مدل‌های رگرسیون براساس تابع مفصل

هادی جباری نوقابی، عاطفه جاوید نظام دوست

گروه آمار، دانشگاه فردوسی مشهد، مشهد، ایران

چکیده: در تجزیه و تحلیل مدل‌های وابستگی از جمله مدل‌های رگرسیونی، پیدا کردن تابعی که وابستگی بین متغیر پاسخ و متغیرهای توضیحی را نشان دهد، اهمیت دارد. وابستگی در این تابع به رفتار کناری متغیرها و رفتار توأم آن‌ها مربوط است. چون در تابع مفصل وابستگی به کناری‌ها بستگی ندارد، در این مقاله روشی برای جدا ساختن این دو نوع وابستگی از یکدیگر از طریق تابع مفصل ارائه می‌شود.

واژه‌های کلیدی: تابع مفصل، رگرسیون مفصل، مفصل FGM.

۱ مقدمه

تجزیه و تحلیل رگرسیون یکی از روش‌های آماری با کاربردهای وسیع است که از نظر تاریخی به سال ۱۸۷۵ و فعالیت‌های پژوهشی فرانسیس گالتون (۱۹۹۸) باز می‌گردد. در سال‌های اخیر برای تسهیل تحلیل رگرسیون از تابع مفصل که دارای ویژگی‌های بارزی است، استفاده شده است. برخی از مؤلفین در که در زمینه رگرسیون مفصل تحقیقاتی انجام داده‌اند عبارتند از: چروبینی (۲۰۰۴)، مک نیل و همکاران (۲۰۰۵)، کولیف و همکاران (۲۰۰۸)، و همچنین کتاب نلسن (۲۰۰۶) یک مرجع بسیار مفید برای مفصل و وابستگی می‌باشد.

ایده اصلی این روش، تشخیص رابطه بین چند متغیر است. ساختار مدل رگرسیونی به این صورت است که یک متغیر وابسته یا پاسخ (Y) با چند متغیر مستقل $(X_n, ..., X_1)$ مرتبط در نظر گرفته می‌شود. مساله اصلی این است که تعیین کنیم توزیع متغیر پاسخ چگونه با متغیرهای مستقل در ارتباط است. در این مقاله با تکیه بر مفهوم تابع توزیع شرطی، تابع رگرسیون مفصل را تعریف می‌کنیم. ساختار مقاله به شرح زیر است:

در بخش دوم پس از بیان تعاریف و مفاهیم اولیه، ویژگی‌های تابع رگرسیون مفصل، تابع رگرسیون خطی و همچنین یک مثال را بیان می کنیم. در بخش سوم روش برآورد با استفاده از مفصل را در تجزیه وتحلیل رگرسیون با ذکر یک مثال بیان می‌کنیم.

۱.۱ تعاریف و مفاهیم اولیه

فرض می‌کنیم X و Y متغیرهای پیوسته، به ترتیب با توابع توزیع کناری G و H و تابع توزیع توام F و مفصل C باشند. آن‌گاه $U = G(X)$ و $V = H(Y)$ دارای توزیع یکنواخت روی $(۰, ۱)$ است و توزیع توام آن‌ها مفصل C است.

تعریف ۱.۱. *اگر بردار تصادفی (U, V) دارای توزیع‌های کناری یکنواخت روی $[۰, ۱]$ و مفصل C باشد، آن‌گاه $E_c[V \mid U = u]$ تابع رگرسیون مفصل V روی U نامیده می‌شود و آن را با $r_c(u)$ نشان می‌دهیم.*

چند نکته چروبینی (۲۰۰۴) را که در ادامه استفاده خواهد شد، در اینجا ذکر می‌کنیم:

۱- تابع توزیع شرطی V به شرط $U = u$ یعنی $C_u(v)$ به صورت زیر است:

$$\mathbb{P}(V \leq v \mid U = u) = \frac{\partial C(u, v)}{\partial u} = C_u(v) \qquad (۱)$$

۲- تابع توزیع شرطی Y به شرط $X = x$ برابر است با :

$$\begin{aligned} \mathbb{P}(Y \leq y \mid X = x) &= \mathbb{P}(V \leq H(y) \mid U = G(x)) \\ &= \frac{\partial C(u, v)}{\partial u} \mid_{u=G(x), v=H(y)} \\ &= C_u(v) \mid_{u=G(x), v=H(y)} \qquad (۲) \end{aligned}$$

۳- تابع رگرسیون مفصل V روی U به صورت زیر است:

$$E[V_u] = r_C(u) = ۱ - \int_۰^۱ C_u(v)dv \qquad (۳)$$

۴- *تابع رگرسیون Y روی X به صورت زیر است:*

$$E[Y \mid X = x] = \hat{y} = H^{[-1]}\left(1 - \int_0^1 C_u(v)dv\right)$$
$$= H^{[-1]}\left(r_C(G(x))\right) \qquad (۴)$$

که در آن $H^{[-1]}$ معکوس تابع توزیع متغیر تصادفی Y است. شایان ذکر است که هر تبدیل اکیداً صعودی و یکنوا از X و Y توزیع حاشیه‌ای را تغیر خواهد داد، بدون اینکه بر رفتار توزیع توام X و Y تاثیری داشته باشد.

گزاره ۱.۱. *فرض کنید C یک مفصل و $r_C(u)$ مطابق رابطه (۳) باشد، بنابراین :*

الف) اگر $C^{\cdot}(u,v) = uv$ آن‌گاه:

$$r_{\cdot}(u) = \frac{1}{2}$$

ب) اگر $C^{+}(u,v) = M(u,v) = min\{u,v\}$ آن‌گاه:

$$r_{+}(u) = r_M(u) = u$$

ج) اگر $C^{-}(u,v) = W(u,v) = max\{u+v-1, 0\}$ آن‌گاه:

$$r_{-}(u) = r_W(u) = 1 - u$$

برهان. الف) از اینکه $C_{\cdot}(v) = \frac{\partial C^{\cdot}(u,v)}{\partial u} = v$ پس به ازای هر $u \in (0,1)$ داریم:

$$r_{\cdot}(u) = 1 - \int_0^1 C_{\cdot}(v)dv = 1 - \int_0^1 vdv = 1 - \frac{1}{2} = \frac{1}{2}$$

ب) چون $C^{+}(u,v) = \begin{cases} v & ; v \le u \\ u & ; v > u \end{cases}$ ، بنابراین: $C_U^{+}(v) = \frac{\partial C^{+}(u,v)}{\partial u} = \begin{cases} 0 & ; v \le u \\ 1 & ; v > u \end{cases}$ و در نتیجه:

$$r_{+}(u) = r_M(u) = 1 - \int_0^1 C^{+}(v)dv = 1 - \int_u^1 1dv = 1 - 1 + u = u$$

ج) با توجه به فرض، $C^-(u,v) = \begin{cases} ۰ & ; v \leq ۱ - u \\ u + v - ۱ & ; v > ۱ - u \end{cases}$ پس:

$$C_U^- = \frac{\partial C^-(u,v)}{\partial u} = \begin{cases} ۰ & ; v \leq ۱ - u \\ ۱ & ; v > ۱ - u \end{cases}$$

و در نتیجه:

$$r_-(u) = ۱ - \int_۰^۱ C^-(v)dv = ۱ - \int_{۱-u}^۱ ۱ dv = ۱ - (۱ - (۱ - u)) = ۱ - u$$

□

۲ تابع رگرسیون خطی

در این بخش تابع رگرسیون خطی را با ارائه یک مثال مورد بررسی قرار می‌دهیم. همچنین ضریب همبستگی پیرسون را محاسبه می‌کنیم. فرض کنید C_t کلاس مفصل‌هایی باشد که تابع رگرسیون خطی دارند، یعنی:

$$C_\iota = \{C : I^۲ \to I, r_c(u) = \alpha + \beta u\}$$

قضیه‌ی ۱.۲. *مفصل C دارای تابع رگرسیون خطی است، $(C \in C_\iota)$ اگر و فقط اگر $r_C(u)$ به صورت $r_C(u) = \alpha + (۱ - ۲\alpha)u$ یا $r_C(u) = \frac{۱-\beta}{۲} + \beta u$ باشد.*

برهان. از اینکه:

$$r_C(u) = ۱ - \int_۰^۱ \frac{\partial C(u,v)}{\partial u} dv = \alpha + \beta u$$

نتیجه می‌شود:

$$\frac{\partial}{\partial u} \int_۰^۱ C(u,v)dv = ۱ - \alpha - \beta u$$

و با انتگرال گیری از طرفین معادله به راحتی این قضیه را می‌توان اثبات کرد. شبیه قضیه ۲ ۰ ۱ به راحتی می‌توان قضیه بعدی را اثبات کرد. □

قضیه‌ی ۲.۲. *اگر مفصل C دارای تابع رگرسیون خطی باشد، آن‌گاه ضریب همبستگی پیرسون $\rho_C = ۱ - ۲\alpha$ است و تابع رگرسیون مفصل برابر با $r_C(u) = \frac{۱-\rho_C}{۲} + \rho_C u$ می‌باشد.*

مثال ۱.۲. *مفصل C از خانواده مفصل‌هایی FGM را به صورت زیر در نظر می‌گیریم:*

$$C(u,v) = uv\,[۱ + \theta(۱-u)(۱-v)]$$

که در آن $\theta \in [-۱, ۱]$. چون $\frac{\partial C(u,v)}{\partial u} = v\theta - v^۲\theta - ۲uv\theta + ۲uv^۲\theta + v$ پس:

$$\begin{aligned} r_C(u) &= ۱ - \int_۰^۱ C_u(v)dv = ۱ - \int_۰^۱ \frac{\partial C(u,v)}{\partial u}dv \\ &= \int_۰^۱ (۱ + \theta - ۲u\theta)vdv + \int_۰^۱ v^۲(\theta + ۲u\theta)dv \\ &= ۱ - \frac{۳ + \theta - ۲u\theta}{۶} = \frac{۳-\theta}{۶} + \frac{\theta}{۳}u \end{aligned}$$

با مقایسه با فرم کلی تابع رگرسیون خطی $\alpha = \frac{۳-\theta}{۶}$ ، تابع رگرسیون مفصل اصلاح شده به صورت زیر است:

$$r_C(u) = \alpha + (۱ - ۲\alpha)u$$

و از این‌که $\rho_C = ۱ - ۲\alpha$ در این مثال داریم:

$$\rho_C = ۱ - ۲\frac{۳-\theta}{۶} = \frac{\theta}{۳}$$

۳ برآورد با استفاده از مفصل در تجزیه وتحلیل رگرسیون

تابع مفصل می‌تواند شامل یک بردار از پارامترها باشد، که درجه آن از وابستگی بین توابع توزیع حاشیه‌ای تک متغیره گرفته می‌شود. در زمینه رگرسیون هر توزیع حاشیه‌ای را می‌توان توسط یک بردار از متغیرهای کمکی مشخص کرد. برآورد حاصله توسط انتخاب (کولیف و پایوا، ۲۰۰۹) مفصل مناسب $C(۰, ۰; \theta)$

که برداری از پارامترهای θ و توزیع‌های حاشیه‌ای $F_1(y_1 \mid x_1, \beta_1)$ و $F_2(y_2 \mid x_2, \beta_2)$ است. که در آن‌ها x_1 و x_2 متغیرهای کمکی و β_1 و β_2 پارامترهای مجهول هستند. سپس طبق روش‌های درستنمایی ماکسیمم و استاندارد کردن، تابع توزیع توام را به‌صورت زیر داریم:

$$F(y_1, y_2 \mid x_1, x_2, \beta_1, \beta_2) = C(F_1(y_1 \mid x_1, \beta_1), F_2(y_2 \mid x_2, \beta_2); \theta).$$

برای یک بردار از پارامترهای $\Theta = (\beta_1, \beta_2, \theta)$، تابع چگالی آن عبارت است از:

$$f(y_{1i}, y_{2i}, \Theta) = c(F_1(y_{1i}, \beta_1), F_2(y_{2i}, \beta_2), \theta) f_1(y_{1i}, \beta_1) f_2(y_{2i}, \beta_2). \qquad (5)$$

مثال ۱.۳. *فرض کنید توابع حاشیه‌ای y_{1i} و y_{2i} دارای توزیع نمایی به‌صورت زیر باشند:*
$Y_{1i} \sim \exp(\beta_1)$ و $Y_{2i} \sim \exp(\beta_2)$ *و فرض کنید مفصل C از خانواده FGM به نحوه زیر باشد:*

$$C(u, v; \theta) = uv + \theta u(1-u)(1-v) \quad \theta \in [-1, 1]$$

و تابع چگالی مفصل

$$c(u, v; \theta) = 1 + \theta(1-2u)(1-2v) \qquad (6)$$

باشد.طبق فرمول (۵) داریم:

$$f(y_{1i}, y_{2i}, \theta, \beta_1, \beta_2) = \beta_1 e^{-\beta_1 y_{1i}} \cdot \beta_2 e^{-\beta_2 y_{2i}} \cdot c(F_1(y_{1i}, \beta_1), F_2(y_{2i}, \beta_2), \theta)$$

با استفاده از رابطه (۶) نتیجه می‌گیریم:

$$f(y_{1i}, y_{2i}, \theta, \beta_1, \beta_2) =$$
$$\beta_1 \beta_2 e^{-\beta_1 y_{1i} - \beta_2 y_{2i}} \left[1 + \theta\left(1 - 2\left(1 - e^{-\beta_1 y_{1i}}\right)\right)\left(1 - 2\left(1 - e^{-\beta_2 y_{2i}}\right)\right)\right].$$

حال می‌خواهیم پارامترهای β_1، β_2 و θ را برآورد کنیم. با توجه به این‌که $\tau_{X,Y} = \frac{2\theta}{9}$ است، می‌توان θ را به‌صورت زیر برآورد کرد:

$$\hat{\theta} = \frac{9}{2}\hat{\tau}$$

بنابراین داریم:

$$f(y_{۱i}, y_{۲i}, \theta, \beta_۱, \beta_۲) =$$
$$\beta_۱\beta_۲ e^{-\beta_۱ y_{۱i} - \beta_۲ y_{۲i}} \left[۱ + \hat{\theta} \left(۱ - ۲ \left(۱ - e^{-\beta_۱ y_{۱i}} \right) \right) \left(۱ - ۲ \left(۱ - e^{-\beta_۲ y_{۲i}} \right) \right) \right].$$

حال با استفاده از روش درستنمایی ماکسیمم می‌خواهیم پارامترهای $\beta_۱$ و$\beta_۲$ را برآورد کنیم. بنابراین داریم:

$$L(\beta_۱, \beta_۲) = \Sigma_{i=۱}^{n} ln(f \left\{ y_{۱i}, y_{۲i}, \hat{\theta}, \beta_۱, \beta_۲ \right\})$$
$$= \Sigma_{i=۱}^{n} \left\{ ln\beta_۱ + ln\beta_۲ - \beta_۱ y_{۱i} - \beta_۲ y_{۲i} + ln \left[۱ + \hat{\theta} \left(۱ - ۲(۱ - e^{-\beta_۱ y_{۱i}}) \right) \left(۱ - ۲(۱ - e^{-\beta_۲ y_{۲i}}) \right) \right] \right\}.$$

حال نسبت به $\beta_۱$ مشتق می‌گیریم:

$$\frac{\partial L(\beta_۱, \beta_۲)}{\partial \beta_۱} = \Sigma_{i=۱}^{n} \frac{\partial ln(f)}{\partial \beta_۱}$$

که در آن:

$$\Sigma_{i=۱}^{n} \frac{\partial ln(f)}{\partial \beta_۱} = \Sigma_{i=۱}^{n} [\frac{\beta_۲ e^{-\beta_۱ y_۱ - \beta_۲ y_۲} - \beta_۱\beta_۲ e^{-\beta_۱ y_۱ - \beta_۲ y_۲} y_۱ \left(۱ + \hat{\theta} \left(-۱ + ۲e^{-\beta_۱ y_۱} \right) \left(-۱ + ۲e^{-\beta_۲ y_۲} \right) \right)}{\beta_۱\beta_۲ e^{-\beta_۱ y_۱ - \beta_۲ y_۲} \left(۱ + \hat{\theta} \left(-۱ + ۲e^{-\beta_۱ y_۱} \right) \left(-۱ + ۲e^{-\beta_۲ y_۲} \right) \right)}$$
$$- \frac{۲\beta_۱\beta_۲ e^{-\beta_۱ y_۱ - \beta_۲ y_۲} t y_۱ e^{-\beta_۱ y_۱} \left(-۱ + ۲e^{-\beta_۲ y_۲} \right)}{\beta_۱\beta_۲ e^{-\beta_۱ y_۱ - \beta_۲ y_۲} \left(۱ + \hat{\theta} \left(-۱ + ۲e^{-\beta_۱ y_۱} \right) \left(-۱ + ۲e^{-\beta_۲ y_۲} \right) \right)}].$$

با توجه به این‌که توزیع متقارن است مشتق لگاریتم تابع درستنمایی نسبت به $\beta_۲$ نیز مشابه $\beta_۱$ است. بنابراین برای برآورد پارامترهای $\beta_۱$ و $\beta_۲$ باید دستگاه معادلات زیر حل شود:

$$\begin{cases} \Sigma_{i=۱}^{n} \frac{\partial ln(f)}{\partial \beta_۱} = ۰ \\ \Sigma_{i=۱}^{n} \frac{\partial ln(f)}{\partial \beta_۲} = ۰ \end{cases} \quad (۷)$$

حال با ارایه یک مثال عددی، نتایج مثال ۳ . ۱ را بکار می‌گیریم. در این مثال داده‌ها را از صفحه اقتصادی روزنامه خراسان جمع آوری کرده‌ایم به این‌صورت که در هر روز از شش ماهه دوم سال ۱۳۹۰ نرخ دلار برحسب ریال استخراج شده است و بیشترین مقدار و کمترین مقدار نرخ دلار در هر هفته نیز مد نظر است. داده‌ها شامل ۲۰۰ زوج مشاهده از بیشترین نرخ دلار $(y_۱)$ با میانگین ۱/ ۰۲۳ و واریانس

۰/۹۸۷ و کمترین نرخ دلار (y_2) با میانگین ۱/۱۲۱ و واریانس ۱/۲۶۲ است.
ابتدا با استفاده از آزمون نیکویی برازش کلموگروف-اسمیرنوف برای متغیر y_1 با $p-value =$ ۰/۹۷۵۳ و y_2 با $p-value =$ ۰/۵۳۵۶ ملاحظه می‌شود که هر دو متغیر دارای توزیع نمایی با پارامتر یک هستند. همچنین مقادیر ضرایب همبستگی پیرسون (۰/۰۳۹۵۵۴۴۴)، کندال (۰/۰۱۱۴۵۷۲۹) و اسپیرمن (۰/۰۱۷۸۳۵۴۵) تاییدی بر وابستگی ضعیف در بین این دو مجموعه داده است، بنابراین مفصل FGM برای این داده‌ها مناسب است. همچنین با انجام آزمون نیکویی برازش $(gofCopula)$ در پکیج $(copula)$ در R، با $p-value =$ ۰/۹۸۶۵ مشخص می‌شود که مفصل FGM با پارامتر θ برای داده‌ها مناسب است.

$$\hat{\theta} = ۰/۰۵۵۸۹۹۰۲$$

بنابراین به راحتی با حل دستگاه (۷) می‌توان پارامترهای β_1 و β_2 را به‌صورت زیر برآورد کرد:

$$\hat{\beta_1} = ۰/۹۷۶۴۴۴۹$$

$$\hat{\beta_2} = ۰/۸۹۱۸۵۳۶.$$

بحث و نتیجه‌گیری

در این مقاله تلاش کردیم تا نگاه عمیق‌تری به معادله رگرسیونی نسبت به حالت تابعی ساده آن داشته باشیم و با مثالی نشان دادیم که فرم تابع خط رگرسیون، به رفتار توأم متغیرها (که مفصل تعیین کننده آن است) و رفتار حاشیه‌ای (که با تابع توزیع حاشیه‌ای شکل می‌گیرد) وابسته است. هدف اصلی در تجزیه وتحلیل رگرسیون، نمایش تابع رگرسیون به صورت خطی است، که در این مقاله نیز بررسی شد. همچنین نحوه برآورد با استفاده از مفصل در تجزیه و تحلیل رگرسیون را با ذکر مثالی نشان دادیم.

مراجع

Cherubini, U. Luciano, E. Vecchiato, W. (2004),*Copula Methods in Finance*. New York: John Wily and Sons, p. 182

Frees, E., and E. Valdez, “Understanding Relationships Using Copulas,” North American Actuarial , *Journal 2, 1112, pp. 1–25.*

Kolev, N., Anjos, U. and Mendes, B. (2006).*Copulas: a review and recent developments*. Statistical Models 22, 617-660.

Kolev, N., and Paiva,D,(2009),*Copula-based regression models .Journal of Statistical Planning and Inference* 139,3847-3856

McNeil, A., Frey, R. and Embrechts, P. (2005).*Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

Nelsen, R. (2006).*An Introduction to Copulas*, 2nd Edition, Springer: New York.