Testowanie warunkowej niezależności

Melka Kamil

Uniwersytet Wrocławski

13.06.2022

Plan prezentacji

Zastosowanie

Warunkowa niezależność znajduje zastosowanie w

- statystyce dostatecznej (Fisher, 1920),
- statystyce swobodnej (Fisher, 1934),
- w modelach grafowych (Koller i Friedman, 2009).

Opis zagadnienia warunkowej niezależności

Rozważmy trzy wektory losowe X, Y i Z, które przyjmują wartości w $\mathbb{R}$, $\mathbb{R}$ i $\mathbb{R}^d$, odpowiednio. Załóżmy, że łączny rozkład jest bezwzględnie ciągły względem miary Lebesgue'a z gęstością p. Mówimy, że X jest warunkowo niezależny z Y przy danym Z i zapisujemy to jako

$$X \perp Y|Z,$$

jeśli dla każdych x, y, z z $p(z) > 0$, mamy $p(x, y|z) = p(x|z)p(y|z)$.

Opis zagadnienia warunkowej niezależności

Niech $L^2_{X,Z}$ opisuje przestrzeń wszystkich funkcji $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$, takich że $\mathbb{E}f(X,Z)^2 < \infty$. $L^2_{Y,Z}$ definiujemy analogicznie. Daudin (1980) udowodnił, że X i Y są warunkowo niezależne przy danym Z wtedy i tylko wtedy, gdy

$$\mathbb{E}f(X,Z)g(Y,Z) = 0 \tag{1}$$

dla wszystkich funkcji $f \in L^2_{X,Z}$ i $g \in L^2_{Y,Z}$, takich że $\mathbb{E}[f(X,Z)|Z] = 0$ i $\mathbb{E}[g(Y,Z)|Z] = 0$, odpowiednio.

Testowanie hipotez statystycznych

Zapiszmy $\mathbb{E}_P(\cdot)$ jako wartość oczekiwaną zmiennej losowej, której rozkład jest określony przez P, podobnie $\mathbb{P}_P(\cdot) = \mathbb{E}_P \mathbb{1}_{\{\cdot\}}$. Niech $\mathcal{H}$ będzie potencjalnym zbiorem hipotez zerowych składających się z kolekcji rozkładów dla (X, Y, Z). Dla $i = 1, 2, \ldots$ niech $(x_i, y_i, z_i) \in \mathbb{R}^{1+1+d}$ będą i.i.d. kopiami (X, Y, Z).

Testowanie hipotez statystycznych

Wtedy $\mathbf{X}^{(n)} \in \mathbb{R}^{1\cdot n}$, $\mathbf{Y}^{(n)} \in \mathbb{R}^{1\cdot n}$ i $\mathbf{Z}^{(n)} \in \mathbb{R}^{d\cdot n}$ będą macierzami z i-tymi wierszami x_i, y_i i z_i, odpowiednio. Niech Ψ_n będzie potencjalnym zrandomizowanym testem, którego można użyć na danych $(\mathbf{X}^{(n)}, \mathbf{Y}^{(n)}, \mathbf{Z}^{(n)})$.

$$\Psi_n : \mathbb{R}^{(1+1+d)\cdot n} \times [0,1] \to \{0,1\}$$

jest mierzalną funkcją, której ostatni argument jest zarezerwowany dla zmiennej losowej $U \sim \mathcal{U}[0,1]$ niezależnej do danych, która odpowiada za zrandomizowanie testu.

Pożądane własności testu

Mając ciąg testów $(\Psi_n)_{n=1}^{\infty}$ jesteśmy zainteresowani w spełnieniu następujących własności. Mając określony poziom istotności $\alpha \in (0,1)$ i hipotezy zerowe $\mathcal{H}$, mówimy, że test Ψ_n ma

odpowiedni poziom dla próby rozmiaru n jeśli $\sup_{P \in \mathcal{H}} \mathbb{P}_P(\Psi_n = 1) \leq \alpha,$

gdzie lewa strona nierówności jest rozmiarem testu. Ciąg $(\Psi_n)_{n=1}^{\infty}$ ma

jednostajnie asymptotyczny poziom jeśli $\limsup_{n\to\infty} \sup_{P \in \mathcal{H}} \mathbb{P}_P(\Psi_n = 1) \leq \alpha,$

punktowo asymptotyczny poziom jeśli $\sup_{P \in \mathcal{H}} \limsup_{n\to\infty} \mathbb{P}_P(\Psi_n = 1) \leq \alpha.$

Pożądane własności testu

Mając ciąg testów $(\Psi_n)_{n=1}^{\infty}$ oraz zbiór hipotez alternatywnych $\mathcal{A}$, jest pożądane, aby moc była jednostajnie duża na zbiorze $\mathcal{A}$. Chcemy mieć spełnione $\inf_{P\in\mathcal{A}} \mathbb{P}_P(\Psi_n = 1) \to 1$. W przypadku testowania hipotez może się zdarzyć, że żaden test nie osiąga mocy przeciwko dowolnej alternatywie, tj. $\sup_{P\in\mathcal{A}} \mathbb{P}_P(\Psi_n = 1) \leq \alpha$. To znaczy, że dla każdych n, testów Ψ_n i rozkładów z alternatywy $P \in \mathcal{A}$, mamy

$$\mathbb{P}_P(\Psi_n = 1) \leq \sup_{Q\in\mathcal{H}} \mathbb{P}_Q(\Psi_n = 1).$$

Wtedy problem testowania hipotez zdefiniowany przez parę $(\mathcal{H}, \mathcal{A})$ jest określany jako nietestowalny.

Nietestowalność warunkowej niezależności

Zdefiniujmy $\mathcal{E}_0$ jako zbiór wszystkich rozkładów (X, Y, Z), które są bezwzględnie ciągłe względem miary Lebesgue'a. Wtedy przez $\mathcal{H}_0 \subset \mathcal{E}_0$ określamy podzbiór rozkładów, dla których $X \perp Y|Z$. Dalej, dla każdego $M \in (0, \infty]$, niech $\mathcal{E}_{0,M} \subseteq \mathcal{E}_0$ będzie podzbiorem wszystkich rozkładów, których nośnik jest ściśle zawarty w ℓ_∞ kuli o promieniu M $(\mathcal{E}_{0,\infty} = \mathcal{E}_0)$. Również definiujemy $\mathcal{A}_0 = \mathcal{E}_0 \setminus \mathcal{H}_0$ oraz zbiory $\mathcal{H}_{0,M} = \mathcal{E}_{0,M} \cap \mathcal{H}_0$ i $\mathcal{A}_{0,M} = \mathcal{E}_{0,M} \cap \mathcal{A}_0$. Rozważmy konfigurację z Sekcji 1.2 z hipotezami zerowymi $\mathcal{H} = \mathcal{H}_{0,M}$.

Nietestowalność warunkowej niezależności

Twierdzenie

Mając dowolne $n \in \mathbb{N}$, $\alpha \in (0,1)$, $M \in (0,\infty]$ i potencjalny zrandomizowany test Ψ_n, który ma odpowiedni poziom α dla hipotez zerowych $\mathcal{H}_{0,M}$, to wtedy $\mathbb{P}_P(\Psi_n = 1) \leq \alpha$ dla wszystkich $P \in \mathcal{A}_{0,M}$. Stąd Ψ_n nie ma mocy przeciwko dowolnej alternatywie.

Nietestowalność warunkowej niezależności

Wniosek

Dla każdego $M \in (0, \infty]$ *i dla dowolnego ciągu* $(\Psi_n)_{n=1}^{\infty}$ *testów mamy*

$$\sup_{P \in \mathcal{A}_{0,M}} \limsup_{n \to \infty} \mathbb{P}_P(\Psi_n = 1) \leq \limsup_{n \to \infty} \sup_{Q \in \mathcal{H}_{0,M}} \mathbb{P}_Q(\Psi_n = 1).$$

Model

Mając dany rozkład P dla (X, Y, Z), możemy zapisać model w postaci

$$X = f_P(Z) + \varepsilon_P, \quad Y = g_P(Z) + \xi_P,$$

gdzie $f_P(z) = \mathbb{E}_P(X|Z = z)$ i $g_P(z) = \mathbb{E}_P(Y|Z = z)$. Podobnie, dla $i = 1, 2, \ldots$, definiujemy $\varepsilon_{P,i}$ i $\xi_{P,i}$ przez $x_i - f_P(z_i)$ i $y_i - g_P(z_i)$, odpowiednio. Również określmy $u_P(z) = \mathbb{E}_P(\varepsilon_P^2|Z = z)$ i $v_P(z) = \mathbb{E}_P(\xi_P^2|Z = z)$.

Statystyka testowa

Niech $\hat{f}^{(n)}$ i $\hat{g}^{(n)}$ będą estymatorami warunkowych wartości oczekiwanych f_P i g_P, np. regresując $\mathbf{X}^{(n)}$ i $\mathbf{Y}^{(n)}$ na $\mathbf{Z}^{(n)}$. Dla $i = 1, \ldots, n$ obliczamy iloczyn między residuami z regresji:

$$R_i = (x_i - \hat{f}(z_i))(y_i - \hat{g}(z_i)).$$

Następnie definiujemy $T^{(n)}$ jako znormalizowaną sumę R_i:

$$T^{(n)} = \frac{\sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} R_i}{(\frac{1}{n} \sum_{i=1}^{n} R_i^2 - (\frac{1}{n} \sum_{j=1}^{n} R_j)^2)^{1/2}}.$$

Asymptotyczny rozkład $T^{(n)}$

Twierdzenie

Zdefiniujmy poniższe wielkości:

$$A_f := \frac{1}{n}\sum_{i=1}^{n}(f_P(z_i)-\hat{f}(z_i))^2, \quad B_f := \frac{1}{n}\sum_{i=1}^{n}(f_P(z_i)-\hat{f}(z_i))^2 v_P(z_i),$$

$$A_g := \frac{1}{n}\sum_{i=1}^{n}(g_P(z_i)-\hat{g}(z_i))^2, \quad B_g := \frac{1}{n}\sum_{i=1}^{n}(g_P(z_i)-\hat{g}(z_i))^2 u_P(z_i).$$

Asymptotyczny rozkład $T^{(n)}$

Twierdzenie

Wtedy mamy następujące rezultaty:

(i) *Jeśli dla* $P \in \mathcal{H}_0$, $A_f A_g = o_P(n^{-1})$, $B_f = o_P(1)$, $B_g = o_P(1)$ *oraz* $0 < \mathbb{E}_P(\varepsilon_P^2 \xi_P^2) < \infty$, *to wtedy*

$$\sup_{t \in \mathbb{R}} |\mathbb{P}_P(T^{(n)} \leq t) - \Phi(t)| \to 0.$$

(ii) *Niech* $\mathcal{H} \subset \mathcal{H}_0$ *będzie klasą rozkładów, takich że* $A_f A_g = o_{\mathcal{H}}(n^{-1})$, $B_f = o_{\mathcal{H}}(1)$, $B_g = o_{\mathcal{H}}(1)$. *Jeśli dodatkowo* $\inf_{P \in \mathcal{H}} \mathbb{E}(\varepsilon_P^2 \xi_P^2) \geq c_1$ *i* $\sup_{P \in \mathcal{H}} \mathbb{E}_P\{|\varepsilon_P \xi_P|^{2+\eta}\} \leq c_2$ *dla pewnych* $c_1, c_2 > 0$ *i* $\eta > 0$, *to wtedy*

$$\sup_{P \in \mathcal{H}} \sup_{t \in \mathbb{R}} |\mathbb{P}_P(T^{(n)} \leq t) - \Phi(t)| \to 0.$$

Moc testu

Statystyka testowa $T^{(n)}$ jest znormalizowaną wersją warunkowej kowariancji $\mathbb{E}_P(\varepsilon_P \xi_P) = \mathbb{E}_P \mathbf{cov}_P(X, Y|Z)$, gdzie $\mathbf{cov}_P(X, Y|Z) = \mathbb{E}_P(XY|Z) - \mathbb{E}_P(X|Z)\mathbb{E}_P(Y|Z)$. Z równania (1), wiemy że przy hipotezie zerowej jest to równe 0. Przy alternatywie niekoniecznie tak musi być, możemy mieć tylko nadzieję, że test będzie miał moc przeciwko alternatywom, dla których warunkowa kowariancja jest niezerowa.

Kopuła

Definicja

Kopuła to dystrybuanta m-wymiarowego rozkładu prawdopodobieństwa na $[0,1]^m$ o jednostajnych rozkładach brzegowych.

Z tego powodu ograniczymy się do zbioru rozkładów $\mathcal{P}_{[0,1]} \subset \mathcal{P}$, których nośnik to $[0,1]^{2+d}$. Wtedy $X, Y \in [0,1]$ oraz $Z \in [0,1]^d$.

Warunkowa dystrybuanta i warunkowa funkcja kwantylowa

Mając dany $z \in [0,1]^d$ określamy przez

$$F_{X|Z}(t|z) := P(X \leq t|Z = z)$$

warunkową dystrybuantę $X|Z = z$ dla $t \in [0,1]$. Przez

$$Q_{X|Z}(\tau|z) := \inf\{t \in [0,1]|F_{X|Z}(t|z) \geq \tau\}$$

określamy warunkową funkcję kwantylową warunkowej dystrybuanty $X|Z = z$ dla $\tau \in [0,1]$.

Estymator warunkowej funkcji kwantylowej

W regresji kwantylowej modelujemy funkcję $z \to Q(\tau|z)$ dla ustalonych $\tau \in [0,1]$. Estymacja funkcji regresji kwantylowej jest realizowana przez rozwiązanie problemu minimalizacji empirycznego ryzyka

$$\hat{Q}(\tau|\cdot) \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L_\tau(X_i - f(Z_i)),$$

gdzie funkcja straty to $L_\tau(u) = u(\tau - \mathbb{1}_{\{u<0\}})$ i $\mathcal{F}$ jest pewną klasą funkcji.

Aproksymacja warunkowej dystrybuanty

Bazując na warunkowej funkcji kwantylowej Q możemy zdefiniować aproksymację $\tilde{F}^{(m)}$ warunkowej dystrybuanty F. Niech $\tau_{\min}$ i $\tau_{\max}$ będą konkretnymi wartościami kwantyli, spełniającymi $0 < \tau_{\min} < \tau_{\max} < 1$ oraz niech $q_{\min,z} := Q(\tau_{\min}|z) > 0$ i $q_{\max,z} := Q(\tau_{\max}|z) < 1$ oznaczają odpowiadające warunkowe kwantyle.
Niech $\mathcal{T} = [\tau_{\min}, \tau_{\max}]$ oznacza zbiór potencjalnych kwantyli. Podział w $\mathcal{T}$ jest ciągiem $(\tau_k)_{k=1}^m$, takim że $\tau_{\min} = \tau_1 < \cdots < \tau_m = \tau_{\max}$ dla $m \geq 2$. Równomierny podział jest takim podziałem $(\tau_k)_{k=1}^m$, dla którego $\tau_{k+1} - \tau_k$ jest stały dla $k = 1, \ldots, m-1$. Również niech $\tau_0 = 0$ i $\tau_{m+1} = 1$ będą wyznaczone.

Aproksymacja warunkowej dystrybuanty

Mając dany podział $(\tau_k)_{k=1}^m$ oznaczamy $q_{k,z} := Q(\tau_k|z)$ dla $k = 1, \ldots, m$ i definiujemy $q_{0,z} := 0$ i $q_{m+1,z} := 1$. Dla każdego $z \in [0,1]^d$ definiujemy funkcję $\tilde{F}^{(m)}(\cdot|z) : [0,1] \to [0,1]$ przez liniową interpolację punktów $(q_{k,z}, \tau_k)_{k=0}^{m+1}$:

$$\tilde{F}^{(m)}(t|z) := \sum_{k=0}^{m} \left(\tau_k + (\tau_{k+1} - \tau_k) \frac{t - q_{k,z}}{q_{k+1,z} - q_{k,z}} \right) \mathbb{1}_{(q_{k,z}, q_{k+1,z}]}(t).$$

Estymator warunkowej dystrybuanty bazujący na regresji kwantylowej

Definicja

Niech $(\tau_k)_{k=1}^m$ *będzie podziałem w* $\mathcal{T}$. *Zdefiniujmy* $\hat{q}_{0,z}^{(n)} := 0$ *i* $\hat{q}_{m+1,z}^{(n)} := 1$ *i niech* $\hat{q}_{k,z}^{(n)} := \hat{Q}^{(n)}(\tau_k|z)$ *dla* $k = 1, \ldots, m$ *będą predykcjami modelu regresji kwantylowej wyznaczonego z i.i.d. próby* $(X_i, Z_i)_{i=1}^n$. *Definiujemy estymator* $\hat{F}^{(m,n)}(\cdot|z) : [0,1] \to [0,1]$ *przez*

$$\hat{F}^{(m,n)}(t|z) := \sum_{k=0}^{m} \left(\tau_k + (\tau_{k+1} - \tau_k) \frac{t - \hat{q}_{k,z}^{(n)}}{\hat{q}_{k+1,z}^{(n)} - \hat{q}_{k,z}^{(n)}} \right) \mathbb{1}_{(\hat{q}_{k,z}^{(n)}, \hat{q}_{k+1,z}^{(n)}]}(t)$$

dla każdego $z \in [0,1]^d$.

Model regresji kwantylowej

W celu uzyskania estymatora kwantyli, będziemy używać modelu regresji kwantylowej, dla którego uzyskano wyniki dotyczące zgodności. Nasz model będzie miał postać

$$Q(\tau|z) = h(z)^T \beta_\tau, \tag{2}$$

gdzie $h : [0,1]^d \to \mathbb{R}^p$ jest znaną i ciągłą transformacją Z, np. wielomian lub funkcja sklejana, które mogą służyć do modelowanie nieliniowych efektów.

Cząstkowa kopuła

Zdefiniujmy parę nowych zmiennych losowych przez transformację Rosenblatta

$$U_1 := F_{X|Z}(X|Z) \text{ i } U_2 := F_{Y|Z}(Y|Z).$$

Stwierdzenie

Spełnione jest to, że $U_l \sim \mathcal{U}[0,1]$ *i* $U_l \perp Z$ *dla* $l = 1, 2$.

Ta transformacja może być rozumiana jako normalizacja, gdzie brzegowe zależności między X na Z oraz między Y na Z zostały odfiltrowane. Łączny rozkład dla U_1 i U_2 jest określany jako cząstkowa kopuła dla X i Y przy danym Z.

Cząstkowa kopuła

Stwierdzenie

Jeśli $X \perp Y|Z$, to wtedy $U_1 \perp U_2$.

Wtedy pytanie o warunkową niezależność może być zamienione na pytanie o niezależność. Należy jednak pamiętać, że $U_1 \perp U_2$ nie musi implikować tego, że $X \perp Y|Z$.

Ogólna procedura testowania

Definicja

Niech $(X_i, Y_i, Z_i)_{i=1}^n$ *będzie i.i.d. próbą z* $P \in \mathcal{P}_0$, *gdzie* $\mathcal{P}_0 \subset \mathcal{P}_{[0,1]}$. *Przez* $\mathcal{H}_0 := \mathcal{P}_0 \cap \mathcal{H}$ *i* $\mathcal{A}_0 := \mathcal{P}_0 \cap \mathcal{A}$ *oznaczamy rozkłady w* $\mathcal{P}_0$, *które spełniają warunkową niezależność i warunkową zależność, odpowiednio. Również niech* ψ_n *oznacza potencjalny niezrandomizowany test na niezależność w ciągłym dwuwymiarowym rozkładzie. Wtedy ogólna procedura przeprowadzania testu wygląda następująco.*

(i) Wyznacz estymatory $\hat{F}^{(n)}_{X|Z}$ *i* $\hat{F}^{(n)}_{Y|Z}$ *oparte na* $(X_i, Y_i, Z_i)_{i=1}^n$.

(ii) Oblicz wyestymowane nieparametryczne residua

$$\hat{U}^{(n)}_{1,i} := \hat{F}^{(n)}_{X|Z}(X_i|Z_i) \text{ i } \hat{U}^{(n)}_{2,i} := \hat{F}^{(n)}_{Y|Z}(Y_i|Z_i)$$

dla $i = 1, \ldots, n$.

(iii) Niech $\hat{\Psi}_n := \psi_n\big((\hat{U}^{(n)}_{1,i}, \hat{U}^{(n)}_{2,i})_{i=1}^n\big)$ *i odrzucamy hipotezę* $X \perp Y|Z$ *jeśli* $\hat{\Psi}_n = 1$.

Uogólniona miara korelacji

Poniżej zostanie zdefiniowana uogólniona miara korelacji przedstawiona w pracy Petersena i Hansena (2021), która posłuży jako baza do testu niezależności pomiędzy nieparametrycznymi residuami U_1 i U_2.

Definicja

Uogólniona korelacja pomiędzy U_1 i U_2 jest zdefiniowana w języku wielowymiarowej funkcji $\varphi = (\varphi_1, \ldots, \varphi_q) : [0,1] \to \mathbb{R}^q$ jako

$$\rho = \mathbb{E}_P(\varphi(U_1)\varphi(U_2)^T),$$

gdzie ρ jest macierzą rozmiaru $q \times q$ z elementami $\rho_{kl} = \mathbb{E}_P(\varphi_k(U_1)\varphi_l(U_2))$ dla $k, l = 1, \ldots, q$.

Uogólniona miara korelacji

Zakładamy, że funkcja $\varphi = (\varphi_1, \ldots, \varphi_q)$ spełnia poniższe założenia.

Założenia

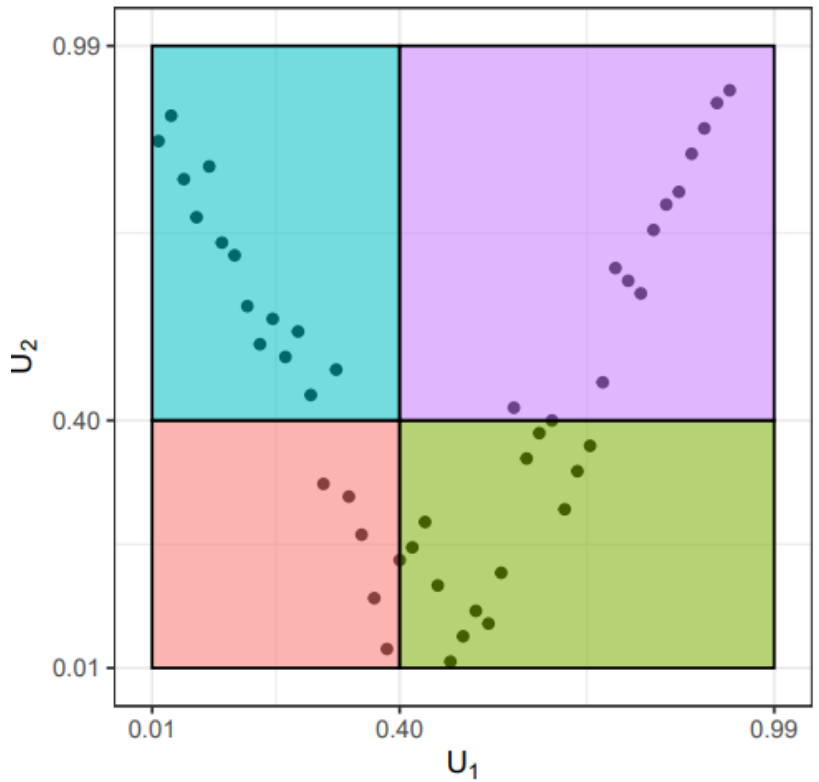
(i) *Nośnik* $\mathcal{T}_k$ *dla każdej funkcji* φ_k *jest zwartym podzbiorem odcinka* $(0, 1)$.

(ii) *Każda funkcja* $\varphi_k : [0, 1] \to \mathbb{R}$ *spełnia warunek Lipschitza.*

(iii) $\int_0^1 \varphi_k(u)du = 0$ *i* $\int_0^1 \varphi_k(u)^2 du = 1$ *dla każdego* $k = 1, \ldots, q$.

(iv) *Funkcje* $\varphi_1, \ldots, \varphi_q$ *są liniowo niezależne.*

Interpretacja ρ_{kl}



Rysunek: Próba z kopuły (U_1, U_2) z widoczną zależnością, gdzie próbkowa korelacja jest bliska 0. Zależność jest wykrywana przez branie korelacji między obserwacjami w danych regionach ogólnej przestrzeni. Źródło: *Fig. 1, L. Petersen and N.R.Hansen, Testing Conditional Independence via Quantile Regression Based Partial Copulas, 2021*

Test bazujący na uogólnionej korelacji

Dla ρ określającego uogólnioną korelację pomiędzy U_1 i U_2, możemy zdefiniować $\rho_n : [0,1]^{2n} \to \mathbb{R}^{q \times q}$ jako empiryczną wersję ρ:

$$\rho_n((u_i, v_i)_{i=1}^n) := \frac{1}{n} \sum_{i=1}^{n} \varphi(u_i)\varphi(v_i)^T.$$

Stąd można uzyskać postać statystyki testowej

$$\hat{\rho}_n := \rho_n\big((\hat{U}_{1,i}, \hat{U}_{2,i})_{i=1}^n\big) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\hat{U}_{1,i})\varphi(\hat{U}_{2,i})^T.$$

Asymptotyczny rozkład statystyki testowej

Określmy normę $||\cdot||_{\mathcal{T},\infty}$ daną przez

$$||f(t,z)||_{\mathcal{T},\infty} = \sup_{z\in[0,1]^d} \sup_{t\in Q_{X|Z}(\mathcal{T}|z)} |f(t,z)|.$$

W podobny sposób określmy normę $||\cdot||'_{\mathcal{T},\infty}$ daną przez

$$||f(t,z)||'_{\mathcal{T},\infty} = \sup_{z\in[0,1]^d} \sup_{t\in Q_{Y|Z}(\mathcal{T}|z)} |f(t,z)|.$$

Asymptotyczny rozkład statystyki testowej

Założenia

Dla każdego rozkładu $P \in \mathcal{P}_0$ *istnieją deterministyczne funkcje szybkości* g_P *i* h_P *zbiegające do zera przy* $n \to \infty$ *i funkcje* $\xi, \xi' : [0,1] \times [0,1]^d \to \mathbb{R}$, *takie że*

(i) $||F_{X|Z} - \hat{F}^{(n)}_{X|Z}||_{\mathcal{T},\infty} \in O_P(g_P(n))$ *i* $||F_{Y|Z} - \hat{F}^{(n)}_{Y|Z}||'_{\mathcal{T},\infty} \in O_P(h_P(n))$.

(ii) $||\xi - \hat{F}^{(n)}_{X|Z}||_{\mathcal{T}^c,\infty} \in O_P(g_P(n))$ *i* $||\xi' - \hat{F}^{(n)}_{Y|Z}||'_{\mathcal{T}^c,\infty} \in O_P(h_P(n))$.

Asymptotyczny rozkład statystyki testowej

Twierdzenie

Załóżmy, że powyższe założenia są spełnione z funkcjami szybkości g_P *i* h_P*, takimi że* $\sqrt{n} g_P(n) h_P(n) \to 0$ *przy* $n \to \infty$ *dla każdego* $P \in \mathcal{P}_0$*. Wtedy statystyka testowa* $\hat{\rho}_n$ *spełnia*

$$\sqrt{n}\hat{\rho}_n \xrightarrow[P]{d} N(0, \Sigma \otimes \Sigma)$$

dla każdego ustalonego $P \in \mathcal{H}_0$*. Asymptotyczna macierz kowariancji jest dana przez*

$$\Sigma_{k,l} = \int_0^1 \varphi_k(u)\varphi_l(u)du$$

dla $k, l = 1, \ldots, q$ *i nie zależy ona od* P*.*

Postać finalnej statystyki testowej

Znając asymptotyczny rozkład $\hat{\rho}_n$ można określić finalną statystykę testową o postaci

$$T_n := ||\Sigma^{-1/2}\hat{\rho}_n\Sigma^{-1/2}||_F^2,$$

gdzie $||\cdot||_F$ oznacza normę Frobeniusa.

Asymptotyczny rozkład finalnej statystyki testowej

Wniosek

Niech założenia z powyższego twierdzenia będą spełnione i niech T_n będzie jak powyżej. Wtedy spełnione jest

$$nT_n \xrightarrow[P]{d} \chi^2_{q^2}$$

dla każdego określonego $P \in \mathcal{H}_0$.

Przycięta korelacja Spearmana

Rozważmy poniższą funkcję

$$\varphi_k(u) = \varphi_l(u) = \sqrt{12}\left(u - \frac{1}{2}\right)$$

dla $u \in [0,1]$, której wynikiem jest ρ_{kl} opisujące warunkową wartość oczekiwaną korelacji Spearmana między X i Y przy danym Z. Na podobieństwo powyższego wzoru można utworzyć klasę funkcji $\varphi : [0,1] \to \mathbb{R}^q$ poprzez

$$\varphi_k(u) = c_k(u - m_k)\sigma_k(u),$$

takie że każda funkcja $\varphi_k : [0,1] \to \mathbb{R}$ jest określona przez funkcję $\sigma_k : [0,1] \to \mathbb{R}$ spełniającą warunek Lipschitza, z nośnikiem $\mathcal{T}_k$ dla σ_k, który jest zwartym odcinkiem w $(0,1)$, $\int_0^1 \sigma_k(u)du = 1$ oraz

$$m_k = \int u\sigma_k(u)du \text{ i } c_k = \left(\int (u - m_k)^2\sigma_k(u)^2 du\right)^{-1/2}.$$

Przycięta korelacja Spearmana

Punktem startowym do wybrania przyciętej funkcji σ jest znormalizowany indykator

$$u \to (\lambda - \mu)^{-1} \mathbb{1}_{[\mu,\lambda]}(u)$$

dla $u \in [0,1]$, gdzie $0 < \mu < \lambda < 1$ są parametrami przycięcia. Jednakże, powyższa funkcja nie spełnia warunku Lipschitza. Zatem rozważymy proste liniowe przybliżenie $\sigma : [0,1] \to \mathbb{R}$ dane przez

$$\sigma(u) = Kf(u) \text{ i } f(u) = \begin{cases} 1 & \text{dla } u \in [\mu + \delta, \lambda - \delta], \\ 0 & \text{dla } u \in [\mu, \lambda]^c, \\ \delta^{-1}(u - \mu) & \text{dla } u \in [\mu, \mu + \delta), \\ \delta^{-1}(\lambda - u) & \text{dla } u \in (\lambda - \delta, \lambda], \end{cases}$$

gdzie $K = (\lambda - \mu - \delta)^{-1}$. Tutaj $0 < \delta < (\lambda - \mu)/2$ jest wyznaczonym parametrem, który określa dokładność przybliżenia.

Praktyczne rozważania

W poprzednich podrozdziałach zakładaliśmy, że wszystkie zmienne losowe przyjmują wartości na odcinku $[0, 1]$. W przypadku, np. $X \in \mathbb{R}$ możemy nałożyć ściśle rosnące, ciągłe przekształcenie $t : \mathbb{R} \to [0, 1]$, aby otrzymać nową zmienną losową $X' = t(X)$ o wartościach $[0, 1]$. Petersen i Hansen (2021) polecają transformację danych poprzez empiryczną dystrybuantę, dla których nowe zmienne są nazywane w literaturze obserwacjami z pseudokopuły. Przekształcenie tworzy zależności podobne do tych utworzonych przez centrowanie czy skalowanie.

Praktyczne rozważania

Do estymacji warunkowych dystrybuant $\hat{F}^{(m,n)}_{X|Z}$ i $\hat{F}^{(m,n)}_{Y|Z}$ sugerowane jest wybranie $\tau_{\min} = 0.01$ i $\tau_{\max} = 0.99$, a następnie utworzenie równomiernego podziału $(\tau_k)_{k=1}^{m}$ w $\mathcal{T} = [\tau_{\min}, \tau_{\max}]$ z liczbą punktów podziału $m = \lceil \sqrt{n} \rceil$. Następnie sugerowane jest użycie modelu w postaci (2) dla obu modeli regresji kwantylowej $Q_{X|Z}(\tau_k|\cdot)$ i $Q_{Y|Z}(\tau_k|\cdot)$ dla każdego $k = 1, \ldots, m$, gdzie bazy h_1 i h_2 mogą być określone np. poprzez addytywną krzywą B-sklejaną dla każdej komponenty Z.

Praktyczne rozważania

Do testowania warunkowej niezależności zalecane jest użycie $\hat{\Psi}_n$ bazującego na nieparametrycznych residuach $(\hat{U}_{1,i}, \hat{U}_{2,i})_{i=1}^n$. Następnie wybieramy $q \geq 1$ i określamy $\tau_{\min} = \lambda_0 < \cdots < \lambda_q = \tau_{\max}$ jako równomierny podział w $\mathcal{T}$. Później definiujemy przyciętą funkcję σ_k z parametrami przycięcia λ_k i λ_{k+1} oraz parametrem aproksymacji $\delta = 0.01 \cdot (\lambda_{k+1} - \lambda_k)$ dla każdego $k = 0, \ldots, q-1$. Na sam koniec określamy $(\sigma_k)_{k=1}^q$, obliczamy statystykę testową $\hat{\rho}_n$ i obliczamy $\hat{\Psi}_n$ dla ustalonego poziomu istotności $\alpha \in (0, 1)$.

Praktyczne rozważania

Pozostają dwa nietrywialne wybory. Pierwszym z nich jest wybór baz h_1 i h_2 dla modelów regresji kwantylowej $Q_{X|Z}(\tau_k|z) = h_1(z)^T\beta_{\tau_k}$ i $Q_{Y|Z}(\tau_k|z) = h_2(z)^T\beta_{\tau_k}$. Drugi wybór, to określenie liczby wymiaru $q \geq 1$ dla uogólnionej korelacji. Petersen i Hansen (2021) sugerują, aby próbować z małymi wartościami, np. $q \in \{1, \ldots, 5\}$ i odrzucać hipotezę o warunkowej niezależności, gdy jeden z testów odrzuca hipotezę. Należy jednak być ostrożnym, ponieważ wtedy mamy do czynienia z wielokrotnym testowaniem.

Literatura

[1] J.J. Daudin. Partial association measures and an application to qualitative regression. Biometrika, 67, 581–590, 1980.

[2] R.A. Fisher. A mathematical examination of the methods of determining the accuracy of observation by the mean error, and by the mean square error. Monthly Notices of the Royal Astronomical Society, 80, 758–770, 1920.

[3] R.A. Fisher. Two new properties of mathematical likelihood. Proceedings of the Royal Society of London Series A, 144, 285–307, 1934.

[4] D. Koller and N. Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

[5] L. Petersen and N. R. Hansen. Testing conditional independence via quantile regression based partial copulas. Journal of Machine Learning Research, 22, 1–47, 2021.

[6] R.D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. The Annals of Statistics, 48, 1514–1538, 2020.

Dziękuję za uwagę.