
بسم الله الرحمن الرحيم

دانشکده علوم ریاضی

پایان‌نامه کارشناسی ارشد آمار

تحلیل رگرسیونی داده‌های نرخ و نسبت با استفاده از توزیع سیمپلکس

نگارنده: سیما ملکی

استاد راهنما

دکتر داود شاهسونی

بهمن ۱۳۹۷

اگر شایستهٔ تقدیم باشد، به

خانواده و دوستانم

که همواره در امر تحصیل مشوق من بوده‌اند.

سپاس گزاری...

منت خدای را عزو جل که طاعتش موجب قربت است و به شکر اندرش مزید نعمت. هر نفسی که فرو می رود ممد حیات است و چون بر می آید مفرح ذات. پس در هر نفسی دو نعمت موجود است و بر هر نعمتی شکری واجب.

سپاس خود را به محضر استاد راهنما، آقای دکتر شاهسونی برای رهنمودهایشان تقدیم می دارم.

از استاد فرزانه، آقای دکتر باغیشنی و دانشجوی گرامی ایشان آقای میعادولی پور کمال قدردانی را دارم. از استاد گرانقدر، آقای دکتر ربیعی که همواره در این راه دلگرمی بنده بودند و مرا به صبر و شکیبایی دعوت نمودند؛ کمال تشکر را دارم.

کمال تشکر را دارم از خانم اسدی، خانم لاله زاده و آقای حسین زاده به پاس کمک های ارزنده شان که هم چون یک دوست و راهنمای بنده بودند.

در پایان از همه دوستان که همواره باعث دلگرمی بودند، نهایت تشکر را دارم.

سیما ملکی
بهمن ۱۳۹۷

تعهد نامه

این‌جانب **سیما ملکی** دانشجوی کارشناسی ارشد رشته **آمار** دانشکده **علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **تحلیل رگرسیونی داده‌های نرخ و نسبت با استفاده از توزیع سیمپلکس** ، تحت راهنمایی **داود شاهسونی** 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.

سیما ملکی
بهمن ۱۳۹۷

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

در مسائل رگرسیون، هنگامی‌که متغیر پاسخ از نوع نرخ یا نسبت در دامنه تغییرات $(0,1)$ است، به‌طور معمول از مدل‌های رگرسیونی لجستیک یا پروبیت استفاده می‌شود. در حالتی که توزیع این داده‌ها چوله باشد، مدل‌های مذکور مناسب نبوده و نمی‌توانند به خوبی به داده‌ها برازش شوند، زیرا توزیع‌های زیر بنایی آن‌ها یعنی لجستیک و نرمال متقارن بوده و با ماهیت چوله بودن در تعارض است. در این پایان‌نامه، مدل رگرسیونی سیمپلکس، به‌عنوان یک راه حل جایگزین، معرفی شده‌است. ابتدا توزیع سیمپلکس، مبتنی بر توزیع گوسی معکوس آمیخته را مطرح نموده و نحوه شبیه‌سازی داده‌های سیمپلکس را بر اساس شبیه‌سازی گوسی معکوس آمیخته بیان می‌کنیم. از آن‌جا که مدل رگرسیونی سیمپلکس در قالب «مدل‌های پراکندگی» قابل طرح است، لذا مدل‌های پراکندگی را معرفی نموده و همانند روش مدل‌های خطی تعمیم‌یافته، پارامترهای مرتبط را برآورد می‌کنیم. چون توزیع سیمپلکس شامل دو پارامتر است، مسئله برآورد را در دو حالت ثابت و متغیر بودن پارامتر واریانس، بررسی کرده و از مثال‌های شبیه‌سازی متنوع برای درک مطلب بهره می‌جوییم. از رگرسیون سیمپلکس برای مدل‌سازی داده‌های طولی نیز استفاده کرده و با استفاده از دو مثال واقعی، کارایی و عملکرد رگرسیون سیمپلکس را با این نوع داده‌ها، بررسی می‌کنیم.

کلمات کلیدی: مدل‌های پراکندگی، داده‌های نرخ و نسبت، تولید متغیر تصادفی، مدل‌های خطی تعمیم‌یافته، توزیع سیمپلکس، مدل رگرسیونی سیمپلکس، داده‌های طولی، معادلات برآوردیابی تعمیم‌یافته.

لیست مقالات مستخرج از پایان‌نامه

۱. ملکی، س.، شاهسونی، د. (۱۳۹۷)، ’’مدل‌بندی داده‌های نرخ و نسبت با استفاده از توزیع سیمپلکس‘‘، اولین سمینار دانشجویی آمار ایران، دانشگاه تربیت مدرس.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

تعاریف و مفاهیم اولیه

۱.۱ مقدمه

مدل‌سازی آماری، روابط بین متغیر پاسخ و متغیرهای تبیینی را بیان می‌کند که صورت کلی این مدل‌ها عبارت است از

مولفه خطای تصادفی + مولفه سیستماتیک = متغیر پاسخ

مولفه سیستماتیک حاوی تغییرات متغیر پاسخ برحسب متغیرهای تبیینی است و مولفه خطای تصادفی حاوی سهمی از تغییرات پاسخ است که توسط مولفه سیستماتیک بیان نشده است. فرمول‌بندی کلی یک مدل آماری برای n مقدار مشاهده‌شده متغیر پاسخ Y، به‌صورت

$$y_i = \eta_i + \epsilon_i, \qquad i = ۱, \cdots, n$$

بیان می‌شود که η_i همان مولفه سیستماتیک و ϵ_i مولفه خطاست. مولفه η_i می‌تواند نسبت به پارامترهای خود، صورت‌های خطی نظیر

$$\eta_i = \beta_۰ + \beta_۱ x_i$$

$$\eta_i = \beta_۰ + \beta_۱ x_i + \beta_۲ x_i^۲$$

$$\eta_i = \beta_۰ + \beta_۱ x_{i۱} + \beta_۲ x_{i۲} + \cdots + \beta_p x_{ip}$$

داشته باشد یا این‌که نظیر

$$\eta_i = \beta_۰ + \exp\{\beta_۱ x_i\}$$

$$\eta_i = x_i/(\beta_۰ + \beta_۱ x_i)$$

نسبت به پارامترها غیرخطی باشد، که در آن $x_{i۱}, \cdots, x_{ip}$ مقادیر متغیرهای تبیینی $X_۱, \cdots, X_p$ هستند.

فرض کنید Y_i به‌صورت خطی به مقادیر این p متغیر تبیینی $X_۱, \cdots, X_p$ وابسته باشد. آن‌گاه با فرض این‌که

$$\eta_i = \beta_۰ + \beta_۱ x_{i۱} + \beta_۲ x_{i۲} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

مدل مفروض را یک مدل خطی کلی[۱]، $\eta_i = \sum_{j=۰}^{p} \beta_j x_{ij}$ را «پیشگوی خطی»[۲] نامند. این مدل طیف وسیعی از مباحث آماری را به خود اختصاص داده به‌طوری‌که در وضعیت متداول هم‌خطی متغیرهای تبیینی، مباحثی از قبیل رگرسیون جریمه‌شده، رگرسیون مولفه‌های اصلی، رگرسیون کمترین توان‌های دوم جزئی خودنمایی می‌کنند.

درمدل‌های خطی، پذیره‌هایی حاکم هستند که موجب تسهیل در استنباط آماری می‌شوند. اغلب فرض می‌شود که ϵ_i ها متغیرهای تصادفی ناهمبسته با میانگین صفر و واریانس مشترک $\sigma^۲$ هستند. یعنی

$$\mu_i = \mathrm{E}(Y_i) = \eta_i, \qquad \mathrm{V}(Y_i) = \sigma^۲.$$

علاوه بر این، با اعمال پذیره توزیعی نرمال برای ϵ ها، $\epsilon_i \sim \mathrm{N}(۰, \sigma^۲)$، داریم

$$Y_i \sim \mathrm{N}(\eta_i, \sigma^۲).$$

موارد فوق را به‌صورت زیر خلاصه می‌کنیم

۱. Y_i ها متغیرهای تصادفی مستقل با میانگین μ_i هستند، $\mathrm{E}(Y_i) = \mu_i$، و هم‌چنین $Y_i \sim \mathrm{N}(\mu_i, \sigma^۲)$.

۲. تاثیر p متغیر تبیینی بر متغیر پاسخ، با مولفه سیستماتیک خطی مدل‌سازی می‌شود، یعنی $\eta_i = \sum_{j=۰}^{p} \beta_j x_{ij}$.

۳. ارتباط بین ۱ و ۲ عبارت است از $\mu_i = \eta_i$.

صور سه‌گانه بالا را می‌توان به‌صورت زیر تعمیم داد

۱. توزیع Y_i لزوما نرمال نیست.

[1] General Linear Model

[2] Linear Predictor

۲. μ_i می‌تواند تابعی از η_i باشد، یعنی $\mu_i = g(\eta_i)$.

۳. مولفه سیستماتیک می‌تواند غیرخطی باشد.

این مفاهیم مبانی «مدل‌های خطی تعمیم‌یافته»[۳] (GLM) (مک کالا و نلدر، ۱۹۸۹) را شکل می‌دهند. در واقع مدل‌های خطی تعمیم‌یافته، تعمیمی از مدل‌های خطی کلی هستند. نظریه مدل‌های خطی تعمیم‌یافته، حاکی از آن است که تحلیل رگرسیون نیازمند تشخیص نوع متغیر پاسخ است. علیرغم معروف بودن و پرکاربرد بودن توزیع نرمال، جرجنسن (۱۹۹۷) به این نکته اشاره داشت که به‌جز برای داده‌های با پراکندگی کم، توزیع نرمال یک استثنا است تا این‌که یک قانون باشد. فیشر (۱۹۵۳)، آماردان برجسته، نیز به ما یادآوری کرد که توصیف داده‌ها باید با توجه به ذات و ماهیت طبیعی آن‌ها صورت گیرد.
نلدر و ودربرن (۱۹۷۲) اولین کسانی بودند که توانستند با معرفی رده GLM نشان دهند که طیف وسیعی از داده‌های غیر نرمال را می‌توان با یک چارچوب واحد تحلیل کرد. در واقع GLM برای خانواده توزیع‌های نمایی بسط داده شده، یعنی

۱. متغیر تصادفی Y_i دارای توزیعی در خانواده توزیع‌های نمایی است.

۲. متغیرهای تصادفی از طریق مولفه سیستماتیک وارد مدل می‌شوند، $\eta_i = \sum_{j=0}^{p} \beta_j x_{ij}$.

۳. میانگین Y_i یعنی μ_i، از طریق تابعی به نام «تابع پیوند»[۴] g به η_i مرتبط می‌شود، یعنی $\eta_i = g(\mu_i)$.

هنگامی‌که متغیر پاسخ از جنس نرخ و نسبت است، مدل‌های معدودی برای برازش چنین داده‌هایی معرفی شده و سودمند هستند. در توزیع بتا که اغلب در آمار بیزی مورد استفاده قرار می‌گیرد، می‌توان با تبدیل مناسبی از پارامترها، آن را برای مدل‌سازی چنین متغیرهای پاسخی به‌کار برد. فراری و سریباری (۲۰۰۴) مدل رگرسیونی بتا را برای پاسخ‌های نرخ و نسبت به‌کار بردند و روش‌های برآورد و استنباط آماری مرتبط، در نرم‌افزار R توسط بسته betareg عملیاتی شد (سریباری و همکاران، ۲۰۱۰).
با این وجود، وقتی پارامتر پراکندگی به مدل‌سازی اضافه می‌شود، مدل GLM بر مبنای توزیع سیمپلکس[۵] (بارندروف، نیلسن و جرجنسن، ۱۹۹۱) برای تحلیل داده‌های پیوسته از جنس نسبت، نیرومندتر شناخته شد (ژنگ و ژو، ۲۰۱۴). توزیع سیمپلکس شامل رده وسیعی از توزیع‌هایی است که دامنه آن‌ها در بازه $(0,1)$ است. در واقع توزیع سیمپلکس یک مدل پراکندگی است و خواص تحلیلی مشترکی با مدل‌های پراکندگی نمایی دارد.

[3] Generalized Linear Models
[4] Link Function
[5] Simplex Distribution

در بحث تحلیل داده‌های طولی نیز، توزیع سیمپلکس نقش موثری را ایفا کرده است. مدل‌های حاشیه‌ای سیمپلکس[۶] (سانگ و تن، ۲۰۰۰) و مدل‌های اثرات آمیخته سیمپلکس (ژو و همکاران، ۲۰۰۸) نیز از جمله این موارد هستند.

در این پایان‌نامه، با معرفی توزیع سیمپلکس در فصل ۲ آغاز کرده و رابطه آن با توزیع گوسی معکوس آمیخته[۷] و نحوه شبیه‌سازی اعداد تصادفی سیمپلکس را شرح می‌دهیم. سپس مدل خطی تعمیم‌یافته سیمپلکس را در فصل ۳ مورد بررسی قرار می‌دهیم. در فصل ۴ به معرفی داده‌های طولی[۸] و روش معادلات برآوریابی تعمیم‌یافته[۹] (GEEs) می‌پردازیم. سرانجام در فصل ۵ به برازش و تحلیل داده‌ها در مدل‌های رگرسیونی سیمپلکس با استفاده از بسته simplexreg موجود در نرم‌افزار R، می‌پردازیم و سپس کاربرد مدل را با دو مثال واقعی نمایش می‌دهیم.
برای ورود به فصل ۳، نیازمند معرفی اجمالی مدل‌های پراکندگی هستیم که در بخش ۲.۱ به آن می‌پردازیم. لازم به ذکر است که فصل ۲، کاملا مستقل از مفهوم رگرسیون و مدل رگرسیونی سیمپلکس است و در صورتی‌که خواننده صرفا به بحث رگرسیونی آن علاقه‌مند باشد می‌تواند فقط با دانستن چگالی سیمپلکس در بخش ۳.۲، بحث رگرسیون سیمپلکس را از فصل ۳ دنبال کند.

۲.۱ مدل‌های پراکندگی

ایده اصلی حاکم بر GLM، توسط جرجنسن (۱۹۹۷) به رده بزرگ‌تری از مدل‌ها که «مدل‌های پراکندگی»[۱۰] (DM) نامیده می‌شوند، تعمیم داده شد. در این مدل‌ها عقیده کلی در مورد «مکان» و «مقیاس» به «موقعیت»[۱۱] و «پراکندگی» تعمیم می‌یابد. در این صورت طیف وسیعی از داده‌های غیر نرمال، پوشش داده می‌شود و گونه‌هایی نظیر داده‌های مثبت، مثبت و صفر، داده‌های شمارشی، داده‌های دودویی و داده‌های جهتی[۱۲] می‌توانند توسط مدل‌های پراکندگی و با حفظ ماهیت ذاتی آن‌ها مورد تحلیل و مدل‌سازی قرار گیرند؛ در واقع رده مهمی از مدل‌های پراکندگی، مدل‌های خطی تعمیم‌یافته است که در آن نقش مجموع توان‌های دوم، به‌عنوان مبنا، به مفهوم دیگری به نام «انحراف»[۱۳] که ابزاری مهم برای برازش و انتخاب مدل است، تعمیم داده می‌شود.
برای تشریح مدل‌های پراکندگی، توزیع نرمال را که نقش اساسی در رگرسیون خطی کلاسیک

[6] Simplex Marginal Models
[7] Mixture Inverse Gaussian Distribution
[8] Longitudinal Data
[9] Generalized Estimating Equations
[10] Dispersion Models
[11] Position
[12] Directional
[13] Deviance

ایفا می‌کند، در نظر بگیرید. در چگالی $\mathrm{N}(\mu,\sigma^۲)$ یعنی

$$f(x)=\frac{۱}{\sqrt{۲\pi\sigma^۲}}\exp\left\{-\frac{(y-\mu)^۲}{۲\sigma^۲}\right\}$$

جمله $(y-\mu)^۲$ یک فاصله اقلیدسی است که تفاوت[۱۴] بین مشاهده y و میانگین μ را اندازه‌گیری می‌کند. این معیار اندازه‌گیری تفاوت، در تحلیل رگرسیون، برای تشکیل آماره F و دستیابی به نیکویی برازش استفاده می‌شود.
با الگوپذیری از چگالی نرمال، جرجنسن (۱۹۹۲) مدل‌های پراکندگی را ارائه داد که در آن فاصله اقلیدسی $(y-\mu)^۲$ به‌صورت کلی‌تری از مفهوم تفاوت و با نماد $d(y;\mu)$ در نظر گرفته می‌شود.

تعریف ۱.۲.۱. مدل پراکندگی $\mathrm{DM}(\mu,\sigma^۲)$ با پارامتر مکان μ و پارامتر پراکندگی $\sigma^۲$، خانواده‌ای از توزیع‌هاست که دارای تابع چگالی به‌صورت زیر است:

$$f(y;\mu,\sigma^۲)=a(y;\sigma^۲)\exp\left\{\frac{-۱}{۲\sigma^۲}d(y;\mu)\right\}\qquad y\in\mathcal{C}\qquad(۱.۱)$$

که در آن $\mu\in\Omega$، $\sigma^۲>۰$ و ثابت نرمال‌ساز $a\geq۰$ و مستقل از μ است و عموما $\Omega\subseteq\mathcal{C}\subseteq\mathbb{R}$.
از آن‌جا که ثابت نرمال‌ساز شامل μ نیست، لذا برآورد μ (یا β در GLMها) به‌طور مجزا از برآورد $\sigma^۲$ انجام می‌شود که سهولت برآورد پارامترها را در بر خواهد داشت. این خاصیت بسیار مفید که به‌عنوان متعامدسازی درستنمایی[۱۵] شناخته می‌شود، در توزیع نرمال برقرار بوده و در مدل‌های پراکندگی نیز پا برجاست.

تابع دو متغیره $d(.,.)$ را تابع انحراف[۱۶] نامیده که $(y,\mu)\in\mathcal{C}\times\Omega$ و دارای دو شرط زیر است:

۱. وقتی‌که مشاهده y و μ مورد انتظار باهم برابر باشند، $y=\mu$، آن‌گاه

$$d(y;y)=۰,\qquad\forall y\in\Omega.$$

۲. وقتی که مشاهده y و μ مورد انتظار باهم برابر نباشند، آن‌گاه

$$d(y;\mu)>۰,\qquad\forall y\neq\mu.$$

تابع انحراف، منظم[۱۷] نامیده می‌شود اگر مشتق‌های مرتبه دوم تابع $d(y;\mu)$ نسبت به (y,μ) با هم برابر باشند

$$\frac{\partial^۲ d}{\partial y^۲}(y;y)=\frac{\partial^۲ d}{\partial\mu^۲}(y;\mu)\Big|_{\mu=y}>۰,\qquad\forall y\in\Omega.$$

[14] Discrepancy
[15] Likelihood Orthogonality
[16] Deviance
[17] Regular

برای تابع انحراف منظم، تابع واریانس[18] $\mathrm{V}:\Omega\longrightarrow(۰,\infty)$ به‌شرح زیر تعریف می‌شود

$$\mathrm{V}(\mu)=\frac{۲}{\frac{\partial^{۲}d}{\partial\mu^{۲}}(y;\mu)\big|_{y=\mu}},\qquad \mu\in\Omega. \tag{۲.۱}$$

برخی از مدل‌های پراکندگی معروف به‌همراه تابع انحراف d و تابع واریانس V متناظرشان در جدول ۱.۱ معرفی شده‌اند.

جدول ۱.۱: تابع انحراف و تابع واریانس برخی مدل‌های پراکندگی

توزیع	d	$\mathcal{C}$	Ω	$\mathrm{V}(\mu)$
نرمال	$(y-\mu)^{۲}$	$(-\infty,\infty)$	$(-\infty,\infty)$	۱
پواسون	$۲(y\log\frac{y}{\mu}-y+\mu)$	$\{۰,۱,\dots\}$	$(۰,\infty)$	μ
گوسی معکوس	$\frac{(y-\mu)^{۲}}{y\mu^{۲}}$	$(۰,\infty)$	$(۰,\infty)$	$\mu^{۳}$
سیمپلکس	$\frac{(y-\mu)^{۲}}{y\,(۱-y)\,\mu^{۲}\,(۱-\mu)^{۲}}$	$(۰,۱)$	$(۰,۱)$	$\mu^{۳}(۱-\mu)^{۳}$

در دو مثال زیر، چگونگی قرار گرفتن توزیع‌های نرمال و پواسون را در قالب مدل‌های پراکندگی بررسی می‌کنیم.

مثال ۱.۲.۱. توزیع نرمال $\mathrm{N}(\mu,\sigma^{۲})$ را در نظر بگیرید که در آن $d(y;\mu)=(y-\mu)^{۲}$ و $y\in\mathcal{C}=\mathbb{R}$ و همچنین $\mu\in\Omega=\mathbb{R}$. واضح است که تابع d نامنفی است و در صورتی‌که $y=\mu$، برابر با صفر است. در نتیجه مشتق‌های مرتبه اول و دوم d به‌صورت زیر است:

$$\frac{\partial d}{\partial\mu}=-۲(y-\mu),\qquad \frac{\partial^{۲}d}{\partial\mu^{۲}}=۲$$

سپس تابع واریانس را با توجه به رابطه (۲.۱) می‌توان به‌دست آورد

$$\mathrm{V}(\mu)=\frac{۲}{۲}=۱$$

مثال ۲.۲.۱. توزیع پواسون با پارامتر μ را در نظر بگیرید

$$f(y;\mu)=\frac{\mu^{y}}{y!}e^{-\mu},\quad y\in\{۰,۱,\dots\};\mu\in\Omega=(۰,\infty),$$

[18] Variance Function

که معادل است با

$$f(y;\mu)=\frac{۱}{y!}\exp\{y\log\mu-\mu\}\cdot$$

در این‌جا $\{y\log\mu-\mu\}$ یک تابع انحراف نیست؛ زیرا با توجه به شرایطی که d باید دارا باشد، وقتی $y=\mu$ داریم $d\neq ۰$. حال رابطه فوق را در عبارت $\exp\{y\log y-y\}$ ضرب و تقسیم می‌کنیم؛ در نتیجه داریم

$$p(y;\mu)=\left\{\frac{۱}{y!}\exp(y\log y-y)\right\}\exp\left\{-\frac{۱}{۲}۲(y\log y-y-y\log\mu+\mu)\right\}.$$

حال با توجه به رابطه بالا و تعریف DM در (۱.۱) داریم

$$d(y;\mu)=۲(y\log\frac{y}{\mu}-y+\mu)$$

$$a(y,\sigma^۲)=\frac{۱}{y!}\exp\{y\log y-y\}$$

$$\sigma^۲=۱۰$$

برای نشان دادن این‌که تابع d یک تابع انحراف منظم است، کافی‌است نشان دهیم که این تابع محدب دارای مینیمم یکتای صفر است. برای هر μ داریم

$$\frac{\partial d}{\partial y}=۲(\log y-\log\mu),\qquad \frac{\partial^۲}{\partial y^۲}=\frac{۲}{y}\cdot$$

واضح است که مشتق اول منفی است اگر $y<\mu$ و مثبت است اگر $y>\mu$، و به این معنی است که تابع انحراف d یک تابع محدب با مینیمم مقدار یکتای صفر در $y=\mu$ است؛ در نتیجه با توجه به تعریف تابع واریانس در رابطه (۲.۱) داریم

$$\mathrm{V}(\mu)=\frac{۲}{\frac{\partial^۲ d}{\partial\mu^۲}}=\frac{۲}{\frac{۲y}{\mu^۲}}=\frac{۲}{\frac{۲}{\mu}}=\mu\cdot$$

۱.۲.۱ خواص مدل‌های پراکندگی

خاصیت ۱. اگر d تابع انحراف منظم باشد آن‌گاه

$$\frac{\partial^۲ d}{\partial y^۲}(y;y)=\frac{\partial^۲ d}{\partial\mu^۲}(y;y)=-\frac{\partial^۲ d}{\partial\mu\partial y}(y;y),\qquad \forall y\in\Omega. \tag{۳.۱}$$

برهان. با استفاده از تعریف انحراف داریم

$$d(y;y)=d(\mu;\mu)=۰,\qquad d(y;\mu)\geq ۰,\forall y,\mu\in\Omega$$

یعنی $d(y;.)$ دارای یک مینیمم یکتا در y است و به‌طور مشابه $d(.;\mu)$ دارای یک مینیمم یکتا در μ است، بنابراین

$$\frac{\partial d}{\partial\mu}(y;y)=\frac{\partial d}{\partial y}(y;y)=۰. \tag{۴.۱}$$

نتیجه (۳.۱) به سادگی با استفاده از مشتق دو معادله (۴.۱) نسبت به y حاصل می‌شود.
□

خاصیت ۲. بسط تیلور انحراف منظم در نزدیکی $(\mu_\circ, \mu_\circ)$ به‌صورت زیر است

$$d(\mu_\circ + x\delta; \mu_\circ + m\delta) = \frac{\delta^۲}{\mathrm{V}(\mu_\circ)}(x-m)^۲ + O(\delta^۲)$$

که در آن $\mathrm{V}(.)$ همان تابع واریانس است.

برهان. با توجه به (۴.۱) داریم

$$\begin{aligned} d(\mu_\circ + x\delta; \mu_\circ + m\delta) &= d(\mu_\circ, \mu_\circ) + \frac{\partial d}{\partial \mu}(x\delta) + \frac{\partial d}{\partial y}(m\delta) \\ &+ \frac{۱}{۲}\frac{\partial d^۲}{\partial \mu^۲}(\mu_\circ, \mu_\circ)(\delta^۲ x^۲) + \frac{۱}{۲}۲\frac{\partial d^۲}{\partial \mu \partial y}(\mu_\circ, \mu_\circ)(\delta m)(\delta x) \\ &+ \frac{۱}{۲}\frac{\partial d^۲}{\partial y^۲}(\mu_\circ, \mu_\circ)(\delta^۲ m^۲) + O(\delta^۲) \end{aligned}$$

از آن‌جا که سه جمله اول سمت راست عبارت بالا صفر هستند، بنابراین داریم:

$$\begin{aligned} d(\mu_\circ + x\delta; \mu_\circ + m\delta) &= \frac{\delta^۲}{\mathrm{V}(\mu_\circ)}x^۲ + \frac{\delta^۲}{\mathrm{V}(\mu_\circ)}۲xm + \frac{\delta^۲}{\mathrm{V}(\mu_\circ)}m^۲ + O(\delta^۲) \\ &= \frac{\delta^۲}{\mathrm{V}(\mu_\circ)}(x-m)^۲ + O(\delta^۲). \end{aligned}$$

□

از آن‌جایی که ثابت نرمال‌ساز $a(.)$ صورت بسته‌ای ندارد و به‌دست آوردن برآورد پارامتر پراکندگی $\sigma^۲$ کار دشواری است؛ تقریبی از ثابت نرمال‌ساز $a(.)$، که تقریب نقطه‌زینی برای چگالی با پراکندگی کوچک است، را با استفاده از گزاره زیر به‌دست می‌آوریم.

خاصیت ۳. (**تقریب نقطه زینی**[19]) هرگاه $\sigma^۲ \longrightarrow \circ$، چگالی یک مدل پراکندگی منظم، DM، می‌تواند به‌صورت زیر تقریب شود

$$f(y; \mu, \sigma^۲) \simeq \{۲\pi\sigma^۲\mathrm{V}(y)\}^{-\frac{۱}{۲}} \exp\{-\frac{۱}{۲\sigma^۲}d(y;\mu)\}$$

که رابطه فوق معادل است با این‌که بگوییم وقتی $\sigma^۲ \longrightarrow \circ$، ثابت نرمال‌ساز به‌صورت زیر تقریب می‌شود:

$$a(y; \sigma^۲) \simeq \{۲\pi\sigma^۲\mathrm{V}(y)\}^{-\frac{۱}{۲}} \qquad (۵.۱)$$

[19] Saddlepoint Approximation

که در آن $\mathrm{V}(.)$ همان تابع واریانس است. برای اثبات به بارندروف، نیلسن و کاکس (۱۹۸۹) رجوع کنید.

از خواص ۲ و ۳ می‌توان خاصیت نرمال مجانبی را که در زیر آمده است، نتیجه گرفت.

خاصیت ۴. (نرمال مجانبی) فرض کنید $Y \sim \mathrm{DM}(\mu_\circ + \sigma\mu, \sigma^\mathrm{۲})$ یک مدل پراکندگی با تقریب نقطه زینی همگرای یکنواخت است، یعنی همگرایی در (۵.۱) یکنواخت است؛ داریم

$$\frac{Y-\mu_\circ}{\sigma} \xrightarrow{d} \mathrm{N}(\mu, \mathrm{V}(\mu_\circ)), \quad \sigma^\mathrm{۲} \longrightarrow \circ. \tag{۶.۱}$$

به‌عبارتی دیگر، برای پراکندگی‌های کوچک داریم

$$\mathrm{DM}(\mu_\circ + \sigma\mu, \sigma^\mathrm{۲}) \simeq \mathrm{N}(\mu_\circ + \sigma\mu, \sigma^\mathrm{۲}\mathrm{V}(\mu_\circ)).$$

۲.۲.۱ برآورد درستنمایی ماکسیمم

در این بخش به برآورد درستنمایی ماکسیمم در GLM بر اساس مدل‌های پراکندگی می‌پردازیم. یک مجموعه داده مقطعی (y_i, x_i)، $i = \mathrm{۱}, \ldots, m$ را در نظر بگیرید، که در آن y_iها مصداقی از متغیرهای تصادفی مستقل و هم توزیع Y_i هستند $Y_i \sim \mathrm{DM}(\mu_i, \sigma^\mathrm{۲})$ و $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$. قرار می‌دهیم $\mathbf{y} = (y_\mathrm{۱}, \ldots, y_m)^T$ و $\mu = (\mu_\mathrm{۱}, \ldots, \mu_m)^T$. درستنمایی برای بردار پارامتر $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^\mathrm{۲})$ به‌صورت زیر است

$$\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=\mathrm{۱}}^{m} a(y_i; \sigma^\mathrm{۲}) \exp\left\{-\frac{\mathrm{۱}}{\mathrm{۲}\sigma^\mathrm{۲}} d(y_i; \mu_i)\right\}, \boldsymbol{\beta} \in \mathbb{R}^{p+\mathrm{۱}}, \sigma^\mathrm{۲} > \circ.$$

لگاریتم درستنمایی نیز به‌صورت زیر بیان می‌شود

$$\begin{aligned} \ell(\boldsymbol{\theta}; \boldsymbol{y}) &= \sum_{i=\mathrm{۱}}^{m} \log a(y_i; \sigma^\mathrm{۲}) - \frac{\mathrm{۱}}{\mathrm{۲}\sigma^\mathrm{۲}} \sum_{i=\mathrm{۱}}^{m} d(y_i; \mu_i) \\ &= \sum_{i=\mathrm{۱}}^{m} \log a(y_i; \sigma^\mathrm{۲}) - \frac{\mathrm{۱}}{\mathrm{۲}\sigma^\mathrm{۲}} D(\boldsymbol{y}; \mu) \end{aligned} \tag{۷.۱}$$

که در آن $\mu_i = \mu_i(\boldsymbol{\beta})$ یک تابع غیرخطی از $\boldsymbol{\beta}$ و $D(\boldsymbol{y}; \mu) = \sum_{i=\mathrm{۱}}^{m} d(y_i; \mu_i)$ فقط وابسته به $\boldsymbol{\beta}$ است. تابع D مشابه مجموع توان‌های دوم باقی‌مانده در مدل رگرسیون خطی است.

تابع امتیاز برای ضریب رگرسیونی β به‌صورت زیر است:

$$\mathrm{U}(\mathbf{y}; \boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = -\frac{\mathrm{۱}}{\mathrm{۲}\sigma^\mathrm{۲}} \sum_{i=\mathrm{۱}}^{m} \frac{\partial d(y_i; \mu_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}.$$

iامین پیشگوی خطی را با $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ و همچنین تابع امتیاز انحراف[۲۰] را با u نمادگذاری می‌کنیم. داریم

$$u(y_i; \mu_i) = -\frac{\mathrm{۱}}{\mathrm{۲}} \frac{\partial d(y_i; \mu_i)}{\partial \mu_i}, \quad i = \mathrm{۱}, \ldots, m. \tag{۸.۱}$$

[20] Deviance Scores Function

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \{g'(\mu_i)\}^{-1}\mathbf{x}_i$$

که در آن $g'(\mu_i)$ مشتق مرتبه اول تابع پیوند، نسبت به μ است. در جدول ۲.۱ برخی از توابع پیوند و مشتق آن‌ها آمده است.

جدول ۲.۱: برخی توابع پیوند و مشتق اول آن‌ها

مدل	تابع پیوند	مشتق تابع پیوند	Ω
سیمپلکس، دوجمله‌ای	$\log(\frac{\mu}{1-\mu})$	$\frac{1}{\mu(1-\mu)}$	$\mu \in (0,1)$
پواسون، گوسی معکوس، دوجمله‌ای منفی	$\log(\mu)$	$\frac{1}{\mu}$	$\mu \in (0,\infty)$
گاما	$\frac{1}{\mu}$	$-\frac{1}{\mu^2}$	$\mu \in (0,\infty)$

حال تابع امتیاز را به‌صورت زیر می‌توان نوشت

$$\mathrm{U}(\mathbf{y};\boldsymbol{\beta}) = \frac{1}{\sigma^2}\sum_{i=1}^{m}\mathbf{x}_i\frac{1}{g'(\mu_i)}u(y_i;\mu_i). \qquad (۹.۱)$$

برآورد درستنمایی ماکسیمم پارامتر $\boldsymbol{\beta}$ از حل معادله زیر حاصل می‌شود

$$\sum_{i=1}^{m}\mathbf{x}_i\frac{1}{g'(\mu_i)}u(y_i;\mu_i) = 0. \qquad (۱۰.۱)$$

باید دقت داشت که معادله (۱۰.۱) شامل پارامتر پراکندگی σ^2 نیست؛ تحت شرایط نظم، برآوردگر درستنمایی ماکسیمم $\hat{\boldsymbol{\beta}}_m$ یک برآوردگر سازگار برای $\boldsymbol{\beta}$ است، یعنی

$$\hat{\boldsymbol{\beta}}_m \xrightarrow{P} \boldsymbol{\beta}, m \longrightarrow \infty$$

و دارای نرمال مجانبی با میانگین صفر و ماتریس کواریانس $\mathrm{I}^{-1}(\boldsymbol{\theta})$ است که $\mathrm{I}(\boldsymbol{\theta})$ ماتریس اطلاع فیشر است که به‌صورت زیر بیان می‌شود

$$\begin{aligned}
\mathrm{I}(\boldsymbol{\theta}) &= -\mathrm{E}\{\mathrm{U}'(\mathbf{Y};\boldsymbol{\beta})\} \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{m}\mathbf{x}_i\frac{1}{\{g'(\mu_i)\}^2}\mathrm{E}\{-u'(Y_i;\mu_i)\}\mathbf{x}_i^T \\
&= \sum_{i=1}^{m}\mathbf{x}_i w_i \mathbf{x}_i^T \\
&= \mathbf{X}^T W \mathbf{X}.
\end{aligned}$$

که در آن X یک ماتریس با بعد $m \times (p+1)$ است که iامین سطر آن x_i می‌باشد. هم‌چنین W یک ماتریس قطری است که iامین عنصر قطری آن w_i است که به‌صورت زیر بیان می‌شود

$$w_i = \frac{1}{\sigma^2} \frac{1}{\{g'(\mu_i)\}^2} \mathrm{E}\{-u'(Y_i, \mu_i)\}, \quad i = 1, \ldots, m.$$

زمانی‌که پارامتر پراکندگی در مدل وجود داشته باشد، یعنی اگر $a(y, \sigma^2)$ به گونه‌ای باشد که مشتق‌گیری آن مقدور باشد، برای برآورد درستنمایی ماکسیمم آن مشابه ثابت نرمال‌ساز، کافی است مانند توزیع نرمال از آن مشتق گرفت؛ اما زمانی که ثابت نرمال‌ساز دارای صورت بسته نباشد، مشتق آن نسبت به σ^2 دارای حل عددی پیچیده‌ای است که برای رفع این مشکل و برآورد پارامتر σ^2 دو روش پیشنهاد می‌شود. اولین روش، استفاده از خاصیت نرمال مجانبی کوچک پراکندگی است که در آن

$$\log a(y; \sigma^2) \simeq -\frac{1}{2} \log \sigma^2.$$

با اعمال این تقریب در رابطه (۷.۱)، تقریب لگاریتم درستنمایی نسبت به σ^2 به‌صورت زیر است

$$-\frac{m}{2\sigma^2} + \frac{1}{2\sigma^4} D(y; \mu) = 0$$

که منجر به برآورد σ^2 به‌صورت زیر می‌شود

$$\widehat{\sigma^2} = \frac{1}{m} D(y; \hat{\mu}) = \frac{1}{m} \sum_{i=1}^{m} d(y_i, \hat{\mu}_i).$$

این برآوردگر را برآوردگر جرجنسن می‌نامند که میانگینی از برآورد انحراف است. برآوردگر جرجنسن در حالت کلی ناریب نیست، حتی اگر روی درجه‌های آزادی آن تصحیح انجام دهیم، یعنی $m - (p+1)$ را جایگزین m کنیم. این رابطه فقط وقتی پیشنهاد می‌شود که پارامتر σ^2 کوچک (< 5) باشد.

روش دوم برای به‌دست آوردن یک برآوردگر ناریب از σ^2 استفاده از ویژگی گشتاوری است که در زیر به آن می‌پردازیم.

گزاره ۱.۲.۱. اگر $Y \sim DM(\mu, \sigma^2)$ و $d(y; \mu)$ یک انحراف منظم باشد، آن‌گاه

$$\mathrm{E}\{u(Y; \mu)\} = 0,$$

$$\mathrm{V}\{u(Y; \mu)\} = \sigma^2 \mathrm{E}\{-u'(Y; \mu)\}$$

که در آن u' مشتق اول نسبت به μ، در رابطه (۸.۱) است.

برهان. می‌دانیم که $\int f(y; \mu, \sigma^2) \mathrm{d}y = 1$، حال نسبت به μ از آن مشتق گرفته و برابر صفر قرار می‌دهیم

$$-\frac{1}{2\sigma^2} \int d'(y; \mu) f(y; \mu, \sigma^2) \mathrm{d}y = 0,$$

یا این‌که می‌توان گفت $\mathrm{E}\{d'(Y;\mu)\}=0$. بار دیگر نسبت به μ مشتق گرفته و نتیجه می‌شود

$$-\frac{1}{2\sigma^2}\int\{d'(y;\mu)\}^2 f(y;\mu,\sigma^2)\mathrm{d}y+\int d''(y;\mu)f(y;\mu,\sigma^2)\mathrm{d}y=0$$

یا به‌طور معادل، گوییم

$$\mathrm{E}\{d''(Y;\mu)\}=\frac{1}{2\sigma^2}\mathrm{E}\{d'(Y;\mu)\}^2=\frac{1}{2\sigma^2}\mathrm{V}\{d'(Y;\mu)\}.$$

با توجه به رابطه (۸.۱)، می‌توان رابطه بالا را بازنویسی کرد؛ که عبارت است از

$$\mathrm{V}\{u(Y;\mu)\}=\sigma^2\mathrm{E}\{-u'(Y;\mu)\}. \qquad (11.1)$$

□

بر اساس این نتیجه می‌توان پارامتر پراکندگی σ^2 را با روش گشتاورها برآورد کرد. این برآوردگر عبارت است از

$$\widehat{\sigma^2}=\frac{\sum_{i=1}^m(u_i-\bar{u})^2}{\sum_{i=1}^m(-u'_i)},$$

که در آن $u_i=u(y_i,\hat{\mu}_i)$، $u'_i=u'(y_i,\hat{\mu}_i)$ و $\bar{u}=\frac{1}{m}\sum_{i=1}^m u_i$.

۳.۱ معیارهای اندازه‌گیری دقت مدل

در آمار معیارهایی برای ارزیابی توانایی مدل‌ها در برآورد متغیر پاسخ وجود دارند که با استفاده از آن‌ها می‌توان نتایج مدل‌های مختلف را با یک‌دیگر مقایسه کرد. این معیارها با اندازه‌گیری خطای برآورد یا دقت برآورد یک مدل، نتایج را ارزیابی و مقایسه می‌کنند. معیارهای متفاوتی برای این هدف وجود دارند که در این بخش، به چند معیار اشاره خواهیم کرد.

میانگین توان دوم خطا

یکی از راه‌های اندازه‌گیری خطای برآورد مدل، معیار میانگین توان دوم خطا [۲۱] (MSE) است. برای مقایسه برآوردگرهای اریب با برآوردگرهای نااریب و یا مقایسه دو برآوردگر یک پارامتر، با اریبی‌های متفاوت، MSE معیار مفیدی است و مقدار آن به‌صورت زیر محاسبه می‌گردد:

$$\mathrm{MSE}\left(\hat{\theta}\right)=\mathrm{E}\left[\left(\theta-\hat{\theta}\right)^2\right]$$

که در آن θ مقدار واقعی پارامتر و $\hat{\theta}$ مقدار برآورد شده پارامتر است. در صورتی‌که $\mathrm{MSE}\,(\theta)$ به سمت صفر میل کند، بیان‌گر سازگار [۲۲] بودن $\hat{\theta}$ می‌باشد.

[21] Mean Squared Error

[22] Consistency

اریبی

آماره $\delta(X)$ که X یک متغیر تصادفی با توزیعی از خانواده توزیع‌های $P_\theta : \theta \in \Theta$ است، برای $\gamma(\theta)$ اریب[۲۳] گفته می‌شود اگر برای یک $\theta \in \Theta$،

$$\mathrm{E}_\theta[\delta(X)] \neq \gamma(\theta)$$

آن‌گاه اریبی برآوردگر $\delta(X)$ در برآورد $\gamma(\theta)$ به‌صورت زیر است:

$$\mathrm{b}(\theta) = \mathrm{E}_\theta[\delta(X)] - \gamma(\theta)$$

معیار اطلاع آکاییک

معیار اطلاع آکاییک[۲۴] (AIC)، معیاری برای انتخاب بهترین مدل آماری از بین چند نامزد است. این معیار، نیکویی برازش مدل را همراه با پیچیدگی، به‌طور توام، در نظر می‌گیرد که توسط آکاییک در سال ۱۹۷۴ پیشنهاد شد. با داده‌های مشاهده‌شده، چند مدل رقیب بر حسب مقدار AIC رتبه‌بندی می‌شوند و مدل با کمترین AIC بهترین است. در حالت کلی، AIC به‌صورت

$$\mathrm{AIC} = -۲\log(\mathrm{L}) + ۲p$$

تعریف می‌شود که در آن p تعداد پارامترهای مدل و L مقدار ماکسیمم تابع درستنمایی برای مدل برآوردشده است. به‌طور مجانبی مینیمم کردن AIC معادل مینیمم کردن آماره CV است. این نتیجه برای هر مدلی، نه فقط مدل خطی، برقرار است. این ویژگی باعث شده است زمانی که هدف پیشگویی است، استفاده از AIC توصیه شود.

معیار اطلاع بیزی

معیار اطلاع بیزی[۲۵] (BIC) به‌صورت زیر

$$\mathrm{BIC} = -۲\log(\mathrm{L}) + p\log(n)$$

تعریف می‌شود که در آن n حجم نمونه است. به دلیل جریمه سنگین‌تر این معیار نسبت به AIC، مدلی که توسط BIC انتخاب می‌شود، یا همان مدل منتخب بر اساس AIC است یا مدلی است با تعداد پارامتر کمتر.

[23] Bias
[24] Akaike Information Criterion
[25] Bayesian Information Criterion

فصل ۲

توزیع سیمپلکس

مقدمه

با توجه به هدف این پایان‌نامه که بررسی مدل رگرسیونی با متغیر پاسخ از نوع سیمپلکس است، این فصل به شناخت توزیع سیمپلکس اختصاص دارد. از آن‌جا که توزیع سیمپلکس، تبدیلی از توزیع گوسی معکوس آمیخته است، ابتدا توزیع گوسی معکوس[۱] و توزیع گوسی معکوس آمیخته را مطرح می‌کنیم و سپس توزیع سیمپلکس، خواص آن و تولید اعداد تصادفی سیمپلکس را تشریح می‌کنیم.

۱.۲ توزیع گوسی معکوس

توزیع گوسی معکوس یکی از پرکاربردترین توزیع‌ها در تحلیل قابلیت اعتماد و هم‌چنین به‌عنوان مدلی برای طول عمر و زمان خرابی است. هم‌چنین این توزیع در تحلیل داده‌های مثبت چوله به راست نیز مورد استفاده قرار می‌گیرد. توزیع گوسی معکوس توسط شرودینگر در سال ۱۹۱۵ به‌دلیل خواص مهم فیزیکی آن تحت عنوان توزیع «اولین لحظه گذر در حرکت براونی» معرفی شد. فولکز و چیکارا (۱۹۸۸) به شباهت‌های قابل توجهی بین توزیع گوسی معکوس و خانواده گوسی اشاره کرده‌اند. جرجنسن (۱۹۹۱) تابع توزیع گوسی معکوس تعمیم‌یافته را

[1] Inverse Gaussian Distribution

معرفی و خواص آن را بیان کرد. هم‌چنین مطالعه‌ای گسترده درباره پیدایش و خواص نظری توزیع گوسی معکوس توسط سشادری (۱۹۹۹) صورت گرفت. مادولکار و ناتارجان (۲۰۰۲) خصوصیات مختلف این توزیع از جمله تقارن، کشیدگی و چولگی را بررسی کردند و نشان دادند که توزیع گوسی معکوس متعلق به خانواده توزیع‌های نمایی است و در صورتی‌که داده‌ها نامنفی و دارای چولگی مثبت باشند، می‌تواند به‌عنوان جایگزینی برای توزیع گوسی استفاده شود. مادولکار و وانگ (۲۰۰۷) درباره توزیع وارون متغیر گوسی معکوس مطالعاتی انجام دادند. بالاکیریشنان و همکاران (۲۰۰۹) به معرفی توزیع گوسی معکوس آمیخته و تبدیل‌ها و گشتاورهای آن پرداختند و یک روش جدید برای محاسبه گشتاورهای توزیع معرفی کردند و از گشتاورهای توزیع برای انتخاب مدل مناسب استفاده کردند.
در تحلیل داده‌ها، در بسیاری از موارد داده‌ها متقارن نبوده و دارای چولگی هستند. یکی از کاربردهای مهم این توزیع در مسئله داده‌های چوله است.

تعریف ۱.۱.۲. اگر متغیر تصادفی پیوسته و مثبت X دارای تابع چگالی احتمال زیر باشد

$$f\left(x;\xi,\epsilon^{۲}\right)=\left(۲\pi\epsilon^{۲}x^{۳}\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-\left(x-\xi\right)^{۲}}{۲\xi^{۲}\epsilon^{۲}x}\right\},\quad x>۰,\epsilon^{۲}>۰,\xi>۰ \tag{۱.۲}$$

آن‌گاه X دارای توزیع گوسی معکوس با نماد $X\sim IG(\xi,\epsilon^{۲})$ است که پارامتر ξ میانگین توزیع و $\epsilon^{۲}$ پارامتر مقیاس است.

نمودار تابع چگالی گوسی معکوس در شکل ۱.۲ به ازای مقادیر مختلفی از ξ و $\epsilon^{۲}$ نشان داده شده‌است. امید ریاضی و واریانس توزیع گوسی معکوس عبارتند از

$$\mathrm{E}\left(X\right)=\xi$$

$$\mathrm{V}\left(X\right)=\xi^{۳}\epsilon^{۲}$$

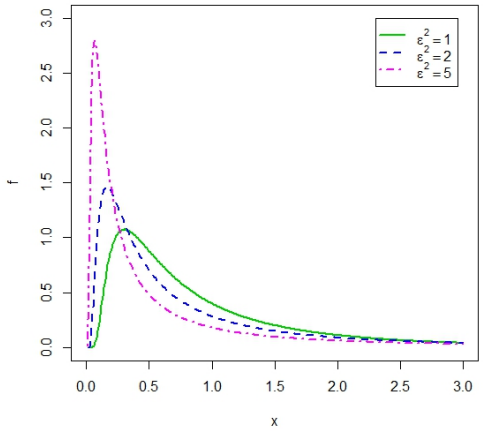
توزیع گوسی معکوس یک خانواده نمایی دو پارامتری با آماره‌های طبیعی $X,\frac{۱}{X}$ و پارامترهای طبیعی $\frac{-۱}{۲\epsilon^{۲}},\frac{-۱}{۲\xi^{۲}\epsilon^{۲}}$ است.
با قرار دادن $\xi=۱$ و $\epsilon=۱$ در چگالی (۱.۲) صورت استاندارد توزیع به‌صورت

$$f\left(x;۱,۱\right)=\left(۲\pi x^{۳}\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-\left(x-۱\right)^{۲}}{۲x}\right\}$$

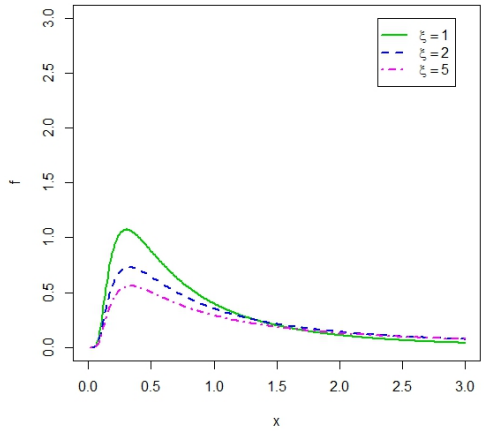
حاصل می‌شود که در واقع شکل حدی توزیع نمونه در یک آزمون نسبت احتمال دنباله‌ای می‌باشد که توسط والد به‌دست آمده است. از این‌رو، گاهی اوقات آن‌را توزیع والد استاندارد گویند.

برای $\epsilon^{۲}=۱$، تابع چگالی و تابع توزیع تجمعی گوسی معکوس به‌ترتیب به‌صورت زیر هستند

$$f_{IG}\left(x;\xi\right)=\left(۲\pi x^{۳}\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-\left(x-\xi\right)^{۲}}{۲\xi^{۲}x}\right\} \tag{۲.۲}$$



(ب) نمودار تابع چگالی گوسی معکوس برای $\xi = ۱$ و مقادیر مختلف $\epsilon^۲$.



(آ) نمودار تابع چگالی گوسی معکوس برای $\epsilon^۲ = ۱$ و مقادیر مختلف ξ.

شکل ۱.۲: نمودار توابع چگالی گوسی معکوس

$$F_{IG}(x;\xi) = \Phi(\frac{۱}{\sqrt{x}}\frac{x-\xi}{\xi}) + \exp(\frac{۲}{\xi})\Phi(-\frac{۱}{\sqrt{x}}\frac{x+\xi}{\xi}) \quad x > ۰, \xi > ۰$$

که در آن $\Phi(.)$ تابع توزیع تجمعی نرمال استاندارد است.

۱.۱.۲ تابع مولد گشتاور توزیع گوسی معکوس

اگر متغیر تصادفی $X \sim IG(\xi, \epsilon^۲)$ باشد، در این‌صورت

$$\mathrm{M}_X(t) = \int_۰^\infty e^{tx} f(x;\xi,\epsilon^۲)\mathrm{d}x = e^{\frac{۱}{\xi\epsilon^۲}} \int_۰^\infty \left\{\frac{۱}{۲\pi\epsilon^۲x^۳}\right\}^{\frac{۱}{۲}} \exp\left\{\frac{-(x^۲+\xi^۲)-۲\xi^۲x^۲t}{۲\xi^۲\epsilon^۲x}\right\}\mathrm{d}x$$

که می‌توان نوشت

$$\begin{aligned}\frac{-(x^۲+\xi^۲)-۲\xi^۲x^۲t}{۲\xi^۲\epsilon^۲x} &= \frac{-\left(\sqrt{۱-۲\xi^۲\epsilon^۲t}x-\xi\right)^۲}{۲\xi^۲\epsilon^۲x} - \frac{۱}{\xi\epsilon^۲}\sqrt{۱-۲\xi^۲\epsilon^۲t} \\ &= \frac{-(x-\xi')^۲}{۲\xi'^۲\epsilon^۲x} - \frac{۱}{\xi\epsilon^۲}\sqrt{۱-۲\xi^۲\epsilon^۲t}\end{aligned}$$

که در آن $\xi' = \dfrac{\xi}{(۱-۲\xi^۲\epsilon^۲t)^{\frac{۱}{۲}}}$. با جایگذاری رابطه بالا در انتگرال نمایش $\mathrm{M}_X(t)$ داریم

$$\mathrm{M}_X(t) = \exp\left\{\frac{۱}{\xi\epsilon^۲}\right\}\exp\left\{\frac{-۱}{\xi\epsilon^۲}\sqrt{۱-۲\xi^۲\epsilon^۲t}\right\}\int_۰^\infty \sqrt{\frac{۱}{۲\pi\epsilon^۲x^۳}}\exp\left\{\frac{-(x-\xi')^۲}{۲\xi'^۲\epsilon^۲x}\right\}\mathrm{d}x$$

تابع زیر انتگرال بالا تابع چگالی احتمال یک توزیع گوسی معکوس با پارامترهای ξ' و $\epsilon^{۲}$ است. بنابراین مقدار انتگرال، برابر با یک است و در نتیجه تابع مولد گشتاور گوسی معکوس به‌صورت زیر است:

$$\mathrm{M}_X(t) = \exp\left\{\frac{۱}{\xi\epsilon^{۲}}\left[۱-\left(۱-۲\xi^{۲}\epsilon^{۲}t\right)^{\frac{۱}{۲}}\right]\right\}, \quad t < \frac{۱}{۲\xi^{۲}\epsilon^{۲}}$$

۲.۱.۲ وجه تسمیه معکوس

اگر $M_X(t)$ نشان دهنده تابع مولد گشتاور متغیر تصادفی X باشد، تابع مولد انباشتگی[۲] متغیر تصادفی X به‌صورت زیر تعریف می‌شود

$$\psi(t) = \log \mathrm{M}_X(t) = \log \mathrm{E}[e^{tX}]$$

حال اگر $\mathrm{L}_X(t)$ نشان دهنده لگاریتم تبدیل لاپلاس متغیر تصادفی نامنفی X باشد، بنابراین داریم

$$\mathrm{L}_X(t) = \log \mathrm{E}[e^{-tX}] = \psi(-t)$$

در سال ۱۹۴۵ تویدی ارتباطی وارون بین تابع مولد انباشتگی جفت‌هایی از متغیر تصادفی دو جمله‌ای و دو جمله‌ای منفی، پواسون و گاما، نرمال و گوسی معکوس ارائه نمود.
در این‌جا ابتدا حالت کلی متغیرهای تصادفی معکوس معرفی می‌شود، سپس این ارتباط وارون در مورد جفت متغیرهای تصادفی که دارای این خاصیت هستند مورد بررسی قرار گیرند.

تعریف ۲.۱.۲. متغیرهای تصادفی X و Y معکوس یکدیگرند اگر و تنها اگر $\mathrm{L}_X(t)$ و $\mathrm{L}_Y(t)$ به ازا هر مقدار t در دامنه مشترک در شرایط زیر صدق کند

$$i)\mathrm{L}_X(t) = \alpha \mathrm{L}(t) \qquad ii)\mathrm{L}_Y(t) = \beta \mathrm{L}^{-۱}(t) \tag{۳.۲}$$

که در آن $\mathrm{L}_X(t)$ نشان دهنده لگاریتم تبدیل لاپلاس متغیر تصادفی X است، α و β مقادیر ثابت و $\mathrm{L}^{-۱}(t)$ وارون تابع $\mathrm{L}(t)$ می‌باشد، یعنی $\mathrm{L}(\mathrm{L}^{-۱}(t)) = t$. توزیع‌های متناظر با متغیرهای تصادفی معکوس را توزیع‌های معکوس می‌نامند.

فرض کنید متغیر تصادفی X دارای توزیع نرمال با میانگین μ و واریانس $\sigma^{۲}$ با تابع چگالی احتمال زیر باشد

$$f(x) = \frac{۱}{\sqrt{۲\pi\sigma^{۲}}}\exp\left\{\frac{-(x-\xi)^{۲}}{۲\sigma^{۲}}\right\}, \quad x \in \mathbb{R}, \xi > ۰, \sigma > ۰$$

آن‌گاه

$$\mathrm{E}[e^{tX}] = \exp\left\{\xi t + \frac{\sigma^{۲}t^{۲}}{۲}\right\} \Longrightarrow \mathrm{L}_X(t) = \xi\left(\frac{\sigma^{۲}t^{۲}}{۲\xi} - t\right), \quad t \in \mathbb{R}$$

[2] Cumulant Generating Function

با فرض $\alpha = \xi$ و $\mathrm{L}(t) = \frac{\sigma^۲ t^۲}{۲\xi} - t$، به‌راحتی می‌توان دید که

$$\mathrm{L}^{-۱}(t) = \frac{\xi}{\sigma^۲}\left[۱ - (۱ + \frac{۲\sigma^۲}{\xi}t)^{\frac{۱}{۲}}\right] \qquad t < \frac{\xi}{۲\sigma^۲}$$

با قرار دادن $\frac{۱}{\epsilon^۲} = \frac{\xi^۳}{\sigma^۲}$ داریم

$$\mathrm{L}^{-۱}(t) = \frac{۱}{\xi^۲\epsilon^۲}\left[۱ - (۱ + ۲\xi^۲\epsilon^۲ t)^{\frac{۱}{۲}}\right] \qquad t < \frac{۱}{۲\xi^۲\epsilon^۲}$$

حال اگر متغیر تصادفی Y دارای تابع چگالی احتمال (۱.۲) باشد، آن‌گاه

$$\mathrm{L}_Y(t) = \frac{۱}{\xi\epsilon^۲}\left[۱ - (۱ + ۲\xi^۲\epsilon^۲ t)^{\frac{۱}{۲}}\right] = \xi\mathrm{L}^{-۱}(t) \qquad t < \frac{۱}{۲\xi^۲\epsilon^۲}$$

با فرض $\beta = \xi$ به ازای تمام مقادیر $t < \frac{۱}{۲\xi^۲\epsilon^۲}$ روابط (۳.۲) در مورد متغیرهای تصادفی X و Y حفظ می‌شود؛ یعنی متغیر تصادفی Y معکوس متغیر تصادفی نرمال است.

۳.۱.۲ برخی خواص توزیع گوسی معکوس

توزیع گوسی معکوس در مقایسه با بسیاری از توزیع‌های کلاسیک از خواصی برخوردار می‌باشد که بعضی از آن‌ها مشابه توزیع نرمال است. در زیر به بررسی برخی از مهم‌ترین ویژگی‌های این توزیع می‌پردازیم.

ویژگی ۱ـ اگر $X \overset{iid}{\sim} IG\left(\xi, \epsilon^۲\right)$ آن‌گاه $kX \sim IG\left(k\xi, k\epsilon^۲\right), k > ۰$

ویژگی ۲ـ اگر $X_i \overset{iid}{\sim} IG\left(\xi, \epsilon^۲\right), i = ۱, \cdots, n$ آن‌گاه $\sum X_i \sim IG\left(n\xi, n^۲\epsilon^۲\right)$

ویژگی ۳ـ اگر $X_i \overset{iid}{\sim} IG\left(\xi, \epsilon^۲\right), i = ۱, \cdots, n$ آن‌گاه $\bar{X} \sim IG\left(\xi, n\epsilon^۲\right)$

ویژگی ۴ـ

قضیه ۱.۱.۲. (شاستر، ۱۹۶۸) اگر متغیر تصادفی X دارای تابع چگالی به‌صورت رابطه (۱.۲) باشد، آن‌گاه

$$\frac{(X-\xi)^۲}{X\xi^۲\epsilon^۲} \sim \chi^۲_{(۱)}$$

ویژگی ۵ـ (وارون توزیع گوسی معکوس[۳]) اگر $X \sim IG(\xi, ۱)$ که تابع چگالی آن در رابطه (۲.۲) داده شد، آن‌گاه تابع چگالی احتمال متغیر تصادفی $X^{-۱}$ (وارون توزیع گوسی معکوس) که آن را با نماد $X \sim RIG(\xi, ۱)$ نشان می‌دهیم، به‌صورت زیر است

$$f_{RIG}(x;\xi) = (۲\pi x)^{\frac{-۱}{۲}} \exp\left\{\frac{-(۱-\xi x)^۲}{۲\xi^۲ x}\right\} \qquad (۴.۲)$$

[۳] Reciprocal Inverse Gaussian Distribution

و تابع توزیع تجمعی آن نیز به‌صورت

$$\begin{aligned} F_{RIG}(x;\xi) &= ۱ - F_{IG}(\frac{۱}{x},\xi) \\ &= ۱ - \Phi(\frac{۱}{\sqrt{x}}\frac{۱-\xi x}{\xi}) - \exp(\frac{۲}{\xi})\Phi(-\frac{۱}{\sqrt{x}}\frac{۱+\xi x}{\xi}) \end{aligned} \tag{۵.۲}$$

است.

برهان. برای اثبات رابطه (۴.۲)، با توجه به رابطه (۲.۲) و بنا به اصول روش تغییر متغیر می‌دانیم

$$f_{RIG}(x;\xi) = \frac{۱}{x^۲} f_{IG}(\frac{۱}{x};\xi) \tag{۶.۲}$$

که $\frac{۱}{x^۲}$ ژاکوبین تبدیل است و لذا

$$\begin{aligned} f_{RIG}(x;\xi) &= \frac{۱}{x^۲}\left[(۲\pi x^{-۳})^{-\frac{۱}{۲}} \exp\left\{\frac{-(x^{-۱}-\xi)^۲}{۲\xi^۲ x^{-۱}}\right\}\right] \\ &= \left[(۲\pi x)^{-\frac{۱}{۲}} \exp\left\{\frac{-(x^{-۱}-\xi)^۲}{۲\xi^۲ x^{-۱}}\frac{x^۲}{x^۲}\right\}\right] \\ &= \left[(۲\pi x)^{-\frac{۱}{۲}} \exp\left\{\frac{-x^۲(x^{-۲}+\xi^۲-۲\xi x^{-۱})}{۲\xi^۲ x}\right\}\right] \\ &= \left[(۲\pi x)^{-\frac{۱}{۲}} \exp\left\{\frac{-(۱+x^۲\xi^۲-۲\xi x)}{۲\xi^۲ x}\right\}\right] \\ &= \left[(۲\pi x)^{-\frac{۱}{۲}} \exp\left\{\frac{-(۱-\xi x)^۲}{۲\xi^۲ x}\right\}\right] \end{aligned}$$

حال برای اثبات رابطه (۵.۲) داریم

$$P(X \leqslant t) = F_X(t) = P(\frac{۱}{X} \geqslant \frac{۱}{t}) = P(Y \geqslant \frac{۱}{t}) = ۱ - F_Y(\frac{۱}{t}).$$

بنابراین

$$F_Y(t) = ۱ - F_X(\frac{۱}{t}).$$

□

لازم به ذکر است که از این ویژگی در توصیف توزیع گوسی معکوس آمیخته در بخش ۲.۲ استفاده خواهیم کرد.

۴.۱.۲ تولید مقادیر تصادفی از توزیع گوسی معکوس

فرض کنید متغیر تصادفی پیوسته X دارای تابع توزیع F باشد، به سادگی می‌توان نشان داد که $Y = F(X)$ دارای توزیع یکنواخت روی بازه $(۰,۱)$ است. حال اگر تابع وارون F به آسانی

قابل محاسبه باشد می‌توان با به‌دست آوردن معکوس F، اعداد تصادفی که از توزیع F پیروی می‌کنند را تولید کرد. کافی است اعدادی تصادفی در بازه $(۰,۱)$ اختیار شوند و آن‌ها را در ضابطه تابع معکوس قرار دهند و مقادیر متغیر متناظر محاسبه شود. در مورد توزیع گوسی معکوس به‌دلیل شکل پیچیده تابع توزیع، محاسبه معکوس F به‌راحتی امکان‌پذیر نیست.
میکائیل و همکاران در سال ۱۹۷۶ روشی برای تولید مشاهدات تصادفی از یک توزیع گوسی معکوس ارائه کردند. فرض کنید $X \sim IG(\xi, \epsilon^۲)$. با توجه به ویژگی ۴ توزیع گوسی معکوس، قضیه ۱.۱.۲، که در آن $Z = \frac{(X-\xi)^۲}{X\xi^۲\epsilon^۲} \sim \chi^۲_{(۱)}$ است، اگر X را بر حسب Z به‌دست آوریم با یک معادله درجه دو بر حسب X مواجه هستیم. فرض کنید $X^{(۱)}$ و $X^{(۲)}$ ریشه‌های این معادله باشند؛ داریم

$$X^{(۱)} = \xi + \frac{\xi^۲\epsilon^۲ Z}{۲} - \frac{\xi\epsilon^۲}{۲}\sqrt{\frac{۴\xi Z}{\epsilon^۲} + \xi^۲ Z^۲}, \quad X^{(۲)} = \frac{\xi^۲}{X^{(۱)}}. \tag{۷.۲}$$

میکائیل و همکارانش احتمال شرطی انتخاب ریشه‌ها با فرض تولید یک متغیر تصادفی Z از توزیع کای‌دو با یک درجه آزادی را محاسبه نمودند و نشان دادند که ریشه کوچک‌تر با احتمال $\frac{\xi}{\xi + X^{(۱)}}$ و ریشه دیگر ، با احتمال $\frac{X^{(۱)}}{\xi + X^{(۱)}}$ انتخاب می‌شود. پس اگر Z را از توزیع کای‌دو با یک درجه آزادی تولید کنیم، آن‌گاه با روابط (۷.۲) می‌توان به مقادیر متغیر گوسی معکوس دست پیدا کرد؛ به این صورت که یک آزمایش برنولی با احتمال موفقیت $\frac{\xi}{\xi + X^{(۱)}}$ انجام می‌دهیم؛ برای اثبات به میکائیل و همکاران، (۱۹۷۶) رجوع کنید.

لازم به ذکر است که برای تولید یک مشاهده از توزیع کای‌دو با یک درجه آزادی، کافی است یک مشاهده از توزیع نرمال استاندارد را مجذور نماییم و برای تولید یک مشاهده نرمال استاندارد از تبدیلات زیر استفاده می‌کنیم

$$Z_۱ = \sqrt{-۲\ln U_۱}\cos ۲\pi U_۲ \quad , \qquad Z_۲ = \sqrt{-۲\ln U_۱}\sin ۲\pi U_۲$$

که در آن $U_۱$ و $U_۲$ دو مشاهده مستقل از توزیع یکنواخت در بازه $(۰,۱)$، $Z_۱$ و $Z_۲$ دو متغیر نرمال استاندارد مستقل هستند.
بنابراین می‌توان الگوریتم تولید اعداد تصادفی گوسی معکوس را به‌صورت زیر ارائه داد.

الگوریتم ۱.۱.۲. الگوریتم تولید اعداد تصادفی از گوسی معکوس

گام ۱ـ اعداد تصادفی Z از توزیع کای‌دو با یک درجه آزادی تولید می‌کنیم.

گام ۲ـ برای هر عدد تصادفی در گام ۱، ریشه کوچک‌تر، $X^{(۱)}$، را براساس رابطه (۷.۲) محاسبه می‌کنیم.

گام ۳ـ یک آزمایش برنولی با احتمال موفقیت $p = \frac{\xi}{\xi + X^{(۱)}}$ اجرا می‌کنیم.

گام ۴ـ اگر نتیجه آزمایش موفقیت باشد ریشه $X^{(۱)}$ را به‌عنوان مشاهده‌ای از توزیع گوسی معکوس انتخاب می‌کنیم و در غیر این‌صورت ریشه بزرگ‌تر $X^{(۲)}$ انتخاب می‌شود.

رابطه بین ریشه‌های معادله $X^{(۱)}X^{(۲)} = \xi^۲$ است.

۲.۲ توزیع گوسی معکوس آمیخته

بالاکیریشنان و همکاران (۲۰۰۹) توزیع گوسی معکوس آمیخته و تبدیل‌ها و گشتاورهای آن را معرفی کرده و یک روش جدید برای محاسبه گشتاورهای این توزیع معرفی کردند. در تحلیل پارامتری داده‌ها، مدل‌های چوله اهمیت بسیاری دارند. توزیع‌های آمیخته نیز بسیار انعطاف‌پذیرند و دارای خصوصیات آماری خوبی نیز می‌باشند. توزیع گوسی معکوس آمیخته یکی از مدل‌های آماری است که کاربرد زیادی دارد. با انجام تبدیلاتی روی توزیع گوسی معکوس آمیخته توزیع‌های پارامتری جدیدی به‌دست می‌آیند که در موقعیت‌های مختلف مفیدند.

تعریف ۱.۲.۲. فرض کنید $f_{X_۱}(x)$ و $f_{X_۲}(x)$ به‌ترتیب توابع چگالی احتمال متغیرهای تصادفی $X_۱$ و $X_۲$ باشند. متغیر تصادفی X را آمیخته گسسته گویند، اگر دارای چگالی زیر باشد:

$$f_X(x) = (۱-p)\, f_{X_۱}(x) + p f_{X_۲}(x), \qquad x > ۰, ۰ \le p \le ۱.$$

تعریف ۲.۲.۲. متغیر تصادفی X دارای توزیع گوسی معکوس آمیخته با پارامترهای مکان ξ، مقیاس $\epsilon^۲$ و پارامتر آمیختگی p است، هرگاه تابع چگالی احتمال آن به‌صورت زیر باشد

$$f_X(x) = \phi\left(\frac{۱}{\sqrt{\xi\epsilon^۲}}\left[\sqrt{\frac{x}{\xi}} - \sqrt{\frac{\xi}{x}}\right]\right)\frac{۱}{\sqrt{\epsilon^۲ x^۳}}\left[۱ - p + \frac{px}{\xi}\right], \quad x > ۰ \qquad (۸.۲)$$

که در آن $\phi(.)$ تابع چگالی احتمال نرمال استاندارد است. این توزیع را با نماد $X \sim MIG(x; \xi, \epsilon^۲, p)$ نشان می‌دهند.

توزیع گوسی معکوس آمیخته، یک خانواده انعطاف‌پذیر از توزیع‌های آمیخته و چوله به راست است.

قضیه ۱.۲.۲. اگر متغیر تصادفی X دارای توزیع گوسی معکوس آمیخته باشد، $X \sim MIG(\xi, \epsilon^۲, p)$، آن‌گاه

الف) $CX \sim MIG(C\xi, C\epsilon^۲, p)$

ب) $R = X^{-۱}$ (وارون توزیع گوسی معکوس آمیخته[۴]) دارای تابع چگالی زیر است، که آن را به اختصار با نماد $\frac{۱}{X} \sim MIG(\frac{۱}{\xi}, \frac{۱}{\xi^۲\epsilon^۲}, ۱-p)$ نشان می‌دهیم

$$f_R(r) = \phi\left(\frac{۱}{\sqrt{\xi\epsilon^۲}}\left[\sqrt{\frac{۱}{r\xi}} - \sqrt{r\xi}\right]\right)\frac{۱}{r^۲}\left[(۱-p)\sqrt{\frac{r^۳}{\epsilon^۲}} + p\sqrt{\frac{r}{\xi^۲\epsilon^۲}}\right], \quad r > ۰.$$

[4] Reciprocal Mixture Inverse Gaussian Distribution

برهان. **الف:** با تغییر متغیر نتیجه مستقیما حاصل می‌شود.
ب: با تغییر متغیر $x = \frac{۱}{r}$ و ژاکوبین تبدیل $|-\frac{۱}{r^۲}|$ در رابطه (۸.۲) داریم

$$f_R(r) = \phi\left(\frac{۱}{\sqrt{\xi\epsilon^۲}}\left[\sqrt{\frac{۱}{\xi r}} - \sqrt{\frac{\xi}{\frac{۱}{r}}}\right]\right)\frac{۱}{\sqrt{\epsilon^۲(\frac{۱}{r})^۳}}\left[۱ - p + \frac{p\frac{۱}{r}}{\xi}\right] \times \frac{۱}{r^۲}$$
$$= \phi\left(\frac{۱}{\sqrt{\xi\epsilon^۲}}\left[\sqrt{\frac{۱}{r\xi}} - \sqrt{r\xi}\right]\right)\frac{۱}{r^۲}\left[(۱-p)\sqrt{\frac{r^۳}{\epsilon^۲}} + p\sqrt{\frac{r}{\xi^۲\epsilon^۲}}\right]$$

و اثبات کامل می‌شود. □

گزاره ۱.۲.۲. اگر $X_۱ \sim IG(\xi, \epsilon^۲)$ و $X_۲^{-۱} \sim IG(\xi^{-۱}, \xi^۲\epsilon^۲)$، آن‌گاه متغیر تصادفی X دارای توزیع گوسی معکوس آمیخته با پارامترهای ξ، $\epsilon^۲$ و p که $X_۱$ با احتمال $p - ۱$ و $X_۲$ با احتمال p است و تابع چگالی احتمال آن به‌صورت

$$f\left(x; \xi, \epsilon^۲, p\right) = \left(۲\pi\epsilon^۲x^۳\right)^{-\frac{۱}{۲}}\left(۱ - p + \frac{px}{\xi}\right)\exp\left\{-\frac{۱}{۲\epsilon^۲}\frac{(x-\xi)^۲}{\xi^۲x}\right\}, \quad x > ۰ \qquad (۹.۲)$$

است.

برهان. بنا به رابطه (۱.۲) می‌دانیم

$$f_{X_۱}(x) = \left(۲\pi\epsilon^۲x^۳\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-(x-\xi)^۲}{۲\xi^۲\epsilon^۲x}\right\}, \quad x > ۰, \epsilon^۲ > ۰, \xi > ۰$$

قرار دهید $X_۲^{-۱} = U$ و $D(x,\xi) = \frac{(x-\xi)^۲}{\xi^۲x}$ و $\frac{D(x,\xi)}{\epsilon^۲} = \frac{D(x^{-۱},\xi^{-۱})}{\xi^۲\epsilon^۲}$. حال چگالی مربوط به آن را با توجه به این‌که $X_۲^{-۱} \sim IG(\xi^{-۱}, \xi^۲\epsilon^۲)$ و بنا به رابطه بالا به‌صورت زیر به‌دست می‌آوریم

$$f_U(u) = \left(۲\pi\xi^۲\epsilon^۲u^۳\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-D\left(u,\xi^{-۱}\right)}{۲\xi^۲\epsilon^۲}\right\}$$

حال نیازمند به‌دست آوردن توزیع خود $X_۲$ هستیم که از رابطه (۶.۲) می‌توان آن را به‌صورت زیر به‌دست آورد

$$f_{X_۲}(u) = \frac{۱}{u^۲}f_U(\frac{۱}{u})$$
$$= \frac{۱}{u^۲}\left(۲\pi\xi^۲\epsilon^۲u^{-۳}\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-D\left(u^{-۱},\xi^{-۱}\right)}{۲\xi^۲\epsilon^۲}\right\}$$
$$\Rightarrow f_{X_۲}(x) = \left(۲\pi\xi^۲\epsilon^۲x\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-D\left(x,\xi\right)}{۲\epsilon^۲}\right\}$$

حال با توجه به تعریف (۱.۲.۲) داریم:

$$f_X(x) = (۱-p)\left(۲\pi\epsilon^۲x^۳\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-(x-\xi)^۲}{۲\xi^۲\epsilon^۲x}\right\} + p\left(۲\pi\xi^۲\epsilon^۲x\right)^{\frac{-۱}{۲}}\exp\left\{\frac{-(x-\xi)^۲}{۲\xi^۲\epsilon^۲x}\right\}$$

بنابراین با فاکتور گرفتن از جمله مشترک و ساده‌سازی، رابطه (۹.۲) به‌دست می‌آید. □

۱.۲.۲ تولید مقادیر تصادفی از توزیع گوسی معکوس آمیخته

متغیرهای تصادفی مستقل از هم $X_۱$ و $X_۲$ را در نظر بگیرید که توزیع آن‌ها $X_۱ \sim IG(\xi, \epsilon^۲)$ و $X_۲^{-۱} \sim IG(\xi^{-۱}, \xi^۲\epsilon^۲)$ است. اکنون متغیر تصادفی X را به‌صورت زیر تعریف می‌کنیم

$$X = \begin{cases} X_۱ & \text{با احتمال } ۱-p \\ X_۲ & \text{با احتمال } p \end{cases}$$

که آمیخته‌ای از $X_۱$ و $X_۲$ است. جانسون و کاتز (۱۹۷۰) نشان دادند که $X_۲$ را می‌توان به‌صورت ترکیب خطی $X_۱ + X_۳$ نوشت (ضمیمه آ.۱ را ببینید)، که $X_۳ \sim \xi^۲\epsilon^۲\chi^۲_{(۱)}$.
پس می‌توان نوشت $X = X_۱ + V$، که در آن

$$V = \begin{cases} ۰ & \text{با احتمال } ۱-p \\ X_۳ & \text{با احتمال } p \end{cases}$$

و هم‌چنین $X_۱$ و V مستقلند.
طبق مطالب فوق یک روش شبیه‌سازی ساده برای تولید مقادیر تصادفی X از توزیع گوسی معکوس آمیخته به شرح زیر قابل بیان است.

الگوریتم ۱.۲.۲. الگوریتم تولید اعداد تصادفی از گوسی معکوس آمیخته

گام ۱ـ تولید مقادیر متغیر تصادفی از توزیع $X_۱ \sim IG(\xi, \epsilon^۲)$ (زیر بخش ۴.۱.۲)

گام ۲ـ تولید مقادیر متغیر تصادفی از تابع توزیع $W \sim \chi^۲_{(۱)}$ و قرار می‌دهیم $X_۳ = \xi^۲\epsilon^۲W$

گام ۳ـ تولید اعداد تصادفی از توزیع یکنواخت، $U \sim U(۰, ۱)$

گام ۴ـ اگر $U < p$، $X = X_۱ + X_۳$، در غیر این‌صورت $X = X_۱$.

در این صورت X متغیر تصادفی تولیدشده از توزیع گوسی معکوس آمیخته با پارامترهای مذکور می‌باشد.

۳.۲ توزیع سیمپلکس

توزیع سیمپلکس برگرفته از توزیع گوسی معکوس تعمیم‌یافته است (بارندروف، نیلسن و جرجنسن، ۱۹۹۱) که تابع چگالی احتمال آن به‌صورت زیر تعریف می‌شود

(۱۰.۲)

$$f\left(y; \alpha_۱, \alpha_۲, \mu, \sigma^۲\right) = c\left(\alpha_۱, \alpha_۲, \mu, \sigma^۲\right) y^{\alpha_۱ - ۱} (۱-y)^{\alpha_۲ - ۱} \exp\left\{-\frac{۱}{۲\sigma^۲} d_{\alpha_۱, \alpha_۲}(y; \mu)\right\}, y \in (۰, ۱)$$

که در آن $\alpha_۱$، $\alpha_۲$، μ و $\sigma^۲$ پارامترهای توزیع و $c\left(\alpha_۱, \alpha_۲, \mu, \sigma^۲\right)$ ثابت نرمال‌ساز و $d_{\alpha_۱, \alpha_۲}(y; \mu)$، انحراف است که به‌صورت

$$d_{\alpha_۱, \alpha_۲}(y; \mu) = \mu^{۲\alpha_۱ - ۱} (۱-\mu)^{۲\alpha_۲ - ۱} \frac{(y-\mu)^۲}{y(۱-y)}$$

تعریف می‌شود و تابع واریانس آن نیز به‌صورت زیر است:

$$V_{\alpha_1,\alpha_2}(\mu) = \mu^{2(1-\alpha_1)}(1-\mu)^{2(1-\alpha_2)}.$$

دامنه پارامترهای توزیع عبارتند از $\mu \in (0,1)$، $(\alpha_1,\alpha_2)^T \in \mathbb{R}^2$ و $\sigma^2 > 0$. چگالی (۱۰.۲) را هم‌چنین توزیع سیمپلکس کلی[5] می‌نامند و آن را با $Y \sim S\left(\alpha_1,\alpha_2,\mu,\sigma^2\right)$ نمایش می‌دهند. اگر $\alpha_1,\alpha_2 > 0$ و $\sigma^2 \longrightarrow \infty$ آن‌گاه Y دارای توزیع بتا با پارامترهای α_1 و α_2، $Y \sim Beta\,(\alpha_1,\alpha_2)$ است.

حالت خاصی از توزیع (۱۰.۲) زمانی‌است که $(\alpha_1,\alpha_2) = (-\frac{1}{2},-\frac{1}{2})$ که آن را سیمپلکس استاندارد گویند و از این پس با توزیع سیمپلکس استاندارد کار خود را پیش برده و برای سهولت آن را «توریع سیمپلکس» می‌نامیم.

تعریف ۱.۳.۲. (متغیر تصادفی سیمپلکس) متغیر تصادفی Y را دارای توزیع سیمپلکس با پارامتر مکان $\mu \in (0,1)$ و پارامتر پراکندگی $\sigma^2 > 0$ گویند و با نماد $Y \sim S^-(\mu,\sigma^2)$ نشان می‌دهند اگر تابع چگالی آن به‌صورت زیر باشد

$$f(y;\mu,\sigma^2) = \left[2\pi\sigma^2\left\{y\,(1-y)\right\}^3\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}d\,(y;\mu)\right\}, \qquad y \in (0,1) \qquad (11.2)$$

که در آن

$$d(y;\mu) = \frac{(y-\mu)^2}{y\,(1-y)\,\mu^2\,(1-\mu)^2}.$$

جرجنسن (۱۹۹۷)، واریانس Y را به‌صورت زیر

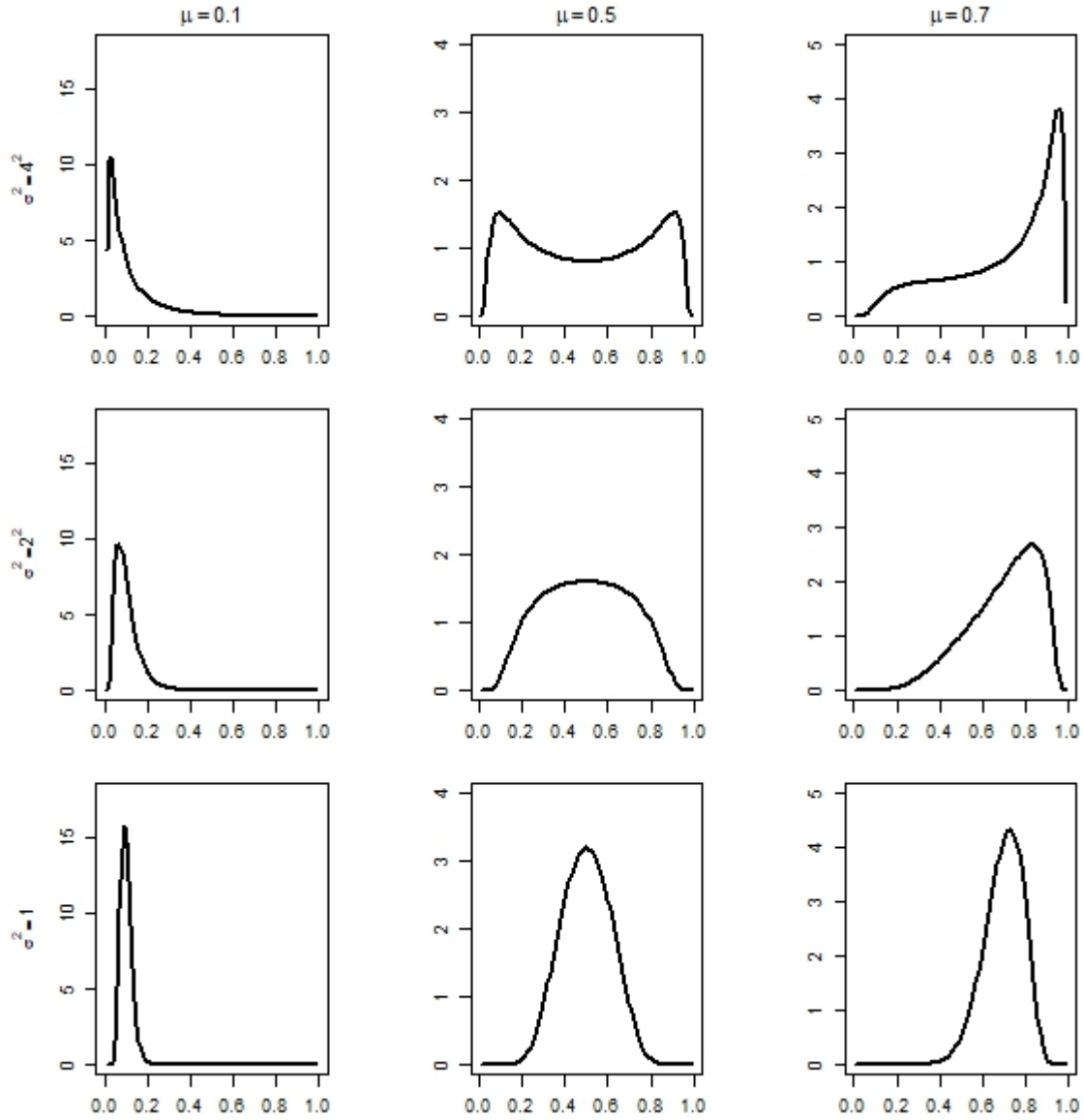
$$\tau^2 = \mu\,(1-\mu) - \frac{1}{\sqrt{2\sigma^2}} \exp\left\{\frac{1}{2\sigma^2\mu^2\,(1-\mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\sigma^2\mu^2\,(1-\mu)^2}\right\}$$

محاسبه کرد که در آن $\Gamma\,(a,b)$، تابع گامای ناکامل، به‌صورت $\Gamma\,(a,b) = \int_b^\infty t^{a-1}e^{-t}\mathrm{d}t$ است.

شکل ۲.۲ نمودار تابع چگالی سیمپلکس را با پارامترهای متفاوت مکان و پراکندگی نشان می‌دهد. همان‌گونه که ملاحظه می‌شود، توزیع سیمپلکس شامل یک کلاس بزرگ از توزیع‌های در محدوده $(0,1)$ است به‌طوری‌که اگر $\sigma \leq \frac{2}{\sqrt{3}}$، توزیع تک مُدی و در غیر این‌صورت چند مُدی خواهد بود. در این شکل، ردیف اول و دوم و سوم به‌ترتیب متناظر با σ^2 برابر با $1,2^2,4^2$ است، هم‌چنین ستون اول و دوم و سوم از راست به‌ترتیب متناظر با μ برابر با $0/1, 0/5, 0/7$ است.

در مقایسه رابطه (۱۱.۲) و (۱.۱) به آسانی می‌توان دریافت که توزیع سیمپلکس متعلق به خانواده مدل‌های پراکندگی است.

[5] General Simplex Distribution



شکل ۲.۲: نمودار تابع چگالی سیمپلکس

با توجه به خاصیت ۴ مدل‌های پراکندگی، رابطه (۶.۱)، و رابطه (۷.۱) (بخش ۲.۲.۱ را ببینید) می‌توان گفت که اگر $\sigma^2 \longrightarrow 0$، آن‌گاه

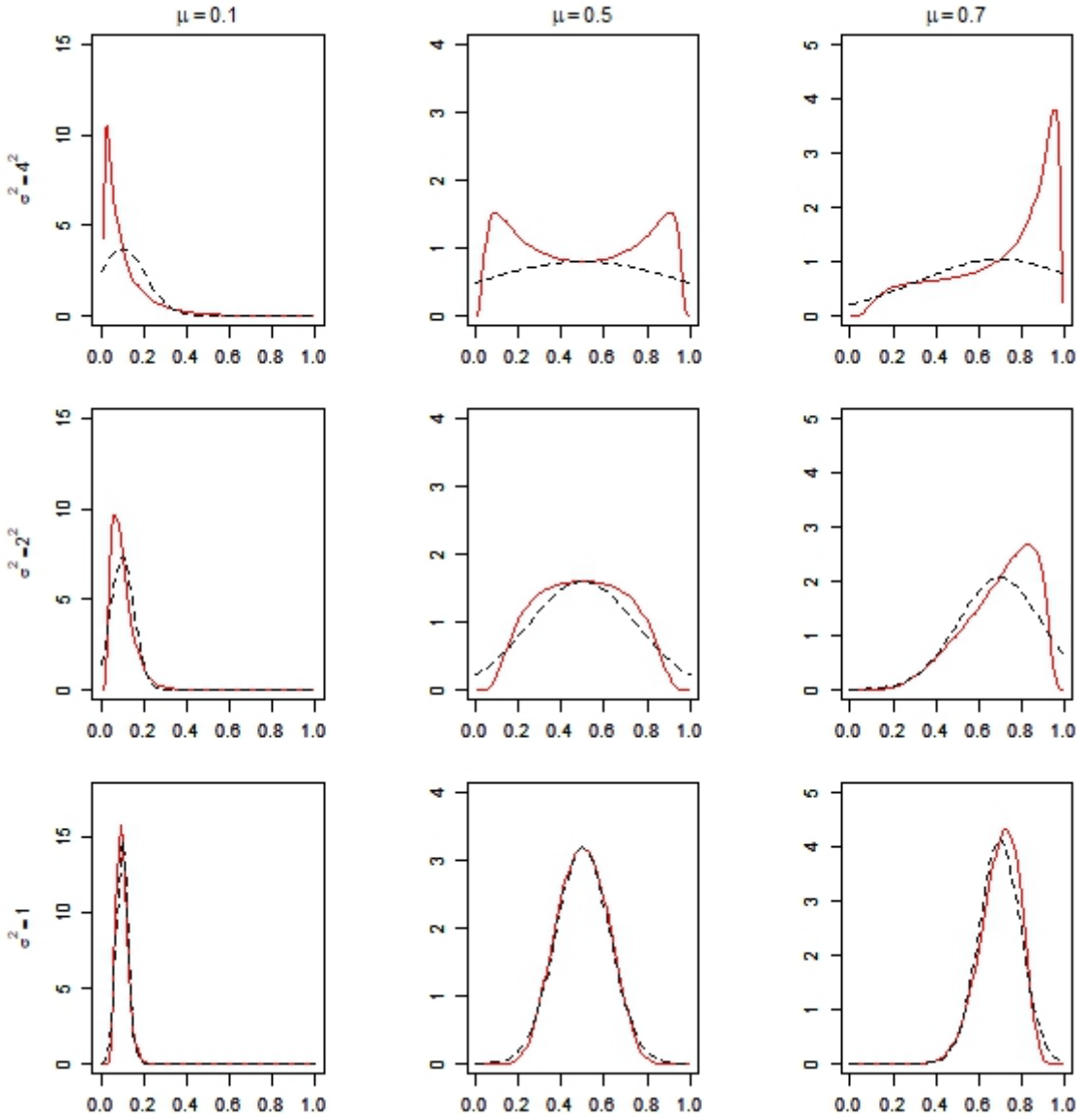
$$\frac{Y-\mu}{\sigma\sqrt{\mathrm{V}(\mu)}} \xrightarrow{d} N(0,1) \qquad (۱۲.۲)$$

که در آن $\mathrm{V}(\mu) = \mu^3(1-\mu)^3$ تابع واریانس است.

برای بررسی این موضوع، نمودار چگالی احتمال سیمپلکس و نرمال $\mathrm{N}(\mu, \sigma^2\mathrm{V}(\mu))$ را در یک قاب و به‌ازای μ و σ^2 های مختلف رسم کرده‌ایم (شکل ۳.۲).

در این شکل نیز همانند قبل، ردیف اول و دوم و سوم به‌ترتیب متناظر با σ^2 برابر است با $1, 2^2, 4^2$. همچنین ستون اول و دوم و سوم به‌ترتیب متناظر با μ برابر با $0.1, 0.5, 0.7$ است. همان‌گونه که مشخص است، به‌ازای σ^2 کوچک‌تر، توزیع نرمال تطابق بیشتری با

توزیع سیمپلکس دارد.



شکل ۳.۲: نمودار توابع چگالی سیمپلکس (منحنی توپر) و نرمال (منحنی مقطع)

۱.۳.۲ رابطه بین توزیع گوسی معکوس آمیخته و توزیع سیمپلکس

اگر $X \sim MIG(\xi, \epsilon^2, p)$ و قرار دهیم

$$\mu = \frac{\xi}{1+\xi}, \quad \sigma^2 = \frac{\epsilon^2}{(1-p)^2} = \epsilon^2 (1-\xi)^2$$

آن‌گاه

$$Y = \frac{X}{1+X} \sim S^-(\mu, \sigma^2).$$

برای اثبات به جرجنسن (۱۹۹۷) و ژو (۲۰۰۱) رجوع کنید. یعنی توزیع سیمپلکس می‌تواند با اعمال تغییر متغیری بر توزیع گوسی معکوس آمیخته، به‌دست آید. این موضوع، کلیدی‌ترین نکته در تولید اعداد تصادفی سیمپلکس است که در بخش بعد آن را دنبال می‌کنیم.

۲.۳.۲ تولید مقادیر تصادفی از توزیع سیمپلکس

با توجه به این‌که اگر $X \sim MIG(\xi, \epsilon^2)$ آن‌گاه $Y = \frac{X}{1+X} \sim S^-(\mu, \sigma^2)$، لذا الگوریتم تولید اعداد تصادفی سیمپلکس دقیقا همان الگوریتم تولید اعداد تصادفی از توزیع MIG است که در انتها، مقادیر Y از رابطه بالا به‌دست می‌آید.
این الگوریتم را می‌توان برای تولید اعداد تصادفی از توزیع سیمپلکس، و به استناد زیر بخش ۱.۲.۲ به‌صورت زیر خلاصه کرد:

الگوریتم ۱.۳.۲. الگوریتم تولید اعداد تصادفی از توزیع سیمپلکس

گام ۱ـ قرار دهید $\epsilon^2 = \sigma^2 (1-\mu)^2$ $\xi = \frac{\mu}{1-\mu}$، $p = \mu$.

گام ۲ـ عدد تصادفی X را از توزیع گوسی معکوس آمیخته، $X \sim MIG(\xi, \epsilon^2, p)$ (زیر بخش ۱.۲.۲)، تولید کنید.

گام ۳ـ برای تولید عدد تصادفی سیمپلکس قرار دهید $Y = \frac{X}{1+X}$.

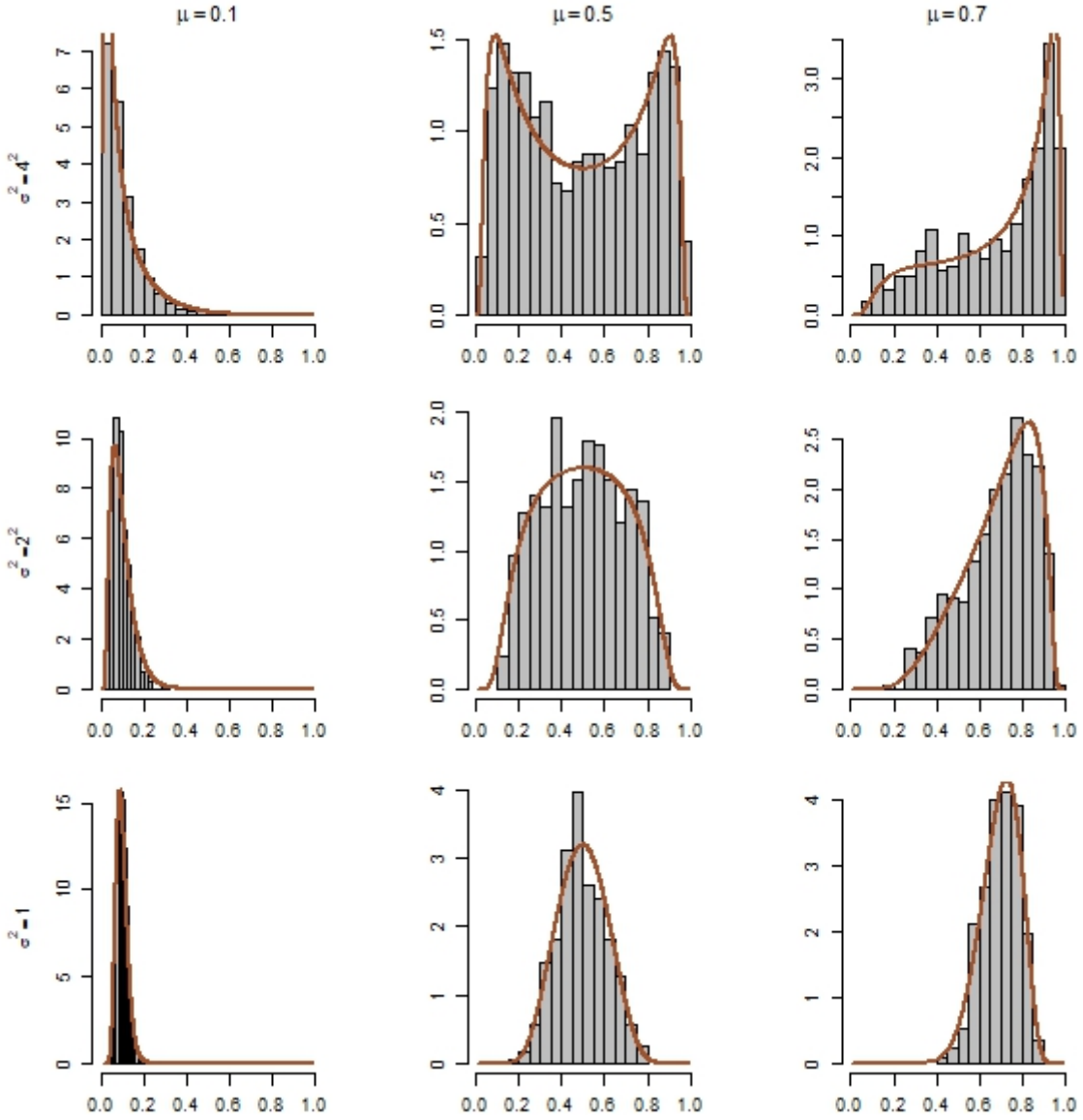
بررسی تطابق داده‌های تولیدشده از توزیع سیمپلکس با چگالی توزیع سیمپلکس

در شکل ۴.۲، هر نمودار بیانگر بافت‌نگار ۵۰۰ عدد تصادفی تولیدشده از توزیع سیمپلکس و هم‌چنین تابع چگالی سیمپلکس $S^-(\mu, \sigma^2)$ است. همان‌طور که ملاحظه می‌شود، تابع چگالی سیمپلکس پوشش مناسبی برای بافت‌نگار ایجاد کرده است. هم‌چنین در شکل ۵.۲ هر نمودار بیانگر هسته چگالی با ۱۵۰۰ عدد تصادفی تولیدشده از توزیع سیمپلکس و تابع چگالی سیمپلکس $S^-(\mu, \sigma^2)$ است.

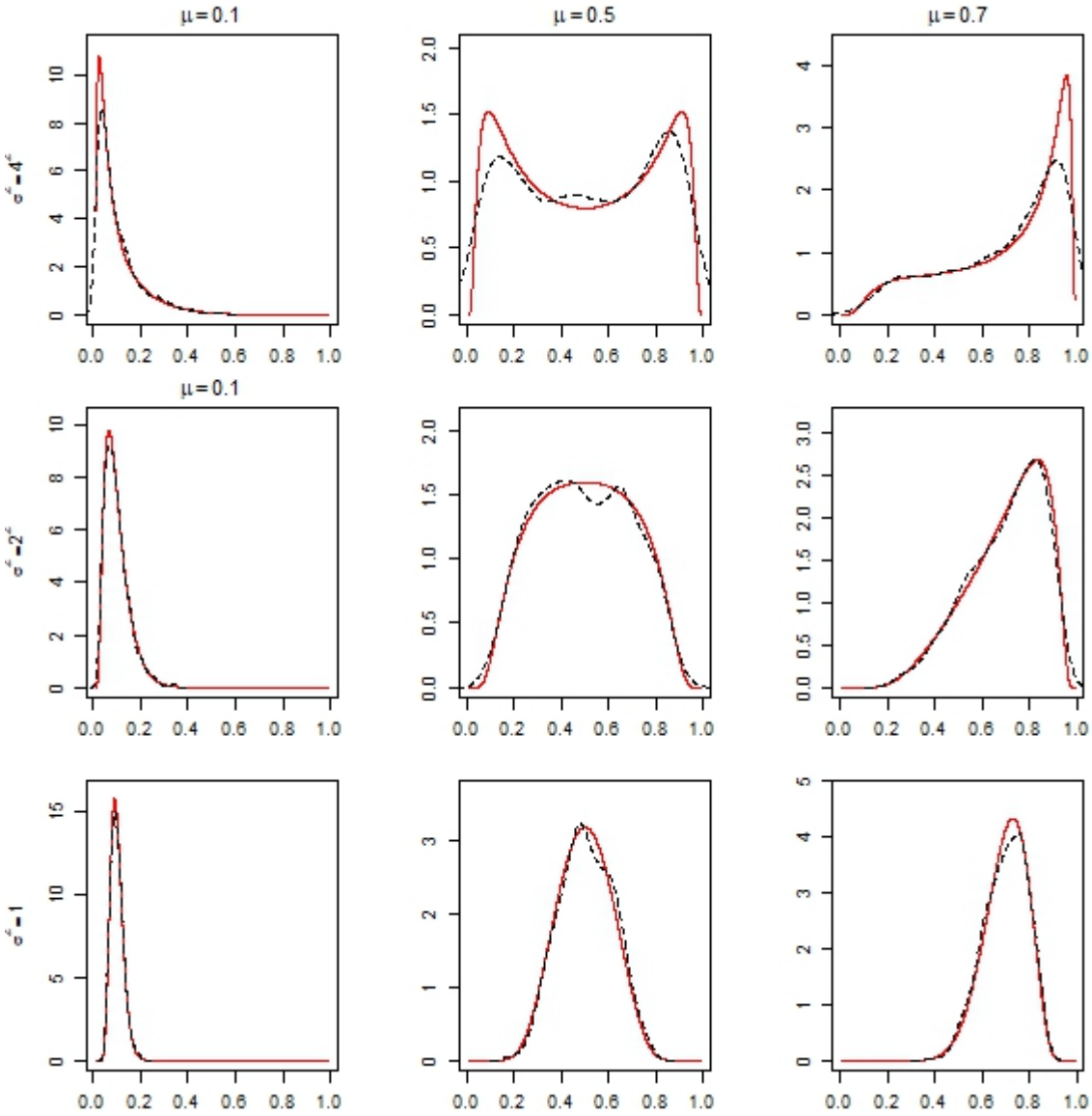
دستورات مربوط به توزیع احتمالی سیمپلکس در بسته‌افزار simplexreg در نرم‌افزار R به شرح زیر هستند:

```
dsimplex(x, mu, sig)
psimplex(q, mu, sig)
qsimplex(p, mu, sig)
rsimplex(n, mu, sig)
```

که به‌ترتیب تابع چگالی احتمال، تابع توزیع تجمعی، تابع چندک و تولید اعداد تصادفی هستند.



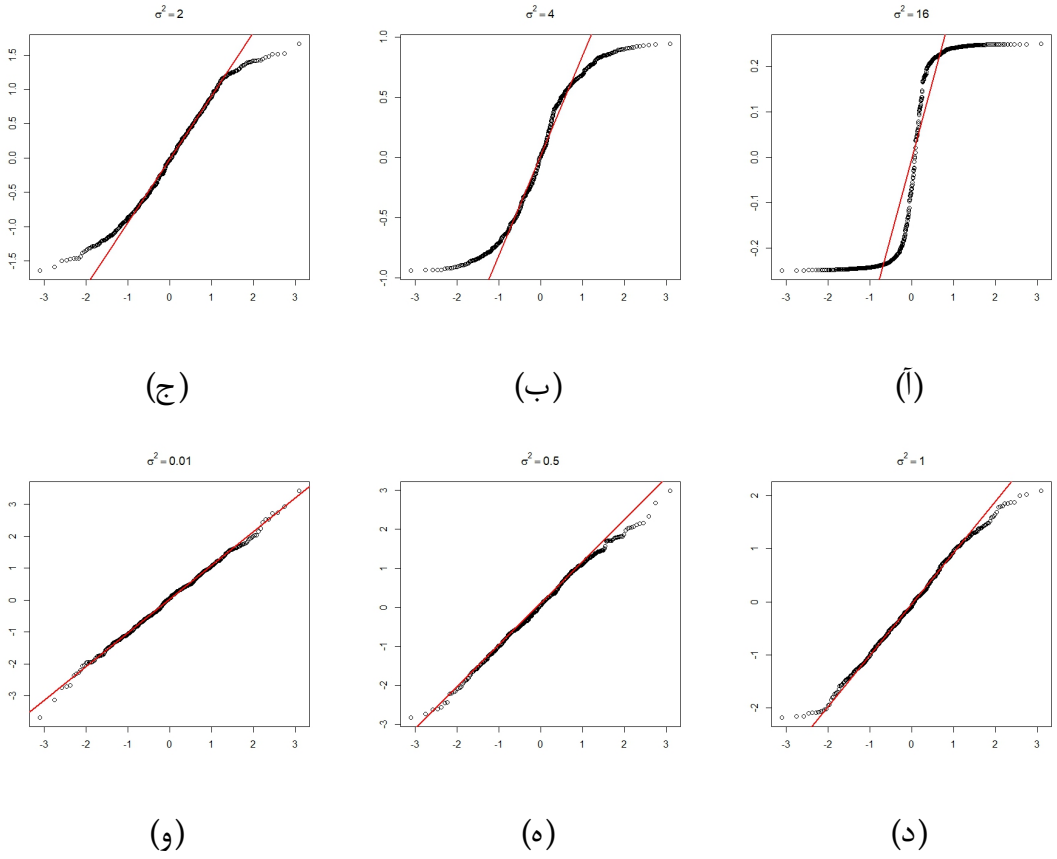
شکل ۴.۲: نمودار چگالی (منحنی توپر) به همراه بافت‌نگار اعداد تصادفی از توزیع سیمپلکس



شکل ۵.۲: نمودار چگالی (منحنی توپر) به همراه هسته چگالی (منحنی مقطع) اعداد تصادفی از توزیع سیمپلکس

بررسی نظریه نرمال مجانبی توسط نمودار Q-Q

برای بررسی تقریب نرمال مندرج در رابطه (۱۲.۲) مبنی بر این‌که اگر $\sigma^2 \longrightarrow 0$ آن‌گاه $\frac{Y-\mu}{\sigma\sqrt{V(\mu)}} \xrightarrow{d} N(0,1)$، تعداد ۵۰۰ عدد تصادفی سیمپلکس را با $\mu = 0.5$ و σ^2 های مختلف $(16, 4, 2, 1, 0.5, 0.01)$ تولید کرده و برای آزمون نرمال بودن، از ابزار بصری نمودار چندک چندک استفاده کرده‌ایم. همان‌طور که در شکل ۶.۲ دیده می‌شود، می‌توان نرمال مجانبی بودن را نتیجه گرفت.



شکل ۶.۲: نمودار چندک_چندک

۳.۳.۲ برخی خواص توزیع سیمپلکس

اگر $Y \sim S^{-}(\mu, \sigma^{۲}, p)$، با $\mathrm{E}[Y] = \mu$ آن‌گاه

$$(۱)\mathrm{E}\left\{d(Y;\mu)\right\} = \sigma^{۲}$$

$$(۲)\mathrm{E}\left\{(Y-\mu)d'(Y;\mu)\right\} = -۲\sigma^{۲}$$

$$(۳)\mathrm{E}\left\{(Y-\mu)d(Y;\mu)\right\} = ۰$$

$$(۴)\mathrm{E}\left\{d'(Y;\mu)\right\} = ۰$$

$$(۵)\frac{۱}{۲}\mathrm{E}\left\{d''(Y;\mu)\right\} = \frac{۳\sigma^{۲}}{\mu(۱-\mu)} + \frac{۱}{\mu^{۳}(۱-\mu)^{۳}}$$

$$(۶)\mathrm{V}\left\{d(Y;\mu)\right\} = ۲(\sigma^{۲})^{۲}$$

$$(۷)\mathrm{V}\left\{u(Y;\mu)\right\} = \frac{۳\sigma^{۴}}{\mu(۱-\mu)} + \frac{\sigma^{۲}}{\mu^{۳}(۱-\mu)^{۳}}.$$

برهان. با اثبات قسمت (۲) شروع می‌کنیم؛ می‌دانیم که

$$۰ = \mathrm{E}[(Y-\mu)] = \int_{۰}^{۱} (y-\mu) f(y;\mu,\sigma^{۲}) \mathrm{d}y,$$

حال از طرفین نسبت به μ مشتق گرفته

$$\begin{aligned}
۰ &= \frac{\mathrm{d}}{\mathrm{d}\mu} \int_{۰}^{۱} (y-\mu) a(y;\sigma^{۲}) \exp\left\{\frac{-۱}{۲\sigma^{۲}} d(y;\mu)\right\} \mathrm{d}y \\
&= \int_{۰}^{۱} a(y;\sigma^{۲}) \left[-۱ \exp\left\{\frac{-۱}{۲\sigma^{۲}} d(y;\mu)\right\}\right] + \int_{۰}^{۱} a(y;\sigma^{۲}) \left[(y-\mu)\frac{d'(y;\mu)}{۲\sigma^{۲}} \exp\left\{\frac{-۱}{۲\sigma^{۲}} d(y;\mu)\right\}\right] \\
&= -۱ - \frac{۱}{۲\sigma^{۲}} \mathrm{E}[(Y-\mu) d'(Y;\mu)]
\end{aligned}$$

در نتیجه اثبات کامل می‌شود؛ یعنی $\mathrm{E}[(Y-\mu)d'(Y;\mu)] = -۲\sigma^{۲}$.

برای اثبات قسمت (۴) از قسمت (۳) و رابطه (۵.۳) استفاده کرده و سپس امید ریاضی می‌گیریم.

برای اثبات قسمت (۵) از قسمت‌های (۱)، (۲) و (۳) استفاده می‌کنیم و می‌توان نوشت

$$\begin{aligned}
\frac{۱}{۲} d''(y;\mu) = &\frac{۱}{\mu(۱-\mu)} d(y;\mu) + \frac{۱-۲\mu}{\mu^{۲}(۱-\mu)^{۲}} (y-\mu) d(y;\mu) \\
&+ \frac{۱}{\mu^{۳}(۱-\mu)^{۳}} + \frac{۱-۲\mu}{\mu^{۴}(۱-\mu)^{۴}} (y-\mu) \\
&- \frac{۱}{\mu(۱-\mu)} (y-\mu) d'(y;\mu) - \frac{۲(۲\mu-۱)}{\mu^{۴}(۱-\mu)^{۴}} (y-\mu).
\end{aligned}$$

حال از رابطه فوق امید ریاضی می‌گیریم و نتیجه می‌شود

$$\begin{aligned}
\frac{۱}{۲} \mathrm{E}\{d''(Y;\mu)\} = &\frac{۱}{\mu(۱-\mu)} \Big[\mathrm{E}\{d(Y;\mu)\} - \mathrm{E}\{(Y-\mu) d'(Y;\mu)\}\Big] \\
&+ \frac{۱-۲\mu}{\mu^{۲}(۱-\mu)^{۲}} \mathrm{E}\{(Y-\mu) d(Y;\mu)\} + \frac{۱}{\mu^{۳}(۱-\mu)^{۳}} \\
= &\frac{۳\sigma^{۲}}{\mu(۱-\mu)} + \frac{۱}{\mu^{۳}(۱-\mu)^{۳}}.
\end{aligned}$$

برای اثبات سایر موارد به جرجنسن (۱۹۹۷)، سانگ و تن (۲۰۰۰)، ژو (۲۰۰۱)، سانگ و همکاران (۲۰۰۴)، ژنگ و ژو (۲۰۱۴) رجوع کنید. □

فصل ۳

مدل رگرسیونی سیمپلکس

مقدمه

در این فصل، یک مدل رگرسیونی مناسب را برای مدل‌بندی متغیر پاسخی که به‌طور پیوسته در فاصله واحد، یعنی $y \in (۰,۱)$، اندازه‌گیری می‌شود، به‌نام رگرسیون سیمپلکس[۱] معرفی می‌کنیم. این مدل رگرسیونی بر اساس پذیره پیروی متغیر پاسخ از توزیع سیمپلکس که در فصل قبل به‌طور مفصل توضیح داده شد، بنا نهاده شده‌است.
از آن‌جایی که توزیع سیمپلکس عضوی از مدل‌های پراکندگی است، روند استنباط برای مدل رگرسیونی سیمپلکس، مشابه مدل‌های پراکندگی و مدل‌های خطی تعمیم‌یافته است. بنابراین برازش مدل را با روش درستنمایی ماکسیمم انجام می‌دهیم. در پایان، با مطالعه شبیه‌سازی، مدل را مورد ارزیابی قرار می‌دهیم.

۱.۳ مدل خطی تعمیم‌یافته سیمپلکس

ما مدل خطی تعمیم‌یافته سیمپلکس[۲] (SGLM) را برای مدل‌بندی متغیرهای وابسته‌ای که به‌طور پیوسته در فاصله واحد مقدار می‌گیرند، معرفی می‌کنیم که مدل بر اساس پذیره پیروی

[1] Simplex Regression
[2] Simplex Generalized Linear Model

متغیر پاسخ از توزیع سیمپلکس است. فرض بر این است که $Y_i \mid x_i \sim S^-(\mu_i, \sigma^۲)$، که تابع چگالی احتمال آن در رابطه (۱۱.۲) با میانگین μ_i و پارامتر پراکندگی $\sigma^۲$ نشان داده شده‌است. با در نظر گرفتن μ_i به‌عنوان پارامتر میانگین، پیشگوی خطی مدل به‌صورت زیر می‌باشد

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

این مدل را مدل سیمپلکس همگن[۳] گویند که در آن $g : (۰, ۱) \longrightarrow (-\infty, \infty)$ تابع پیوند بوده و $\boldsymbol{x}_i$ یک بردار p عنصری شامل مشاهده iام برای p متغیر تبیینی است که $i = ۱, ۲, \ldots, n$ و $\boldsymbol{\beta}$ بردار پارامترهای مجهول رگرسیون است. هدف ما مدل‌سازی متغیر پاسخ Y بر حسب تابعی از $\boldsymbol{x}$ است. این یک نمونه از مدل‌های خطی تعمیم‌یافته است که توسط جرجنسن (۱۹۹۷) معرفی شد؛ در این‌جا از تابع پیوند لجیت استفاده می‌کنیم، گرچه سایر تابع‌های پیوند مانند پروبیت نیز می‌توانند استفاده شوند.

۱.۱.۳ برازش مدل

برای برآورد پارامترهای $\boldsymbol{\beta}$ در SGLM از روش کمترین توان‌های دوم بازموزون تکراری[۴] (IRWLS) استفاده می‌کنیم. رابطه (۱۱.۲) را در نظر بگیرید؛ لگاریتم تابع درستنمایی برای یک نمونه به حجم n به‌صورت زیر است

$$\begin{aligned}
\ell(\beta, \sigma^۲) &= \sum_{i=۱}^{n} \ell_i(\beta_i, \sigma^۲) \\
&= -\frac{۱}{۲} \sum_{i=۱}^{n} \left\{ \frac{d(y_i; \mu_i)}{\sigma^۲} - \log \sigma^۲ + h(y_i) \right\} \\
&= -\frac{۱}{۲\sigma^۲} \sum_{i=۱}^{n} d(y_i; \mu_i) - \frac{n}{۲} \log \sigma^۲ - \frac{۱}{۲} \sum_{i=۱}^{n} h(y_i)
\end{aligned}$$

که در آن $h(y_i) = \log \left\{ ۲\pi [y_i(۱ - y_i)]^۳ \right\}$. تابع ℓ_i از طریق $\mu_i \in (۰, ۱)$ توسط تابع لجیت زیر، وابسته به $\boldsymbol{\beta}$ می‌شود

$$\eta_i = g(\mu_i) = \log \frac{\mu_i}{۱ - \mu_i} = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

برای یافتن برآورد پارامترهای مجهول، مشتق‌های جزئی تابع درستنمایی را در نظر بگیرید

$$\frac{\ell_i(\boldsymbol{\beta})}{\partial \beta_j} = -\frac{۱}{۲\sigma^۲} \sum_{i=۱}^{n} \frac{\partial d(y_i; \mu_i)}{\partial \beta_j}$$

که بنا به قاعده زنجیره‌ای داریم

$$\frac{\ell(\boldsymbol{\beta})}{\partial \beta_j} = -\frac{۱}{۲\sigma^۲} \sum_{i=۱}^{n} \frac{\partial d(y_i; \mu_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

[3] Homogeneous Simplex Model

[4] Iteratively Re-Weighted Least Squares

اکنون هر یک از سه جمله در مجموع فوق را به‌صورت زیر محاسبه و خلاصه می‌کنیم

$$\frac{\partial d(y_i;\mu_i)}{\partial \mu_i} = d'(y_i;\mu_i) \tag{۱.۳}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{۱}{g'(\mu_i)} \tag{۲.۳}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j}\boldsymbol{x}_i^T\boldsymbol{\beta} = \frac{\partial}{\partial \beta_j}(x_{i۱}\beta_۱ + x_{i۲}\beta_۲ + \cdots + x_{ij}\beta_j + \cdots + x_{ip}\beta_p) = x_{ij} \tag{۳.۳}$$

در نهایت با توجه به روابط (۱.۳)، (۲.۳) و (۳.۳) داریم

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = -\frac{۱}{۲\sigma^۲}\sum_{i=۱}^{n} d'(y_i;\mu_i)\frac{۱}{g'(\mu_i)}x_{ij} = \sum_{i=۱}^{n} c_i x_{ij}$$

و از آن‌جا که

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial}{\partial \beta_۱}\ell(\boldsymbol{\beta}) \\ \\ \frac{\partial}{\partial \beta_۲}\ell(\boldsymbol{\beta}) \\ \vdots \\ \\ \frac{\partial}{\partial \beta_p}\ell(\boldsymbol{\beta}) \end{pmatrix}$$

داریم

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \sum_{i=۱}^{n} c_i x_{i۱} \\ \sum_{i=۱}^{n} c_i x_{i۲} \\ \vdots \\ \sum_{i=۱}^{n} c_i x_{ip} \end{pmatrix} = c_۱ \begin{pmatrix} x_{۱۱} \\ x_{۱۲} \\ \vdots \\ x_{۱p} \end{pmatrix} + \ldots + c_n \begin{pmatrix} x_{n۱} \\ x_{n۲} \\ \vdots \\ x_{np} \end{pmatrix}.$$

در نتیجه داریم

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=۱}^{n} \mathbf{x}_i c_i = -\sum_{i=۱}^{n} \mathbf{x}_i \frac{d'(y_i;\mu_i)}{۲\sigma^۲ g'(\mu_i)}.$$

برای استفاده از روش IRWLS بایستی تابع اطلاع فیشر را محاسبه کنیم

$$\mathrm{I}(\boldsymbol{\beta}) = -\mathrm{E}\left(\frac{\partial^۲ \ell\,(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right)$$

اما

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}^T}\left(\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right) &= \frac{\partial}{\partial\boldsymbol{\beta}^T}\left[-\sum_{i=1}^{n}\mathbf{x}_i\frac{d'(y_i;\mu_i)}{2\sigma^2 g'(\mu_i)}\right]\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\frac{\partial}{\partial\boldsymbol{\beta}^T}\left[\frac{d'(y_i;\mu_i)}{g'(\mu_i)}\right]\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\frac{\partial}{\partial\mu_i}\left(\frac{d'(y_i;\mu_i)}{g'(\mu_i)}\right)\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\boldsymbol{\beta}^T}\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\left[\frac{d''(y_i;\mu_i).g'(\mu_i)}{\{g'(\mu_i)\}^2}-\frac{g''(\mu_i).d'(y_i;\mu_i)}{\{g'(\mu_i)\}^2}\right]\frac{1}{g'(\mu_i)}\mathbf{x}_i^T\end{aligned}$$

بنابراین تابع $\mathrm{I}(\boldsymbol{\beta})$ عبارت است از

$$\mathrm{I}(\boldsymbol{\beta}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathbf{x}_i\left(\frac{E[d''(y_i;\mu_i)]}{g'(\mu_i)}-\frac{E[d'(y_i;\mu_i)]g''(\mu_i)}{(g'\{\mu_i)\}^2}\right)\frac{1}{g'(\mu_i)}\mathbf{x}_i^T$$

بنا به خاصیت شماره ۴ توزیع سیمپلکس از زیربخش (۳.۳.۲)، $\mathrm{E}\{d'(Y;\mu)\}=0$ و با در نظر گرفتن

$$w_i = \frac{\mathrm{E}[d''(y_i;\mu_i)]}{2\sigma^2\{g'(\mu_i)\}^2}$$

در نهایت داریم

$$\mathrm{I}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{x}_i w_i\mathbf{x}_i^T \qquad (۴.۳)$$

که با توجه به خاصیت شماره ۵ توزیع سیمپلکس از زیربخش ۳.۳.۲، یعنی $\frac{1}{2}\mathrm{E}\{d''(Y;\mu)\} = \frac{3\sigma^2}{\mu(1-\mu)}+\frac{1}{\mu^3(1-\mu)^3}$، داریم

$$w_i = \frac{1}{\sigma^2\{g'(\mu_i)\}^2}\left(\frac{3\sigma^2}{\pi_i}+\frac{1}{\pi_i^3}\right)$$

که در آن $\pi_i = \mu_i(1-\mu_i)$.

حال تابع امتیاز متناظر با β_jها را نیز به‌دست می‌آوریم

$$\begin{aligned}\mathrm{U}(\boldsymbol{\beta}) = \frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= -\sum_{i=1}^{n}\mathbf{x}_i^T\frac{d'(y_i;\mu_i)}{2\sigma^2 g'(\mu_i)}\\ &= -\sum_{i=1}^{n}\mathbf{x}_i^T\frac{\mathrm{E}[d''(y_i;\mu_i)]}{2\sigma^2\{g'(\mu_i)\}^2}\frac{d'(y_i;\mu_i)g'(\mu_i)}{\mathrm{E}[d''(y_i;\mu_i)]}.\end{aligned}$$

می‌دانیم

$$\begin{aligned}u(y;\mu) &= -\frac{1}{2}d'(y;\mu)\\ &= \frac{y-\mu}{\mu(1-\mu)}\left\{d(y;\mu)+\frac{1}{\mu^2(1-\mu)^2}\right\}.\end{aligned} \qquad (۵.۳)$$

بنابراین

$$\begin{aligned}\mathrm{U}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{x}_i^T w_i \frac{2u(y_i;\mu_i)g'(\mu_i)}{2\sigma^2 w_i\{g'(\mu_i)\}^2} \\ &= \sum_{i=1}^{n} \mathbf{x}_i^T w_i \left(\frac{u(y_i;\mu_i)}{\sigma^2 w_i g'(\mu_i)}\right) = \sum_{i=1}^{n} \mathbf{x}_i^T w_i \left(s_i - g(\mu_i)\right) \\ &= \sum_{i=1}^{n} \mathbf{x}_i^T w_i \left(s_i - \eta_i\right)\end{aligned} \tag{۶.۳}$$

که در آن $s_i = g(\mu_i) + \frac{u(y_i,\mu_i)}{\sigma_i^2 w_i g'(\mu_i)}$.

این روند، نشان‌دهنده یک الگوریتم تکراری است، که برآوردگر $(k+1)$امین $\boldsymbol{\beta}$ می‌باشد که به‌صورت زیر است:

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k + \mathrm{I}^{-1}\left(\boldsymbol{\beta}^k\right)\mathrm{U}\left(\boldsymbol{\beta}^k\right) \tag{۷.۳}$$

حال با توجه به روابط (۴.۳) و (۶.۳) در رابطه (۷.۳) داریم

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \Big(\sum_{i=1}^{m} \mathbf{x}_i^T w_i^{(k)} \mathbf{x}_i\Big)^{-1} \sum_{i=1}^{k} \mathbf{x}_i^T w_i^{(k)} \big(s_i^{(k)} - \eta_i^{(k)}\big).$$

۲.۱.۳ آماره انحراف

برای مقایسه مجموعه‌ای از مدل‌های رگرسیونی آشیانه‌ای[5] که متغیر پاسخ آن‌ها دارای توزیع نرمال است، عموما از روش تحلیل واریانس[6] استفاده می‌شود که بر اساس مجموع توان دوم مختلف بیان کننده تغییر پذیری داده‌ها هستند. برای GLMs این به تحلیل انحراف تعمیم می‌یابد.

فرض کنید $\hat{\mathrm{L}}_c$ و $\hat{\mathrm{L}}_f$ به‌ترتیب نشان دهنده مقدار ماکسیمم تابع درستنمایی برای مدل مورد نظر[7] و مدل اشباع‌شده[8] (مدلی با تمام متغیرهای تبیینی) باشند. آماره انحراف مقیاس‌بندی‌شده[9] برای یک مدل به‌صورت زیر تعریف می‌شود

$$\begin{aligned}D^* &= -2\log\left(\frac{\hat{\mathrm{L}}_c}{\hat{\mathrm{L}}_f}\right) \\ &= -2\left\{\log\hat{\mathrm{L}}_c - \log\hat{\mathrm{L}}_f\right\}\end{aligned} \tag{۸.۳}$$

که شکل مشابه‌ای با نسبت دو تابع درستنمایی دارد و در احتمال به‌طور مجانبی دارای توزیع کای‌دو با $n-p$ درجه آزادی است، که در آن n تعداد کل پارامترها و p تعداد پارامترهای

[5] Nested
[6] Analysis of Variance (ANOVA)
[7] Current Model
[8] Full Model
[9] Scaled Deviance

رگرسیونی نامعلومی است که در برازش مدل برآورد شده‌اند. برای داده‌ها با فرض داشتن توزیع نرمال وقتی واریانس معلوم و تابع پیوند همانی باشد این آماره دقیقا دارای توزیع کای‌دو است. اما در حالت کلی به دلیل ثابت نبودن تعداد پارامترها، ممکن است تقریب کای‌دو مفید نباشد. زمانی‌که یك مدل مناسب برازش شود و تقریب $\chi^{۲}_{n-p}$ برقرار باشد، مقدار تقریبی امید ریاضی رابطه (۸.۳) برابر $n-p$ است.
در GLMs با توجه به لگاریتم تابع درستنمایی زیر

$$\log \mathrm{L}(\beta) = \sum_{i=۱}^{n} \left\{ \frac{۱}{\phi} \left[\theta_i y_i + c\left(\theta_i\right) \right] + d\left(\phi, y_i\right) \right\}$$

داریم

$$\log \hat{\mathrm{L}}_c = \sum_{i=۱}^{n} \left\{ \frac{۱}{\phi} \left[\hat{\theta}_i y_i + c\left(\hat{\theta}_i\right) \right] + d\left(\phi, y_i\right) \right\}$$

که در آن $\hat{\theta}_i$، برآورد درستنمایی ماکسیمم θ_i است.
در مدل اشباع‌شده برآورد $y_i = \mu_i$ است و $\tilde{\theta}_i$ متناظر با برآورد θ_i است؛ حال با استفاده از این‌که $\mu_i = -c'\left(\theta_i\right)$، می‌توان $\tilde{\theta}_i$ را به‌صورت زیر محاسبه کرد

$$-c'\left(\tilde{\theta}_i\right) = y_i$$

سپس

$$\log \hat{\mathrm{L}}_f = \sum_{i=۱}^{n} \left\{ \frac{۱}{\phi} \left[\tilde{\theta}_i y_i + c\left(\tilde{\theta}_i\right) \right] + d\left(\phi, y_i\right) \right\}$$

و در آخر

$$D^* = \frac{۲}{\phi} \sum_{i=۱}^{n} \left\{ \left(\tilde{\theta}_i - \hat{\theta}_i\right) y_i + c\left(\tilde{\theta}_i\right) - c\left(\hat{\theta}_i\right) \right\}$$

حال آماره انحراف را می‌توان به‌صورت زیر بیان کرد

$$D = \phi D^* \qquad (۹.۳)$$

که در آن D آماره انحراف و ϕ پارامتر پراکندگی است. در حقیقت آماره انحراف برابر با حاصلضرب مقداری ثابت در انحراف مقیاس‌بندی‌شده است و زمانی که این مقدار ثابت بزرگ‌تر یا کوچك‌تر از یك شود، تمایز بین این دو مقدار نقش مهمی در تحلیل‌های رگرسیونی ایفا می‌کند. از این رو، آماره انحراف مستقل از پارامتر پراکندگی ϕ تعریف می‌شود. جدول ۱.۳ مقدار آماره انحراف مقیاس‌بندی‌شده را برای برخی از توزیع‌های معروف متعلق به خانواده نمایی نشان می‌دهد. در واقع آماره انحراف، آماره لگاریتم نسبت درستنمایی برای مقایسه مدل مورد نظر c، با مدل اشباع‌شده f است و یا به‌عبارتی آماره‌ای برای آزمون فرضی است که بیان می‌کند تمام پارامترهایی که در مدل اشباع‌شده وجود دارند و در مدل مورد نظر نیستند، برابر صفر هستند. در حقیقت هدف از محاسبه آماره انحراف تصمیم‌گیری نسبت به سودمندی پارامترهای اضافه

جدول ۱.۳: آماره انحراف مقیاس‌بندی‌شده برخی از توزیع‌های متعلق به خانواده نمایی

توزیع	آماره انحراف مقیاس بندی‌شده
نرمال	$\sum_i (y_i - \mu_i)^2$
پواسون	$2\sum_i \left(y_i \ln\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i) \right)$
گاما	$2\sum_i \left(\frac{y_i - \mu_i}{\mu_i} - \ln\left(\frac{y_i}{\mu_i}\right) \right)$
گوسی معکوس	$\sum_i \frac{(y_i - \mu_i)^2}{y_i \mu_i^2}$

شده به مدل مورد نظر است. این آماره برای مدل رگرسیون پواسون همان آماره آزمون نیکویی برازش برای این مدل است که این آزمون به وسیله مقایسه مقادیر مشاهده‌شده با مقادیر برازش‌شده، انجام می‌شود.

۳.۱.۳ برآورد پارامتر پراکندگی

عموما در تحلیل داده‌ها فرض آن‌که پارامتر پراکندگی σ^2 معلوم باشد، دور از ذهن است و در اکثر موارد این پارامتر نامعلوم است و محقق ملزم به برآورد آن می‌باشد. دراین حالت پارامتر پراکندگی را می‌توان توسط روش گشتاوری و با استفاده از عبارت زیر برآورد نمود

$$\widehat{\sigma^2} = \frac{1}{n-p} \sum_{i=1}^{n} w_i \hat{r}_i \qquad (۱۰.۳)$$

که در آن نیز n تعداد کل مشاهدات و p تعداد پارامترهای رگرسیونی نامعلومی است که در برازش متداول مدل به کمك باقی‌مانده‌های پیرسون به‌شکل $\hat{r}_i = \frac{(Y_i - \hat{\mu}_i)}{\sqrt{\mathrm{V}(\hat{\mu}_i)}}$، برآورد شده‌اند و w_i یك وزن معلوم است که محقق برای برخی مقادیر از Y_i یا تمام آن‌ها، به دلخواه انتخاب می‌کند.

براي برآورد پارامتر پراکندگی، برآوردگر دیگری مبنی بر آماره انحراف نیز وجود دارد. بر مبنای روابط (۸.۳) و (۹.۳) یك برآورد گشتاوری برای σ^2 به‌شکل زیر تعریف می‌شود

$$\widehat{\sigma^2} = \frac{1}{n-p} \sum_{i=1}^{n} d(y_i; \mu_i) \qquad (۱۱.۳)$$

این برآوردگر گشتاوری به‌طور مجانبی ناریب است (کیدیم و فوکیانوز، ۲۰۰۵).

۴.۱.۳ تحلیل باقی‌مانده‌ها برای مدل رگرسیونی سیمپلکس

در این بخش، به تحلیل باقی‌مانده‌ها می‌پردازیم و انواع باقی‌مانده‌ها را معرفی می‌کنیم. هدف از تحلیل باقی‌مانده‌ها، ارزیابی نیکویی برازش مدل و تشخیص تاثیر متغیرهای تبیینی بر روی متغیر پاسخ است. در حقیقت انحراف یك مقدار مشاهده‌شده از مقدار برآوردشده آن را باقی‌مانده می‌نامند. در یك مدل که در آن متغیرهای پاسخ Y به‌صورت $\mathrm{Y} = \mathrm{X}(\boldsymbol{\beta}) + \epsilon$ مدل‌سازی می‌شوند و در آن مولفه‌های ϵ مستقل و هم توزیع، $\epsilon \sim \mathrm{N}(۰, \sigma^۲)$، هستند، باقی‌مانده‌های پاسخ متناظر برای مقادیر $i = ۱, \cdots, n$ به‌صورت زیر تعریف می‌شوند:

$$\hat{e}_i = y_i - \hat{\mu}_i$$

که در آن $\hat{\mu}_i$ مقدار برازش‌شده حاصل از برآورد ML بردار ضرایب رگرسیون $\boldsymbol{\beta}$ است و ساده‌ترین نوع باقی‌مانده است.
یکی دیگر از انواع باقی‌مانده‌های مورد استفاده، باقی‌مانده‌های پیرسونی است که فراری و سریباری (۲۰۰۴) آن را باقی‌مانده‌های معمولی استانداردشده نامیدند و به‌صورت زیر بیان می‌شود:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mathrm{V}}(y_i)}} = \frac{y_i - \hat{\mu}_i}{\hat{\tau}_i}. \qquad (۱۲.۳)$$

از آن‌جایی که $\mathrm{V}(y_i)$ صورت بسته‌ای ندارد، برای داده‌های بیش‌پراکنده که پارامتر پراکندگی $\sigma^۲$ خیلی بزرگ است و واریانس متغیر پاسخ به سمت $\mu(۱ - \mu)$ میل می‌کند، باقی‌مانده پیرسون مجانبی برابر است با

$$r_i^a = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(۱ - \hat{\mu}_i)}}. \qquad (۱۳.۳)$$

حال برای ارزیابی نیکویی برازش مدل مد نظر با استفاده از باقی‌مانده‌ها، اگر مدل درست انتخاب شده باشد، آن‌گاه باقی‌مانده‌ها باید دارای توزیع نرمال استاندارد مجانبی باشند و نقاط در نمودار پراکنش مقادیر برآوردشده در مقابل باقی‌مانده‌ها، بدون نمایش هیچ الگوی منظمی عمدتا در بازه $(-۲, ۲)$ قرار دارند.

۵.۱.۳ مطالعه شبیه‌سازی

در این بخش، ویژگی‌های برآوردگرهای پارامترهای مدل رگرسیونی سیمپلکس که در این فصل، به‌طور نظری بیان کردیم، برای حجم‌های نمونه متفاوتی مورد ارزیابی قرار می‌دهیم. شبیه‌سازی‌ها در محیط نرم‌افزار R و با کمک بسته simplexreg اجرا شده‌اند. نحوه برازش یک مدل رگرسیونی سیمپلکس با استفاده از بسته simplexreg را در فصل ۵ به‌طور کامل تشریح خواهیم کرد.

مثال ۱.۱.۳. مثال اول

برای مثال اول شبیه‌سازی، در مدل (۱۱.۲)، فرض می‌کنیم پیشگوی خطی به‌صورت

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{1i}, \qquad i = 1, \cdots, n. \tag{۱۴.۳}$$

باشد، که در آن $(\beta_0, \beta_1) = (1/5, 0/5)$ و پارامتر پراکندگی $\sigma^2 = 2$ انتخاب شدند. مقادیر متغیر تبیینی x_1 را از توزیع برنولی با احتمال موفقیت ۰/۶ تولید کردیم؛ هم‌چنین برای بررسی عملکرد روش، پنج حجم نمونه $n = 30, 50, 100, 200, 400$ را در نظر گرفتیم. در جدول ۲.۳ برآوردهای نهایی هر پارامتر که با میانگین‌گیری از برآوردهای حاصله در ۵۰۰ تکرار، محاسبه شده، نشان داده شده‌اند.

جدول ۲.۳: خلاصه نتایج شبیه‌سازی برای برآورد ضرایب مدل (۱۱.۲) با پیشگوی خطی (۱۴.۳)

حجم نمونه	پارامترها	مقدار واقعی	برآورد	Stdev	MSE	اریبی
۳۰	β_0	۱/۵	۱/۵۱۴۰	۰/۱۶۳۴	۰/۰۲۶۸	۰/۰۱۴
	β_1	۰/۵	۰/۴۷۴۲	۰/۱۹۶۳	۰/۰۳۹	−۰/۰۲۵۷
	σ^2	۲	۱/۶۰۹۶	۰/۲۶۲۸	۰/۲۲۱۲	−۰/۳۹۰۳
۵۰	β_0	۱/۵	۱/۵۱۳۱	۰/۱۱۳۶	۰/۰۱۳	۰/۰۱۳۱
	β_1	۰/۵	۰/۴۷۲۸	۰/۱۳۹۳	۰/۰۲	−۰/۰۲۷۱
	σ^2	۲	۱/۷۰۸۲	۰/۲۳۰۱	۰/۱۳۷۹	−۰/۲۹۱۷
۱۰۰	β_0	۱/۵	۱/۵۰۵۷	۰/۰۷۶۲	۰/۰۰۵۸	۰/۰۰۵۷
	β_1	۰/۵	۰/۴۸۵۸	۰/۰۹۳۵	۰/۰۰۸۹	−۰/۰۱۴۱
	σ^2	۲	۱/۷۹۳۲	۰/۱۶۲۶	۰/۰۶۹	−۰/۲۰۶۷
۲۰۰	β_0	۱/۵	۱/۵۰۲۲	۰/۰۵۹۷	۰/۰۰۳۵	۰/۰۰۲۲
	β_1	۰/۵	۰/۴۹۳۱	۰/۰۷۳۲	۰/۰۰۵۳	−۰/۰۰۶۸
	σ^2	۲	۱/۸۳۹۸	۰/۱۱۳۶	۰/۰۳۸۵	−۰/۱۶۰۱
۴۰۰	β_0	۱/۵	۱/۵۰۱۱	۰/۰۴۰۳	۰/۰۰۱۶	۰/۰۰۱۱
	β_1	۰/۵	۰/۴۹۷۱	۰/۰۴۸۷	۰/۰۰۲۳	−۰/۰۰۲۸
	σ^2	۲	۱/۸۸۲۹	۰/۰۸۶۲	۰/۰۲۱۱	−۰/۱۱۷

همان‌گونه که مشخص است میانگین مقادیر برآورد پارامترها، بسیار نزدیک به مقادیر واقعی است و با افزایش حجم نمونه، دقت این برآوردها به واسطه کاهش خطای معیار، افزایش می‌یابد. هم‌چنین با افزایش حجم نمونه، MSE برآوردها در حال کاهش و همگرا به صفر است و مقادیر اریبی برآوردگرها نیز ناچیز است. بنابراین، می‌توان سازگاری برآوردگرها را برای حجم‌های نمونه کوچک و متوسط نیز نتیجه گرفت.

مثال ۲.۱.۳. مثال دوم

برای شبیه‌سازی در مدل (۱۱.۲)، فرض می‌کنیم که پیشگوی خطی به‌صورت

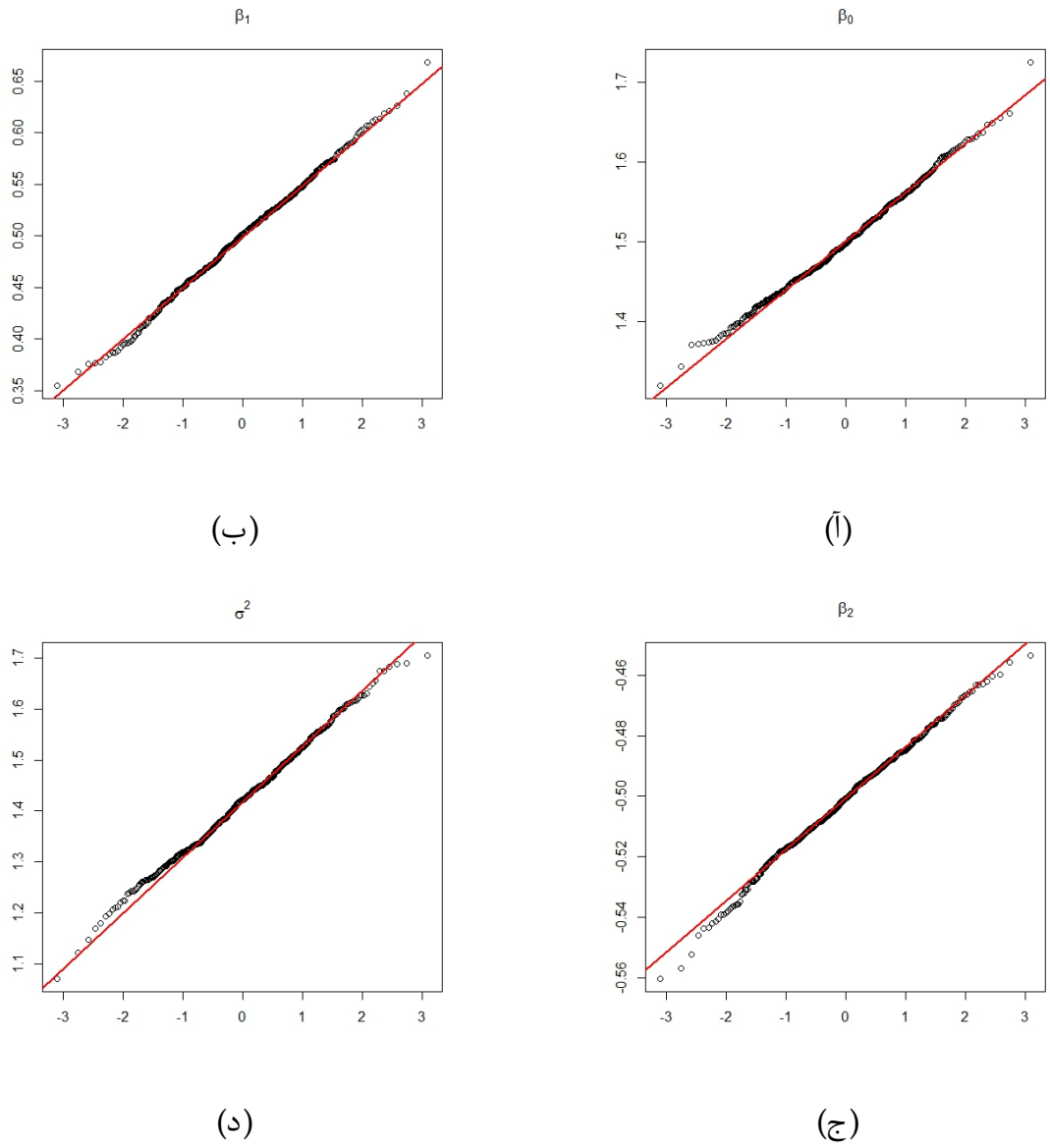
$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \qquad i = 1, \cdots, n. \tag{۱۵.۳}$$

باشد، که در آن $(\beta_0, \beta_1, \beta_2) = (۱/۵, ۰/۵, -۰/۵)$ و پارامتر پراکندگی $\sigma^2 = ۲$ انتخاب شدند. مقادیر متغیر تبیینی x_1 را از توزیع برنولی با احتمال موفقیت ۰/۶ و مقادیر x_2 را از توزیع $U(0,5)$ تولید کردیم. هم‌چنین برای بررسی عملکرد روش، پنج حجم نمونه $n = ۳۰, ۵۰, ۱۰۰, ۲۰۰, ۴۰۰$ را در نظر گرفتیم. در جدول ۳.۳ برآوردهای نهایی هر پارامتر که با میانگین‌گیری از برآوردهای حاصله در ۵۰۰ تکرار، محاسبه شده، نشان داده شده‌اند.

جدول ۳.۳: خلاصه نتایج شبیه‌سازی برای برآورد ضرایب مدل (۱۱.۲) با پیشگوی خطی (۱۵.۳)

حجم نمونه	پارامترها	مقدار واقعی	برآورد	Stdev	MSE	اریبی
۳۰	β_0	۱/۵	۱/۳۸۴	۰/۲۱۸۷	۰/۰۶۰۷	-۰/۱۱۶
	β_1	۰/۵	۰/۴۷۶۴	۰/۱۷۵۱	۰/۰۳۰۸	-۰/۰۲۳۶
	β_2	-۰/۵	-۰/۴۵۳۷	۰/۰۶۳	۰/۰۰۶	۰/۰۴۶۳
	σ^2	۲	۱/۳۷۷۸	۰/۲۸	۰/۰۹۵۷	-۰/۶۲۲۲
۵۰	β_0	۱/۵	۱/۴۱۳۴	۰/۱۷۶۷	۰/۰۳۸۱	-۰/۰۸۶۶
	β_1	۰/۵	۰/۴۶۳۲	۰/۱۶۲۲	۰/۰۲۷۱	-۰/۰۳۶۸
	β_2	-۰/۵	-۰/۴۶۲۸	۰/۰۵۴۷	۰/۰۰۴۳	۰/۰۳۷۲
	σ^2	۲	۱/۴۶۶۵	۰/۱۷۵۷	۱/۱۵۹۶	-۰/۵۳۳۵
۱۰۰	β_0	۱/۵	۱/۴۰۸۷	۰/۱۴۶۳	۰/۰۲۸۶	-۰/۰۹۱۳
	β_1	۰/۵	۰/۴۸۷۴	۰/۱۳۴۴	۰/۰۱۷۳	-۰/۰۱۲۶
	β_2	-۰/۵	-۰/۴۵۷۸	۰/۰۴۶۵	۰/۰۰۳۸	۰/۰۴۲۲
	σ^2	۲	۱/۵۴۲	۰/۱۵۱۳	۰/۵۷۲۵	-۰/۴۵۸
۲۰۰	β_0	۱/۵	۱/۴۲۵	۰/۰۲۵۵	۰/۰۰۵۹	-۰/۰۷۵
	β_1	۰/۵	۰/۳۹۲۲	۰/۰۲۷	۰/۰۱۱۹	-۰/۱۰۷۸
	β_2	-۰/۵	-۰/۴۴۹۵	۰/۰۰۶۳	۰/۰۰۲۵	۰/۰۵۰۵
	σ^2	۲	۱/۶۹۵	۰/۰۲۲۹	۰/۰۲۴۸	-۰/۳۰۵
۴۰۰	β_0	۱/۵	۱/۵۰۱۶	۰/۰۶۸۳	۰/۰۰۴۶	۰/۰۱۶
	β_1	۰/۵	۰/۴۹۸۷	۰/۰۵۸۲	۰/۰۰۳۳	-۰/۰۰۱۳
	β_2	-۰/۵	-۰/۵۰۱۵	۰/۰۲	۰/۰۰۰۴	-۰/۰۰۱۵
	σ^2	۲	۲/۰۰۶۹	۰/۱۴۶۵		۰/۰۰۶۹

روند تغییرات انحراف معیار برآوردگرها، مشابه مثال اول هستند و در این‌جا نیز می‌توان سازگاری برآوردگرها را نتیجه گرفت. شکل ۱.۳، نمودارهای چندک‌ـ‌چندک پارامترهای برآورد شده در ۵۰۰ مجموعه داده شبیه‌سازی‌شده را برای پیشگوی خطی (۱۵.۳) نمایش می‌دهد. با توجه به این نمودارها، می‌توان نرمال مجانبی برآوردگرها را نتیجه گرفت، زیرا توزیع نرمال تقریب مناسبی برای توزیع برآوردگرها است. با توجه به نتایج به‌دست آمده می‌توان گفت که روش پیشنهادی قابل اعتماد است.



شکل ۱.۳: نمودار چندک‌ـ‌چندک برآوردهای پارامترهای مدل (۱۱.۲)، با پیشگوی خطی (۱۵.۳)

۲.۳ مدل رگرسیونی سیمپلکس با پارامتر پراکندگی متغیر

در بخش قبل مدل رگرسیونی سیمپلکس را برای حالتی که پارامتر پراکندگی، $\sigma^۲$، ثابت بود، مورد بررسی قرار دادیم.
در این بخش، مدل را برای حالتی که پارامتر $\sigma_i^۲$ متغیر و خود تابعی از متغیرهای تبیینی است، تعمیم می‌دهیم. هم‌چنین به برازش مدل با استفاده از روش درستنمایی ماکسیمم می‌پردازیم.

۱.۲.۳ معرفی مدل

همان‌طور که در فصل قبل بیان کردیم، متغیر پاسخ y دارای توزیع سیمپلکس به‌صورت (۱۱.۲) است. اکنون فرض می‌کنیم پارامتر پراکندگی $\sigma^۲$ نیز تابعی از مشاهدات است، به‌طوری‌که متغیرهای تبیینی $z_i, i = ۱, \ldots, m$ که می‌توانند زیرمجموعه‌ای از x_i ها است، بر روی آن تاثیر دارند. فرض کنید میانگین و پارامتر پراکندگی برای y_i به‌صورت زیر مدل‌بندی شوند:

$$g(\mu_i) = \eta_{۱i} = \boldsymbol{x}_i^T \boldsymbol{\beta}, \qquad h(\sigma_i^۲) = \eta_{۲i} = \boldsymbol{z}_i^T \gamma. \tag{۱۶.۳}$$

که در آن $g : (۰,۱) \longrightarrow (-\infty, \infty)$ و $h : (۰, \infty) \longrightarrow (-\infty, \infty)$. این مدل را سیمپلکس ناهمگن[۱۰] گویند که در آن هر دو تابع پیوند، اکیدا یکنوا و مشتق‌پذیر می‌باشند؛ $\boldsymbol{\beta}$ و γ نیز بردارهای پارامترهای نامعلوم برای توابع میانگین و پراکندگی است؛ هم‌چنین $\eta_{۱i}$ و $\eta_{۲i}$، پیشگوهای خطی‌اند و $\boldsymbol{x}_i$ و z_i متغیرهای تبیینی معلوم می‌باشند.
تابع پیوند $g(.)$ همان تابع پیوندی است که در فصل ۲ معرفی شده، $g(\mu) = \log \frac{\mu}{۱ - \mu}$، اما تابع پیوندی که برای $h(.)$ در نظر گرفته می‌شود، تابع پیوند لگاریتمی است، یعنی $h(\sigma_i^۲) = \log \sigma_i^۲$.

۲.۲.۳ برازش مدل

تابع لگاریتم درستنمایی برای یک نمونه به حجم n، برای این مدل رگرسیون سیمپلکس به‌صورت زیر است

$$\ell(\boldsymbol{\beta}, \gamma) = \sum_{i=۱}^{n} \ell_i(\mu_i, \sigma_i^۲) \tag{۱۷.۳}$$

که در آن

$$\ell_i(\beta_i, \gamma_i) = -\frac{۱}{۲} \sum_{i=۱}^{m} \left\{ \frac{d(y_i; \mu_i)}{\sigma_i^۲} - \log \sigma_i^۲ \right\}.$$

توابع $\mu_i = g^{-۱}(\eta_{۱i})$ و $\sigma_i^۲ = h^{-۱}(\eta_{۲i})$ که در رابطه (۱۶.۳) تعریف شده‌اند، به‌ترتیب توابعی از β و γ می‌باشند. برای برآورد پارامترها، مشابه زیربخش ۱.۱.۳، باید تابع امتیاز متناظر با γ را

[10] Simplex Heterogeneous Model

نیز به‌دست آوریم.
با استفاده از قاعده زنجیره‌ای، برای محاسبه تابع امتیاز U، با مشتق‌گیری از تابع لگاریتم درستنمایی (۱۷.۳)، نسبت به γ_j، $j = ۱, \cdots, n$ داریم

$$\mathrm{U}(\gamma) = \frac{\partial \ell(\boldsymbol{\beta}, \gamma)}{\partial \gamma_j} = -\frac{۱}{۲} \sum_{i=۱}^{n} \left\{ \frac{\partial d(y_i; \mu_i) - \log(\sigma_i^۲)}{\partial \sigma_i^۲} \frac{\partial \sigma_i^۲}{d\eta_{۲i}} \frac{d\eta_{۲i}}{\partial \gamma_j} \right\}. \tag{۱۸.۳}$$

بنابراین برای محاسبه (۱۸.۳)، باید مشتق‌های جزئی مجموع سمت راست آن را به‌ترتیب محاسبه و جایگذاری کنیم (ژانگ و ژو، ۲۰۱۴). می‌توان نوشت

$$\frac{\partial d(y_i; \mu_i)}{\partial \sigma_i^۲} = \left(\frac{-d(y_i; \mu_i)}{(\sigma_i^۲)^۲} + \frac{۱}{\sigma_i^۲}\right) \tag{۱۹.۳}$$

حال با توجه به این‌که $\sigma_i^۲ = h^{-۱}(\eta_{۲i})$، داریم

$$\frac{\partial \sigma_i^۲}{d\eta_{۲i}} = \frac{۱}{h'(\sigma_i^۲)}, \tag{۲۰.۳}$$

و

$$\frac{d\eta_{۲i}}{\partial \gamma_j} = z_{ij} \tag{۲۱.۳}$$

در نهایت با جایگذاری (۱۹.۳)، (۲۰.۳) و (۲۱.۳) در رابطه (۱۸.۳)، نتیجه می‌شود

$$\mathrm{U}(\gamma) = -\frac{۱}{۲} \sum_{i=۱}^{n} \left(\frac{-d(y_i; \mu_i)}{(\sigma_i^۲)^۲} + \frac{۱}{\sigma_i^۲}\right) \frac{۱}{h'(\sigma_i^۲)} z_{ij}$$

قرار می‌دهیم

$$v_i = \frac{d(y_i; \mu_i) - \sigma_i^۲}{۲(\sigma_i^۲)^۲ h'(\sigma_i^۲)}, \quad i = ۱, \ldots, n.$$

حال می‌توان رابطه را به‌صورت زیر بازنویسی کرد

$$\mathrm{U}(\gamma) = \sum_{i=۱}^{n} z_i^T v_i \tag{۲۲.۳}$$

برای حل معادله‌های تابع درستنمایی، باید از مشتق دوم تابع لگاریتم درستنمایی، امید ریاضی بگیریم. به‌همین منظور می‌توان نوشت:

$$\mathrm{I}(\gamma) = -\mathrm{E}\left(\frac{\partial^۲ \ell(\gamma)}{\partial \gamma \partial \gamma^T}\right).$$

ابتدا عبارت داخل امید ریاضی را محاسبه کرده سپس امید ریاضی می‌گیریم.

$$\begin{aligned}
\frac{\partial^۲ \ell(\gamma)}{\partial \gamma \partial \gamma^T} &= \frac{\partial}{\partial \gamma}\left[-\frac{۱}{۲} \sum_{i=۱}^{n} \left(\frac{-d(y_i; \mu_i)}{(\sigma_i^۲)^۲} + \frac{۱}{\sigma_i^۲}\right) \frac{۱}{h'(\sigma_i^۲)} \mathbf{z}_i\right] \\
&= -\frac{۱}{۲} \sum_{i=۱}^{n} \frac{\partial}{\partial \sigma_i^۲}\left[\frac{۱}{h'(\sigma_i^۲)}\left(\frac{-d(y_i; \mu_i)}{(\sigma_i^۲)^۲} + \frac{۱}{\sigma_i^۲}\right)\right] \frac{\partial \sigma_i^۲}{\partial \eta_{۲i}} \frac{\partial \eta_{۲i}}{\partial \gamma^T} \mathbf{z}_i
\end{aligned}$$

با استفاده از قاعده زنجیره‌ای فرمول بالا را باز نویسی می‌کنیم

$$\begin{aligned}
\frac{\partial^۲ \ell(\gamma)}{\partial\gamma\partial\gamma^T} &= -\frac{۱}{۲}\sum_{i=۱}^{n}\frac{\partial}{\partial\sigma_i^۲}\left[\frac{-d(y_i;\mu_i)}{h'(\sigma_i^۲)(\sigma_i^۲)^۲}+\frac{۱}{h'(\sigma_i^۲)\sigma_i^۲}\right]\frac{۱}{h'(\sigma_i^۲)}\mathbf{z}_i^T\mathbf{z}_i \\
&= -\frac{۱}{۲}\sum_{i=۱}^{n}\left[\frac{(h''(\sigma_i^۲)(\sigma_i^۲)^۲+۲\sigma_i^۲h'(\sigma_i^۲))d(y_i;\mu_i)}{h'^۲(\sigma_i^۲)(\sigma_i^۲)^۴}-\frac{h''(\sigma_i^۲)\sigma_i^۲+h'(\sigma_i^۲)}{h'^۲(\sigma_i^۲)(\sigma_i^۲)^۲}\right]\frac{۱}{h'(\sigma_i^۲)}\mathbf{z}_i^T\mathbf{z}_i \\
&= -\sum_{i=۱}^{n}\left[\frac{\mathbf{z}_i^T\mathbf{z}_i(h''(\sigma_i^۲)\sigma_i^۲d(y_i;\mu_i)+۲h'(\sigma_i^۲)d(y_i;\mu_i))}{۲h'^۳(\sigma_i^۲)(\sigma_i^۲)^۳}\right] \\
&+\sum_{i=۱}^{n}\left[\frac{\mathbf{z}_i^T\mathbf{z}_i(h''(\sigma_i^۲)\sigma_i^۲+h'(\sigma_i^۲))}{۲h'^۳(\sigma_i^۲)(\sigma_i^۲)^۲}\right] \\
&= -\sum_{i=۱}^{n}\left[\frac{\mathbf{z}_i^T\mathbf{z}_ih''(\sigma_i^۲)\sigma_i^۲d(y_i;\mu_i)}{۲h'^۳(\sigma_i^۲)(\sigma_i^۲)^۳}\right]-\sum_{i=۱}^{n}\left[\frac{۲\mathbf{z}_i^T\mathbf{z}_id(y_i;\mu_i)}{۲h'^۲(\sigma_i^۲)(\sigma_i^۲)^۳}\right]+\sum_{i=۱}^{n}\left[\frac{\mathbf{z}_i^T\mathbf{z}_ih''(\sigma_i^۲)}{۲h'^۳(\sigma_i^۲)\sigma_i^۲}\right] \\
&+\sum_{i=۱}^{n}\left[\frac{\mathbf{z}_i^T\mathbf{z}_i}{۲h'^۲(\sigma_i^۲)(\sigma_i^۲)^۲}\right] \\
&= \sum_{i=۱}^{n}\frac{\mathbf{z}_i\mathbf{z}_i^Th''(\sigma_i^۲)}{۲h'^۳(\sigma_i^۲)}\left(-\frac{d(y_i;\mu_i)}{(\sigma_i^۲)^۲}+\frac{۱}{\sigma_i^۲}\right)-\sum_{i=۱}^{n}\frac{\mathbf{z}_i\mathbf{z}_i^T}{۲h'^۲(\sigma_i^۲)}\left(\frac{۲d(y_i;\mu_i)}{(\sigma_i^۲)^۳}-\frac{۱}{(\sigma_i^۲)^۲}\right)
\end{aligned}$$

حال از طرفین امید ریاضی می‌گیریم؛ با توجه به ویژگی ۴ در زیربخش ۳.۳.۲ داریم

$$\mathrm{E}[-\frac{\partial^۲\ell(\gamma)}{\partial\gamma\partial\gamma^T}]=\sum_{i=۱}^{n}\frac{\mathbf{z}_i\mathbf{z}_i^T}{۲h'^۲(\sigma_i^۲)}\left(\frac{۱}{(\sigma_i^۲)^۲}\right)$$

می‌توان رابطه بالا را به‌صورت زیر بازنویسی می‌کنیم

$$= \mathrm{E}[-\frac{\partial^۲\ell(\gamma)}{\partial\gamma\partial\gamma^T}]=\sum_{i=۱}^{n}\mathbf{z}_i^Tq_i\mathbf{z}_i \qquad (۲۳.۳)$$

که در آن $q_i=\frac{۱}{۲(\sigma_i^۲)^۲h'^۲(\sigma_i^۲)}$. این روند نشان‌دهنده یک الگوریتم تکراری است، که برآوردگر $(k+۱)$امین γ با توجه به رابطه (۷.۳) و روابط (۲۲.۳) و (۲۳.۳) به‌صورت زیر است:

$$\gamma^{(k+۱)}=\gamma^{(k)}+\Big(\sum_{i=۱}^{n}\mathbf{z}_i^Tq_i\mathbf{z}_i\Big)^{-۱}\sum_{i=۱}^{n}\mathbf{z}_i^Tv_i^{(k)}$$

۳.۲.۳ مطالعه شبیه‌سازی

روند کار به‌همان صورتی است که در بخش قبل انجام شد، با این تفاوت که در این بخش پارامتر پراکندگی مدل به‌صورت یک تابع است. یعنی باید برای پارامتر پراکندگی، همانند پارامتر میانگین، μ، عمل کنیم. در نتیجه مطالعه شبیه‌سازی را با استفاده از تابع پیوند لجیت برای میانگین و تابع پیوند لگاریتمی برای پارامتر پراکندگی، انجام دادیم. مدل مورد نظر را به‌صورت

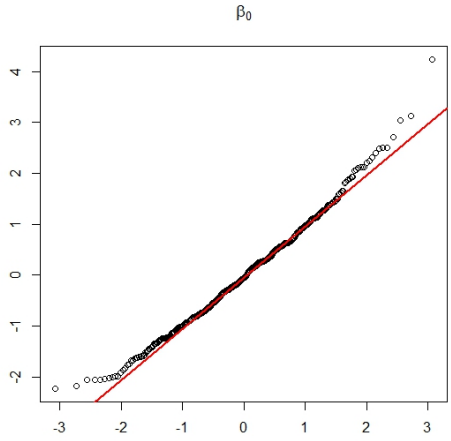
$$\mathrm{logit}(\mu_i)=\eta_{۱i}=\beta_۰+\beta_۱x_{۱i}+\beta_۲x_{۲i},\quad \log(\sigma_i^۲)=\eta_{۲i}=\gamma_۰+\gamma_۱z_{۱i},\quad i=۱,\cdots,n \qquad (۲۴.۳)$$

در نظر گرفتیم، که مقادیر واقعی ضرایب عبارتند از $(\beta_0, \beta_1, \beta_2) = (1, 0.5, -1)$ و $(\gamma_0, \gamma_1) = (5, 0.3)$. مقادیر متغیرهای تبیینی x_1 و x_2 را به‌ترتیب از توزیع $N(0, 1)$ و توزیع برنولی با احتمال موفقیت ۰/۶، تولید کردیم. مقادیر متغیر تبیینی z_1 را نیز از توزیع $N(0, 1)$ تولید کردیم. هم‌چنین برای بررسی عملکرد روش، پنج حجم نمونه $n = 30, 50, 100, 200, 400$ را در نظر گرفتیم. در جدول ۴.۳ برآوردهای نهایی پارامترها که با میانگین‌گیری از برآوردهای حاصله در ۵۰۰ تکرار، محاسبه شده نشان داده شده‌اند که با افزایش حجم نمونه MSE کاهش یافته و هم‌چنین اریبی برآوردگرها ناچیز است.

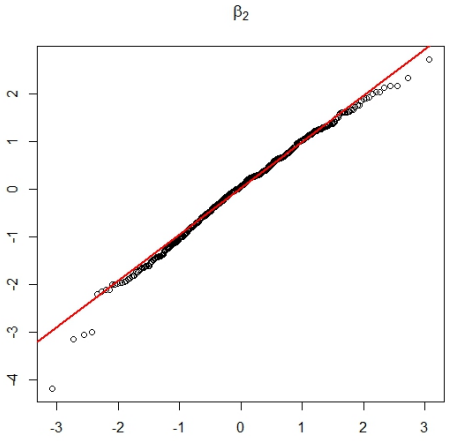
شکل ۲.۳ نیز منحنی چگالی مقادیر برآوردشده از مجموعه داده شبیه‌سازی‌شده را برای ضرایب مدل (۲۴.۳) با $n = 400$ نمایش می‌دهد. این شکل نیز تایید دیگری بر خوب بودن تقریب نرمال برای توزیع برآوردگرهای پارامتر مدل است. با توجه به نتایج به‌دست آمده، می‌توان عملکرد مناسب روش برازش پیشنهادی را نتیجه گرفت.

جدول ۴.۳: خلاصه نتایج شبیه‌سازی برای برآورد ضرایب مدل (۱۱.۲) با پیشگوی خطی (۲۴.۳)

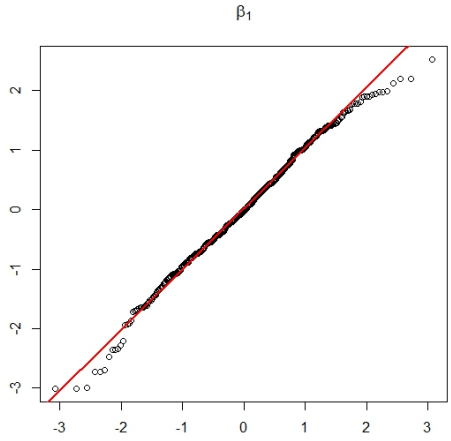
حجم نمونه	پارامترها	مقدار واقعی	میانگین برآورد	Stdev	MSE	اریبی
۳۰	β_0	۱	۱/۰۶۹۳	۰/۴۴۲۸	۰/۲۰۰۵	۰/۰۶۹۳
	β_1	۰/۵	۰/۵۳۲	۰/۳۱۴۳	۰/۰۹۹۶	۰/۰۳۲
	β_2	−۱	−۱/۰۴۷۶	۰/۵۳۸۵	۰/۲۹۱۶	−۰/۰۴۷۶
	γ_0	۵	۴/۸۱۰۶	۰/۲۷۲۹	۰/۱۱۰۲	−۰/۱۸۹۴
	γ_1	۰/۳	۰/۲۷۳۳	۰/۳۲۱۵	۰/۱۰۳۸	−۰/۰۲۶۷
۵۰	β_0	۱	۱/۰۵۳۱	۰/۳۳۳۳	۰/۱۱۳۷	۰/۰۵۳۱
	β_1	۰/۵	۰/۵۵۶	۰/۲۳۲۳	۰/۰۵۷	۰/۰۵۶
	β_2	−۱	−۱/۰۵۹۹	۰/۴۲۴۹	۰/۱۸۳۷	−۰/۰۵۹۹
	γ_0	۵	۴/۸۶۹۳	۰/۲۰۸۱	۰/۰۶۰۲	−۰/۱۳۰۶
	γ_1	۰/۳	۰/۲۹۸۹	۰/۲۱۱۴	۰/۰۴۴۶	−۰/۰۰۱
۱۰۰	β_0	۱	۱/۰۶۲	۰/۲۶۴۴	۰/۰۷۳۶	۰/۰۶۲
	β_1	۰/۵	۰/۵۲۶۵	۰/۱۴۸۹	۰/۰۲۲۸	۰/۰۲۶۵
	β_2	−۱	−۱/۰۵۸۹	۰/۳۰۷۷	۰/۰۹۷۹	−۰/۰۵۸۹
	γ_0	۵	۴/۹۴۰۳	۰/۱۴۳۴	۰/۰۲۴	−۰/۰۵۹۶
	γ_1	۰/۳	۰/۲۸۴۵	۰/۲۸۴۵	۰/۰۲۲۵	−۰/۰۱۵۴
۲۰۰	β_0	۱	۱/۰۰۷۹	۰/۱۵۷۳	۰/۰۲۴۷	۰/۰۰۷۹
	β_1	۰/۵	۰/۵۱۴۹	۰/۰۹۳۵	۰/۰۰۸۹	۰/۰۱۴۹
	β_2	−۱	−۱/۰۰۶	۰/۱۹۸۴	۰/۰۳۹۳	−۰/۰۰۶
	γ_0	۵	۴/۹۷۰۶	۰/۰۹۹	۰/۰۱۰۶	−۰/۰۲۹۳
	γ_1	۰/۳	/۳۰۲۶	۰/۱۰۵۵	۰/۰۱۱۱	۰/۰۰۲۶
۴۰۰	β_0	۱	۱/۰۰۱۳	۰/۱۰۶۸	۰/۰۱۱۴	۰/۰۰۱۳
	β_1	۰/۵	۰/۵۰۷۹	۰/۰۶۶۵	۰/۰۰۴۴	۰/۰۰۷۹
	β_2	−۱	−۰/۹۹۳	۰/۱۳۱۶	۰/۰۱۷۳	۰/۰۰۶۹
	γ_0	۵	۴/۹۸۵۹	۰/۰۷۳۷	۰/۰۰۵۶	−۰/۰۱۴۱
	γ_1	۰/۳	۰/۲۹۳۶	۰/۰۶۸	۰/۰۰۴۶	−۰/۰۰۶۳



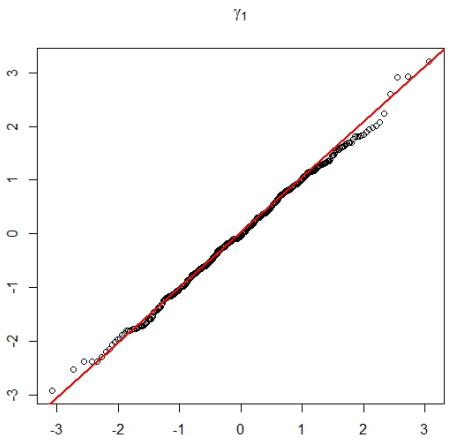
(آ)



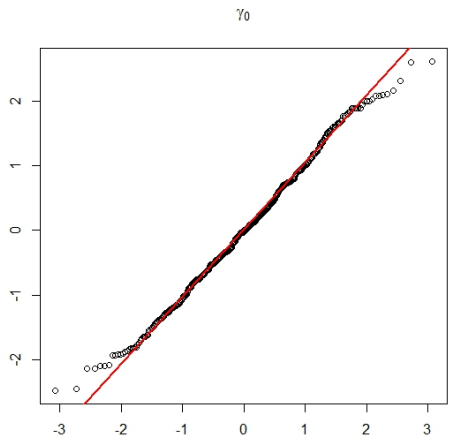
(ج)



(ب)



(ه)



(د)

شکل ۲.۳: نمودار چندک‌ـچندک برآوردهای ضرایب رگرسیونی میانگین و پارامتر پراکندگی در مدل (۲۴.۳) برای $n = ۴۰۰$ و تعداد تکرار ۵۰۰.

فصل ۴

مدل حاشیه‌ای برای تحلیل داده‌های نسبت طولی

مقدمه

امروزه استفاده از مطالعات طولی در بسیاری از علوم، نظیر علوم اجتماعی و اقتصادی به خوبی رواج یافته است. مطالعات طولی در قیاس با مطالعات مقطعی و سری زمانی، به دلیل این‌که واحدهای مورد نظر را در طول زمان بررسی می‌نمایند، از قابلیت بیشتری جهت مطالعه پدیده‌های مختلف برخوردارند.

در موقعیت‌هایی که با اندازه‌های تکراری در داده‌ها مواجه هستیم، به‌دلیل وجود همبستگی برای هر واحد بین اندازه‌ها، نمی‌توانیم از مدل‌های خطی استفاده نماییم و بایستی معادلات برآوردیابی دیگری را به کار بریم، زیرا در یك مدل رگرسیون خطی کلاسیك پذیره‌های استقلال مشاهدات، ثابت بودن واریانس و وجود ارتباط خطی بین مشاهدات و متغیرهای تصادفی از الزامات تعریف مدل، برای صحیح بودن نتایج تحلیل است و در GLM تنها دانستن واریانس مشاهدات، به‌عنوان تابعی از میانگین و وابسته بودن میانگین مشاهدات به تابعی خطی یا غیرخطی از متغیرهای تبیینی توسط تابع پیوند، کافی است. در حالی‌که استفاده از تحلیل‌هایی بدون در نظر گرفتن همبستگی موجب صحیح نبودن برآورد مولفه‌های واریانس و نیز استنباط اشتباه نسبت به بردار ضرایب رگرسیونی می‌شود. به همین دلیل در فصل ۱ به معرفی انواع

مدل‌های خطی و GLM و مباحث مربوط به آن‌ها پرداختیم و به منظور بررسی و تحلیل این نوع داده‌های تکراری که همبستگی بین مشاهدات آن‌ها وجود دارد، روش شبه‌درستنمایی[1] بر مبنای GLM با عنوان GEEs در بخش‌های بعدی معرفی می‌شود که این روش توسط GLM برآوردهایی کارا و ناریب برای پارامترهای مدل ارائه می‌دهد.

۱.۴ معرفی داده‌های طولی

داده‌های طولی که در اقتصاد و جامعه‌شناسی با عنوان داده‌های پانلی[2] شناخته می‌شوند، به مجموعه داده‌هایی گفته می‌شود که در زمان‌های متوالی از واحدهای مشخص به‌دست می‌آیند. بنابراین این داده‌ها دارای دو بعد هستند: بعد مقطعی (واحدها) و بعد زمانی. در نتیجه داده‌های مقطعی و داده‌های سری زمانی حالت‌های خاصی از داده‌های طولی هستند.

۱.۱.۴ ساختار داده‌های طولی

مطالعات طولی، واحدهای مشخصی را در زمان‌های متوالی مورد بررسی قرار داده و در آن متغیر پاسخ و متغیرهای تبیینی اندازه‌گیری می‌شوند. چنین داده‌هایی را می‌توان به‌صورت جدول ۱.۴ نمایش داد: در این جدول y_{it} مقدار متغیر پاسخ مربوط به واحد iام در زمان tام

جدول ۱.۴: ساختار داده‌ها در مطالعات طولی

		واحد				
		۱	$\cdots$	i	$\cdots$	n
	۱	$(y_{۱۱}, x_{۱۱۱}, \cdots, x_{۱۱k})$	$\cdots$	$(y_{i۱}, x_{i۱۱}, \cdots, x_{i۱k})$	$\cdots$	$(y_{n۱}, \cdots, x_{۱n۱}, x_{۱nk})$
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
زمان	t	$(y_{۱t}, x_{۱t۱}, \cdots, x_{۱tk})$	$\cdots$	$(y_{it}, x_{it۱}, \cdots, x_{itk})$	$\cdots$	$(y_{nt}, \cdots, x_{nt۱}, x_{ntk})$
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
	T	$(y_{۱T}, x_{۱T۱}, \cdots, x_{۱Tk})$	$\cdots$	$(y_{iT}, x_{iT۱}, \cdots, x_{iTk})$	$\cdots$	$(y_{nT}, \cdots, x_{nT۱}, x_{nTk})$

و x_{itr} مقدار rامین متغیر تبیینی مربوط به زمان tام و واحد iام می‌باشد. T و n نیز به‌ترتیب

[1] Quasi-Likelihood
[2] Panel Data

معرف تعداد دوره‌های زمانی و تعداد واحدهایی هستند که بررسی شوند. جدول ۱.۴ مربوط به یک مطالعه طولی متعادل می‌باشد زیرا تمام واحدها به‌طور مساوی و در T دوره زمانی، مورد ارزیابی قرار گرفته‌اند. هم‌چنین مواردی وجود دارند که در آن واحد iام در T_i دوره زمانی اندازه‌گیری می‌شود. چنین مطالعه‌ای را نامتعادل گویند.
در جدول ۱.۴ سطرها (زمان) ماهیت مقطعی داده‌ها و ستون‌ها (واحدها) ماهیت زمانی داده‌ها را نشان می‌دهند. به‌عنوان مثال، اگر داده‌هایی را که در زمان tام جمع‌آوری شده‌اند در نظر بگیریم در واقع تمام n واحد را تنها در یک مقطع از زمان مورد بررسی قرار داده‌ایم. بنابراین برای هر دوره زمانی، داده‌ها مقطعی می‌باشند. اما اگر مثلا داده‌های مربوط به واحد iام را در نظر بگیریم، تنها یک واحد را در زمان‌های متوالی مورد بررسی قرار داده‌ایم؛ بنابراین داده‌ها سری زمانی می‌باشند.

۲.۱.۴ اهداف و ضرورت تحلیل داده‌های طولی

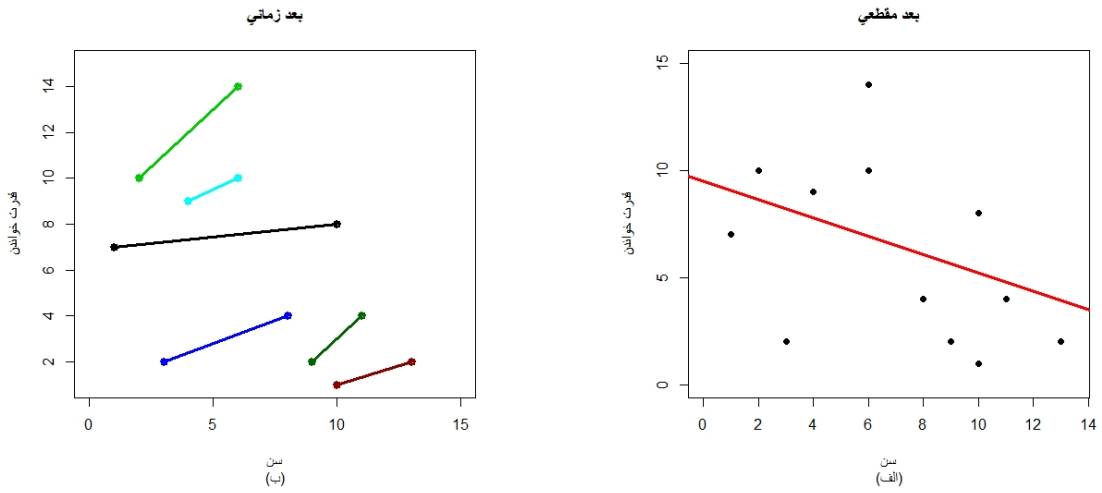
با جمع‌آوری مشاهدات واحدها در طول زمان، علاوه بر دستیابی به رابطه بین متغیرها، می‌توان تغییرات واحدها در طول زمان را نیز مورد مطالعه قرار داد. در نتیجه در تحلیل داده‌های طولی با سه هدف زیر روبرو هستیم:

۱. مطالعه تغییرات واحدها در طول زمان.

۲. مطالعه رابطه بین متغیرها (اثر متغیرهای پیشگو بر روی متغیر پاسخ).

۳. مطالعه تاثیر زمان بر رابطه بین متغیرها.

برای درک بهتر به مثال زیر از دیگل و همکاران (۲۰۰۲) اشاره می‌شود.
در شکل ۱.۴ داده‌ها از یک مطالعه طولی مربوط به قدرت خواندن کودکان به‌دست آمده‌اند که در آن شاخص توانایی مطالعه هر کودک، دو بار اندازه‌گیری شده است. بعد مقطعی داده‌ها در قسمت (آ)، کاهش قدرت خواندن کودکان با گذشت زمان را نشان می‌دهد؛ در حالی‌که در قسمت (ب)، بیان‌کننده افزایش توانایی هر کودک با گذشت زمان است. در مجموع می‌توان گفت در حالی‌که توانایی خواندن در کودکان جوان‌تر سطح بالاتری دارد؛ همه کودکان توانایی خود را با گذشت زمان بهبود می‌بخشند. بنابراین تحلیل داده‌های طولی که با هدف مشاهده تغییر رفتار پدیده‌ها در طول زمان انجام می‌گیرد، نسبت به داده‌های صرفا مقطعی یا صرفا زمانی از اهمیت ویژه‌ای برخوردارند.
علاوه بر حجیم بودن اطلاعات حاصل از داده‌های طولی و ماهیت دو بعدی آن‌ها که پیچیدگی داده‌ها را سبب می‌شوند، دلیل دیگری نیز وجود دارد که تحلیل این داده‌ها را با چالش رو برو می‌سازد. بر خلاف پذیره‌های روش‌های آماری متداول مبتنی بر پایه استقلال مشاهدات مربوط به هر واحد دارای همبستگی درونی بوده و از مشاهدات سایر واحدها مستقل می‌باشند. واضح است که جهت رسیدن به استنباط علمی، در نظر گرفتن این همبستگی امری ضروری



(آ) (ب)

شکل ۱.۴: نمودار قدرت خواندن کودکان

است.

دلایل بالا از جمله مهم‌ترین مواردی است که روش‌های آماری کلاسیک از قبیل رگرسیون و سری‌های زمانی به تنهایی قادر به تحلیل این داده‌ها نیستند.

۲.۴ روش معادلات برآوردیابی تعمیم‌یافته

روش معادلات برآوردیابی تعمیم‌یافته توسط لیانگ و زیگر (۱۹۸۶) به‌عنوان یك روش برآوردیابی مناسب در GLM با مشاهدات وابسته، معرفی شد. معادله برآوردیابی تعمیم‌یافته، تعمیمی از روش‌های برآوردیابی متداول در مدل‌های رگرسیونی کلاسیك بر اساس شبه‌درستنمایی است. این روش در تحلیل داده‌های طولی بسیار کاراتر از روش‌های متداول در مدل‌های رگرسیونی کلاسیك است.

در این دیدگاه همبستگی بین مشاهدات با در نظر گرفتن ماتریس‌های همبستگی مبنای[۳] مختلف مدل‌سازی می‌شود که تعیین درستی از این ساختارهای همبستگی در بهبود کارایی برآوردیابی ضرایب رگرسیونی موثر خواهد بود. برآورد ضرایب رگرسیونی و مولفه‌های واریانس در این روش با توجه به فرض ساختارهای متفاوت از ماتریس همبستگی مبنا، توسط امید ریاضی و واریانس توزیع فرضی جامعه به‌دست می‌آیند و لذا نیازی به تعیین توزیع احتمال حاشیه‌ای کامل مشاهدات یا تابع درستنمایی نیست.

[3] Working Correlation Matrix

۱.۲.۴ خانواده توزیع‌های نمایی

خانواده‌ای از توزیع‌ها را عضو خانواده توزیع‌های نمایی گوییم، اگر بتوان تابع چگالی احتمال آن‌را به‌صورت زیر نوشت:

$$f_Y(y) = \exp\left\{\frac{1}{\phi}[\theta y + c(\theta)] + d(\phi, y)\right\}$$

که در آن θ پارامتر طبیعی است که تابعی از امید ریاضی متغیر پاسخ است و ϕ پارامتر پراکندگی یا مقیاس است که برای تمامی متغیرهای پاسخ یکسان است و $c(\theta)$ و $d(\phi, y)$ توابع معلوم متناظر با نوع چگالی مربوط به خانواده توزیع‌های نمایی می‌باشند. برای بسیاری از اعضای این خانواده ϕ برابر با σ^2 یا ۱ است.
اگر متغیر تصادفی Y داری توزیعی از خانواده توزیع‌های نمایی باشد، امید ریاضی و واریانس آن به‌صورت زیر محاسبه می‌شوند:

$$\mathrm{E}(Y) = \mu = -c'(\theta) \qquad (۱.۴)$$

$$\mathrm{V}(Y) = \sigma^2 = -c''(\theta)\phi. \qquad (۲.۴)$$

از جمله توزیع‌های نمایی معروف می‌توان به توزیع نرمال، پواسون، دوجمله‌ای، نمایی، گاما، بتا و گوسی معکوس اشاره کرد. به‌عنوان مثال، توزیع پواسون را در نظر بگیرید:

$$f(y) = \frac{e^{-\mu}\mu^y}{y!}$$

که می‌توان آن‌را به شکل خانواده توزیع‌های نمایی نوشت:

$$\exp\{[y\log\mu - \mu] - \log y!\}$$

که در آن با توجه به روابط (۱.۴) و (۲.۴) داریم:

$$\phi = 1, \quad \theta = \log\mu, \quad c(\theta) = -e^{\theta}, \quad d(\phi, y) = -\log y!$$

$$\mathrm{E}(Y) = \mu = -c'(\theta) = e^{\theta}, \quad \mathrm{V}(Y) = \mu = -c''(\theta)\phi = e^{\theta}$$

۲.۲.۴ تابع شبه‌درستنمایی

هدف از برازش مدل‌های مختلف رگرسیون بررسی تأثیر متغیرهای تبیینی بر متغیر پاسخ است که در آن نتایج تحلیل داده‌ها انجام شود. استنباط در مورد پارامترهای مدل به محقق کمك می‌کند تا تشخیص دهد کدام متغیر تبیینی روی متغیر پاسخ اثر گذار می‌باشد، در حالی‌که سایر متغیرهای تبیینی که می‌توانستند اثر تداخلی داشته باشند، ثابت در نظر گرفته شده‌اند. در حقیقت اندازه پارامترهای برآوردشده مدل، توانایی و اهمیت اثر را مشخص می‌سازند. دو دیدگاه متمایز برای استنباط پارامترها وجود دارند.

۱ـ کمترین توان‌های دوم کلاسیك: این دیدگاه برآورد بهینه از نقطه نظر توابع برآورد است که توسط گودامب (۱۹۶۰) معرفی شده است.

۲ـ روشی مشابه با ML: دیدگاهی که توسط ودربرن (۱۹۷۴) معرفی شده و به‌عنوان پایه‌ای برای تحلیل GLMs است که روش شبه‌درستنمایی نامیده شد (هید ۱۹۹۷؛ هاردین و هیلب، ۲۰۰۲).

در برخی از تحلیل‌ها از جمله زمانی که مقادیر پاسخ وابسته‌اند و به خانواده توزیع‌های نمایی تعلق دارند، دسترسی به شکل صحیح و کاملی از تابع درستنمایی غیرممکن است. به این دلیل دیدگاه ودربرن (۱۹۷۴) می‌تواند به‌عنوان حالتی ویژه از دیدگاه تابع برآورد بهینه، در جایی که فضای تابع برآورد به رده‌ای خاص محدود شده، در نظر گرفته شود. وی پذیره‌های روی تابع احتمال را با تعریف تابعی تنها براساس رابطه بین میانگین و واریانس جامعه، جایگزین کرد که معادلات و ویژگی‌هایی مشابه با تابع لگاریتم درستنمایی دارند.
برای معرفی تابع شبه‌درستنمایی در یك مدل رگرسیونی تابع امتیاز $\mathrm{U} = \frac{\partial \ell}{\partial \mu}$ را به‌وسیله معادله مشتق در نظر بگیرید که با استفاده از قاعده زنجیره‌ای و روابط (۱.۴) و (۲.۴) به‌شکل زیر ساده می‌شود:

$$\frac{\partial \ell}{\partial \mu} = \frac{\partial \ell}{\partial \theta}\frac{\partial \theta}{\partial \mu} = \frac{y + c'(\theta)}{\phi}\frac{-1}{c''(\theta)} = \frac{y-\mu}{V(y)}.$$

یادآوری می‌کنیم که روابط $\mathrm{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = 0$ و $\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \theta_i^2} + \left(\frac{\partial \ell}{\partial \theta_i}\right)^2\right] = 0$ برای این تابع برقرار است. از این‌رو معادله $\mathrm{U} = 0$ مانند مشتق یك تابع امتیاز رفتار می‌کند و این یك تابع برآورد شبه‌امتیاز[4] نامیده می‌شود. از این رو شبه‌درستنمایی به کمك رابطه بین گشتاورهای مرتبه اول و دوم توزیع فرضی جامعه به برآورد پارامترها می‌پردازد که به‌صورت زیر تعریف شده‌است:

$$\mathrm{Q}(y;\mu) = \int_y^\mu \frac{y-\mu}{V(\mu)}\mathrm{d}\mu.$$

هم‌چنین واریانس پاسخ به‌عنوان تابعی از میانگین به‌صورت $V(Y_i) = \phi V(\mu_i)$ تعریف می‌شود و ϕ پارامتر پراکندگی توزیع مورد نظر است. وی هم‌چنین اشاره می‌کند که این معادلات برای هر انتخابی از تابع توزیع، حتی زمانی که انتگرال معادلات متناظر با عضوی از خانواده توزیع‌های نمایی نباشند یا حتی زمانی که توزیع پارامتر معلوم نباشد، حل خواهد شد. ضوابط تابع واریانس و شبه‌درستنمایی برای چند توزیع از خانواده توزیع‌های نمایی در جدول ۲.۴ فهرست شده‌اند.

[4] Quasi Score

جدول ۲.۴: ضوابط تابع واریانس و شبه‌درستنمایی برای چند توزیع از خانواده توزیع‌های نمایی

توزیع	$V(\mu)$	$Q(\mu)$
نرمال	۱	$-\frac{۱}{۲}\sum(y-\mu)^۲$
برنولی	$\mu(۱-\mu)$	$y\ln(\mu/(۱-\mu))+\ln(۱-\mu)$
پواسون	μ	$y\ln(\mu)-\mu$
گاما	$\mu^۲$	$-(y/\mu+\ln(\mu))$
گوسی معکوس	$\mu+\mu^۲$	$y(\ln(\mu)-۲\ln(\mu+۱))$

۳.۴ مدل حاشیه‌ای سیمپلکس برای تحلیل داده‌های نسبت طولی

حالت خاصی از روش GEEs که توسط لیانگ و زیگر (۱۹۸۶) و پرنتیس (۱۹۸۸) برای مشاهدات وابسته ارائه شده‌است؛ روش شبه‌درستنمایی ودربرن (۱۹۷۴) است. به‌عبارت دیگر، در این روش وابستگی بین مشاهدات نیز در تحلیل داده‌ها، توسط یك ساختار ماتریسی به نام ساختار همبستگی مبنا منظور می‌شود. این روش برآورد پارامترها را توسط گشتاورهای مرتبه اول و دوم توزیع فرضی جامعه و نه بر اساس توزیع احتمال حاشیه‌ای کامل مشاهدات یا روش درستنمایی کامل انجام می‌دهد که از ویژگی‌های آن در کاربرد برای GLM است و به‌طور مستقیم از گشتاور مرتبه اول جامعه برای برآورد ضرایب رگرسیونی و از گشتاور مرتبه دوم برای برآورد مولفه‌های واریانس و کوواریانس و در واقع برای برآورد پارامترهای همبستگی بین واحدها در یك گروه، بهره می‌گیرد و به این دلیل در برآوردیابی ضرایب رگرسیونی، کارایی بیشتری دارد.

۱.۳.۴ مدل حاشیه‌ای

فرض کنید y_{ij} مقادیر متغیر پاسخ برای $(j=۱,\cdots,n_i, i=۱,\cdots,m)$ و نشان‌دهنده jامین اندازه روی iامین واحد باشد و هم‌چنین $t_{ij}, j=۱,\cdots,n_i$ زمان‌های متوالی متناظر با هر

واحد است که در آن زمان‌ها اندازه‌گیری انجام شده است. علاوه بر این $x_{ijk}, k = ۱, \cdots, p$ متشکل از متغییرهای تبیینی است که با هر مقدار y_{ij} (متغیر پاسخ) متناظر است. حال فرض کنید y_{ij} که مصداق متغیرهای تصادفی Y_{ij} هستند از توزیع سیمپلکس ($Y_{ij} \sim S^{-}(\mu_{ij}, \sigma^{۲}_{ij})$) که تابع چگالی آن در رابطه (۱۱.۲) آمده است، پیروی کنند. در توزیع مذکور $\mu_{ij} \in (۰, ۱)$ پارامتر میانگین و $\sigma^{۲}_{ij} > ۰$ پارامتر پراکندگی هستند و هر دو پارامتر می‌توانند تابعی از متغیرها باشند. قرار می‌دهیم $\boldsymbol{Y}_i = (Y_{i۱}, \cdots, Y_{in_i})^T, \boldsymbol{x}_{ij} = (x_{ij۱}, \cdots, x_{ijp})^T$ که در آن $Y_۱, \cdots, Y_m$ از هم مستقل‌اند.

مشخصات کلی مدل حاشیه‌ای سیمپلکس به‌صورت زیر است:

۱ـ مولفه اول میانگین حاشیه‌ای یا احتمال پاسخ برای نمونه iام به‌صورت زیر که مدلی برای توصیف اثرات متوسط جامعه است، تعریف می‌شود

$$g(\mu_{ij}) = \boldsymbol{x}^T_{ij}\boldsymbol{\beta}$$

که در آن $\boldsymbol{\beta} = (\beta_۰, \cdots, \beta_{p-۱})^T$ بردار p بعدی از ضرایب رگرسیونی و $g : (۰, ۱) \longrightarrow (-\infty, \infty)$ تابع پیوند لجیت است.

۲ـ مولفه دوم، مدلی برای توصیف پارامتر پراکندگی است؛ با در نظر گرفتن $\sigma^{۲}_{ij} > ۰$ به‌عنوان پارامتر پراکندگی، پیشگوی خطی که تابعی از متغیرهای z_{ij} است (که ممکن است z_{ij} زیرمجموعه‌ای از $\boldsymbol{x}_{ij}$ها باشند) به‌صورت زیر تعریف می‌شود:

$$h(\sigma^{۲}_{ij}) = \boldsymbol{z}^T_{ij}\gamma$$

که در آن $\gamma = (\gamma_۰, \cdots, \gamma_{r-۱})^T$ که $\gamma_۰$ عرض از مبدا است و $h : (۰, \infty) \longrightarrow (-\infty, \infty)$ تابع پیوند لگاریتمی است. هم‌چنین $\sigma^{۲}_{ij}$ را می‌توان به‌صورت زیر نوشت:

$$\sigma^{۲}_{ij} = \exp(\boldsymbol{z}^T_{ij}\gamma) = \prod_{k=۰}^{r-۱}(e^{\gamma_k})^{z_{ijk}} = e^{\gamma_۰}\prod_{k=۱}^{r-۱}(e^{\gamma_k})^{z_{ijk}}.$$

جایی‌که z_{ij} داده‌های رسته‌ای[۵] هستند و ساختار جمعی[۶] می‌تواند مورد توجه قرار گیرد. برای مثال، در مثال داده‌های جراحی چشم در فصل ۵، مدل‌سازی پراکندگی متغیر در ۳ تیمار متفاوت (تاثیر سطح غلظت گاز در داده‌ها) به‌صورت زیر است

$$\sigma^{۲}_{ij} = \sigma_۰ + \sigma_۱ z_{ij۱} + \sigma_۲ z_{ij۲}$$

که در آن z_{ijk}، متغیرهای ظاهری مربوط به تیمار ۱ و ۲ هستند؛ به‌طوری‌که $\sigma_۰$ نشان‌دهنده پراکندگی تیمار ۳ است.

[5] Categorical Covariates
[6] Additive Structure

۳ـ مولفه سوم برای مدل‌سازی ساختار همبستگی است؛ همبستگی بین Y_{ij} و Y_{ik} که تابعی از میانگین‌های حاشیه‌ای و احتمالا پارامترهای فرعی $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_q)^T$ است، به‌صورت زیر تعریف می‌شود

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \boldsymbol{\alpha})$$

که در آن ρ تابعی معلوم و $\boldsymbol{\alpha}$ برداری از پارامترهای نامعلوم است. انواع مختلفی از ساختارهای همبستگی ممکن است برای تابع ρ مورد استفاده قرار گیرند که سه مورد رایج در تحلیل داده‌های طولی، تبادل‌پذیر[۷]، اتورگرسیو مرتبه اول (AR(1))[۸] و ساختار همبستگی وابسته مرتبه m[۹] هستند.

۲.۳.۴ روش GEEs برای برآورد پارامترها

با توجه به بردار میانگین برای هر i، $\boldsymbol{\mu_i} = (\mu_{i1}, \cdots, \mu_{in_i})^T$، بردار امتیاز به‌صورت زیر تعریف می‌شود

$$u_i = (u_{i1}, \cdots, u_{in_i})^T, \quad u_{ij} = -\frac{1}{2} d'(y_{ij}; \mu_{ij})$$

که تحت شرایط نظم، $\mathrm{E}(u_{ij}) = 0$ که می‌توان نتیجه گرفت $\mathrm{E}(\boldsymbol{u}_i) = 0$ (سانگ و تان، ۲۰۰۰)؛ حال واریانس u_{ij} طبق رابطه (۱۱.۱) برابراست با

$$\mathrm{var}(u_{ij}) = \frac{\sigma_{ij}^2}{2} \mathrm{E}\left\{d''(Y_{ij}; \mu_{ij})\right\} = \frac{3\sigma_{ij}^4}{\mu_{ij}(1-\mu_{ij})} + \frac{\sigma_{ij}^2}{\mathrm{v}(\mu_{ij})}.$$

ماتریس کوواریانس بردار پاسخ iام برای مقادیر $i = 1, \cdots, n$، در روش GEEs به‌صورت زیر تعریف می‌شود

$$\boldsymbol{V}_i = diag^{\frac{1}{2}}\{\mathrm{var}(s_{ij})\} \boldsymbol{R}(\boldsymbol{\alpha}) diag^{\frac{1}{2}}\{\mathrm{var}(s_{ij})\}$$

که در آن $diag^{\frac{1}{2}}\{\mathrm{var}(s_{ij})\}$ برای مقادیر $j = 1, \cdots, n_i$ یك ماتریس قطری $(n_i \times n_i)$ است. هم‌چنین، کمیت $\boldsymbol{R}(\boldsymbol{\alpha})$ یك ماتریس $(n_i \times n_i)$ است که ماتریس همبستگی مبنا نامیده می‌شود و به‌طور کامل با بردار پارامتر فرعی α که برداری $(q \times 1)$ است، مشخص می‌شود که این بردار را پارامتر همبستگی می‌نامند و می‌بایست در روند برازش مدل برآورد شود. بردار مبنا را به‌صورت $\boldsymbol{s_i} = diag\{\mathrm{v}(\mu_{ij})\}\boldsymbol{u}_i$ تعریف می‌کنیم.

روش GEEs برای برآورد بردار ضرایب رگرسیون $\boldsymbol{\beta}$، شامل یك سیستم غیرخطی از معادلات بر حسب $\boldsymbol{\beta}$ است که عموما به‌طور تکراری با الگوریتم نیوتن‌ـرافسون حل می‌شود؛ این معادلات به‌صورت زیر هستند

$$\Psi_1(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \boldsymbol{D}_i^T \boldsymbol{A}_i \boldsymbol{V}_i^{-1} \boldsymbol{s}_i = 0 \qquad (۳.۴)$$

[7] Exchangeable Correlation
[8] Autoregressive Correlation(AR)
[9] M-Dependent Correlation

که در آن $\boldsymbol{A}_i = diag\{\sigma_{ij}^{-2}\mathrm{v}(\mu_{ij})\mathrm{var}(u_{ij})\}$ و $\boldsymbol{D}_i^T = \dfrac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}}$ ماتریس $(p \times n_i)$ از مشتقات جزئی میانگین iامین واحد نسبت به ضرایب رگرسیون و $\boldsymbol{V}_i$ ماتریس کوواریانس بردار پاسخ Y_i است. در این روش تنها به پذیره‌هایی از گشتاورهای مرتبه اول نیاز است که در صورت برقراری آن‌ها، برآوردهای ضرایب رگرسیون سازگار خواهند بود. مزیت روش GEEs در آن است که تنها لازم است پارامترهای میانگین به‌طور صحیح مشخص شده باشند که در این صورت برآورد ضرایب رگرسیونی سازگارند و این موضوع حتی اگر ساختار ماتریس همبستگی مبنا به‌صورت صحیح تعیین نشده باشد، نیز برقرار است (پارك و شین، ۱۹۹۹).

چنان‌چه گفته شد، از ویژگی‌های روش GEEs مدل‌سازی همبستگی بین مشاهدات در یك واحد با ماتریس همبستگی مبنا است و انتخاب صحیح ساختار همبستگی در بهبود کارآمدی برآوردیابی ضرایب رگرسیونی موثر است. به این منظور لازم است ماتریس همبستگی صحیحی از بردار پاسخ iام داشته باشیم تا بتوانیم ماتریس کوواریانس درستی از بردار Y_i محاسبه کنیم. در واقع باید کمیت $\boldsymbol{R}_i(\alpha)$ صحیح انتخاب شود تا $\boldsymbol{V}_i$ مناسبی برای تحلیل به‌دست آید (هورتن و لیپسیتز، ۱۹۹۹ و هاردین و هیلب، ۲۰۰۳).

ماتریس همبستگی مبنا با بردار پارامتر فرعی α تعیین می‌شود که عموما نامعلوم است و لازم است در روند برازش تکراری به‌همراه ضرایب رگرسیونی $\boldsymbol{\beta}$ برآورد شود.

در برآورد بردار پارامتر همبستگی α برای مقادیر $i = 1, \cdots, m$ و $j = 1, \cdots, n_i$ یك معادله برآوردیابی به‌صورت زیر تعریف می‌شود

$$\Psi_3(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\xi}_i^T}{\partial \boldsymbol{\alpha}} \right) \boldsymbol{H}_i^{-1}(\boldsymbol{r}_i - \xi_i) = [\circ]_{q \times 1} \qquad (۴.۴)$$

که در آن $\xi_i = \mathrm{E}(r_i)_{q \times 1}$ و $\boldsymbol{H_i}$، ماتریس کواریانس که ماتریسی است قطری $(q \times q)$ از مقادیر واریانس و $q = \binom{n_i}{2}$ است که در آن $\boldsymbol{r}_i = (r_{i1}, r_{i2}, \cdots, r_{in_i-1}, r_{in_i})$ است که در آن r_{ij}، ijامین مانده پیرسون و $\binom{n_i}{2}$ است. این معادله برآوردیابی نتیجه می‌دهد که پارامتری کردن ماتریس همبستگی مبنا، از طریق معادله $\xi_i = \mathrm{E}(r_i)$ است و برای برآورد پارامترهای همبستگی α بایستی مقادیر $\hat{r}_{ij}$ به‌دست آمده از برآورد متداول $\hat{\boldsymbol{\beta}}$، جانشین مانده‌های پیرسون شوند.

دیدگاه‌های یکسانی در ارتباط با ساختارهای ماتریس متفاوت همبستگی مبنا وجود دارند. ممکن است مشاهدات درون یك واحد از مرتبه مشخصی پیروی نکنند یا به‌طور یکسان همبسته باشند یا فرض‌های پیچیده‌تری بر روی مشاهدات درون یك واحد در نظر گرفته شوند. در نتیجه در هر حالتی ساختاری متفاوت مد نظر محقق است و شاید یك بردار از پارامترهای اضافی تا حدود $(\max n_i) - 1$ پارامتر یا ماتریس کاملی از پارامترهای با $\binom{\max n_i}{2} - n_i$ پارامتر اضافی لازم باشد.

بر این اساس محققین انتخاب‌هایی را از ساختارهای ماتریس همبستگی مبنا با توجه به موضوع مورد بررسی و مدل‌های مد نظر فراهم کرده‌اند که در قسمت‌های بعدی تعدادی از

آن‌ها را همراه با فرمول‌هایی ساده برای برآورد مولفه‌های همبستگی α مربوط به آن‌ها، معرفی می‌کنیم. در تمام حالات فرمول‌ها به‌طور مستقیم از معادله برآوردیابی ارائه‌شده در عبارت (۴.۴) نتیجه می‌شود.

ساختار همبستگی اتورگرسیو مرتبه اول

این ساختار یك همبستگی زمانی را برای مشاهدات مختلف درون واحدها در نظر می‌گیرد. در واقع اگر همبستگی مشاهدات تکراری در درون واحدها، مرتبه‌ای طبیعی داشته باشند این ساختار به خوبی می‌تواند آن‌را نشان دهد. برای مثال، در مطالعات پزشکی، واحدها می‌توانند بیمارانی با اندازه‌های تکراری باشند که در ماه‌های متوالی مقدار ویژگی مشخصی از آن‌ها اندازه‌گیری شده‌است. در حالت کلی ساختار همبستگی اتورگرسیو مرتبه اول برای مقادیر $i = ۱, \cdots, n$ و زمان‌های $t = ۱, \cdots, n_i$ به‌صورت زیر تعریف می‌شود

$$\text{Corr}(Y_{it}, Y_{it'}) = \begin{cases} \alpha^{|t-t'|} & t \neq t' \\ ۱ & t = t' \end{cases} \qquad (۵.۴)$$

در این تعریف مقدار برآورد پارامتر همبستگی α توسط معادله برآوردیابی (۴.۴)، مانده‌های پیرسون $\hat{r}_{it}$ و رابطه (۱۰.۳) به‌صورت زیر است

$$\hat{\alpha} = \frac{۱}{\hat{\phi}} \sum_{i=۱}^{n} \left[\frac{\sum_{t=۱}^{n_t - ۰} \hat{r}_{it}\hat{r}_{i,t+۰}}{n_i}, \cdots, \frac{\sum_{t=۱}^{n_t - \ell} \hat{r}_{it}\hat{r}_{i,t+\ell}}{n_i} \right]$$

نکته قابل توجه آن است که یك فرآیند اتورگرسیو مرتبه ℓ، خود همبستگی غیر صفر، برای بیش از ℓ تاخیر دارد و به‌عنوان نمونه برای زمانی که چهار متغیر تبیینی مورد توجه محقق باشد، ساختار (۵.۴) به‌شکل ماتریسی زیر باز نویسی می‌شود

$$\text{AR}(1) = \begin{bmatrix} ۱ & \alpha & \alpha^۲ & \alpha^۳ \\ \alpha & ۱ & \alpha & \alpha^۲ \\ \alpha^۲ & \alpha & ۱ & \alpha \\ \alpha^۳ & \alpha^۲ & \alpha & ۱ \end{bmatrix}$$

ساختار همبستگی تبادل‌پذیر

ساده‌ترین شکل ساختار ماتریس همبستگی یك ماتریس همانی است. در تعمیم ساده‌ای از این ساختار همبستگی مستقل می‌توان فرض کرد که برخی مشاهدات در درون واحدها همبستگی مشترك داشته باشند و در حقیقت یك پارامتر فرعی به ماتریس اضافه می‌شود و مقدار (j, l)امین عنصر از این ماتریس همبستگی مبنا برای مقادیر $i = ۱, \cdots, n$ و $j = ۱, \cdots, n_i$ در حالت کلی

به شکل زیر تعریف می‌شود

$$\text{Corr}(Y_{ij}, Y_{il}) = \begin{cases} ۱ & j = l \\ \alpha & j \neq l \end{cases} \quad \Leftrightarrow \quad \boldsymbol{R}(\alpha) = \begin{bmatrix} ۱ & \alpha & \cdots & \alpha \\ \alpha & ۱ & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & ۱ \end{bmatrix}$$

این ساختار همبستگی برای مجموعه داده‌هایی که در آن‌ها اندازه‌های تکراری وابستگی زمانی ندارند و هم‌چنین برای هر جایگشتی از آن‌ها، معتبر است. در این حالت پارامتر فرعی α یك اسکالر است که به کمك معادله برآوردیابی (۴.۴) و توسط معادله زیر برآورد می‌شود

$$\hat{\alpha} = \frac{۱}{\hat{\phi}} \sum_{i=۱}^{n} \left[\frac{\sum_{j=۱}^{n_i} \sum_{l=۱}^{n_i} \hat{r}_{ij}\hat{r}_{il} - \sum_{j=۱}^{n_i} \hat{r}_{ij}^{۲}}{n_i(n_i - ۱)} \right]$$

که در آن از برآورد مانده‌های پیرسون $\hat{r}_{ij}$ و رابطه (۱۰.۳) برای برآورد پارامتر پراکندگی ϕ استفاده می‌شود. این ساختار تحت چندین نام دیگر از جمله ساختار همبستگی مساوی[10] و همبستگی مشترك[11] نیز در مباحث مختلف آماری شناخته می‌شود.

ساختار همبستگی وابسته مرتبه m

در مباحث سری زمانی، ممکن است همبستگی بین مشاهدات برای تعداد معینی از واحدهای زمانی و نه برای تمامی آن‌ها مد نظر محقق باشد. در این حالت ساختار ماتریس همبستگی برای مقادیر $i = ۱, \cdots, n$ و $j = ۱, \cdots, n_i$ به شکل زیر تعریف می‌شود

$$\text{Corr}(Y_{ij}, Y_{i,j+t}) = \begin{cases} ۱ & t = ۰ \\ \alpha_t & t = ۱, ۲, \cdots, m \\ ۰ & t > m \end{cases}$$

مشابه قبل برای مقادیر $t = ۱, \cdots, m$ می‌توان پارامترهای همبستگی $\hat{\alpha}_t$ را توسط مانده‌های پیرسون برآورد شده $\hat{r}_{ij}$، معادله برآوردیابی (۴.۴) و رابطه (۱۰.۳) به‌صورت زیر برآورد نمود

$$\hat{\alpha}_t = \frac{۱}{(N_t - p)\hat{\phi}} \sum_{i=۱}^{n} \sum_{j \leq n_i - t} \hat{r}_{ij}\hat{r}_{i,j+t}$$

که در آن p بعد بردار ضرایب رگرسیونی و $N_t = \sum_{i=۱}^{n}(n_i - t)$ است.
برای مثال، اگر در مدل‌سازی ماتریسی با بعد ۳ مورد استفاده قرار گیرد، آن‌گاه آن ماتریس

[10] Equal Correlation
[11] Common Correlation

ساختار زیر را دارد

$$\boldsymbol{R}(\alpha)=\begin{bmatrix} ۱ & \alpha_۱ & \alpha_۲ & \alpha_۳ \\ \alpha_۱ & ۱ & \alpha_۱ & \alpha_۲ \\ \alpha_۲ & \alpha_۱ & ۱ & \alpha_۱ \\ \alpha_۳ & \alpha_۲ & \alpha_۱ & ۱ \end{bmatrix}$$

۳.۳.۴ روش معادلات برآوردیابی تعمیم‌یافته مرتبه اول (GEE1) و مرتبه دوم (GEE2)

در مدل‌سازی به روش GEEs اگر مقدار ماتریس همبستگی مبنا انتخاب‌شده در تحلیل داده‌ها، بسیار متفاوت از ماتریس همبستگی حقیقی باشد ممکن است این روش در برآورد پارامترهای مدل ناکارآمد باشد. به این منظور محققان مدل‌سازی ساختاری دیگر را توصیه کرده‌اند که در آن یك ساختار همبستگی توسط گشتاورهای مرتبه بالاتر نیز در تحلیل اعمال می‌شود. این تغییرپذیری شامل در نظر گرفتن یك معادله برآوردیابی اضافی برای پارامتر همبستگی α است که آن‌را معادله برآوردیابی تعمیم‌یافته مرتبه دوم (GEE2) می‌نامند و شکل ساده‌تر آن، معادله برآوردیابی تعمیم‌یافته مرتبه اول (GEE1) است. در واقع این دو روش تعمیم‌هایی از روش GEEs هستند (بالمی و لی، ۲۰۰۹ ؛ زورن، ۲۰۰۱).

روش GEE1 تقریبا به‌همان کارایی روش GEE2 برای برآورد بردار ضرایب رگرسیونی $\boldsymbol{\beta}$ است و تنها زمانی که بهبود کارایی قابل توجه‌ای مد نظر محقق باشد، استفاده از روش GEE2 توصیه می‌شود. در حالت کلی مجموعه معادلات ساده GEE1 ترجیح داده می‌شوند و در تحلیل داده‌های وابسته کاربرد بیشتری دارند.

با توجه به پرنتیس و ژائو (۱۹۹۱)، در روش GEE2، با اضافه کردن یک مجموعه از معادلات برآورد شده برای پارامترهای همبستگی براساس باقی‌مانده‌های امتیاز استاندارد شده[۱۲]، به‌صورت زیر تعریف می‌شود:

$$r_{ij}=\frac{u_{ij}}{\sqrt{\operatorname{var}(u_{ij})}}=\frac{u_{ij}}{\sqrt{\frac{\sigma_{ij}^{۲}}{۲}\mathrm{E}\left\{d''(Y_{ij};\mu_{ij})\right\}}}. \qquad (۶.۴)$$

به‌راحتی می‌توان دید که خواص گشتاوری $\mathrm{E}(r_{ij})=۰$، $\operatorname{var}(r_{ij})=۱$ و $\mathrm{E}(r_{ij}r_{ij'})=\operatorname{Corr}(u_{ij},u_{ij'})=\operatorname{Corr}(s_{ij},s_{ij'})$ برای باقی‌مانده‌های امتیاز برقرار است.

حال برای برآورد ضرایب رگرسیون γ، معادله برآوردیابی به‌صورت زیر تعریف می‌شود

$$\Psi_۲(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})=\sum_{i=۱}^{m}\left(\frac{\partial\boldsymbol{\sigma}_i^T}{\partial\gamma}\right)\boldsymbol{\Sigma}_i^{-۱}(\boldsymbol{d}_i-\boldsymbol{\sigma}_i)=۰ \qquad (۷.۴)$$

[12] Standardised Score Residuals

که در آن $\boldsymbol{d}_i = (d(y_{i1};\mu_{i1}),\cdots,d(y_{in_i};\mu_{in_i}))^T$، $\boldsymbol{\Sigma}_i$ ماتریس کواریانس و $\boldsymbol{\sigma}_i = E(\boldsymbol{d}_i) = (\sigma^2_{i1},\cdots,\sigma^2_{in_i})^T$.

با تعمیم روش GEEs توسط بسط معادلات برآوردیابی برای پارامترهای $\boldsymbol{\theta} = (\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})$، معادلات GEE2 به‌صورت زیر حاصل می‌شوند

$$\Upsilon(\boldsymbol{\theta}) = \Upsilon(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}) = \begin{bmatrix}\Psi_1(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})\\ \Psi_2(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})\\ \Psi_3(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})\end{bmatrix} = 0. \qquad (۸.۴)$$

معادله برآوردیابی $\Upsilon(\boldsymbol{\theta}) = 0$ ناریب است، زیرا $\mathrm{E}(\Upsilon(\boldsymbol{\theta})) = 0$. از این رو طبق نظریه استاندارد[13] معادلات برآوردیابی تحت شرایط نظم، برآوردگر $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}},\hat{\gamma},\hat{\boldsymbol{\alpha}})$، برآوردگری سازگار است و $m^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ به‌طور مجانبی گاوسی چند متغیره با میانگین صفر و ماتریس کواریانس $m\boldsymbol{J}^{-1}(\boldsymbol{\theta})$ است که $\boldsymbol{J}(\boldsymbol{\theta})$ ماتریس اطلاع گودامب[14] به‌صورت $\boldsymbol{J}(\boldsymbol{\theta}) = \boldsymbol{S}^T\boldsymbol{R}^{-1}\boldsymbol{S}$ است؛ که در آن ماتریس حساسیت $\boldsymbol{S} = \mathrm{E}\{\partial\Upsilon(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\}$ و ماتریس تغییرپذیری برابر با $\boldsymbol{R} = \mathrm{E}\{\Upsilon(\boldsymbol{\theta})\Upsilon^T(\boldsymbol{\theta})\}$ است. برای مشاهده جزئیات بیشتر به سانگ و همکاران (۲۰۰۴) رجوع شود (ضمیمه را آ.۲ ببینید). با استفاده از الگوریتم نیوتن‌ـرافسون، حل $\hat{\boldsymbol{\theta}}$ برای معادله (۸.۴) را می‌توان به‌صورت زیر به‌دست آورد

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \boldsymbol{S}^{-1}\Upsilon(\boldsymbol{\theta}^{(k)}).$$

۴.۳.۴ انواع معادلات برآوردیابی تعمیم‌یافته

برخی محققان دو رده‌بندی از مدل‌ها را برای ساختار واحدهای درون داده‌ها پیشنهاد کرده‌اند که در حقیقت این تقسیم‌بندی بر مبنای نحوه محاسبه میانگین و به کارگیری اثر تصادفی در مدل است که در این بخش به معرفی آن‌ها به‌عنوان دو نوع گوناگون از مجموعه روش‌های GEEs می‌پردازیم (هاردین و هیلب، ۲۰۰۳).

مدل متوسط جامعه[15]

مدلی که مقدار همبستگی درون واحدها را به واسطه اثرات متوسط روی تمام واحدها در نظر می‌گیرد و به‌عنوان یك مدل حاشیه‌ای مطرح است را مدل متوسط جمعیت می‌نامند. در حقیقت این مدل یك مشخصه‌سازی را برای ساختار کوواریانس سطوح واحدها فرض می‌کند و آن‌گاه مقدار کوواریانس یا همبستگی سطوح را با میانگین گرفتن در میان اطلاعات تمام واحدها برآورد می‌کند. مدل‌های انتقالی[16] و پاسخ‌های شرطی از نوع مدل‌های متوسط جامعه

[13] Standard Theory
[14] Godambe
[15] Population Averaged
[16] Transitional

هستند و در تحلیل مطالعات طولی زمانی که بایستی مقدار همبستگی متداول مقادیر پاسخ روی پاسخ‌های قبلی نشان داده شود، استفاده می‌شوند.

مدل واحد ویژه[۱۷]

این مدل همبستگی داخل واحدها را توسط مطرح کردن مولفه‌های تصادفی سطوح آن‌ها بیان می‌کند و به واسطه معرفی اثر یك واحد، در حالی‌که به یك کوواریانس سطح اشاره می‌کند، به‌دست می‌آید. در این مدل هر اثر با استفاده از اطلاعات تنها یك واحد خاص، برآورد می‌شود. مدل‌ها با اثرات ثابت و تصادفی از این نوع مدل هستند.

۵.۳.۴ تحلیل باقی‌مانده‌ها

پیشنهاد می‌شود از دو نوع باقی‌مانده استفاده شود که تشخیص برای پذیره‌های مدل کلیدی، یعنی

۱) توزیع‌های حاشیه‌ای

۲) توابع پیوند

۳) ساختار همبستگی

ایجاد شود. اولی، باقی‌مانده‌های امتیاز استانداردشده r_{ij} در (۶.۴) است و دیگری باقی‌مانده‌های معمولی استاندارد شده پیرسون $e_{ij} = (y_{ij} - \mu_{ij})/\sqrt{\text{var}(Y_{ij})}$ است. از آن‌جایی که $\text{var}(Y_{ij})$ شامل تابع گامای ناکامل است، صورت بسته‌ای ندارد. هماورد نمونه از r_{ij} یا e_{ij} با جایگزین کردن پارامترها با تخمین‌های متناظر به‌دست می‌آید که آن‌ها را با $\hat{r}_{ij}$ یا $\hat{e}_{ij}$ نشان می‌دهند. تحلیل باقی‌مانده‌ها برای شناسایی سیگنال‌های قوی مربوط به نقض فرض مدل خاص در زیر آورده شده است.

پذیره توزیع سیمپلکس می‌تواند با نمودار $\hat{e}_{ij}$ در مقابل $\hat{m}_{ij}$ که هدف آن بررسی رابطه واریانس‌ـ میانگین است، کنترل شود. اگر این پذیره درست باشد، "$\text{var}(e_{ij}) = ۱$ مستقل از میانگین μ_{ij} است. بنابراین، نقاط موجود در نمودار باید به‌طور تصادفی در اطراف خط افقی در صفر (امید باقی‌مانده‌ها) با تقریب ٪۹۵ در نوار افقی بین ۲ و ۲− پخش شوند. هر گونه واکنش ظاهری از این امر تخطی از پذیره توزیع یا احتمالا برازش یک مدل ضعیف را نشان می‌دهد. تحقیقات بیشتری برای شناسایی عامل چنین خروجی مورد نیاز است. از آن‌جایی که $\sigma^۲$ بزرگ است این روش قابل اعتماد است زیرا رابطه واریانس میانگین، تحت کنترل $\mu(۱ - \mu)$ یک حالت مشابه با توزیع دوجمله‌ای دارد.

[17] Subject Specific

با توجه به مک‌کالا و نلدر(۱۹۸۹)، نمودار متغیر وابسته تعدیل‌یافته a_{ij} در مقابل پیشگوی خطی $\hat{\eta}_{ij}$ برای بررسی تابع پیوند انتخاب‌شده، استفاده می‌کنیم. حال تعریف می‌کنیم:

$$a_i = g(\mu_i) + \left\{ \frac{\mathfrak{۳}\sigma_{ij}^{\mathfrak{۴}}}{\mu_{ij}(\mathfrak{۱} - \mu_{ij})} + \frac{\sigma_{ij}^{\mathfrak{۲}}}{\mathrm{v}(\mu_{ij})} \right\}^{\frac{-\mathfrak{۱}}{\mathfrak{۲}}} u(y_i; \mu_i) \qquad (۹.۴)$$

واضح است که $\mathrm{E}(a_i) = g(\mu_{ij})$ زیرا $\mathrm{E}(u_{ij}) = \mathfrak{۰}$ و $\mathrm{var}(a_i) = \mathrm{E}\left\{a_i - g(\mu_{ij})^{\mathfrak{۲}}\right\} = \mathfrak{۱}$ است. اگر تابع پیوند مناسب باشد، نمودار برآورد $\hat{a}_i$ در مقابل $\hat{\boldsymbol{\eta}}_{ij} = \boldsymbol{x}_{ij}^T\hat{\boldsymbol{\beta}}$ باید یک خط مستقیم را که با تقریب ٪۹۵ نقاط در یک باند با محدودیت‌های بالا و پایین حدود $\hat{\boldsymbol{\eta}}_{ij} \pm \mathfrak{۲}$ قرار می‌گیرند، نشان دهد. این نمودار مانند مدل‌های خطی تعمیم‌یافته بهترین تابع پیوند را برای مدل پیشنهاد نمی‌دهد بلکه فقط به بررسی غیر رسمی برای هر گونه نقض شدید پیوند مورد استفاده، می‌پردازد.

اگر چه مدل‌سازی حقیقی ساختار همبستگی داده‌های طولی دشوار است اما ساختارهای همبستگی صحیح به ضرایب رگرسیون، این امکان را می‌دهند که به‌طور موثر تخمین زده شوند. بنابراین، ارزیابی تناسب همبستگی استفاده‌شده در GEEs از طریق آنالیز باقی‌مانده‌ها مهم است. توجه داشته باشید که

$$\mathrm{Corr}(r_{ij}, r_{ij'}) = \mathrm{Corr}(a_i, a_{i'})$$

این امر حاکی از آن است که همبستگی واقعی متغیر a_i برابر با باقی‌مانده امتیاز استانداردشده r_{ij} است. برای ارزیابی همبستگی داده‌ها، می‌توان برخی از روش‌های اکتشافی ارائه شده در بخش ۳.۴ دیگل و همکاران (۲۰۰۲) را به‌کار برد.

فصل ۵

کاربرد مدل رگرسیونی سیمپلکس و برازش آن با R

در این فصل، نحوه استخراج استنباط‌های مطرح‌شده در فصل‌های قبلی را برای مدل‌های رگرسیونی سیمپلکس با استفاده از بسته simplexreg شرح می‌دهیم. سپس کاربرد مدل را با دو مثال واقعی شامل داده‌های سلول بنیادی خون محیطی و داده‌های جراحی چشم نشان می‌دهیم.

۱.۵ مقدمه

تابع برازش مدل رگرسیونی سیمپلکس، $simplexreg()$، طوری طراحی شده‌است که تا حد امکان مشابه تابع $glm()$ استاندارد (تیم مرکزی توسعه R[1]، ۲۰۰۹) که برای برازش GLM به‌کار می‌رود، باشد. یک تفاوت اساسی بین این دو تابع آن است که در $simplexreg()$ می‌توان معادله‌هایی که معرف میانگین و پارامتر پراکندگی هستند را در این تابع لحاظ کرد و در نتیجه دو ماتریس متفاوت از متغیرهای تبیینی، دو پیشگوی خطی و دو مجموعه از ضرایب به‌دست

[1] R Development Core Team

آورد. روش استفاده از $simplexreg()$ در R به‌صورت زیر است:

```
simplexreg(formula, data, subset, na.action,
link = c("logit", "probit", "cloglog", "neglog"), corr = "Ind", id = NULL,
control = simplexreg.control(...), model = TRUE, y = TRUE, x = FALSE, ...)
```

چهار آرگومان اول، شامل مشخصات مدل استاندارد (چمبرز و هستی، ۱۹۹۲) هستند. از $formula$ برای مشخص کردن مدل رگرسیون و پیشگوهای خطی، مشابه تابع GLM، استفاده می‌شود. داده‌ها توسط $data$ مشخص می‌شوند. آرگومان $link$ معرف نوع تابع پیوند و $corr$ ساختار همبستگی را تعیین می‌کند. آرگومان $model$ معرف یک مقدار منطقی است و تعیین می‌کند که از مقادیر خروجی تابع $simplexreg.fit$ مختصات مدل ظاهر شود. در بسته simplexreg و بسته‌های مرتبط با آن می‌توان به توابع زیر که برای استخراج استنباط‌ها در مدل رگرسیونی سیمپلکس مفید هستند، اشاره کرد:

- $summary()$ خلاصه‌ای از مدل رگرسیونی برازش داده‌شده را ارائه می‌دهد.
- $coef()$ برای استخراج ضرایب برآوردشده مدل، مورد استفاده قرار می‌گیرد.
- $vcov()$ ماتریس کواریانس برآوردگرها را ارائه می‌دهد.
- $predict()$ برای پیش‌گویی داده‌های جدید می‌توان از آن بهره برد.
- $fitted()$ برازش میانگین‌ها برای داده‌های مشاهده‌شده را انجام می‌دهد.
- $residuals()$ برای استخراج باقی‌مانده‌ها مورد استفاده قرار می‌گیرد.
- $estfun()$ تخمین مشاهدات و برآورد پارامترها را نمایش می‌دهد.
- استخراج مولفه‌های مدل، مولفه‌های ماتریس مدل و مولفه‌های اصلی مدل به ترتیب با توابع $terms()$، $model.matrix()$ و $model.frame()$ صورت می‌گیرد.
- $loglik()$ لگاریتم درستنمایی برازش‌شده را ارائه می‌دهد.
- $plot()$ برای رسم نمودارهای مناسب مدل می‌توان از آن بهره برد.

تابع $simplexreg.fit$ یک ابزار دیگری برای محاسبات مدل است که به‌صورت زیر است:

```
simplexreg.fit(y, x, z = NULL, t = NULL, link = "logit", corr = "Ind",
id = NULL, control = simplexreg.control())
```

آرگومان $corr$ ساختار همبستگی را در مدل حاشیه‌ای مشخص می‌کند و شامل سه گزینه مستقل، تبادل‌پذیر و اتورگرسیو مرتبه اول است که آن‌ها را به‌ترتیب با "Ind"، "Exc" و "AR(1)" نشان می‌دهند. به‌طور پیش‌فرض، آرگومان $corr$ مقدار "Ind" را می‌گیرد که در این‌صورت مدل حاشیه‌ای معادل مدل خطی تعمیم‌یافته، SGLM، است. در تابع $simplexreg()$ اگر به‌عنوان مثال $formula$ به‌صورت $y \sim x۱ + x۲$ باشد، آن‌گاه متغیرهای تبیینی $x_۱$ و $x_۲$ تنها برای مدل‌بندی میانگین رگرسیون سیمپلکس در نظر گرفته می‌شوند. در این حالت $\sigma^۲$ ثابت فرض می‌شود. به‌عبارت دیگر، $z_i = ۱$ و $h(.)$ تابع پیوند همانی است که همان مدل معرفی‌شده در فصل ۳ بخش ۱.۳ است. اما چنان‌چه $formula$ شامل دو بخش به‌صورت $y \sim x۱+x۲ \mid z۱+z۲$ باشد، در این حالت $\sigma^۲$ نیز تابعی از متغیرهای تبیینی $z_۱$ و $z_۲$ است که مدل معرفی‌شده در فصل ۳ بخش ۲.۳ می‌باشد. برای داده‌های نسبت طولی، مدل‌های سیمپلکس حاشیه‌ای از سه جز، اثرات میانگین جامعه، الگوی پراکنش[۲] و ساختار همبستگی تشکیل می‌شود. حال اگر $formula$ به‌صورت سه قسمتی $y \sim x۱ + x۲|۱|t$ در نظر گرفته شود، معرف مدل با پارامتر پراکندگی ثابت است و اگر به‌صورت $y \sim x۱ + x۲|z۱ + z۲|t$ باشد برای مدل‌های حاشیه‌ای سیمپلکس با پارامتر پراکندگی متغیر (corr = "Exc" یا "AR(1)") به‌کار می‌رود. پارامتر t در تابع $simplexreg.fit$ و در $formula$ مربوط به متغیر کمکی زمان است.

تابع پیوند پیش‌فرض برای میانگین $g(\mu) = \text{logit}$ و برای پارامتر پراکندگی $h(\sigma^۲) = \log$ است. تابع پیوند برای میانگین μ را می‌توان با انتخاب $link$، تغییر داد.
توابع مجاز برای μ همان توابع پیوند مجاز در $glm()$ شامل "$logit$","$probit$","$cloglog$","$neglog$" و برای $\sigma^۲$، توابع همانی، لگاریتمی و ریشه دوم هستند.

```
plot(x, type = c("residuals", "corr", "GOF"), res = "adjvar", lag = 1, ...)
```

آرگومان $type$ نوع نمودار را مشخص می‌کند. اگر $type =$ "residuals" تحلیل باقی‌مانده‌ها صورت می‌گیرد که چهار انتخاب برای آن وجود دارد

۱. اگر res : stdPerr آن‌گاه باقی‌مانده استانداردشده پیرسون محاسبه می‌شود که در رابطه (۱۲.۳) آمده است.

۲. اگر res : appstdPerr آن‌گاه تقریبی از باقی‌مانده پیرسون محاسبه می‌شود که در رابطه (۱۳.۳) آمده است.

۳. اگر res : stdscor آن‌گاه باقی‌مانده استانداردشده امتیاز محاسبه می‌شود که در رابطه (۶.۴) آمده است.

۴. اگر res : adjvar آن‌گاه متغیر وابسته تعدیل‌یافته، محاسبه می‌شود که در رابطه (۹.۴) آمده است.

[2] Pattern of Dispersion

سه مورد اول به منظور بررسی رابطه میانگین واریانس و نیز تشخیص خطای مدل، در مقابل میانگین μ رسم می‌شوند. در حالی‌که با استفاده از مورد چهارم، متغیر وابسته تعدیل‌یافته در مقابل پیشگوی خطی η برای بررسی تابع پیوند رسم می‌شود.
آرگومان $type = \text{"corr"}$ ساختار همبستگی را نشان می‌دهد. در lag ، برای بررسی خودهمبستگی از باقی‌مانده‌های استانداردشده امتیاز استفاده می‌شود و هرگاه آرگومان $type = \text{"GOF"}$ باشد انحراف جزئی در مقابل متغیرهای زمان رسم می‌شود.

۲.۵ کاربرد مدل در دو مثال واقعی

این بخش شامل مثال‌های کاربردی است که توسط آن‌ها مباحث نظری بیان‌شده در فصل‌های سوم و چهارم را به‌صورت کاربردی تشریح می‌کنیم.

۱.۲.۵ داده‌های سلول بنیادی خون محیطی

پیوندهای سلول بنیادی خون محیطی[۳] (PBSC) به‌طور گسترده‌ای برای بهبود سریع هماتولوژیک پس از میلوبلیتیو تراپی[۴] برای انواع اختلالات خونی بدخیم استفاده می‌شود. بازسازی مجدد هماتوپوئتیک[۵]، به‌طور گسترده‌ای به تزریق تعداد کافی سلول‌های بنیادی برای پیوند در مغز استخوان محیط میکرو[۶] بستگی دارد. همان‌طور که در مطالعه آلان و همکاران (۲۰۰۲) نشان داده شد، مقدار سلول‌های زنده CD34+ به‌عنوان یک نشانه مهم برداشت کافی PBSC و هم‌چنین به‌عنوان یک پیش‌بینی کننده پیوند هماتوپوئتیک در نظر گرفته می‌شود. در این مطالعه ۲۴۲ بیمار موافقت کردند که پیوند اتولوگ با PBSC را پس از دوزهای شیمی‌درمانی در سال‌های ۲۰۰۳ و ۲۰۰۸ در آزمایشگاه سلول‌های بنیادی هماتوپوئید در مؤسسه سرطان ـ خدمات درمانی آلبرتا، دریافت و اطلاعات را ثبت کنند. این داده‌ها شامل اطلاعاتی در مورد سن، جنس و ویژگی‌های بالینی بیماران است. داده‌های مربوط به پیوند PBSC، با نام sdac ، در بسته simplexreg گنجانده شده‌است. این مجموعه داده شامل ۲۳۹ مشاهده (بیمار) و ۵ متغیر تبیینی به شرح زیر است:

[3] Peripheral Blood Stem Cell
[4] Myeloablative Therapy
[5] Hematopoietic
[6] Micro-Environment

- age: سن بیماران
- gender: جنسیت بیماران که اکثر آن‌ها یعنی ۱۷۱ نفر، مرد هستند
- rcd: میزان سلول‌های زنده CD34+
- chemo: متغیر ظاهری؛ chemo = ۰ معرف پروتکل یک روزه و chemo = ۱ معرف پروتکل سه روزه است
- ageadj: متغیر سن تعدیل یافته؛ age < ۴۰ به‌عنوان سن پایه در نظر گرفته و سایر سنین از عدد ۴۰ کم می‌شوند

بخشی از داده‌های sdac موجود در بسته simplexreg به شرح زیر است.

	age	gender	rcd	ageadj	chemo
1	62	M	0.75	22	0
2	39	M	0.83	0	1
3	43	M	0.94	3	1
4	58	M	0.86	18	0
5	43	M	0.54	3	0
6	51	M	0.7	11	1
7	64	F	0.59	24	0
8	64	M	0.82	24	0
9	48	M	0.67	8	1
10	51	M	0.57	11	0
11	57	M	0.75	17	0
12	60	M	0.74	20	1
13	53	M	0.75	13	1
14	47	M	0.94	7	0
15	51	M	0.87	11	1
16	29	M	0.99	0	1
17	43	M	0.83	3	1
18	57	M	0.82	17	1
19	62	F	0.81	22	0

پارامتر پراکندگی ثابت

ابتدا مدل را با فرض ثابت بودن پارامتر پراکندگی σ^2 در نظر گرفتیم. پیشگوی خطی برای μ_i به‌صورت

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \mathbf{ageadj} + \beta_2 \mathbf{chemo} \qquad (۱.۵)$$

است. با دستورات زیر می‌توان مدل (۱.۵) را بر روی داده‌ها برازش داد

```
library(simplexreg)
data("sdac", package="simplexreg")
sim.glm1 <- simplexreg(rcd ~ ageadj + chemo, data = sdac)
summary(sim.glm1)
```

نتایج برازش مدل (۱.۵) در جدول ۲.۵ گزارش شده‌اند. نتایج مذکور با ۶ تکرار در مرحله بهینه‌سازی و با انحراف ۲۳۵ به‌دست آمده که لگاریتم درستنمایی آن ۱۰۵/۳ و pـ مقدار مربوطه برابر ۰/۴۹ است. همان‌طور که ملاحظه می‌شود عرض از مبدا و ضرایب متغیرهای **ageadj** و **chemo** معنی‌دار هستند.

جدول ۱.۵: باقی‌مانده استاندارد پیرسون مدل (۱.۵)

بزرگترین مقدار	چارک سوم	میانه	چارک اول	کوچکترین مقدار
۱/۳۹۶	۰/۴۹۷	−۰/۰۰۸	−۰/۵۸۵	−۲/۸۲۵

جدول ۲.۵: برآورد پارامترهای مدل (۱.۵)

پارامترها	برآورد	Std.Err	z-value	Pr(>|z|)
عرض از مبدا	۱/۱۰۰۲	۰/۱۴۱	۷/۸۲۱	$۵/۲۶e^{-۱۵}$***
ageadj	۰/۰۱۳	۰/۰۰۶	۲/۰۸۲	۰/۰۳۷۳*
chemo	۰/۲۶۶	۰/۱۲۴	۲/۱۲۹	۰/۰۳۳۳*

'***' سطح ۰، '**' سطح ۰/۰۰۱، '*' سطح ۰/۰۵ و '.' سطح ۰/۱ را نشان می‌دهد.

پارامتر پراکندگی متغیر

در این قسمت، مدل را برای حالتی که $\sigma^۲$ تابعی از متغیرهای تبیینی باشد نیز برازش دادیم که در آن پیشگوهای خطی به‌صورت زیر هستند:

$$logit(\mu_i) = \beta_۰ + \beta_۱\mathbf{ageadj} + \beta_۲\mathbf{chemo} \qquad \log(\sigma_i^۲) = \gamma_۰ + \gamma_۱\mathbf{age}. \qquad (۲.۵)$$

کد برازش این مدل در R به‌صورت زیر است

```
data("sdac", package="simplexreg")
sim.glm2 <- simplexreg(rcd~ageadj+chemo|age, link = "logit", data = sdac)
summary(sim.glm2)
```

نتایج برازش پارامترهای مدل (۲.۵) با ۶ تکرار و با انحراف ۲۳۹ در جدول ۴.۵ گزارش شده است. لگاریتم درستنمایی برابر ۹۹/۵۱ و pـ مقدار مربوطه معادل ۰/۴۳ است.

pـ مقدار آزمون کای‌دو برای هر دو مدل به‌ترتیب ۰/۴۹ و ۰/۴۳ است و این به معنای نقص برازش نیست.

جدول ۳.۵: باقی‌مانده استاندارد پیرسون مدل (۲.۵)

بزرگترین مقدار	چارک سوم	میانه	چارک اول	کوچکترین مقدار
۱/۴۴۸	۰/۴۹۶	−۰/۰۰۳	−۰/۵۳۹	−۳/۰۸۲

جدول ۴.۵: برآورد پارامترهای مدل (۲.۵)

پارامترها	برآورد	Std.Err	z-value	Pr(>|z|)
عرض از مبدا	۱/۱۱۵	۰/۱۴۱	۷/۸۹	$۳/۰۳e^{-۱۵}$***
ageadj	۰/۰۱۳	۰/۰۰۶۴	۲/۰۱۷	۰/۰۴۳۷*
chemo	۰/۲۵۱	۰/۱۲۱	۲/۰۶۸	۰/۰۳۸۶*
عرض از مبدا	۲/۶۰۷	۰/۳۶۶	۷/۱۰۷	$۱/۱۸e^{-۱۲}$***
age	−۰/۰۱۵	۰/۰۰۶	−۲/۱۸۱	۰/۰۲۹*

'***' سطح ۰، '**' سطح ۰/۰۰۱، '*' سطح ۰/۰۵ و '.' سطح ۰/۱ را نشان می‌دهد.

جدول ۵.۵ و ۶.۵، ملاک‌های AIC و BIC برای دو مدل با $\sigma^۲$ ثابت و متغیر را نشان می‌دهند. با توجه به این دو مقدار، مدل (۱.۵)، یعنی مدل با پارامتر پراکندگی ثابت بهتر عمل کرده است.

جدول ۵.۵: مقایسه AIC برای دو مدل

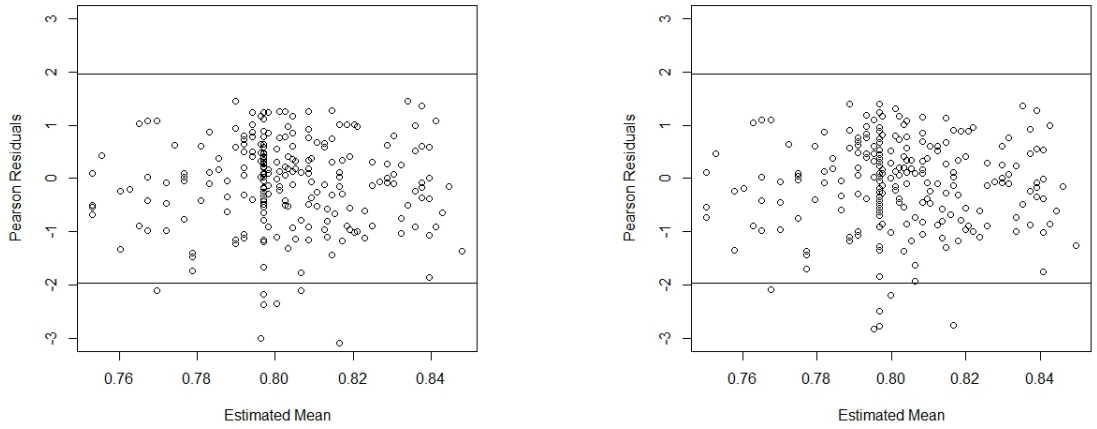
مدل	درجه آزادی	AIC
با پراکندگی ثابت (۱.۵)	۴	−۲۰۲/۵۴۷
با پراکندگی متغیر (۲.۵)	۵	−۱۸۹/۰۳

جدول ۶.۵: مقایسه BIC برای دو مدل

مدل	درجه آزادی	BIC
با پراکندگی ثابت (۱.۵)	۴	−۱۸۸/۶۴۰۸
با پراکندگی متغیر (۲.۵)	۵	−۱۷۱/۶۴۷۷

بررسی فرض ثابت بودن واریانس باقی‌مانده‌ها توسط رسم نمودار باقی‌مانده‌ها در برابر میانگین مقادیر برازش داده شده، صورت می‌گیرد. چنان‌چه این نمودار، ساختار خاصی را نشان ندهد، مثلا به شکل قیفی نباشد، می‌توان فرض ثابت بودن واریانس باقی‌مانده‌ها

را پذیرفت. همان‌طور که در شکل ۱.۵ ملاحظه می‌شود این نمودارها که برای دو مدل با پراکندگی ثابت و متغیر هستند، ساختار خاصی را نشان نمی‌دهند و حدود ٪۹۷ از نقاط در بازه $(-۱/۹۶, ۱/۹۶)$ قرار دارند. بنابراین می‌توان فرض ثابت بودن واریانس باقی‌مانده‌ها را پذیرفت.



(آ) بررسی توزیع دقیق باقی‌مانده‌های پیرسون برای مدل با پارامتر پراکندگی ثابت

(ب) بررسی توزیع دقیق باقی‌مانده‌های پیرسون برای مدل با پارامتر پراکندگی متغیر

شکل ۱.۵: بررسی نیکویی برازش مدل‌های رگرسیون سیمپلکس با باقی‌مانده‌های پیرسون، در مثال PBSC

۲.۲.۵ داده‌های جراحی چشم

سانگ و همکاران (۲۰۰۲) داده‌های مربوط به گاز C_3F_8 را برای بررسی تاثیر استفاده از تزریق این گاز درون چشم، در جراحی‌های ترمیم شبکیه که توسط میرز و همکاران (۱۹۹۲) را مجددا بررسی کردند. داده‌های مربوط به شبکیه، retinal، در بسته simplexreg در نرم‌افزار R گنجانده شده‌است. متغیر پاسخ، درصد حجم گاز باقی‌مانده در چشم است. گاز با سه سطح غلظت متفاوت، ٪۱۵، ٪۲۰ و ٪۲۵ به داخل چشم ۳۱ بیمار قبل از عمل جراحی تزریق شده است. سپس در یک دوره سه ماهه، ۳ تا ۸ بار (به‌طور متوسط ۵ بار) پیگیری شدند و حجم گاز موجود در چشم در دفعات بعدی، به‌عنوان درصد ابتدایی حجم گاز ثبت می‌شود. هدف اولیه، تاثیرگذاری سطح غلظت گاز تزریق‌شده در چشم بیماران بر نرخ فاسد شدن گاز داخل چشم است. این مجموعه داده شامل ۱۸۱ مشاهده (بیماران) و ۶ متغیر تبیینی به شرح زیر است:

- Gas: درصد حجم گاز باقی‌مانده در چشم
- Time: زمان‌بندی روزهای پس از تزریق گاز
- LogT: لگاریتم زمان (log(Time))
- LogT2: توان دوم LogT
- Level: متغیر ظاهری؛ سطوح غلظت گاز اولیه داخل چشم؛ اگر Level $= -۱$ یعنی سطح غلظت گاز ۱۵٪، Level $= ۰$ سطح غلظت گاز ۲۰٪ و Level $= ۱$ سطح غلظت گاز ۲۵٪
- ID: یک عامل نشان‌دهنده بیماران

بخشی از داده‌های retinal موجود در بسته simplexreg به شرح زیر است. قبل از تعیین مدل، ابتدا ساختار همبستگی را از طریق نمودار خودهمبستگی در تاخیرهای متفاوت بررسی می‌کنیم.

	Gas	Time	LogT	LogT2	Level	ID
1	0.99	1	0	0	0	1
2	0.95	2	0.6931472	0.480453	0	1
3	0.95	3	1.098612	1.206949	0	1
4	0.5	8	2.079442	4.324077	0	1
5	0.375	16	2.772589	7.687248	0	1
6	0.225	33	3.496508	12.22557	0	1
7	0.05	61	4.110874	16.89928	0	1
8	0.999	1	0	0	-1	2
9	0.95	2	0.6931472	0.480453	-1	2
10	0.9	3	1.098612	1.206949	-1	2
11	0.9	4	1.386294	1.921812	-1	2
12	0.55	12	2.484907	6.174761	-1	2
13	0.475	21	3.044522	9.269117	-1	2
14	0.25	29	3.367296	11.33868	-1	2
15	0.05	44	3.78419	14.32009	-1	2
16	0.99	1	0	0	1	3
17	0.99	2	0.6931472	0.480453	1	3
18	0.95	3	1.098612	1.206949	1	3
19	0.9	4	1.386294	1.921812	1	3

پارامتر پراکندگی ثابت

ابتدا مدل خطی تعمیم‌یافته سیمپلکس بدون همبستگی پیاپی[۷] را برای داده‌های جراحی چشم در نظر می‌گیریم؛ jامین سطح غلظت گاز برای iامین بیمار در روز t_{ij} را با نماد y_{ij} نشان می‌دهیم. متغیرهای ظاهری به‌صورت زیر بیان می‌شود:

$$x_{ij} = \frac{\text{gas}_{ij} - ۲۰}{۵} \begin{cases} ۱ & \text{گاز با سطح غلظت } ۱۵\% \\ ۰ & \text{گاز با سطح غلظت } ۲۰\% \\ -۱ & \text{گاز با سطح غلظت } ۲۵\% \end{cases}$$

[7] Serial Correlation

با جایگزینی ij توسط k، متغیر پاسخ ما به‌صورت $y_k \sim S^{-}(\mu_k, \sigma^2)$ نشان داده می‌شود که در آن $k = 1, \cdots, N$ و $N = \sum_{i=1}^{n} n_i = 181$ تعداد کل مشاهدات موجود است. پیشگوی خطی برای μ_k به‌صورت زیر است

$$\text{logit}(\mu_k) = \beta_0 + \beta_1 \log(t_k) + \beta_2 \log^2(t_k) + \beta_3 x_k \qquad (۳.۵)$$

که t_k تعداد روزها پس از تزریق گاز تا زمان مشاهده است. با دستورات زیر می‌توان مدل (۳.۵) را بر روی این داده‌ها برازش داد؛ در این‌جا Level همان x_k است.

```
data("retinal", package = "simplexreg")
sim.glm3 <- simplexreg(Gas ~ LogT + LogT2 + Level,
data = retinal, id = ID)
summary(sim.glm3)
```

نتایج برازش مدل (۳.۵) در جدول ۸.۵ گزارش شده‌اند. نتایج مذکور با ۲۳ تکرار و با انحراف ۱۷۷ به‌دست آمده‌اند که لگاریتم درستنمایی آن برابر با ۳۹۱/۲و pـ مقدار ۰/۴۸۶است. همان‌طور که ملاحظه می‌شود، عرض از مبدا و ضرایب متغیرهای LogT2 و Level معنی‌دار هستند.

جدول ۷.۵: باقی‌مانده استاندارد پیرسون مدل (۳.۵)

بزرگترین مقدار	چارک سوم	میانه	چارک اول	کوچکترین مقدار
۳/۶۵۹	۰/۴۴۱	۰/۰۷	−۰/۳۲۷	−۴/۰۱۳

جدول ۸.۵: برآورد پارامترهای مدل (۳.۵)

پارامترها	برآورد	Std.Err	z-value	Pr(>|z|)
عرض از مبدا	۲/۶۸۵	۰/۳۲۶	۸/۲۳۱	$< 2e-16^{***}$
LogT	۰/۰۶۵	۰/۳۴۶	۰/۱۸۸	۰/۸۵۱
LogT2	−۰/۳۳۵	۰/۰۸۴۷	−۳/۹۵۹	$7/53e-05^{***}$
Level	۰/۳۲۵	۰/۱۴۷	۲/۲۰۸	۰/۰۲۷*

'***' سطح ۰، '**' سطح ۰/۰۰۱، '*' سطح ۰/۰۵ و '.' سطح ۰/۱ را نشان می‌دهد.

پارامتر پراکندگی متغیر

در این قسمت، مدل را برای حالتی که σ^2 تابعی از متغیرهای ظاهری است برازش می‌دهیم. پیشگوهای خطی به‌صورت زیر هستند:

$$\text{logit}(\mu_k) = \beta_0 + \beta_1 \log(t_k) + \beta_2 \log^2(t_k) + \beta_3 x_k \qquad \sigma_k^2 = \sigma_0 + \sigma_1 \log(t_k) + \sigma_2 x_k \quad (۴.۵)$$

برازش این مدل در R با کد زیر انجام می‌شود

```
sim.glm4 <- simplexreg(Gas ~ LogT + LogT2 + Level | LogT + Level,
 data = retinal, id = ID)
summary(sim.glm4)
```

نتایج مذکور با ۲۰ تکرار و با انحراف ۱۸۱ به‌دست آمده‌اند که لگاریتم درستنمایی آن ۳۷۰/۲ و pـ مقدارش ۰/۴۰۲۶ است. در جدول ۱۰.۵ برآورد هر پارامتر، برای مدل حاشیه‌ای سیمپلکس با پارامتر پراکندگی متغیر نشان داده شده‌است. نتایج نشان می‌دهد که ضریب LogT در ساختار پیشگوی میانگین معنی‌دار نبوده و بایستی حذف شود.

جدول ۹.۵: باقی‌مانده استاندارد پیرسون مدل (۴.۵)

بزرگترین مقدار	چارک سوم	میانه	چارک اول	کوچکترین مقدار
۴/۴۲	۰/۳۹۱	۰/۰۶۵	−۰/۳۱۹	−۴/۳۶۷

جدول ۱۰.۵: برآورد پارامترهای مدل (۴.۵)

پارامترها	برآورد	Std.Err	z-value	Pr(>|z|)
عرض از مبدا	۲/۷۱۷	۰/۳۵	۷/۷۶۳	$۸/۳۲e-۱۵^{***}$
LogT	−۰/۰۲	۰/۳۴۶	−۰/۰۵۹	۰/۹۵۲
LogT2	−۰/۳۱۲۶۳	۰/۰۸۱۱۷	−۳/۸۵۱	$۰/۰۰۰۱^{***}$
Level	۰/۴۰۵	۰/۱۴۵	۲/۷۹۲	$۰/۰۰۵^{**}$
عرض از مبدا	۶/۱۵۵	۰/۱۹۸	۳۰/۹۵۷	$<۲e-۱۶^{***}$
LogT	−۰/۴۵۸	۰/۰۸۰۴	−۵/۷۰۱	$۱/۱۹e-۰۸^{***}$
Level	−۰/۴۹۳	۰/۱۴۲	−۳/۴۵۶	$۰/۰۰۰۵^{***}$

'***' سطح ۰، '**' سطح ۰/۰۰۱، '*' سطح ۰/۰۵ و '.' سطح ۰/۱ را نشان می‌دهد.

مدل حاشیه‌ای سیمپلکس با پارامتر پراکندگی متغیر و ساختار همبستگی Exc

در این قسمت، پارامتر پراکندگی مدل حاشیه‌ای سیمپلکس را به‌صورت متغیر در نظر می‌گیریم. هم‌چنین با توجه به ساختار همبستگی تبادل‌پذیر (Exc) برای بررسی تاثیر زمان بر پارامتر پراکندگی، پیشگوهای خطی به‌صورت زیر معرفی می‌شوند:

(۵.۵)

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \log(t_{ij}) + \beta_2 \log^2(t_{ij}) + \beta_3 x_{ij} \quad \sigma^2_{ij} = \sigma_0 + \sigma_1 \log(t_{ij}) + \sigma_2 \log^2(t_{ij}).$$

با اجرای کد زیر، نتایج برازش مدل در ۱۲.۵ آمده‌است. نتایج مذکور با ۱۷ تکرار و با انحراف ۱۷۷ به‌دست آمده است.

```
sim.gee1 <- simplexreg(Gas ~ LogT + LogT2 + Level | LogT + Level | Time,
corr = "Exc", id = ID, data = retinal)
summary(sim.gee1)
```

جدول ۱۱.۵: باقی‌مانده استاندارد پیرسون مدل (۵.۵)

بزرگترین مقدار	چارک سوم	میانه	چارک اول	کوچکترین مقدار
۵/۰۴۲	۰/۴۷۲	۰/۱۳۶	−۰/۲۵۱	−۳/۸

جدول ۱۲.۵: برآورد پارامترها با ساختار همبستگی Exc در مدل (۵.۵)

پارامترها	برآورد	Std.Err	z-value	Pr(>|z|)
عرض از مبدا	۲/۶۵۴	۰/۳۰۳	۸/۷۶	$< 2e-16^{***}$
LogT	۰/۱۶	۰/۲۴۴	۰/۶۵۷	۰/۵۱۱
LogT2	−۰/۳۷۹	۰/۰۶۴	−۵/۹۱۵	$3/31e-09^{***}$
Level	۰/۲۵	۰/۱۹۹	۱/۲۵۳	۰/۲۱
α	−۱/۳۸۸	۰/۴۸۹	−۲/۸۳۴	۰/۰۰۴۶**
ρ	۰/۲۴۹	۰/۱۲۲	۲/۰۴۱	۰/۰۴۱*

'***' سطح ۰، '**' سطح ۰/۰۰۱، '*' سطح ۰/۰۵ و '.' سطح ۰/۱ را نشان می‌دهد.

نتایج جدول ۱۲.۵ نشان می‌دهد که ضرایب LogT و Level معنی‌دار نیستند و بایستی از مدل حذف شده و مدل مذکور مجددا برازش شود.

مدل حاشیه‌ای سیمپلکس با پارامتر پراکندگی متغیر و ساختار همبستگی AR(1)

اکنون مدل (۵.۵) را مجددا با استفاده از ساختار همبستگی اتورگرسیو مرتبه اول در داده‌های طولی، توسط کد زیر برازش می‌دهیم

```
sim.gee2 <- simplexreg(Gas ~ LogT + LogT2 + Level | LogT + Level | Time,
corr = "AR1", id = ID, data = retinal)
summary(sim.gee2)
```

نتایج مذکور با ۲۵ تکرار و با انحراف ۱۸۱ به‌دست آمده‌اند. در جدول ۱۴.۵ برآورد هر پارامتر، برای مدل حاشیه‌ای سیمپلکس با پارامتر پراکندگی متغیر و ساختار همبستگی AR(1) نشان داده شده‌است. در این‌جا همانند قبل ضرایب LogT و Level در پیشگوی میانگین و Level در پیشگوی پراکندگی معنی‌دار نیستند و این متغیرها باید از مدل مذکور حذف شوند.

جدول ۱۳.۵: باقی‌مانده استاندارد پیرسون مدل (۵.۵) با ساختار همبستگی AR(1)

بزرگترین مقدار	چارک سوم	میانه	چارک اول	کوچکترین مقدار
۴/۶۳۷	۰/۳۹۱	۰/۰۵۹	−۰/۳۴۵	−۴/۵۸

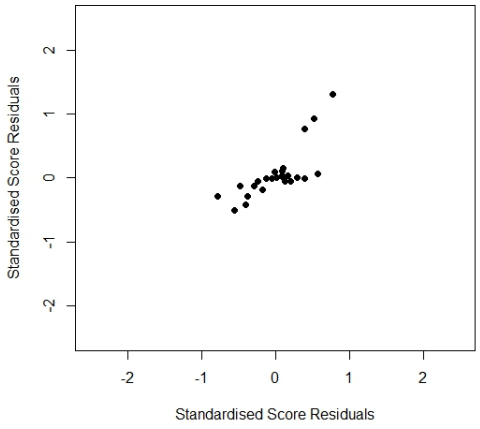
جدول ۱۴.۵: برآورد پارامترها با ساختار همبستگی AR(1) در مدل (۵.۵)

پارامترها	برآورد	Std.Err	z-value	Pr(>|z|)
عرض از مبدا	۲/۷۲۱	۰/۲۰۳	۱۳/۴۲۵	$< 2e-16^{***}$
LogT	۰/۰۳۴	۰/۳۱۲	۰/۱۰۹	۰/۹۱۳
LogT2	−۰/۳۲۹	۰/۰۸۵	−۳/۸۶۹	۰/۰۰۰۱***
Level	۰/۴۰۹	۰/۲۱۶	۱/۸۸۷	۰/۰۵۹.
عرض از مبدا	۶/۱۵۳	۰/۳۵۱	۱۷/۵۱۱	$< 2e-16^{***}$
LogT	−۰/۴۵۷	۰/۱۶۹	−۲/۶۹۹	۰/۰۰۷**
Level	−۰/۴۹۲	۰/۳۵۶	−۱/۳۸۱	۰/۱۶۷
α	−۰/۳۴۹	۰/۱۸۶	−۱/۸۷۲	۰/۰۶۱.
ρ	۰/۷۰۵	۰/۱۳۱	۵/۳۶۳	$8/17e-08^{***}$

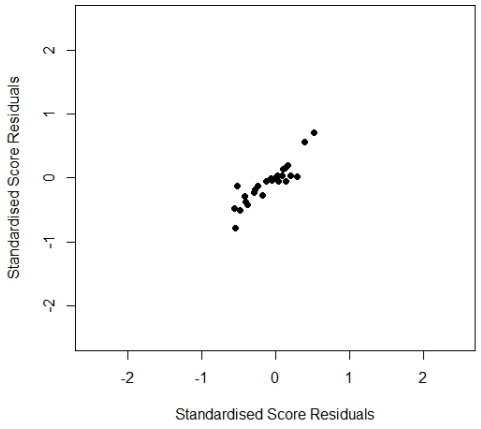
'***' سطح ۰، '**' سطح ۰/۰۰۱، '*' سطح ۰/۰۵ و '.' سطح ۰/۱ را نشان می‌دهد.

برای بررسی صحت ساختار همبستگی، میزان خودهمبستگی‌ها در چهار تاخیر محاسبه و نتایج

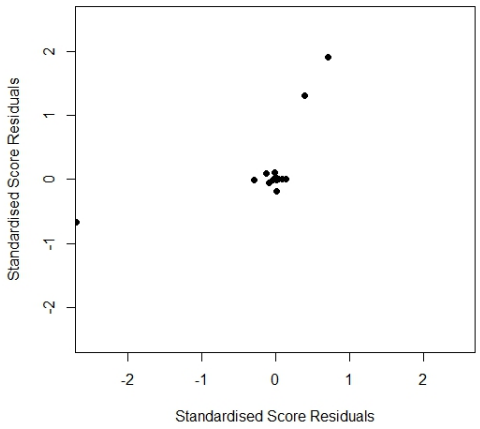
آن در نمودار شکل ۲.۵ رسم شده‌اند؛ که می‌توان دریافت که خودهمبستگی در تاخیر اول، ۱ =lag ، قوی‌تر است و با افزایش تاخیر، خودهمبستگی کاهش می‌یابد، به‌طوری‌که در تاخیر چهارم، خودهمبستگی ناچیز می‌شود. در نتیجه مشخص است که ساختار همبستگی AR(1) برای داده‌ها مناسب است. در شکل (آ) ۳.۵، نمودار متغیر $\hat{a}_{ij}$ در برابر پیشگوی خطی η_i نشان



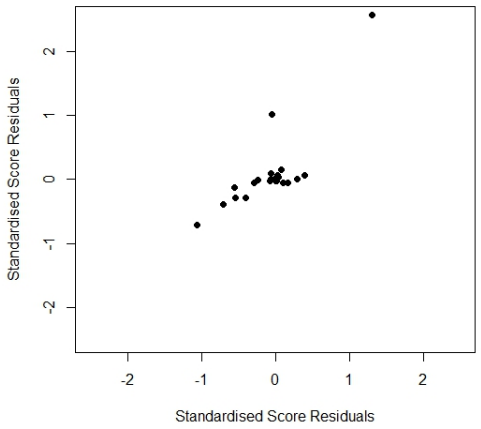
(آ) بررسی خودهمبستگی $lag = ۱$



(ب) بررسی خودهمبستگی $lag = ۲$



(ج) بررسی خودهمبستگی $lag = ۳$

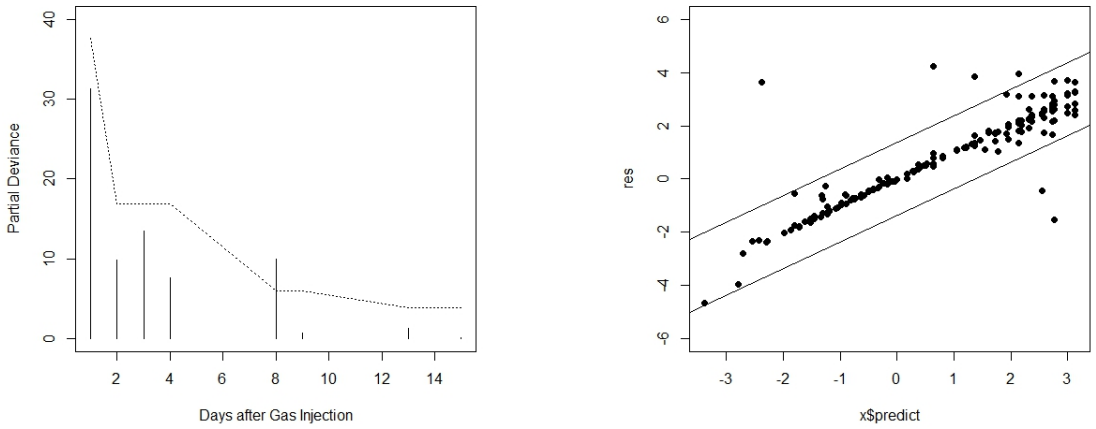


(د) بررسی خودهمبستگی $lag = ۴$

شکل ۲.۵: بررسی خودهمبستگی در چهار تاخیر مختلف، در مثال جراحی چشم.

داده شده است و ٪۹۷ از نقاط در نوار اطمینان ٪۹۵ قرار دارند که نشان‌دهنده یک روند خطی است. هم‌چنین گویای عملکرد تابع پیوند لجیت است که یک انتخاب مناسب برای مدل‌بندی داده‌هاست. شکل (ب) ۳.۵، نشان‌دهنده آزمون نیکویی برازش مدل حاشیه‌ای سیمپلکس با پارامتر پراکندگی متغیر و ساختار همبستگی اتورگرسیو مرتبه اول است که بیانگر خوب بودن و قابل اعتماد بودن مدل است؛ حتی با توجه به این‌که در تاخیر هشتم مدل خوب عمل نکرده

است. نتیجه این‌که مدل حاشیه‌ای سیمپلکس با پارامتر پراکندگی متغیر و ساختار همبستگی اتورگرسیو مرتبه اول عملکرد مناسبی داشته است.



(آ) بررسی تابع پیوند (ب) آزمون نیکویی برازش

شکل ۳.۵: بررسی تابع پیوند و نیکویی برازش مدل در مثال جراحی چشم

۳.۵ نتیجه‌گیری

در این پایان‌نامه، یک مدل رگرسیونی مناسب برای پاسخ‌هایی که به‌طور پیوسته در فاصله واحد تغییر می‌کنند، ارائه شد. این مدل رگرسیونی بر اساس پذیره پیروی متغیر پاسخ از توزیع سیمپلکس است.

طبق مطالعات شبیه‌سازی انجام‌شده، برآورد پارامترهای مدل در هر دو حالتی که پارامتر پراکندگی ثابت و متغیر باشد، برآوردی سازگار و قابل اعتماد است؛ دقت این برآوردها به واسطه کاهش خطای معیار، افزایش می‌یابد.

ما مدل را علاوه بر مدل‌بندی پارامتر میانگین، برای حالتی که پارامتر پراکندگی نیز مدل‌بندی می‌شود، بررسی کردیم. در این پایان‌نامه، مدل‌های متنوعی مورد بررسی قرار گرفتند. در برخی لز آن‌ها، پذیره استقلال بین مشاهدات مد نظر قرار گرفت و در برخی دیگر به تحلیل داده‌های طولی که همبستگی جز لاینفک ساختار آن‌ها است، پرداختیم.

به کمک بسته simplexreg در R، تابع چگالی، تابع توزیع تجمعی، تابع چندک و تولید اعداد تصادفی را مورد بررسی قرار داده و به تحلیل باقی‌مانده‌های رگرسیون سیمپلکس با استفاده از روش درستنمایی ماکسیمم و معادلات برآوردیابی تعمیم‌یافته پرداختیم.

پیوست آ

۱.آ تولید عدد تصادفی از توزیع MIG

برای تولید متغیر تصادفی از توزیع گوسی معکوس آمیخته از روش پیشنهادی جرجنسن (۱۹۹۱) استفاده می‌کنیم که در آن تابع مولد گشتاور X به‌صورت زیر می‌باشد:

$$m(s;\xi,\epsilon^{۲},p)=(۱-p)m_{۱}(s;\xi,\epsilon^{۲})+pm_{۲}(s;\xi,\epsilon^{۲})$$

که در آن $m_{۱}$ و $m_{۲}$ به ترتیب تابع مولد گشتاورهای $X_{۱}$ و $X_{۲}$ می‌باشند.
حال از سشادری (۱۹۹۳) داریم

$$m_{۱}(s;\xi,\epsilon^{۲})=\exp\left\{\frac{۱}{\epsilon}\left[\frac{۱}{\xi\epsilon}-\left(\frac{۱}{\xi^{۲}\epsilon^{۲}}-۲s\right)^{\frac{۱}{۲}}\right]\right\}.$$

و از جانسون و کاتز (۱۹۷۰) داریم

$$m_{۲}(s;\xi,\epsilon^{۲})=m_{۱}(s;\xi,\epsilon^{۲})m_{۳}(s;\xi,\epsilon^{۲}),$$

که در آن $m_{۳}$ برابر است با

$$m_{۳}(s;\xi,\epsilon^{۲})=(۱-۲\xi^{۲}\epsilon^{۲}s)^{-\frac{۱}{۲}}.$$

$m_{۳}$ همان تابع مولد گشتاور $X_{۳}\sim\xi^{۲}\epsilon^{۲}\chi^{۲}_{(۱)}$ است که $X_{۱}$ و $X_{۳}$ مستقل از هم به‌صورت زیر می‌باشند

$$m(s;\xi,\epsilon^{۲},p)=m_{۱}(s;\xi,\epsilon^{۲})[۱-p+pm_{۳}(s;\xi^{۲}\epsilon^{۲})].$$

که می‌توان نوشت $X = X_۱ + V$، که در آن

$$V = \begin{cases} ۰ & ۱-p \quad \text{با احتمال} \\ X_۳ & p \quad \text{با احتمال} \end{cases}$$

و هم‌چنین $X_۱$ و V از هم مستقل و $X_۱ \sim IG(\xi, \epsilon^۲)$.

۲.آ ماتریس اطلاع گودامب

ماتریس حساسیت، یک ماتریس (۳ × ۳) به‌صورت زیر است

$$\boldsymbol{S} = \mathrm{E}\{\frac{\partial \Upsilon(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\} = \begin{pmatrix} \boldsymbol{S}_{۱۱} & \boldsymbol{S}_{۱۲} & \boldsymbol{S}_{۱۳} \\ \boldsymbol{S}_{۲۱} & \boldsymbol{S}_{۲۲} & \boldsymbol{S}_{۲۳} \\ \boldsymbol{S}_{۳۱} & \boldsymbol{S}_{۳۲} & \boldsymbol{S}_{۳۳} \end{pmatrix}$$

که در آن $\boldsymbol{S}_{۱۲} = ۰$ ، $\boldsymbol{S}_{۱۳} = ۰$ و $\boldsymbol{S}_{۲۳} = ۰$. مولفه چهارم، $\boldsymbol{S}_{۲۱}$، نیز برابر با صفر است زیرا $\mathrm{E}[u_{ij}] = ۰$؛ در نتیجه می‌توان ماتریس $\boldsymbol{S}$ را به‌شکل زیر بازنویسی کرد

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{۱۱} & ۰ & ۰ \\ ۰ & \boldsymbol{S}_{۲۲} & ۰ \\ \boldsymbol{S}_{۳۱} & \boldsymbol{S}_{۳۲} & \boldsymbol{S}_{۳۳} \end{pmatrix}$$

حال وارون ماتریس فوق به‌صورت زیر است

$$\boldsymbol{S}^{-۱} = \begin{pmatrix} \boldsymbol{S}_{۱۱}^{-۱} & ۰ & ۰ \\ ۰ & \boldsymbol{S}_{۲۲}^{-۱} & ۰ \\ -\boldsymbol{S}_{۳۳}\boldsymbol{S}_{۳۱}\boldsymbol{S}_{۱۱}^{-۱} & -\boldsymbol{S}_{۳۳}^{-۱}\boldsymbol{S}_{۳۲}\boldsymbol{S}_{۲۲}^{-۱} & \boldsymbol{S}_{۳۳}^{-۱} \end{pmatrix}$$

باید توجه داشت که تمام مولفه‌های روی قطری اصلی، وارون هستند و زمانی که توزیع $r_{ij}r_{ij'}$ مستقل از پارامترهای میانگین و پراکندگی باشند، مولفه‌های $\boldsymbol{S}_{۳۱}$ و $\boldsymbol{S}_{۳۲}$ نیز برابر با صفرند. مولفه‌های روی قطر اصلی ماتریس $\boldsymbol{S}$ برابرند با

$$\boldsymbol{S}_{۱۱} = -\sum_{i=۱}^{n} \boldsymbol{D}_i^T \boldsymbol{A}_i \boldsymbol{V}_i^{-۱} \boldsymbol{A}_i \boldsymbol{D}_i$$

$$\boldsymbol{S}_{۲۲} = -\sum_{i=۱}^{n} (\frac{\partial \boldsymbol{\sigma}_i^T}{\partial \boldsymbol{\gamma}}) \boldsymbol{\Sigma}_i^{-۱} (\frac{\partial \boldsymbol{\sigma}_i}{\partial \boldsymbol{\gamma}^T})$$

$$\boldsymbol{S}_{۳۳} = -\sum_{i=۱}^{n} (\frac{\partial \boldsymbol{\xi}_i^T}{\partial \boldsymbol{\alpha}}) \boldsymbol{H}_i^{-۱} (\frac{\partial \boldsymbol{\xi}_i}{\partial \boldsymbol{\alpha}^T})$$

ماتریس متغیر $\boldsymbol{R}$ یک ماتریس ۳ × ۳ به‌صورت زیر است

$$\boldsymbol{R}=\mathrm{E}\{\Upsilon(\boldsymbol{\theta})\Upsilon^{T}(\boldsymbol{\theta})\}\begin{pmatrix}\boldsymbol{V}_{11} & \boldsymbol{V}_{12} & \boldsymbol{V}_{13}\\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} & \boldsymbol{V}_{23}\\ \boldsymbol{V}_{31} & \boldsymbol{V}_{32} & \boldsymbol{V}_{33}\end{pmatrix}$$

که مولفه‌های آن در زیر آمده است

$$\boldsymbol{V}_{11}=\mathrm{E}\{\Psi_1\Psi_1^{T}\}=\sum_{i=1}^{n}\boldsymbol{D}_i^{T}\boldsymbol{A}_i\boldsymbol{V}_i^{-1}\mathrm{cov}(\boldsymbol{s}_i)\boldsymbol{V}_i^{-1}\boldsymbol{A}_i\boldsymbol{D}_i$$

$$\boldsymbol{V}_{12}=\mathrm{E}\{\Psi_1\Psi_2^{T}\}=\sum_{i=1}^{n}\boldsymbol{D}_i^{T}\boldsymbol{A}_i\boldsymbol{V}_i^{-1}\mathrm{cov}(\boldsymbol{s}_i,\boldsymbol{d}_i)\boldsymbol{\Sigma}_i^{-1}\left(\frac{\partial\boldsymbol{\sigma}_i}{\partial\boldsymbol{\gamma}^{T}}\right)$$

$$\boldsymbol{V}_{13}=\mathrm{E}\{\Psi_1\Psi_3^{T}\}=\sum_{i=1}^{n}\boldsymbol{D}_i^{T}\boldsymbol{A}_i\boldsymbol{V}_i^{-1}\mathrm{cov}(\boldsymbol{s}_i,\boldsymbol{r}_i)\boldsymbol{H}_i^{-1}\left(\frac{\partial\boldsymbol{\xi}_i}{\partial\boldsymbol{\alpha}^{T}}\right)$$

$$\boldsymbol{V}_{22}=\mathrm{E}\{\Psi_2\Psi_2^{T}\}=\sum_{i=1}^{n}\left(\frac{\partial\boldsymbol{\sigma}_i^{T}}{\partial\boldsymbol{\gamma}}\right)\boldsymbol{\Sigma}_i^{-1}\mathrm{cov}(\boldsymbol{d}_i)\boldsymbol{\Sigma}_i^{-1}\left(\frac{\partial\boldsymbol{\sigma}_i}{\partial\boldsymbol{\gamma}^{T}}\right)$$

$$\boldsymbol{V}_{23}=\mathrm{E}\{\Psi_2\Psi_3^{T}\}=\sum_{i=1}^{n}\left(\frac{\partial\boldsymbol{\sigma}_i^{T}}{\partial\boldsymbol{\gamma}}\right)\boldsymbol{\Sigma}_i^{-1}\mathrm{cov}(\boldsymbol{d}_i,\boldsymbol{r}_i)\boldsymbol{H}_i^{-1}\left(\frac{\partial\boldsymbol{\xi}_i}{\partial\boldsymbol{\alpha}^{T}}\right)$$

$$\boldsymbol{V}_{33}=\mathrm{E}\{\Psi_3\Psi_3^{T}\}=\sum_{i=1}^{n}\left(\frac{\partial\boldsymbol{\xi}_i^{T}}{\partial\boldsymbol{\alpha}}\right)\boldsymbol{H}_i^{-1}\mathrm{cov}(\boldsymbol{r}_i)\boldsymbol{H}_i^{-1}\left(\frac{\partial\boldsymbol{\xi}_i}{\partial\boldsymbol{\alpha}^{T}}\right).$$

می‌دانیم که $\boldsymbol{V}_{21}=\boldsymbol{V}_{12}^{T}$، $\boldsymbol{V}_{31}=\boldsymbol{V}_{13}^{T}$ و $\boldsymbol{V}_{32}=\boldsymbol{V}_{23}^{T}$.

با توجه به این‌که $\mathrm{cov}(\boldsymbol{w_i})=diag\{\mathrm{v}(\boldsymbol{\mu_{ij}})\}\mathrm{cov}(\boldsymbol{u}_i)diag\{\mathrm{v}(\boldsymbol{\mu_{ij}})\}$ که برآورد $\mathrm{cov}(\boldsymbol{u}_i)$ از طریق برآورد $\hat{\mu}_{ij}$ها به‌دست می‌آید و در این عبارت می‌توان $\mathrm{cov}(\boldsymbol{u}_i)$ را با $\hat{\boldsymbol{u}}_i\hat{\boldsymbol{u}}_i^{T}$ جایگذاری کرد.

آ.۳ دستورات نرم‌افزار R

شکل ۱.۲

```
x<-seq(0,3,0.01)
f<-(2*pi*x^3)^(-1/2)*(exp((-(x-1)^2)/(2*x)))
f1<-(2*pi*x^3)^(-1/2)*(exp((-(x-2)^2)/(8*x)))
f2<-(2*pi*x^3)^(-1/2)*(exp((-(x-5)^2)/(50*x)))
plot(x, f, type="l", ylim=c(0, 3), col=3, lwd=2)
lines(x, f1, lty=2, col=4, lwd=2)
lines(x, f2, lty=4, col=6, lwd=2)
legend(2.3,3, c(expression(xi==1), expression(xi==2), expression(xi==5)),
```

```
col=c(3,4,6), text.col="black", lty=c(1,2,4) ,lwd=2,
merge=TRUE, bg="white")
xlab='x'
f<-(2*pi*x^3)^(-1/2)*(exp((-(x-1)^2)/(2*x)))
f1<-(2*(2*pi*x^3))^(-1/2)*(exp((-(x-1)^2)/(4*x)))
f2<-(5*(2*pi*x^3))^(-1/2)*(exp((-(x-1)^2)/(10*x)))
plot(x, f, type="l", ylim=c(0,3), col=3, lwd=2)
lines(x, f1, lty=2, col=4, lwd=2)
lines(x, f2, lty=4, col=6, lwd=2)
legend(2.3,3, c(expression(epsilon^2==1), expression(epsilon^2==2),
expression(epsilon^2==5)), col=c(3,4,6), text.col="black",
lty=c(1,2,4), merge=TRUE, bg="white", lwd=2)
xlab='x'
```

شکل ۲.۲

```
library(simplexreg)
f=function(x,a,b){
dsimplex(x,a,b)
}
curve(f(x,0.1,4), c(0,1), col="red", xlim=c(0,1), ylim=c(0,18),
main=expression(mu==0.1), xlab=' ', ylab=expression(sigma^2==4^2))
```

شکل ۳.۲

```
library(simplexreg)
f=function(x,a,b){
dsimplex(x,a,b)
}
curve(f(x,0.1,4), c(0,1),col="red", xlim=c(0,1), ylim=c(0,15),
main=expression(mu==0.1), xlab=' ', ylab=expression(sigma^2==4^2))
par(new=TRUE)
v=sqrt((0.1*0.9)^3)
curve(dnorm(x,0.1,v*4), xlim=c(0,1), lty=2, ylim=c(0,15),
main=' ', xlab=' ', ylab=' ')
```

شکل ۴.۲

```
## or plotting (histogram):
rsimplot <- function(n, mu, sigma) {
x <- seq(0.01, 0.99, 0.01)
for (j in 1:length(sigma)) {
for (i in 1:length(mu)) {
hist(rsimplex(n, mu[i], sigma[j]), density = 50,
xlim = c(0, 1), ylim = c(0, max(dsimplex(x, mu[i], sigma[j])) + 0.5),
freq = F, breaks  =  20, xlab = " ", ylab = " ", main = NULL)
lines(x, dsimplex(x, mu[i], sigma[j]), type  =  "l", col = "red")
}
}
}
par(mfrow = c(3, 3), mar = c(2, 2, 2, 2), cex = 0.55)
rsimplot(5000, c(0.1, 0.5, 0.7), c(4, 2, 1))
```

شکل ۵.۲

```
## plotting (Kernel density):
rsimplot <- function(n, mu, sigma) {
x <- seq(0.01, 0.99, 0.01)
for (j in 1:length(sigma)) {
for (i in 1:length(mu)) {
plot(x, dsimplex(x, mu[i], sigma[j]),
ylim = c(0, max(dsimplex(x, mu[i], sigma[j])) + 0.5),
type = "l", col = "red")
lines(density(rsimplex(n, mu[i], sigma[j])), lty = 2)
}
}
}
par(mfrow = c(3, 3), mar = c(2, 2, 2, 2), cex = 0.55)
rsimplot(5000, c(0.1, 0.5, 0.7), c(4, 2, 1))
```

شکل ۶.۲

```
library(simplexreg)
z<-array(0,dim=c(500,1))
mu=0.5
sigma=0.01
xx=rsimplex(500,mu,sigma)
for (i in 1:500){
z[i]=(xx[i] -mu)/(sigma*sqrt((mu^3)*(1-mu)^3))
}
qqnorm(z,main=expression(sigma^2==0.01),xlab="",ylab="")
qqline(z, col = 2,lwd=2)
```

شکل ۱.۵

```
plot(sim.glm1, type="residuals", res="stdPerr", ylim=c(-3,3))
plot(sim.glm2, type="residuals", res="stdPerr", ylim = c(-3,3))
```

شکل ۲.۵

```
plot(sim.glm3, type="corr", xlim=c(-2.5,2.5), ylim=c(-2.5,2.5), pch=16)
plot(sim.glm3, type="corr", lag=2, xlim=c(-2.5,2.5),
ylim=c(-2.5,2.5), pch=16)
plot(sim.glm3, type="corr", lag=3, xlim=c(-2.5,2.5),
ylim=c(-2.5,2.5), pch=16)
plot(sim.glm3, type="corr", lag=4, xlim=c(-2.5,2.5),
ylim=c(-2.5,2.5), pch=16)
```

شکل ۳.۵

```
plot(sim.gee2, type="residuals", ylim=c(-6,6), pch=16)
plot(sim.gee2, type="GOF", xlab="Days after Gas Injection", ylim=c(0,40))
```

کد مربوط به مثال دوم شبیه‌سازی فصل ۳

```
library(Formula)
```

```
library(simplexreg)
remove(list=ls())
M=500
param = matrix (NA, nrow = M, ncol = 3)
param1 = matrix (NA, nrow = M, ncol = 1)
par = matrix (NA, nrow = M, ncol = 4)
beta0=rep(0,M)
beta1=rep(0,M)
beta2=rep(0,M)
sigma1=rep(0,M)
for (j in 1:M)
{
#Data generation from differnt seeds
set.seed(12345+j)
print(j)
n = 200
beta=c(1.5,0.5,-0.5)
sigma=sqrt(4)
x1 = rep(1,length(n))
x2 = rbinom(n,1,0.6)
x3=runif(n, 0, 5)
X = cbind(x1,x2,x3)
eta = X%*%beta
mu = exp(eta)/(1+exp(eta))
y <- rep(0, n)
for (i in 1:n){
y[i] <- rsimplex(1, mu[i],sqrt(sigma))
}
data=data.frame(cbind(y,X))
try({
# Error in svd(w) : infinite or missing values in 'x'
fit=simplexreg(y ~ x2+x3 , data=data, link="logit")
param[j,] = as.vector(unlist(fit$fixef[,1]))
param1[j,] = as.vector(unlist(fit$Dispersion))
```

```
})
summary(fit)
beta0[j]=fit$fixef[1]
beta1[j]=fit$fixef[2]
beta2[j]=fit$fixef[3]
sigma1[j]=fit$Dispersion
}
#All coefficients in one matrix
A<-matrix(NA,nrow=M,ncol=4)
for(j in 1:3){
A[,1]<-beta0
A[,2]<-beta1
A[,3]<-beta2
A[,4]<-sigma1
}
index<-which((rowSums(abs(A))>4))
A<-A[-index,]
A
#Final estimation of coefficients
beta.and.gamma.hat<-colMeans(A)
beta.and.gamma.hat
#Standard deviation of all coefficients after remove outliers
sd<-GMCM:::colSds(A)
sd
# Computing bias
b0=beta.and.gamma.hat[1]-1.5
b1=beta.and.gamma.hat[2]-0.5
b2=beta.and.gamma.hat[3]+0.5
s=beta.and.gamma.hat[4]-2
#Computing MSE
betaa=c(1.5,0.5,-0.5,2)
Mse=numeric(4)
for (k in 1:4){
mse=function(x) mean((x-betaa[k])^2)
```

```
Mse[k]=mse(A[,k])
}
Mse
#Computing variances
var1=mean((beta0-mean(beta0))^2)
var2=mean((beta1-mean(beta1))^2)
var3=mean((beta2-mean(beta2))^2)
var4=mean((sigma1-mean(sigma1))^2)
f1=(beta0-mean(beta0))/sqrt(var1)
f2=(beta1-mean(beta1))/sqrt(var2)
f3=(beta2-mean(beta1))/sqrt(var3)
f4=(sigma1-mean(sigma1))/sqrt(var4)
#Computing qq-plot
qqnorm(f1,main=expression(beta[0]), xlab="", ylab="")
qqline(f1, col=2, lwd=2)
qqnorm(f2, main=expression(beta[1]), xlab="", ylab="")
qqline(f2, col = 2, lwd=2)
qqnorm(f3, main=expression(beta[2]), xlab="", ylab="")
qqline(f3, col = 2, lwd=2)
qqnorm(f4, main=expression(sigma^2), xlab="", ylab="")
qqline(f4, col = 2, lwd=2)
```

کد مربوط به مثال سوم شبیه‌سازی فصل ۳

```
library(Formula)
library(simplexreg)
remove(list=ls())
M=500
param = matrix (NA, nrow = M, ncol = 3)
param1 = matrix (NA, nrow = M, ncol = 2)
par = matrix (NA, nrow = M, ncol = 5)
beta0=rep(0,M)
beta1=rep(0,M)
beta2=rep(0,M)
```

```
gamma0=rep(0,M)
gamma1=rep(0,M)
for (j in 1:M)
{
#Data generation from differnt seeds
set.seed(12345+j)
print(j)
n = 400
beta=c(1,0.5,-1)
sigma=sqrt(2)
x1 = rep(1,length(n))
x2 = rnorm(n)
x3 = rbinom(n,1,0.6)
X = cbind(x1,x2,x3)
eta = X%*%beta
mu = exp(eta)/(1+exp(eta))
gamma=c(5,0.3)
z1=rep(1,length(n))
z2=rnorm(n)
Z = cbind(z1,z2)
dispersion =Z%*%gamma
sigma1 = exp(dispersion)
y <- rep(0, n)
for (i in 1:n){
y[i] <- rsimplex(1, mu[i], sqrt(sigma1[i]))
}
data=data.frame(cbind(y,X,Z))
try({
#Error in svd(w) : infinite or missing values in 'x'
fit=simplexreg(y ~ x2+x3 | z2, data=data, link="logit")
param[j,] = as.vector(unlist(fit$fixef[,1]))
param1[j,] = as.vector(unlist(fit$dispar[,1]))
})
summary(fit)
```

```
beta0[j]=fit$fixef[1]
beta1[j]=fit$fixef[2]
beta2[j]=fit$fixef[3]
gamma0[j]=fit$dispar[1]
gamma1[j]=fit$dispar[2]
}
#All coefficients in one matrix
A<-matrix(NA,nrow=M,ncol=5)
for(j in 1:5){
A[,1]<-beta0
A[,2]<-beta1
A[,3]<-beta2
A[,4]<-gamma0
A[,5]<-gamma1
}
index<-which((rowSums(abs(A))>10))
A<-A[-index,]
A
#Final estimation of coefficients
beta.and.gamma.hat<-colMeans(A)
beta.and.gamma.hat
# Standard deviation of all coefficients after remove outliers
sd<-GMCM:::colSds(A)
sd
# Computing bias
b0=beta.and.gamma.hat[1]-1
b1=beta.and.gamma.hat[2]-0.5
b2=beta.and.gamma.hat[3]+1
g0=beta.and.gamma.hat[4]-5
g1=beta.and.gamma.hat[5]-0.3
#Computing MSE
betaa=c(1, 0.5,- 1, 5, 0.3)
Mse=numeric(5)
for (k in 1:5){
```

```
mse=function(x) mean((x-betaa[k])^2)
Mse[k]=mse(A[,k])
}
Mse
#Computing variances
var1=mean((A[,1]-mean(A[,1]))^2)
var2=mean((A[,2]-mean(A[,2]))^2)
var3=mean((A[,3]-mean(A[,3]))^2)
var4=mean((A[,4]-mean(A[,4]))^2)
var5=mean((A[,5]-mean(A[,5]))^2)
f1=(A[,1]-mean(A[,1]))/sqrt(var1)
f2=(A[,2]-mean(A[,2]))/sqrt(var2)
f3=(A[,3]-mean(A[,3]))/sqrt(var3)
f4=(A[,4]-mean(A[,4]))/sqrt(var4)
f5=(A[,5]-mean(A[,5]))/sqrt(var5)
#Computing qq-plot
qqnorm(f1, main=expression(beta[0]), xlab="", ylab="")
qqline(f1, col=2, lwd=2)
qqnorm(f2, main=expression(beta[1]), xlab="", ylab="")
qqline(f2, col=2, lwd=2)
qqnorm(f3, main=expression(beta[2]), xlab="", ylab="")
qqline(f3, col=2, lwd=2)
qqnorm(f4, main=expression(gamma[0]), xlab="", ylab="")
qqline(f4, col=2, lwd=2)
qqnorm(f5, main=expression(gamma[1]), xlab="", ylab="")
qqline(f5, col=2, lwd=2)
```

مراجع

[1] Allan, D. S., Keeney, M., Howson-Jan, K., Popma, J., Weir, K., Bhatia, M., Sutherland, D. & Chin-Yee, I. H. (2002). Number of viable CD34+ cells reinfused predicts engraftment in autologous hematopoietic stem cell transplantation. *Bone marrow transplantation*, 29(12), 967.

[2] Balakrishnan, N., Leiva, V., Sanhueza, A., & Cabrera, E. (2009). Mixture inverse Gaussian distributions and its transformations, moments and applications. *Statistics*, 43(1), 91-104.

[3] Balemi, A., & Lee, A. (2009). Comparison of GEE1 and GEE2 estimation applied to clustered logistic regression. *Journal of Statistical Computation and Simulation*, 79(4), 361-378.

[4] Barndorff-Nielsen, O. E., & Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis*, 39(1), 106-116.

[5] Chambers, J. M., & Hastie, T. J. (Eds.). (1992). *Statistical models in S*, (251). Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

[6] Chhikara, R. (1988). *The Inverse Gaussian Distribution: Theory: Methodology, and Applications*, (95). CRC Press.

[7] Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika*, 74(3), 591-597.

[8] Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K. Y., & Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

[9] Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7), 799-815.

[10] Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130), 295-305.

[11] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31(4), 1208-1211.

[12] Hardin, J. W., & Hilbe, J. M. (2002). *Generalized Estimating Equations*. Chapman and Hall/CRC, London.

[13] Heyde, C. C.(1997). *Quasi-likelihood and its Application: A General Approach to Optimal Parameter Estimation*. Springer, New York.

[14] Horton, N. J., & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *The American Statistician*, 53(2), 160-169.

[15] Johnson, N. L., & Kotz, S. (1970). *Distributions in Statistics: Continuous Unvariate Distributions-1*. Houghton Mifflin.

[16] Jørgensen, B., Seshadri, V., & Whitmore, G. A. (1991). On the mixture of the inverse Gaussian distribution with its complementary reciprocal. *Scandinavian Journal of Statistics*, 77-89.

[17] Jørgensen, B. (1992). Exponential dispersion models and extensions: A review. *International Statistical Review/Revue Internationale de Statistique*, 5-20.

[18] Jorgensen, B. (1997). *The Theory of Dispersion Models*. CRC Press.

[19] Kedem, B., & Fokianos, K. (2005). *Regression Models for Time Series Analysis* (488). John Wiley & Sons.

[20] Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.

[21] McCullagh, P., Nelder, JA. (1989),. *Generalized Linear Models*. Chapman and Hall, London.

[22] Meyers, S. M., Ambler, J. S., Tan, M. I. N. G., Werner, J. C., & Huang, S. S. (1992). Variation of perfluoropropane disappearance after vitrectomy. *Retina (Philadelphia, Pa.)*, 12(4), 359-363.

[23] Michael, J. R., Schucany, W. R., & Haas, R. W. (1976). Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2), 88-90.

[24] Mudholkar, G. S., & Natarajan, R. (2002). The inverse Gaussian models: analogues of symmetry, skewness and kurtosis. *Annals of the Institute of statistical Mathematics*, 54(1), 138-154.

[25] Mudholkar, G. S., & Wang, H. (2007). IG-symmetry and R-symmetry: interrelations and applications to the inverse Gaussian theory. *Journal of Statistical Planning and Inference*, 137(11), 3655-3671.

[26] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370-384.

[27] Paik, M. C. (1992). Parametric variance function estimation for nonnormal repeated measurement data. *Biometrics*, 19-30.

[28] Park, T., & Shin, D. Y. (1999). On the use of working correlation matrices in the GEE approach for longitudinal data. *Communications in Statistics-Simulation and Computation*, 28(4), 1011-1029.

[29] Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 1033-1048.

[30] Prentice, R. L., & Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 825-839.

[31] Qiu, Z. (2001). *Simplex Mixed Models for Longitudinal Proportional Data*, Ph.D. Dissertation, York University

[32] Qiu, Z., Song, P. X. K., & Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, 35(4), 577-596.

[33] Seshadri, V. (1993),. *The Inverse Gaussian Distribution*. New York.

[34] Seshadri, V .(1999),. *The Inverse Gaussian Distribution: Statistical Theory and Applications*(137). Springer Science& Business Media New York

[35] Shuster, J. (1968). On the inverse Gaussian distribution function. *Journal of the American Statistical Association*, 63(324), 1514-1516.

[36] Song, P. X. K., & Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics*, 56(2), 496-502.

[37] Song, P. X. K., Qiu, Z., & Tan, M. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. Biometrical Journal: *Journal of Mathematical Methods in Biosciences*, 46(5), 540-553.

[38] Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, 61(3), 439-447.

[39] Yang, H., Acker, J. P., Cabuhat, M., Letcher, B., Larratt, L., & McGann, L. E. (2005). Association of post-thaw viable CD34+ cells and CFU-GM with time to hematopoietic engraftment. *Bone Marrow Transplantation*, 35(9), 881.

[40] Zeileis, A., Cribari-Neto, F., Grün, B., & Kos-midis, I. (2010). Beta regression in R. *Journal of Statistical Software*, 34(2), 1-24.

[41] Zhang, P., & Qiu, Z. (2014). Regression analysis of proportional data using simplex distribution. *Scientia Sinica Mathematica*, 44(1), 89-104.

[42] Zhang, P., Qiu, Z., & Shi, C. (2016). simplexreg: An r package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Software*, 71(11), 1-21.

[43] Zorn, C. J. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 470-490.

Abstract

In regression problems, Logistic and Probit models are commonly used for proportional response variable in unit interval $(0, 1)$. But when the distribution of proportional response is highly skewed, these two models are not suitable, because their underlying distribution *i.e.* Logistic and Normal are symmetric, hence a contradiction is made. In this thesis, "Simplex regression" is introduced as an alternative solution. Firstly, Simplex distribution based on mixture inverse Gaussian distribution is introduced. Secondly, Simplex simulation studies are made to assert the capabilities of the distribution. As Simplex regression can be defined on the framework of "Dispersion models", we explain it first, thereafter, similar to generalized linear models methods, the regression parameters are estimated in two ways: heterogeneous and homogeneous variance. Modeling the longitudinal data by Simplex regression is the last subject in this note. All models are also evaluated by simulation and real examples.

Keywords: Dispersion models, Proportional data, Random variable generation, Generalized linear models, Simplex distribution, Simplex regression models, Longitudinal data, Generalized estimating equations.

Faculty Of Mathematical Sciences

MSc Thesis in Statistics

Regression Analysis of Proportion Data Using the Simplex Distribution

By: Sima Maleki

Supervisor:

Davood Shahsavani

January 2019