دومین کارگاه قابلیت اعتماد و کاربردهای آن، ۱۸ و ۱۹ خرداد ۱۳۹۰، قطب علمی داده‌های ترتیبی و فضایی

برآورد ضریب همبستگی کندال برای داده‌های بقای وابسته از طریق مدل‌های مفصل ارشمیدسی دومتغیره تحت سانسور تصادفی

هادی جباری نوقابی *
استادیار گروه آمار دانشگاه فردوسی مشهد Jabbarinh@um.ac.ir
اعظم مختاری
کارشناس ارشد آمار دانشگاه آزاد اسلامی واحد مشهد
حسنعلی آذرنوش
استاد گروه آمار دانشگاه آزاد اسلامی واحد مشهد، قطب علمی داده های ترتیبی و فضایی دانشگاه فردوسی مشهد

چکیده

بردار زمان بقای (T_1, T_2) را که به وسیله مدل مفصل ارشمیدسی مدل‌بندی می‌شود، در نظر می‌گیریم. در این مقاله با برازش مفصل کلایتون به داده‌ها و تعمیم ضریب همبستگی کندال برای داده‌های سانسور شده از راست، یک برآورد گشتاوری ساده برای پارامتر وابستگی مدل‌های مفصل ارشمیدسی به دست می آوریم.

مفصل‌های ارشمیدسی، مفصل کلایتون، ضریب همبستگی τ کندال، داده‌های بقای وابسته

۱ مقدمه

تاکنون مدل‌های زیادی برای بررسی داده‌های بقای چند متغیره‌ی وابسته پیشنهاد شده‌اند. در میان آن‌ها، مدل‌های ارشمیدسی بسیار متداول هستند. در این زمینه باندین روچ ولیانگ (۱۹۹۶)، ونگ و ول (a و b ۲۰۰۰) و هوگارد(۲۰۰۰) کار کرده‌اند. براین اساس تابع بقای دو متغیره $S(t_1, t_2)$ با توابع حاشیه‌ای $S(t_2) = S(0, t_2)$ و $S(t_1) = S(t_1, 0)$ را می‌توانیم به وسیله مفصل ارشمیدسی با درنظر گرفتن وابستگی بین دو زمان بقا مدل‌بندی کنیم. در این صورت داریم:

$$S(t_1, t_2) = p[q(S_1(t_1)) + q(S_2(t_2))]$$

که در آن، تابع $q(.)$ یک تابع محدب و نزولی بر بازه $[0, 1]$ است و $q(1) = 0$. همچنین $q(.)$ معکوس تابع $p(.)$ است. جنست و ریوست (۱۹۹۳) و اکس (۱۹۸۹) در این زمینه تحقیق کرده‌اند. در این مقاله برای مدل‌بندی وابستگی بین داده‌های بقا از مفصل کلایتون استفاده می‌کنیم که در آن $p(S) = (1 + S)^{\frac{-1}{\alpha}}$. با توجه به مفصل کلایتون تابع بقای دو متغیره به صورت زیر خواهد بود:

$$S(t_1, t_2) = [\frac{1}{S_1(t_1)^{-\alpha} + S_2(t_2)^{-\alpha} - 1}]^{\frac{1}{\alpha}}.$$

*سخنران

متغیرهای تصادفی و مستقل V و U به صورت زیر تعریف می‌کنیم:

$$U = \frac{q(S_1(T_1))}{q(S_1(T_1)) + q(S_2(T_2))} \quad (1)$$

$$V = S(T_1, T_2) = p[q(S_1(T_1)) + q(S_2(T_2))] \quad (2)$$

جنست و ریوست (۱۹۹۳) ثابت کردند که اگر (T_1, T_2) از مفصل ارشمیدسی با توابع بقای حاشیه‌ای $S_1(t_1)$ و $S_2(t_2)$ پیروی کند، آن‌گاه U دارای توزیع یکنواخت بر فاصله‌ی $[0, 1]$ و V دارای توزیع کندال با تابع چگالی $k(v) = \phi(v)\frac{\phi''(v)}{\phi'(v)^2}$ است که بر فاصله‌ی $[0, 1]$ تعریف می‌شود و در آن $\phi'(.)$ و $\phi''(.)$ به ترتیب مشتقات مرتبه اول و دوم تابع $\phi(.)$ است.

فرض می‌کنیم بردار (C_1, C_2) نشان‌دهنده‌ی سانسور تصادفی راست با تابع توزیع دلخواه در ازای بردار (T_1, T_2) باشد. در بخش دوم، تابع توزیع شرطی $V = S(T_1, T_2)$ هنگامی که T_1 و T_2 هر دو یا یکی از آن‌ها سانسور شده‌اند، بررسی می‌شود. در بخش سوم برآوردگر مناسبی برای ضریب همبستگی τی کندال در معرض داده‌های سانسور شده معرفی می‌شود. در نهایت در بخش چهارم نتایج حاصل از برآوردگر ضریب همبستگی τی کندال که با استفاده از شبیه‌سازی بدست آمده است، ارایه می‌گردد.

۲ مواد و روش‌ها

در این بخش دو قضیه و یک نتیجه در مورد خواص متغیرهای تصادفی زمان بقا در معرض مشاهدات سانسور شده از راست بیان می‌شود.

فرض کنید (T_1, T_2) یک جفت متغیر تصادفی باشد طوری که تابع توزیع آن‌ها توسط یک مفصل ارشمیدسی تولید شده است. همچنین فرض کنید (T_1, T_2) بوسیله بردار (C_1, C_2) از راست سانسور شده باشد که (C_1, C_2) دارای توزیع توأم پیوسته دلخواهی است. قضیه زیر توزیع شرطی $V = S(T_1, T_2)$ را با فرض زمان‌های بقای T_1 و T_2 ارایه می‌کند.

قضیه ۱.:(ونگ و اکس، ۲۰۰۸)

۱. تابع توزیع شرطی $(V \mid T_1 > c_1, T_2 > c_2)$ برابر است با

$$F_1(v, c_1, c_2) = \frac{1}{S(c_1, c_2)}[v - \frac{\phi(v) - \phi(S(c_1, c_2))}{\phi'(v)}]; \quad 0 \le v \le S(c_1, c_2)$$

۲. تابع توزیع شرطی $(V \mid T_1 > c_1, T_2 = t_2)$ برابر است با

$$F_2(v, c_1, t_2) = \frac{\phi'(S(c_1, t_2))}{\phi'(v)}; \quad 0 \le v \le S(c_1, t_2)$$

۳. تابع توزیع شرطی $(V \mid T_1 = t_1, T_2 > c_2)$ برابر است با

$$F_2(v, t_1, c_2) = \frac{\phi'(S(t_1, c_2))}{\phi'(v)}; \quad 0 \le v \le S(t_1, c_2)$$

: تحت شرایط قضیه ۱ نتایج زیر بدست می آید:

$$i)E(V \mid T_1 > c_1, T_2 > c_2) = \frac{S(c_1,c_2)}{2} - q[S(c_1,c_2)] \int_0^1 \frac{du}{\acute{q}[uS(c_1,c_2]} + \int_0^1 \frac{q[uS(_1,c_2]}{\acute{q}[uS(c_1,c_2)]} du.$$

$$ii)E(V \mid T_1 > c_1, T_2 = t_2) = S(c_1,t_2) - S(c_1,t_2)\acute{q}[S(c_1,t_2)] \int_0^1 \frac{du}{\acute{q}[uS(c_1,t_2)]}.$$

$$iii)E(V \mid T_1 = t_1, T_2 > c_2) = S(t_1,c_2) - S(t_1,c_2)\acute{q}[S(t_1,c_2)] \int_0^1 \frac{du}{\acute{q}[uS(t_1,c_2)]}.$$

برهان:: تحت شرایط قضیه۱، برای هر $0 < v < S(c_1,c_2)$ در مورد قسمت (i) داریم:

$$\begin{aligned} E(V \mid T_1 > c_1, T_2 > c_2) &= \int_v [1 - F(v,c_1,c_2)]dv \\ &= \int_0^{S(c_1,c_2)} dv - \int_0^{S(c_1,c_2)} \frac{v}{S(c_1,c_2)} dv \\ &+ \frac{\int_0^{S(c_1,c_2)} \frac{\phi(v)}{\acute{\phi}(v)} dv}{S(c_1,c_2)} \\ &- \frac{\int_0^{S(c_1,c_2)} \frac{\phi(S(c_1,c_2)}{\acute{\phi}(v)} dv}{S(c_1,c_2)} \end{aligned} \quad (3)$$

باتغییر متغیر$u = \frac{v}{S(c_1,c_2)}$ داریم:

$$\begin{aligned} E(V \mid T_1 > c_1, T_2 > c_2) &= v \mid_0^{S(c_1,c_2)} - \frac{v^2}{2S(c_1,c_2)} \mid_0^{S(c_1,c_2)} + \int_0^1 \frac{\phi(uS(c_1,c_2))}{\acute{\phi}(uS(c_1,c_2))} du \\ &-\phi(S(c_1,c_2)) \int_0^1 \frac{du}{\acute{\phi}(uS(c_1,c_2))} \\ &= S(c_1,c_2) - \frac{S(c_1,c_2)}{2} + \int_0^1 \frac{\phi(uS(c_1,c_2))}{\acute{\phi}(uS(c_1,c_2))} du \\ &-\phi(S(c_1,c_2)) \int_0^1 \frac{du}{\acute{\phi}(uS(c_1,c_2))}. \end{aligned}$$

برای اثبات (ii) داریم:

$$\begin{aligned} E(V \mid T_1 > c_1, T_2 = t_2) &= \int_v [1 - F(v,c_1,t_2)]dv \\ &= \int_0^{S(c_1,c_2)} dv - [\int_0^1 \frac{du}{\acute{\phi}(uS(c_1,t_2))}] \acute{\phi}(S(c_1,t_2)S(c_1,t_2). \end{aligned} \quad (4)$$

برای اثبات (iii) داریم:

$$E(V \mid T_1 = t_1, T_2 > c_2) = \int_v [1 - F(V, t_1, c_2)]dv$$
$$= \int_0^{S(c_1,c_2)} dv - [\int_0^1 \frac{du}{\dot{\phi}(uS(t_1,c_2))}]\dot{\phi}(S(t_1,c_2)S(t_1,c_2) \quad (5)$$

قضیه ۲.:(ونگ و اکس، ۲۰۰۸) فرض کنید (T_1, T_2) یک بردار تصادفی باشد که توسط مفصل کلایتون ۱ مدل‌بندی می‌شود. آن‌گاه بنا به قضیه ۱ و فرع ۱ روابط زیر را حاصل می‌شوند:

$$i) \quad E(V \mid T_1 > c_1, T_2 > c_2) = \frac{S(c_1,c_2)}{2}(\frac{\alpha+1}{\alpha+2})$$
$$ii) \quad E(V \mid T_1 > c_1, T_2 = t_2) = S(c_1,t_2)(\frac{\alpha+1}{\alpha+2})$$
$$iii) \quad E(V \mid T_1 = t_1, T_2 > c_2) = S(t_1,c_2)(\frac{\alpha+1}{\alpha+2}) \quad (6)$$

پس در مدل کلایتون داریم:

$$E(V \mid T_1 > x_1, T_2 > x_2) = \frac{E(V \mid T_1 > x_1, T_2 = x_2)}{2}$$

و

$$E(V \mid T_1 > x_1, T_2 > x_2) = \frac{E(V \mid T_1 = x_1, T_2 > x_2)}{2}$$

واضح است که وقتی $\alpha = 0$ و T_1 و T_2 مستقل باشند، روابط (6) به صورت زیر نوشته می‌شوند:

$$i) \quad E(V \mid T_1 > c_1, T_2 > c_2) = \frac{S(c_1,c_2)}{4} = \frac{S_1(c_1)S_2(c_2)}{4}$$
$$ii) \quad E(V \mid T_1 > c_1, T_2 = t_2) = \frac{S_1(c_1)S_2(t_2)}{2}$$
$$iii) \quad E(V \mid T_1 = t_1, T_2 > c_2) = \frac{S_1(t_1)S_2(c_2)}{2}$$

بنابراین هنگامی که α افزایش می‌یابد، امید ریاضی‌های شرطی نیز افزایش می‌یابند. دلیل این خاصیت این است که تابع بقا $S(t_1, t_2)$ در مدل مفصل کلایتون یک تابع صعودی از α است. همچنین هنگامی که $\alpha \to \infty$

$$E(V \mid T_1 > c_1, T_2 > c_2) \to \frac{S(c_1,c_2)}{2}$$

و

$$E(V \mid T_1 > c_1, T_2 = t_2) \to S(c_1,t_2)$$

و

$$E(V \mid T_1 = t_1, T_2 > c_2) \to S(t_1,c_2).$$

Clayton ۱

۳ کاربرد

یک راه ساده برای برآورد پارامترهای مدل‌های مفصل ارشمیدسی، استفاده از ضریب همبستگی τی کندال است. این ضریب همبستگی به صورت زیر تعریف می‌شود:

$$\tau = E[sign\{T_{1i} - T_{1j}\}sign\{T_{2i} - T_{2j}\}], \circ < \tau < 1$$

که در آن (T_{1i}, T_{2i}) و (T_{1j}, T_{2j}) مشاهدات مستقل از (T_1, T_2) بوده و $sign$ تابع علامت بوده و به صورت زیر تعریف می‌شود:

$$sign(T_{1i} - T_{1j}) = \begin{cases} 1 & (T_{1i} - T_{1j}) > \circ \\ \circ & (T_{1i} - T_{1j}) = \circ \\ -1 & (T_{1i} - T_{1j}) < \circ \end{cases}$$

اگر δ_1 و δ_2 به ترتیب وضعیت سانسور T_1 و T_2 را نشان دهد، رابطه‌ی بین τ و $V = S(T_1, T_2)$ با استفاده از فرمول زیر بیان می‌شود:

$$\begin{aligned} \tau &= g(\theta) = 4E[V] - 1 \\ &= 4E[V(1-\delta_1)(1-\delta_2)] + E[V\delta_1(1-\delta_2)] + E[V(1-\delta_1)\delta_2] \\ &+ E[V\delta_1\delta_2]\} - 1 \end{aligned}$$

که در آن $\delta_1 = I(T_1 \le C_1)$ و $\delta_2 = I(T_2 \le C_2)$.
با استفاده از روابط (۳) ، (۴) و (۵) داریم:

$$\begin{aligned} g(\theta) &= \frac{4}{n}\sum_i [\frac{S(C_{1i}, C_{2i})}{2} - q(S(C_{1i}, C_{2i})\int_\circ^1 \frac{du}{\acute{q}(uS(C_{1i}, C_{2i}))} \\ &+ \int_\circ^1 \frac{q(uS(C_{1i}, C_{2i}))}{\acute{q}(uS(C_{1i}, C_{2i}))}du](1-\delta_{1i})(1-\delta_{2i}) \\ &+ \frac{4}{n}\sum_i [S(C_{1i}, T_{2i}) - S(C_{1i}, T_{2i})\acute{q}(S(C_{1i}, T_{2i})) \\ &\times \int_\circ^1 \frac{du}{\acute{q}(uS(C_{1i}, T_{2i}))}]\delta_{1i}(1-\delta_{2i}) \\ &+ \frac{4}{n}\sum_i [S(T_{1i}, C_{2i}) - S(T_{1i}, C_{2i})\acute{q}(S(T_{1i}, C_{2i})) \\ &\times \int_\circ^1 \frac{du}{\acute{q}(uS(T_{1i}, C_{2i}))}](1-\delta_{1i})\delta_{2i} \\ &+ \frac{4}{n}\sum_i S(T_{1i}, T_{2i})\delta_{1i}\delta_{2i} - 1 \end{aligned} \quad (۷)$$

بنا به نتایج جنست و ریوست (۱۹۹۳) در مدل مفصل کلایتون داریم: $\tau = \frac{\alpha}{\alpha+2}$. پس با تعریف $X_{1i} = min(T_{1i}, C_{1i})$ و $X_{2i} = min(T_{2i}, C_{2i})$ و جایگذاری آن‌ها در رابطه (۷) داریم:

$$
\begin{aligned}
(\frac{\alpha}{\alpha+۲})+۱ &= \frac{۴}{n}\sum_i S(X_{۱i},X_{۲i})(۱-\delta_{۱i})(۱-\delta_{۲i})[\\
&\quad (\frac{۱}{۲})+\frac{\delta_{۱i}}{(۱-\delta_{۱i})}+\frac{\delta_{۲i}}{(۱-\delta_{۲i})}+\frac{\delta_{۱i}\delta_{۲i}}{(۱-\delta_{۱i})(۱-\delta_{۲i})}(\frac{\alpha+۲}{\alpha+۱})] \\
&= \frac{۴}{n}(\frac{\alpha+۱}{\alpha+۲})\sum_i S(X_{۱i},X_{۲i})(۱-\delta_{۱i})(۱-\delta_{۲i})[\\
&\quad \frac{(\alpha+۱)+(\alpha+۱)\delta_{۱i}+(\alpha+۱)\delta_{۲i}+(۱-\alpha)\delta_{۱i}\delta_{۲i}}{۲(۱-\delta_{۱i})(۱-\delta_{۲i})(\alpha+۱)}]
\end{aligned}
$$

در نتیجه با کمی محاسبات ساده داریم:

$$\sum_i S(X_{۱i},X_{۲i})[۱+\delta_{۱i}+\delta_{۲i}+\frac{۱-\alpha}{۱+\alpha}\delta_{۱i}\delta_{۲i}]-n=۰ \qquad (۷)$$

حال اگر در رابطه‌ی (۷) به جای α از برآورد آن به صورت $\hat{\alpha}_n=\frac{۲\hat{\tau}_n}{۱-\hat{\tau}_n}$ استفاده کنیم، می توانیم برآوردی برای ضریب همبستگی τ ی کندال به دست آوریم زمانی که داده‌ی سانسور شده داریم. با استفاده از برآورد تابع بقای دو متغیره نیز که توسط دابروسکا(۱۹۸۸) بیان شده است، داریم:

$$\hat{\tau}_n=\frac{\sum_i \hat{S}(X_{۱i},X_{۲i})(۱+\delta_{۱i}+\delta_{۲i}+\delta_{۱i}\delta_{۲i})-n}{\sum_i \hat{S}(X_{۱i},X_{۲i})(-۱-\delta_{۱i}-\delta_{۲i}+۳\delta_{۱i}\delta_{۲i})+n}$$

بدیهی است هنگامی که هیچ داده‌ی سانسورشده‌ای نداشته باشیم(یعنی به ازای هرi، $\delta_{۱i}=\delta_{۲i}=۱$)، خواهیم داشت:

$$\hat{\tau}_n=\frac{۴}{n}\sum_{i=۱}^{n}\hat{S}(T_{۱i},T_{۲i})-۱=\frac{۴}{n}\sum_{i=۱}^{n}\hat{V}_i-۱$$

۴ بحث و نتیجه گیری

در این مقاله با استفاده از قضیه ۱ و نتیجه‌ی حاصل از آن و مدل مفصل کلایتون یک برآورد گشتاوری برای τی کندال زمانی که داده‌ی سانسورشده داریم، ارایه شد. همچنین با استفاده از قضیه ۱ می‌توان پارامتر مجهول مدل‌های مفصل ارشمیدسی را برآورد کرد.

در مثال زیر کابرد برآوردگر ضریب همبستگی کندال را زمانی که با مشاهدات سانسورشده‌ی تصادفی از راست سرو کار داریم، توضیح می‌دهیم.

مثال ۱.: در یک مطالعه شبیه‌سازی شده، ابتدا یک نمونه‌ی تصادفی ۱۰۰ تایی از مدل مفصل کلایتون با پارامترهای $\alpha=۰.۵, ۱.۳۳, ۳, ۸$ $(\tau=۰.۲, ۰.۴, ۰.۶, ۰.۸)$، بردارهای زمان را تولید می کنیم. بردارهای سانسور را نیز از توزیع یکنواخت در بازه $(۰, ۱)$ تولید می کنیم. پس از تولید داده‌ها، مشاهده می کنیم که ۲۰ درصد آن‌ها سانسور شده‌اند. برای ارزیابی دقت برآوردگر ضریب همبستگی τی کندال در حالت داده‌های سانسور شده، اریبی و انحراف معیار برآوردها را به صورت تجربی با تکرار مستقل فرآیند شبیه‌سازی به اندازه‌ی ۱۰۰۰ بار محاسبه می کنیم. جدول ۱ خلاصه‌ی نتایج شبیه‌سازی با نرم افزار R را نشان می دهد.

جدول ۱: برآورد اریبی و انحراف معیار برآوردهای ضریب همبستگی τ ی کندال

α	۰/۵	۱/۳۳	۳	۸
اریبی برآورد شده	−۰/۷۵۹۸	−۰/۵۸۴۸	−۰/۳۹۲۴	−۰/۱۹۷
انحراف معیار برآورد شده	۰/۰۲۷۴	۰/۰۰۹۲	۰/۰۰۴۹	۰/۰۰۱۷

با مقایسه داده‌های جدول ۱ مشاهده می‌کنیم که با افزایش α، قدر مطلق اریبی و انحراف معیار برآوردهای ضریب همبستگی کندال کاهش می‌یابد.

۵ مراجع

Bandeen-Roche, K. J., Liang, K.Y. (1996). Modelling failure-timeassociations in data with multiple levels of clustering. *Biometrika.*, **83**, 29–39.

Dabrowska, D.M., (1988). Kaplan - Meier estimate on the plane. *Ann. Statist* , **16**, 1475–1489.

Genest, C., Rivest, L.P., (1993). Statistical inference procedures for bivariate Archimedean copulas. *J.Amer. Statist. Assoc.*, **88**, 1034–1043.

Hougaard, P., (2000). Analysis of Multivariate Survival Data. Springer-Verlag, New York, Inc.

Wang, A., Oakes, D., (2008). Some properties of the Kendall distribution in bivariate Archimedean copula models under censoring. *Statistics. Probability.*, **78**, 2578–2583.

Wang, J., Zafra, P., (2009). Estimating Bivariate Survival Function by Volterra Estimator Using Dynamic Programming Techniques. *Journal of Data Science.*, **7**, 365–380.

Wang, W., Wells, M.T., (2000a). Model selection and semiparametric inference for bivariate failure-time data. *J. Amer. Statist. Assoc.*, **449**, 62–72.

Wang, W., Wells, M.T., (2000b). Estimation of kendalls tau under censoring. *Statist. Sinica.*, **10**, 1199–1218.

Archimedean Copula, Clayton Copula,Kendall τ

Jabbari, H[1].Mokhtari, A[2].andAzarnoosh, H.A[3]

Department of Statistics, Ferdowsi University of Mashhad, mashhad, Iran.

Department of Statistics, Azad Islamic University of Mashhad, mashhad, Iran.

Department of Statistics, Azad Islamic University of Mashhad, mashhad, Iran.

Abstract: Survival times vector (T_1, T_2) *can be Modelled by an Archimedean Copula Model. In this paper by fitting clayton copula for data and kendall correlation coefficients for censored data and We obtaind, asimple Moment estimator of the dependence parameter in Archimedean Copula Models is proposed.*

Keywords: Archimedean Copula, Clayton Copula,Kendall τ.

Mathematics Subject Classification (2000): 62N02