مجله علوم آماری، بهار و تابستان ۱۳۹۷
جلد ۱۲، شماره ۱، ص ۷۳ – ۹۶
DOI: 10.29252/jss.12.1.73

رگرسیون ضرایب متغیر طولی حاشیه‌ای

حسین بهرامی چشمه‌علی، آرش اردلان

گروه ریاضی، دانشگاه یاسوج
تاریخ دریافت: ۱۳۹۵/۱۱/۱۴ تاریخ آخرین بازنگری: ۱۳۹۷/۱/۲۶

چکیده: مدل‌های رگرسیونی ناپارامتری و نیمه‌پارامتری در زمینه داده‌های مستقل توسعه چشمگیری پیداکرده‌اند، اما رشد آن‌ها در زمینه داده‌های طولی، محدود به چند سال اخیر است. از آنجا که روش‌های رگرسیونی معمول برای داده‌های همبسته نسبت به داده‌های مستقل توانایی کمتری دارند، باید از مدل‌هایی استفاده شود، که همبستگی بین داده‌ها را نیز در نظر بگیرند. در این میان مدل‌های آمیخته و حاشیه‌ای که عامل همبستگی بین داده‌ها را نیز در نظر می‌گیرند، مدل‌هایی هستند که برای برازش داده‌های طولی مورد استفاده قرار می‌گیرند. همچنین با توجه به انعطاف‌پذیری مدل‌های نیمه‌پارامتری نسبت به مدل‌های پارامتری و ناپارامتری، مدل رگرسیون نیمه‌پارامتری طولی حاشیه‌ای با برآوردهای اسپلاین تاوانیده مدل مناسبی برای تحلیل داده‌های طولی است. در این مقاله رگرسیون نیمه‌پارامتری با ضرایب متغیر که در آن ارتباط بین متغیر پاسخ و یک متغیر پیش‌بین بر مبنای متغیر پیش‌بین دیگر مشخص می‌شود، بررسی شده است. همچنین استنباط بیزی برای مدل ناپارامتری روی‌داده‌های شبیه‌سازی شده و برای مدل نیمه‌پارامتری طولی حاشیه‌ای روی داده‌های واقعی، با نرم‌افزارهای استاندارد انجام شده است که نشان‌دهنده عملکرد قابل‌قبول این استنباط است.

واژه‌های کلیدی: نمونه‌گیری گیبز، مدل آمیخته، مدل حاشیه‌ای، مدل گرافیکی، مدل بیزی سلسله مراتبی.

آدرس الکترونیکی نویسنده مسئول مقاله: آرش اردلان، a.ardalan@yu.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62F15، 62M05، 62J12

۱ مقدمه

مدل رگرسیونی یکی از پرکاربردترین ابزارهای آماری است که به دلیل تنوع زیاد مطالعات، روش‌های گوناگونی برای انجام آن پدید آمده است. از یک منظر رگرسیون را می‌توان به سه بخش پارامتری، ناپارامتری، نیمه‌پارامتری تقسیم کرد. رگرسیون پارامتری بیشتر وقتی مورداستفاده قرار می‌گیرد که اطلاعات تحقیق به‌قدر کافی زیاد باشد تا آنجایی که بتوان ارتباط بین متغیر پاسخ و پیش‌بین را به وسیله یک مدل خطی یا قابل‌تبدیل به خطی بیان کرد. رگرسیون ناپارامتری در مواقعی که نتوان نوع ارتباط بین متغیرهای پیش‌بین و پاسخ را با یک شکل تابعی معلوم و مشخص بیان کرد، مورد استفاده قرار می‌گیرد، که در آن می‌توان از برآوردهای موضعی مانند اسپلاین استفاده کرد. مدل‌های نیمه‌پارامتری که در این پژوهش تمرکز بیشتر بر آنها است، ترکیبی از مدل‌های پارامتری و ناپارامتری است، به این معنی که بعضی از متغیرها دارای روند معلوم تغییرات نسبت به متغیر پاسخ بوده و برای بعضی از متغیرها نمی‌توان شکل تابعی خاصی را در نظر گرفت. مدل‌های نیمه‌پارامتری ازآنجایی‌که شکل تابعی پیچیده‌تری نسبت به مدل‌های پارامتری و ناپارامتری دارند، مسئله برآورد یابی برای آن‌ها نیز پیچیدگی بیشتری دارد که برای مثال می‌توان به آمیخته شدن این مسئله با مفاهیم محض ریاضی همچون فضای هیلبرت اشاره کرد. ازآنجا که فرم‌های تابعی نامعلوم در مدل‌های ناپارامتری و نیمه‌پارامتری اغلب با روش اسپلاین برآورد می‌شوند و نوع داده‌ها نیز طولی است، یک روش منطقی برای بررسی این مدل‌ها استفاده از ارتباط مدل‌های آمیخته و مدل‌های اسپلاین است. گرچه این مسئله تا حدودی برآوردیابی را در حوزه مدل‌های نیمه پارامتری آسان می‌کند اما باید توجه داشت که برآوردیابی برای مدل‌هایی با فرض‌های ناگوسی با این روش نیز کمی دچار مشکل است و امکان دارد برای حجم نمونه پایین اطلاعات حاصل اریب نیز باشد . در علوم نوین آمار و همچنین در آمار بیزی برای حل این مشکل از الگوریتم‌های نمونه‌گیری مانند مونت‌کارلوی زنجیر مارکوفی[1] (MCMC) استفاده می‌کنند. در این مقاله نیز، ماتریس کوواریانس مدل حاشیه‌ای نیمه‌پارامتری به این روش برآورد شده است. یک مسئله قابل‌توجه ادغام مدل‌های حاشیه‌ای و نیمه‌پارامتری برای بررسی داده‌های طولی است که از مزایای آن می‌توان به موارد زیر اشاره کرد:

۱- برآورد انعطاف‌پذیر از توابع رگرسیونی که با مدل‌های ساده رگرسیونی امکان‌پذیر نیست.

۲- برآورد حاشیه‌ای از ماتریس کوواریانس، که در بحث داده‌های طولی دارای اهمیت زیادی است.

این مسئله با استفاده از روش برآوردیابی اسپلاین تاوانیده، توسط الخدیری و همکاران (۲۰۱۰) مورد بررسی

[1] Markov Chain Monte Carlo

قرارگرفته است. همچنین زگر و دیگل (۱۹۹۴) مرجع اولیه برای رگرسیون ناپارامتری طولی حاشیه‌ای فراهم ساخته‌اند. کارل و همکاران (۲۰۰۹) بر روی تأثیر برآوردهای حاشیه‌ای نیمه‌پارامتری برای مدل خطی جمعی جزئی درداده‌های طولی کارکرده‌اند. فان و هوانگ (۲۰۰۷) تحلیل داده‌های طولی با برآورد نیمه‌پارامتری تابع کوواریانس را موردمطالعه قرار داده‌اند. وانگ و کارل (۲۰۰۵) کارآیی برآورد حاشیه‌ای نیمه‌پارامتری بر روی‌داده‌های دسته‌بندی‌شده طولی را مطالعه کرده‌اند. ولهام و همکاران (۲۰۰۷) مقایسه‌ای از مدل‌های آمیخته با اسپلاین تاوانیده برای منحنی برازش داده‌شده انجام داده‌اند. ژائو و همکاران (۲۰۰۶) در مورد طرح‌های بیزی برای مدل‌های آمیخته‌ی خطی به بحث کرده‌اند.

در بخش ۲ مروری بر دیدگاه بیز سلسله مراتبی خواهد شد. در بخش ۳ به معرفی مدل ناپارامتری طولی حاشیه‌ای و داده‌های شبیه‌سازی شده برای آن خواهیم پرداخت. در بخش ۴ مدل طولی حاشیه‌ای ضرایب متغیر که اساس کار این پژوهش است، موردبررسی قرار می‌گیرد. در بخش ۵ برآورد ماکسیمم درستنمایی برای برآورد ضرایب، مدل را مورد مطالعه قرار خواهیم داد و در بخش ۶ بابیان محدودیت‌های این برآورد، مقدمه‌ای برای استفاده از استنباط بیزی را فراهم خواهیم آورد. در بخش‌های ۷ و ۸ ابتدا با استفاده از مدل ناپارامتری و داده‌های شبیه‌سازی شده، دقت برآورد بیزی مؤلفه‌های واریانس ارزیابی می‌شود، سپس مدل نیمه‌پارامتری طولی حاشیه‌ای تحلیل خواهد شد.

۲ مدل بیز سلسله مراتبی و مدل گرافیکی

مدل رگرسیون خطی ساده به صورت

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2) \qquad (1)$$

را در نظر بگیرید. اگر بخواهیم اطلاعات اولیه مدل (۱) برای یک استنباط بیز سلسله مراتبی بیان شود، می‌توان آن را به‌صورت

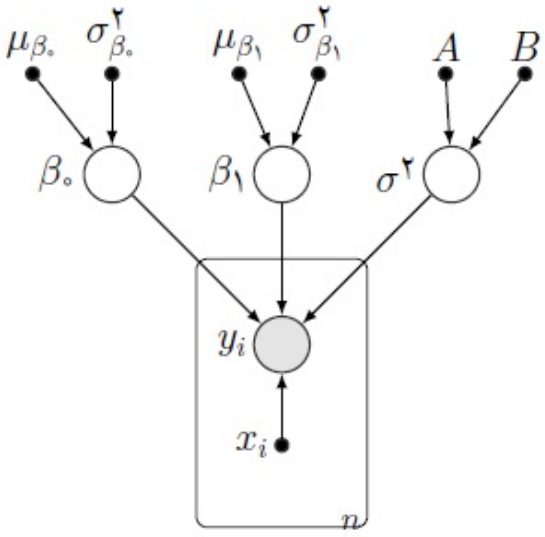
$$\begin{aligned}
&[y_i|\beta_0, \beta_1, \sigma^2] \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \\
&[b_0] \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0}) \quad [b_1] \sim N(\mu_{\beta_1}, \sigma^2_{\beta_1}) \\
&[\sigma^2] \sim IG(A, B)
\end{aligned}$$

نوشت، که در آن $\{\mu_{\beta_0}, \sigma^2_{\beta_0}, \mu_{\beta_1}, \sigma^2_{\beta_1}, A, B\}$ ابر پارامترهای مدل هستند. با افزایش پارامترهای مدل، مدل بیز سلسله مراتبی شکل پیچیده‌تری پیدا می‌کند. برای سهولت در فهم ارتباطات مدل‌های بیز سلسله مراتبی می‌توان از مدل‌های گرافیکی بهره برد. مدل‌های گرافیکی شاخه‌ای در علوم کامپیوتر و ریاضیات به‌حساب می‌آیند که ترکیبی از نظریه گراف و احتمال حاصل می‌شود. این مدل‌ها تجسم مدل احتمالی را ساده‌تر کرده و برای تعیین روابط بین متغیرها و بیان کارآمدی مدل، مورد استفاده قرار می‌گیرند. دو روش معمول مدل‌های گرافیکی عبارتنداز:

(۱) گراف غیر مدور مستقیم[۲] (DAGs) که با شبکه بیزی نیز شناخته می‌شود.

(۲) گراف غیرمستقیم که با میدان مارکوف تصادفی نشان داده می‌شود.

برای مدل‌های سلسله مراتبی بیزی اغلب از مدل گرافیکی DAGs استفاده می‌شود، که فرم آن در شکل ۱ نشان داده شده است.



شکل ۱. مدل گرافیکی برای مدل بیزی سلسله مراتبی (۲)

[2]Directed acyclic graph

دایره‌های توپر نشان‌دهنده مقادیر ثابت مدل، در اینجا x_iها و ابر پارامترها، هستند. دایره سایه زده شده مربوط به متغیر y، نشان دهنده این موضوع است که این متغیر دارای مقادیر مشاهده شده است. درنهایت نیز دایره‌های توخالی، نشان‌دهنده پارامترهای مدل هستند که بر اساس روش‌های بیزی مورد استنباط قرار می‌گیرند. همچنین n نشان‌دهنده تعداد مشاهدات برای این مدل است. در ادامه از این مدل‌ها برای روشن نمایی مدل‌های بیز سلسله مراتبی در مدل‌های طولی حاشیه‌ای ناپارامتری و نیمه پارامتری استفاده خواهیم کرد.

۳ مدل ناپارامتری طولی حاشیه‌ای

برای $1 \le i \le m$ آزمودنی، $1 \le j \le n$ $(n << m)$ مشاهده متناظر با متغیر پاسخ y_{ij} و متغیر پیش‌بین x_{ij} را در نظر بگیرید. همچنین فرض کنید $\boldsymbol{y_i}$ و $\boldsymbol{x_i}$ به ترتیب بردار پاسخ و بردار پیش‌بین‌ها برای آزمودنی iام است. یک رگرسیون ناپارامتری طولی حاشیه‌ای[3] به‌صورت

$$E(Y_{ij}) = f(x_{ij}), \quad Cov\{\boldsymbol{Y_i}|f(\boldsymbol{x_i})\} = \boldsymbol{\Sigma} \ , 1 \le i \le m, \ 1 \le j \le n \qquad (۲)$$

تعریف می‌شود، که در آن f تابع هموار و $\boldsymbol{\Sigma}$ ماتریس کوواریانس است. شکل ۲ نشان‌دهنده داده‌های شبیه‌سازی شده برای مدل (۲)، با $m = 100$ ، $n = 2$ و

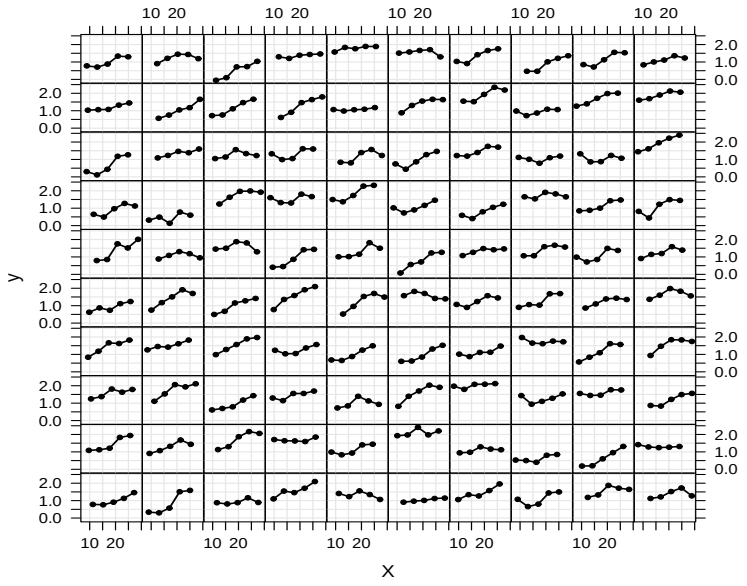
$$f(x) = 1 + \frac{1}{2}\phi((2x - 36)/5) \qquad (۳)$$

و ماتریس کوواریانس

$$\boldsymbol{\Sigma} = \begin{pmatrix} ۰/۱۲۲ & ۰/۰۹۸ & ۰/۰۷۳ & ۰/۰۶۳ & ۰/۰۵ \\ ۰/۰۹۸ & ۰/۱۲۲ & ۰/۰۹۸ & ۰/۰۷۳ & ۰/۰۶۳ \\ ۰/۰۷۳ & ۰/۰۹۸ & ۰/۱۲۲ & ۰/۰۹۸ & ۰/۰۷۳ \\ ۰/۰۶۳ & ۰/۰۷۸ & ۰/۰۹۸ & ۰/۱۲۲ & ۰/۰۹۸ \\ ۰/۰۵ & ۰/۰۶۳ & ۰/۰۷۸ & ۰/۰۹۸ & ۰/۱۲۲ \end{pmatrix} \qquad (۴)$$

است، که در آن ϕ تابع چگالی نرمال استاندارد است.

[3] Marginal longitudinal nonparametric regression



شکل ۲. داده‌های شبیه‌سازی شده برای مدل (۲) با در نظر گرفتن فرم تابعی (۳) و ماتریس کوواریانس (۴)

مسئله اصلی پیدا کردن برآوردی برای f و در نهایت برای $\boldsymbol{\Sigma}$ است. یک روش مناسب برای برآورد f، استفاده از اسپلاین به‌صورت

$$f(x) = \beta_{\circ} + \beta_{1} x + \sum_{k=1}^{K} u_k Z_k(x) \qquad (۵)$$

است، که در آن $Z_1, \ldots, Z_k$ یک مجموعه مناسب از توابع پایه‌ای اسپلاین هستند. یک‌پایه ساده با قرار دادن

$$Z_k(x) = (x - c_k)_+ = \begin{cases} x - c_k & \text{اگر} \quad x - c_k > \circ \quad \text{یا} \quad x > c_k \\ \circ & \text{اگر} \quad x - c_k < \circ \quad \text{یا} \quad x < c_k \end{cases}$$

حاصل می‌شود، به‌طوری‌که $c_1, ..., c_k$ یک نمایش از مکان گره‌ها روی دامنه x_iها هستند که در بیشتر اوقات $k = ۲۵$ برای مدل کفایت می‌کند (واند و اورمرود، ۲۰۰۸). برای پرهیز از بیش برازشی احتیاج به جریمه ضرایب مدل اسپلاین داریم. یک جریمه مناسب در نظر گرفتن ضرایب اسپلاین، یعنی u_kها،

به‌عنوان متغیرهای تصادفی از یک توزیع نرمال با میانگین صفر و واریانس σ^2 است. حال با اعمال این جریمه بر روی معادله (۵)، یک نمایش ماتریسی از مدل (۲) را می‌توان به‌صورت

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \tag{۶}$$

نشان داد، که در آن

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y_m} \end{pmatrix} \ , \ \boldsymbol{X} = \begin{pmatrix} 1 & \boldsymbol{x}_1 \\ \vdots & \vdots \\ 1 & \boldsymbol{x_m} \end{pmatrix} \ , \ \boldsymbol{Z} = \begin{pmatrix} Z_1(\boldsymbol{x}_1) & \dots & Z_K(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ Z_1(\boldsymbol{x_m}) & \dots & Z_K(\boldsymbol{x_m}) \end{pmatrix}$$

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_K \end{pmatrix} \ , \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \ , \ \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon_m} \end{pmatrix} \tag{۷}$$

بردارهای تصادفی در سمت راست معادله (۶) دارای میانگین صفر و ماتریس کوواریانس زیر است:

$$Cov \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon_m} \end{pmatrix} = \begin{pmatrix} \sigma^2 I & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Sigma} & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{\Sigma} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \boldsymbol{\Sigma} \end{pmatrix} = \begin{pmatrix} \sigma^2 I & 0 \\ 0 & I_m \otimes \boldsymbol{\Sigma} \end{pmatrix}$$

که در آن نماد $\otimes$ نشان دهنده ضرب کرونکر است، همچنین اگر σ^2 و $\boldsymbol{\Sigma}$ معلوم در نظر گرفته شوند، می‌توان از روش بهترین پیش‌بینی نااریب خطی [۴] (BLUP) برای برآورد $\boldsymbol{\beta}$ و $\boldsymbol{u}$ استفاده کرد. در عمل σ^2 و $\boldsymbol{\Sigma}$ نیز هر دو باید برآورد شوند و فرض مناسب برای رسیدن به این هدف این است که

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \ , \ \begin{pmatrix} \sigma^2 I & 0 \\ 0 & I_m \otimes \boldsymbol{\Sigma} \end{pmatrix} \right)$$

[4] Best linear unbiased prediction

۴ مدل ضرایب متغیر طولی حاشیه‌ای

یک نوع از رگرسیون نیمه‌پارامتری با پیش‌بین‌های چندگانه یک مدل با ضرایب گوناگون [۵] است. فرض کنید برای هر y_{ij} پیش‌بین‌های جداگانه‌ای موجود باشد. در این قسمت با حفظ نمادهای بخش قبل، مدل را به دو متغیر کمی برای jامین تکرار از آزمودنی iام که به‌صورت x_{ij} و s_{ij} نمایش داده می‌شود، محدود می‌کنیم. مدل ضرایب متغیر طولی حاشیه‌ای برای داده‌های تعریف‌شده در بخش قبل به‌صورت

$$E(Y_{ij}) = f_{\circ}(s_{ij}) + f_{\text{۱}}(s_{ij})x_{ij}\ , \quad Cov\{\boldsymbol{Y_i}|f_{\circ}(\boldsymbol{s_i}), f_{\text{۱}}(\boldsymbol{s_i})\} = \boldsymbol{\Sigma} \qquad (\text{۸})$$

تعریف می‌شود، که در آن $f_{\circ}$ و $f_{\text{۱}}$ توابع هموار هستند. همان‌طور که از ساختار مدل مشخص است، این مدل به‌گونه‌ای است که برای هر مقدار ثابت s، یک ارتباط خطی بین y و x بیان می‌کند. حال اگر توابع $f_{\circ}$ و $f_{\text{۱}}$ با روش اسپلاین تخمین زده شود، مدل (۸) به‌صورت

$$\begin{aligned} y_{ij} \quad = \quad & \alpha_{\circ} + \alpha_{\text{۱}} s_{ij} + \sum_{k=\text{۱}}^{K_{\text{۱}}} u_{\text{۱}k}(s_{ij} - c_k)_+ \\ & + \{\beta_{\circ} + \beta_{\text{۱}} s_{ij} + \sum_{k=\text{۱}}^{K_{\text{۲}}} u_{\text{۲}k}(s_{ij} - c_k)_+\} x_{ij} + \epsilon_{ij} \end{aligned} \qquad (\text{۹})$$

حاصل می‌شود. اگر $u_{\text{۱}}$ و $u_{\text{۲}}$ متغیرهای تصادفی با توزیع‌های

$$u_{\text{۱}k} \overset{iid}{\sim} N(\text{۰}, \sigma_{\text{۱}}^{\text{۲}}), \quad u_{\text{۲}k} \overset{iid}{\sim} N(\text{۰}, \sigma_{\text{۲}}^{\text{۲}})$$

در نظر گرفته شوند، می‌توان مدل (۹) را به‌صورت مدل آمیخته

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$$

در نظر گرفت، که در آن
$\boldsymbol{\beta} = [\alpha_{\circ}, \alpha_{\text{۱}}, \beta_{\circ}, \beta_{\text{۱}}]^T$، $\boldsymbol{X} = [\text{۱}, \boldsymbol{s}, \boldsymbol{x}, \boldsymbol{sx}]$، $\boldsymbol{u} = [u_{\text{۱۱}}, ..., u_{\text{۱}K_{\text{۱}}}, u_{\text{۲۱}}, ..., u_{\text{۲}K_{\text{۲}}}]^T$ و

[5] Varying cofficient models

$$\boldsymbol{Z} = [\underset{۱\leq k\leq K_۱}{(s_{ij}-c_k)_+} \quad \underset{۱\leq k\leq K_۲}{(s_{ij}-c_k)_+ x_{ij}}]_{۱\leq i\leq m, ۱\leq j\leq n}.$$

ماتریس کوواریانس برای ضرایب اسپلاین و خطای مدل به‌صورت

$$\begin{pmatrix} \boldsymbol{u}_۱ \\ \boldsymbol{u}_۲ \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} ۰ \\ ۰ \\ ۰ \end{pmatrix} , \begin{pmatrix} \sigma_۱^۲ I & ۰ & ۰ \\ ۰ & \sigma_۲^۲ I & ۰ \\ ۰ & ۰ & I_m \otimes \boldsymbol{\Sigma} \end{pmatrix} \right)$$

تعریف می‌شود، که در آن $\boldsymbol{u}_۱$ یک بردار $K_۱ \times ۱$ بعدی شامل $u_{۱k}$ها است و $\boldsymbol{u}_۲$ نیز به‌صورت مشابه تعریف می‌شود.

۵ برآورد ماکسیمم درستنمایی و بهترین پیش‌بینی

در مبحث مدل‌های حاشیه‌ای مسئله اصلی برآورد مؤلفه‌های واریانس است که اغلب به روش ماکسیمم درستنمایی مقید انجام می‌شود. در حالت کلی تعمیم مدل‌های (۲) و (۸) به مدل‌هایی با d تابع هموارساز را می‌توان به مدل آمیخته گاوسی به صورت

$$\boldsymbol{y}|\boldsymbol{u} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zu}, I_m \otimes \boldsymbol{\Sigma}), \;\; \boldsymbol{u} \sim N(۰, \underset{۱\leq r\leq d}{blockdiag}(\sigma_r^۲ I_{k_r})) \qquad (۱۰)$$

تبدیل کرد، که در آن K_r متناظر با تعداد توابع پایه اسپلاین مورد کاربرد در rامین تابع هموارساز برآورد شده است. فرض کنید $\boldsymbol{\sigma}^۲ = (\sigma_۱^۲, ..., \sigma_d^۲)$ بردار پارامترهای واریانس باشد. لگاریتم درستنمایی مدل (۱۰) به‌صورت

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^۲, \boldsymbol{\Sigma}) = -\frac{۱}{۲}\{n\log(۲\pi) + \log|V| + (\boldsymbol{y}-\boldsymbol{X\beta})^T V^{-۱}(\boldsymbol{y}-\boldsymbol{X\beta})\}$$

حاصل می‌شود، به‌طوری‌که

$$V = Cov(\boldsymbol{Y}) = \sum_{r=۱}^{d} \sigma_r^۲ Z_{[r]} Z_{[r]}^T + I_m \otimes \boldsymbol{\Sigma}$$

که در آن $Z_{[1]}, ..., Z_{[d]}$ افرازهایی از Z هستند که متناظر با توابع پایه برای هر تابع هموارساز است. با توجه به برآوردیابی در مدل‌های آمیخته برای مقادیر ثابت $\boldsymbol{\sigma}^2$ و $\boldsymbol{\Sigma}$، برآورد ضرایب ثابت به‌صورت

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) = (\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V^{-1} \boldsymbol{y}$$

حاصل می‌شود. بر اساس این برآورد لگاریتم درستنمایی نیمرخ به‌صورت

$$\ell_p(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) = -\frac{1}{2}\{\log|V| + \boldsymbol{y}\, V^{-1}\{I - \boldsymbol{X}(\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T V^{-1}\}\boldsymbol{y}\} - \frac{n}{2}\log(2\pi)$$

خواهد شد. البته در عمل اغلب لگاریتم درستنمایی مقید به‌صورت

$$\ell_R(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) = \ell_p(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) - \frac{1}{2}\log|\boldsymbol{X}^T V^{-1} \boldsymbol{X}|$$

مورد استفاده قرار می‌گیرد، زیرا برآورد حاصل از روش ماکسیمم درستنمایی مقید ،(REML) این نکته که $\boldsymbol{\beta}$ها برآورد شده‌اند را نیز در نظر می‌گیرد (روپرت و همکاران، ۲۰۰۳). پس مسئله برآوردیابی برای پارامترهای مدل، یعنی $\boldsymbol{\beta}$ و $\boldsymbol{\sigma}^2$ و $\boldsymbol{\Sigma}$ به شرح زیر است:

۱- به دست آوردن برآورد REML برای $\boldsymbol{\sigma}^2$ و $\boldsymbol{\Sigma}$ با استفاده از بیشینه کردن ℓ_R.

۲- به دست آوردن برآورد ماکسیمم درستنمایی (MLE) برای $\boldsymbol{\beta}$ به‌صورت $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\Sigma}})$

و درنهایت مسئله برآورد ضرایب اسپلاین، یعنی $\boldsymbol{u}$، مطرح است. چون $\boldsymbol{u}$ یک بردار تصادفی است، نمی‌توان از برآوردهایی مانند ماکسیمم درستنمایی استفاده کرد و معمولاً روش بهترین پیش‌بینی مورد استفاده قرار می‌گیرد. بنا بر مدل (۱۰) این برآورد به‌صورت

$$\tilde{\boldsymbol{u}}(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) \equiv E(\boldsymbol{Y}|\boldsymbol{u}) \equiv \boldsymbol{G}_{\sigma^2} \boldsymbol{Z}^T V(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma})^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}))$$

است، به‌طوری‌که $\boldsymbol{G}_{\sigma^2} = \underset{1 \le r \le d}{blockdiag}(\sigma_r^2 I_{k_r})$ ازآنجایی‌که مقادیر $\boldsymbol{\beta}$ و $\boldsymbol{\sigma}^2$ و $\boldsymbol{\Sigma}$ نامعلوم هستند و باید مورد برآورد قرار گیرند، یک برآورد مناسب برای $\tilde{\boldsymbol{u}}$، بهترین پیش‌بینی تجربی به‌صورت

$$\hat{\boldsymbol{u}} = \boldsymbol{G}_{\hat{\sigma}^2} \boldsymbol{Z}^T V(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\Sigma}})^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(\hat{\boldsymbol{\sigma}}^2, \hat{\boldsymbol{\Sigma}}))$$

است. حال با استفاده از این برآوردها می‌توان توابع رگرسیونی f را به‌سادگی به دست آورد و برازش رگرسیونی را انجام داد.

ممکن است این تصور پیش آید که شاید بتوان مدل (۱۰) را به‌عنوان یک مدل آمیخته در نظر گرفت و آن را با توابعی همانند lme() در نرم‌افزار R برآورد کرد که در جواب باید گفت این موضوع با محدودیت‌های تابع lme() قابل انجام نیست. بنابراین با تمامی محدودیت‌های موجود در نظر گرفتن استنباط بیزی و اجرای آن توسط فرایند گیبز شاید اجتناب‌ناپذیر باشد. اگرچه برآورد نقطه‌ای دارای مفهومی ساده‌تر و فرایند انجام آن سریع‌تر است، اما باید به این نکته توجه کرد که این برآورد به مدل‌های آمیخته گاوسی وابستگی شدید داشته و در مواردی که با مدل‌های ناگوسی روبرو هستیم، ازلحاظ محاسباتی بسیار مشکل خواهد بود. همچنین برای به دست آوردن توزیع برآوردهای REML، جهت ساختن بازه اطمینان با توجه به قانون اعداد بزرگ نیازمند مشاهدات زیادی هستیم. این محدودیت سبب پایین آمدن دقت برآورد و نامعتبر بودن بازه اطمینان می‌گردد. از طرف دیگر روش MCMC به توزیع گاوسی وابسته نیست و با افزایش تعداد تکرار الگوریتم، دقت برآورد توزیع پسین و استنباط‌های مورد نظر بالاتر رفته و بازه اطمینان نیز قابل‌اعتمادتر می‌شوند (وست و همکاران، ۲۰۰۶)

۶ استنباط بیزی

در این قسمت سعی بر آن است که با استفاده از برآورد بیزی ماتریس کوواریانس مدل حاشیه‌ای را به دست آورده و استنباط‌های موردنظر را انجام دهیم. بر این اساس برای پارامترهای مدل آمیخته (۱۰) توزیع‌های پیشین

$$\boldsymbol{\beta} \sim N(\circ, \quad \boldsymbol{F}), \;\; \sigma_r^{\tau} \sim IG(A_r, B_r), \;\; \boldsymbol{\Sigma} \sim IW(a, \boldsymbol{B})$$

پیشنهاد می‌شود، که در آن A_r و B_r $(\text{۱} \leq r \leq d)$ مقادیر ثابت و مثبت، $\boldsymbol{F}$ و $\boldsymbol{B}$ ماتریس‌های معین مثبت ابرپارامتر و IG و IW بترتیب نمادهای توزیع‌های گامای وارون و ویشارت وارون با توابع چگالی

$$[\sigma_r^{\tau}] \;=\; \frac{{B_r}^{A_r}}{\Gamma(A_r)} \, (\sigma_r^{\tau})^{-A-\text{۱}} \, e^{-B_r/\sigma_r^{\tau}}$$

$$[\boldsymbol{\Sigma}] \;=\; C_{n,a}^{-1}\,|\boldsymbol{B}|^{a/2}\,|\boldsymbol{\Sigma}|^{-(a+n+1)/2}\exp\{-\frac{1}{2}tr(\boldsymbol{B}\boldsymbol{\Sigma}^{-1})\}\;,\quad a>0$$

هستند، به‌طوری‌که $C_{n,a}=2^{an/2}\,\pi^{n(n-1)/4}\,\prod_{i=1}^{n}\,\Gamma(\frac{a+1-i}{2})$. برای انجام استنباط بیزی نیاز به تعیین توزیع‌های پسین

$$[\boldsymbol{\beta}|\boldsymbol{y}],\quad [\boldsymbol{u}|\boldsymbol{y}],\quad [\boldsymbol{\Sigma}|\boldsymbol{y}],\quad [\boldsymbol{\sigma}_{\boldsymbol{r}}^{2}|\boldsymbol{y}]$$

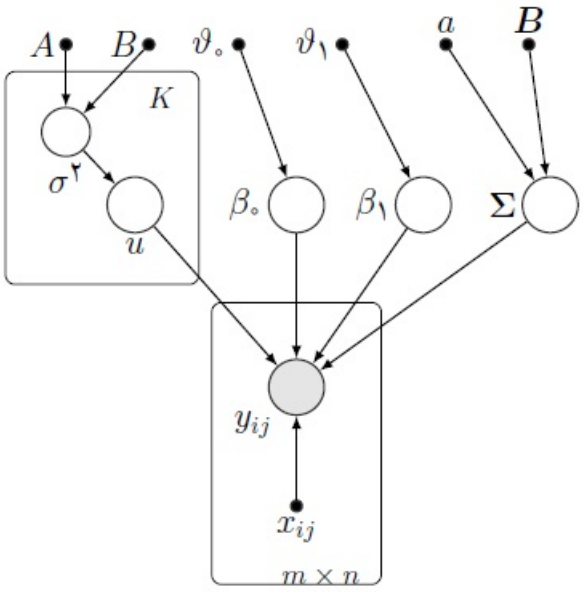
است که به دست آوردن آن‌ها مشکل و حتی شاید غیرممکن است. معمولاً این توزیع‌ها را می‌توان با روش‌های محاسباتی تصادفی و گاهی غیر تصادفی تخمین زد (مینیستاسا و واند، ۲۰۱۳). دراین‌بین روش MCMC یکی از روش‌های معمول تصادفی محاسباتی است که در ادامه با نرم‌افزارهای OpenBUGS و R انجام داده خواهد شد. برای انجام محاسبات بیزی انتخاب ابر پارامترها که تأثیر زیادی بر توزیع پسین خواهد گذاشت، بسیار مهم است. اگر در مورد توزیع پیشین اطلاعاتی در دسترس باشد، می‌توان ابر پارامترها را بر مبنای این اطلاعات انتخاب نمود. اما اغلب این اطلاعات موجود نیست، بنابراین انتخاب ابر پارامترها به‌گونه‌ای است که منجر به توزیع‌های پیشین ناآگاهی‌بخش شوند. بر این اساس ابر پارامترها به‌صورت زیر پیشنهاد می‌شوند:

$$\boldsymbol{F}=10^{8}I\;\;,\;\;A_r=B_r=0/01\;\;,\;\;a=n,\;\;\boldsymbol{B}=0/01I_n$$

حال برای استفاده از الگوریتم نمونه‌گیری گیبز به توزیع‌های تمام شرطی پارامترها نیاز است که با توجه به توزیع‌های پیشین، به‌صورت زیر به دست می‌آیند:

$$\begin{aligned}
&\boldsymbol{\Sigma}|rest\sim IW(a+m,\;\boldsymbol{B}+(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u})^{T})\\
&\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}|rest\sim N((\boldsymbol{C}(I_m\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{C}+\boldsymbol{G}_{\sigma^2})^{-1}\boldsymbol{C}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{y},(\boldsymbol{C}(I_m\otimes\boldsymbol{\Sigma}^{-1})\boldsymbol{C}+\boldsymbol{G}_{\sigma^2})^{-1})\\
&\sigma_r^2|rest\;\sim\;IG(A_r+\frac{1}{2}K_r,B_r+\frac{1}{2}||\boldsymbol{u}_r||^2)
\end{aligned}$$

که در آن $\boldsymbol{C}=[\boldsymbol{X}\;\boldsymbol{Z}]$.



شکل ۳. مدل گرافیکی برای مدل بیزی سلسله مراتبی (۱۱)

۷ مدل ناپارامتری برای داده‌های شبیه‌سازی‌شده

در این بخش با استفاده از اطلاعات بخش‌های گذشته و داده‌های شبیه‌سازی شده در بخش ۳ دقت برآورد بیزی مؤلفه‌های واریانس مدل حاشیه‌ای، سنجیده می‌شود. برای بیان مدل ناپارامتری (۲) به‌صورت یک مدل بیزی سلسله مراتبی، داریم

$$
\begin{aligned}
&[\boldsymbol{y_i}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}, \boldsymbol{\Sigma}] \sim N(f(\boldsymbol{x_i}), \boldsymbol{\Sigma}) \\
&[\beta_0] \sim N(0, \vartheta_0) \ , \ [\beta_1] \sim N(0, \vartheta_1) \ , \ \boldsymbol{\Sigma} \sim IW(a, \boldsymbol{B}) \\
&u_1, \ldots, u_K \overset{iid}{\sim} N(0, \sigma^2) \ , \ [\sigma^2] \sim IG(A, B)
\end{aligned} \qquad (۱۱)
$$

همان‌طور که در بخش ۲ بیان شد، هدف استنباط بیزی تخمین توزیع پسین برای پارامترهای مدل است. در اینجا چون با مدل‌های حاشیه‌ای روبرو هستیم، تمرکز ما بر روی برآورد $\boldsymbol{\Sigma}$ است. در اینجا از داده‌های شبیه‌سازی‌شده‌ی شکل ۲ ، که در آن مقادیر x_{ij} با فواصل یکسان تولید شدند با این ویژگی که

اولین مقدار برای هر آزمودنی، برای مثال x_{i1} ، از توزیع یکنواخت در فاصله‌ی (۸, ۱۲) به‌طور تصادفی انتخاب‌شده است، استفاده می‌کنیم. حال الگوریتم نمونه‌گیری گیبز را برای ۵۰۰۰ تکرار، با میزان دورریز برابر ۵۰۰۰ و با میزان معیار تقلیل برابر ۵، برای این مدل ناپارامتری انجام می‌دهیم. برای این کار از بسته BRugs در نرم‌افزار R، که پل ارتباطی بین نرم‌افزارهای OpenBUGS و R است، استفاده شده است. خروجی‌های حاصل در شکل‌های ۴ و ۶ سطح مطلوبی از برآورد بیزی را نشان می‌دهد. همان‌طور که ملاحظه می‌شود این خروجی‌ها شامل ۵ قسمت اصلی به شرح زیر هستند:

۱- $trace$: این قسمت نمودار تکرار الگوریتم در مقابل مقادیر حاصل از فرایند گیبز را نشان می‌دهد. نبودن جهش در داده‌ها در طول تکرار الگوریتم می‌تواند به‌منزله همگرایی داده‌ها به یک توزیع خاص باشد.

۲- lag۱ : این نمودار به بررسی تصادفی بودن داده‌ها می‌پردازد. برای داده‌های تصادفی نمودار پراکندگی داده‌ها در مقابل تکرار نباید از روند خاصی پیروی کند.

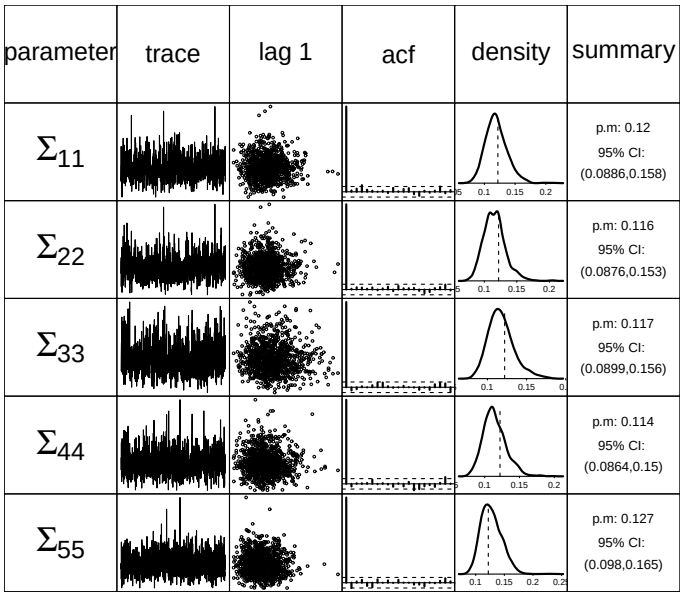
۳- acf : مقدار خودهمبستگی بین داده‌ها را نشان می‌دهد. در عمل خودهمبستگی کمتر از ۰/۱ برای اولین lag قانع‌کننده است (اولین lag منظور داده‌هایی که برای مبنای پارامتر تقلیل مشخص‌شده، به‌دست آمده‌اند). خط‌چین‌ها در نمودار نشان دهنده خط‌های ۰/۱ و ۰/۱− است.

۴- $density$: این قسمت نمودار نشان دهنده شکل تابع توزیع پسین است که با استفاده از نمونه تولید شده از توزیع پسین در فرایند گیبز حاصل شده است.

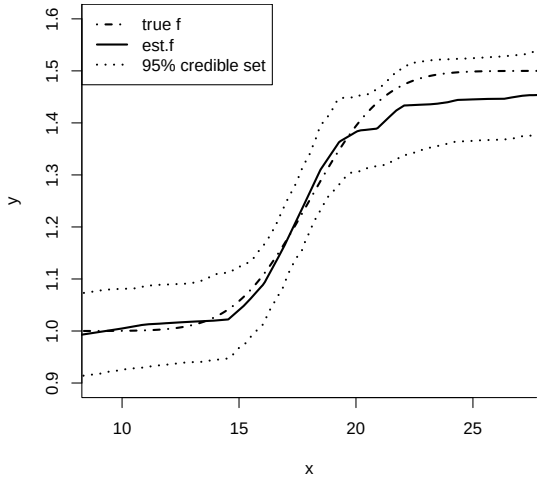
۵- $summary$: این قسمت از دو بخش میانگین پسین و بازه مؤثق [۶] تشکیل شده است. میانگین پسین، میانگین نمونه‌ای از مقادیر تولید شده توسط فرایند گیبز است. بازه مؤثق نیز یک بازه اطمینان برآورد شده برای پارامتر موردبررسی است که به‌صورت نمونه‌ای حاصل‌شده است. برای مثال حد پایین یک بازه مؤثق ۹۵ درصدی، چندک ۰/۰۲۵ داده‌ها و حد بالای آن چندک ۰/۹۷۵ داده‌ها است.

نمودار $trace$ همگرایی داده‌ها به یک توزیع مانا را تأیید می‌کند و نمودارهای lag۱ و acf عدم وجود همبستگی بین داده‌ها را نشان می‌دهند. نمودار $density$ نشان‌دهنده این موضوع است که بیشترین احتمال حول مقدار واقعی پارامتر موردنظر قرارگرفته است (خط بریده عمودی نشان‌دهنده مقدار دقیق

[6] Credible interval

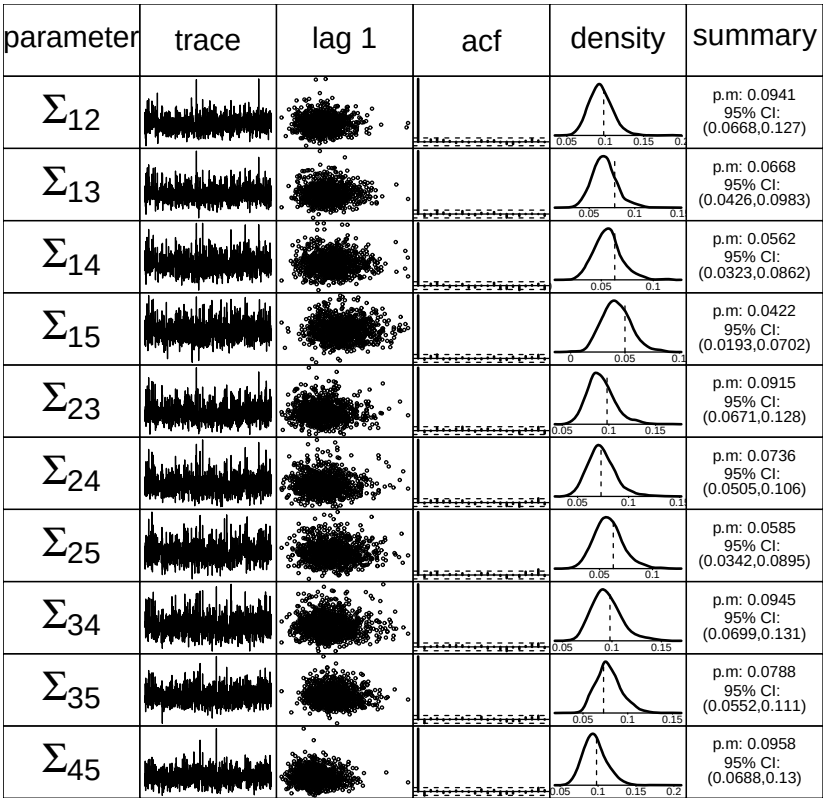


شکل ۴. جزئیات استنباط بیزی از مقادیر قطر اصلی ماتریس Σ برای مدل ناپارامتری (۱۱) (مقادیر درست پارامترها با خطهای بریده افقی نمایش داده شده است)



شکل ۵. نمودار f (خط ممتد)، برآورد تابع با الگوریتم گیبز (خط‌چین) و بازه مؤثق (نقطه‌چین)

پارامتر است). همچنین در قسمت $summary$ می‌بینیم که میانگین نمونه‌ای به مقدار اصلی پارامتر نزدیک بود و بازه مؤثق نیز این مقدار را در برمی‌گیرد. شکل ۵ نیز نشان دهنده مقایسه برآورد تابع f (از روش ،(MCMC با مقدار واقعی آن است. همان‌طور که ملاحظه می‌شود، فرایند گیبز تا حد قابل‌قبولی این برآورد را درست انجام داده است. همچنین در این شکل بازه مؤثق ۹۵ درصد برای این برآورد نیز به نمایش درآمده است. به‌طورکلی می‌توان گفت که استنباط بیزی عملکرد قابل‌قبولی را نشان می‌دهد (لون و همکاران، ۲۰۰۰).



شکل ۶. جزئیات استنباط بیزی از مقادیر کوواریانس ماتریس Σ برای مدل ناپارامتری (۱۱) (مقادیر درست پارامترها با خط‌های بریده افقی نمایش داده‌شده است)

۸ مدل ضرایب متغیر برای داده‌های همه گیرشناسی

در این بخش رگرسیون ضرایب متغیر (۸) نمونه‌ای از داده‌های اپیدمیولوژِی تغذیه برازش داده می‌شود. این داده‌ها بر اساس گزارش ۲۴ ساعته بر روی نشانگرهای انرژی و پروتئین (نیتروژن ادرار)، همراه با پرسشنامه مصرف مواد غذایی در ۲۴ ساعت انجام‌شده است. متغیرهای موجود در این مدل به‌صورت زیر تعریف می‌شوند:

y: لگاریتم مصرف پروتئین اندازه‌گیری شده با نشانگر نیتروژن ادرار

s: شاخص توده‌ی بدن

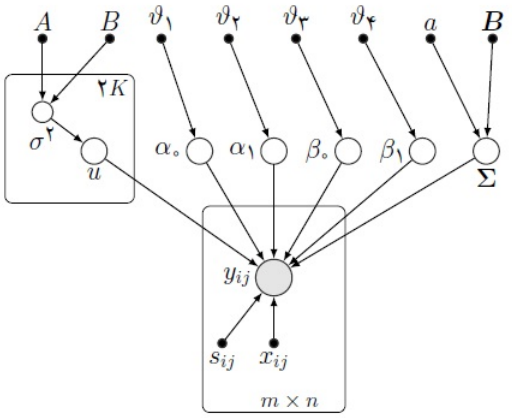
x: لگاریتم مصرف پروتئین اندازه‌گیری شده در گزارش ۲۴ ساعته

در اینجا $m = ۲۹۴$ ، $n = ۲$ و تعداد گره‌ها برابر با ۳۰ اختیار شده است. در این قسمت نیز با استفاده از روش نمونه‌گیری گیبز برآوردی برای $\boldsymbol{\Sigma}$ به دست آورده می‌شود. ازاین‌رو مدل بیز سلسله مراتبی برای مدل (۸) را می‌توان به‌صورت

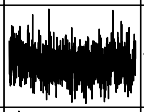
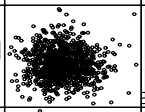
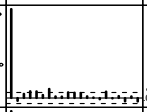
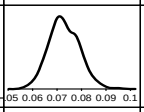
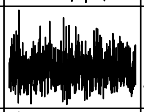
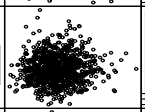
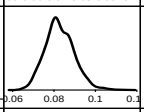
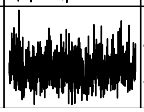
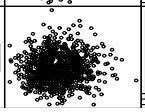
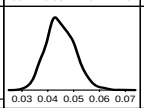
$$
\begin{aligned}
&[\boldsymbol{y_i}|\boldsymbol{\beta}, \gamma, \sigma^۲, \boldsymbol{u}, \boldsymbol{\Sigma}] \sim N(f_۰(\boldsymbol{s_i}) + f_۱(\boldsymbol{s_i})\boldsymbol{x_i}, \boldsymbol{\Sigma}) \\
&[\beta_۰] \sim N(۰, \vartheta_۲) \ , \ [\beta_۱] \sim N(۰, \vartheta_۳) \\
&[\alpha_۰] \sim N(۰, \vartheta_۱) \ , \ [\alpha_۱] \sim N(۰, \vartheta_۴) \\
&u_{۱۱}, \ldots, u_{۱K}, u_{۲K} \overset{iid}{\sim} N(۰, \sigma^۲) \\
&\boldsymbol{\Sigma} \sim IW(a, \boldsymbol{B}) \ , \ [\sigma^۲] \sim IG(A, B)
\end{aligned}
\qquad (۱۲)
$$

در نظر گرفت. همچنین مدل گرافیکی این روابط در شکل ۷ آمده است. باید دقت شود که در این مدل $f_۰$ و $f_۱$ مانند معادله (۹) تعریف شده است. همچنین تعداد گره‌ها به‌صورت $K_۱ = K_۲ = ۳۰$ در نظر گرفته شده و مکان گره‌ها نیز در فواصل یکسان از هم انتخاب‌شده‌اند. برای این مدل نیز الگوریتم نمونه‌گیری گیبز برای ۵۰۰۰ تکرار، با میزان دورریز ۵۰۰۰ و با میزان معیار تقلیل ۵ انجام شده است.

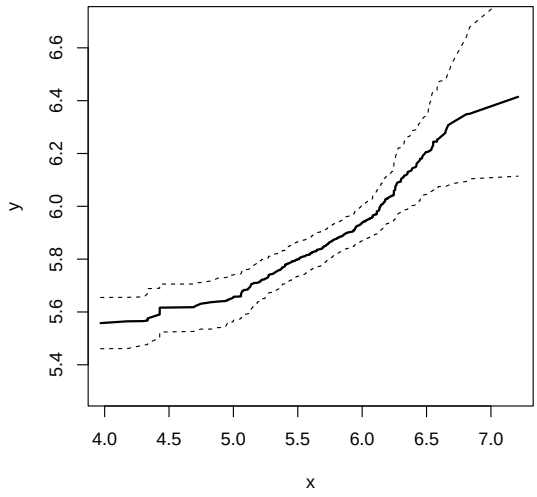
در شکل ۸ جزئیات استنباط بیزی انجام شده توسط فرایند گیبز نمایش داده شده است. همان‌طور که ملاحظه می‌شود نمودار $trace$ همگرایی مطلوبی را نشان می‌دهد و بر مبنای نمودارهای $lag۱$ و acf ، مقدار همبستگی بین نمونه تولید شده توسط فرایند گیبز، ناچیز است. همچنین توزیع مؤلفه‌های واریانس، توزیع تقریباً متقارن برآورد شده است. لازم به ذکر است از آنجایی‌که داده‌ها در دو زمان تکرار شده‌اند ماتریس $\boldsymbol{\Sigma}$ ، یک ماتریس ۲×۲ است.



شکل ۷. مدل گرافیکی برای مدل بیزی سلسله مراتبی (۱۲)

parameter	trace	lag 1	acf	density	summary
Σ_{11}					p.m: 0.0729 95% CI: (0.0616,0.0859)
Σ_{22}					p.m: 0.0828 95% CI: (0.0703,0.0971)
Σ_{12}					p.m: 0.045 95% CI: (0.0352,0.0566)

شکل ۸. استنباط بیزی برای مؤلفه‌های ماتریس Σ ، مدل ضرایب متغیر (۱۲)



شکل ۹. برآورد بیزی تغییرات متغیر پاسخ y در مقابل متغیر پیش‌بین x

همچنین در شکل ۹ برآورد بیزی تغییرات متغیر پاسخ y در مقابل متغیر پیش‌بین x (حاصل از فرایند گیبز) به نمایش درآمده است. همچنین خط‌چین‌ها در این شکل نشان‌دهنده بازه مؤثق ۹۵ درصدی برای این برآورد است.

۱.۸ بررسی همگرایی مدل با آماره گلمن-روبین

از مسائل مهم در برآورد یابی به روش MCMC، بررسی همگرایی فرایند گیبز به یک توزیع مانا یا توزیع پسین پارامتر موردعلاقه است. روش‌های گوناگونی برای بررسی همگرایی وجود دارد که همگی در مجموعه روش‌های تعقیبی[۷] قرار دارند. یکی از راه‌های تشخیص همگرایی، روش گلمن-روبین[۸] است، که آماره آن طی گام‌های زیر به دست می‌آید:

۱- اجرا کردن فرایند گیبز برای $m \geq ۲$ زنجیره با طول $۲n$ و با مقادیر اولیه متفاوت.

۲- کنار گذاشتن n نمونه اول از هر زنجیره.

[7] Post hoc
[8] Gelman-Rubin

۳- به دست آوردن واریانس‌های بین و درون زنجیر. واریانس درون زنجیر برای پارامتر θ به‌صورت زیر حاصل می‌شود:

$$W = \frac{۱}{m}\sum_{j=۱}^{m} s_j^۲ \qquad \text{به طوری که} \qquad s_j^۲ = \frac{۱}{n-۱}\sum_{i=۱}^{n}(\theta_{ij} - \bar{\theta}_j)^۲$$

اگر داده‌های به‌دست آمده توسط فرایند گیبز، به یک توزیع مانا همگرا نشده باشند، مقدار W، واریانس پارامتر را درست برآورد نمی‌کند. همچنین واریانس بین گروهی برای زنجیره‌های متفاوت به‌صورت زیر حاصل می‌شود:

$$B = \frac{n}{m-۱}\sum_{j=۱}^{m}(\bar{\theta}_j - \bar{\bar{\theta}})^۲ \qquad \text{به طوری که} \qquad \bar{\bar{\theta}} = \frac{۱}{m}\sum_{j=۱}^{m}\bar{\theta}_j \qquad (۱۳)$$

۴- برآورد واریانس پارامتر مورد نظر به‌وسیله واریانس بین و درون زنجیره به‌صورت زیر به دست می‌آید:

$$\widehat{Var}(\theta) = (۱ - \frac{۱}{n})W + \frac{۱}{n}B \qquad (۱۴)$$

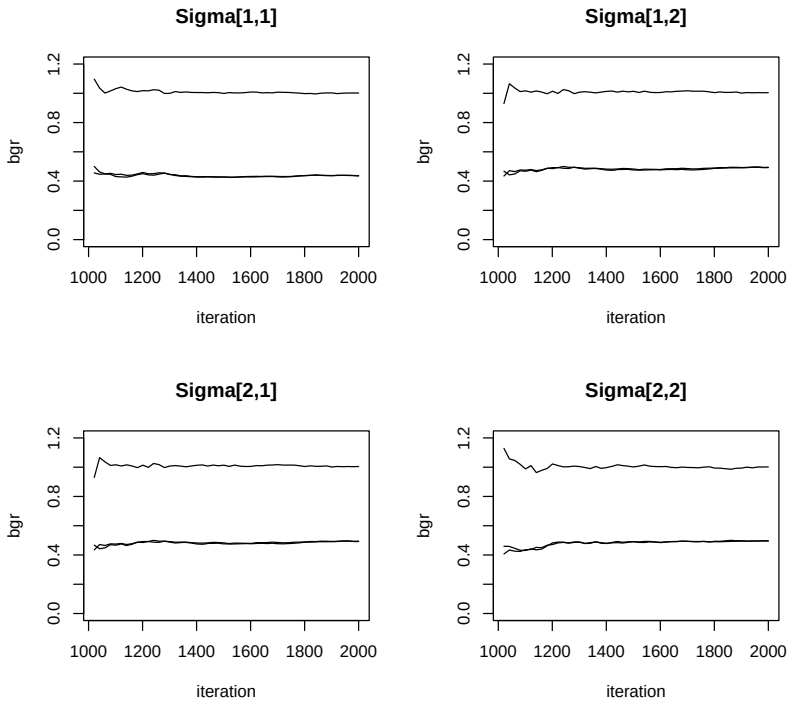
۵- به دست آوردن ضریب کاهش بالقوه[9] به‌صورت زیر:

$$R = \frac{\widehat{Var}(\theta)}{W} \qquad (۱۵)$$

حال اگر تمامی زنجیره‌های تولید شده به توزیع پسین همگرا شده باشند، تمام زنجیره‌ها هم‌پوشانی داشته و $B = W$ است و $R = ۱$ به دست می‌آید. ولی اگر این مقدار بیشتر از ۱/۱ یا ۱/۲باشد، باید تعداد تکرارها را تا زمانی که مقدار R به سطح مطلوب (حول ۱، با اختلاف کمتر از ۱/۰ تا حداکثر ۲/۰) برسد افزایش می‌دهیم. در شکل ۱۰ نمودار حاصل از این آماره، برای مؤلفه‌های ماتریس Σ به تصویر کشیده شده است. این نمودار برحسب سه زنجیره که هرکدام با میزان دورریز ۵۰۰۰ و تکرار ۵۰۰۰ و با تقلیل ۵ اجراشده‌اند، حاصل شده است. در هر یک از این نمودارها دو خط افقی پایین نشان دهنده صورت و مخرج

[9] Potential scale reduction factor

معادله (۱۵) هستند. دو خط پایین به علت قرار گرفتن آن‌ها روی هم در اکثر نقاط، شاید به‌خوبی قابل مشاهده نباشند). مقدار R یا آماره گلمن-روبین نیز در خط بالای نمودار مشخص شده است که از همگرایی مدل به یک توزیع مانا حمایت می‌کند (کراسچکه، ۲۰۱۴).



شکل ۱۰. بررسی آماره گلمن-روبین برای مؤلفه‌های ماتریس $\boldsymbol{\Sigma}$ ، دو خط پایین صورت و مخرج معادله (۱۵) هستند.

همچنین در جدول ۱ نیز مقادیر حاصل‌شده برای آماره گلمن-روبین گزارش شده است. همان‌طور که ملاحظه می‌شود، تکرارها به ۵۰ بلوک تقسیم شده و برای هر بلوک واریانس‌های بین و درون زنجیر به‌دست آمده و در نهایت مقدار R طبق معادله (۱۵) حاصل‌شده است.

جدول ۱. مقادیر آماره گلمن-روبین برای مؤلفه‌های ماتریس Σ

بلوک	تکرار	$R_{\Sigma_{11}}$	$R_{\Sigma_{12}}$	$R_{\Sigma_{22}}$
۱	۱۰۰۱–۱۰۲۰	۰۹۷۱/۱	۹۸۹۸/۰	۱۲۸۹/۱
۲	۱۰۲۱–۱۰۴۰	۰۳۵۰/۱	۰۶۵۲/۱	۰۵۶۶/۱
۳	۱۰۴۱–۱۰۶۰	۰۰۱۲/۱	۰۳۶۲/۱	۰۴۶۰/۱
۴	۱۰۶۱–۱۰۸۰	۰۱۶۷/۱	۰۱۱۷/۱	۰۱۹۴/۱
۵	۱۰۸۱–۱۲۰۰	۰۴۲۴/۱	۰۱۶۵/۱	۹۸۸۱/۰
۶	۱۲۰۱–۱۲۲۰	۰۱۶۵/۱	۰۰۸۱/۱	۹۶۳۰/۰
⋮	⋮	⋮	⋮	⋮
۴۴	۱۸۶۱–۱۸۸۰	۰۰۰۸/۱	۰۰۸۸/۱	۹۹۲۵/۰
۴۵	۱۸۸۱–۱۹۰۰	۰۰۲۴/۱	۰۰۱۲/۱	۹۹۳۶/۰
۴۶	۱۹۰۱–۱۹۲۰	۹۹۷۸/۰	۰۰۴۹/۱	۰۰۰۷/۱
۴۷	۱۹۲۱–۱۹۴۰	۰۰۰۶/۱	۰۰۳۴/۱	۹۹۵۵/۰
۴۸	۱۹۴۱–۱۹۶۰	۰۰۲۰/۱	۰۰۴۶/۱	۰۰۱۲/۱
۴۹	۱۹۶۱–۱۹۸۰	۰۰۱۹/۱	۰۰۳۵/۱	۰۰۱۱/۱
۵۰	۱۹۸۱–۲۰۰۰	۰۰۱۹/۱	۰۰۴۵/۱	۰۰۱۹/۱

بحث و نتیجه‌گیری

مدل‌های طولی حاشیه‌ای مدل‌های مناسبی برای تحلیل داده‌های طولی هستند. در این مقاله مدل را بر اساس روش مدل‌های آمیخته بر پایه اسپلاین‌های تاوانیده که کمتر به‌صورت عمیق به آن پرداخته شده است را مورد بررسی قرار داده‌ایم. همچنین با استفاده از BUGS و بسته نرم‌افزاری $BRugs$ استنباط‌های بیزی مورد نظر انجام شده است. علاوه بر آن با استفاده از شبیه‌سازی نشان داده شده است که استنباط‌های موردنظر از دقت نسبتاً بالائی برخوردارند. لازم به ذکر است که از دیگر مزیت‌های این روش می‌توان به کاربرد آن در مدل‌های پیچیده‌تر دیگر مانند مدل‌های لگ خطی، رگرسیون چندکی و یا حتی مدل‌ها با داده‌های گم‌شده نیز اشاره کرد که در این مقاله به آن پرداخته نشده است و می‌تواند پیشنهادی برای پژوهش‌های بعدی در نظر گرفته شوند.

تقدیر و تشکر

نویسندگان از داوران گرامی برای ارائه پیشنهادهای سازنده در راستای بهبود این پژوهش و از دقت نظر ویراستار محترم مجله بسیار سپاس‌گزارند.

مراجع

Al Kadiri, M., Carroll, R. J. and Wand, M. P., (2010), Marginal Longitudinal Semiparametric Regression via Penalized Splines, *Statistics and Probability Letters*, **80**, 1242–1252.

Carroll, R. J., Maity, A., Mammen, E., Yu, K., (2009), Efficient Semiparametric Marginal Estimation for the Partially Linear Additive Model for Longitudinal/Clustered Data, *Statistics in Biosciences*, **1**, 10-31.

Fan, J., Huang, T. and Li, R., (2007), Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function, *Journal of the American Statistical Association*, **102**, 632-641.

Kruschke, K. J., (2014), *Donig Bayesian Data Analysis , a Tutarial with R and BUGS* , Academic Press ,Elsevier.

Lunn, D. J., Thomas, A., Best, N., Spiegelhalter, D., (2000), WinBUGS-a Bayesian Modellingramework: Concepts, Structure, and Extensibility, *Journal of Statistics and Computing*, Stat. **10**, 325-337.

Menictasa, M., and Wand, M. P., (2013), Variational Inference for Marginal Longitudinal Semiparametric Regression, *Dissemination of Statistics Research*, **2**, 61–71.

Ruppert, D., Wand, P. and Carroll, R. J., (2003), *Semiparametric Regression*, Cambridge, New York.

West, B. T., Galecki, A. T. and Welch, K. (2006), *Linear Mixed Models A Practical Guide Using Statistical Software*, Taylor & Francis Group New York.

Wand, M. P., and Ormerod, J., (2008), On Semiparametric Regression With O'Sullivan Penalized Splines, *Australian Statistical Publishing Association*, Stat. **50**, 179–198.

Wand, M. P., (2008), *Semiparametric Regression and Graphical Models*, Centre for Statistical and Survey Methodology, The University of Wollongong.

Wang, N., Carroll, R. J. and Lin, X., (2005), Efficient Semiparametric Earginal Estimation for Longitudinal/Clustered Data, *Journal of the American Statistical Association*, **100**, 147-157.

Welham, S. J., Cullis, B. R., Kenward, M. G. and Thompson, R., (2007), A Comparison of Mixed Model Splines for Curve Fitting, *Australian and New Zealand Journal of Statistics*, **49**, 1-23.

Zeger, S. and Diggle, P. J., (1994), Semiparametric Moels for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters, *Biometrics*, **50**,689-699.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P., (2006), General Design Bayesian Generalized Linear Mixed Models, *Statistical Science*, **21**, 35-51.