^T$ نشان داده می‌شود، که $\boldsymbol{\lambda}$ بردار پارامترهای حاشیه‌ای و $\boldsymbol{\tau}$ بردار پارامترهای وابسته را نام‌گذاری می‌کنند.

مدل ۳.۲ و ۴.۲ تفسیر نرمالی بر اساس نظریه مفصل ارائه می‌کنند که در آن امتیازات نرمال ϵ_i به‌عنوان تبدیلات خطی متغیرهای Y_i در نظر گرفته می‌شوند. (جویی . اچ (۱۹۹۷)) رویکردی که در اینجا بررسی می‌شود با بسیاری از دیگر رویکردها در بررسی‌های گذشته در مورد مدل‌سازی مفصل‌ها متفاوت است به‌طوری‌که برعکس حاشیه‌ها، به‌عنوان مولفه‌های مزاحم در نظر گرفته می‌شوند و نکته قابل توجه آن‌ها این است که در در پارامترهای وابسته‌ی مفصل قرار گرفته‌اند. به‌عنوان مثالی از کاربرد مفصل نیمه‌پارامتری گاوسی می‌توانید به (هوف (۲۰۰۷)) مراجعه کنید.

همان‌طور که در بخش ۲.۲ گفته شد، نتایج ایده‌آل برای $\boldsymbol{\lambda}$ باید براساس مدل نیمه‌پارامتری کلیه مفصل‌های احتمالی در حاشیه‌های $p_i(y_i; \boldsymbol{\lambda})$ باشند. مدل مفصلی که در اینجا مورد بحث قرار می‌گیرد مدل مفصل گاوسی است.

یک محدودیت پذیره نرمال بودن تابع مفصل، این است که ساختار کامل چندمتغیره توسط وابستگی‌های دومتغیره القا می‌شود. برای فرار از این محدودیت، می‌توان از سایر توابع مفصل استفاده کرد. اما مشکل استفاده از سایر توابع مفصل سختی تفسیرپذیری نتایج حاصل از آن‌هاست. سونگ (۲۰۰۰) مدل‌های خطی تعمیم‌یافته مفصل‌های گاوسی را برای تحلیل داده‌های طولی معرفی کرد. برای توسعه پاسخ‌های طولی چندمتغیره به سونگ و همکاران (۲۰۰۹) مراجعه کنید. این مدل‌ها، مثالی از GCMR هستند.

۴.۲ مدل‌های وابسته

ساختار وابستگی در GCMR از طریق مشخصه یک ماتریس همبستگی مناسب Ω بردار خطاهای $\boldsymbol{\epsilon}$ مدل‌سازی می‌شود. با این‌که، هر ماتریس همبستگی Ω مجاز است، در اینجا برخی از انواع مدل‌های خاصی را که کاربرد گسترده‌تری دارند معرفی خواهیم کرد.

۱.۴.۲ داده‌های طولی و خوشه‌ای

فرض کنید مشاهدات در m خوشه و ابعاد $r = 1 \ldots, m, n_r$ با $\sum_{r=1}^{m} n_r = n$ گروه‌بندی شده‌اند. این مورد داده‌های طولی یا پانلی است که دراینجا m آزمودنی در n_r موقعیت مشاهده شدند. براساس فرض

استاندارد وابستگی بین آزمودنی‌های مختلف یا گروه‌ها، ماتریس‌های همبستگی مناسب برای خطاهای قطری بلوکی هستند. داریم:

$$\Omega = \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_m \end{pmatrix},$$

دراینجا Ω_r یک ماتریس هم‌بسته $n_r \times n_r$ است. بطور مشابه با روش تعمیم برآورد معادلات (لی یانگ و زگر (۱۹۸۶))، می‌توانیم برخی ساختارهای هم‌بسته مفید را برای بلوک عمومی Ω_r شناسایی کنیم. اندیس‌های i و j نشان‌دهنده دو مشاهده متعلق به خوشه مشابه r هستند، یعنی

$$(n_1 + \cdots + n_{r-1}) + 1 \leqslant i < j \leqslant (n_1 + \cdots + n_r).$$

پس برای ساختارهای همبستگی مختلف خواهیم داشت:

۱. تبادل‌پذیر: $corr(\epsilon_i, \epsilon_j) = \tau$

۲. اتورگرسیو مرتبه یک: $corr(\epsilon_i, \epsilon_j) = \tau^{|i-j|}$

۳. میانگین متحرک مرتبه q. $corr(\epsilon_i, \epsilon_j) = \tau_{|i-j|}$ برای $|i-j| \leqslant q$;

۴. بدون ساختاری که متناظر است با یک ماتریس همبستگی بدون محدودیت.

سونگ (۲۰۰۷) مدل‌های این نوع را برای موارد خاص حاشیه‌ای که توسط مدل‌های خطی تعمیم‌یافته مفروض شده‌اند مورد بررسی قرار داد و آن‌ها را بردار مدل‌خطی تعمیم‌یافته نامید. با این‌حال، مفصل‌های گاوسی را می‌توان برای پیوند به هر نوع حاشیه‌ای استفاده کرد حتی آن‌هایی که متعلق به خانواده نمایی هستند. به‌عنوان مثالی از این مورد، پاسخ‌های خوشه‌بندی‌شده سانسور شده بعداً توسط مدل رگسیون ویبول[۳] در بخش ۱.۵.۳ مورد تجزیه و تحلیل قرار می‌گیرند. نوشته سونگ بر روی داده‌های‌طولی با خوشه‌هایی در اندازه کوچک متمرکز شده است.

۲.۴.۲ سری‌های زمانی

مدل‌های رگرسیون حاشیه‌ای با خطاهای سری‌های‌زمانی برای مشاهدات فواصل برابر ممکن است با فرض این‌که Ω یک ماتریس هم‌بسته است و توسط یک اتورگرسیو و فرایند میانگین متحرک مرتبه‌های p و q ایجاد شده است.

[۳] Weibull regression

۳.۴.۲ داده‌های فضایی

تحلیل رگرسیون با پاسخ‌های فضایی_زمانی وابسته را می‌توان با فرض اینکه خطاها از میدان تصادفی گاوسی مانا ایجاد شده‌اند مدل‌سازی نمود (کرسی(۱۹۹۳)). یک انتخاب انعطاف‌پذیر برای Ω در مورد داده‌های‌فضایی، تابع هم‌بسته ایزوتروپیک ماترن [۴] است

$$corr(\epsilon_i, \epsilon_j) = \frac{1}{2^{r_2-1}\Gamma(\tau_2)} \left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\| 2}{\tau_1} \right)^{\tau_2} B_{r_2} \left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\| 2}{\tau_1} \right), \qquad (۵.۲)$$

در اینجا s_i مختصات iامین مشاهده و B_{τ_2} تابع بسل تعدیل‌کننده مرتبه τ_2 است. پارامترهای τ_1 و τ_2 هر دو نیاز دارند که به‌شدت مثبت باشند. نمایش رگرسیون‌فضایی در بخش ۲.۵.۳ ارائه شده است.

۴.۴.۲ ارائه چند مثال از کاربرد مفصل

در این بخش به‌منظور بررسی ساختار وابستگی، چند مثال مطرح می‌کنیم که در مثال اول اهمیت کشف و تحلیل ساختار وابستگی بین متغیرها را بیان نموده، در مثال دوم اهمیت مفصل‌ها را در تحلیل ساختار وابستگی در نظر می‌گیریم و در پایان مفصل ارشمیدسی مناسب با قیمت سهام چند شرکت نفت را برآورد و به‌وسیله مفاصل قیمت سهام هریک را پیش‌بینی می‌کنیم. بعد از ارائه این مثال‌ها به‌بررسی ویژگی‌های مدل و شکل‌های توزیعی آن می‌پردازیم.

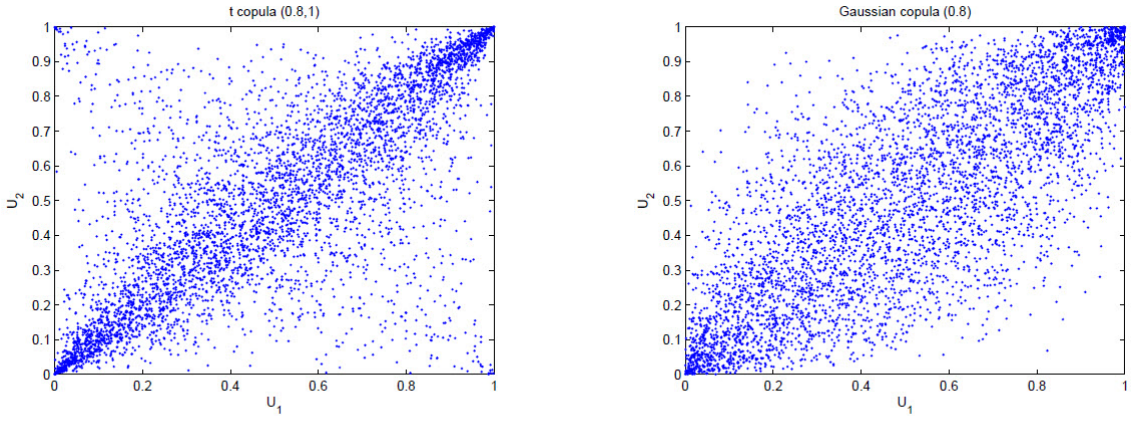
مثال ۱: دو پراکنش بین دو متغیر که هر کدام دارای نمونه‌های ۵۰۰۰ تایی هستند را در نظر می‌گیریم (شکل ۱.۲) همبستگی‌خطی پیرسون در هر دو نمودار پراکنش ۰/۸ می‌باشد، اما در نمودار (b) ساختار وابستگی نقاط توام در کرانه‌ها و نزدیک قطر بسیار زیاد می‌باشد. در حالی‌که در نمودار (a) چنین وضعیتی وجود ندارد.

بنابراین علاوه بر ضریب‌همبستگی خطی پیرسون، پی‌بردن به ساختار وابستگی بین متغیرها از اهمیت ویژه‌ای برخوردار است. به‌عبارتی این بررسی باید فراتر از معیارهای همبستگی مثل همبستگی خطی پیرسون باشد. یکی از این مدل‌ها برای بررسی ساختار وابستگی جامع‌تر توابع مفصل می‌باشد.

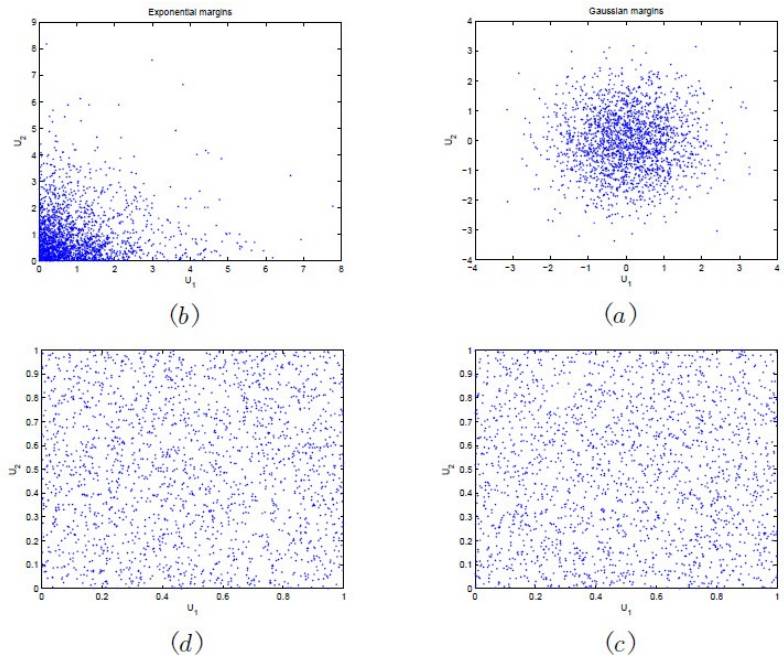
مثال ۲: برای پی‌بردن به اهمیت توابع مفصل در نشان دادن ساختار وابستگی، برای بردار تصادفی (X_1, X_2) دو توزیع دومتغیره را در نظر بگیرید. توزیع دومتغیره اول را توزیع نرمال دومتغیره با بردار میانگین صفر، واریانس یک و ضریب همبستگی صفر و توزیع دومتغیره دیگر را حاصل‌ضرب دو توزیع نمایی پیش‌بین با پارامتر یک می‌باشد.

توجه کنید که هر دو توزیع دومتغیره پیش‌بین و دارای حاشیه‌های متفاوت می‌باشند. اما همان‌طور که سطر اول شکل ۲.۲ نشان می‌دهد، این استقلال به‌وسیله رسم نموار پراکنش مشاهدات قابل نمایش نمی‌باشد، زیرا اثر توزیع‌های حاشیه‌ای در آن‌ها وجود دارد. اما این استقلال با کمک نمودارهای پراکنش رتبه‌ای آن‌ها یعنی نمودار $F_1(x_1)$ در مقابل $F_2(x_2)$ که همان عنصر دامنه تابع مفصل می‌باشند و به‌ترتیب توابع توزیع متغیرهای X_1 و X_2 هستند، به‌خوبی نمایان است (سطر دوم نمودار ۲.۲) در همه مثال‌های زیر، سه متغیر داریم: یک متغیر وابسته و دو پیش‌بین، همچنین حجم نمونه در همه

[۴] Mat´ern isotropic correlation function



شکل ۱.۲: نمودار پراکنش بین متغیرها با ضریب همبستگی پیرسون یکسان ۰/۸ و ساختار وابستگی متفاوت



شکل ۲.۲: نمودار پراکنش مشاهدات توزیع نرمال پیش‌بین (a) و توزیع نمایی (b) به‌همراه پراکنش رتبه ای آن‌ها یعنی (c) و (d)

مثال‌ها ۵۰ مورد است. از همان ماتریس همبستگی در تمام مثال‌های زیر استفاده می‌کنیم.

$$R = \begin{bmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix}$$

برای شبیه‌سازی ۵۰ سه‌گانه داده به‌کاررفته دراین مثال از روش‌شناسی سلمن و رایلی[۵] (۱۹۹۰) توصیف شده استفاده می‌کنیم.

قرار دهید $F_i(x_i), i = 1, 2, 3$ که با ماتریس همبستگی مشخص می‌شود.

۱. ابتدا بردار (u_1, u_2, u_3) را از یک توزیع نرمال چندمتغیره صفر و یک در نظر می‌گیریم.

۲. مقدار $s_i = \Phi(u_i)$ برای هر یک از سه متغیر محاسبه کنید.

۳.خطا را با استفاده از $F_{X_i}^{-1}(s_i), i = 1, 2, 3.$ محاسبه کنید.

مثال ۳: توزیع سه متغیر که X_1 متغیر وابسته است. که در این مثال متغیر وابسته، گسسته است. توزیع X_2 توزیع تقریبی است. در این مثال مدل رگرسیون مفصل بهتر از مدل OLS است. نتایج در جدول ۱.۲ آمده است.

جدول ۱.۲: برآورد خطا و پارامتر برای مثال ۳

متغیر	X_1	X_2	X_3	مجموع توان‌های دوم خطا	
مدل	پواسن	نمایی	گاما	مفصل	OLS
خطاها/MLEs	5.65	119.38907	3.67 ، 88.98	109.072	114.6649

مثال ۴: فرضیات این مثال نیز همانند مثال قبل است با این تفاوت که توزیع X_1 و X_2 به‌صورت تجربی برآورد شده‌اند و داده با استفاده از مدل پواسن از روش GLM تجزیه و تحلیل شدند. مدل رگرسیون مفصل بهترین کارایی را دارد. نتایج در جدول ۲.۲ نشان داده شده است.

جدول ۲.۲: برآورد خطا و پارامتر برای مثال ۴

متغیر	X_1	X_2	X_3	مجموع توان‌های دوم خطا		
مدل	پواسن	نمایی	گاما	مفصل	OLS	GLM
خطاها/MLEs	5.65	N/A	N/A	110.379	114.6649	119.518

با توجه به مثال‌های ۳ و ۴ رگرسیون مفصل در مقایسه با GLM، رویکرد بهتری ارائه می‌کند. قدرت آن در توانایی انتخاب توزیع برای متغیرهای وابسته‌ای که عضو خانواده نمایی نیستند، مشهود است.

همچنین، به محقق اجازه می‌دهد توزیع دلخواه خود را برای حاشیه‌ها انتخاب کند (GLM تنها به‌توصیف توزیع متغیر وابسته نیاز دارد). همان‌طور که توزیع‌های سنگین و مبهم در بیمه رایج هستند، این روش راهی برای تخصیص توزیع‌های دلخواه به مدل‌های رگرسیون را فراهم می‌کند. مانند GLM، این روش امکان مدل‌سازی وابستگی غیر خطی را فراهم می‌کند. از آنجا که همبستگی اندازه مفیدی

[۵] Clemen and Riley

از وابستگی در دنیای غیرنرمال نیست، برای توصیف ارتباط بین متغیرها ضروری است. در این راستا رگرسیون مفصل جایگزین خوبی برای GLM است.

۵.۲ مشخصه‌های مدل

در این بخش به‌بررسی شکل‌های توزیعی مدل و ویژگی‌های وابستگی می‌پردازیم.

۱.۵.۲ شکل‌های توزیعی

مدل‌های $GCMR$ با مدل‌های حاشیه‌ای دیگر در شکل دوبعدی متفاوت هستند. در ادامه، نگاشت ۳.۲ بین ϵ_i و Y_i یک‌به‌یک است، از این رو توزیع‌های حاشیه‌ای پاسخ ها از طریق قوانین تبدیل استاندارد از توزیع خطاهای متناظر به دست می‌آیند. به‌عنوان مثال، در توزیع دو متغیره زیر داریم:

$$p_{ij}(y_i, y_j; \boldsymbol{\theta}) = p_i(y_i; \boldsymbol{\lambda}])p_j(y_j; \boldsymbol{\lambda})q(\epsilon_i, \epsilon_j; \boldsymbol{\theta}), \tag{۶.۲}$$

که در این‌جا

$$q(\epsilon_i, \epsilon_j; \boldsymbol{\theta}) = \frac{p(\epsilon_i, \epsilon_j; \boldsymbol{\theta})}{p(\epsilon_i; \boldsymbol{\lambda})p(\epsilon_j; \boldsymbol{\lambda})}, \tag{۷.۲}$$

چگالی مفصل گاوسی دومتغیره است. با در نظر گرفتن فرضیات مدل، $p(\epsilon_i; \boldsymbol{\lambda})$ چگالی نرمال استاندارد تک‌متغیره است، درحالی‌که $p\left(\epsilon_i, \epsilon_j; \boldsymbol{\theta}\right)$ چگالی نرمال دومتغیره با میانگین‌های صفر است، واریانس واحد و همبستگی توسط یکی از اجزا در موقعیت (i, j) در ماتریس Ω مفروض شده‌اند.

در موارد رسته‌ای و مجزا، نگاشت ۶.۲ چندبه‌یک است. که نشان‌دهنده مسئله است که توزیع‌های حاشیه‌ای مشترک توسط انتگرال‌های نرمال چندمتغیره بیان می‌شوند. به‌عنوان مثال، توزیع حاشیه‌ای تک‌متغیره یک انتگرال دوبعدی است

$$p_{ij}(y_i, y_j; \boldsymbol{\theta}) = \int_{D_i(y_i;\boldsymbol{\lambda})} \int_{D_j(y_j;\boldsymbol{\lambda})} p(\epsilon_i, \epsilon_j; \boldsymbol{\theta}) d\epsilon_i d\epsilon_j, \tag{۸.۲}$$

دامنه آن‌ها، حاصلضرب دکارتی بازه‌ها به‌شکل زیر است:

$$D_i(y_i; \boldsymbol{\lambda}) = \left[\phi^{-1}\left\{F_i(y_i^-; \boldsymbol{\lambda})\right\}, \phi^{-1}\left\{F_i(y_i; \boldsymbol{\lambda})\right\}\right], \tag{۹.۲}$$

است که در این‌جا $F_i(y_i^-; \boldsymbol{\lambda})$ حد چپ $F_i(\cdot; \boldsymbol{\lambda})$ در Y_i است. اگر Y_i شامل N باشد، مانند توزیع‌های دوجمله‌ای و حاشیه‌ای پواسن، پس اگر گسسته باشد $y_i^- = y_i - 1$.

۲.۵.۲ ویژگی‌های وابستگی

در مورد خاص مدل‌های رگرسیون خطی با خطاهای نرمال توزیع شده، همبستگی بین جفت پاسخ‌ها، متغیر کمکی متناظر ایجاد می‌کند که هم‌زمان با همبستگی خطاهای نرمال متناظر $corr(\epsilon_i, \epsilon_j)$ است.

در غیراین‌صورت، همبستگی میان Y_i و Y_j یک تابع غیرخطی از همبستگی ϵ_i و ϵ_j است. این تابع غیرخطی را می‌توان با روش‌های عددی یکپارچه مختلف محاسبه نمود. اخیراً، کوگیومتزیس و بوراسنتا (۲۰۱۰) پیشنهاد دادند که تقریب این همبستگی‌ها در تقریب خطی تکه‌ای استفاده می‌شود.

در تئوری مفصل، اندازه‌های جایگزین مرتبط مبتنی بر رتبه، اغلب مورد استفاده قرار می‌گیرند. دو اندازه مرتبه بسیار محبوب، τ کندال و ρ اسپیرمن هستند. یک بررسی جدید در مورد کاربرد τ کندال در تحلیل رگرسیون طولی توسط پارزن و همکاران (۲۰۱۱) ارائه شده است. سونگ بحث دقیقی در مورد شایستگی‌های مرتبط با اندازه‌ها در مدل مفصل‌های گاوسی ارائه نمود.

اگرچه، محاسبه شکل بسته برای دومتغیره‌ها، و مرتبه بالاتر، در حال حاضر وجود ندارد، برخی از وجوه کلیدی ساختار وابسته مدل ۶.۲ و ۴.۲ به‌سادگی مشتق شده‌اند.

ویژگی ۱: اگر خطاهای ϵ_i و ϵ_j ناهم‌بسته باشند، پس جفت پاسخ‌های متناظر Y_i و Y_j با در نظر گرفتن متغیرهای کمکی X_i و X_j ، پیش‌بین هستند. به‌عبارت دیگر، ناهمبستگی خطاها به‌معنی کوواریانس صفر است و تنها در توزیع نرمال در صورتی که همبستگی صفر باشد X_i و X_j مستقل از یکدیگر خواهند بود و برای جلوگیری از تاثیر دو متغیر بر میانگین، دو متغیر X_i و X_j بایستی ثابت در نظر گرفته شوند.

کیفیت انسجام از بیان توزیع دو متغیره در موارد پیوسته ۶.۲ و غیرپیوسته ۸.۲ مشخص است. به بدنار و همکاران (۲۰۱۰) مراجعه کنید که ویژگی‌های مدل‌های مفصل گاوسی را به‌خوبی مورد بررسی قرار داده‌اند. برای درک ویژگی ۱، مورد خاصی از رگرسیون را با خطاهای مانای سری زمانی در نظر بگیرید. این ویژگی بیان‌کننده این‌است که اگر خطاها به‌دنبال پروسه میانگین متحرک مرتبه q باشند، پس پاسخ‌هایی که بیشتر از واحدهای q هستند مستقل می‌باشند.

ویژگی ۲: جهت ارتباط بین هر جفت پاسخ‌های Y_i و Y_j با در نظر گرفتن متغیرهای کمکی X_i و X_j، هم‌زمان با نشانه همبستگی بین جفت خطاهای متناظر ϵ_i و ϵ_j هستند.

این پیامد مستقیم نگاشت ۴.۲ است که غیرافزایشی شده است. این ویژگی اطمینان می‌دهد که ساختار همبستگی خطاهای نرمال، جهت وابستگی بین پاسخ‌هایی که متغیر کمکی را مفروض کرده‌اند تعدیل نمی‌کند. همان‌طور که در مثال‌های بخش ۵.۳ نشان داده می‌شود، این نتیجه ساده برای تفسیر مدل برازش بسیار کاربردی است.

ویژگی ۳: اگر بردار خطای ϵ یک روند مارکوفی با مرتبه p را دنبال کند و کلیه توزیع‌های حاشیه‌ای پیوسته باشند، بردار پاسخ Y با مفروض داشتن مدل ماتریس X فرایند مارکوفی با مرتبه p را دنبال خواهد کرد.

این ویژگی به‌سادگی در مورد خاص سری‌های زمانی یا داده‌های طولی قابل تایید است. در واقع، چگالی شرطی یک پاسخ پیوسته Y_i موارد ماقبل خود را به‌صورت زیر مفروض دارد

$$p_i(y_i\,|y_{i-1}\,,\ldots,y_1;\boldsymbol{\theta}) = \frac{p_i(y_i;\boldsymbol{\lambda})}{p(\varepsilon_i;\boldsymbol{\lambda})}p(\varepsilon_i\,|\varepsilon_{i-1},\ldots,\varepsilon_1;\boldsymbol{\theta}). \qquad (۱۰.۲)$$

اگر خطاها از نوع مارکوفی مرتبه p باشند، چگالی شرطی فوق به‌چگالی شرطی محدود حافظه کاهش

می‌یابد، که قطعاً

$$\begin{aligned} p_i(y_i \,|y_{i-1} \,, \ldots, y_1; \boldsymbol{\theta}) &= \frac{p_i(y_i; \boldsymbol{\lambda})}{p(\varepsilon_i; \boldsymbol{\lambda})} p(\varepsilon_i \,|\varepsilon_{i-1}, \ldots, \varepsilon_1; \boldsymbol{\theta}) \\ &= p_i(y_i \,|y_{i-1}, \ldots, y_{1-p+1}; \boldsymbol{\theta}) \end{aligned} \tag{۱۱.۲}$$

به‌طور مشابه، اما با پیچیدگی‌های نمادی بیش‌تر، می‌توان نشان داد که اگر خطاهای ϵ، تحقق از یک میدان تصادفی گاوسی مارکوف[۶] هستند، سپس پاسخ‌های چندمتغیره مستمر Y تحقق از یک میدان تصادفی مارکوفی می‌باشند. برای سری‌های‌زمانی، اگر ویژگی ۱ بیان کند که یک نوع توازی بین خطاهای میانگین متحرک و پاسخ‌ها وجود دارد، پس ویژگی ۳ این توازی را تنها برای پاسخ‌های مستمر به فرآیندهای اتورگرسیو نیز گسترش می‌دهد.

[۶] Gaussian Markovian random field

فصل ۳

استنباط آماری مدل و معیارهای برازش

۱.۳ مقدمه

دانشمندان علم آمار، این علم را به دو شاخه نظری و تجربی تقسیم می‌کنند. آمار تجربی به روش‌هایی اطلاق می‌شود که با استفاده از روش‌های شبیه‌سازی، خواص برآوردکننده‌ها را مطالعه می‌نمایند. این روش‌ها به شبیه‌سازی مونت‌کارلو موسوم شده است. روش‌های مونت‌کارلو به روش‌هایی اطلاق می‌شوند که براساس دنباله‌ای از اعداد تصادفی به‌بررسی مسائل می‌پردازند. نام مونت‌کارلو اولین بار توسط متروپلیس[۱] به‌دلیل شباهت شبیه‌سازی آماری به بازی‌های شانسی و به‌دلیل این‌که شهر مونت‌کارلو مرکز کشور کوچک موناکو، مرکز بازی‌های شانسی بود، به‌کارگرفته شد.

امروزه این روش در بسیاری از علوم مورد استفاده قرار می‌گیرد. کاربرد روش مونت‌کارلو در مدل‌سازی مسائل فیزیکی به ما این امکان را می‌دهد که سیستم‌های پیچیده‌تری را که در عمل بررسی آن‌ها بسیار مشکل است مطالعه کنیم. حل معادلاتی که بیانگر کنش و واکنش اتم‌های یک سیستم‌اند، محاسبه انتگرال‌های پیچیده‌ای که از راه‌های کلاسیک قابل محاسبه نیستند و مسائل بسیار پیچیده دیگر، نمونه‌هایی از کاربرد این روش‌اند.

هم‌چنین موضوع بوت‌استرپ[۲] (بازنمونه‌گیری) (افرون (۱۹۷۳)) به‌عنوان رویکرد ارتقای قابلیت

[۱] Metropolis
[۲] Bootstrap

یافته‌های آماری به‌شدت مورد توجه محققان مختلف است. بوت‌استرپ روشی ساده اما درعین حال قوی از روش نمونه‌گیری مونت‌کارلو است که برای تعیین دقت آماری یا برآوردکردن توزیع از روی آماره‌های نمونه است. در واقع یک نوع نمونه‌گیری است که با تبدیل از یک نمونه به‌دست می‌آید؛ نمونه‌گیری درون یک نمونه. اصل برابری بوت‌استرپ نشان می‌دهد که برآوردگر نمونه گرفته‌شده با روش بوت‌استرپ، برابر است با برآورد نمونه‌ی اصلی. روش بوت‌استرپ دارای دو پیش‌فرض است:

۱ــ نمونه شما، نمونه معتبری از جمعیت است.

۲ــ بوت استرپ نمونه‌ای را با جایگزینی از نمونه اصلی خواهد گرفت به‌طوری که هر نمونه گرفته‌شده با این روش مستقل، ولی دارای توزیع برابرند. به‌عبارت دیگر؛ نمونه‌های گرفته شده با روش بوت استرپ دارای توزیع جمعیت برابرند ولی هر نمونه، مستقل از نمونه‌های دیگر است.

در این فصل، ابتدا برازش پیشنهاد شده مدل‌های $GCMR$ با روش درستنمایی ماکزیمم را بیان کرده، سپس نمونه‌گیری نقاط مهم را به‌عنوان تقریب مناسبی از روش مونت‌کارلو برای داده‌هایی با ابعاد بزرگ‌تر مانند سری‌های زمانی، داده‌های‌فضایی و مطالعات‌طولی ارائه می‌دهیم. در ادامه آزمون مشخصات هاسمن را برای تعیین و بررسی درستی انتخاب مدل چندمتغیره به‌کار می‌گیریم.

۲.۳ رگرسیون مفصل گاوسی

یک بردار با n متغیر وابسته $Y_1, Y_2, ..., Y_n$ را در نظر بگیرید. توزیع تجمعی حاشیه‌ای تک‌متغیره y_i توسط $F(\cdot|\,\mathbf{x}_i)$ و به یک بردار p ــ بعدی از متغیر x_i بستگی دارد. فرض می‌کنیم $F(\cdot|\,\mathbf{x}_i)$ براساس پارامتر موقعیت μ_i پارامتربندی شده است، و معمولا به مقدار مورد انتظار $E(y_i|\,\mathbf{x}_i)$ مربوط است، که به x_i از طریق رابطه زیر بستگی دارد

$$g_1(\mu_i) = \mathbf{x}_i{}^T\boldsymbol{\beta} \qquad (۱.۳)$$

برای یک تابع پیوند مناسب $g_1(\cdot)$ و یک بردار p ــ بعدی از ضرایب رگرسیون $\boldsymbol{\beta}$. این تنظیمات شامل انواع کلاس‌های محبوب مدل مانند، مدل‌های‌خطی تعمیم‌یافته (مک‌کالاغ و نلدر(۱۹۸۹)) یا رگرسیون بتا (کریباری و زیلیس(۲۰۱۰)) است. اگر توزیع Y_i شامل پارامتر پراکندگی باشد، مدل را می‌توان برای پراکندگی متغیر با مدل رگرسیون دوم گسترش داد.

$$g_2(\psi_i) = z_i{}^T\gamma \qquad (۲.۳)$$

که $g_2(\cdot)$ تابع پیوند پراکندگی است، ψ_i پارامتر پراکندگی مربوط به y_i است، z_i بردار q ــ بعدی از متغیر پراکندگی است و γ بردار متناظر با ضرایب رگرسیون است. به‌خاطر ساده بودن نماد، از این پس، توزیع‌تجمعی یکنواخت حاشیه‌ای Y_i به‌عنوان $F(\cdot|\,\mathbf{x}_i)$ و حتی در مورد متغیر پراکندگی، که در واقع مدل $F(\cdot|\,\mathbf{x}_i, \mathbf{z}_i)$ است، نشان داده می‌شود.

در رگرسیون مفصل گاوسی، وابستگی بین متغیرها با مفصل گاوسی مدل‌سازی می‌شود به‌طوری که تابع توزیع تجمعی توام داده‌ها از رابطه زیر به‌دست می‌آید

$$Pr(Y_1 \leqslant y_1, ..., Y_n \leqslant y_n) = \Phi_n(\epsilon_1, ..., \epsilon_n; \mathrm{P}),$$

که در آن $\epsilon_i = \Phi^{-1}\{F(y_i|\, x_i)\}$، با $\Phi(\cdot)$ نشان‌دهنده تابع توزیع تجمعی نرمال استاندارد تک‌متغیره است و $\Phi_n(\cdot, \mathrm{P})$ تابع توزیع‌تجمعی نرمال استاندارد چندمتغیره با ماتریس همبستگی P است.

فرمول معادل مدل مفصل گاوسی که بر تنظیم رگرسیون تأکید دارد، در ماساروتو و وارین (۲۰۱۲) بررسی داده شده است. یک مدل رگرسیون را در نظر بگیرید که هر متغیر y_i را به یک بردار از متغیر x_i توسط رابطه عمومی مرتبط می‌کند

$$y_i = h(\mathbf{x}_i, \epsilon_i),$$

که ϵ_i خطای تصادفی است. در بین بسیاری از ویژگی‌های ممکن از عملکرد $h(\cdot)$ و خطای ϵ_i، مدل رگرسیون مفصل‌گاوسی فرض می‌کند

$$h(\mathbf{x}_i, \epsilon_i) = F^{-1}\{\,\Phi(\epsilon_i)|\, x_i\},$$

و بردار خطا $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ دارای توزیع نرمال استاندارد چندمتغیره با ماتریس همبستگی P است. به‌بیان دیگر، مفصل گاوسی یک مدل رگرسیون ساخته‌شده در مسیر دستیابی به (i) حفظ توزیع یکنواخت حاشیه‌ای و (ii) داشتن خطاهای نرمال چندمتغیره است.

یکی از ویژگی‌های جالب توجه مفصل گاوسی این است که شکل‌های مختلف وابستگی را می‌توان با پارامتری کردن مناسب ماتریس همبستگی P بیان کرد. برای مثال، داده‌های‌طولی با فرض اینکه مدل‌های همبستگی در نظر گرفته شده در معادلات برآوردشده تعمیم داده شوند می‌توانند مدل‌سازی شوند (سونگ(۲۰۰۷)). وابستگی سریالی در سری‌های زمانی می‌تواند با ماتریس همبستگی مربوط به روند متناوب اتورگرسیو و میانگین متحرک (گولو و وارین(۲۰۱۴)) توصیف شود، درحالی‌که وابستگی‌فضایی را می‌توان با یک ماتریس همبستگی القا شده توسط یک منطقه تصادفی گاوسی توصیف کرد (بای(۲۰۱۴)).

۱.۲.۳ استنباط آماری

بسته $gcmr$ از حداکثر نتیجه‌گیری درست برای مدل رگرسیون مفصل گوسی پیروی می‌کند. θ را بردار پارامترهای مدل قرار دهید که شامل پارامترهای حاشیه‌ای یکنواخت و پارامترهای وابسته به ماتریس همبستگی مفصل گاوسی است. تابع احتمال θ در حالت پیوسته دارای فرم بسته است (به‌عنوان مثال، سونگ(۲۰۰۰))

$$L(\theta) = \phi_n(\epsilon_1, ..., \epsilon_n; \mathrm{P}) \prod_{i=1}^{n} \frac{f(\,y_i|\, x_i)}{\phi(\epsilon_i)},$$

که $\phi(\epsilon(\cdot))$ نشان‌دهنده چگالی نرمال استاندارد، $\phi(\epsilon_n(\cdot, \mathrm{P}))$ چگالی نرمال استاندارد n ـبعدی با ماتریس همبستگی P و $f(\cdot|\, x_i)$ چگالی Y_i با توجه به x_i است. حالت گسسته به‌طور قابل توجهی پیچیده‌تر است، زیرا احتمال آن توسط انتگرال نرمال n ـبعدی ارائه شده است.

$$L(\theta) = \int_{D_1} ... \int_{D_n} \phi_n(\epsilon_1, ..., \epsilon_n; \mathrm{P}) d_{\epsilon_1} ... d_{\epsilon_n}, \qquad (۳.۳)$$

که دراین رابطه دامنه انتگرال حاصل‌ضرب‌دکارتی بازه‌ها است.

$$\mathfrak{D}_i = \left[\Phi^{-1}\{F(\,y_i - 1|\, x_i)\}, \Phi^{-1}\{F(\,y_i|\, x_i)\}\right].$$

انتخاب‌های مختلفی برای ارزیابی عدم اطمینان برآورد درستنمایی ماکزیمم وجود دارد. اول، رویکرد کلاسیک که عدم اطمینان را با معکوس اطلاعات فیشر مشاهده شده ارزیابی می‌کند. به‌همین ترتیب، براوردگر واریانس تقریبی درستنمایی ماکزیمم را می‌توان با محصول بیرونی نمرات حاصل از تجزیه پیش‌بینی احتمال براورد نمود.

انتخاب‌های بالا اگر مدل مفصل‌گاوسی به‌درستی مشخص شود، معتبر هستند. با توجه به نگرانی‌های بالقوه در مورد فرضیه مفصل‌گاوسی، مقایسه خطاهای استاندارد مبتنی بر مدل با برآوردگرهای ساندویچ قوی توصیه می‌شود.

۳.۳ استنباط مبتنی بر درستنمایی

برازش پیشنهاد شده مدل‌های GCMR از طریق روش درستنمایی ماکزیمم انجام می‌گیرد. یک مزیت صریح این روش این است که می‌توان ابزارهای استانداردی را برای تست‌های فرضی و انتخاب مدل، مانند معیارهای اطلاعات و آمار نسبت درستنمایی مورد استفاده قرار داد. در خطوط زیر، جزئیات محاسبات درستنمایی مورد بحث قرار می‌گیرد. برای این هدف، بهتر است که موارد پیوسته، گسسته یا رسته‌ای به‌صورت جداگانه بررسی شوند. ما از ابتدا شروع می‌کنیم، زیرا ساده‌تر است و نسبت به موارد بعدی مقدماتی‌تر به‌حساب می‌آید. رابطه یک‌به‌یک بین پاسخ‌های $\mathbf{Y}$ و خطاهای $\boldsymbol{\epsilon}$ در حالت پیوسته منجربه درستنمایی برای $\boldsymbol{\theta}$ به‌صورت زیر می‌شود:

$$L(\boldsymbol{\theta}; y) = L_{ind}(\boldsymbol{\lambda}; y) q(\epsilon; \boldsymbol{\theta}), \tag{۴.۳}$$

که در آن چگالی مفصل به‌صورت زیر است

$$q(\epsilon; \boldsymbol{\theta}) = \frac{p(\epsilon_1, \ldots, \epsilon_n; \boldsymbol{\theta})}{p(\epsilon_1; \boldsymbol{\lambda}) \times \cdots \times p(\epsilon_n; \boldsymbol{\lambda})}$$

که می‌تواند به‌عنوان نسبت درستنمایی بین مدل نرمال چندمتغیره فرض شده برای خطاها و تحت فرضیه عدم وابستگی تفسیر شود. از این‌رو، درستنمایی $L(\boldsymbol{\theta}, y)$ از طریق تیز[۳] کردن درستنمایی مستقل $L_{ind}(\boldsymbol{\lambda}, y)$ از طریق یک معیار $q(\epsilon, \boldsymbol{\theta})$ که دلیلی برای وابستگی بین خطاها است به دست می‌آید.

حالت سخت‌تر حالت گسسته یا پاسخ‌های رسته‌ای است. در این حالت، ارزیابی درستنمایی نیازمند محاسبه‌ی انتگرال مستطیلی n بعدی است:

$$L(\boldsymbol{\theta}, y) = \int_{D_1(y_1;\boldsymbol{\lambda})} \cdots \int_{D_n(y_n;\boldsymbol{\lambda})} p(\epsilon_1, ..., \epsilon_n; \boldsymbol{\theta}) d\epsilon_1 ... d\epsilon_n \tag{۶.۳}$$

بیان مجدد انتگرال بالا با در نظر گرفتن تغییر متغیر از $\mathfrak{D}_1(y_1; \boldsymbol{\lambda}) \times ... \times \mathfrak{D}_n(y_n; \boldsymbol{\lambda})$ به $(0,1)^n$ با وارون سازی مؤلفه‌ای زیر راحت‌تر است:

$$\epsilon_i(u_i) = \Phi\{F_i(y_i; \boldsymbol{\lambda}) - u_i p_i(y_i; \boldsymbol{\lambda})\}, \quad i = 1, ..., n. \tag{۷.۳}$$

[۳] Sharp

سپس، درستنمایی ۶.۳ فرم زیر را در نظر می‌گیرد

$$L(\boldsymbol{\theta}; y) = L_{ind}(\boldsymbol{\lambda}; y) \int_{[0,1]} \cdots \int_{[0,1]} q\{\epsilon_1(u_1), ..., \epsilon_n(u_n); \boldsymbol{\theta}\} du_1 ... du_n, \qquad (۸.۳)$$

که شرح آن مشابه حالت پیوسته ۴.۳ است، جز در ترم تنظیم‌کننده که یک میانگین در نسبت‌های درستنمایی نوع $q(\epsilon; \boldsymbol{\theta})$ است اما در خطاهای تصادفی $\epsilon_i(u_i)$ داده‌شده توسط عبارت ۷.۳ محاسبه شده است.

همچنین برای محاسبه تابع درستنمایی مفصل‌های گاوسی با حاشیه‌های گسسته به‌دلیل اینکه فاقد عبارت شکل بسته است (برخلاف موردی که حاشیه‌ها پیوسته هستند) محاسبه عددی کارآمد موردنیاز است. فرض کنید که $Y(\cdot)$ در n مکان نمونه‌گیری گسسته مشاهده شده است که منجر به شمارش $y = (y_1, \ldots, y_n)$ می‌شود که در اینجا y_i تحقق Y_i است.

پس تابع درستنمایی $\eta = (\lambda^T, \theta^T)$ مبتنی بر داده y است و به‌صورت زیر در نظر گرفته شده

$$\begin{aligned} L(\eta; y) &= P_\eta(Y_1 = y_1, \ldots, Y_n = y_n) \\ &= \sum_{j_1=1}^{2} \cdots \sum_{j_n=1}^{2} (-1)^{\sum_{i=1}^{n} j_i} \Phi_n(\Phi^{-1}(F_1(t_1; \psi)), \ldots, \Phi^{-1}(F_n(t_n; \psi)); \Omega) \end{aligned} \qquad (۹.۳)$$

با $t_i = y_i + 1 - j_i$. ارزیابی معادله (۹.۳) به لحاظ محاسباتی حل‌نشدنی است حتی برای n متوسط زیرا نیازمند محاسبه 2^n است که هر یک شامل یک انتگرال nبعدی هستند. توجه داشته باشید که:

$$Y \overset{\mathrm{d}}{=} F^{-1}(\Phi(Z(s)); \lambda) \qquad (۱۰.۳)$$

در اینجا $F^{-1}(u; \lambda) = inf\{x : F(x; \lambda) \geq u\}, \ u \in (0, 1)$ تابع چندک توزیع است و $Z(\cdot)$ میدانی تصادفی گاوسی است. برای شمارش تابع توزیع تجمعی $F^{-1}(\cdot; \lambda)$ که یک تابع پله‌ای است و برای هر $y \in N_0$ آن‌گاه $Y = y$ در صورتی‌که

$$F(y - 1; \lambda) < \Phi(Z(s)) \leq F(y; \lambda)$$

باشد.پس

$$\begin{aligned} L(\eta; y) &= P_\eta(\zeta(y_1 - 1; \lambda) < Z(s_1) \leq \zeta(y_1; \lambda), \ldots, \zeta(y_n - 1; \lambda) < Z(s_n) \leq \zeta(y_n; \lambda)) \\ &= \int_{\zeta(y_1-1;\lambda)}^{\zeta(y_1;\lambda)} \cdots \int_{\zeta(y_n-1;\lambda)}^{\zeta(y_n;\lambda)} \phi_n(z_1, \ldots, z_n; \Omega) dz_1 \ldots dz_n \end{aligned} \qquad (۱۱.۳)$$

در اینجا $\zeta(y; \lambda) = \Phi^{-1}(F(y; \lambda))$ و $\phi_n(\cdot; \Omega)$ تابع چگالی احتمال توزیع $N_n(0, \Omega)$ است. عبارت اخیر برای محاسبات عددی بهتر پاسخ می‌دهد و در بخش بعد به‌عنوان نقطه شروع تقریب درستنمایی استفاده خواهد شد. یک نمایش دیگر از درستنمایی در ماساراتو و وارین (۲۰۱۲) ارائه شده که مشخص شد موافق با مورد پیدا شده توسط مادسن و فانگ (۲۰۱۱) است.

تغییر متغیرهای $(z_1, \ldots, z_n) \to (u_1, \ldots, u_n)$ را در نظر بگیرید که توسط معادله زیر مفروض است

$$u_i = \frac{F(y_i; \lambda) - \Phi(z_i)}{f(y_i; \lambda)}, \quad i = 1, \ldots, n \qquad (۱۲.۳)$$

برای معکوس مدل (۱۲.۳) خواهیم داشت:

$$\frac{d}{du_i}z_i(u_i) = \frac{-f(y_i;\lambda)}{\phi(\Phi^{-1}(F(y_i;\lambda) - u_i f(y_i;\lambda)))} = \frac{-f(y_i;\lambda)}{\phi(z_i(u_i))}, \quad i = 1,\ldots,n$$

از آن‌جا که z_i تابعی از u_i است ماتریس ژاکوبین $J = \frac{\partial(z_1(u_1),\ldots,z_n(u_n))}{\partial(u_1,\ldots,u_n)}$ در مدل معکوس قطری است پس

$$|det(J)| = \prod_{i=1}^{n} \frac{f(y_i;\lambda)}{\phi(z_i(u_i))}$$

از آن‌جا که (۱۲.۳) $(\zeta(y_1-1),\zeta(y_1)) \times \cdots \times (\zeta(y_n-1),\zeta(y_n))$ را به $(0,1)^n$ تبدیل می‌کند، از طریق فرمول‌های تغییر متغیر برای انتگرال‌های چندمتغیره به‌صورت زیر خواهد بود:

$$\begin{aligned} L(\eta;y) &= \int_0^1 \cdots \int_0^1 \phi_n(z_1(u_1),\ldots,z_n(u_n);\Omega)|det(J)|\, du_1 \ldots du_n \\ &= \prod_{i=1}^{n} f(y_i;\lambda). \int_0^1 \cdots \int_0^1 \frac{\phi_n(z_1(u_1),\ldots,z_n(u_n);\Omega)}{\prod_{i=1}^{n}\phi(z_i(u_i))} du_1 \ldots du_n \\ &= \prod_{i=1}^{n} f(y_i;\lambda). \int_0^1 \cdots \int_0^1 |\Omega|^{-\frac{1}{2}} exp\left(-\frac{1}{2}z^T(u)(\Omega^{-1} - I_n)z(u)\right) du \end{aligned}$$

که در این‌جا $z(u) = (z_1(u_1),\ldots,z_n(u_n))^T$ است. پس انتگرال (۱۱.۳) را می‌توان به‌صورت زیر بازنویسی کرد

$$L(\eta;y) = \prod_{i=1}^{n} f(y_i;\lambda). \int_0^1 \cdots \int_0^1 |\Omega|^{-\frac{1}{2}} exp\left(-\frac{1}{2}z(u)^T(\Omega^{-1} - I_n)z(u)\right) du \qquad (۱۳.۳)$$

۱.۳.۳ محاسبه تابع درستنمایی

در حالت غیر پیوسته، درستنمایی برحسب انتگرال احتمال گاوسی ۶.۳ بیان شده است. هرگاه این انتگرال همانند مطالعات طولی با مشاهدات کمتر به‌ازای هر موضوع، به ترم‌های با بعد پایین تجزیه شود، ما پیشنهاد می‌کنیم که از تقریب‌های معین قطعی و دقیق همانند روش جو (۱۹۹۵) یا روش‌های عددی اخیر توسط میوا و همکاران (۲۰۰۳) و کریگ (۲۰۰۸) استفاده کنید. الگوریتم‌های جو و میوا هردو به‌صورت عمومی از طریق R بسته‌های *mprobit* در دسترس هستند، که به‌ترتیب توسط جو، چو و ژانگ (۲۰۰۱) بسته *mvtnorm* تألیف شده‌اند.

برای ابعاد بزرگ‌تر که با سری‌های‌زمانی، داده‌های‌فضایی و مطالعات طولی طولانی‌تر همراه است، هزینه‌های محاسباتی تقریب‌های قطعی برای استفاده عملی بسیار زیاد است. از این‌رو باید روش‌های تصادفی مورد توجه قرار گیرد. به‌عنوان مثال روش شبه‌مونت‌کارلو تصادفی از گنز و برتز (۲۰۰۱) موجود است که در R بسته *mvtnorm* که در بالا ذکر شد گنجانده شده است. این روش عددی دارای هدف کلی است و ممکن است برای مدل‌های خاصی که در این پایان‌نامه در نظر گرفته می‌شود کارآمد نباشد. به‌همین دلیل، در بخش بعدی ما یکی دیگر از تقریب‌های مونت‌کارلوی درستنمایی را برای پاسخ‌های غیرپیوسته توصیف می‌کنیم، که با کلاس مدل $GCMR$ متناسب است.

۲.۳.۳ درستنمایی شبیه‌سازی شده جوک‌ـ هاجیواسیلیو‌ـ کین

ماساروتو و وارین (۲۰۱۲) متوجه شدند که شباهت بین تابع درستنمایی ۱۲.۳ و تابع مدل‌های پروبیت چندمتغیره وجود دارد به همین دلیل الگوریتم پیشنهادی توسط جوک (۱۹۹۱)، هاجیواسیلیو و همکاران (۱۹۹۶) و کین (۱۹۹۴) را برای تقریب زدن احتمالات مستطیل نرمال چندمتغیره انتخاب کردند. این مورد براساس شبیه‌ساز جوک‌ـ هاجیواسیلیو و کین[۴] مشخص شده و به‌آن روش GHK نامیده می‌شود. انتگرال (۱۱.۳) توسط نمونه‌گیری نقاط مهم تقریب زده میشود با استفاده از چگالی نمونه‌گیری نقاط مهم با تکیه‌گاه

$$(\zeta(y_1-1),\zeta(y_1))\times\cdots\times(\zeta(y_n-1),\zeta(y_n))$$

توسط $g_\eta(z)=\prod_{i=1}^n p_\eta(z_i|z_{i-1},\ldots,z_1;y_i)$ فرض می‌شود که در این‌جا

$$p_\eta(z_i|z_{i-1},\ldots,z_1;y_i)$$

چگالی شرطی $Z(s_i)$ با توجه به $Z(s_{i-1}),\ldots,Z(s_1)$ است، در زمانی‌که $Y_i=y_i$ و $Z(\cdot)$ میدانی تصادفی گاوسی هستند که پس از (۱۰.۳) تعریف می‌شوند. از آن‌جا که

$$Z(s_i)|Z(s_{i-1}),\ldots,Z(s_1)\sim N(m_i,v_i^2)$$

با

$$m_i=E_\vartheta(Z(s_i)|Z(s_{i-1}),\ldots,Z(s_1))=m_i(Z(s_{i-1}),\ldots,Z(s_1);\vartheta)$$

و

$$v_i^2=var_\vartheta(Z(s_i)|Z(s_{i-1}),\ldots,Z(s_1))=v_i^2(\vartheta)$$

به‌دنبال (۱۰.۳) مشخص می‌شود که $p_\eta(z_i|z_{i-1},\ldots,z_1;y_i)$ چگالی توزیع $N(m_i,v_i^2)$ است که بریده شده و به

$$(\zeta(y_i-1;\lambda),\zeta(y_i;\lambda))$$

تبدیل شده است. سپس روش GHK را توسط

$$\begin{aligned}\hat{L}^{GHK}(\eta;y)&=\frac{1}{M}\sum_{k=1}^M\frac{\phi_n(Z^{(k)};\Omega)}{g_\eta(Z^{(k)})}\\&=\frac{1}{M}\sum_{k=1}^M\left\{\prod_{i=1}^n\frac{p_\theta(Z_i^{(k)}|Z_{i-1}^{(k)},\ldots,Z_1^{(k)})}{p_\eta(Z_i^{(k)}|Z_{i-1}^{(k)},\ldots,Z_1^{(k)};y_i)}\right\}\\&=\frac{1}{M}\sum_{k=1}^M\left\{\prod_{i=1}^n\left[\Phi\left(\frac{\zeta(y_i;\lambda)-m_{ki}}{v_{ki}}\right)-\Phi\left(\frac{\zeta(y_i-1;\lambda)-m_{ki}}{v_{ki}}\right)\right]\right\}\end{aligned}\qquad(۱۴.۳)$$

[۴] Geweke-Hajivassiliou-Keane

تقریب می‌زند که در این‌جا $Z^1, \ldots, Z^{(M)}$ متغیرهای تصادفی پیش‌بین با توزیع یکسان به‌دست آمده از $g_\eta(z)$ هستند همراه با $Z^{(K)} = (Z_1^{(k)}, \ldots, Z_n^{(k)})$ و $m_{ki} = m_i(Z_{i-1}^{(k)}, \ldots, Z_1^{(k)}; \theta)$ و $v_{ki}^2 = v_i^2(\theta)$. شبیه‌سازی هر $Z^{(k)}$ متعاقبا انجام می‌شود، $Z_1^{(k)}, Z_2^{(k)}, \ldots, Z_n^{(k)}$ ، با استفاده الگوریتم استاندارد برای شبیه‌سازی از توزیع‌های نرمال بریده شده: برای $k = 1, \ldots, M$ و $i = 1, \ldots, n$ در نظر بگیرید

$$Z_i^{(k)} = m_{ki} + v_{ki}\Phi^{-1}\left[(1-U_{ki})\Phi\left(\frac{\zeta(y_i-1;\lambda)-m_{ki}}{v_{ki}}\right) + U_{ki}\Phi\left(\frac{\zeta(y_i;\lambda)-m_{ki}}{v_{ki}}\right)\right]$$

که در این‌جا U_{ki} متغیرهای تصادفی پیش‌بین با توزیع unif(0,1) هستند. روش GHK در بسته $gcmr$ اجرا شده است (ماساروتو و وارین ۲۰۱۶). هاجیواسیلیو و همکاران (۱۹۹۶) روش GHK را با چند روش دیگر برای تقریب‌زنی احتمال‌های مستطیل نرمال چندمتغیره مقایسه نمود و متوجه شد که روش GHK کارایی بیشتری دارد.

تذکر ۱: روش GHK نیازمند مشخص کردن ترتیب مکان‌های نمونه‌گیری است. اگرچه یک ترتیب طبیعی در مورد داده‌های سری‌زمانی وجود دارد، هیچ ترتیب طبیعی در مورد داده‌های‌فضایی وجود ندارد پس هر انتخابی تا حدی اختیاری خواهد بود. در اکتشافات عددی متوجه شدیم زمانی‌که مکان‌های نمونه‌گیری هیچ تئوری منظمی ندارند، پارامتر می‌تواند حساسیت اندک نمایش داده شده را برای انتخاب ترتیب برآورد کند. از سوی‌دیگر، زمانی‌که مکان‌های نمونه‌گیری روی یک تئوری غیرمنظم هستند یک موقعیت مرسوم با داده‌های زمین‌آماری، پارامتر می‌تواند حساسیت نمایش داده شده بیشتری را برای برخی ترتیب‌های انتخابی برآورد کند اما نه برای همه، به‌ویژه با در نظر گرفتن برآوردهای پارامتر کوواریانس. براساس تجربه محدودی که داریم، حدس می‌زنیم حساسیت برآوردهای پارامتر به‌ترتیب انتخابی مکان‌های نمونه‌گیری اندک است و ممکن است با افزایش اندازه نمونه کاهش پیدا کند.

۳.۳.۳ نمونه‌گیری نقاط مهم

تعریف ۱.۳.۳. در آمار، نمونه‌گیری نقاط مهم یک روش کلی برای براورد خواص یک توزیع خاص است، در حالی‌که تنها نمونه‌هایی از توزیع‌های مختلف از توزیع مورد نظر تولید می‌شود. این روش را به‌این دلیل که تمرکز نمونه‌گیری بر نواحی چگال‌تر(و در نتیجه انتخاب نقاط مهم‌تر) قرار دارد نمونه‌گیری نقاط مهم نامیده‌اند.

عبارت ۸.۳ تقریب مونت‌کارلو ساده زیر از درستنمایی را پیشنهاد می‌دهد:

۱. برای $k = 1, \ldots, K$

(a) n متغیر از توزیع یکنواخت و مستقل $(0,1)$ ایجاد کنید $u_1^{(k)}, \ldots, u_n^{(k)}$؛

(b) خطاهای تصادفی $\epsilon_i^{(k)} = \epsilon(u_i^{(k)})$ از معادله ۷.۳ قابل محاسبه است،

(c) چگالی مفصل‌گاوسی را به‌شیوه زیر محاسبه کنید

$$q^{(k)}(\boldsymbol{\epsilon};\boldsymbol{\theta}) = \frac{p(\epsilon_1^{(k)},\ldots,\epsilon_n^{(k)};\boldsymbol{\theta})}{p(\epsilon_1^{(k)};\boldsymbol{\lambda}) \times \cdots \times p(\epsilon_n^{(k)};\boldsymbol{\lambda})}; \qquad (۱۵.۳)$$

۲. درستنمایی را با استفاده از معادله زیر برآورد کنید

$$\tilde{L}(\boldsymbol{\theta};y) = L_{ind}(\lambda;y)\frac{1}{\mathrm{K}}\sum_{k=1}^{K} q^{(k)}(\epsilon;\boldsymbol{\theta}). \qquad (۱۶.۳)$$

متأسفانه، نتیجه این برآورد تقریباً ناکارآمد است. درواقع، تقریب نمونه‌گیری نقاط مهم درستنمایی را در نظر بگیرید

$$L^{IS}(\boldsymbol{\theta};\mathbf{y}) = \frac{1}{\mathrm{K}}\sum_{k=1}^{K}\frac{p(y,\epsilon^{(k)};\boldsymbol{\theta})}{p^{IS}(\epsilon^{(k)}\,|y;\boldsymbol{\theta})}, \qquad (۱۷.۳)$$

که در آن $\epsilon^{(k)}$ یک بردار از خطاهای تصادفی رسم شده از توزیع نمونه‌گیری نقاط مهم $p^{IS}(|y;\theta)$ است. این بدین معنی است که برآورد ۱۶.۳ متناظر با چگالی نقاط مهم ایجادشده تحت فرض کاری است که خطاهای ϵ_i با توجه به نتایج داده‌شده y_i مستقل هستند:

$$p^{IS}(\boldsymbol{\epsilon}\,|\mathbf{y};\boldsymbol{\theta}) = \prod_{i=1}^{n}\frac{p(\epsilon_i;\boldsymbol{\lambda})}{p_i(y_i;\boldsymbol{\lambda})},$$

اگرچه معتبر است، اما این چگالی نقاط مهم به دلیل فرضیه عدم وابستگی قوی می‌تواند بسیار ناکارآمد باشد. بدیهی است که این چگالی نقاط مهم ایده‌آل می‌تواند دقیقاً برابر با چگالی مشروط باشد $p(\boldsymbol{\epsilon}\,|\mathbf{y};\boldsymbol{\theta})$، با این انتخاب، هر عبارت در جمع ۱۷.۳ دقیقاً برابر است با $p(\mathbf{y};\boldsymbol{\theta})$ و بنابراین یک حالت $(k=1)$ می‌تواند برای دستیابی به درستنمایی دقیق کافی باشد. متأسفانه، استفاده از این توزیع ایده‌آل امکان‌پذیر نیست زیرا وزن‌های نمونه‌گیری نقاط مهم وابسته به $p(\mathbf{y};\boldsymbol{\theta})$ هستند.

با توجه به این‌که یک حالت از فرم ایده آل $p(\boldsymbol{\epsilon}\,|\mathbf{y};\boldsymbol{\theta})$ را می‌توان با نمونه‌گیری دنباله‌ای از

$$p_i(\epsilon_i\,|y_i\,,\ldots,y_1;\boldsymbol{\theta}), \quad i=1,\ldots,n,$$

به دست آورد، ما این چگالی نقاط مهم غیرکنترل‌پذیر را با نمونه‌گیری دنباله‌ای از چگالی زیر جایگزین می‌کنیم

$$p_i(\epsilon_i\,|y_i,\epsilon_{i-1}\,,\ldots,\epsilon_1;\boldsymbol{\theta}), \quad i=1,\ldots,n, \qquad (۱۸.۳)$$

برای حالت خاص رگرسیون پروبیت چند متغیره، چگالی نمونه‌گیری نقاط مهم فوق متناظر با شبیه‌ساز (GHK) است، برای مثال به کین (۱۹۹۴) رجوع کنید.

بسط شبیه‌ساز GHK برای عمل کردن در $GCMR$ بی‌واسطه است. تحت فرضیه‌های مدل، $p(\epsilon_i\,|\epsilon_{i-1}\,,\ldots,\epsilon_1;\theta)$ یک چگالی نرمال با میانه $m_i = \mathrm{E}(\epsilon_i\,|\epsilon_{i-1}\,,\ldots,\epsilon_1;\theta)$و واریانس

$$v_i^2 = \mathrm{var}(\epsilon_i\,|\epsilon_{i-1}\,,\ldots,\epsilon_1)$$

است. بنابراین، (۱۸.۳) یک چگالی نرمال بریده‌شده بر بازه $\mathfrak{D}_i(y_i;\boldsymbol{\lambda})$ است و یک حالت از آن از طریق جایگذاری‌های زیر به‌دست می‌آید

$$\epsilon_i(u_i) = m_i + v_i\phi^{-1}\left\{(1-u_i)\,a_i + u_i b_i\right\}, \quad i = 1,\ldots,n, \tag{۱۹.۳}$$

که در آن، $u_1,\ldots,u_n$، n برابر با n حالت مستقل از یک متغیر تصادفی یکنواخت $(0,1)$ هستند و

$$a_i = \Phi\left[\frac{\Phi^{-1}\left\{F_i(y_i^-;\boldsymbol{\lambda})\right\} - m_i}{v_i}\right], b_i = \Phi\left[\frac{\Phi^{-1}\left\{F_i\left(y_i;\boldsymbol{\lambda}\right)\right\} - m_i}{v_i}\right], \quad i = 1,\ldots,n.$$

بنابراین، الگوریتم نمونه‌گیری دنباله‌ای نتیجه‌گیری شده برای برآورد کردن درستنمایی برابر است با

۱. برای $k = 1,\ldots,K$

(a) n متغیر $(0,1)$ مستقل یکنواخت تولید می‌کند $u_1^{(k)},\ldots,u_n^{(k)}$

(b) خطاهای تصادفی $\varepsilon_i^{(k)} = \epsilon(u_i^{(k)})$ را از معادله ۱۹.۳ محاسبه کنید

۲. درستنمایی را با استفاده از معادله زیر برآورد کنید

$$L^{IS}(\theta;y) = \frac{1}{\mathrm{K}}\sum_{k=1}^{K}\left\{\prod_{i=1}^{n}\frac{p\left(\epsilon_i^{(k)}\left|\epsilon_{i-1}^{(k)},\ldots,\epsilon_1^{(k)};\theta\right.\right)}{p_i\left(\epsilon_i^{(k)}\left|y_i\,,\epsilon_{i-1}^{(k)},\ldots,\epsilon_1^{(k)};\theta\right.\right)}\right\}.$$

چند نظر درباره جنبه‌های عددی به‌شرح زیر هستند. کمیت‌های m_i و v_i^2 را می‌توان به‌خوبی با استفاده از تجزیه چولسکی از Ω محاسبه نمود که در هر حالتی، یک مؤلفه انتگرال از محاسبات درستنمایی است. صرفه‌جویی در محاسبات اساسی از طریق استفاده از این واقعیت به‌دست می‌آید که ماتریس همبستگی خطا Ω برای تمام بردارهای خطای شبیه‌سازی شده $\epsilon^{(k)}, k = 1,\ldots,K$ مشابه است.

دیگر چگالی‌های نقاط مهم را نیز می‌توان در نظر گرفت، به‌عنوان‌مثال، با استفاده از ایده‌های دوربین و کوپمن (۲۰۰۱). بااین‌وجود، افزایش هزینه‌های محاسباتی ممکن است از لحاظ دقت عددی، با توجه به سادگی و نتایج خوب حاصل‌شده با استفاده از نمونه‌گیری پیشنهادشده، توجیه نداشته باشد. برای بررسی مطالعه شبیه‌سازی به گیدو و کریستیانو (۲۰۱۲) مراجعه کنید.

به‌طورکلی، پیچیدگی کلی برای یک تقریب درستنمایی با نمونه‌گیری نقاط مهم موردبحث، همانند دیگر تقریب‌های درستنمایی برای مدل‌های $GCMR$ از مرتبه $\mathrm{O}\left(n^3\right)$ است، که این امر به‌دلیل معکوس سازی لازم برای ماتریس همبستگی Ω است. بااین‌وجود، برای مسائل خاص هزینه محاسباتی بسیار کمتر است. به‌طور مثال، در مدل‌های سری‌های زمانی با ϵ در پی یک پروسه میانگین متحرک اتورگرسیو، تجزیه چولسکی را می‌توان به‌صورت کارآمد از طریق فیلتر کالمن پیاده‌سازی کرد و تنها محاسبات $\mathrm{O}\left(n\right)$ موردنیاز است. اگر n بزرگ باشد، هزینه محاسباتی کلی $\mathrm{O}\left(n^3\right)$ غیرعملی است.

براوردهای درستنمایی ماکزیمم از طریق درستنمایی لگاریتمی بهتر محاسبه می‌شوند. به‌هرحال، نمونه‌گیری نقاط مهم برای ارائه برآوردهای نارایب درستنمایی طراحی شده‌اند، درحالی‌که $L^{IS}\left(\theta;y\right)$

یک براوردگر با اندکی اریبی از لگاریتم درستنمایی $l(\theta; y)$ است. در بسته $gcmr$ ما، یک براوردگر نااریب از لگاریتم درستنمایی توسط دوربین و کوپمن به دست می‌آید.

اندازه مونت‌کارلو مورد نیاز برای نتایج استنباطی صحیح عموماً یک تابع از درجه جدایی است، و سخت‌ترین حالت آن پاسخ‌های دودویی است. به‌عنوان یک پیشنهاد کلی، می‌توان توصیه کرد که آنالیز را برای بررسی کفایت اندازه مونت‌کارلو تکرار کرد، برای مثال، با شروع از یک اندازه مونت‌کارلو کوچک و سپس افزایش آن تا زمانی که تفاضل در برآوردهای پارامتری عملاً ناچیز بشوند.

۴.۳ آزمون مشخص‌سازی هاسمن

در اکثر حالت‌ها، یک علاقه علمی به پارامترهای حاشیه‌ای λ یا یک زیرمجموعه از پارامترهای رگرسیون β وجود دارد. مرزهای مفروض را می‌توان از طریق یک طیف گسترده از روش‌های عددی و گرافیکی توسعه‌یافته بررسی کرد، و به‌این‌ترتیب، مشخصات دقیق درستنمایی مستقل به‌صورت کلی به دست می‌آید. این امر منجر به یک برهان قاطع دووجهی از استنباط مبنا برروی λ در یک درستنمایی مستقل امن اما ناکارآمد و یا در نظر گرفتن یک مدل کامل اما با ریسک وجود مشخصات نادرست مفصل می‌شود. به‌عبارت‌دیگر، ما می‌خواهیم فرضیه صفر زیر را در نظر بگیریم

H_0: مدل چند متغیره مفروض به‌درستی مشخص شده است.

در مقابل

H_1: توزیع‌های حاشیه‌ای صحیح هستند اما توزیع نرمال چند متغیره مفروض برای خطاها صحیح نیست (مفصلی نادرست).

برآوردگر درستنمایی مستقل λ_{int} تحت هر دو فرضیه صفر و جایگزین سازگار است، درحالی‌که براوردگر درستنمایی ماکزیمم $\hat{\theta} = \left(\hat{\lambda}^T, \hat{\tau}^T\right)^T$ تحت فرضیه صفر سازگار و کارآمد است اما تحت فرضیه جایگزین ناسازگار است. این امر تأیید کردن مشخصات مدل صحیح توسط آمار نوع هاسمن را پیشنهاد می‌کند.

$$h(Y) = \left(\hat{\lambda}_{ind} - \hat{\lambda}\right)^T D^{-1} \left(\hat{\lambda}_{\text{ind}} - \hat{\lambda}\right),$$

با واریانس $D = \text{var}\left(\lambda_{\text{ind}} - \hat{\lambda}\right)$ که تحت شرایط فرضیه صفر محاسبه شده است که در آن $h(Y)$ به‌عنوان یک متغیر تصادفی کای‌ـدو[۵] با درجه‌های آزادی $\dim\left(\hat{\lambda}\right)$ توزیع می‌شود.

توجه: چهارچوب فوق، با آزمون متداول هاسمن متفاوت است، که بر روی پارامتر کامل θ متمرکز است و نه بر روی زیرمجموعه خودش λ.

متعاقبا، نتیجه تعامد هاسمن صادق نیست، که یعنی $\text{cov}(\lambda_{ind} - \lambda, \lambda) \neq 0$ می‌باشد. به این ترتیب، D به‌صورت تفاضل واریانس $\text{var}\left(\hat{\lambda}_{\text{ind}}\right) - \text{var}\left(\hat{\lambda}\right)$ خلاصه نمی‌شود.

تحت فرضیه صفر، بردار براوردگرها $\left(\hat{\lambda}_{ind}^T, \hat{\theta}^T\right)^T$ دارای ماتریس واریانس مجانبی است

$$\text{V} = \begin{pmatrix} H_1^{-1} J_1 H_1^{-1} & H_1^{-1} J_{12} H_2^{-1} \\ H_2^{-1} J_{21} H_1^{-1} & H_2^{-1} \end{pmatrix}$$

[۵] Chi-squared

که در آن،

$$H_1 = E\left\{-\nabla^2 \ell_{ind}\left(\lambda; Y\right)\right\}$$

$$J_{12} = \operatorname{cov}\left\{\nabla \ell_{ind}\left(\lambda; Y\right), \nabla \ell\left(\theta; Y\right)\right\}$$

$$H_2 = E\left\{-\nabla^2 \ell\left(\theta; Y\right)\right\}$$

$$J_1 = \operatorname{var}\left\{\nabla \ell_{ind}\left(\lambda; Y\right)\right\}$$

$$J_{21} = J_{12}^T$$

به‌این‌ترتیب داریم $\mathrm{D} = C^T V C$ برای یک ماتریس تضاد $C = (I^T, -I^T, 0^T)$، که در آن ماتریس‌های واحد I، دارای بعد λ هستند. بطور کلی، مولفه های ماتریس V در شکل بسته در دسترس نیستند. آن‌ها را می‌توان از طریق شبیه‌سازی مونت‌کارلو از مدل مفروض برآورد کرد.

به‌صورت جایگزین، می‌توان از طریق یک اتورگرسیو پارامتری یک براورد مستقیم از D که دقیق‌تر اما دارای هزینه محاسباتی بالاتر باشد در نظر گرفت. آزمون هاسمن را می‌توان برای بررسی درستی استنباط درباره پارامتر حاشیه‌ای سرتاسری λ و همچنین برای بررسی زیرمجموعه‌های λ مانند رگرسورهای منفرد/ مجزا β_i یا ترکیبی از رگرسورها، در مطابقت با توجه علمی استفاده کرد.

علی‌رغم مطالب ارائه شده در این بخش، اعتبارسنجی کامل و درست از فرض نرمال چندمتغیره برای خطاهای ϵ، بسیار دشوار است. همچنین، طبیعی است که پرسشی در مورد استنباط درست برای λ براساس بدمشخص شدن توزیع چندمتغیره برای خطاها مطرح شود که پاسخ دادن کلی برای این پرسش، نیز کار سختی است. با این حال، گیدو و کریستیانو (۲۰۱۲) بررسی‌های شبیه‌سازی برای استواری در مقابل بدمشخص‌سازی توزیع خطا را بررسی نموده‌اند.

۵.۳ مثال‌ها

در این بخش، انعطاف‌پذیری مدل را با ارائه دو مثال کاربردی از داده‌های‌طولی و فضایی بررسی می‌کنیم. به‌منظور تسهیل مقایسه با سایر مدل‌های دیگر و روش‌های برازش، کلیه مثال‌ها مبتنی بر مجموعه داده‌های شناخته شده هستند. همچنین در بخش ۳.۵.۳ نظراتی در مورد وجوه محاسباتی ارائه شده‌اند.

۱.۵.۳ داده‌های مربوط به زمان تا پیشامد

برای نشان دادن آنالیز داده‌های‌بقا، داده‌های مانتل (۱۹۷۷) و همکاران را در مورد دفعات بروز تومور در موش‌ها مجددا آنالیز کردیم. سه جفت موش از بین ۵۰ موش هم شکم/ زایمان مشاهده قرار گرفتند. یک جفت از موش‌ها تحت درمان قرار گرفت درحالیکه دو جفت دیگر به‌عنوان گروه‌های کنترل در نظر گرفته شدند. پاسخ‌ها شامل هفته وقوع تومور یا هفته مرگ پیش از شدت گرفتن هر تومور می‌باشد. در هر مورد، کلیه موش‌ها پس از ۱۰۴ هفته از قربانی شدند. در این بررسی به‌لحاظ علمی ارزیابی ارتباط احتمالی بین درمان‌ها و زمان وقوع تومور مورد توجه قرار گرفته است.

جدول ۱.۳: داده‌های جفت‌گیری سوسمارها: براورد احتمال جفت‌گیری و فاصله اطمینان ٪۹۵ از مدل حاشیه‌ای در مقایسه با برآوردهای ارائه شده توسط زگر و کریم

	تابستان		پاییز	
احتمال‌های واقعی	برآورد	فاصله اطمینان ۹۵%	برآورد	فاصله اطمینان ۹۵%
π_{RR}	۰/۷۳	$(0.60, 0.83)$	۰/۶۴	$(0.52, 0.74)$
π_{RW}	۰/۶۳	$(0.49, 0.75)$	۰/۵۳	$(0.41, 0.64)$
π_{WR}	۰/۲۷	$(0.17, 0.41)$	۰/۲۰	$(0.12, 0.31)$
π_{WW}	۰/۷۴	$(0.62, 0.84)$	۰/۶۶	$(0.54, 0.76)$
کریم و زگر				
	تابستان		پاییز	
	est.	۹۵% CI	est.	۹۵% CI
π_{RR}	۰/۷۳	$(0.58, 0.84)$	۰/۶۴	$(0.51, 0.76)$
π_{RW}	۰/۶۲	$(0.46, 0.76)$	۰/۵۲	$(0.39, 0.65)$
π_{WR}	۰/۲۴	$(0.13, 0.41)$	۰/۱۸	$(0.10, 0.28)$
π_{WW}	۰/۷۳	$(0.58, 0.84)$	۰/۶۴	$(0.51, 0.76)$

قرار دهید $\mathrm{Y}_i = \min\{\mathrm{T}_i, c_i\}$ که نشان‌دهنده پاسخ با T_i در زمان وقوع تومور می‌باشد و c_i زمان سانسور برای $i = 1, \ldots, 150$ است. مدل رگرسیون وایبل را برای زمان بقا T_i فرض کردیم که $\mathrm{F}_i(t_i; \lambda) = 1 - \exp\{-(t_i/\eta_i)^\alpha\}$ است، که α نشان‌دهنده پارامتر شکل، $\eta_i = \exp(\beta_0 + \beta_1 x_i)$ به‌همراه X_i به‌عنوان شاخص درمان و $\lambda = (\alpha, \beta_0, \beta_1)^T$ است. این مدل متناظر با مدل عمر شتاب‌دهنده و مدل خطر متناسب است.

زمان‌های بقا برای موش‌های مختلف در صورتی مستقل در نظر گرفته می‌شدند که موش‌ها هم‌شکم نبودند و زمانی وابسته بود که آن‌ها هم‌شکم بودند. به‌این‌ترتیب ماتریس همبستگی خطاها توسط حاصل‌ضرب کرونکر، مدل‌سازی شدند $\Omega = \mathrm{I}_{50} \otimes \Omega_1$ که در اینجا Ω_1 یک ماتریس همبستگی تبادل‌پذیر با پارامتر همبستگی یکسان T می‌باشد. درستنمایی، حاصل‌ضرب سهم از شکم‌های مختلف است. هریک از آن‌ها یک چگالی سه‌متغیره هستند که شکل متفاوتی را فرض می‌کنند که به تعداد مشاهدات سانسور شده بستگی دارند. برای ساده کردن نمادین، سه‌متغیره اول $p(y_1, y_2, y_3; \theta)$ را در نظر بگیرید. احتمالات زیر را داریم:

۱. هیچ مشاهده سانسورشده‌ای وجود ندارد،

$$p(y_1, y_2, y_3; \theta) = p_1(t_1; \lambda)\, p_2(t_2; \lambda)\, p_3(t_3; \lambda)\, q(\epsilon_1, \epsilon_2, \epsilon_3; \theta),$$

در اینجا $p_i(0; \lambda)$ تابع چگالی مدل رگرسیون وایبل است.

۲. یک مشاهده سانسورشده وجود دارد، مثلا مورد سوم،

$$\begin{aligned} p\left(y_1, y_2, y_3; \theta\right) &= \int_{c_3}^{\infty} p\left(t_1, t_2, t_3; \theta\right) \mathrm{dt}_3 \\ &= p_1\left(t_1; \lambda\right) p_2\left(t_2; \lambda\right) q_{12}\left(\epsilon_1, \epsilon_2; \theta\right) \int_{\phi^{-1}\{\mathrm{F}_3(c_3;\lambda)\}}^{\infty} p\left(\epsilon_3 \left| \epsilon_1 \right., \epsilon_2; \theta\right) \mathrm{d}\, \epsilon_3 \end{aligned}$$

که در آن $p\left(\epsilon_3 \left| \epsilon_1 \right., \epsilon_2; \theta\right)$ چگالی متغیر نرمال با میانگین $\tau/(1+\tau)\left(\varepsilon_1+\varepsilon_2\right)$ و واریانس $1-2\tau^2/(1+\tau)$ است.

۳. دو مشاهده سانسورشده وجود دارد، مثلا دو مورد آخر را در نظر بگیرید،

$$\begin{aligned} p\left(y_1, y_2, y_3; \theta\right) &= \int_{c_2}^{\infty} \int_{c_3}^{\infty} p\left(t_1, t_2, t_3; \theta\right) \mathrm{d}\, t_2\, \mathrm{d}\, t_3 \\ &= p_1\left(t_1; \lambda\right) \int_{\phi^{-1}\{\mathrm{F}_2(c_2;\lambda)\}}^{\infty} \int_{\phi^{-1}\{\mathrm{F}_3(c_3;\lambda)\}}^{\infty} p\left(\epsilon_2, \epsilon_3 \left| \epsilon_1 \right.; \theta\right) \mathrm{d}\, \epsilon_2\, \mathrm{d}\, \epsilon_3; \end{aligned}$$

که در آن $p(\epsilon_2, \epsilon_3 \left| \epsilon_1 \right.; \theta)^{\mathrm{T}}$ چگالی نرمال دومتغیره با بردار میانگین $(\tau\epsilon_1, \tau\epsilon_1)$ و ماتریس واریانس $(1-\tau)\begin{pmatrix} 1+\tau & \tau \\ \tau & 1+\tau \end{pmatrix}$ است.

۴. کلیه مشاهدات سانسورشده باشند

$$\begin{aligned} p\left(y_1, y_2, y_3; \theta\right) &= \int_{c_1}^{\infty} \int_{c_2}^{\infty} \int_{c_3}^{\infty} p\left(t_1, t_2, t_3; \theta\right) \mathrm{d}\, t_2\, \mathrm{d}\, t_3 \\ &= \int_{\phi^{-1}\{\mathrm{F}_1(c_1;\lambda)\}}^{\infty} \int_{\phi^{-1}\{\mathrm{F}_2(c_2;\lambda)\}}^{\infty} \int_{\phi^{-1}\{\mathrm{F}_3(c_3;\lambda)\}}^{\infty} p\left(\epsilon_1, \epsilon_2, \epsilon_3; \theta\right) \mathrm{d}\, \epsilon_1\, \mathrm{d}\, \epsilon_2\, \mathrm{d}\, \epsilon_3 \end{aligned}$$

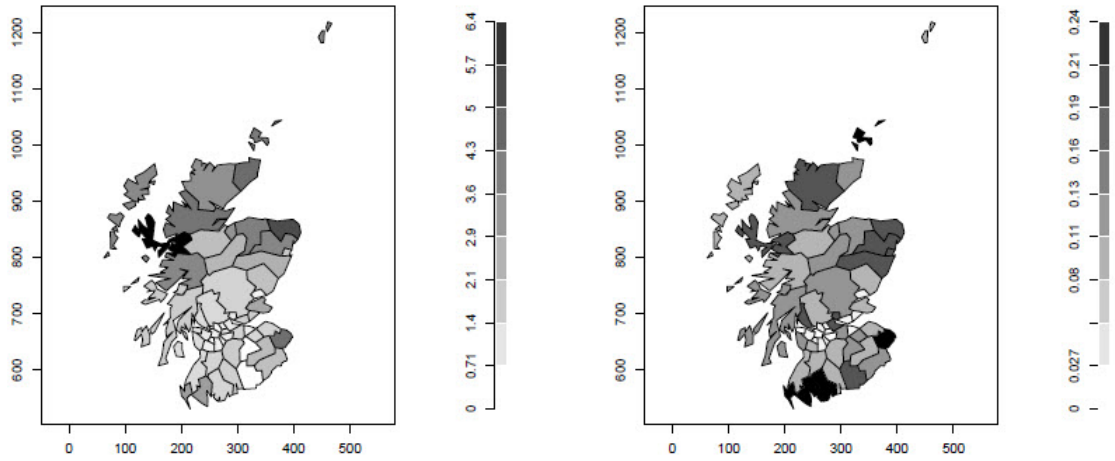
در اینجا $p\left(\epsilon_1, \epsilon_2, \epsilon_3; \theta\right)$ چگالی نرمال سه‌متغیره با میانگین صفر و ماتریس واریانس Ω_1 است.

بنابراین، محاسبات درستنمایی نیازمند تقریب احتمالات نرمال راست‌گوشه با بعد دو و سه است. برای چنین انتگرال‌هایی با بعد کم،به‌کارگیری الگوریتم نمونه‌گیری نقاط مهم که در بخش ۳.۳.۳ گفته شد ضروری نیست. در عوض، از تقریب قطعی دقت که توسط الگوریتم میوا (۲۰۰۳) و همکاران ارائه شده استفاده کردیم. پارامترهای حاشیه‌ای برآوردزده شده عبارتند از $\hat{\alpha} = -3.79\,(0.55)$ ، $\hat{\beta}_0 = 4.98\,(0.08)$ و $\hat{\beta}_1 = -0.24\,(0.09)$؛ دومین عامل تأثیر درمان است. براورد پارامتر $\hat{\tau} = 0.53\,(0.15)$ با توجه‌به ویژگی ۲ در ویژگی‌های وابستگی۲.۵.۲ وجود رابطه مثبت بین زمان بقای موش‌ها از یک شکم را تایید می‌کند.

۲.۵.۳ رگرسیون فضایی با داده‌های شمارشی

داده‌ها در مورد بروز سرطان لب در مردان اسکاتلندی طی سال‌های ۱۹۷۵ تا ۱۹۸۰ توسط محققان مختلف برای نشان دادن روش‌های مختلف نقشه برداری/نگاشت بیماری مورد بررسی و تحلیل قرار

گرفت، برای مطالعه جزیئات بیشتر به والر و گاتوی (۲۰۰۴) و ویکفیلد (۲۰۰۷) مراجعه کنید. داده‌ها شامل Y_i مشاهده شده و مورد انتظار در هر ۵۶ شهر اسکاتلند است، از طریق وب‌سایت $www.sph.emory.edu$ می‌توانید به اطلاعات بیشتری دست پیدا کنید.



شکل ۱.۳: داده‌های سرطان اسکاتلند: SMR (پنل چپ) و AFF (پنل راست) نقشه ها.

مدل استاندارد غیرفضایی برای این داده‌ها شامل این فرض است که موارد مشاهده شده Y_i به‌عنوان متغیر پواسن با میانگین $\mu_i = \phi_i e_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$ توزیع شده‌اند که در اینجا x_{1i} متغیر کمکی AFF و x_{2i} مختصات عرضی (تقسیم‌شده با ۱۰۰) است. مقدار ϕ_i، پارامتر بیش‌پراکنش به عنوان یک متغیر گاما با میانگین ۱ و مقیاس k توزیع شده است. به‌عبارت دیگر، یک مدل حاشیه‌ای دوجمله‌ای منفی برای Y_i فرض شده است.

وابستگی فضایی باقی‌مانده با فرض اینکه خطاهای ϵ_i با صفر پیوسته میانگین مدل میدان‌تصادفی گاوسی مدل‌سازی می‌شود تحقق می‌یابد.

براساس فرض توزیع‌فضایی یکنواخت در هر شهر، همبستگی بین خطاهای شهر i و شهر j به‌واسطه میانگین ارائه می‌شوند

$$\operatorname{corr}(\epsilon_i, \epsilon_j) = \frac{1}{|A_i|\,|A_j|} \int_{A_i} \int_{A_j} \rho\left(\|s_i - s_j\|_2; \tau\right) \mathrm{d}s_i ds_i, \qquad (۲۰.۳)$$

که در اینجا A_i نشان‌دهنده i امین شهر است، $|A_i|$ نشان‌دهنده منطقه و $\rho(\cdot;\tau)$ یک تابع همبستگی فضایی است. با در نظر گرفتن انعطاف‌پذیری به تابع همبستگی‌فضایی ماترن که در معادله ۵.۲ تعریف شده است می‌پردازیم.

آنالیز تقریبی $\operatorname{corr}(\epsilon_i, \epsilon_j)$ را با تابع همبستگی فضایی بین دو حوزه آغاز می‌کنیم،

$$\operatorname{corr}(\epsilon_i, \epsilon_j) \approx \rho\left(\|\tilde{s}_i - \tilde{s}_j\|_2\right) \qquad (۲۱.۳)$$

و $\tilde{s}_i$نشان دهنده منطقه i امین شهر است. دیگل و ریبیرو (۲۰۰۷) تایید کردند که شکل پارامتر τ_2 تابع‌همبستگی ماترن به‌سختی قابل شناسایی است و پیشنهاد دادند که مقدار آن را از مجموعه مجزای ۰.۵،۱.۵،۲.۵ انتخاب کنیم که نشان‌دهنده سطوح مختلف تقسیم‌بندی میانگین مربع فرایند سیگنال زیر است. برای این داده‌ها، تفاوت اندکی بوسیله تفاوت در شکل پارامتر τ_2 به‌دست آوردیم. در ادامه، براوردهای به‌دست آمده از مدل با $\tau_2 = 0.5$ ارائه می‌شوند، که متناظر با تابع همبستگی نمایی $\text{corr}\,(\epsilon_i, \epsilon_j) = \exp(-\|\hat{s}_i - \hat{s}_j\|_2/\tau_1)$ است. پارامترهای حاشیه‌ای براورد شده، پاسخ میانگین را به‌دست می‌آورند

$$\text{E}\,(\text{Y}_i\,|x_{1i}, x_{2i}\,) = e_i \exp\left(-20.80_{(4.58)} + 4.31_{(1.43)}x_{1i} + 36.74_{(8.06)}x_{2i}\right)$$

و پارامتر غیرفضایی ϕ_i به‌عنوان متغیر گاما با میانگین یک و پارامتر مقیاس براورد شده $\hat{\kappa} = 0.17\,(0.06)$ است. متغیر کمکی $\text{AFF}\,(x_1)$ به‌شکل معنادار و مثبتی با تعداد بیش‌از حد وقوع تومور ارتباط دارند. ضریب رویکرد فضایی براوردشده (x_2) نیز اهمیت زیادی دارد اگرچه این ارتباط باید با دقت زیادی تفسیر شود زیرا بیشتر مناطق شمالی اسکاتلند دارای جمعیت پراکنده (کم‌جمعیت) هستند که منجر به امید ریاضی صفر می‌شود.

شواهدی از برخی از همبستگی‌های فضایی باقی‌مانده محلی نیز وجود دارد. براورد پارامتر همبستگی تابع همبستگی نمایی $Km\,(6.19Km)$ 14.36 است.

همچنین دو شهر که دورتر از $3 \times 14.36 = 43.08Km$ باشند دارای خطایی با همبستگی پایین‌تر از 0.05 هستند. ۱۷۵ جفت شهر مجزا با مناطق مختلف در فاصله‌ای کمتر از ۴۳ کیلومتر با $56!\,(2!54!) =$ 1.540 جفت مشاهده شده اند. با توجه‌به ویژگی ۱.۲ موضوع وابستگی‌فضایی باقی‌مانده در پاسخ‌ها، هنگام وجود روند جنوب–شمال برای متغیر AFF وجود دارد.

آنالیز دقیق‌تری باید با درنظر گرفتن جغرافیای اسکاتلند صورت بگیرد که نیازمند محاسبه همبستگی بین خطاهای تولیدشده است که با یک فرایند پیوسته در فرمول ۲۰.۳ صورت می‌گیرد. بااین‌حال این آنالیز دقیق‌تر تا حدودی غیرضروری است زیرا وابستگی ضعیف متغیر کمکی $AIFF$ و روند فضایی نیز باید لحاظ شود.

۳.۵.۳ جزئیات محاسباتی

مدل ارائه‌شده در ۲.۵.۳ با الگوریتم GHK متناسب هستند. مدل در ابتدا با k=۱۰۰ تکرار مونت‌کارلو، با استفاده از ماکزیمم درستنمایی استقلال براوردها به عنوان مقادیر اولیه برای پارامترهای حاشیه‌ای مطابقت دارد. بنابراین، مدل مجددا با مقدار بزرگتر تکرار $k = 1000$ با استفاده از براوردهای اولیه مقادیر آغازین برازش شد. تجربه نشان می‌دهد $k = 1000$ یک اندازه کافی برای به دست آوردن براوردهای پایدار ثابت آماری با داده‌ها که به چند صد مشاهده تنظیم شده است. زمان محاسباتی مورد نیاز برای برازش مدل‌های مختلف به عوامل متفاوتی بجز مشاهدات بستگی دارد. از جمله این عوامل مهم عبارتند از تعداد پارامترهای وابسته، درجه وابستگی و سطح مجزا بودن پاسخ‌ها. به‌عنوان مثال، با نوت‌بوک، مک‌بوک ایر اپل که مجهز به پردازنده $1.8Ghz\ Intel\ Core\,i7$ و مموری ۴ گیگ است

ما تنها به ۰.۰۸ ثانیه زمان نیاز داریم تا مدل فضایی با داده‌های سرطان لب در اسکاتلند برازش شود که شامل ۵۶ مشاهده است که تنها یک پارامتر وابسته دارد و درجه وابستگی بسیار پایین است. زمان محاسبه برای برازش مدل $ARMA(2,1)$ با داده‌های فلج اطفال در حدود ۱.۲ دقیقه است. این زمان طولانی‌تر به‌دلیل بزرگ‌تر بودن اندازه نمونه $(n = 168)$ است، حضور سه پارامتر وابستگی و وابستگی سریالی معنادار است. تحلیل داده‌های جفت‌گیری سوسمارها به دلیل مدل برازش خاص آن به ۴.۷۸ ثانیه زمان نیاز دارد.

فصل ۴

ارزیابی مدل با شبیه‌سازی و مثال‌های کاربردی

۱.۴ مقدمه

در این فصل بسته gcmr برای مدل‌های رگرسیون حاشیه‌ای مفصل گاوسی را در نرم‌افزار R بررسی می‌کنیم. همانطور که در فصل‌های قبل گفته شد، مفصل گاوسی یک چهارچوب ریاضی مناسب برای به‌کارگیری انواع مختلف وابستگی در مدل‌های رگرسیون ارائه می‌دهد. به‌عنوان مثال، در سری‌زمانی، مطالعات‌طولی و داده‌های‌فضایی. بسته gcmr برای ارائه مشخصات خاصی از مدل رگرسیون و ساختار وابستگی طراحی شده است. مثال‌های ارائه‌شده شامل مدل‌سازی دوگانه منفی داده‌های شمارشی‌طولی، رگرسیون بتا برای سری‌زمانی نرخ‌ها و رگرسیون لجستیک برای داده‌های دوجمله‌ای همبستگی‌فضایی است.

۲.۴ اجرا مثال‌های کاربردی در R

تابع اصلی بسته gcmr دستور gcmr است که اجازه می‌دهد مدل‌های مفصل گاوسی، با درست‌نمایی ماکزیمم، در حالت پیوسته و با درست‌نمایی شبیه‌سازی شده ماکزیمم، در حالت گسسته، اندازه‌گیری شوند. آرگومان‌های ورودی تابع gcmr به صورت زیر هستند

```
${gcmr}$(formula, data, subset, offset, marginal, cormat, start,
```

```
fixed, options = ${gcmr}$.options(...), model = TRUE, ...)
```

تابع دارای آرگومان‌هایی استاندارد برای چهارچوب مشخصات مدل است (چامبرز و هاستین ۱۹۹۳ را ببنید). از قبیل: آرگومان (Formula)، احتمال محدود کردن تجزیه و تحلیل به‌یک زیرمجموعه (subset) از داده‌ها (data) است. همچنین به‌منظور تعیین تضاد برای عوامل، آرگومان (offset) تنظیم می‌شود.

آرگومان‌های ویژه در دستور (gcmr) شامل دو آرگومان کلیدی (marginal) و (cormat) هستند که به‌ترتیب بخش حاشیه‌ای مدل و ماتریس همبستگی مفصل را مشخص می‌کنند. در نهایت، سه آرگومان اختیاری نیز وجود دارد که برای تعیین کردن مقادیر شروع (start)، مقادیر بعضی از پارامترها (fixed) و تنظیم گزینه‌های برازش (option) در نظر گرفته شده‌اند. در ادامه این بخش، مولفه‌های دستور (gcmr) و روش‌های وابسته به آن را توصیف می‌کنیم.

۱.۲.۴ فرمول‌های دو بخشی

فرمول پایه‌ای در دستور gcmr، از نوع $y \sim x1 + x2$ است و رگرسیون مدل را برای پاسخ میانگین رابطه ۱.۳.۴ با تابع لینک $g_1(\cdot)$ که در آرگومان marginal تعریف شده است (همان‌طور که در ادامه توضیح داده می‌شود)، مشخص می‌کند. در ادامه اجرای رگرسیون بتا در بسته betareg (کریباری‌ـ نتو و زلیس(۲۰۱۰)) می‌توان گفت: بسته $gcmr$ امکان تشخیص یک مدل رگرسیون مرتبه دوم را برای پراکندگی، از طریق یک فرمول دوبخشی از نوع $y \sim x1 + x2 \mid z1 + z2$ فراهم می‌کند و این عمل را با استفاده از عوامل موروثی بسته Formula (زلیس و کراسنت(۲۰۱۰))

۲.۲.۴ مشخصات مدل حاشیه‌ای

مدل حاشیه‌ای $F(\cdot \mid x_i)$ به‌وسیله‌ی یک شی از کلاس $gcmr$ مشخص و در آرگومان marginal در تابع gcmr تنظیم می‌شود. توزیع‌های حاشیه‌ای موجود در $gcmr$، بتا، دوجمله‌ای، گاما، گوسی، دوجمله‌ای‌منفی، پواسون و وایبول هستند (برای هریک‌از این توزیع‌ها جدول ۳.۲.۴ را ببینید). می‌توان مشابه آنچه که در مدل‌های خطی تعمیم‌یافته سنتی است، یک تابع لینک که وابسته به پاسخ میانگین پیش‌گویی‌کننده خطی است، انتخاب کرد. تمام توابع لینک موجود در کلاس $\texttt{glm} - \texttt{link}$ مجاز هستند. حاشیه‌ای‌های گوسی به‌منظور کامل بودن بسته، در $gcmr$ وجود دارند، اما استفاده از $gcmr$ برای برازش مدل‌های نرمال چندمتغیره که به صورت ناقص از ترکیب حاشیه‌ای‌های گوسی با یک مفصل گوسی به‌دست می‌آیند، توصیه نمی‌شود. در واقع، بسته $gcmr$ به‌منظور کار با مدل‌های مفصل گوسی با توزیع‌های حاشیه‌ای تک‌متغیره کلی طراحی می‌شود و بنابراین از نظر عددی برای استنباط در مدل‌های گوسی خطی چندمتغیره کارآمد نیست، در جایی که دسترسی به نتایج تحلیلی، تسریع قابل‌توجه محاسبات را در پی دارد.

۳.۲.۴ مشخصات ساختار همبستگی

ماتریس همبستگی P در مفصل گوسی به‌وسیله یک شی از کلاس cormat.gcmr تعیین و در آرگومان cormat در تابع *gcmr*.تنظیم می‌شود. نسخه ۰.۷.۵ *gcmr* شامل چهار ساختار همبستگی با قابلیت اجرا گسترده است (جدول ۳.۲.۴ را ببینید). امکان همبستگی استقلال کاری، در عمل مشابه معادلات برآورد تعمیم‌یافته است.

جدول ۱.۴: مدل‌های حاشیه‌ای موجود در نسخه gcmr با جزئیات تابع لینک. ستون ''پراکندگی'' توزیع‌ها را با یک پارامتر پراکندگی مشخص می‌کند

پراکندگی	توزیع	marginal.gcmr
بله	بتا	beta.marg(link = ''logit'')
نه	دوجمله‌ای	binomial.marg(link = ''logit'')
بله	گاما	Gamma.marg(link = ''inverse'')
بله	گاوسی	gaussian.marg(link = ''identity'')
بله	دوجمله‌ای منفی	negbin.marg(link = ''log'')
نه	پواسن	poisson.marg(link = ''log'')
بله	وایبل	weibull.marg(link = ''log'')

سه ساختار همبستگی دیگر امکان استفاده از سری‌های زمانی، داده‌های‌طولی یا خوشه‌ای و داده‌های خاص را فراهم می‌کند. داده‌های‌طولی و خوشه‌ای را می‌توان با cluster.cormat(id, type) تحلیل نمود که از روی توابع موروثی در بسته ulme (پینهرو و همکاران ۲۰۱۴) ساخته شده است.

ورودی‌های cluster.cormat() برداری از شی id و نوع (type) همبستگی، با امکانات

```
\mathtt{"independence"}, \mathtt{"ar1"},
\mathtt{"ma1"}, \mathtt{"exchangeable"},
\mathtt{"unstructured"}
```

هستند.

شی id برداری است به‌طول یکسان که نشان‌دهنده تعداد مشاهدات است. داده‌ها به‌صورت مرتب‌شده در نظر گرفته می‌شوند، به‌این ترتیب که، مشاهدات از شی (یا خوشه) یکسان در کنار یکدیگر هستند، در غیراین‌صورت *gcmr* توقف کرده و یک پیغام خطا برمی‌گرداند. وابستگی پیاپی در سری‌های‌زمانی می‌تواند با تابع arma.cormat(p, q) توصیف شود و رتبه‌های p و q را به‌عنوان ورودی، از فرآیند $ARMA(p, q)$ دریافت می‌کند. داده‌های‌همبسته فضایی با در نظر گرفتن تابع همبستگی فضایی ماترن مدل می‌شود، همچنین با تابع matern.cormat(D, alpha) تنظیم می‌شود که در آن D ماتریس فواصل بین مشاهدات و alpha پارامتر شکل (دیگل و ریبریو ۲۰۰۷) است. تابع matern.cormat() از روی تابع matern() در بسته geoR (ریبریو و دیگل ۲۰۱۵) ساخته می‌شود. مقدار پیش‌فرض برای پارامتر alpha برابر ۰.۵ است. همچنین متناظر با یک مدل همبستگی نمایی می‌باشد.

جدول ۲.۴: مدل همبستگی موجود در بسته gcmr

همبستگی	cormat.gcmr
ARMA(p،q)	arma.cormat(p، q)
خوشه‌ای طولی/ داده‌های	cluster.cormat(id، type)
استقلال	ind.cormat()
مترن همبستگی	matern.cormat(D، alpha)

۴.۲.۴ امکانات برازش

امکانات برازش در gcmrبا آرگومان options و یا با نوشتن توابع زیر تنظیم می‌شود.

```
${gcmr}$.options(seed = round(runif(1, 1, 1e+05)),
nrep = c(100, 1000), no.se = FALSE,
method = c("BFGS", "Nelder-Mead", "CG"), ...)
```

آرگومان‌های موجود عبارتند از: seed، که به‌منظور تعیین دانه‌های شبه‌ـتصادفی که در الگوریتم GHK استفاده می‌َشود و تابع درست‌نمایی را با پاسخ‌های گسسته تقریب می‌زند، در نظر گرفته شده است، nrep ، که به‌منظور تنظیم تعداد تکرار آزمایش مونت کالو در الگوریتم GHK، تعبیه شده، no.se، به‌منظور اینکه آیا خطاهای استاندارد محاسبه شوند یا خیر و در نهایت method، به‌منظور انتخاب روش بهینه‌سازی که از دستور optim() به‌دست می‌آید، مورد استفاده قرار می‌گیرند.

الگوریتم بهینه‌سازی پیش‌فرض، الگوریتم شبه‌ـنیوتن BFGS است. ممکن است برداری از تکرار آزمایش‌های مونت کارلو به آرگومان nrep داده شود، به‌طوری که مدل با دنباله‌ای از اندازه‌های مونت کارلو متفاوت برازش شود. در این حالت، مقادیر آغازین برای بهینه‌سازی درست‌نمایی، از برازش‌های قبلی گرفته می‌شود. یک استراتژی منطقی، برازش مدل با اندازه مونت‌کارلو کوچک است تا مقدار آغازین معقولی به‌دست آید و سپس برازش دوباره با اندازه مونت‌کارلو بزرگ‌تر می‌باشد. اندازه مونت‌کارلو پیش‌فرض برای اولین بهینه‌سازی برابر با ۱۰۰ و برای دومین بهینه‌سازی و بهینه‌سازی نهایی برابر با ۱۰۰۰ می‌باشد. اگر پاسخ‌ها پیوسته باشند، در این‌صورت، تابع درست‌نمایی دارای یک عبارت فرم بسته است و مقادیر seed و nrep در نظر گرفته نمی‌شوند.

۵.۲.۴ روش‌ها

مدل برازش‌شده و برگردانده شده از کلاس $gcmr$ لیستی شامل، برآوردهای درست‌نمایی ماکزیمم، برآوردهای عددی و لگاریتم‌ـدرست‌نمایی ماکزیمم‌شده هسین و ژاکوبین لگاریتم‌ـدرست‌نمایی که در برآورد درست‌نمایی‌ماکزیمم محاسبه شده است، می‌باشد. یک مجموعه از روش‌های استاندارد به‌منظور به‌دست آوردن اطلاعات از مدل برازش‌شده در دسترس است (جدول ۴.۲.۴ را ببینید). اکثر توابع و روش‌ها دارای کد دستور یکسان در بسته‌های R که متمایل به تحلیل رگرسیون هستند، می‌باشند (به‌عنوان مثال betareg (سریباری‌ـنتو و زلیس ۲۰۱۰) را ببینید).

جدول ۳.۴: توابع و روش‌های موجود برای اشیا کلاس $gcmr$

Function	پراکندگی
print()	شده براورد ضرایب ساده نمایش
summary()	رگرسیون استاندارد خروجی
coef()	ضرایب براورد
vcov()	ضرایب براورد کوواریانس ماتریس
fitted()	داده‌ها مشاهده برای داده‌شده برازش میانگین
residuals()	باقیمانده‌ها دنباله
estfun()	۲۰۰۶) (زیلیس ساندویچ براوردگرهای برای براوردگر توابع
bread()	۲۰۰۶) (زیلیس ساندویچ براوردگرهای برای ”بِرد” ماتریس
terms()	مدل اجزای شرایط
model.frame()	مدل چهارچوب
model.matrix()	مدل ماتریس
logLik()	لگاریتمی درستنمایی ماکزیمم
plot()	باقیمانده دنباله تشخیص برای نمودار
profile()	تمرکز ضریب درستنمایی پروفایل
coeftest()	والد جزئی ضرایب تست
waldtest()	توپی مدل‌های از والد آزمایشات
lrtest()	توپی مدل‌های درستنمایی نسبت آزمون
AIC()	اطلاعات معیارهای

روش $gcmr.plot$نمودارهای تشخیصی گوناگونی را از شی برازش‌شده $gcmr$ تولید می‌کند، که از آن‌جمله می‌توان قطعه‌های پراکندگی مانده‌های چندک در برابر شاخص‌های مشاهدات و یا در مقابل پیش‌گویی کننده خطی، نمودار احتمال نرمال با بازه اطمینان مبتنی بر اجرا در بسته car (فوکس و ویسبرگ[1] ۲۰۱۱)، قطعه پراکنده مقادیر پیش‌بینی شده در برابر مقادیر مشاهده‌شده، نمودارهای خودهمبستگی و خودهمبستگی جزئی مانده‌ها را نام برد.

رفتار پیش‌فرض $gcmr.plot$ از نوعی ماتریس همبستگی به‌همین روش گرفته شده است. به‌عنوان مثال: نمودارهای خودهمبستگی به‌طور خودکار برای $ARMA(p,q)$نمایش داده می‌شوند که در واقع همبستگی مشخص شده با تابع arma.cormat(p, q) است.

چندک‌های مانده با روش زیر محاسبه می‌شوند:

```
residuals.${gcmr}$(object, type = c("conditional", "marginal"),
method = c("random", "mid"), ...)
```

که در آن آرگومان type امکان انتخاب بین چندک مانده conditional یا marginal را می‌دهد.

[1] Fox and Weisberg

آرگومان method تنها در حالت گسسته به‌منظور انتخاب بین مانده‌های چندک random یا مانده‌های چندک بازه mid، فعال است. نمودار لگاریتم‌ـ درست‌نمایی می‌تواند با استفاده از روش زیر به‌دست آید:

```
profile.${gcmr}$(fitted, which, low, up, npoints = 10,
display = TRUE, alpha = 0.05, progress.bar = TRUE, ...)
```

که در آن آرگومان which اندیس پارامتر نمودار شده است. low و up حدود بالایی و پایینی در محاسبات هستند. npoints، تعداد نقاط استفاده شده در محاسبات درست‌نمایی نمودار است. Alpha، سطح معناداری، display کنترل می‌کند که آیا درستنمایی نمودار باید رسم شود یا خیر. در نهایت progress.bar به‌منظور تجسم پیشرفت محاسبات درست‌نمایی نمودار زمان‌ـ مصرف، یک progressbar تنظیم می‌کند. اگر حدود low و up مقداردهی نشده باشند، در این صورت آن‌ها به‌ترتیب برابر با پارامتر برآورد به‌اضافه سه برابر خطای استاندارد و منهای سه برابر خطای استاندارد، در نظر گرفته می‌شوند.

۳.۴ کاربردها

استفاده از $gcmr$ با سه مجموعه داده متفاوت در زیر نشان‌داده شده است که شکل‌های گوناگونی از عدم استقلال را پوشش می‌دهد و در دنیای واقعی کاربرد دارد.

۱.۳.۴ داده‌های شمارشی طولی

اولین مثال مطالعات‌طولی شناخته شده روی حملات ناگهانی صرع را نشان می‌دهد که توسط دیگل و همکاران (۲۰۰۲) توصیف شده است.

داده‌ها شامل اطلاعاتی درباره ۵۹ فرد مشاهده شده در پنج موقعیت متفاوت، است. مبنای مشاهدات متشکل از تعداد حملات ناگهانی صرع در یک بازه هشت هفته‌ای است و بر اساس چهار اندازه‌گیری که در ملاقات‌های صورت‌گرفته در هر دو هفته جمع‌آوری شده‌اند، شکل می‌گیرد. متغیرهای موجود، شناسه بیمار id، سن بیمار age، شناساگر trt که اگر بیمار با دارو تراگابایت تحت مراقبت قرار گرفته باشد ($\texttt{trt} = 1$) و اگر نباشد ($\texttt{trt} = 0$) است، تعداد حملات ناگهانی صرع counts، دوره مشاهدات در هفته‌ها time، که برای مبنا $\texttt{time} = 8$ و برای ملاقات‌های صورت‌گرفته بعدی $\texttt{time} = 2$ است، شناساگر visit که اگر مشاهدات متناظر با یک ملاقات باشد ($\texttt{visit} = 1$) و یا اگر به صورت مبنا باشد ($\texttt{visit} = 0$) است. دیگل و همکاران داده‌های حمله ناگهانی را با روش معادلات برآورد شده تعمیم‌یافته که یک مدل رگرسیون لگاریتم‌ـ خطی برای counts با لگاریتم time به عنوان متعادل‌کننده در نظر می‌گیرد و همچنین متغیرهای کمکی visit و trt را به همراه فعل و انفعالات آن‌ها مورد بررسی قرار دادند. افزون بر این، به‌منظور حذف بیمار ''خارجی'' (در اینجا بیماری که شناسه آن $\texttt{id} = 49$ است)، تعداد حملات ناگهانی بسیار بالایی (۱۵۱ حمله) را که حتی بعد از معالجه تعداد آن دو برابر می‌شود (۳۰۲ حمله بعد از ۸ هفته) پیشنهاد دادند. در واقع، اگر بیمار به‌طور جداگانه در نظر گرفته

شود، ضرایب مدل برآورد شده به‌طور قابل ملاحظه‌ای تغییر می‌کند. تحلیل مفصل گوسی متناظر توصیف شده در زیر یک توزیع دوجمله‌ای حاشیه‌ای با میانگین تعیین شده مشابه کار دیگل و همکاران (۲۰۰۲)، در نظر می‌گیرد. ما برای مفصل گوسی، با در نظر گرفتن ماتریس همبستگی استقلال کاری، تحلیل را آغاز می‌کنیم.

```
Call:
gcmr(formula = counts ~ offset(log(time)) + visit + trt + visit:trt,
    data = epilepsy, subset = (id != 49), marginal = negbin.marg,
     cormat = cluster.cormat(id, type = "ind"))


Coefficients marginal model:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.34759    0.16649   8.094 5.77e-16 ***
visit        0.11187    0.18802   0.595    0.552
trt         -0.10685    0.23057  -0.463    0.643
visit:trt   -0.30237    0.26118  -1.158    0.247
dispersion   0.73421    0.07153  10.264  < 2e-16 ***

No coefficients in the Gaussian copula
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

log likelihood = 948.06,  AIC = 1906.1
```

تابع $gcmr.summary$ خطاهای استاندارد سنتی را که از وارون اطلاعات فیشر مشاهده شده به‌دست آمده است، محاسبه می‌کند. یک انتخاب مناسب‌تر برای این داده‌های طولی با برآوردگر ساندویچ که می‌تواند با استفاده از بسته sandwich (زلیس[۲] ۲۰۰۴، ۲۰۰۶) محاسبه شود، فراهم می‌شود و به‌راحتی با تابع coeftest() از بسته lmtest (زلیس و هتورن[۳] ۲۰۰۲) تجسم می‌شود.

```
z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  1.347586   0.157997  8.5292 < 2.2e-16 ***
```

[۲] Zeileis
[۳] Zeileis and Hothorn

```
visit         0.111869   0.115634  0.9674    0.3333
trt          -0.106846   0.194159 -0.5503    0.5821
visit:trt    -0.302373   0.169183 -1.7873    0.0739 .
dispersion    0.734208   0.095039  7.7253 1.116e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

خطاهای استاندارد قدرتمند ساندویچ، اساسا نتایج قبلی را تایید می‌کنند. معناداری قوی پارامتر پراکندگی، پشتیبانی از انتخاب توزیع حاشیه‌ای دوجمله‌ای را به‌جای توزیع پواسن، به‌دست می‌دهد.

به‌هرحال، یک توصیف دقیق‌تر از داده‌ها هم‌بستگی پیاپی مشاهدات را از شی یکسان، بیان می‌کند. به‌عنوان مثال: مدل می‌تواند با ماتریس همبستگی مفصل گوسی $AR(1)$ دوباره برآورد شود.

دستور قبلی استفاده از امکانات برازش $gcmr$ را نشان می‌دهد. عدد تصادفی مولد seed به‌منظور تضمین قابلیت تکثیر نتایج، تعیین می‌شوند. در صورتی‌که تعداد تکرار آزمایش‌های مونت‌کارلو (nrep) با عددی کوچک‌تر از حالت پیش‌فرض تنظیم می‌شود، در واقع احتمال آن می‌رود که در طول مرحله مشخصات مدل سودمند باشند.

خطاهای استاندارد قدرتمند ساندویچ، حضور قابل ملاحظه خودهمبستگی بین مشاهدات یکسان بیمار را تایید می‌کند. در واقع، ضریب $AR(1)$ برآورد شده برابر با ۰.۶۳ با خطای استاندارد ۰.۰۵ است:

```
z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  1.307160   0.162154  8.0612 7.553e-16 ***
visit        0.156902   0.108164  1.4506   0.14689
trt         -0.010332   0.202333 -0.0511   0.95927
visit:trt   -0.420571   0.164879 -2.5508   0.01075 *
dispersion   0.636966   0.077159  8.2552 < 2.2e-16 ***
ar1          0.628785   0.048897 12.8594 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

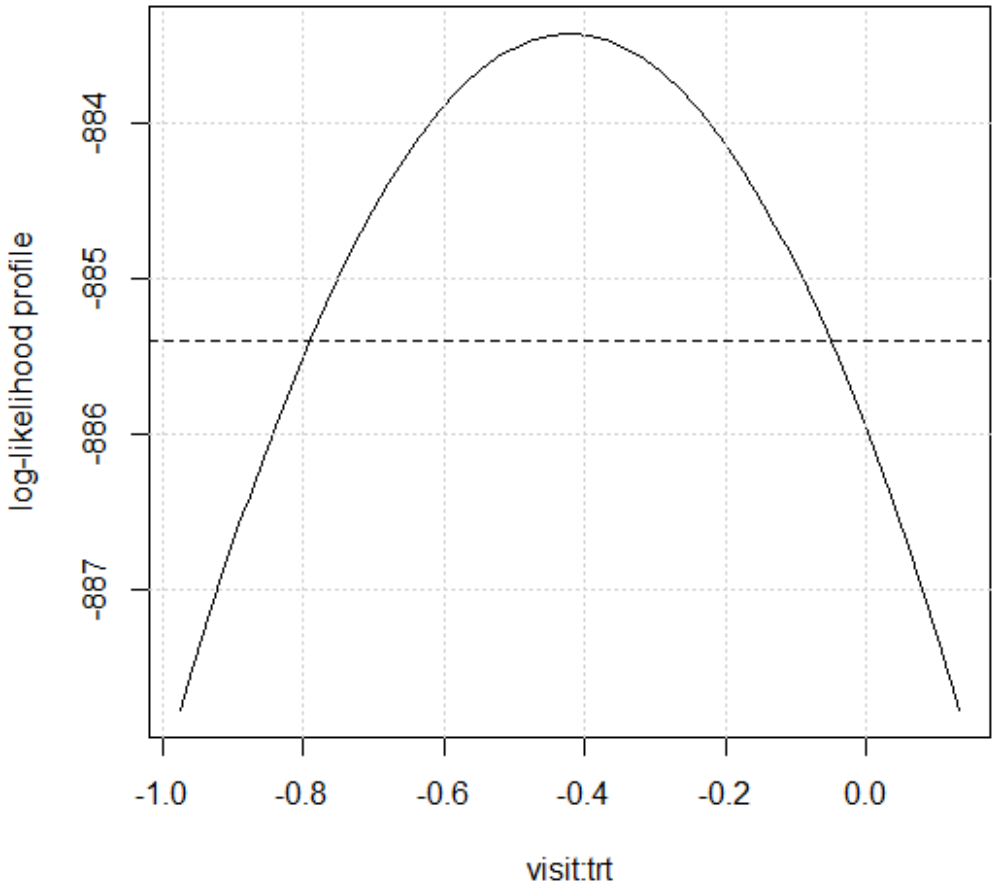
بر خلاف مدل استقلال کاری، مدل خودبرگشتی یک اثر قابل ملاحظه را از فعل و انفعالات بین ملاقات و معالجه که توسط مدل استقلال کاری کشف نشده بود، مشخص می‌کند. نتایج از نظر کیفی با آنچه که از تحلیل معادله برآورد تعمیم‌یافته توسط دیگل و همکاران (۲۰۰۲) به‌دست آمده، تطابق دارد و می‌تواند با استفاده از بسته geepack (یان[۴] ۲۰۰۲، هوجسگارد[۵] و همکاران ۲۰۰۶) بازتولید شود.

[۴] Yan
[۵] Hojsgaard

```
Call:
geeglm(formula = counts ~ offset(log(time)) + visit + trt + visit:trt,
    family = poisson, data = epilepsy, subset = (id != 49), id = id,
    corstr = "ar1")

 Coefficients:
            Estimate  Std.err   Wald Pr(>|W|)
(Intercept)  1.31383  0.16159 66.103 4.44e-16 ***
visit        0.15094  0.11077  1.857   0.1730
trt         -0.07973  0.19831  0.162   0.6877
visit:trt   -0.39872  0.17454  5.218   0.0223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)    10.61    2.35

Correlation: Structure = ar1  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha   0.7831 0.05192
Number of clusters:   58   Maximum cluster size: 5
```

از میان مزایای تحلیل درست‌نمایی اجراشده درgcmr نسبت‌به روش‌های غیر درست‌نمایی از جمله معادلات برآورد تعمیم‌یافته، می‌توان به احتمال محاسبه نمودار درست‌نمایی‌ـ‌لگاریتم اشاره کرد.

به‌عنوان مثال: نمودار درست‌نمایی‌ـ‌لگاریتم را برای فعل و انفعالات اثر ملاقات با معالجه در نظر بگیرید، که می‌تواند با استفاده از دستور $gcmr.profile$با آرگومان $\texttt{which} = 4$، به‌دست آید، زیرا اثر فعل و انفعالات متناظر با پارامتر مدل چهارم به‌صورت زیر است: نمودار درست‌نمایی‌ـ‌لگاریتم گزارش داده‌شده در شکل ۱.۴، ضریب منفی قابل ملاحظه‌ای را که متناظر با فعل و انفعالات ملاقات با معالجه است، نشان می‌دهد.

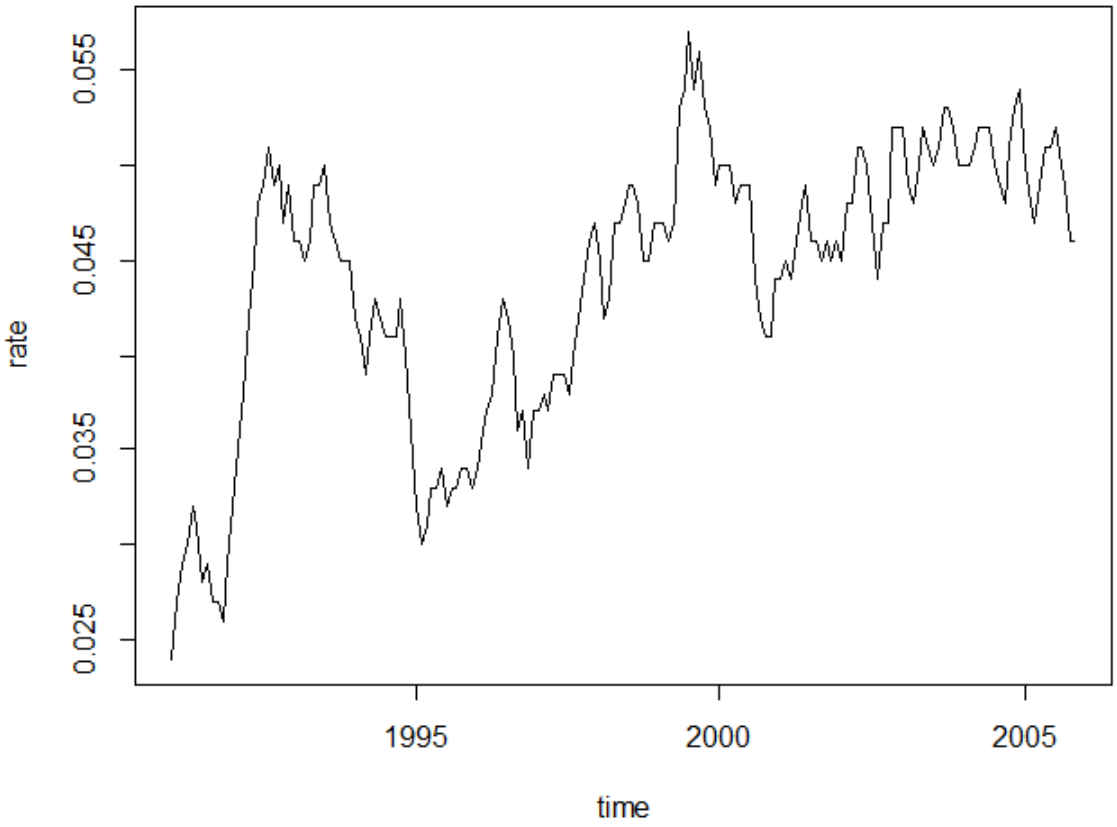


شکل ۱.۴: اطلاعات تشنج. درستنمایی لگاریتمی پروفایل برای تعامل بین ویزیت و درمان

۲.۳.۴ سری زمانی نرخ‌ها

مثال دوم، که دربردارنده سری‌های‌زمانی نرخ بیکاری پنهان (HUR) در سائوپائولوی برزیل است، از پایگاه داده موسسه تحقیقات اقتصادی کاربردی (IPEA) در دولت فدرال برزیل $(www.ipea.gov.br)$ به‌دست آمده است: اطلاعات نشان‌داده شده در شکل ۲.۴، به‌وسیله روکا و کریباری نتو[۶] با یک مدل میانگین متحرک و اتورگرسیو بتای مشاهده‌محور[۷] تحلیل شده است. به‌عنوان یک جایگزین برای تحلیل‌های انجام‌شده به‌وسیله روکا و کریباری نتو (۲۰۰۹)، یک مدل مفصل گاوسی را با توزیع بتای‌حاشیه‌ای و همبستگی مفصل $ARMA(p,q)$ در نظر می‌گیریم. میانگین و دقت حاشیه‌های بتا هر دو با فرض وابستگی به یک روند خطی می‌باشند. جهت جلوگیری از ناپایداری‌های عددی، روند متمرکزسازی و مقیاس بندی می‌شود:

[۶] Roca and Cribari-Neto
[۷] observation-driven



شکل ۲.۴: اطلاعات مربوط به‌نرخ بیکاری پنهان در سائوپائولو، برزیل. منبع داده موسسه تحقیقات اقتصادی کاربردی ،(IPEA) دولت فدرال برزیل (www.ipea.gov.br)

در ادامه مدل را با خطای $ARMA(1,3)$ نشان می‌دهیم. این مدل انتخاب شده است زیرا دارای حداقل AIC در میان شانزده مدل $ARMA(p,q)$ به‌دست‌آمده با مرتبه‌های p و q در محدوده ۰ تا ۳ است: دستور قبلی نشان‌دهنده استفاده از فرمول تعمیم‌یافته $HUR\ \ trend \mid trend$ است که مشخص می‌کند هم میانگین و هم پراکندگی به روند (مقیاس‌بندی‌شده) بستگی دارند. خلاصه‌ای از مدل برازش داده‌شده، یک روند آماری قابل‌توجه است:

```
Call:
gcmr(formula = HUR ~ trend | trend, marginal = beta.marg,
cormat = arma.cormat(1, 3))
```

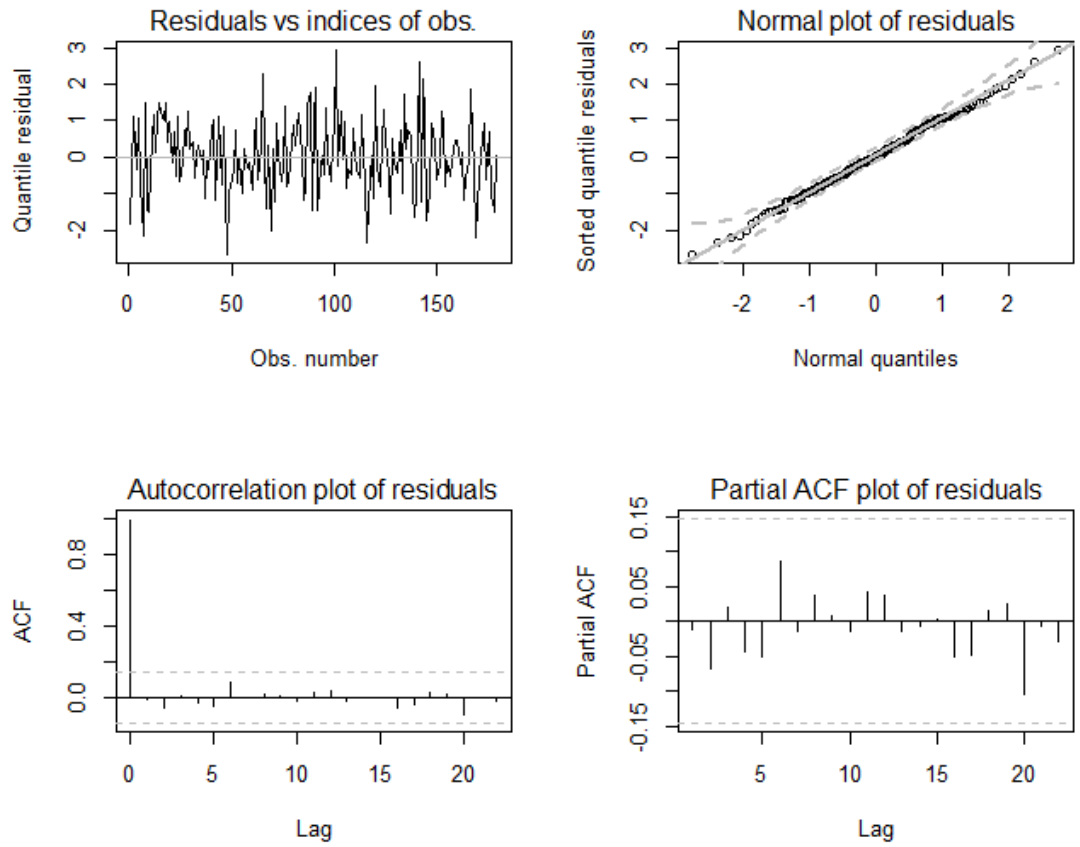
```
Coefficients marginal model:
                     Estimate Std. Error z value Pr(>|z|)
mean.(Intercept)      -3.1077     0.0461  -67.39   <2e-16 ***
mean.trend             0.1151     0.0394    2.92   0.0035 **
precision.(Intercept)  7.2487     0.3720   19.49   <2e-16 ***
precision.trend        0.3611     0.1137    3.18   0.0015 **

Coefficients Gaussian copula:
    Estimate Std. Error z value Pr(>|z|)
ar1   0.9103     0.0459   19.84  < 2e-16 ***
ma1   0.3415     0.0947    3.61  0.00031 ***
ma2   0.4715     0.0857    5.50  3.8e-08 ***
ma3  -0.4290     0.1030   -4.17  3.1e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

log likelihood = -895.06,  AIC = -1774
```

شواهدی که فرضیات مدل فوق را برآورده می‌کند، به‌وسیله پنل گرافیکی باقی‌مانده‌های چارکی گزارش‌شده در شکل ۳.۴ ارائه شده است:

۳.۳.۴ داده دوجمله‌ای همبسته فضایی

آخرین مثال در مورد شیوع مالاریا در کودکان گزارش‌شده در ۶۵ روستای گامبیا است. که با داده‌های اصلی ارائه‌شده در تامسون[۸] و همکاران (۱۹۹۹) متفاوت است در بسته GeoR (ریبیرو و دیگل[۹] ۲۰۱۵) وجود دارد. در اینجا داده‌های جمع‌آوری‌شده در سطح روستا را که از طریق gcmr با قاب داده مالاریا در دسترس هستند در نظر می‌گیریم: داده‌ها شامل اطلاعات مربوط‌به مختصات دهکده (x, y)، تعداد کودکان نمونه‌گیری‌شده $(size)$ دارای مالاریا $(cases)$ در هر روستا، میانگین سن کودکان نمونه‌گیری‌شده در هر روستا (age)، فراوانی کودکان نمونه‌گیری‌شده که به‌طور مرتب زیر توری (پشه‌بند) می‌خوابند در هر روستا $(netus)$، فراوانی کودکان نمونه‌گیری‌شده که توری آن‌ها مورد بررسی و کار قرار گرفته $(treated)$، اندازه‌گیری ماهواره‌ای از سبز بودن پوشش گیاهی در نزدیکی روستا $(green)$، متغیر شاخص نشان‌دهنده وجود (۱) یا عدم وجود (۰) یک مرکز بهداشت در روستا (pch) و یک شاخص

[۸] Thomson
[۹] Thomson



شکل ۳.۴: داده‌های نرخ بیکاری پنهان. نمودارهای تشخیصی استاندارد برای داده‌های سری‌های زمانی تولیدشده به وسیله plot.gcmr.

از مناطق جغرافیایی با مقادیر متفاوت از خطرهای بالقوه مالاریا $(area)$ است. برای جزئیات بیشتر به دیگل و ریبیرو (۲۰۰۷) مراجعه می‌کنیم. هدف این است که رابطه بین تعداد موارد و متغیرهای تصادفی کمکی مختلف را مدل‌سازی کنیم، درحالی‌که حضور بالقوه وابستگی فضایی شیوع مالاریا بین روستاها را در نظر می‌گیریم. اولین گام تحلیل داده ایجاد ماتریسی از فاصله بین روستاها است، که این امر به‌طور مثال بااستفاده از تابع $(spDists)$ از بسته sp صورت می‌گیرد (پبسما و بیوند[۱۰] ۲۰۰۵؛ بیوند[۱۱] و همکاران ۲۰۱۳). فاصله‌ها برحسب کیلومتر از طریق مقیاس‌بندی با ضریب ۱۰۰۰، بیان می‌شود. مقیاس‌بندی برای اجتناب از ناپایداری‌های عددی بالقوه در برآورد پارامتر وابستگی فضایی مفید است. مدل اول، موارد مالاریا را با یک مدل رگرسیون لجستیک مفصل گاوسی فضایی توصیف می‌کند. متغیرهای کمکی تصادفی pch netuse، و green هستند که با ضریب ۱۰۰ مقیاس‌بندی شده‌اند. وابستگی فضایی با یک ماتریس همبستگی نمایی متناظر با مقدار پیش‌فرض پارامتر شکل $alpha = (0.5)inmatern.cormat(D, alpha)$ مدل‌سازی شده‌اند:

[۱۰] Pebesma and Bivand
[۱۱] Bivand

```
Call:
gcmr(formula = cbind(cases, size - cases) ~ netuse + I(green/100) + phc,
   data = malaria, marginal = binomial.marg, cormat = matern.cormat(D),
    seed = 12345)


Coefficients marginal model:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.828      0.406   -2.04    0.042 *
netuse         -1.176      0.161   -7.32  2.4e-13 ***
I(green/100)    2.949      0.750    3.93  8.4e-05 ***
phc            -0.405      0.102   -3.98  7.0e-05 ***

Coefficients Gaussian copula:
    Estimate Std. Error z value Pr(>|z|)
tau    1.509      0.377       4  6.3e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

log likelihood = 252.68,  AIC = 515.4
```

متغیرهای کمکی netuse و phc متناسب با یک کاهش قابل‌توجه در موارد مالاریا هستند درحالی‌که green با خطر بالاتری از بیماری متناسب است. برآورد پارامتر وابستگی tau نشان‌دهنده حضور وابستگی فضایی بسیار موضعی است. درواقع، مقدار برآورد شده برای tau نشان‌دهندهٔ یک دامنه عملی (برای مثال دیگل و ریبیرو ۲۰۰۷) در حدود ۴.۵ کیلومتر $(= 1.51 \times 3)$ است. تنها ۳۷ جفت روستا وجود دارند که فاصله آن‌ها از یکدیگر کمتر از ۴.۵ کیلومتر است.
مدل دوم شامل یک اثر اضافی ناشی از متغیر کمکی area است:

```
Call:
gcmr(formula = cbind(cases, size - cases) ~ netuse + I(green/100) + phc +
area, data = malaria, marginal = binomial.marg, cormat = matern.cormat(D),
seed = 12345)


Coefficients marginal model:
             Estimate Std. Error z value Pr(>|z|)
```

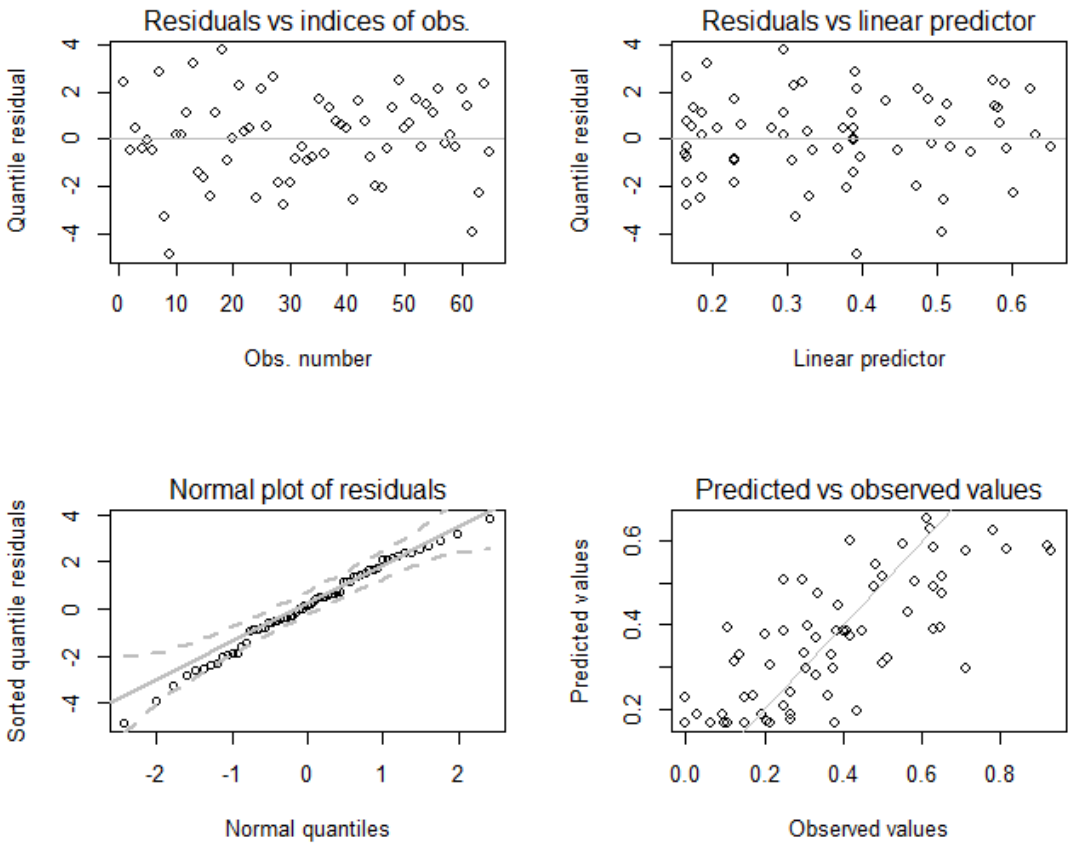
```
(Intercept)    0.2046     0.6398    0.32  0.74907
netuse        -0.6387     0.1789   -3.57  0.00036 ***
I(green/100)  -0.0611     1.4057   -0.04  0.96533
phc           -0.4081     0.1073   -3.80  0.00014 ***
area2         -0.6133     0.1792   -3.42  0.00062 ***
area3         -0.7515     0.1945   -3.86  0.00011 ***
area4          0.3441     0.2432    1.41  0.15710
area5          0.6840     0.2316    2.95  0.00314 **

Coefficients Gaussian copula:
    Estimate Std. Error z value Pr(>|z|)
tau    0.682      0.362    1.88     0.06 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

log likelihood = 222.15,  AIC = 462.3
```

دربرداشتن area دراین مدل منجر به یک افت بزرگ در مقادیر AIC می‌شود:

```
         df   AIC
mod       5 515.4
mod.area  9 462.3
```

به‌طور خلاصه می‌توان تأیید کرد که متغیر کمکی area شامل اطلاعات مناسب درباره تغییرات جغرافیایی خطر مالاریا در ناحیه مورد مطالعه است. درواقع، برآورد پارامتر وابستگی فضایی tau در مدل $mod.area$ نشان می‌دهد که وابستگی‌فضایی باقی‌مانده ناچیز است.

در پایان، شیوه‌های تشخیص گزارش‌شده در شکل ۴.۴ بیان می‌کند که شرایط مدل برقرار هستند:



شکل ۴.۴: داده‌های مالاریا. نمودارهای تشخیصی استاندارد تولیدشده با plot.gcmr.

۴.۴ نتیجه‌گیری

در این فصل، پیاده‌سازی رگرسیون حاشیه‌ای مفصل گاوسی موجود در بسته gcmr در نرم‌افزار R ارائه شد. مثال‌های موردبحث نشان‌دهنده توانایی بسته برای اداره انواع مختلف داده‌ها و ساختارهای وابستگی است. مدل‌ها با روش درستنمایی ماکزیمم (شبیه‌سازی‌شده) منطبق هستند که نیازمند تجزیه چولسکی متوالی از ماتریس همبستگی مفصل گاوسی می‌باشد. در نسخه فعلی gcmr، مرتبه محاسبات مورد نیاز برای ارزیابی درستنمایی، $O\left(n^3\right)$ است؛ که n نشان‌دهنده تعداد مشاهدات است. در حالت مجموعه داده‌های‌بزرگ، که به‌عنوان‌مثال، شامل چندین هزار مشاهده می‌باشند، هزینه محاسباتی می‌تواند از استفاده معمول از gcmr جلوگیری کند. بااین‌حال، تجزیه چولسکی را می‌توان برای برخی از ساختارهای وابستگی خاص به‌صورت کارآمدتر پیاده کرد. به‌عنوان‌مثال، ماتریس‌های همبستگی متوسط متحرک و اتورگرسیو را می‌توان در تعداد خطی از محاسبات با استفاده از فیلتر کالمن از طریق نمایش فضای حالت تجزیه کرد.

پژوهش‌های آینده بر روی پیاده‌سازی روش‌های محاسباتی مناسب برای رسیدگی به انواع خاص وابستگی در چارچوب کلی رگرسیون مفصل گاوسی تمرکز خواهد کرد. رویکردهای امیدوارکننده شامل

درستنمایی های ترکیبی برای کاهش تلاش محاسباتی از طریق تجزیه‌های درستنمایی مناسب (وارین و همکاران ۲۰۱۱) و روش‌های پراکنده طراحی‌شده برای تقریب ماتریس همبستگی مفصل گاوسی با یک ماتریس قطری قطعه‌ای قابل‌کنترل‌تر است.

نویسنده‌های متعددی از مفصل‌های t و گاوسی برای ساخت مدل‌های رگرسیون مشترک برای پاسخ‌های چندگانه، و همچنین از نوع ترکیبی آن استفاده کرده‌اند (به‌طور مثال فریز و والدز ۲۰۰۸؛ سانگ و همکاران ۲۰۰۹؛ وو و دی لئون ۲۰۱۴؛ ییریای و همکاران ۲۰۱۵). روش‌های برای به‌کار بردن پاسخ‌های چندگانه در نسخه‌های آینده gcmr قرار خواهد گرفت.

پیوست آ

کدهای مربوط به شبیه‌سازی داده‌ها

در این بخش برنامه شبیه‌سازی داده‌های مورد استفاده در فصل ۴ ارائه شده است.

۱.آ کدهای شبیه سازی مثال ۱.۳.۴

```
data("epilepsy", package = "gcmr")
## independence model
mod.ind <- gcmr(counts ~ offset(log(time)) + visit
  + trt + visit:trt, data = epilepsy,
  subset = (id != 49), marginal = negbin.marg,
  cormat = cluster.cormat(id, type = "ind"))
summary(mod.ind)
## sandwich standard errors
library("sandwich")
library("lmtest")
coeftest(mod.ind, vcov. = sandwich(mod.ind))
```

```
## AR(1) correlation
mod.ar1 <- update(mod.ind, cormat = cluster.cormat(id, "ar1"),
  seed = 12345, nrep = 100)
coeftest(mod.ar1, vcov. = sandwich(mod.ar1))
## alternative analysis with GEE
library("geepack")
gee.ar1 <- geeglm(counts ~ offset(log(time)) + visit
   + trt + visit:trt, data = epilepsy,
id = id, subset = (id != 49), family = poisson, corstr = "ar1")
summary(gee.ar1)
## profile likelihood (time consuming)
profile(mod.ar1, which = 4)
```

آ.۲ کدهای شبیه سازی مثال ۲.۳.۴

```
data("HUR", package = "gcmr")
plot(HUR, ylab = "rate", xlab = "time")
## beta regression with trend and ARMA(1,3) errors
trend <- scale(time(HUR))
mod <- gcmr(HUR ~ trend | trend, marginal = beta.marg,
       cormat = arma.cormat(1, 3))
summary(mod)
## diagnostic plots
par(mfrow = c(2, 2))
plot(mod)
```

آ.۳ کدهای شبیه سازی مثال ۳.۳.۴

```
data("malaria", package = "gcmr")
## spatial distances
library("sp")
D <- spDists(cbind(malaria$x, malaria$y))/1000
## model without area
mod <- gcmr(cbind(cases, size - cases) ~ netuse + I(green/100)
     + phc, data = malaria,
```

```
marginal = binomial.marg, cormat = matern.cormat(D), seed = 12345)
summary(mod)
## model with area
mod.area <- update(mod, . ~ . + area)
summary(mod.area)
AIC(mod, mod.area)
## diagnostic plots
par(mfrow = c(2, 2))
plot(mod.area)
```

مراجع

[۱] امیدی، م. و محمدزاده، م. (۱۳۹۱) «تحلیل داده‌های فضایی با استفاده از توابع مفصل»، مجموعه مقالات دومین کارگاه آموزشی نظریه مفصل و انواع وابستگی‌ها، گروه آمار دانشگاه فردوسی مشهد و انجمن آمار ایران.

[2] Akaike, Hirotugu (1974), "A new look at the statistical model identification". **IEEE Transactions on Automatic Control**, 19(6), pp 716–723.

[3] Bai Y, Kang J, Song PXK (2014). "Efficient Pairwise Composite Likelihood Estimation for Spatia Clustered Data", **Biometrics**, 70(3), 661-670.

[4] Bivand RS, Pebesma E, Gomez-Rubio V (2013). "Applied Spatial Data Analysis with R", **Springer-Verlag**, New York.

[5] Bodnar, O. Bodnar, T., and Gupta, A.K. (2010). "Estimation and inference for dependence in multivariate data". **Journal of Multivariate Analysis**, 101, 869–881.

[6] Chambers JM, Hastie TJ (1993). "Statistical Models in S", **Chapman Hall**, London.

[7] Craig, P. (2008). "A new reconstruction of multivariate normal orthant probabilities". **Journal of the Royal Statistical Society**, Series B 70, 227–243.

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

[9] Cribari-Neto F, Zeileis A (2010). "Beta Regression in R", **Journal of Statistical Software**, 34(2), 1-24.

[10] Durbin, J. and Koopman, S.J. (2001). "**Time Series Analysis by State Space Methods**", **Oxford University Press**.

[11] Diggle, P.J. and Ribeiro, P.J.J. (2007). "**Model-based Geostatistics**", **Springer**, New York.

[12] Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). "Analysis of longitudinal data. Second edition". **Oxford University Press**, Oxford.

[13] Diggle PJ, Heagerty P, Liang KY, Zeger SL (2002). "Analysis of Longitudinal Data", 2nd edition. **Oxford University Press**, Oxford.

[14] Embrechts P, McNeil A and Straumann D. (1999). "Correlation and dependence in riskmanagement:Properties and pitfalls. RISK". No. 69-71

[15] Fox J,Weisberg S (2011). "An R Companion to Applied Regression", Second edition. **Sage**, Thousand Oaks CA.

[16] Guido, M. and Cristiano, V. (2012). "Gaussian copula marginal regression", **Electronic Journal of Statistics**. Vol. 6. 1517–1549

[17] Guolo A, Varin C (2014). "Beta Regression for Time Series Analysis of Bounded Data, with Application to Canada Google Flu Trends", **The Annals of Applied Statistics**, 8(1), 7488.

[18] Hausman, J.A. (1978). "Specification tests in econometrics". **Econometrica**, 46, 1251–1271.

[19] Hoff, P.D. (2007). "Extending the rank likelihood for semiparametric copula estimation". **The Annals of Applied Statistics**, 1, 265–283.

[20] Hojsgaard S, Halekoh U, Yan J (2006). "The R Package geepack for Generalized Estimating Equations", **Journal of Statistical Software**, 15(2), 1-11.

[21] Hothorn, A., Bertz, F., and Genz, A. (2001). "On multivariate T and Gaussian probabilities in R". **R News** 1, 27–29.

[22] Joe, H. (1995). "Approximation to multivariate normal rectangle probabilities based on conditional expectations", **Journal of the American Statistical Association** 90, 957–964.

[23] Joe, H. (1997). "Multivariate Models and Dependence Concepts". **Chapman Hall** London.

[24] Keane, M.P. (1994). "A computationally practical simulation estimator for panel data". **Econometrica** 62, 95–116.

[25] Kugiumtzis, D. and Bora-Senta, E. (2010). "Normal correlation coefficient of non-normal variables using piece-wise linear approximation". **Computational Statistics**, 25, 645–662.

[26] Liang, K.-L. and Zeger, S.L. (1986). "Longitudinal data analysis using generalized linear models". **Biometrika**, 73, 13–22.

[27] Mantel, N., Bohidar, N.R. and Ciminera, J.L. (1977). "Mantel Haenszel analysis of litter-matched time-to-response data, with modifications to recovery of interlitter information". **Cancer Research** 37, 3863–3868.

[28] Mari D.D, Kotz S. (2001). "Correlation and Dependence Concepts". **Imperial College Press**, London.

[29] Masarotto G, Varin C (2012). "Gaussian Copula Marginal Regression", **Electronic Journal of Statistics**, 6, 1517-1549.

[30] Miwa, T., Hayter, A.J. and Kuriky, S. (2003). "The evaluation of general non-centred orthant probabilities", **Journal of the Royal Statistical Society**, Series B 65, 223-234.

[31] Molenberghs, G. and Verbeke, G. (2005). "Models for Discrete Longitudinal Data", **Springer**. Springer, 2005, ISBN 0-387-25144-8. New York

[32] Nelsen R.B. (2006). "An Introduction to Copulas. Springer Series in Statistics". **second ed**. Springer, NewYork

[33] Nikoloulopoulos, A.K., Joe, H. and Chaganty, N.R. (2011). "Extreme value properties of multivariate t copulas". **Extremes**, 12, 129–148.

[34] Parzen, M., Ghosh, S., Lipsitz, S., Sinha, D., Fitzmaurice, G.M., Mallick, B.K., Ibrahim, J.G. (2011). "A generalized linear mixed model for longitudinal binary data with a marginal logit link function". **The Annals of Applied Statistics**, 5, 449–467.

[35] McCullagh P, and Nelder J (1989). "Generalized Linear Models. Second edition", **Chapman and Hall/CRC**, Boca Raton.

[36] Pebesma EJ, Bivand RS (2005). "Classes and Methods for Spatial Data in R", **R News**, 5(2), 9-13.

[37] Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2014). "nlme: Linear and Nonlinear Mixed Effects Models", **R Package Version 3.1**, -118.

[38] Ribeiro PJJ, Diggle PJ (2015). "geoR: Analysis of Geostatistical Data", **R package version 1.7-5.1**.

[39] Roca V, Cribari-Neto F (2009). "Beta Autoregressive Moving Averag Models", **Test**, 18, 529-545.

[40] Song, P.X.-K. (2000). "Multivariate dispersion models generated from Gaussian copula". **Scandinavian Journal of Statistics**, 27, 305–320.

[41] Song, P.X-K. (2007). "Correlated Data Analysis: Modeling, Analytics and Applications". **Springer-Verlag**.

[42] Song, P.X.-K., Li, M. and Yuan, Y. (2009). "Joint regression analysis of correlated data using Gaussian copulas". **Biometrics**, 65, 60–68.

[43] Thomson M, Connor S, D'Alessandro U, Rowlingson B, Diggle PJ, M Cresswell BG (1999). "Predicting Malaria Infection in Gambian Children from Satellite Data and Bednet Use Surveys: The Importance of Spatial Correlation in the Interpretation of Results", **American Journal of Tropical Medicine and Hygiene**, 61, 2-8.

[44] Wakefield, J. (2007). "Disease mapping and spatial regression with count data", **Biostatistics** 8, 158–183.

[45] Waller, L.A. and Gotway, C.A. (2004). "Applied Spatial Statistics for Public Health Data", **John Wiley and Sons**. New York.

[46] Zeileis A (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators", **Journal of Statistical Software**, 11(10), 1-17.

[47] Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses", **Journal of Statistical Software**, 34(1), 1-13.

[48] Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships", **R News**, 2(3), 7-10.

[49] Zeger, S.L. and Karim, M.R. (1991). "Generalized linear models with random effects: a Gibbs sampling approach", **Journal of the American Statistical Association** 86, 79–86.

[50] Joe, H. (1997). "**Multivariate Models and Dependence Concepts**". Chapman and Hall.

[51] White, H. (1994). "**Estimation, Inference and Specification Analysis**". Cambridge University Press.

واژه‌نامه فارسی به انگلیسی

اتورگرسیو ..... Autoregressive

اسپایک ..... Spike

رگرسیون حاشیه‌ای مفصل گاوسی ..... Gaussian copula marginal regression (GCMR)

الگوی خطی تعمیم یافته ..... Generalized linear model (GLM)

الگوهای نقطه‌ای ..... Point patterns

باکس ـ جنکینز ..... Box-Jenkins

باقی‌مانده ترکیبی ..... Hybrid Residual

برنشتاین ..... Bernestine

بوت‌استرپ ..... Bootstrap

تابع بسل ..... Bessel function

تابع پیوند ..... Link function

تابع همبسته ایزوتروپیک ماترن ..... Mat´ern isotropic correlation function

تغییرات بزرگ‌مقیاس ..... Large scale variation

تغییرات کوچک‌مقیاس ..... Small scale variation

توانی فضایی ..... Spatial Power

خوشه‌ای ..... Cluster

چگال ..... Dense

داده‌های رسته‌ای ..... Categorical data

داده‌های زمین‌آماری ..... Geostatistical data

داده‌های طولی ..... Longitudinal data

داده‌های فضایی ..... Spatial data

داده‌های مشاهده‌ای ..... Observational data

داده‌های مشبکه‌ای ..... Lattice data

دم‌ـسبک ..... light-tailed

دودویی ..... Binary

دیکشنری لاتین کاسل ..... Cassell's Latin Dictionary

رگرسیون پروبیت ..... Probit regression

رگرسیون پواسن ..... Poisson regression

رگرسیون چندجمله ای ..... Ranking regression

رگرسیون خطی چندگانه ..... Multiple linear regression

رگرسیون خطی ساده ..... Simple linear regression

رگرسیون لجستیک ..... Logistic regression

رگرسیون لجستیک چندگانه ..... Multiple logistic regression

رگسیون ویبول ..... Weibull regression

روند ..... Trend

زمین‌آمار ..... Geostatistics

زمین‌خورده ..... Grounded

زنجیره مارکوف ..... Markovian process

ساختار خطی ..... Spatial Linear

ساختار درجه‌دوم ..... Spatial Rational Quadratic

ساختار کروی ..... Spatial Spherical

سانسور ..... Censor

شبیه‌ساز جوک ـ هاجیواسیلیو و کین ..... Geweke-Hajivassiliou-Keane

عدم وابستگی به مقیاس ..... scale-free

فرانک ..... Frank

فرمول‌های ساندویچی ..... Sandwich-type formulas

فضایی ـ زمانی ..... Spatio-temporal

کامل تصادفی‌فضایی ..... Complete spatial randomness

کای–دو ..... Chi-squared

کلایتون ..... Clayton

گاوسی فضایی ..... Spatial Gaussian

گامبل ..... Gumbel

گرادیان ..... Gradient

مانایی ..... Stationary

متروپلیس ..... Metropolis

مدل خطي لگاریتمی ..... Log-linear model

مدل خودرگرسیو میانگین متحرک ..... Autoregressive moving average model

مشاهده محور ..... observation-driven

معادلات برآوردگر تعمیم‌یافته ..... Generalized estimating equations

معیار آکائیک ..... Akaike information criterion

مفصل ..... Copula

مفصل ارشمیدسی ..... Archimedean Copula

منظم ..... Regular

مونت‌کارلو ..... Montecarlo

مونت کارلوی زنجیر مارکوفی ..... Markov Chain Monte Carlo

میانگین متحرک ..... Moving Average

میدان تصادفی..... Random field

میدان تصادفی گاوسی مارکوف ..... Gaussian Markovian random field

نمایی فضایی ..... Spatial Eponential

همبستگی خودبرگشت مرتبه اول ..... first order autoregressive

همسانگردی ..... Isotropic

واژه‌نامه انگلیسی به فارسی

معیار آکائیک ..... Akaike information criterion

مفصل ارشمیدسی ..... Archimedean Copula

اتورگرسیو ..... Autoregressive

مدل خودرگرسیو میانگین متحرک ..... Autoregressive moving average model

برنشتاین ..... Bernestine

تابع بسل ..... Bessel function

دودویی ..... Binary

بوت‌استرپ ..... Bootstrap

باکس–جنکینز ..... Box-Jenkins

دیکشنری لاتین کاسل ..... Cassell's Latin Dictionary

داده‌های رسته‌ای ..... Categorical data

رگرسیون حاشیه‌ای مفصل گاوسی ..... Gaussian copula marginal regression (GCMR)

سانسور ..... Censor

کای–دو ..... Chi-squared

کلایتون ..... Clayton

خوشه‌ای ..... Cluster

کامل تصادفی فضایی ..... Complete spatial randomness

مفصل ..... Copula

چگال ..... Dense

همبستگی خودبرگشت مرتبه اول ..... First order autoregressive

فرانک ..... Frank

میدان تصادفی گاوسی مارکوف ..... Gaussian Markovian random field

زمین‌آمار ..... Geostatistics

داده‌های زمین‌آماری ..... Geostatistical data

معادلات برآوردگر تعمیم‌یافته ..... Generalized estimating equations

الگوی خطی تعمیم‌یافته ..... Generalized linear model (GLM)

شبیه‌ساز جوک ـ هاجیواسیلیو و کین ..... Geweke-Hajivassiliou-Keane

گرادیان ..... Gradient

زمین‌خورده ..... Grounded

گامبل ..... Gumbel

باقی‌مانده ترکیبی ..... Hybrid Residual

همسانگردی ..... Isotropic

تغییرات بزرگ‌مقیاس ..... Large scale variation

داده‌های مشبکه‌ای ..... Lattice data

دم‌ـسبک ..... Light-tailed

تابع پیوند ..... Link function

مدل خطي لگاریتمی ..... Log-linear model

رگرسیون لجستیک ..... Logistic regression

داده‌های طولی ..... Longitudinal data

مونت‌کارلوی زنجیر مارکوفی ..... Markov Chain Monte Carlo

زنجیره مارکوف ..... Markovian process

تابع همبسته ایزوتروپیک ماترن ..... Mat´ern isotropic correlation function

متروپلیس ..... Metropolis

مونت‌کارلو ..... Montecarlo

میانگین متحرک ..... Moving Average

رگرسیون خطی چندگانه ..... Multiple linear regression

رگرسیون لجستیک چندگانه ..... Multiple logistic regression

مشاهده‌محور ..... Observation-driven

داده‌های مشاهده‌ای ..... Observational data

الگوهای نقطه‌ای ..... Point patterns

رگرسیون پواسن ..... Poisson regression

رگرسیون پروبیت ..... Probit regression

میدان تصادفی ..... Random field

رگرسیون چندجمله ای ..... Ranking regression

منظم ..... Regular

فرمول‌های ساندویچی ..... Sandwich-type formulas

عدم وابستگی به‌مقیاس ..... Scale-free

رگرسیون خطی ساده ..... Simple linear regression

تغییرات کوچک‌مقیاس ..... Small scale variation

داده‌های فضایی ..... Spatial data

نمایی فضایی ..... Spatial Eponential

گاوسی فضایی ..... Spatial Gaussian

ساختار خطی ..... Spatial Linear

توانی فضایی ..... Spatial Power

ساختار درجه دوم ..... Spatial Rational Quadratic

ساختار کروی ..... Spatial Spherical

فضایی_زمانی ..... Spatio-temporal

اسپایک ..... Spike

مانایی ..... Stationary

روند ..... Trend

رگسیون ویبول ..... Weibull regression

Abstract

Regression models are mostly used in various sciences, including medicine, natural sciences, social sciences, economics, and environmetrics. In cases where the response variable is non-normal, generalized linear models are usually used instead of linear ones. A basic assumption in these models is the independence between observations. However, in many situations we are encountered with some sort of dependency structure; e.g. longitudinal data, time series, and spatial data analysis. In these cases, the dependence of responses should be introduced into the model. Different approaches have been proposed for this task. The usual approach is to use generalized linear mixed models. Statistical inference in this class of models (in both frequentist and Bayesian approaches) faces serious computational problems. A new alternative solution to account for the dependency of responses is to use copula regression models.

In this thesis, we first introduce Gaussian copula marginal regression model. Then, we establish the statistical inference of the model and goodness of fit criteria, and we use Hunsman's test to determine the correctness of the copula model selection as well. Our inference approach is based on the likelihood function. We evaluate the performance of model by using simulated and real examples.

Key word: *Copula function, Gaussian copula, Hausman-type specification test, likelihood inference, longitudinal data, marginal regression, spatial data, time series.*

Shahrood University of Technology

Faculty Of Mathematical Sciences

MSc Thesis in: Mathematical Statistics

Modeling dependent responses using Gaussian copula

By: Khadijeh Soltani

Supervisors

Dr. Hossein Baghishani
Dr. Mohamad Arashi

August 2017