وابستگی رگرسیونی و توابع مفصل

محمد حسین علامت‌ساز[۱]　　فائزه شفیعی‌فر[۱]

چکیده

مفاهیم وابستگی از ابزارهای مهم درک روابط بین رویدادهای تصادفی چندمتغیره هستند. در مقاله حاضر، به بررسی مفهوم وابستگی رگرسیونی خواهیم پرداخت. به ویژه، نشان خواهیم داد که چگونه می‌توان از توابع مفصل برای بررسی این نوع وابستگی سود جست. همچنین، به نوعی آزمون فرض در این مورد اشاره خواهیم کرد و با مثال‌های واقعی و عددی موضوع را تشریح خواهیم کرد.

واژه‌های کلیدی : وابستگی رگرسیونی، توابع مفصل، استقلال.

۱. مقدمه

مفاهیم اولیه وابستگی مثبت و منفی، توسط لهمن (۱۹٦٦) بیان شد. مفاهیمی که توسط وی ارایه شد، همگی در حالت دو متغیره طرح شده بودند. جو[٤] و در سال‌های اخیر بزرگ‌نیا و امینی[۱۰] مفاهیم وابستگی را در حالت چندمتغیره مورد بررسی قرار دادند.

توابع مفصل توسط اسکلار در سال ۱۹۵۹ معرفی شدند. مفصل تابعی است که تابع توزیع چندمتغیره را به تابع توزیع‌های حاشیه‌ای تک‌متغیره آن پیوند می‌دهد. امروزه، توابع مفصل به عنوان راهی برای مطالعه اندازه‌های آزاد از مقیاس و وابستگی پیشنهاد می‌شوند، زیرا ویژگی‌های ناپارامتری بودن، ناوردایی و وابستگی توابع توزیع متغیرهای تصادفی را به خوبی بیان می‌کنند.

در بخش ۲، وابستگی رگرسیونی مثبت و منفی را معرفی و تشریح می‌کنیم. در بخش ۳، توابع مفصل مورد بررسی قرار گرفته و برخی خواص مهم آنها بیان خواهد شد. کاربرد توابع مفصل را در تعیین وابستگی رگرسیونی در بخش ٤ مورد توجه قرار خواهیم داد و بالاخره در بخش ۵ به آزمون فرضیه لی‌من در مورد وابستگی رگرسیونی در مقابل فرضیه استقلال اشاره خواهیم کرد.

[۱] گروه آمار، دانشگاه اصفهان

۲. وابستگی رگرسیونی

۲-۱- وابستگی رگرسیونی دو متغیره

تعریف ۲-۱: زوج تصادفی (X,Y) را وابسته رگرسیونی منفی گوییم، هرگاه $P[Y \le y | X = x]$ تابعی غیر نزولی بر حسب x باشد. در این صورت، X و Y را وابسته رگرسیونی منفی[۲] (NRD) گوییم و از نماد $NRD(Y \mid X)$ برای نمایش آن استفاده می‌کنیم. خانواده تمام توزیع‌های (X,Y) را که در آنها (X,Y) وابسته رگرسیونی منفی هستند، با g_2 نشان می‌دهیم.

[۲] Negative Regression Dependence

تعریف ۲-۲: X و Y را وابسته رگرسیونی مثبت[۳] (PRD) گوییم، هرگاه تابع $P[Y \le y \mid X = x]$ به ازای هر $x, y \in \Re$ بر حسب x غیر صعودی باشد؛ و از نماد $PRD(Y \mid X)$ برای نمایش آن استفاده می‌کنیم. توزیع‌های (X,Y) را که در آنها (X,Y) وابسته رگرسیونی مثبت هستند، با $F_۲$ نشان می‌دهیم.

در اینجا به بیان چند خاصیت مقدماتی وابستگی رگرسیونی که توسط امینی [۱۰] اثبات شده می‌پردازیم.

قضیه ۲-۱:

الف) زوج (X,Y)، NRD است، اگر و تنها اگر $(-X,Y)$، PRD باشد.

ب) زوج (X,Y)، NRD است، اگر و تنها اگر $(X,-Y)$، PRD باشد.

مثال ۲-۱: فرض کنید $Y = \alpha + \beta X + U$ و X و U مستقل باشند، آنگاه،

الف: $(X,Y) \in g_۲ \Leftrightarrow \beta \le ۰$

ب: $(X,Y) \in f_۲ \Leftrightarrow \beta \ge ۰$

اثبات:

$$\forall x, y \in \Re :$$

$$P[Y \le y | X = x] = P[\alpha + \beta X + U \le y | X = x]$$
$$= P[U \le y - \alpha - \beta X | X = x]$$

(بنابر استقلال U و X)

$$= P[U \le y - \alpha - \beta x]$$
$$= F_u(y - \alpha - \beta x)$$

مشاهده می‌کنیم که $F_u(y - \alpha - \beta x)$ تابعی غیر نزولی از x است، اگر و تنها اگر $۰ \le \beta$. در نتیجه (الف) ثابت می‌شود و به همین ترتیب تابعی غیر صعودی از x است، اگر و تنها اگر $\beta \ge ۰$. در نتیجه (ب) هم ثابت می‌شود.

مثال ۲-۲: هرگاه $Y = \alpha + \beta X + U$، $\beta \le ۰$ و $(X,U) \in g_۲$؛ آنگاه $(X,Y) \in g_۲$.

اثبات:

چون $(X,U) \in g_۲$، لذا $P[U \le u \mid X = x]$ تابعی غیر نزولی بر حسب x است. برای اثبات کافی، نشان دهیم که $P[Y \le y | X = x]$ نیز تابعی غیر نزولی از x است، اما

$$\forall x_۱ < x_۲ :$$

$$P[Y \le y \mid X = x_۱] = P[U \le y - \alpha - \beta X \le y \mid X = x_۱]$$
$$= P[U \le y - \alpha - \beta x_۱ \mid X = x_۱]$$
$$\le P[U \le y - \alpha - \beta x_۲ \mid X = x_۱]$$

با توجه به اینکه $(X,Y) \in g_۲$ و $\beta \le ۰$

$$\le P[U \le y - \alpha - \beta x_۲ \mid X = x_۲]$$
$$= P[U \le y - \alpha - \beta X | X = x_۲]$$
$$= P[Y \le y | X = x_۲]$$

در نتیجه $P[Y \le y \mid X = x]$ تابعی غیر نزولی از x است [۱۰].

۲-۲- تعمیم چندمتغیره

سه مفهوم وابستگی که می‌توانند به عنوان تعمیم چندمتغیره وابستگی رگرسیونی فرض شوند، عبارتند از

۱- وابستگی مثبت از طریق ترتیب تصادفی[٤] (PDS)

۲- صعودی شرطی در دنباله[٥] (CIS)

۳- نزولی شرطی در دنباله[٦] (CDS)

تعریف ۲-۳: بردار تصادفی $(X_۱,\dots,X_n)$ را (PDS) گوییم، اگر $\{X_i : i \ne j\}$ مشروط بر $X_j = x$ به ازای هر j با افزایش یافتن x به طور تصادفی افزایش یابد. به عبارت ساده‌تر، یعنی اگر بدانیم یکی از Xها در بردار $(X_۱,\dots,X_n)$ افزایش یافته است، آنگاه سایر Xها هم افزایش می‌یابند.

تعریف ۲-٤: بردار تصادفی $(X_۱,\dots,X_n)$ را (CIS) گوییم، هرگاه برای هر $x_i \in \Re$:

$$P(X_i > x_i \mid X_j = x_j, \quad j = ۱,\dots,i-۱) \quad i = ۲,\dots,n \quad (۲)$$

تابعی صعودی بر حسب $X_۱,\dots,X_{i-۱}$ باشد. به عبارت ساده‌تر، یعنی اگر بدانیم $X_۱,\dots,X_{i-۱}$ افزایش یافته‌اند، احتمال اینکه x_i افزایش

---

[۳] Positive Regression Dependence

[٤] Positive Dependence through the Stochastic ordering

[٥] Conditional Increasing in sequence

[٦] Conditinal Decrasing in sequence

یابد بیشتر می‌شود. توجه داشته باشید در حالت دو متغیره مفهوم صعودی شرطی در دنباله معادل با وابستگی رگرسیونی مثبت می‌باشد.

تعریف ۲-۵: بردار تصادفی $(X_1,...,X_n)$ را (CDS) گوییم، هرگاه برای هر $x_i \in \Re$:

$$P(X_i > x_i \mid X_j = x_j, \quad j = 1,...,i-1) \quad i = 2,...,n \qquad (3)$$

تابعی نزولی بر حسب $X_1,...,X_{i-1}$ باشد در حالت $n = 2$ مفهوم نزولی شرطی در دنباله معادل با وابستگی رگرسیونی منفی می‌باشد [۴].

۳. تابع مفصل[7]

مفصل‌ها، تابع‌هایی هستند که تابع‌های توزیع یک متغیره را به فرم تابع‌های توزیع چندمتغیره پیوند می‌دهند. کلمه مفصل برای اولین بار توسط اسکلار[8] به کار رفت و از کلمه لاتین «به هم پیوستن[9]» گرفته شده است. جو[۴] و نلسن [۹، ۸، ۷ و ۶] و امبرچتز [۳] گامهای مهمی را برای اشاعه کاربرد توابع مفصل در زمینه مفاهیم وابستگی دو متغیره و چندمتغیره برداشتند.

۴-۱- مفصل‌های دو متغیره

یک مفصل دو بعدی تابعی مانند $C : I^2 \to I = [0,1]$ با ویژگی‌های زیر می‌باشد؛

(الف)

$$C(0,t) = C(t,0) = 0 \ , \ C(1,t) = C(t,1) = t \quad \forall t \in [0,1]$$

(ب) $u_1 \le u_2 \ , \ v_1 \le v_2 \quad \forall u_1, u_2, v_1, v_2 \in I$

$$C(u_1,v_1) - C(u_1,v_2) - C(u_2,v_1) + C(u_2,v_2) \ge 0 \qquad (4)$$

به سادگی ثابت می‌شود که تابع مفصل C نسبت به هر یک از متغیرهای خود، غیر نزولی و پیوسته است[۷].

اهمیت مفصل‌ها در آمار ریاضی از قضیه‌ای موسوم به قضیه اسکلار مشخص می‌شود که بیان می‌دارد؛ هرگاه H یک تابع توزیع دو بعدی با تابع توزیع‌های حاشیه‌ای F و G باشد، در نتیجه یک مفصل C وجود دارد، به طوری که $H(x,y) = C(F(x),G(y))$.

[7] Copula function

[8] Skalr, A.

[9] Copulare

برعکس، برای هر دو تابع توزیع تک‌متغیره F و G و هر مفصل C، تابع H بالا یک تابع توزیع دو بعدی با توزیع حاشیه‌ای F و G تعریف می‌کند. همچنین، اگر F و G پیوسته باشد، C یکتاست.

فرض کنید X و Y متغیرهای تصادفی پیوسته با تابع توزیع توأم $H(x,y)$ و توزیع‌های حاشیه‌ای به ترتیب $F(x)$ و $G(y)$ باشند. متغیرهای تصادفی $U = F(X)$ و $V = G(Y)$ را در نظر می‌گیریم. بدیهی است که هر کدام از متغیرهای U و V دارای توزیع یکنواخت بر [۰و۱] تابع توزیع توأم $C(u,v)$ هستند. ایده اساسی و در عین حال ساده تبدیل هر کدام از متغیرهای تصادفی به این دلیل ارایه شده است که توزیع‌های حاشیه‌ای، توزیع یکنواخت بر فاصله‌ای به طول یک باشد و نیز گشتاورها و سایر پارامترهای توزیع دو متغیره «ناوردای مقیاس» باشند. بنابراین،

$$C(u,v) = H\left(F^{-1}(u), G^{-1}(v)\right) \quad , \quad u,v \in [0,1] = I \qquad (5)$$

بدیهی است که بنابر تعریف و بر اساس نتایج قضیه اسکلار داریم: X و Y مستقل هستند، اگر و تنها اگر $C(u,v) = uv$. در این صورت آن را مفصل حاصلضرب گوییم و با نماد $\prod(u,v)$ نشان می‌دهیم.

کران‌های تابع مفصل توسط فرشه به صورت زیر مشخص گردید،

$$W(u,v) = \max\{u+v-1, 0\} \qquad (6)$$

$$M(u,v) = \min\{u,v\}$$

$$W(u,v) \le C(u,v) \le M(u,v)$$

این کران‌ها که به ترتیب، کران بالا و پایین فرشه گفته می‌شوند، خود توابع مفصل دو متغیره هستند.

طبق قضیه اسکلار مفصل C یک تابع توزیع می‌باشد. اگر آن را با c نشان دهیم، داریم؛

$$c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v} \qquad (7)$$

در نتیجه

$$H(x_1,x_2) = C(F(x_1), G(x_2)) \qquad (8)$$

$$H(x_1,x_2) = C(u,v)$$

$$\frac{\partial^2 H(x_1,x_2)}{\partial x_1 \partial x_2} = \frac{\partial^2 C(u,v)}{\partial u \partial v} \frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_2}$$

$$h(x_1,x_2) = c(u,v) f(x_1) g(x_2)$$

۴-۲- مفصل چندمتغیره

برای توزیع‌های چندمتغیره پیوسته، حاشیه‌های تک‌متغیره و چند متغیره با ساختار وابستگی توسط مفصل‌ها قابل نمایش می‌باشند. در این حالت،

مفصل یک توزیع چندمتغیره است که تمام حاشیه‌های تک‌متغیره آن دارای توزیع یکنواخت [۰،۱] می‌باشند. بنابراین، اگر C یک مفصل باشد، تابع توزیع یک بردار تصادفی یکنواخت چندمتغیره است.

یک توزیع n متغیره H که iامین تک‌متغیره آن دارای تابع توزیع حاشیه‌ای F_j است را در نظر بگیرید. مفصل مربوط به H یک تابع توزیع به صورت $C:[0,1]^n \to [0,1]$ است که در رابطه زیر صدق می‌کند،

$$H(x)=C(F_1(x_1),\dots,F_n(x_n)) \qquad (۹)$$
$$x=(x_1,x_2,\dots,x_n)\in\Re^n$$

قضیه ۳-۱: اگر H یک تابع توزیع n متغیره پیوسته با حاشیه‌ای‌های تک‌متغیره $F_1,\dots,F_n$ باشد و تابع چندک به صورت $F_1^{-1},\dots,F_n^{-1}$ باشد، آنگاه یک انتخاب یکتا برای (۹) به صورت زیر است.

$$C(u)=H\left(F_1^{-1}(u_1),\dots,F_n^{-1}(u_n)\right) \qquad (۱۰)$$
$$u=(u_1,\dots,u_n)\in[0,1]^n$$
$$F^{-1}(u_i)=\inf\{x|F(x_i)\ge u_i\} \quad , \quad u_i\in[0,1]$$

اثبات:

برای $i=1,\dots,n$ فرض می‌کنیم $X_i\sim F_i$، لذا

$$U_i=F_i(X_i)\sim U(0,1)$$
$$C(u_1,\dots u_n)=P[U_1\le u_1,\dots,U_n\le u_n]$$
$$=P[F_1(X_1)\le u_1,\dots,F_n(X_n)\le u_n]$$

بنابر یکنوا بودن توابع توزیع

$$=P\left[X_1\le F_1^{-1}(u_1),\dots,X_n\le F_n^{-1}(u_n)\right]$$
$$=F\left(F_1^{-1}(u_1),\dots,F_n^{-1}(u_n)\right)$$

قضیه ۳-۲: فرض کنید $(X_1,\dots,X_n)^t$ برداری از متغیرهای تصادفی پیوسته با مفصل C باشد و $T_1,\dots,T_n$ توابعی پیوسته و صعودی روی دامنه $X_1,\dots,X_n$ باشند. در این صورت، $(T_1(X_1),\dots,T_n(X_n))^t$ نیز دارای مفصل C است.

اثبات: فرض کنید $F_1,\dots,F_n$ به ترتیب توابع توزیع متغیرهای تصادفی $X_1,\dots,X_n$ و $G_1,\dots,G_n$ به ترتیب تابع توزیع‌های تبدیل‌های $(T_1(X_1),\dots,T_n(X_n))^t$ باشند. همچنین، فرض کنید که $(X_1,\dots,X_n)^t$ دارای مفصل C بوده و $(T_1(X_1),\dots,T_n(X_n))^t$ مفصل C_T داشته باشد. داریم

$$G_i(x)=P[T_i(X_i)\le x]=P[X_i\le T_i^{-1}(x)]=F_i\left(T_i^{-1}(x)\right)$$
$$C_T=C_T[G_1(x_1),\dots,G_n(x_n)] \quad , \quad x\in\Re$$
$$=P[T_1(X_1)\le x_1,\dots,T_n(X_n)\le x_n]$$
$$=P[X_1\le T_1^{-1}(x_1),\dots X_n\le T_n^{-1}(x_n)]$$
$$=C[F_1(T_1^{-1}(x_1)),\dots F_n(T_n^{-1}(x_n))]$$
$$=C[G_1(x_1),\dots,G_n(x_n)]$$

چون $X_1,\dots,X_n$ پیوسته هستند و دامنه G_iها $[0,1]$ است، می‌توان نتیجه گرفت که در $[0,1]^n$، $C_T=C$ [۳].

فرض کنید $X_1,\dots,X_n$ متغیرهای تصادفی و به ترتیب دارای توابع توزیع $F_1,\dots,F_n$ و چگالی‌های $f_1,\dots,f_n$ باشند. همچنین تابع چگالی توأم آنها برابر h و تابع مفصل چگالی برابر c باشد. آنگاه با توجه به (۹) داریم،

$$h(x_1,\dots,x_n)=c(u_1,\dots,u_n)f_1(x_1)\dots f_n(x_n) \qquad (۱۱)$$

٤. تابع مفصل و مفهوم وابستگی رگرسیونی

قضیه ٤-۱: فرض کنید X و Y دو متغیر تصادفی با تابع توزیع توأم $H(x,y)$ و به ترتیب حاشیه‌های $F(x)$ و $G(y)$ باشند و C تابع مفصل باشد. آنگاه

الف: $PRD(Y|X)$ اگر و تنها اگر $\frac{\partial}{\partial u}C(u,v)$ به ازای تمام vها بر حسب u نزولی باشد.

ب: $NRD(Y|X)$ اگر و تنها اگر $\frac{\partial}{\partial u}C(u,v)$ به ازای تمام vها بر حسب u صعودی باشد.

اثبات: فرض کنید تابع C یک مفصل متغیرهای تصادفی X و Y باشد و همچنین $F(x)=u$ و $G(y)=v$. آنگاه

$$P[Y\le y|X=x]=\frac{\frac{\partial}{\partial x}H(x,y)}{\frac{d}{dx}F(x)}$$
$$=\frac{\frac{\partial}{\partial u}C(u,v)\frac{du}{dx}}{\frac{du}{dx}}=\frac{\partial}{\partial u}C(u,v)$$

با توجه به روابط بالا و تعریف وابستگی رگرسیونی هر دو حکم ثابت می‌شود [۸ و ٦].

با استفاده از ویژگی‌های هندسی تابع مفصل، برای مفاهیم وابستگی می‌توان یک تفسیر هندسی ارایه داد. بدیهی است که اگر رویه $z = C(u,v)$ منطبق بر رویه $z = \prod(u,v)$ قرار گیرد، X و Y مستقل هستند.

قضیه ٤-۲: $PRD(Y \mid X)$ اگر و تنها اگر $z = C(u,v)$ به ازای تمام vها تابعی مقعر از u باشد.

اثبات: چون تابع $f : [0,1] \to \Re$ مقعر است، اگر و تنها اگر مشتق‌های یک طرفه آن f'_+ و f'_- نزولی باشند. در نتیجه، تابع $C(u,v)$ نسبت به u مقعر است، اگر و تنها اگر $\frac{\partial C(u,v)}{\partial u}$ بر حسب u، به ازای تمام vها نزولی باشد، لذا با توجه به قضیه ٤-۱ حکم ثابت می‌شود [٦].

مثال ٤-۱: در سال ۱۹۹۷ یک شرکت هواپیمایی تصمیم به خرید یک هواپیمای جدید گرفت. صاحبان این شرکت می‌خواستند بدانند خرید هواپیما سود بیشتری به همراه می‌آورد یا یک سرمایه‌گذاری پولساز دیگر؟ آنها به آماردانی به نام ریلی مراجعه کردند تا سیستم ریسک و تصمیم را مدیریت کند و یک نمودار از عوامل مؤثر در تصمیم‌گیری و جدولی که سطح پایین، پایه و بالای متغیرهای ورودی را نشان می‌داد به او ارایه کردند. این نمودار و جدول در شکل (۲) و جدول (۱) نمایش داده شده است. در جدول ۱۰ امین، ۵۰ امین و ۹۰ امین درصد از توزیع احتمال هر کدام به عنوان سطوح متغیرها معرفی شده است [۲].

برای ریلی اولین مرحله ارایه یک تحلیل حساس، تعیین متغیرهای بحرانی بود. او پس از مطالعه و بررسی، چهار متغیر بحرانی و با نفوذ معرفی کرد. روش انتخاب این چهار متغیر، بر اساس فرآیند تصمیم حساس دو مرحله‌ای بوده است که مورد بحث ما نمی‌باشد و از آن صرفنظر می‌کنیم. این چهار متغیر عبارتند از؛

(H): ساعت پرواز ، (O): هزینه عملیات

(P): قیمت ، (C): گنجایش

اطلاعات مربوط به این متغیرهای در جدول (۲) نمایش داده شده است. مرحله دوم، مدل‌بندی توزیع توأم این متغیرها بود. او ابتدا چگالی‌های حاشیه‌ای متغیرها را مدل‌بندی کرد. برای اثبات صحت برازش، کسرهای ۰/۱۰، ۰/۵۰ و ۰/۹۰ چگالی‌های تقریبی را با جدول (۲) مقایسه کرد و نتایج به دست آمده را در جدول (۳)، شکل (۳) خلاصه نمود.

مرحله بعدی، انتخاب یک تابع مفصل مناسب است که وابستگی متغیرهای تصادفی را دربرمی‌گیرد. اگرچه در ابتدا ارتباط متغیرها بسیار پیچیده به نظر می‌رسید، اما بررسی بر روی بسیاری از یافته‌های شهودی و داده‌های قبلی نشان داد که رابطه قوی وابستگی رگرسیونی بین متغیرها برقرار است. بررسی ریلی نشان داد که اگر X_i و X_j متغیرهای بحرانی باشند، $F(x_i \mid x_j)$ غیر صعودی (غیر نزولی) بر حسب x_j برای تمام x_iها می‌باشد. در این مدل، وابستگی رگرسیونی مثبت و منفی وجود دارد و ریلی اصطلاح وابستگی رگرسیونی یکنوا را به کار برد. تعدادی از خانواده‌های مفصل را می‌شناسیم که می‌توانند وابستگی رگرسیونی را دربرگیرند. ولی انتخاب مفصل چندمتغیره نرمال بسیار معقول می‌باشد. زیرا اولا برای مدلی که وابستگی از طریق ضریب همبستگی‌ای مثل اسپیرمن بیان شده است، مناسب می‌باشد. ثانیا تغییرپذیری و راحتی تحلیل مفصل چندمتغیره نرمال، امتیاز بالایی را برای انتخاب شدن به این مفصل می‌دهد. در ادامه، چگونگی یافتن تابع مفصل برای داده‌های شرکت هوایی را شرح می‌دهیم.

خانواده گاووسی[۱۰] (نرمال)

مفصل توزیع نرمال n متغیره با ماتریس همبستگی R به صورت زیر است.

$$C_R^{Ga}(u) = \Phi_R^n\left(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n)\right) \qquad (12)$$

Φ_R^n نماد تابع توزیع توأم n متغیر از تابع توزیع نرمال استاندارد با ماتریس همبستگی R است. Φ^{-1} نماد معکوس تابع توزیع نرمال استاندارد تک‌متغیره می‌باشد. در این صورت داریم،

$$C_n(F_1, \dots, F_n) = \Phi_{R_n}^n\left(\Phi^{-1}(F_1), \dots, \Phi^{-1}(F_n)\right) \qquad (13)$$

به طوری که $\Phi_{R_n}^n$ تابع توزیع n متغیره نرمال است که متغیرهای تصادفی آن دارای همبستگی خطی R_n می‌باشند. R_n ماتریس مربعی است که عناصر قطر ۱ آن و عناصر غیر قطری برابر $2Sin(\pi\rho_{ij}/6)$ می‌کنیم که ρ_{ij} ضریب همبستگی رتبه‌ای اسپیرمن می‌باشد و F_i تابع توزیع متغیر iام است [۳].

---

[۱۰] Gaussian Capulas

فرض کنید تابع چگالی متغیر P با $f_1(P)$ و تابع توزیع این متغیر با $F_1(P)$ نمایش داده شوند. به همین ترتیب $f_2(H)$، $f_3(C)$ و $f_4(O)$ توابع چگالی و $F_2(H)$، $F_3(C)$ و $F_4(O)$ توابع توزیع H، C و O هستند. $u_1 = F_1(P)$، $u_2 = F_2(H)$، $u_3 = F_3(C)$ و $u_4 = F_4(O)$ را در نظر بگیرید. مفصل نرمال عبارت است از؛

$$C^4(u) = \Phi_R^4\left(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_4)\right) \tag{۱۴}$$

به طوری که Φ_R^4 نمایانگر تابع توزیع توأم ٤ متغیره نرمال استاندارد با ماتریس همبستگی خطی R می‌باشد و Φ^{-1} معکوس تابع توزیع نرمال استاندارد تک‌متغیره است. برای به دست آوردن R از رابطه $r_{ij} = 2Sin(\pi\rho_{ij}/6)$ استفاده می‌شود. بنابر رابطه (۱۲) می‌دانیم،

$$f(y_1,\ldots,y_4) = f_1(y_1)\ldots f_4(y_4)c(F_1(y_1),\ldots,F_4(y_4)) \,. \tag{۱۵}$$

در اینجا، یک ترفند ریاضی به کار گرفته می‌شود. $y_i = \Phi^{-1}(u_i)$ دارای توزیع نرمال استاندارد است و می‌توان نوشت

$$f_N(y_1,\ldots,y_4) = f_N(y_1)\ldots f_N(y_4)c(F_N(y_1),\ldots,F_N(y_4)) \tag{۱۶}$$

که F_N و f_N به ترتیب تابع توزیع و چگالی نرمال استاندارد تک‌متغیره هستند. در نتیجه،

$$c(F_N(y_1),\ldots,F_N(y_4)) = \frac{f_N(y_1,\ldots,y_4)}{f_N(y_1)\ldots f_N(y_4)} \tag{۱۷}$$

$$= \frac{\frac{1}{(2\pi)^2 |R|^{1/2}}\exp\left\{-\frac{1}{2}Y^T R^{-1} Y\right\}}{\prod_{i=1}^{4}\frac{1}{\sqrt{2\pi}\exp\left\{-\frac{1}{2}y_i^2\right\}}}$$

$$= \frac{\frac{1}{|R|^{1/2}}\exp\left\{-\frac{1}{2}Y^T R^{-1} Y\right\}}{\frac{1}{\exp\left\{-\frac{1}{2}Y^T Y\right\}}}$$

$$= \frac{1}{|R|^{1/2}}\exp\left\{-\frac{1}{2}Y^T\left(R^{-1} - I\right)Y\right\}$$

به طوری که $Y = (y_1,\ldots,y_4)'$ و I ماتریس همانی ٤×٤ می‌باشد. دوباره عبارت (۱۲) را به کار می‌بریم و می‌نویسیم،

$$f(P,H,C,O|R) = f_1(P)\ldots f_4(O) \tag{۱۸}$$

$$\times \frac{1}{|R|^{1/2}}\exp\left\{-\frac{1}{2}Y^T\left(R^{-1} - I\right)Y\right\}$$

$$= \frac{f_1(P)\ldots f_4(O)}{|R|^{1/2}}$$

$$\times \exp\left\{-\frac{1}{2}[\Phi^{-1}(F_1(P)),\ldots,\Phi^{-1}(F_4(O))]^T\right.$$

$$\left.\times (R^{-1} - I)(\Phi^{-1}(F_1(P))),\ldots,\Phi^{-1}(F_4(O))]\right\}$$

این تابع چگالی توأم دارای حاشیه‌های مشخص $f_1(P)$ و ... و $f_4(O)$ می‌باشد. به دلیل آنکه ضریب همبستگی اسپیرمن تحت تبدیل یک به یک ناوردا است، بنابراین R همان ضریب همبستگی متغیرهای P، H، C و O می‌باشد. ریلی با محاسبه تابع چگالی توأم، به راحتی توانست توابع چگالی شرطی و همچنین امید ریاضی شرطی را به دست آورد و وارد مرحله آخر، یعنی مرحله تصمیم‌گیری شود[۲].

۵. آزمون استقلال در مقابل PRD

لی‌من [۵] آزمونی را برای فرض استقلال در مقابل فرض (PRD) ارایه داد. آزمون پیشنهادی لی‌من به ترتیب زیر می‌باشد

$$\begin{cases} H_0: & \text{X و Y مستقل هستند} \\ H_1: & PRD(Y|X) \end{cases}$$

فرض کنید نمونه تصادفی $(x_1, y_1),\ldots,(x_n, y_n)$ در دست باشد، آماره آزمون عبارت است از؛

$$T = \frac{1}{n(n-1)}\sum_{i\neq j} Sgn(x_i - x_j)Sgn(y_i - y_j) \tag{۱۹}$$

مشخص است که آماره آزمون، برآورد τ ی کندال می‌باشد. فرض کنید تحت فرض استقلال ناحیه رد با اندازه α به صورت زیر باشد.

$$P\left(|T| \geq t_{\alpha/2}|H_0\right) = \alpha$$

که نماد $t_{\alpha/2}$ نمایانگر نقطه بحرانی توزیع دقیق T تحت فرض صفر می‌باشد. می‌توان ثابت کرد، تحت فرض استقلال آماره T دارای توزیع پارامتری متقارن است. این توزیع را می‌توان با محاسبه تعداد

جایگشت‌های رتبه‌های X و Y بــه دسـت آورد. دو مفـهوم پایـه‌ای هماهنگی و ناهماهنگی را یادآوری می‌کنیم.

تعریف ۵-۱: دو زوج مستقل $\left(X_i,Y_i\right)$ و $\left(X_j,Y_j\right)$ از متغیرهـای تصادفی که از یک توزیع دو متغیره پیروی می‌کننـد را در نظر بگیریـد. اگر داشـتـه باشیم؛ $X_i < X_j$ وقتی که $Y_i < Y_j$ یا $X_i > X_j$ وقتی که $Y_i > Y_j$ رابطه فوق یک هماهنگی (توافق) نامیده می‌شود.

تعریف ۵-۲: برای هر دو زوج مستقل، اگر داشته باشیم: $X_j < X_j$ وقتی که $Y_i > Y_j$ و یا $X_i > X_j$ وقتی که $Y_i < Y_j$، این رابطه را یک ناهماهنگی (عدم توافق) نامیده می‌شود.

در جدول پیوست کتاب آمار ناپـارامتری[۱] چندک‌هـای توزیـع دقیـق آماره $W = N_c - N_d$ آورده شده است که متغیر N_c معرف تعداد زوج‌های هماهنگی و N_d معرف تعداد زوج‌های ناهماهنگ می‌باشد.

مثال ۵-۱: توزیع دقیق آماره کندال را برای n=۳ تحت فرض H_0 به دست می‌آوریم. تعداد جایگشت‌های متمایز رتبه xها و yهـا برابـر ۳!=۶ می‌باشد. این جایگشت‌ها را در جدول ٤ نشان می‌دهیم و بـرآورد آمـاره (t) را محاسبه می‌کنیم. که در آن R_{Xi} و R_{Yi} به ترتیـب رتبـه x_i در بین xها و رتبه y_i در بین yها و n_c و n_d به ترتیـب تعـداد زوج‌هـای هماهنگ و ناهماهنگ می‌باشند. توزیـع دقیـق آمـاره کنـدال بـرای n=۳ عبارت است از

$P[\tau = t]$	t
۱/٦	۱
۲/٦	۱/۳
۲/٦	-۱/۳
۱/٦	-۱

توزیع احتمال T برای زوج‌های نمونه از هر جامعه دو متغیره به طـور مجانبی نرمال است. بنابراین، اگر بتـوان میـانگین و واریـانس T را تعیـین کرد، T در نمونه‌های بزرگ برای استنباط در مورد فرض استقلال بسیار مناسب است. چون تحت فرض استقلال X و Y آماره T متقارن بـوده و همواره $-1 \le T \le +1$ است، پس $E(T) = 0$. می‌توان ثابت کرد،

$$Var(T) = \frac{2(2n+5)}{9n(n-1)} \qquad (20)$$

و برای nهای بزرگ، تحت فرض استقلال X و Y داریم [۱]

$$Z = \frac{T-0}{\sqrt{Var T}} = \frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}\, T \xrightarrow{d} N(0,1) \qquad (21)$$

مثال ۵-۲: به عنـوان یـک مثـال سـاده بـرای آزمـون وابسـتگی مثبـت، داده‌های بیمه حوادث آتش سوزی [۱۱ و صفحات ۱٤۲-۱۳۸] را مـورد بررسی قرار می‌دهیم. در این بیمه، دو نوع هزینه وجـود دارد. هزینـه‌های نوع A [11] که هزینه‌هایی هستند که مختص عملیات شرکت بیمه می‌باشند و شامل دستمزد مشـاور حقوقـی، هزینـه‌های رسـیدگی بـه ادعاهـا و ... می‌باشند.

هزینه‌های نـوع L [12] که هزینـه‌های شـامل خسـاراتی کـه بیمـه ادعـا می‌کند، می‌باشد. در اینجا، از داده‌های واقعی بیمه آتش‌سوزی که دفـتر خدمـات بیمـه‌ای در اختیـار مـا قـرار داده، استفاده کرده‌ایـم و از آنـــها نمونه‌های تصادفی به حجم ۱۵۰۰ مورد در نظر گرفته‌ایم. بـا اسـتفاده از نـرم‌افزار S-Plus آزمون لی‌من را بر این نمونه تصـادفی انجام داده کـه برنامه و خروجـی آن در ادامـه می‌بـاشد. بـــرآورد τ کنـدال عبـارت است از $T = 0/3109967$. مقدار آمـاره آزمـون بـــرابر اســـت بـا $.Z = 18/0499$

همچنین مقدار احتمال عبارت است از؛

$$\text{P-V=P}(\tau \text{ کندال جامعه حداقل به بزرگی t مشاهده شده} \mid H_0) = 0$$

مشـاهده می‌کنیم، مقـدار احتمـال از هـر سـطح معنـی‌دار دلخواهـی کوچک‌تر باشد. بنابراین، فرض اسـتقلال بـه نفـع وابسـتگی رگرسـیونی مثبت رد می‌شود و نتیجه می‌گیریم متغیرهـای A و L در داده‌هـای بیمـه آتش‌سوزی وابستگی رگرسیونی مثبت می‌باشند.

---

[11] Allocated Loss Adjestment Exenses (ALAE's)

[12] Losses

جدول ۱

متغیر (X)	: درصد	پایین ۰/۱۰	پایه ۰/۵۰	بالا ۰/۹۰
نسبت مؤسس		٪۴۵	٪۵۰	٪۷۰
گنجایش		۴۰۵	٪۵۰	٪۶۰
سطح قیمت		$۹۵	$۱۰۰	$۱۰۸
ساعت پرواز		۵۰۰	۸۰۰	۱۰۰۰
هزینه عملیات در هر ساعت		$۲۳۰	$۲۴۵	$۲۶۰
درصد سرمایه‌گذارها		٪۳۰	٪۴۰	٪۵۰
نرخ بهره		٪۱۰/۵۰	٪۱۱/۵۰	٪۱۳/۰۰
بیمه		$۱۸۰۰۰	$۲۰۰۰۰	$۲۵۰۰۰
مالیات		$۸۵۰۰۰	$۸۷۵۰۰	$۹۰۰۰۰

جدول ۲

		پایین	پایه	بالا	ضریب همبستگی		
متغیر	:درصد	۰/۱۰	۰/۵۰	۰/۹۰	P	H	C
P		$۹۵	$۱۰۰	$۱۰۸			
H		۵۰۰	۸۰۰	۱۰۰۰	-۰/۵۰		
C		٪۴۰	٪۵۰	٪۶۰	-۰/۲۵	۰/۵۰	
O		$۲۳۰	$۲۴۵	$۲۶۰	۰	۰	۰/۲۵

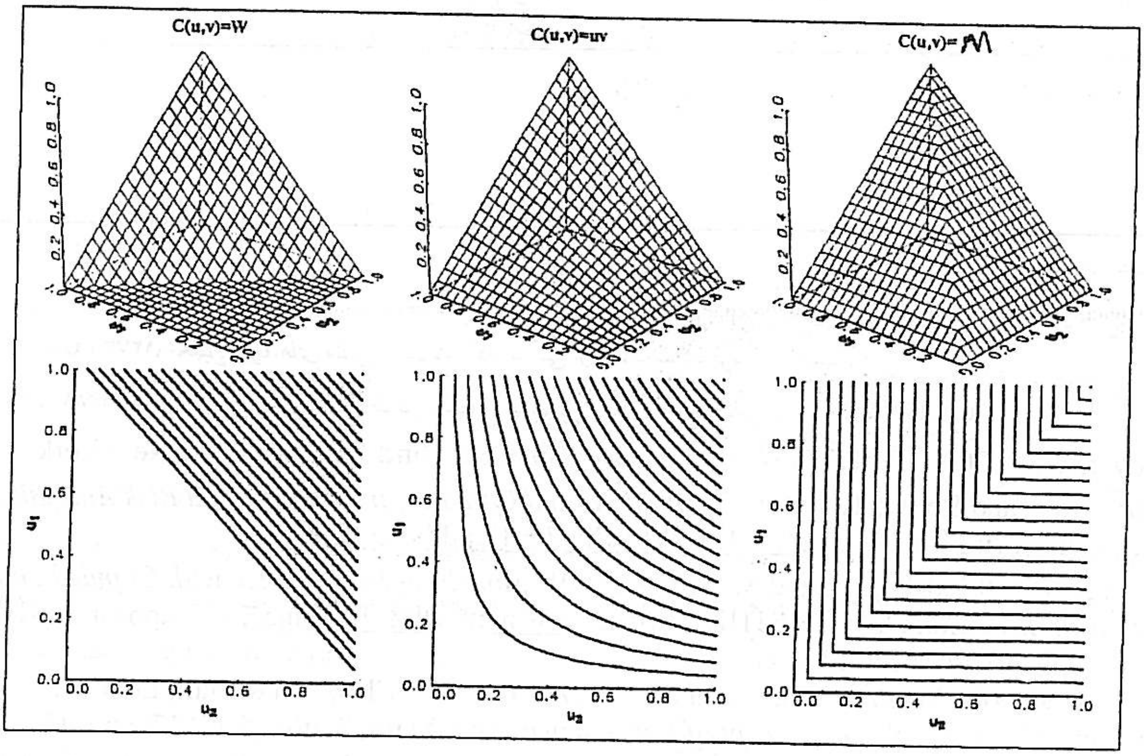
جدول ۳

متغیر	توزیع	پارامترها	دامنه
P	Scailed beta	=۹ ، =۱۵	[$۸۱/۹۴،$۱۳۳/۹۶]
H	Scailed beta	=٤ ، =۲	[۶۶/۹۱،۱۱۵۳/۲۶]
C	Beta	=۲۰ ، =۲۰	[۰،۱]
O	Normal	=۲٤٥ ، =۱۱/۷۲	$(-\infty,\infty)$

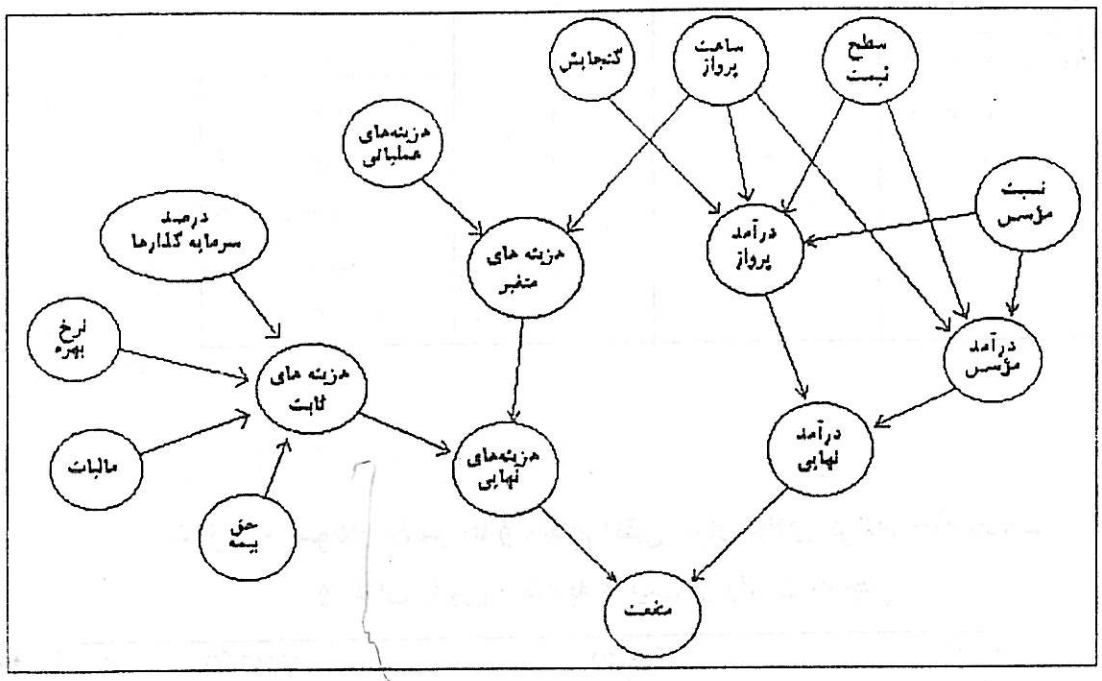
جدول ٤

$i =$ ۱ ۲ ۳	n_c	n_d	$t = (n_c - n_d) / ۳$
$R_{Xi} =$ ۱ ۲ ۳			
$R_{Yi} =$ ۱ ۲ ۳	۳	۰	۱
۱ ۳ ۲	۲	۱	۱/۳
۲ ۱ ۳	۲	۱	۱/۳
۲ ۳ ۱	۱	۲	-۱/۳
۳ ۱ ۲	۱	۲	-۱/۳
۳ ۲ ۱	۰	۱	-۱

شکل ۱- نمودار مفصل‌ها و مقطع افقی کران بالای فرشه، حاصلضرب و کران پایین فرشه به ترتیب از راست به چپ



شکل ۲



مراجع

[۱] امینی دهک، محمد؛ ۱۳۷۳، *تحلیل پیوندهای منفی*، پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد.

[۲] شفیعی‌فر، فائزه؛ ۱۳۸۱، *اندازه‌های وابستگی و ارتباط در مدلهای تصادفی*، پایان‌نامه کارشناسی ارشد، دانشگاه اصفهان.

[3] Conover, W.T., 2000. *Practical Nonparametric Statistics,* Third Edition, wiley, New York.

[4] Clemen, R.T. and Reilly, T., 1999. *Correlations and Copulas for Decision and Risk Analysis,* www.bus.utexas.edu/Faculty/jim.Dyer/DA-WP/WP970012.pdf .

[5] Emberechts, P., Lindskog, F. and Mcneil, A., 2001. *Modelling Dependence with Copulas and Applications to Risk Management,* ETHZ, Zurich, www.math.ethz.ch/finance, will appear in: Handbook of Heavy Tailed Distributions in Finance.

[6] Joe, H., 1997. *Multivariate Models and Dependence Conepts,* Chapman & Hall, London.

[7] Lehmann, E.L., 1966. *Some Concept of Dependence,* Ann. Math. Statist., Vol. 37, pp. 1137-1153.

[8] Nelsen, R.B., 1992. *On Measures of Association as Measures of Positive Dependence,* Statsistics and Probability Letters, Vol. 14, pp. 269-274.

[9] Nelsen, R.B., 1995. *Copulas, Characterization, Correlation and Counterexamples, Mathematics Magazine,* Vol. 68, pp. 193-198.

[10] Nelsen, R.B., 1999. *An Introduction to Copulas,* Springer, New York.

[11] Nelsen, R.B., 2001. *Concordance and Copulas: A Servey. Distributions with Given Marginals and Statistical Modelling,* (Cuadras, C., Fortiana, J., Rodriguez Lallena, J.A., Eds.), Kluwer Academic publishers, Dordrecht, in press.