

مجله علوم آماری، بهار و تابستان ۱۳۹۵

جلد ۱۰، شماره ۱، ص ۱۱۳-۱۲۸

DOI: 10.7508/jss.2016.01.007

مدل سازی داده های آمیخته بقا و گسسته با استفاده از تابع مفصل

مینا گذازی، محمد رضا آخوند، عبدالرحمن راسخ

گروه آمار، دانشگاه شهید چمران اهواز

تاریخ دریافت: ۱۳۹۴/۶/۲۰ تاریخ آخرین بازنگری: ۱۳۹۳/۶/۲۹

چکیده: از جمله روش هایی که در سال های اخیر توجه بسیاری از محققان را برای مدل سازی داده های چند متغیره آمیخته به خود جلب کرده است، استفاده از تابع مفصل می باشد. در این مقاله مدلی رگرسیونی برای پاسخ های آمیخته بقا و گسسته بر اساس تابع مفصل ارائه می شود که در آن متغیر پیوسته از نوع زمان بوده و امکان وقوع مشاهده سانسور شده در آن وجود دارد. برای انجام این کار فرض شد که توزیع های حاشیه ای مشخص هستند و متغیری پنهان برای تبدیل حاشیه گسسته به پیوسته مورد استفاده قرار گرفت. سپس با استفاده از تابع مفصل تابع توزیع توانام برای دو متغیر تشکیل و در پایان مدل به دست آمده بر روی داده های فاصله بین تولد ها در شهر اهواز مورد استفاده قرار گرفت.

واژه های کلیدی: تابع مفصل، مشاهده سانسور شده، داده های آمیخته، متغیر پنهان، فاصله بین تولد ها.

آدرس الکترونیک مسئول مقاله: Mr.Akhoond@Scu.ac.ir

کد موضوع بنای ریاضی (۲۰۱۰) ۶۲N۰۱ و ۶۲J۱۲

۱ مقدمه

در آمار تحلیل رگرسیونی یک روش برای برآورد روابط بین متغیرهاست که شامل روش‌های بسیاری برای مدل‌بندی و تحلیل متغیرهای متعدد می‌شود. در شرایطی که بیش از یک متغیر پاسخ وجود داشته باشد، اگر هر دو متغیر پاسخ کمی یا هر دو کیفی باشند از روش‌های آماری استاندارد نظری رگرسیون دومتغیره یا رگرسیون لوژستیک دو متغیره برای مدل‌بندی روابط بین متغیرهای پاسخ و کمکی می‌توان استفاده نمود. اما در حالتی که متغیرها به صورت آمیخته هستند بدین معنی که تعدادی از متغیرهای پاسخ کمی و تعدادی دیگر کیفی باشند، ساده‌ترین شیوه مدل‌بندی، بررسی هر پاسخ به صورت مجزا در یک چارچوب یک متغیره است. از آن جا که این روش همبستگی بین پاسخ‌ها را نادیده می‌گیرد، از کارایی لازم برخوردار نیست. از طرفی زمانی که با پاسخ‌های آمیخته سر و کار داریم، با توجه به محدودیتی که برای در نظر گرفتن توزیع احتمال مناسب برای بردار پاسخ در این گونه موارد وجود دارد، روش‌های دومتغیره معمول کارایی لازم را ندارند (سدھی و همکاران، ۱۳۸۹). در سال‌های اخیر برای مدل‌سازی و پیش‌بینی پاسخ‌های آمیخته، تحقیقاتی در آمار کلاسیک صورت گرفته است. برای تعیین مدل رگرسیونی چندمتغیره برای پاسخ‌های آمیخته گسسته و پیوسته، یکی از ابتدایی‌ترین پیشنهادها روش فاكتورگیری^۱ است. این روش توزیع توام را به وسیله تجزیه آن به یک توزیع شرطی از یک مجموعه از پاسخ‌ها و یک توزیع حاشیه‌ای از یک مجموعه دیگر تعیین می‌کند. یک اشکال روش فاكتورگیری این است که این مدل‌ها در شرایطی که پاسخ‌های آمیخته چندمتغیره هستند، به آسانی گسترش نمی‌یابند. اشکال دیگر این است که چون پارامترها بسته به تجزیه مورد استفاده تفسیرهای متفاوتی دارند، مدل‌های به دست آمده قابل مقایسه نیستند و حتی ممکن است تحت تجزیه‌های مختلف، برآوردهای کاملاً متفاوتی از همبستگی به دست آید و یا حتی عبارت ساده‌ای برای همبستگی بین پاسخ‌ها محاسبه نشود (دلنو و کریم‌کاچ، ۲۰۱۰؛ قره آقاجی اصل و همکاران، ۱۳۸۶).

مدل‌سازی داده‌های آمیخته با دیدگاه متغیرهای پنهان نیز توسط کاتالانو و ریان (۱۹۹۲)، رگان و کاتالانو (۲۰۰۲) و یانگ و همکاران (۲۰۰۷) مورد مطالعه قرار گرفته است. مولنیرگز و همکاران (۲۰۰۱) نیز به جای استفاده از توزیع نرمال چندمتغیره برای متغیر پنهان از توزیع پلاکت- دیل^۲ استفاده کرده‌اند. یکی دیگر از روش‌های مورد استفاده، مدل با اثرات تصادفی^۳

^۱ Factorization

^۲ Plackett- Dale

^۳ Random Effects

است که اثرات تصادفی همبسته^۴ یا مشترک^۵ را وارد مدل می‌کند تا همیستگی بین پاسخ‌های آمیخته در مدل توان را نشان دهد. مونکین و تریوودی (۱۹۹۹) این روش را در پژوهشی به کار بردند. کاربردهای این روش برای تحلیل داده‌های آمیخته با ابعاد زیاد نیز توسط فیس و همکاران (۲۰۰۸) ارائه شده است.

از جمله روش‌هایی که در سال‌های اخیر توجه بسیاری از محققان را برای مدل‌سازی داده‌های چندمتغیره به خود جلب کرده است، مدل‌سازی داده‌ها با استفاده از تابع مفصل می‌باشد. مطالعه تابع مفصل و کاربردهای آن در آمار یک پدیده نسبتاً جدید است و در سال‌های اخیر با افزایش توجه به استفاده از توابع مفصل کتاب‌هایی در این زمینه نوشته شده است که از مهم‌ترین آن‌ها می‌توان جو (۱۹۹۷) و نلسن (۲۰۰۶) را نام برد.

از توابع مفصل برای مدل‌سازی داده‌ها زمانی که پاسخ‌ها گسسته، پیوسته و همچنین هنگامی که پاسخ‌ها به صورت آمیخته هستند، می‌توان استفاده کرد. سانگ و همکاران (۲۰۰۹)، تریوودی و زیمر (۲۰۰۶) و کامرون و همکاران (۲۰۰۴) توابع مفصل را برای مدل‌سازی داده‌های گسسته چندمتغیره مورد استفاده قرار دادند. همچنین دلثون و وو (۲۰۱۱)، کرایو و ثابتی (۲۰۱۲) و کرامر و همکاران (۲۰۱۳) این توابع را برای مدل‌سازی داده‌های چندمتغیره آمیخته‌ی گسسته و پیوسته به کار گرفتند. دلثون و وو (۲۰۱۱) در مقاله خود دو حالت را مورد بررسی قرار دادند. در حالت اول متغیر گسسته، متغیری دودویی و متغیر پیوسته دارای توزیع نرمال بود و در حالت دوم متغیر گسسته را به صورت رتبه‌ای و دارای سه رده و توزیع نمایی را برای متغیر پیوسته به کار گرفتند. در این مدل‌ها متغیر پیوسته فاقد مشاهدات سانسور شده است. اما در بسیاری از کاربردها متغیر پاسخ مورد نظر از نوع زمان می‌باشد و به عنوان مثال طول عمر افراد یا اشیا را نشان می‌دهد. اگر در بین مشاهدات، داده‌های سانسور شده داشته باشیم، استفاده روش‌های رگرسیونی معمول به دلیل وجود مشاهدات سانسور شده کارایی لازم را نخواهد داشت. در چنین شرایطی برای برآورده پارامترها لازم است تا در تشکیل تابع درستنمایی مشاهدات سانسور شده نیز در نظر گرفته شوند. با توجه به این که هنگامی که داده‌ها به صورت پیوسته و از نوع زمان باشند امکان وقوع مشاهدات سانسور شده در داده‌ها وجود دارد، در این تحقیق، مطالعه دلثون و وو (۲۰۱۱) برای پاسخ‌های آمیخته گسسته-پیوسته به حالتی گسترش داده می‌شود که متغیر پیوسته از نوع زمان بوده و در آن امکان وقوع مشاهدات سانسور شده وجود داشته باشد. در

^۴ Correlated

^۵ Shared

۱۱۶ مدل‌سازی داده‌های آمیخته بقا و گسسته با استفاده از تابع مفصل

این مدل که به طور کامل توسط توزیع‌های حاشیه‌ای و تابع مفصل مشخص می‌شود، پارامترهای حاشیه‌ای α و β تفسیرهایی مشابه با تفسیرهای حاشیه‌ای دارند. همچنین به دست آوردن توزیع شرطی پاسخ‌ها که یک نیاز مهم برای موقعیت‌هایی است که رفتار شرطی پاسخ‌ها مورد توجه است، امکان‌پذیر می‌شود. در این مقاله علاوه بر استفاده از تابع مفصل نرمال که در مطالعه دلشون و وو (۲۰۱۱) مورد استفاده قرار گرفته است از تابع مفصل فرانک نیز برای مدل‌سازی توابع داده‌ها استفاده خواهد شد. همچنین نتایج به دست آمده با مفصل حاصل ضرب که نشان دهنده استقلال میان متغیرها است مورد مقایسه قرار خواهد گرفت. بدین منظور در بخش ۲ به تعریف تابع مفصل پرداخته می‌شود. سپس تابع درستنمایی برای تحلیل داده‌های آمیخته همبسته در حالتی که یکی از متغیرها گسسته و دیگری پیوسته باشد و امکان وقوع مشاهده سانسور شده برای متغیر پیوسته وجود داشته باشد با استفاده از توابع مفصل نرمال، فرانک و حاصل ضرب، تشکیل داده می‌شود. به دلیل وجود وابستگی منفی میان متغیرهای فاصله تولد فرزند اول و تعداد مطلوب فرزندان از دیدگاه مادر از توابع مفصل جامع نرمال و فرانک برای مدل‌سازی داده‌ها استفاده خواهد شد. در بخش ۳ برآوردهای پارامترها با استفاده از روش ماکسیمم درستنمایی به دست می‌آید. در بخش ۴ روش پیشنهاد شده، برای تحلیل داده‌های فاصله بین تولدات در شهر اهواز مورد استفاده قرار می‌گیرد و در بخش ۵ به بحث در مورد نتایج به دست آمده و نتیجه‌گیری در مورد مدل پرداخته خواهد شد.

۲ مدل‌سازی رگرسیونی بر اساس تابع مفصل

براساس قضیه اسکلار برای هر تابع توزیع توانم (\cdot, \cdot) از یک جفت متغیر تصادفی $F_{X,Y}(\cdot, \cdot)$ با توابع توزیع حاشیه‌ای (\cdot) و $F_X(\cdot)$ ، تابع مفصل دو متغیره C وجود دارد، به طوری که برای هر دو عدد حقیقی x و y رابطه

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

برقرار است. علاوه بر این اگر X و Y متغیرهای تصادفی پیوسته باشند، آن گاه تابع مفصل C یکتاست و بر عکس آن نیز برقرار است، یعنی برای هر دو توزیع یک متغیره (\cdot) و $F_X(\cdot)$ و $F_Y(\cdot)$ هر تابع مفصل C ، تابع (\cdot, \cdot) یک تابع توزیع دو متغیره با توزیع‌های حاشیه‌ای (\cdot) و $F_X(\cdot)$ است.

خانواده‌های مختلفی از توابع مفصل در تحلیل‌ها مورد استفاده قرار می‌گیرند که دو خانواده مهم آن‌ها، تابع مفصل ارشمیدسی و تابع مفصل بیضوی هستند.

تابع درستنمایی برای داده‌های آمیخته: متغیرهای پاسخ آمیخته همبسته X_i و Y_i را در نظر بگیرید، که Y_i پیوسته و X_i گسسته هستند. فرض کنید $X_i \sim F_{X_i}$ و $Y_i \sim F_{Y_i}$ باشد. به علاوه فرض کنید $X_i \sim F_{X_i}$ از $i = 1, \dots, N$ مقدار متفاوت دارد مثلاً s_0, \dots, s_E که می‌توانند رتبه‌های اعداد را نشان دهند. برای مدل‌بندی توزیع توانام F_{X_i, Y_i} از X_i و Y_i ، فرض کنید $Y_i^* \sim F_{Y_i^*}$ متغیر پنهان پیوسته مشاهده نشده برای X_i باشد، به طوری که ارتباط بین X_i و Y_i^* با استفاده از رابطه

$$X_i = \begin{cases} s_0 & Y_i^* \in (-\infty, \gamma_1) \\ \vdots & \vdots \\ s_k & Y_i^* \in [\gamma_k, \gamma_{k+1}) \\ \vdots & \vdots \\ s_E & Y_i^* \in [\gamma_E, \infty) \end{cases} \quad (1)$$

مشخص می‌شود، که در آن $\gamma_E < \gamma_1 < \dots < \gamma_{E+1} = +\infty$ و $\gamma_0 = -\infty$ می‌باشد.

در حالت دو متغیره زمانی که متغیر پیوسته زمان بوده و امکان وقوع مشاهدات سانسور شده از راست در آن وجود داشته باشد، تابع درستنمایی برای بردار پارامتر $(\theta_1, \theta_2; \alpha) = \theta$ به صورت

$$\ell(\theta) = \prod_{i=1}^N f(x_i, y_i; \theta_1, \theta_2)^{c_i} f_1(x_i, y_i; \theta_1, \theta_2)^{(1-c_i)} \quad (2)$$

به دست می‌آید، که در آن c_i نشان‌دهنده وضعیت سانسور شدن است، برای افراد با مشاهدات سانسور نشده $c_i = 1$ و در غیر این صورت $c_i = 0$ در نظر گرفته می‌شود. همچنین $f(x_i, y_i; \theta_1, \theta_2) = \partial P(X_i = x, Y_i \leq y) / \partial y$ و $f_1(x_i, y_i; \theta_1, \theta_2) = P(X_i = x, Y_i > y)$ می‌باشد. در تمام حالت‌های زیر فرض می‌شود $u_1^k = F_{Y_i^*}(\gamma_k; \mathbf{Z}_{1i}, \beta, \theta_1)$ و $u_2^k = F_{Y_i^*}(\gamma_k; \mathbf{Z}_{2i}, \beta, \theta_2)$ برای $i = 1, \dots, N$ (برای $x_i, y_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}$) مشاهده شده با بردارهای متغیرهای کمکی \mathbf{Z}_{1i} و \mathbf{Z}_{2i} هستند. همچنین θ به عنوان بردار پارامترها شامل پارامترهای رگرسیونی α و β و پارامتر وابستگی θ و پارامترهای حاشیه‌ای θ_1 و θ_2 به ترتیب از $F_{Y_i^*}$ و F_{Y_i} در نظر گرفته می‌شود.

تابع درستنمایی با استفاده از مفصل نرمال: تابع مفصل نرمال، عضوی از خانواده توابع مفصل بیضوی است و مدل‌های چندمتغیره‌ای را تولید می‌کند که بسیاری خصوصیات مشابه

با توزیع نرمال چندمتغیره را دارا هستند. مفصل نرمال به صورت

$$\begin{aligned} C(u_1, u_2; \rho) &= \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho) \\ &= \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \exp\left\{-\frac{(s^2 - 2\rho st + t^2)}{2(1-\rho^2)}\right\} ds dt \end{aligned} \quad (3)$$

تعریف می‌شود، که در آن Φ تابع توزیع نرمال استاندارد، Φ_2 تابع توزیع نرمال استاندارد دومتغیره و ρ پارامتر وابستگی است که ضریب همبستگی پیرسون بین نمره‌های نرمال (u_1, u_2) و $\Phi^{-1}(u_1), \Phi^{-1}(u_2)$ را نشان می‌دهد و در فاصله $[0, 1]$ قرار دارد.

با توجه به معادله (3) تابع توزیع توان $F_{Y_i^*, Y_i}(y)$ با استفاده از مفصل نرمال به صورت

$$F_{Y_i^*, Y_i}(\gamma_k, y_i) = \Phi_2(\Phi^{-1}\{F_{Y_i^*}(\gamma_k)\}, \Phi^{-1}\{F_{Y_i}(y_i)\}; \rho)$$

به دست می‌آید. با توجه به معادله (2) تابع درستنمایی شامل دو بخش، برای داده‌های سانسور نشده و سانسور شده است. برای افرادی که سانسور رخ نداده چگالی توان $f_{X_i, Y_i}(x, y)$ با استفاده از معادله زیر به دست می‌آید:

$$f_{X_i, Y_i}(x, y) = \begin{cases} \frac{\partial F_{Y_i^*, Y_i}(\gamma_1, y)}{\partial y} & x = s_0 \\ \vdots & \vdots \\ \frac{\partial F_{Y_i^*, Y_i}(\gamma_{k+1}, y)}{\partial y} - \frac{\partial F_{Y_i^*, Y_i}(\gamma_k, y)}{\partial y} & x = s_k \\ \vdots & \vdots \\ f_{Y_i}(y) - \frac{\partial F_{Y_i^*, Y_i}(\gamma_E, y)}{\partial y} & x = s_E \end{cases}$$

در نتیجه:

$$f_{X_i, Y_i}(x, y) = \begin{cases} \Phi\left(\frac{q_1^k - \rho q_2}{\sqrt{1-\rho^2}}\right) f_{Y_i}(y) & x = s_0 \\ \vdots & \vdots \\ \{\Phi\left(\frac{q_{k+1}^k - \rho q_2}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{q_k^k - \rho q_2}{\sqrt{1-\rho^2}}\right)\} f_{Y_i}(y) & x = s_k \\ \vdots & \vdots \\ \{1 - \Phi\left(\frac{q_E^k - \rho q_2}{\sqrt{1-\rho^2}}\right)\} f_{Y_i}(y) & x = s_E \end{cases} \quad (4)$$

که در آن (u_{2i}, u_{1i}) همچنین برای افرادی که سانسور رخ داده

است تابع چگالی با استفاده از معادله زیر به دست می‌آید:

$$P(X_i = x, Y_i > y) = \begin{cases} F_{Y_i^*}(\gamma_1) - F_{Y_i^*, Y_i}(\gamma_1, y) & x = s_0 \\ \vdots & \vdots \\ F_{Y_i^*}(\gamma_{k+1}) - F_{Y_i^*, Y_i}(\gamma_k, y) + F_{Y_i^*, Y_i}(\gamma_k, y) & x = s_k \\ -F_{Y_i^*, Y_i}(\gamma_{k+1}, y) & \\ \vdots & \vdots \\ 1 - F_{Y_i^*}(\gamma_E) + F_{Y_i^*, Y_i}(\gamma_E, y) - F_{Y_i}(y) & x = s_E \end{cases} \quad (5)$$

در نتیجه تابع درستنمایی برای داده‌های کامل (هم سانسور شده و هم سانسور نشده) براساس

معادله (۲) و با استفاده از معادلات (۴) و (۵) به صورت

$$\begin{aligned} l(\theta) &= \prod_{k=0}^E \prod_{\forall x_i=s_k} \left(\left[\left(\Phi \left\{ \frac{q_{\gamma_i}^{k+1}(Z_{\gamma_i}, \beta, \theta_1) - \rho q_{\gamma_i}(Z_{\gamma_i}, \alpha, \theta_2)}{\sqrt{1-\rho^2}} \right\} \right. \right. \right. \\ &\quad \left. \left. \left. - \Phi \left\{ \frac{q_{\gamma_i}^k(Z_{\gamma_i}, \beta, \theta_1) - \rho q_{\gamma_i}(Z_{\gamma_i}, \alpha, \theta_2)}{\sqrt{1-\rho^2}} \right\} \right) f_{Y_i}(y_i) \right]^{(c_i)} \right. \\ &\quad \times \left. \left. \left. [u_{\gamma_i}^{k+1} - u_{\gamma_i}^k + F_{Y_i^*, Y_i}(\gamma_k, y_i) - F_{Y_i^*, Y_i}(\gamma_{k+1}, y_i)]^{(1-c_i)} \right) \right] \end{aligned}$$

به دست می‌آید، که در آن $(F_{Y_i^*, Y_i}(\gamma_0, y_i) = 0, F_{Y_i^*, Y_i}(\gamma_{E+1}, y_i) = F_{Y_i}(y_i))$

تابع چگالی حاشیه‌ای $f_{Y_i}(y_i)$ و $\Phi(\frac{q_{\gamma_i}^{E+1}-\rho q_{\gamma_i}}{\sqrt{1-\rho^2}}) = 1, \Phi(\frac{q_{\gamma_i}^0-\rho q_{\gamma_i}}{\sqrt{1-\rho^2}}) = 0, u_{\gamma_i}^0 = 0$

در نتیجه لگاریتم تابع درستنمایی عبارتست از:

$$\begin{aligned} \ell(\theta) &= \sum_{k=0}^E \sum_{\forall x_i=s_k} [(c_i) \log \left(\Phi \left\{ \frac{q_{\gamma_i}^{k+1}(Z_{\gamma_i}, \beta, \theta_1) - \rho q_{\gamma_i}(Z_{\gamma_i}, \alpha, \theta_2)}{\sqrt{1-\rho^2}} \right\} \right. \right. \\ &\quad \left. \left. - \Phi \left\{ \frac{q_{\gamma_i}^k(Z_{\gamma_i}, \beta, \theta_1) - \rho q_{\gamma_i}(Z_{\gamma_i}, \alpha, \theta_2)}{\sqrt{1-\rho^2}} \right\} \right) + \log \{f_{Y_i}(y_i)\}] \\ &\quad + (1-c_i) [\log \{u_{\gamma_i}^{k+1} - u_{\gamma_i}^k + F_{Y_i^*, Y_i}(\gamma_k, y_i) - F_{Y_i^*, Y_i}(\gamma_{k+1}, y_i)\}] \end{aligned}$$

تابع درستنمایی با استفاده از مفصل فرانک: تابع توزیع توان $F_{Y_i^*, Y_i}(\cdot, \cdot)$ با استفاده از مفصل فرانک به صورت

$$F_{Y_i^*, Y_i}(\gamma_k, y_i) = C(u_{\gamma_i}^k, u_{\gamma_i}; \theta) = -\frac{1}{\theta} \log \left\{ 1 + \frac{(e^{-\theta u_{\gamma_i}^k} - 1)(e^{-\theta u_{\gamma_i}} - 1)}{e^{-\theta} - 1} \right\}$$

است. در نتیجه لگاریتم تابع درستنمایی برای کل داده‌ها و بر اساس روش مورد استفاده در بخش قبل با استفاده از مفصل فرانک به صورت

$$\ell(\theta) = \sum_{k=0}^E \sum_{\forall x_i=s_k} [(c_i) \log \left[\frac{f_{Y_i}(y_i) e^{-\theta u_{\gamma_i}} [e^{-\theta u_{\gamma_i}^{k+1}} - 1]}{(e^{-\theta} - 1) + (e^{-\theta u_{\gamma_i}^{k+1}} - 1)(e^{-\theta u_{\gamma_i}} - 1)} \right] \right]$$

$$-\frac{f_{Y_i}(y_i)e^{-\theta u_{\gamma_i}}[e^{-\theta u_{\gamma_i}^k}-1]}{(e^{-\theta}-1)+(e^{-\theta u_{\gamma_i}^k}-1)(e^{-\theta u_{\gamma_i}}-1)} \\ + (1-c_i)[\log \{u_{\gamma_i}^{k+1}-u_{\gamma_i}^k+F_{Y_i^*, Y_i}(\gamma_k, y_i)-F_{Y_i^*, Y_i}(\gamma_{k+1}, y_i)\}]$$

قابل محاسبه است، که در آن $F_{Y_i^*, Y_i}(\gamma_0, y_i) = F_{Y_i^*}(y_i)$ و $u_{\gamma_i}^{E+1} = 1$

تابع درستنمایی با استفاده از مفصل حاصل ضرب: مفصل حاصل ضرب به دلیل اینکه معادل استقلال میان دو متغیر است به عنوان یک معیار مناسب مقایسه مورد استفاده قرار می‌گیرد.

تابع توزیع توان $F_{Y_i^*, Y_i}$ با استفاده از مفصل حاصل ضرب به صورت

$$F_{Y_i^*, Y_i}(\gamma_k, y_i) = C(u_{\gamma_i}^k, u_{\gamma_i}; \theta) = u_{\gamma_i}^k u_{\gamma_i}$$

به دست می‌آید. در نتیجه لگاریتم تابع درستنمایی برای کل داده‌ها و بر اساس روش مورد استفاده در بخش قبل با استفاده از مفصل حاصل ضرب به صورت

$$\ell(\theta) = \sum_{k=0}^E \sum_{\forall x_i=s_k} [(c_i)[\log \{u_{\gamma_i}^{k+1}-u_{\gamma_i}^k\} + \log \{f_{Y_i}(y_i)\}] \\ + (1-c_i)[\log \{u_{\gamma_i}^{k+1}-u_{\gamma_i}^k+F_{Y_i^*, Y_i}(\gamma_k, y_i)-F_{Y_i^*, Y_i}(\gamma_{k+1}, y_i)\}]]$$

به دست می‌آید، که در آن $F_{Y_i^*, Y_i}(\gamma_0, y_i) = F_{Y_i^*}(y_i)$ و $u_{\gamma_i}^{E+1} = 1$

برآورده

در این مطالعه حالتی بررسی می‌شود که در آن متغیر پیوسته دارای توزیع واپیول با پارامترهای μ_{γ_i} و δ است و $(Y_i \sim N(\mu_{\gamma_i}, \sigma^2))$ است که X_i متوسط رابطه (۱) با y_i مرتبط می‌شود. بنابراین تابع چگالی حاسیه‌ای $f_{Y_i}(y)$ و $f_{Y_i^*}(y^*)$ برای متغیرهای Y_i و Y_i^* به صورت

$$f_{Y_i}(y) = \frac{\delta}{\mu_{\gamma_i}} \left(\frac{y}{\mu_{\gamma_i}} \right)^{\delta-1} \exp \left\{ -\left(\frac{y}{\mu_{\gamma_i}} \right)^\delta \right\} \\ f_{Y_i^*}(y^*) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y^* - \mu_{\gamma_i})^2}{2\sigma^2} \right\}$$

همستند، که در آن $\hat{\theta} = \text{exp}(\mathbf{Z}_{\gamma_i}^T \boldsymbol{\beta})$ و $\mu_{\gamma_i} = \mathbf{Z}_{\gamma_i}^T \boldsymbol{\alpha}$. برآورد ماکسیمم درستنمایی $\hat{\theta}$ از حل معادله $\ell'(\theta) = 0$ به روش نیوتون-رافسون به دست می‌آید. با توجه به خواص برآورده ماکسیمم درستنمایی سازگار است و به طور مجانبی دارای توزیع نرمال چندمتغیره

با میانگین θ و کواریانس ماتریسی است که توسط معکوس ماتریس اطلاع فیشر $I(\theta) = E\{-\ell''(\theta)\}$ به دست می‌آید. همچنین انحراف معیار $\hat{\theta}$ با استفاده از عناصر روی قطره $\{\ell'(\hat{\theta})\ell'^T(\hat{\theta})\}^{-1}$ یا $\{\ell''(\hat{\theta})\ell'^T(\hat{\theta})\}^{-1}$ به دست می‌آید.

۳ مثال کاربردی

باروری یکی از عوامل مهم نوسانات جمعیت در سطح منطقه‌ای، ملی و بین‌المللی است که تعداد و ساختار جمعیت یک کشور را تعیین می‌کند. فاصله بین تولدات، یک معیار و شاخص مهم در موضوع باروری است که علاوه بر تضمیم زوجین، از عوامل اقتصادی، فرهنگی و سیاسی نیز تاثیر می‌پذیرد. مطالعات بسیاری در رابطه با فاصله بین تولدات انجام شده است و عوامل متفاوتی به عنوان عوامل موثر بر فاصله بین تولدات گزارش شده‌اند. بهمنظور استفاده از مدل ارائه شده در بخش قبل بر روی داده‌های مطالعه فاصله بین تولدات در شهر اهواز استفاده شده است. در این داده‌ها فاصله میان ازدواج تا تولد اولین فرزند زنده (بدون توجه به سقط شده‌ها، مرده‌زایی‌ها و بچه‌های مرد در این فاصله) به عنوان متغیر پاسخ پیوسته که امکان وقوع سانسور در آن وجود دارد و تعداد مطلوب فرزندان از دیدگاه مادر به عنوان متغیر پاسخ گسترش دارد. برای زنانی که ازدواج کرده‌اند ولی هنوز کودک زنده‌ای را به دنیا نیاورده‌اند فاصله بین ازدواج تا پایان زمان نمونه‌گیری را به عنوان فاصله تولد تعریف کرده و این فاصله سانسور شده در نظر گرفته شد.

پژوهش از طریق دادن پرسشنامه به زنان متاهل صورت گرفته است و داده‌ها به کمک نرم افزار R و با استفاده از برنامه نوشته شده توسط محقق، مورد تحلیل قرار گرفته‌اند. برای بررسی اثر فاکتورهای مختلف ابتداء لازم است توزیع مناسب برای متغیر پاسخ در نظر گرفته شود. یک مدل رگرسیونی وایبول برای فاصله تولد و یک مدل رگرسیونی نرمال برای متغیر پنهان تعداد مطلوب فرزندان در نظر گرفته شد. بهمنظور بررسی نیکویی برآش توزیع وایبول به داده‌های فاصله تولد از آزمون کلموگروف- اسمیرنوف استفاده شد و p -مقدار^۶ برابر ۰/۱۸۶۱ به دست آمد که برآش مناسب توزیع وایبول به داده‌ها را نشان می‌دهد. همچنین واریانس توزیع پنهان به دلیل امکان عدم شناسایی‌پذیر بودن مدل، یک در نظر گرفته شده است. در مورد متغیرهای کمکی اسمی یکی از رده‌ها به عنوان سطح مرجع در نظر گرفته شده است و سایر رده‌ها با آن مورد مقایسه قرار گرفته‌اند.

^۶ P-value

۱۲۲ مدل‌سازی داده‌های آمیخته بقا و گسسته با استفاده ازتابع مفصل

یکی از روش‌های انتخاب مناسب‌ترین مدل، استفاده از روش پیشرو^۷ با استفاده از معیار اطلاع آکائیک^۸ است. در این روش مدل دارای کمترین آکائیک به عنوان مناسب‌ترین مدل انتخاب می‌شود. در این تحقیق ابتدا با استفاده از روش پیشرو از میان متغیرها، متغیرهای دارای p -مقدار کمتر از ۱٪ برای مدل رگرسیونی حاشیه‌ای مربوط به فاصله تولد اول انتخاب شدند. سپس این متغیرهای معنی‌دار را در مدل توام ساخته شده از مفصل وارد کرده و در حالت توام با استفاده از روش پیشرو متغیرهای معنی‌دار برای تعداد مطلوب فرزندان از دیدگاه مادر به دست آمدند.

مقایسه مدل‌ها

برای مقایسه دوتابع مفصل، آزمون نسبت درستنمایی وونگ^۹ برایفرض‌های غیرآشیانه‌ای^{۱۰} استفاده می‌شود. این آزمون وقتی مناسب است که مدل‌ها غیرآشیانه‌ای باشند، یعنی مدل رگرسیونی به دست آمده از یک تابع مفصل نتواند توسط یک محدودیت، از مدل رگرسیونی برای تابع مفصل دیگر به دست آید. فرض کنید^(۱) ℓ و^(۲) ℓ' به ترتیب نشان‌دهنده بردارهای لگاریتم درستنمایی نقطه‌ای^{۱۱} برای یک مدل با تابع مفصل ۲ و ۱ باشند. در اینجا فرض می‌شود هر دو مدل درجه آزادی‌های مشابه دارند یعنی، تعداد پارامترهای آن‌ها یکسان است. حال تفاوت لگاریتم درستنمایی نقطه‌ای به صورت

$$m_i = \ell_i^{(1)} - \ell_i^{(2)}, i = 1, \dots, n$$

محاسبه می‌شود. در این حالت با استفاده از میانگین^{۱۰} m آماره آزمون به صورت

$$T_V = \frac{\sqrt{n} \bar{m}}{\sqrt{\sum_{i=1}^n (m_i - \bar{m})^2}}$$

به دست می‌آید که به طور مجانبی دارای توزیع نرمال با میانگین صفر و واریانس یک است (کرامر و همکاران، ۲۰۱۳). در داده‌های موجود^(۱) ℓ و^(۲) ℓ' به ترتیب بردارهای لگاریتم درستنمایی نقطه‌ای برای مدل با تابع مفصل نرمال و فرانک در نظر گرفته می‌شود. به ازای $n = ۹۱۵$ داده موجود^(۱) $\bar{m} = ۰/۰۰۳۳$ و در نتیجه^(۲) $T_V = ۰/۰۶۵$ به دست آمد. اگر سطح معنی‌داری^(۱) $\alpha = ۰/۰۵$ و فرض صفر برابری هر دوتابع مفصل باشد، چون p -مقدار برابر

^۷ Forward

^۸ Akaike Information Criterion

^۹ Vuong

^{۱۰} Non-nested

^{۱۱} Pointwise Loglikelihood

۰/۹۴۸۲ به دست می‌آید، در نتیجه فرض صفر مبنی بر تساوی هر دوتابع مفصل پذیرفته می‌شود. برای انتخاب بهترین تابع مفصل از معیار اطلاع آکائیک (AIC) استفاده شده است.

جدول ۱: مقادیر آکائیک برای مدل رگرسیونی توان بر اساس مفصل‌های مختلف

نرمال	تابع مفصل	AIC
۹۲۳۰/۸۴۲		
۹۲۳۰/۰۰۰	فرانک	
۹۲۳۶/۳۹۲	حاصل ضرب	

بر اساس نتایج جدول ۱ بیشترین میزان آکائیک مربوط به تابع مفصل حاصل ضرب است، در نتیجه مدل‌سازی با در نظر گرفتن ساختار وابستگی میان متغیرها مناسب‌تر از مدل‌سازی بر اساس فرض استقلال است. با توجه به اینکه تابع مفصل فرانک دارای کمترین میزان معیار اطلاع آکائیک بود، در ادامه برآورد پارامترهای این تابع ارائه می‌شود.

در جدول ۲ ضرایب همبستگی ناپارامتری (ρ_s) اسپیرمن و τ تاو-کندال برای توابع مفصل فرانک و نرمال آورده شده است. این ضرایب برای تابع مفصل فرانک به صورت $[1 - D_1(\theta)] - [D_2(\theta)] = 1 - \frac{4}{\theta}$ است، که در آن‌ها $D_k(x) = \frac{k}{x^k} \int_0^x \frac{t^k}{(e^t - 1)} dt$ همچنین برای تابع مفصل نرمال ضرایب همبستگی ناپارامتری اسپیرمن و τ تاو-کندال به صورت $\rho_s = 6 \sin^{-1}(\rho/\sqrt{2})/\pi$ و $\tau = 2 \sin^{-1}(\rho/\sqrt{2})/\pi$ محاسبه می‌شوند. انحراف استاندارد این ضرایب همبستگی نیز با استفاده از روش دلتا به دست آمده‌اند.

همان‌طور که مقادیر ضرایب همبستگی ارائه شده در این جدول نشان می‌دهند، همبستگی منفی بین فاصله تولد اول و تعداد مطلوب فرزندان از دیدگاه مادر وجود دارد. همان‌طور که در جدول‌های ۳ و ۴ ملاحظه می‌شود، افزایش سن مادر در زمان ازدواج ($HR=1/075$ و $P=0/001$)، دیدگاه والدین در مورد افزایش سایر فاصله تولددها ($HR=1/008$ و $P=0/001$)، شاغل نبودن مادر ($HR=1/323$ و $P=0/034$)، همچنین قومیت (مادر) عرب ($HR=1/259$ و $P=0/027$) و لر ($HR=1/283$ و $P=0/002$) نسبت به فارس، منزل مسکونی سایر نسبت به شخصی ($HR=1/279$ و $P=0/009$) و وضعیت تولد ناخواسته ($HR=1/455$ و $P=0/028$)، نسبت به خواست مادر موجب افزایش خطر وقوع تولد اول (کاهش فاصله تولد) و بیماری داشتن مادر (بیماری‌هایی مانند نازایی که مانع بارداری می‌شوند) ($HR=0/332$ و $P=0/001$)، وجود سقط و مرده‌زایی ($HR=0/001$ و $P=0/001$)

۱۲۴ مدل‌سازی داده‌های آمیخته بقا و گسسته با استفاده ازتابع مفصل

جدول ۲: برآورد و انحراف استاندارد پارامترهای همبستگی در مدل توام

تابع مفصل	پارامتر	برآورد	انحراف استاندارد (SE)
فرانک	روی اسپیرمن (ρ_s)	-۰/۸۳۹	۰/۳۳۱
	تاو کندال (τ)	-۰/۰۹۳	۰/۰۵۶
نرمال	همبستگی پیرسون (ρ)	-۰/۱۷۷	۰/۰۵۴
	روی اسپیرمن (ρ_s)	-۰/۱۱۳	۰/۰۳۵
	تاو کندال (τ)	-۰/۱۶۹	۰/۰۵۲

($HR=۰/۴۲۶$ و $HR=۰/۰۰۹$) و تحصیلات دانشگاهی مادر نسبت به بی‌سواد ($P<۰/۰۰۶$) موجب کاهش خطر وقوع تولد اول (افزایش فاصله تولد) می‌شود. همچنین دیدگاه مادر در مورد افزایش سایر فاصله‌های تولد، جنسیت مطلوب فرزند از دیدگاه مادر (دختر) نسبت به پسر و همچنین افرادی که جنسیت برای آن‌ها فرقی نمی‌کند نسبت به پسر) و افزایش سن مادر در زمان ازدواج موجب کاهش تعداد مطلوب فرزندان می‌شوند.

۴ بحث و نتیجه‌گیری

در این مطالعه مدل رگرسیونی برای تحلیل داده‌های آمیخته همبسته در حالتی که یکی از متغیرها گسسته و دیگری پیوسته باشد و امکان وقوع مشاهده سانسور شده برای متغیر پیوسته وجود داشته باشد، ارائه شد. برای این کار از تابع مفصل نرمال و فرانک استفاده و نتایج به دست آمده با مفصل حاصل ضرب از نظر معیار آکائیک مورد مقایسه قرار گرفت. که بر اساس معیار آکائیک به دست آمده مدل برآذش شده با استفاده از مفصل فرانک دارای مقدار آکائیک پایین‌تر و در نتیجه برآذش بهتری نسبت به مفصل حاصل ضرب به داده‌ها بود. از جمله محدودیت‌های استفاده از تابع مفصل برای مدل‌سازی توام داده‌ها این است که با توجه به اینکه امکان تعریف متغیر پنهان برای متغیرهای اسمی وجود ندارد، روش معروفی شده در این مقاله را نمی‌توان برای زمانی که متغیر گسسته از نوع اسمی باشد مورد استفاده قرار داد. همچنین بر اساس مقاله دلشون و وو (۲۰۱۱) واریانس توزیع پنهان به دلیل عدم شناسایی پذیر بودن مدل باید یک در نظر گرفته شود. از مزایای مدل ارائه شده امکان گسترش آن به حالت‌های مختلف بسیاری می‌باشد که می‌توان به عنوان نمونه به گسترش مدل به مطالعات طولی که در آن متغیر گسسته بیش از یکبار اندازه‌گیری شده است و یا داده‌های بقای دنباله‌ای که برای هر فرد دنباله‌ای از زمان‌های بقا مشاهده می‌گردد، اشاره نمود.

جدول ۳: برآورد و انحراف استاندارد پارامترهای مدل فاصله تولد و نسبت خطر در مدل توان

متغیرها	برآورد	نسبت خطر (HR)	انحراف استاندارد (SE)	
مقدار ثابت	۴/۵۲۳	-	۰/۱۸۴	
سن مادر در زمان ازدواج	-۰/۰۴۱**	۱/۰۷۵	۰/۰۰۵	
بي سواد	-۰/۰۴۳	۱/۰۷۹	۰/۰۲۵	سطح مرجع
ابتدائي	-۰/۱۳۷	۱/۱۷۴	۰/۰۷۴	
راهنمايی	-۰/۰۵۰	۱/۰۹۳	۰/۰۶۶	
تحصيلات مادر	-۰/۰۹۷*	۰/۰۷۶	۰/۰۷۶	دانشگاهي
فارس	-۰/۱۳۰*	۱/۱۵۹	۰/۰۵۹	سطح مرجع
عرب	-۰/۱۸۳*	۱/۱۳۸	۰/۰۶۱	
لر	-۰/۱۷۰	۱/۱۳۵	۰/۰۹۹	
سایر	-۰/۰۴۲**	۰/۰۴۲۶	۰/۰۵۸	وجود سقط و مردهزابي
خيار	-۰/۰۶۲۳**	۰/۰۳۳۲	۰/۰۸۲	سطح مرجع
بله	-۰/۰۴۸۲**	۰/۰۴۲۶	۰/۰۵۸	
شخيصي	-۰/۰۰۸۸	۰/۰۸۵۶	۰/۰۵۰	سطح مرجع
اجاره‌اي	-۰/۰۱۳۹*	۱/۱۲۷۹	۰/۰۵۴	
سایر	-۰/۰۰۴۶	۱/۰۸۵	۰/۱۵۴	
نوع منزل مسکونی	-۰/۰۰۸۸	-۰/۰۰۴۷	۰/۰۱۰	دیدگاه مادر در مورد ساير فاصله تولدها
شاغل	-۰/۰۱۵۸*	۱/۱۳۲۳	۰/۰۷۴	وضعیت اشتغال مادر
خانه‌دار	-۰/۰۱۸۱	۱/۱۳۷۸	۰/۱۳۶	سطح مرجع
خواست پدر	-۰/۰۲۱۲*	۱/۱۴۵۵	۰/۰۹۹	وضعیت تولد فرزند
ناخواسته	-۰/۰۱۱	۱/۰۲۱	۰/۰۸۲	خواست هر دو
	۱/۰۷۷۰	-	۰/۰۴۴	δ

جدول ۴: برآورد و انحراف استاندارد پارامترهای مدل تعداد مطلوب فرزندان در

		مدل توأم	
		متغیرها	
انحراف		برآورد	
(SE)			
استاندارد (SE)		۳/۵۰۹	مقدار ثابت
۰/۲۱۴			
۰/۰۰۹	-۰/۰۲۳**		سن مادر در زمان ازدواج
۰/۰۹۲	-۰/۲۴۵*	پسر دختر	جنسیت مطلوب فرزند (مادر)
۰/۰۸۳	-۰/۲۲۷*	فرقی نمی‌کند	دیدگاه مادر در مورد سایر فاصله تولددها
۰/۰۱۷	-۰/۰۴۰*		

تقدیر و تشکر

نویسنده‌گان مقاله از نظرات ارزشمند داوران محترم در بهبود کیفیت مقاله و همچنین از سردبیر و هیأت تحریریه به خاطر مطالعه دقیق و ویرایش ادبی مقاله تقدیر و تشکر به عمل می‌آورند.
به علاوه از حمایت معنوی دانشگاه شهید چمران اهواز تشکر می‌شود.

مراجع

سدھی، م.، محراجی، ی.، کاظم‌نژاد، ا.، جوهری مجد، و.، حدائقی، ف. (۱۳۸۹)، طراحی شبکه عصبی مصنوعی برای مدل‌بندی پاسخ‌های دومتغیره آمیخته و کاربرد آن در داده‌های پزشکی، مجله تخصصی اپیدمیولوژی ایران، ۶، ۳۹-۲۸.

قره آقاجی اصل، ر.، مشکانی، م.، فقیه‌زاده، س.، کاظم‌نژاد، ا.، بابایی، غ.، زایری، ف. (۱۳۸۶)، تحلیل بیزی مدل دومتغیره تربیتی بر پایه متغیر پنهان، مجله علوم آماری، ۱، ۱۵۵-۳۹.

Cameron, A. C., Li, T., Trivedi, P. K. and Zimmer, D. M. (2004), Modelling the Differences in Counted Outcomes Using Bivariate Copula Models with Application to Mismeasured Counts, *The Econometrics Journal*, 7, 566-587.

۱۲۷ مینا گدازی و همکاران

- Catalano, P. and Ryan, L. (1992), Bivariate Latent Variable Models for Clustered and Continuous Outcomes, *Journal of the American Statistical Association*, **87**, 651-658.
- Craiu, R. V. and Sabeti, A. (2012), In Mixed Company: Bayesian Inference for Bivariate Conditional Copula Models with Discrete and Continuous Outcomes, *Journal of Multivariate Analysis*, **110**, 106-120.
- De Leon, A. R. and Carriere Chough, K. (2010), Mixed-Outcomes Data, In Chow S-C (Ed.), *Encyclopedia of Biopharmaceutical Statistics*, Taylor and Francis, London.
- De Leon, A. R. and Wu, B. (2011), Copula- Based Regression Models for a Bivariate Mixed Discrete Continuous Outcome, *Statistics in Medicine*, **30**, 175-185.
- Faes, C., Aerts, M., Molenberghs, G., Geys, H., Tteuns, G. and Bijnens, L. (2008), A High dimensional Joint Model for Longitudinal Outcomes of Different Nature, *Statistics in Medicine*, **27**, 4408-4427.
- Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall, London.
- Kramer, N., Brechmann, E. C., Silvestrini, D. and Czado, C. (2013), Total Loss Estimation Using Copula- Based Regression Models, *Insurance: Mathematics and Economics*, **53**, 829-839.
- Molenberghs, G., Geys, H. and Buyse, M. (2001), Evaluation of surrogate End-points in Mixed Discrete and Continuous Outcomes, *Statistics in Medicine*, **20**, 3023-3038.
- Munkin, M. and Trivedi, P. (1999), Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, with Applications, *Econometrics Journal*, **2**, 29-48.

۱۲۸ مدل‌سازی داده‌های آمیخته بقا و گسسته با استفاده ازتابع مفصل

Nelsen, R. B. (2006), *An Introduction to Copulas*, 2nd Ed., Springer, NewYork.

Regan, M. and Catalano, P. (2002), Combined Continuous and Discrete Outcomes, In Aerts, Geys, M., Molenberghs, H. and Ryan, L. (Eds.), *Topics in Modelling of Clustered data*, Chapman and Hall/CRC, London.

Song, P. X. K., Li, M. and Yuan, Y. (2009), Joint Regression Analysis of Correlated Data Using Gaussian Copula, *Biometrics*, **65**, 60-68.

Trivedi, P. K. and Zimmer, D. M. (2006), Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand, *Journal of Business and Economic Statistics*, **24**, 63-76.

Yang, Y., Kang, J., Mao, K. and Zhang, J. (2007), Regression Models for Mixed Poisson and Continuous Longitudinal Data, *Statistics in Medicine*, **26**, 3782-3800.