مجله مدل‌سازی پیشرفته ریاضی
سال ۱۴۰۱، دوره ۱۲، شماره ۱، ص ۱۳-۲۳

شاپا چاپی: ۲۲۵۱-۸۰۸۸
DOI: 10.22055/JAMM.2022.34594.1841

مدل‌بندی همزمان میانگین و دقت در مدلهای آمیخته رگرسیون بتای افزوده

زهره فلاح محسن‌خانی[۱]، پروین اژدری[۲]×

[۱] استادیار، پژوهشکده آمار، تهران، ایران
[۲] استادیار، دانشکده علوم پایه، دانشگاه آزاد اسلامی واحد تهران شمال، تهران، ایران

دبیر مسئول: محمد رضا زادکرمی

تاریخ دریافت: ۱۳۹۹/۵/۱۹ تاریخ پذیرش: ۱۴۰۰/۸/۹

چکیده: مدل رگرسیون بتای افزوده برای مدل‌بندی داده‌هایی از جنس نرخ، نسبت یا درصد استفاده می‌شود. این مدل ازآمیختن توریع بتا روی بازه $(۰,۱)$ و دو توزیع تباهیده در صفر و یک ایجاد می‌شود. با بازپارامتریدن توزیع بتا، پارامترهای میانگین و دقت این مدل با ساختاری شامل اثرات ثابت و تصادفی مدل‌بندی می‌شود. عموما برای راحتی در مطالعات، پارامتر دقت ثابت در نظر گرفته می‌شود و مدل بندی فقط بر اساس پارامتر میانگین انجام می‌شود. در این مقاله مدل بندی همزمان میانگین و دقت در مدل‌های آمیخته رگرسیون بتای افزوده ارائه و کارایی مدل در مطالعات شبیه سازی با رهیافت بیزی مورد بررسی قرار می‌گیرد. سپس نحوه کاربست این مدل برای تحلیل مدل بندی سهم شاغلین درخانوار بر اساس نتایج آمارگیری نیروی کار مرکز آمار نشان داده می‌شود و در انتها بحث و نتیجه‌گیری ارائه خواهد شد.

واژه‌های کلیدی: رگرسیون بتای افزوده، پارامتر دقت، مدل آمیخته، تحلیل بیزی، آمارگیری نیروی کار.

رده‌بندی ریاضی: 62F15, 62J12

۱ مقدمه

برخی از مطالعات در حوزه‌های مختلف شامل داده‌هایی هستند که به صورت نرخ یا نسبت در بازه‌ی (۱و۰) اندازه گیری می‌شوند. مدل رگرسیون بتا یک انتخاب مناسب برای تحلیل متغیرهای پاسخ پیوسته روی بازه واحد است. پائولینو (۲۰۰۱) برای اولین بار به منظور مدل‌بندی متغیرهای پاسخ از جنس نسبت، فرض کرد متغیر پاسخ از توزیع بتا $B(a,b)$ با تابع چگالی

$$\pi(y;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}y^{a-1}(1-y)^{b-1} \qquad 0<y<1, \quad a,b>0$$

×نویسنده مسئول مقاله
رایانامه: zoherhf@yahoo.com (Z. Falah), p_azhdari@iau-tnb.ac.ir (P. Azhdari)

پیروی می‌کند، که میانگین و واریانس آن به ترتیب عبارتند از $E(Y)=\frac{a}{a+b}$ و $Var(Y)=\frac{ab}{(a+b)^{۲}(a+b+۱)}$. وی پارامترهای توزیع بتا، را نیز مدل‌بندی نموده و برآورد ماکسیمم درستنمایی آن‌ها را به‌دست آورد این روش از لحاظ محاسباتی با دشواری بسیاری همراه بود. فراری و کریباری (۲۰۰۴) توزیع بتای بازپارامتریده را برای مطالعه این نوع داده‌ها پیشنهاد کردند. آن‌ها پارامترهای توزیع بتا را به‌گونه‌ای بازنویسی کردند که مدل رگرسیونی براساس میانگین متغیر پاسخ بیان شود. برای این منظور با قرار دادن $\mu=\frac{a}{a+b}$ و $\phi=a+b$ توزیع بتا به‌صورت:

$$\pi(y;\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\left((۱-\mu)\phi\right)}y^{\mu\phi-۱}(۱-y)^{(۱-\mu)\phi-۱}\quad ۰<y<۱ \tag{۱.۱}$$

بازپارامتریده می‌شود، که در آن $\Gamma(.)$ تابع گاما و μ میانگین و ϕ پارامتر دقت است $(۰<\mu<۱ و \phi>۰)$. سپس با در نظر گرفتن الگویی خطی برای متغیرهای تبیینی به برازش مدل پرداختند. در این مدل‌ها متغیرهای تبینی و میانگین متغیر پاسخ از طریق یک تابع پیوند مناسب به صورت $g(\mu_i)=\sum_{j=۱}^{k}x_{ij}\beta_j$ به هم مربوط می‌شوند، که در آن β بردار پارامترهای رگرسیونی است.

چپیدا و گامرمن (۲۰۰۵۲) و اسمیتسون و ورکولین (۲۰۰۶۲)، مدل رگرسیون بتا را با متغیر درنظرگرفتن پارامتر دقت بررسی کردند. آن‌ها به طور همزمان لوجیت میانگین و لگاریتم پارامتر دقت را به صورت الگوی خطی از متغیرهای تبینی لحاظ نمودند. سپس برانسکام و همکاران (۲۰۰۷) یک مدل رگرسیون بتا بیزی را ارائه کردند و کاربرد آن را در داده‌های ژنتیک مطالعه نمودند. مدل آمیخته رگرسیون بتا اولین بار توسط زیمپریچ (۲۰۱۰) در مطالعه داده‌های طولی از طریق وارد کردن اثر تصادفی در مدل میانگین برای بررسی تغییر سرعت واکنش افراد نسبت به محرک‌ها با افزایش سن، مطرح شد، ورکولین و اسمیتسون (۲۰۱۲) با در نظر گرفتن توزیع بتای آمیخته برای متغیر پاسخ، برآورد پارامترهای مدل را با رهیافت بیزی و روش ماکسیمم درستنمایی ارایه نمودند. فیگیورا-زونیکا و همکاران (۱۳۰۲)، پارامترهای میانگین و دقت مدل آمیخته خطی تعمیم‌یافته بتا را با رهیافت بیزی ارائه کردند. همچنین بونات و همکاران (۲۰۱۳) مدل آمیخته رگرسیون بتا را برای داده‌های واقعی استفاده و با فرض ثابت بودن پارامتر دقت، برآوردهای ماکسیمم درستنمایی پارامترهای مدل را ارائه کردند. چپیدا و همکاران (۲۰۱۴) مدل رگرسیون بتا را با در نظر گرفتن یک اثر تصادفی در مدل‌بندی پارامتر دقت مطالعه کرده و برآورد بیزی پارامترهای مدل را به‌دست آوردند.

در مسائل کاربردی ممکن است با مواردی مواجه شویم که نرخ‌ها هر یک از مقادیر مجموعه $\{۰و۱\}$ را نیز اختیار کنند، یک راه حل استفاده از آمیختن گسسته–پیوسته توزیع است که توسط اوسپینا و فراری (۲۰۱۰) پیشنهاد شد که بسته به شرایط، جرم احتمال‌های صفر، یک یا هر دو را نیز شامل می‌شود، همچنین کاربرد این مدل برای داده‌های واقعی در ۲۰۱۲ توسط ورکویلین و سیمتسون ارایه گردید. گالویس و همکاران (۲۰۱۴) مدل رگرسیون بتای افزوده را برای داده‌های مشاهده شده در بازه $[۰,۱]$ پیشنهاد کردند که از آمیختن دو توزیع تباهیده در نقاط صفر و یک و توزیع بتا با تکیه‌گاه $(۰,۱)$ به‌دست می‌آید. پارکر و همکاران (۲۰۱۴)، اثرات فضایی را در مدل‌بندی رگرسیون بتای افزوده در برآورد پارامتر میانگین بررسی و در داده‌های دندانپزشکی بکار گرفتند. فلاح و همکاران (۲۰۱۹) اثرات طولی را در مدل‌بندی پارامتر میانگین این مدل بررسی کردند. نوگاروتو و همکاران (۲۰۲۰) برآوردهای روش بیزی و روش ماکسیمم درستنمایی را در مدل رگرسیون بتای افزوده با یکدیگر مقایسه کردند. نتایج تحقیق آنها نشان داده است که رهیافت بیزی به اندازه روش ماکسیمم درستنمایی دقیق است.

در سال‌های اخیر استفاده از مدل رگرسیون بتای افزوده برای داده‌های نسبتی و لحاظ اثرات ثابت و تصادفی با استفاده از روش‌های گوناگون و با لحاظ رهیافت‌های متفاوت در مدل‌بندی پارامتر میانگین، بحث و بررسی شده است، در این مقاله مدل آمیخته رگرسیون بتای افزوده برای حالتی که پارامتر دقت توزیع بتا نیز متغیر باشد بررسی می‌شود. به عبارتی به طور همزمان یک الگوی خطی برای پارامتر میانگین و یک الگوی خطی برای پارامتر دقت در نظر گرفته، و با استفاده از رهیافت بیزی برآورد پارامترها ارائه می‌شود. سپس مدل معرفی شده برای این نوع داده‌ها در مطالعات شبیه‌سازی مورد بررسی قرار می‌گیرد. همچنین پیشین‌های مناسب برای پارامترهای این مدل معرفی می‌شوند و عملکرد آن‌ها در حالت‌های مختلف بررسی می‌شوند. توزیع‌های پیشین انتخاب شده برای مدل‌بندی سهم شاغلین درخانوار براساس نتایج آمارگیری نیروی کار مرکز آمار که داده‌هایی نسبتی هستند، بکار گرفته و در پایان بحث و نتیجه گیری ارائه می‌شود.

۲ مدل‌بندی داده‌های نسبتی در دامنه بسته [۰و۱]

توزیع بتا اغلب انتخابی مناسب برای برازش داده‌های پیوسته در بازه (۰و۱) است. تابع چگالی احتمال بتای بازپارامتریده شده برای متغیر تصادفی Y برحسب پارامترهای میانگین و دقت به صورت رابطه (۱.۱) است،

$$E(Y)=\mu \text{ و } Var(Y)=(\mu(۱-\mu))/(۱+\phi))$$

علی‌رغم سازگاری مدل رگرسیون بتا با این‌گونه داده‌ها، گاهی وجود تعداد قابل ملاحظه‌ای صفر یا یک در مشاهدات، تطبیق‌پذیری این مدل را به چالش می‌کشاند، زیرا تکیه‌گاه توزیع بتا بازه (۰و۱) است. راه حل مواجهه با این مشکل افزودن احتمال‌های صفر و یک به تابع

چگالی بتا و ایجاد توزیع آمیخته است. گالویس و همکاران (۲۰۱۴) مدل بتای افزودهصفر و یک را پیشنهاد کردند که شامل یک توزیع سه قسمتی است به‌گونه‌ای که در نقاط صفر و یک تباهیده و در بازه (۰،۱) دارای چگالی بتا به صورت زیر است:

$$f(y|p_0,p_1,\mu,\phi)=\begin{cases}p_0 & y=0\\ p_1 & y=1\\ (1-p_0-p_1)\pi(y|\mu,\phi) & 0<y<1\end{cases}$$

که در آن $\pi(y|\mu,\phi)$ تابع چگالی بتا و $p_0,\ p_1\geq 0$ و $0\leq p_0+p_1\leq 1$ میانگین و واریانس این توزیع به ترتیب عبارتند از:

$$E(Y)=(1-p_0-p_1)\mu+p_1,$$
$$Var(Y)=p_1(1-p_1)+(1-p_0-p_1)\left[\frac{\mu(1-\mu)}{(1+\phi)}+(p_0+p_1)\mu^2-2\mu p_1\right]$$

۱.۲ مدل رگرسیون آمیخته بتای افزوده:

فرض کنید $y_i=(y_{i1},\ldots,y_{in_i})$ برداری به طول n_i برای واحد نمونه i ام است که از توزیعی به صورت:

$$f(y_{ij}|p_0,p_1,\mu_{ij},\phi_{ij})=\begin{cases}p_0 & y_{ij}=0\\ p_1 & y_{ij}=1\\ (1-p_0-p_1)\pi(y|\mu_{ij},\phi_{ij}) & 0<y_{ij}<1\end{cases}\qquad(۱.۲)$$

پیروی می‌کند. که در آن $\pi(y|\mu_{ij},\phi_{ij})$ تابع چگالی بتای (۱.۱) است، $p_0=P(Y_{ij}=0)$ و $p_1=P(Y_{ij}=1)$. بنابراین تابع چگالی y_{ij} برحسب تابع نشانگر برابر است با:

$$\pi(y_{ij};p_0,p_1,\mu_{ij},\phi_{ij})=p_0^{I(y_{ij}=0)}p_1^{I(y_{ij}=1)}\{(1-p_0-p_1)\pi(y_{ij};\mu_{ij};\phi_{ij})\}^{(1-I(y_{ij}=0))(1-I(y_{ij}=1))}$$

در مدل رگرسیون بتا وقتی پارامتر دقت ثابت باشد، شرط همگنی واریانس که از جمله شرط‌های مدل‌های خطی تعمیم یافته است، برقرار می‌باشد؛ کریباری و زیلیس (۲۰۱۰). بنابراین متغیرهای تبیینی و اثرات تصادفی را می‌توان براساس یک تبدیل مناسب پارامتر متغیر میانگین به صورت:

$$g(E(\boldsymbol{Y_i}|\boldsymbol{b_i}))=g(\boldsymbol{\mu_i})=\boldsymbol{X_i^T}\beta+\boldsymbol{Z_i^T}\boldsymbol{b_i}$$

مدل‌بندی کرد. که در آن $\boldsymbol{X_i}$ ماتریس طرح $p\times n_i$، $(\beta_1,\ldots,\beta_p)^T=\beta$ بردار اثرات ثابت، $\boldsymbol{Z_i}$ ماتریس طرح $q\times n_i$ و $(b_{i1},\ldots,b_{iq})^T=\boldsymbol{b_i}$ بردار اثرات تصادفی است. برای تابع پیوند $g(.)$ انتخاب‌های متفاوتی از جمله لوجیت را می‌توان اختیار نمود.

۲.۲ مدل بندی همزمان میانگین و دقت

برای مدل بندی همزمان پارامترهای میانگین و دقت، لازم است تابع پیوند مناسبی برای پارامتر دقت نیز در نظر گرفته شود.
فرض کنید $\boldsymbol{y_i}=(y_{i1},\ldots,y_{in_i})$ برداری به طول n_i برای واحد نمونه iام است که از توزیعی براساس رابطه (۱.۲) با ثابت در نظر گرفتن پارامترهای p_0 و p_1 تبعیت می‌کند، مدل بندی پارامترهای میانگین و دقت را می‌توان به‌صورت زیر در نظر گرفت:

$$g(\boldsymbol{\mu_i})=\boldsymbol{X_i^T}\beta+\boldsymbol{Z_i^T}\boldsymbol{b_i}$$
$$h(\boldsymbol{\phi_i})=\boldsymbol{W_i^T}\delta+\boldsymbol{U_i^T}\boldsymbol{d_i}$$

که در آن $\boldsymbol{U_i}$ ماتریس طرح $q^*\times n_i$ بعدی، $(d_{i1},\ldots,d_{iq*})^T=\boldsymbol{d_i}$ بردار اثرات تصادفی می‌باشند. فیوگیورا–زونیگا و همکاران (۲۰۱۳) برآوردهای پارامترهای مدل آمیخته خطی تعمیم‌یافته بتا را در وضعیت متغیر بودن پارامتر دقت با رهیافت بیزی مطالعه نمودند و

برای مدل بندی پارامتر میانگین، از پیوند لوجیت و برای مدل بندی پارامتر دقت، از پیوند لگاریتم استفاده کردند. با در نظر گرفتن پیوند لوجیت برای مدل بندی پارامتر میانگین و پیوند لگاریتم برای پارامتر دقت و صرف‌نظر کردن از اثرات تصادفی در مدل‌بندی پارامتر دقت داریم:

$$log\,it(\boldsymbol{\mu_i}) = \boldsymbol{X_i^T}\beta + \boldsymbol{Z_i^T}\boldsymbol{b_i}$$
$$ln(\boldsymbol{\phi_i}) = \boldsymbol{W_i^T}\delta$$

بعبارتی پارامترها را می‌توان به صورت زیر در نظر گرفت:

$$\mu_{ij} = \frac{\exp\left(\boldsymbol{X_i^T}\beta + \boldsymbol{Z_i^T}\boldsymbol{b_i}\right)}{1 - \exp\left(\boldsymbol{X_i^T}\beta + \boldsymbol{Z_i^T}\boldsymbol{b_i}\right)} = \frac{\exp\left(\eta_{ij}\right)}{1 - \exp\left(\eta_{ij}\right)}$$
$$\phi_{ij} = \exp\left(\boldsymbol{W_i^T}\delta\right) = \exp\left(\tau_{ij}\right)$$

که در آن β پارامترهای رگرسیونی و b_i بردار اثرات تصادفی در مدل‌بندی پارامتر میانگین و δ پارامترهای رگرسیونی در مدل‌بندی پارامتر دقت رگرسیون آمیخته بتای افزوده است. همچنین i نشان‌دهنده فرد مورد نظر و اندیس j معرف تعداد تکرار فرد iام است. در تحلیل بیزی برای اثرات تصادفی توزیع نرمال با میانگین صفر و واریانس σ_b^2 و برای پارامترها توزیع‌های پیشین ناآگاهی بخش یا کم‌آگاهی بخش درنظر گرفته می‌شود، بدین منظور در مواردی از بیز سلسله مراتبی نیز استفاده شده است. برای پارامترهای رگرسیونی در مدل‌بندی پارامتر دقت رگرسیون آمیخته بتای افزوده، گروه‌های متفاوتی از توزیع‌های پیشین را در نظر می‌گیریم. گروه اول، توزیع نرمال $\delta \sim N(0, \Sigma_\delta)$ است، که در آن Σ_δ ماتریس مثبت معین کوواریانس است. گروه دوم پیشین سلسله مراتبی معرفی شده توسط فیوگیورا-زونیگا و همکاران (۱۳۰۲) در مدل‌های آمیخته رگرسیون بتا می‌باشد، آنها برای δ توزیع $t-$ استودنت بصورت $\delta \sim t(\upsilon_d, 0, \Sigma_\delta)$ و همچنین برای υ توزیع نمایی $\upsilon \sim exp(a)$ به‌ازای $a = 0/1$ در نظر گرفتند. گروه سوم مفروضات گروه دوم را داراست علاوه بر آن یک بیز سلسله مراتبی برای پارامتر Σ_δ نیز در نظر گرفته می‌شود، فونگ و همکاران (۲۰۱۰) برای مدل‌های آمیخته خطی تعمیم‌یافته واریانس توزیع $t-$ استودنت، Σ_δ را دارای توزیع ویشارت معکوس $\Sigma_\delta \sim IWishart(\xi, \iota)$ در نظر گرفتند.

در سه مدل معرفی شده، برای بعضی از پارامترها، توزیع‌های پیشین یکسان است که به صورت زیر فرض می‌شوند:

$$\begin{aligned} &\beta \sim N_p(0, \Sigma_\delta) \quad \Sigma_\beta = diag(100, 100, 100) \\ &p_0 \sim U(0, 1) \quad p_1 \sim U(0, 1 - p_0) \\ &\sigma_b^2 \sim I\Gamma(0/1, 0/1) \end{aligned} \tag{۲.۲}$$

در مدل اول توزیع پیشین پارامتر δ به صورت زیر در نظر گرفته شده است:

$$\delta \sim N_q(0, \Sigma_\delta) \quad \Sigma_\delta = diag(100, 100, 100) \tag{۳.۲}$$

برای مدل دوم توزیع پیشین پارامتر δ به صورت زیر تعریف می‌شود:

$$\begin{aligned} &\delta \sim t_q(0, \Sigma_\delta) \quad \Sigma_\delta = diag(100, 100, 100) \\ &\upsilon \sim exp(a),\ a = 0, 1 \end{aligned} \tag{۴.۲}$$

در مدل سوم توزیع پیشین به صورت زیر به‌کار گرفته شده است:

$$\begin{aligned} &\delta \sim t_q\left(\upsilon, 0, \Sigma_\delta\right) \\ &\upsilon \sim \exp\left(a\right),\ a = 0, 1 \\ &\Sigma_\delta \sim IWishart(\xi, \iota),\ \xi = diag(0/439, 0/591),\ \iota = 5 \end{aligned} \tag{۵.۲}$$

مقادیر پیشنهادی فونگ و همکاران (۲۰۱۰) بر اساس شبیه سازی های انجام شده $\xi = diag(0/439, 0/591),\ \iota = 5$ می باشد، که در این مقاله نیز همان مقادیر در نظر گرفته می شود.

با فرض استقلال پارامترها، چگالی پسین توام عبارت است از:

$$f(\boldsymbol{\eta}, \boldsymbol{\tau}, \beta, \delta, p_0, p_1; \boldsymbol{y}) \propto$$
$$\prod_{i=1}^{n_i} \pi(\boldsymbol{y_i}|\boldsymbol{\tau_i}, \boldsymbol{\eta_i}) \prod_{i=1}^{n_i} \pi(\boldsymbol{\eta_i}|\beta) \prod_{i=1}^{n_i} \pi(\boldsymbol{\tau_i}|\delta)\pi(\beta)\pi(\delta)\pi(p_0)\pi(p_1)$$

از آن جا که به‌دست آوردن توزیع‌های پسین حاشیه‌ای بسیار پیچیده است از الگوریتم MCMC و نمونه‌گیر گیبز استفاده می‌شود. مراحل MCMC برای برآورد پارامترها با استفاده از بسته R2WinBUGS می باشد که دو نرم‌افزار R و WinBUGS را به هم متصل می‌کند.

۳.۲ انتخاب مدل بیزی

معیارهای مختلفی برای انتخاب مدل مناسب در استنباط‌های بیزی تعریف شده است. یکی از معیارهای معمول، آمار ترتیب پیش‌بینی شرطی (CPO)است. فرض کنید Y مجموعه داده‌هاو Y^{-i} به تمام داده‌ها بجز مشاهده i ام اشاره داشته باشد، پارامترهای نامعلوم به صورت Θ تعریف شوند و $\pi(\Theta|Y^{(-i)})$ توزیع پسین بردار پارامتر مدل، یعنی Θ به شرط $Y^{(-i)}$ باشد. برای هر مشاهده i، معیار

$$CPO_i = \int_\Theta f\left(y_i|\theta\right)\pi(\theta|Y^{(-i)})d\theta$$

می‌تواند با

$$CPO_i = \left(\frac{۱}{m}\sum_{k=۱}^{m}\frac{۱}{f(y_i|\theta^{(k)},Y)}\right)^{-۱}$$

به‌دست آید که در آن m تعداد تکرارها بعد از دوره دورریز است (گلفن و دی، ۱۹۹۴). یک آماره CPO_i با لگاریتم درستنمایی حاشیه‌ای‌نما (LPML) تعریف می‌شود به‌طوری‌که:

$$LPML = \sum_{i=۱}^{n} log(CPO_i).$$

مقادیر بزرگ‌تر LPML برازش بهتر به مدل را نشان می‌دهند. برای مقایسه مدل‌ها معیارهای تشخیصی دیگری مانند EAIC و EBIC نیز می‌توان در نظر گرفت (کارلین و لوئیس، ۲۰۰۸). فرض کنید $\theta^{(k)}$ نمونه پسین MCMC تولید شده در تکرار k الگوریتم، η تعداد پارامترها، n تعداد مشاهدات و

$$\bar{D} = \frac{۱}{m}\sum_{k=۱}^{m}\sum_{i=۱}^{n} log\, f(y_i|\theta^{(k)},Y), \quad k = ۱,\ldots,m$$

باشد آنگاه داریم:

$$EAIC = -۲\bar{D} + ۲\eta$$
$$EBIC = -۲\bar{D} + \eta log\, n$$

برعکس LPML مقادیر کوچکتر EAIC و EBIC بیانگر مناسب بودن مدل مربوط است.

۳ مطالعه شبیه سازی

در این قسمت برآورد بیزی پارامترهای مدل و دقت توزیع پیشین‌های معرفی شده مورد ارزیابی و مقایسه قرار می‌گیرند. در تمام حالت‌ها، ۱۰۰ مجموعه داده با اندازه نمونه $n = ۱۰۰$ برای ۵ تکرار تولید شده‌است. برای پارامتر μ_{ij}، پیوند لوجیت به‌صورت:

$$log\, it\left(\mu_{ij}\right) = \beta_۱ + \beta_۲x_{۲ij} + \beta_۳x_{۳ij} + b_i, \quad i = ۱,\ldots,۱۰۰, \quad j = ۱,\ldots,۵$$

و برای پارامتر ϕ_{ij} پیوند لگاریتم به‌صورت:

$$log(\phi_{ij}) = \delta_۱ + \delta_۲x_{۲ij}, \quad i = ۱,\ldots,۱۰۰, \quad j = ۱,\ldots,۵$$

در نظر گرفته شده است، که در آن‌ها b_i دارای توزیع نرمال استاندارد، پارامترهای $p_۰$ و $p_۱$ ثابت و برابر با ۰/۱ و $x_{۲ij}$ و $x_{۳ij}$ متغیرهای مستقل یکنواخت $U(۰,۱)$ در نظر گرفته شده‌اند. برای پارامترهای رگرسیونی $\beta_۱$، $\beta_۲$ و $\beta_۳$ به‌ترتیب مقادیر ۰/۵، ۰/۴ و ۰/۶ و برای پارامترهای رگرسیونی $\delta_۱$ و $\delta_۲$ به ترتیب مقادیر ۰/۳ و ۰/۷ در نظر گرفته شده است.
الگوریتم تولید مقادیر y_{ij} به صورت زیر است:

گام اول- یک مقدار تصادفی d از توزیع برنولی با احتمال ۰/۸ تولید شود.
گام دوم- اگر $d = ۱$، مقدار y_{ij} از توزیع بتای $B(\mu_{ij}\phi_{ij}, (۱ - \mu_{ij})\,\phi_{ij})$ تولید شودو اگر $d = ۰$، مقدار y_{ij} از توزیع برنولی با احتمال موفقیت ۰/۵ تولید شود.
با تکرار این الگو، دنباله‌ای از y_{ij} هایی به‌دست می‌آید که با احتمال ۰/۸ دارای توزیع بتا، با احتمال ۰/۱، مقدار صفر و با احتمال ۰/۱، مقدار یک را اختیار می‌کند. تحلیل‌های بیزی با لحاظ پیشین‌های (۲.۲) و سه مدل معرفی شده، انجام می‌شوند.
برای هر مدل، دو زنجیر با مقادیر اولیه متفاوت، با ۱,۰۰۰,۰۰۰ تکرار برای هر زنجیر اجرا می‌شود. نتایج برای ۵۰۰,۰۰۰ تکرار آخر ارائه می‌شود. به‌علاوه برای اجتناب از مسئله‌ی همبستگی زنجیرهای تولید شده، فاصله زنجیره‌ها ۱۰۰ در نظر گرفته می‌شود.

جدول ۱: برآوردهای بیزی پارامترها پس از برازش مدل در اندازه نمونه‌های متفاوت

پارامتر	مقدار واقعی	مدل ۱ .Est(SD)	مدل ۲ .Est(SD)	مدل ۳ .Est(SD)
$\beta_۱$	۰/۵	۰/۳۶۰ (۰/۱۴۵)	۰/۴۴۶ (۰/۱۳۰)	۰/۴۴۳ (۰/۱۳۵)
$\beta_۲$	۰/۴	۰/۴۸۳ (۰/۲۱۸)	۰/۴۴۰ (۰/۲۳۳)	۰/۴۲۸ (۰/۲۰۵)
$\beta_۳$	۰/۶	۰/۵۴۵ (۰/۱۶۷)	۰/۵۵۴ (۰/۱۷۹)	۰/۵۵۶ (۰/۱۷۳)
$\delta_۱$	۰/۳	۰/۳۱۵ (۰/۱۴۹)	۰/۲۸۸ (۰/۱۴۳)	۰/۳۱۲ (۰/۱۴۲)
$\delta_۲$	۰/۷	۰/۶۷۹ (۰/۲۲۳)	۰/۶۸۹ (۰/۲۶۷)	۰/۶۸۷ (۰/۲۵۸)
$p_۰$	۰/۱	۰/۱۰۳ (۰/۰۱۶)	۰/۱۰۳ (۰/۰۱۳)	۰/۱۰۳ (۰/۰۱۱)
$p_۱$	۰/۱	۰/۱۲۶ (۰/۰۱۷)	۰/۱۱۲ (۰/۰۱۵)	۰/۱۱۱ (۰/۰۱۵)
$\sigma_b^۲$	۱	۰/۷۰۶ (۰/۱۴۹)	۰/۷۳۲ (۰/۱۴۲)	۰/۷۴۳ (۰/۱۳۲)

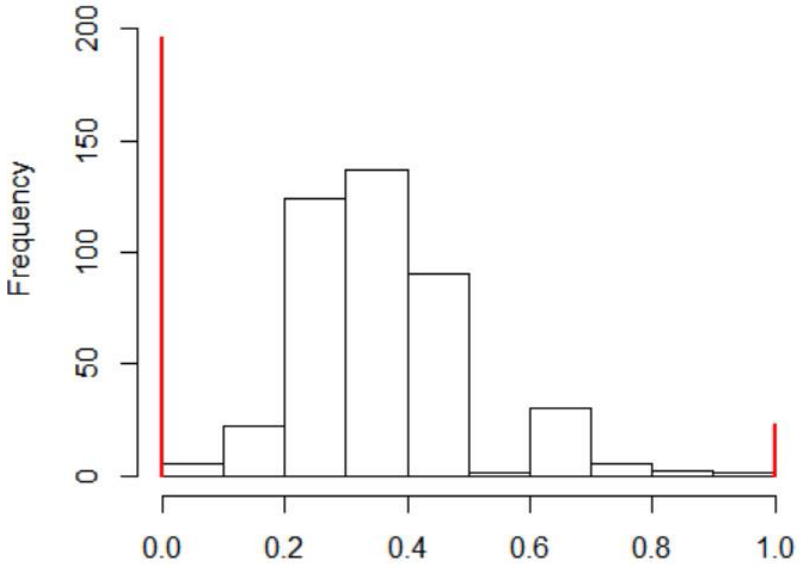
آزمون‌های همگرایی تشخیصی گلمن روبین (۱۹۹۲) و همگرایی تشخیصی هیدل برگ و ولچ (۱۹۸۱) دلالت بر همگرایی زنجیرهای تولید شده دارند. همچنین معیار همگرایی گلمن و روبین توام نیز حدود ۱ به‌دست آمد که این معیار هم همگرایی زنجیرها را نشان می‌دهد. همان طور که در جدول ۱ ملاحظه می‌شود، سه مدل برآوردهای قابل قبولی برای پارامترها ارائه می‌نمایند. البته مدل های ۲ و ۳ نسبت به مدل ۱ از عملکرد بهتری برخوردار هستند و برآوردهای مناسب‌تری برای پارامترها ارائه می‌کنند و مدل ۳ نیز نسبت به مدل ۲ از مقبولیت بیشتری برخوردار است.

۴ تحلیل داده‌های نیروی کار

آمارگیری نیروی کار یکی از مهمترین نمونه‌گیری‌های خانواری دنیا است که جامعه آن همه اعضای خانوارهای معمولی ساکن هستند. در این نمونه‌گیری هر فصل از حدود ۶۰۰۰۰ خانوار آمارگیری می‌شود. برای هر فرد ۱۰ ساله و بیشتر خانوار منتخب نمونه، یک فرم وضع فعالیت تکمیل می‌شود. نمونه‌ها بر اساس الگوی چرخش ۲-۲-۲ انتخاب می‌شوند. یعنی هر خانوار ۲ فصل پی در پی در آمارگیری حضور دارد، در فصل‌های سوم و چهارم در نمونه ها قرار نمی‌گیرد و در دو فصل بعدی پنجم و ششم دوباره به‌عنوان عضو نمونه محسوب می‌شوند. بنابراین برای هر خانوار نمونه ۴ تکرار وجود دارد.برای برآورد پارامترهای جامعه، داده‌هایی که از این آمارگیری برای هر عضو خانوارهای نمونه جمع‌آوری شده است با سه مرحله وزن‌دهی بر اساس وزن پایه، وزن بی‌پاسخی و وزن پیش‌بینی‌های جمعیتی تعدیل می‌شوند. در این آمارگیری بر اساس وضع فعالیت تمام افراد ۱۰ ساله و بیشتر خانوارهای نمونه شاخص‌های مربوط به اشتغال و بیکاری و نسبت شاغلین در هر خانوار تعیین می‌شود. اگر تعداد اعضای شاغل خانوار کمتر از تعداد اعضای خانوار باشد، این نسبت، عددی بین صفر و یک، اگر تمام اعضای خانوار شاغل باشند، عدد یک و در صورتی که هیچ یک از اعضای خانوار شاغل نباشند، عدد صفر را اختیار می‌کند. بنابراین متغیر پاسخ یعنی نسبت شاغلین در خانوار، متغیری پیوسته با تحقق‌هایی در بازهٔ بسته $[۰, ۱]$ است. در این مطالعه داده‌های مربوط به خانوارهای مشترک در فصل‌های اول و دوم سال‌های ۱۳۹۲ و ۱۳۹۳ آمارگیری نیروی کار شهر تهران درنظر گرفته شده است که شامل اطلاعات ۱۵۹ خانوار است. همان طور که در شکل ۱ ملاحظه می‌شود، نسبت شاغلان در خانوار بر اساس نتایج ۴ نوبت آمارگیری، متغیری پیوسته با تحقق‌هایی در بازه بسته $[۰, ۱]$ است.

مدل آمیخته رگرسیون بتای افزوده با $n = ۱۵۹$، برای ۴ تکرار را برای این داده‌ها درنظر می‌گیریم. با فرض ثابت بودن پارامترهای $p_۰$ و $p_۱$ برای پارامتر μ_{ij} پیوند لوجیت به‌صورت

$$log\,it(\mu_{ij}) = \beta_۱ + \beta_۲ x_{۲ij} + \beta_۳ x_{۳ij} + b_i \quad ۱, \ldots, ۱۵۹; \quad j = ۱, \ldots, ۴$$
$$b_i \sim N(۰, \sigma_b^۲)$$



شکل ۱: هیستوگرام نسبت شاغلان درخانوارهای نمونه در فصل‌های اول و دوم سال‌های ۱۳۹۲ و ۱۳۹۳ آمارگیری نیروی کار شهر تهران.

و به‌طور همزمان برای پارامتر ϕ_{ij} نیز پیوند لگاریتم به‌صورت

$$log(\phi_{ij}) = \delta_۱ + \delta_۲ x_{۲ij} \quad i = ۱, \ldots, ۱۵۹; \quad j = ۱, \ldots, ۴$$

درنظر گرفته شده است، که در آن y_{ij} نسبت افراد شاغل در خانوار، $x_{۲ij}$ تعداد افراد ۱۰ ساله و بیشتر در خانوار و $x_{۳ij}$ بعد خانوار است.
بر اساس توزیع‌های پیشین (۲.۲)، سه مدل معرفی شده به داده‌ها برازش شده‌اند. نتایج برای هر مدل، دو زنجیر مستقل با مقادیر اولیه متفاوت، با ۱,۰۰۰,۰۰۰ تکرار و با استفاده از ۵۰۰,۰۰۰ تکرار آخر و احتساب فاصله ۱۰۰ ارائه شده است.

جدول ۲: ارزیابی مدل های ۱ و ۲ و ۳ برای داده‌های نیروی کار

مدل	LPML	EAIC	EBIC
۳	−۲۴٫۱۰	−۴۸٫۱۸	−۱۸٫۱۵
۲	−۲۵٫۲۱	−۴۴٫۱۹	−۱۵٫۵۸
۱	−۲۸٫۱۱	−۳۳٫۷۸	−۱۱٫۲۳

در جدول ۲ عملکرد مدل‌های ۱ و ۲ و ۳ برای داده‌های نیروی کار با استفاده از معیارهای انتخاب مورد ارزیابی و مقایسه قرار گرفته‌اند. همان طور که ملاحظه می‌شود عملکرد مدل ۳، نسبت به سایر مدل‌ها بهتر است.

جدول ۳: برآورد و خطای استاندارد پارامترهای مدل‌های ۱، ۲ و ۳ بر روی داده های آمارگیری نیروی کار شهر تهران برای سالهای ۱۳۹۲ و ۱۳۹۳

پارامتر	مدل ۳				مدل ۲		مدل ۱	
	Est.	(SD)	۲٫۵٪	۹۷٫۵٪	Est.	(SD)	Est.	(SD)
$\beta_۱$	۰٫۶۳۹	۰٫۱۵۱	۰٫۳۴۱	۰٫۷۴۲	۰٫۶۳۸	۰٫۱۹۹	۰٫۴۸۹	۰٫۲۰۳
$\beta_۲$	۰٫۱۸۸	۰٫۰۵۹	۰٫۰۶۷	۰٫۲۲۸	۰٫۶۰۳	۰٫۲۲۸	۰٫۵۷۳	۰٫۲۵۲
$\beta_۳$	−۰٫۵۱۵	۰٫۰۶۷	−۰٫۶۴۵	−۰٫۴۶۹	−۰٫۶۰۶	۰٫۲۱۷	−۰٫۶۴۳	۰٫۲۵۲
$\delta_۱$	۰٫۳۱۰	۰٫۰۱۸	۰٫۲۷۷	۰٫۳۲۲	۰٫۴۳۰	۰٫۱۵۱	۰٫۲۱۲	۰٫۱۵۵
$\delta_۲$	۰٫۳۱۵	۰٫۱۵۱	۰٫۱۸۷	۰٫۶۰۸	۰٫۳۱۰	۰٫۱۵۹	۰٫۸۹۹	۰٫۲۵۸
$p_۰$	۰٫۳۰۹	۰٫۰۱۸	۰٫۲۷۵	۰٫۳۲۱	۰٫۳۰۹	۰٫۰۱۹	۰٫۳۱۱	۰٫۰۲۱
$p_۱$	۰٫۰۳۷	۰٫۰۰۷	۰٫۰۲۴	۰٫۰۴۲	۰٫۰۳۸	۰٫۰۰۸	۰٫۰۳۸	۰٫۰۰۹
$\sigma_b^۲$	۰٫۱۲۶	۰٫۰۲۲	۰٫۰۸۹	۰٫۱۳۹	۰٫۱۵۶	۰٫۱۰۳	۰٫۲۵۵	۰٫۲۱۴

جدول ۳ برآوردهای پارامترهای پسین بدست آمده از برازش مدل‌ها را بر داده‌های آمارگیری نیروی کار نشان می‌دهد. بازه باور ۹۵٪ تنها برای مدل برتر ارائه شده است. نتایج نشان می‌دهند که افراد ۱۰ ساله و بیشتر خانوار با اثر مثبت و متغیر مستقل بعد خانوار بالعکس با اثر منفی در مدل میانگین وارد می‌شود، همچنین افراد ۱۰ ساله و بیشتر خانوار با اثر مثبت بر مدل دقت تاثیرگذار است. پارامترهای p_0 و p_1 که بیانگر خانوارهای فاقد شاغل و خانوارهایی با تمام اعضای شاغل هستند نیز به ترتیب برابر ۰/۳۱ و ۰/۰۴ برآورد شدند. همبستگی داده‌ها طی ۴ فصل، با اثر تصادفی در نظر گرفته شده است و مقدار واریانس آن ۰/۱۳ برآورد شده است.

۵ ارزیابی و مقایسه مدل‌بندی همزمان میانگین و دقت با مدل‌بندی میانگین

حال برآوردهای حاصل از مدل منتخب که برای مدل‌بندی همزمان میانگین و دقت در مدل‌های آمیخته رگرسیون بتای افزوده انتخاب شد با برآوردهایی که از مدل‌بندی میانگین در مدل آمیخته رگرسیون بتای افزوده با اثر تصادفی نرمال به‌دست آمد، ارزیابی و مقایسه می‌شوند.

$$\begin{aligned} &log\,it\,(\mu_{ij}) = \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + b_i,\ \ b_i \sim N\left(0, \sigma_b^2\right),\ i = 1, \ldots, 100,\ j = 1, \ldots, 5, \\ &log\,(\phi_{ij}) = \delta_1 + \delta_2 x_{2ij} \end{aligned} \tag{۱.۵}$$

$$logit\,(\mu_{ij}) = \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + b_i,\ b_i \sim N\left(0, \sigma_b^2\right),\ i = 1, \ldots, 100,\ j = 1, \ldots, 5 \tag{۲.۵}$$

در جدول ۴ عملکرد دو مدل برای برازش به داده‌های نیروی کار براساس معیارهای انتخاب نشان داده می‌شود. مقادیر بزرگ‌تر LPML مربوط به مدل‌بندی همزمان میانگین و دقت با لحاظ اثر تصادفی نرمال است که برازش بهتر مدل را نشان می‌دهد. همین مدل مقادیر کوچکتر EAIC و EBIC که نشان دهنده برازش بهتر به داده‌ها است را نیز ارائه می‌کند. بنابراین مدل‌بندی همزمان میانگین و دقت برازش بهتری نسبت به مدل‌بندی میانگین در مدل‌های آمیخته رگرسیون بتای افزوده بر داده‌های نیروی کار ارائه می‌نماید.

جدول ۴: ارزیابی مدل همزمان میانگین و دقت

مدل بندی	LPML	EAIC	EBIC
میانگین	−۳۰/۵۶۱	−۲۸/۷۶	−۴/۲۱
همزمان میانگین و دقت	−۲۴/۱۰	−۴۸/۱۸	−۱۸/۱۵

۶ نتیجه گیری و پیشنهادات

در این مقاله مدل‌بندی همزمان میانگین و دقت در مدل‌های آمیخته رگرسیون بتای افزوده با پیشین‌های پیشنهادی معرفی شدند و در مطالعات شبیه‌سازی، با لحاظ اریبی و خطای استاندارد برآورد پارامترها، همچنین براساس معیارهای مدل‌های بیزی، بهترین مدل معرفی شد. برای تحلیل داده‌های آمارگیری نیروی کار مرکز آمار ایران، داده‌های مربوط به خانوارهای مشترک در فصل‌های اول و دوم سال‌های ۱۳۹۲ و ۱۳۹۳ شهر تهران، درنظر گرفته و نسبت شاغلین در خانوار براساس دو مدل معرفی شده مدل‌بندی گردید. از آن جا که مهمترین مشکل برآورد وضعیت نیروی کار در سطح کمتر از استان عدم وجود متغیرهای کمکی مناسب است یکی از موارد کاربرد این مدل ساختن متغیر تبیینی مناسب برای برآورد وضعیت نیروی کار نواحی کوچک است زیرا با ضرب سهم شاغلین در خانوار در تعداد خانوار این نواحی، تعداد شاغلین این نواحی قابل برآورد است.

استفاده از الگوریتم های MCMC در مدل های آمیخته رگرسیون بتای افزوده خصوصا وقتی تعداد پارامترها از حالت معمول بیشتر می‌شوند، بسیار زمان‌بر است. بنابراین استفاده از روش‌های تقریب برای کوتاه کردن زمان اجرای برنامه‌ها و همچنین مدل‌بندی همزمان میانگین و دقت در مدل‌های آمیخته رگرسیون بتای افزوده با اثرات تصادفی وابسته فضایی می‌توانند موضوعاتی برای مطالعات بعدی باشند تا بتوان مدل‌های جامع‌تری ارائه نمود.

فهرست منابع

[۱] نتایج آمارگیری نیروی کار ، (۱۳۹۲)، تهران، مرکز آمار ایران.

[2] Branscum, A.J., Johnson, W.O. and Thurmond, M. (2007), *Baysian Beta Regression Applications to Houshold Expenditure Data and Genetic Distance Between Food and Mouth Dieseas Viruses*, Australian & New Zealand Journal of Statistics, 49, 287-301.

[3] Bonat, W.H., Ribeiro, P.J. and Zeviani, W.M. (2013), *Likelihood Analysis for a Class of Beta Mixed*, Models Cornell University Library, arXiv Preprint arXiv: 1312.2413.

[4] Cepeda, E. D., and Gamerman, D. (2005), *Bayesian Methodology for Modeling Parameters in the Two Parameter ExponentialF amily*, Revista Estadística, 57, 168-169.

[5] Cepeda, E. .D, Migon, H. S., Garrido, L. and Achcar, J. A. (2014), *Generalized Linear Models with Random Effects in the Two-ParameterE xponentialF amily*, Journal of Statistical Computation and Simulation, 84, 513-525.

[6] Carlin, B. P. and Louis, T. A. (2008), *Bayesian Methods for Data Analysis*, Mineapolis, CRC Press.

[7] Ferrari, S. and Cribari, F. (2004), *Beta Regression for Modelling Rates and Proportions*, Journal of Applied Statistics,31, 799-815.

[8] Figueroa-Zúñiga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. (2013), *Mixed Beta Regression: A Bayesian Perspective*, Computational Statistics & Data Analysis,61, 137– 147.

[9] Fallah Mohsenkhani, Z., Mohammadzadeh, M. and Baghfalaki, T. (2019), *Augmented Mixed Beta Regression Models with Skew-Normal Independent Distributions: Bayesian Analysis of Labor Force Data*, Communications in Statistics-Simulation and Computation,Volume 48, Issue 7, 2147-2164.

[10] Fong, Y., Rue, H. and Wakefield, J. (2010), *Bayesian Inference for Generalized Linear Mixed Models*, Biostatistics, 11, 397-412.

[11] Galvis, M. D., Dipankar, B. and Victor, H. L. (2014), *Augmented Mixed Beta Regression Models for Periodontal Proportion Data*, Preprinted, (In Press), Statistics in Medicine.

[12] Gelman, A., Rubin, D. B., (1992). *Inference from Iterative Simulation Using Multiple Sequences*, Statistical Science7, 457–511.

[13] Gelfand, A. E. and Dey, D. K. (1994), *Bayesian Model Choice: Asymptotics and Exact Calculations*, Journal of the Royal Statistical Society, Series B (Methodological), 501-514.

[14] Heidelberger, P. and Welch, P. D. (1981), *A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations*, Communications of the ACM, 24, 233-245.

[15] Nogarotto, D, Azevedo, C, Bazan, J.Bayesian (2020), *modeling and prior sensitivity analysis for zero–one augmented beta regression models with an application to psychometric data*, Brazilian Journal of Probability and Statistics, 304-322.

[16] Ospina, R. and Ferrari, S. L. (2010), *Inflated Beta Distributions*, Statistical Papers. 51,111- 126.

[17] Paolino, P. (2001), *Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables*, Political Analysis, 9, 325-346.

[18] Parker, A, Bandyopadhyay, D and Slate, E. (2014), *A spatial augmented beta regression model for periodontal proportion data*, Statistical Modelling, vol. 14, 503-521.

[19] Smithson, M. and Verkuilen, J. (2006), *A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables*, Psychological Methods, 11, 54-71.

[20] Verkuilen, J. and Smithson, M. (2012), *Mixed and Mixture Regression Models for Continuous Bounded Responses Using the Beta Distribution*, Journal of Educational and Behavioral Statistics, 37,82-113.

[21] Zimprich, D. (2010), *Modeling Change in Skewed Variables Using Mixed Beta Regression Models*, Research in Human Development, 7, 9-26.

Simultaneous Modeling of Mean and Variance In Augmented Mixed Beta Regression

Zohreh Fallah Mohsenkhani[1], **Parvin Azhdari**[2] †

(1) Statistical Research and Training Center, Tehran, Iran
(2) Department of Statistics, North Tehran Branch, Islamic Azad University, Tehran, Iran

Communicated by: Mohammad Reza Zadkarami

Received: 2020/8/9 Accepted: 2021/10/31

Abstract: Augmented Beta Regression models are used for modeling data such as rate, ratio or percentage. This model is made by combining the Beta distribution on the interval (0,1) and two degenerate distributions at 0 and 1. By reparameterizing the beta distribution, the mean and precision parameters are modeled with a structure including fixed and random effects. In this paper, simultaneous modeling of mean and precision the augmented mixed beta regression models is presented and the model efficiency in simulation studies by Bayesian approach is investigated. Next, the application of this model to analyze the proportions of employed persons in every household based on the results of the Statistical Centre of Iran is shown and at the end, conclusion and results are presented.

Keywords: Augmented Beta Regression, Precision Parameter, Mixed Model, Bayesian Analysis, Labour Force Survey.

†Corresponding author.
E-mail addresses: zoherhf@yahoo.com (Z. Falah), p_azhdari@iau-tnb.ac.ir (P. Azhdari).