EÖTVÖS LORÁND TUDOMÁNYEGYETEM
TERMÉSZETTUDOMÁNYI KAR
Matematika BSc, Matematikai elemző szakirány

A kopula regresszió

Készítette:
Puskás Teodóra

Témavezető:
Pröhle Tamás
Intézet és Tanszék neve:
Matematikai Intézet, Valószínűségelméleti és Statisztika Tanszék

Budapest, 2018.

Nyilatkozat

Név: Puskás Teodóra
ELTE Természettudományi Kar, szak: Matematika BSc
Szakdolgozat címe: Kopula regresszió

A szakdolgozat szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló munkám eredménye, saját szellemi termékem, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2018 május 31.

_____

Bevezetés

A dolgozatomban a kopula alapú regresszió módszerével foglalkozom.

A statisztikában gyakori probléma annak meghatározása, hogy egy vagy több független változó milyen hatással van a függő változóra, illetve ennek a kapcsolatnak a leírása. Előbbit korreláció-, utóbbit regressziószámításnak nevezzük. A függés szokványos modellje esetén egyetlen szám, a korreláció jellemzi a változópárok függését. Együttes normális eloszlás esetén a korrelációs kapcsolat erősségét a lineáris korrelációs együtthatóval mérjük. Azonban gyakran előfordul, hogy nem elég információnk az együttes eloszlásról. A kopula egy olyan technika, amelynek segitségével a változók eloszlása és függésének karakterisztikája egymástól függetlenül kezelhető.

Az első fejezet a kopulákkal kapcsolatos alapfogalmakat és tételeket foglalja össze, illetve ismerteti a legfontosabb kopula osztályokat és azok generátorfüggvényeit. A második fejezet felidézi a lineáris korreláció alapjait, majd felvázolja a változók közötti kapcsolat kopula alapú mérését. A harmadik fejezet bevezetést nyújt a kopula alapú regresszió alapjaiba. A negyedik és ötödik fejezet bemutatja az R programcsomagot, ezen belül pedig a Gauß Copula Marginal Regression csomagot, majd ennek alkalmazását különböző mintákon.

Tartalomjegyzék

1.

Kopulák

1.1. Alaptulajdonságok

A kopula modell napjaink egyik legnépszerűbb matematikai eszköze, mely segítségével két vagy több valószínűségi változó együttes eloszlásának elemzését végezhetjük el. Alkalmazása elterjedt a kvantitatív pénzügyek, biostatisztika, építészet és egészségügyi területeken. A kifejezés a latin copulare szóból ered, jelentése: összeköt, kapcsol. Matematika vonatkozásban először Abe Sklar használta 1959-ben, mint eszköz, mely segítségével együttes eloszlásokat modellezhetünk adott peremeloszlásokra.

Legyen X és Y folytonos valószínűségi változó eloszlásfüggvénye rendre $F(x) = P(X \leq x)$ és $G(y) = P(Y \leq y)$ és $H(x, y) = P(X \leq x, Y \leq y)$ az együttes eloszlásfüggvényük. $\forall (x, y) \in [-\infty, \infty]^2$-re vegyük azt az I^3-beli ($I = [0, 1]$) pontot, amelynek koordinátái $(F(x), G(y), H(x, y))$. Ezt a $I^2 \to$ I leképzést nevezzük kopulának.

1. Definíció (kopula)**.** *Kopulán egy d-dimenziós, egyenletes eloszlású peremértékekkel rendelkező valószínűségi vektor eloszlásfüggvényét értjük.*
A $C : [0,1]^d \to [0,1]$ *kopulafüggvényre az alábbi tulajdonságok teljesülnek:*

- $C(u_1, u_2, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0 \quad \forall u_1, u_2, \ldots, u_d \in [0,1]$
- $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i \quad \forall u_i \in [0,1]$, *ahol* $i = 1, \ldots, d$
- $\forall (a_1, \ldots, a_d), (b_1, \ldots, b_d) \in [0,1]^d$ *és* $a_i \leq b_i$ *esetén*

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\ldots+i_d} C(u_{1i_1}, \ldots, u_{di_d}) \geq 0$$

ahol $u_{j1} = a_j$ *és* $u_{j2} = b_j$, $\forall j \in \{1, \ldots, d\}$

Az első két tulajdonság miatt egyenletesek a peremeloszlások. Míg a harmadik tulajdonság biztosítja, hogy bármely $(U_1, \ldots, U_d)$ vektorra a $P(U_1 \in [a_1, b_1], \ldots, U_d \in [a_d, b_d])$ valószínűség nem negatív.

Bármely függvény, amelyre igazak az előző feltételek kopulának minősül. Továbbá, ha C egy d-dimenziós kopula, akkor $C'(1, u_1, \ldots, u_{d-1})$ is kopula. Azaz C minden k-dimenziójú pereme kopula, ahol $2 \leq k < d$.

2. Definíció (2-dimenziós kopula)**.** *A fenti definíció speciális esete a 2-dimenziós kopula, amelyre teljesül, hogy:*

- $C(0, y) = 0$ *és* $C(x, 0) = 0$, $\forall x, y \in [0,1]$
- $C(1, y) = y$ *és* $C(x, 1) = x$, $\forall x, y \in [0,1]$
- C *2-növő, azaz* $\forall (a_1, a_2), (b_1, b_2) \in [0,1]^2$ *és* $a_i \leq b_i$ *esetén*

$$\begin{array}{rcl}
\sum_{i_1=1}^{2} \cdots \sum_{i_2=1}^{2} (-1)^{i_1+i_2} C(u_{1i_1}, u_{2i_2}) & \geq & 0 \\
\Updownarrow & & \\
\sum_{i_1=1}^{2} \left[(-1)^{i_1+1} C(u_{1i_1}, u_{21}) + (-1)^{i_1+2} C(u_{1i_1}, u_{22}) \right] & \geq & 0 \\
\Updownarrow & & \\
(-1)^2 C(u_{11},u_{21}) + (-1)^3 C(u_{11},u_{22}) + (-1)^3 C(u_{12},u_{21}) + (-1)^4 C(u_{12},u_{22}) & \geq & 0 \\
\Updownarrow & & \\
C(u_{11}, u_{21}) - C(u_{11}, u_{22}) - C(u_{12}, u_{21}) + C(u_{12}, u_{22}) & \geq & 0
\end{array}$$

Mivel $u_{j1} = a_j$ *és* $u_{j2} = b_j$, $\forall j \in 1, 2$, *ezért:*

$$C(a_1, a_2) - C(a_1, b_2) - C(b_1, a_2) + C(b_1, b_2) \geq 0$$

1.2. Sklar-tétel

A kopulák alaptételét Abe Sklar adta meg[1], melynek lényege, hogy minden többváltozós eloszlásfüggvény kifejezhető a marginálisai függvényében egy kopula segítségével.

1. Tétel (Sklar)**.** *Legyen* F *egy* d*-dimenziós eloszlásfüggvény* $F_1, \ldots, F_d$ *peremeloszlásokkal. Ekkor létezik egy* $C : [0,1]^d \to [0,1]$ *kopula, amelyre* $\forall x \in \mathbb{R}^d$ *esetén igaz, hogy:*

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$$

és ha F_i *folytonos* $\forall i \in \{1, \ldots, d\}$*-re, akkor* C *egyértelmű.*
Megfordítva, ha C *egy* d*-dimenziós kopula és* $F_1, \ldots, F_d$ *egyváltozós eloszlásfüggvények, akkor* F *egy* d*-dimenziós eloszlásfüggvény az* $F_1, \ldots, F_d$ *marginálisokkal.*

A bizonyításhoz szükség van eloszlás transzformáció fogalmára és a hozzá kapcsolódó tételre, mely szerint ha beírunk egy adott valószínűségi változót a saját eloszlásfüggvényébe, akkor egy $[0,1]$-en egyenletes eloszlású változót kapunk.

3. Definíció (Eloszlás transformált)**.** *Legyen* Y *valószínűségi változó eloszlásfüggvénye* F *és* V $(0,1)$*-en egyenletes eloszlású,* Y*-tól független valószínűségi változó, ekkor U:=F(Y,V) módosított eloszlásfüggvényt, ahol*

$$F(x, \lambda) := P(X < x) + \lambda P(Y = x)$$

az Y *eloszlás transzformáltjának nevezzük.*

Ha F folytonos, akkor $F(x, \lambda) = F(x)$ $\forall\lambda$-ra és $U = F(Y) \stackrel{d}{=} U(0,1)$, ahol $\stackrel{d}{=}$ eloszlásbeli egyenlőséget jelöl.

2. Tétel (Eloszlás transzformáció)**.** *Legyen* $U = F(Y, V)$ Y *valószínűségi változó eloszlás transzformáltja. Ekkor* $U \stackrel{d}{=} U(0,1)$*, azaz* U *egyenletes eloszlású a* $(0,1)$ *intervallumon és* $X = F^{-1}(U)$ *majdnem mindig.*

Térjünk vissza Sklar tételének bizonyítására.

Bizonyítás. Legyen $X = (X_1, \ldots, X_n)$ egy tetszőleges vektor a (Ω, A, P) valószínűségi mezőn és F az eloszlásfüggvénye. Legyen V $(0,1)$-en egyenletes eloszlású, X-től független valószínűségi változó. Tekintsük az $U_i := F_i(X_i, V)$ eloszlástranszformációt, $1 \leq i \leq n$. Ekkor 2 miatt $U_i \stackrel{d}{=} U(0,1)$ és $X_i = F_i^{-1}(U_i)$ majdnem mindig, $\forall 1 \leq i \leq n$. Így az $U = (U_1, \ldots, U_n)$ eloszlásfüggvényeként definiált C segítségével felírható, hogy:

$$F(x) = P(X \leq x) = P(F_i^{-1}(U_i) \leq x_i, 1 \leq i \leq n)$$

$$= P(U_i \leq F_i(x_i), 1 \leq i \leq n) = C(F_1(x_1), \ldots, F_n(x_n)),$$

vagyis F-nek kopulája C. □

A tételből látható, hogy minden folytonos, többváltozós eloszlásfüggvény esetén külön tudjuk választani a peremeloszlásokat a függőségi struktúrától, és ez a stuktúra egy kopulával reprezentálható.

Legyen F egy d-dimenziós eloszlásfüggvény folytonos $F_1, \ldots, F_d$ peremeloszlásokkal. Ekkor egyértelműen létezik egy C d-dimenziós kopula, amely $\forall x = (x_1, \ldots, x_d) \in \mathbb{R}^d$-re:

$$F(x) = C(F_1(x_1), \ldots, F_d(x_d)).$$

C kopulát a következő formulával határozzuk meg $\forall u \in [0,1]^d$:

$$C(u) = F(F_1^{[-1]}(u_1), \ldots, F_d^{[-1]}(u_d)),$$

ahol $F_i^{[-1]}(t) := \inf\{x : F_i(x) \geq t\}$.

3. Tétel. *Ha X és Y folytonos valószínűségi változó és $C : [0,1]^2 \to [0,1]$ az együttes eloszlásuknak megfelelő kopulafüggvény, akkor:*

- *C(u,v) növekvő mindegyik változójában*
- *Ha α és β két szigorúan növekvő transzformációja X-nek illetve Y-nak, akkor az X és Y-hoz tartozó kopula egyenlő az $\alpha(X)$ és $\beta(Y)$-hoz hozzárendelt kopulával:*

$$C_{X,Y} = C_{\alpha(X),\beta(Y)}$$

- *Minden kopulára érvényes a Lipschitz-féle egyenlőtlenség:*

$$\forall (u_1, u_2), (v_1, v_2) \in [0,1]^2 \ : \qquad |C(u_1, u_2) - C(v_1, v_2)| \leq |u_1 - u_2| + |v_1 - v_2|$$

1.3. Fréchet határok

A kopulák minden esetben alulról és felülről is korlátosak. Definiáljuk M^n, Π^n és W^n függvényeket a $[0,1]^n$-en a következőképpen:

$$\begin{aligned} M^n(u) &= \min(u_1, \ldots, u_n), \\ \Pi^n(u) &= u_1 \ldots u_n, \\ W^n(u) &= \max(u_1 + \ldots + u_n - n + 1, 0). \end{aligned}$$

M^n és Π^n n-kopula $\forall n \geq 2$-re, míg W^n nem kopula $n \geq 3$ esetén, ahogy az alábbi példából látható.

1. Példa. *Vegyük az $[1/2, 1]^n \subset [0,1]^n$ n-kockát.*

$$\begin{aligned} V_{W^n}([1/2,1]^n) &= \max(1 + \ldots + 1 - n + 1, 0) \\ &\quad - n \max(1/2 + 1 + \ldots + 1 - n + 1, 0) \\ &\quad + \tbinom{n}{2} \max(1/2 + 1/2 + 1 + \ldots + 1 - n + 1, 1) \\ &\quad \ldots \\ &\quad + \max(1/2 + \ldots + 1/2 - n + 1) \\ &= 1 - n/2 + 0 \ldots + 0. \end{aligned}$$

Ennélfogva W^n nem kopula, ha $n \geq 3$.

4. Tétel (Fréchet-Hoeffding határok)**.** *Legyen C egy n-dimenziós kopula. Minden $u \in [0,1]^n$-ra fennáll, hogy:*

$$W^n(u) \leq C(u) \leq M^n(u).$$

Azaz, minden kopula minden pontban nagyobb mint a W függvény és kisebb mint az M kopula.

Ezeket a határokat hívjuk Fréchet-határoknak.

A tétel könnyen belátható 2-dimenziós kopula esetében,
azaz ha $C : [0,1]^2 \to [0,1]$.

Legyen $(u,v) \in [0,1]^2$
A kopula definíciójából következik, hogy:

$$\left.\begin{array}{l} C(u,v) \leq C(u,1) = u \\ C(u,v) \leq C(1,v) = v \end{array}\right\} \Rightarrow C(u,v) \leq \min(u,v) = M(u,v)$$

X Legyen $(u,v) \in [0,1]^2$, ekkor mert
$V_C([u,1] \times [v,1]) = C(1,1) - C(1,v) - C(u,1) + C(u,v) \geq 0$
és $C(1,1) = 1$, $C(1,v) = v$, $C(u,1) = u$ érvényes, hogy $1 - v - u + C(u,v) \geq 0$:

$$\left.\begin{array}{ll} \text{Az előbbi szerint} & C(u,v) \geq u + v - 1 \\ \text{továbbá, mert} & C(u,v) \geq 0 \end{array}\right\} \Rightarrow C(u,v) \geq \max(u + v - 1, 0)$$

Ha n darab valószínűségi változó együttes eloszlásfüggvénye C, akkor $\overline{C}$-vel jelöljük a változók együttes túlélési függvényét, azaz ha $(U_1, \ldots, U_n)^T$ eloszlásfüggvénye C, akkor $\overline{C}(u_1, \ldots, u_n) = P\{U_1 > u_1, \ldots, U_n > u_n\}$.

4. Definíció. *Legyen C_1 és C_2 kopula. Ekkor C_1 kisebb, mint C_2 ($C_1 \prec C_2$), ha:*

$$C_1(u) \leq C_2(u) \qquad \text{és} \qquad \overline{C}_1(v) \leq \overline{C}_2(u)$$

fenáll minden $u \in [0,1]^n$-ra.

1.4. Kopulák osztályozása

Röviden bemutatom az ismertebb kopula osztályokat. A kopulákat két főbb osztályba soroljuk, az archimédeszi kopulák és az elliptikus kopulák osztályába.

1.4.1. Elliptikus kopulák

Az elliptikus kopulák osztályába tartozik a Gauß kopula és a t-kopula. Előbbit többváltozós normális eloszlás, míg utóbbit Student t-eloszlás esetén alkalmazzuk.

5. Definíció (Elliptikus eloszlás)**.** *Ha X egy n-dimenziós valószínűségi vektor, $\mu \in \mathbb{R}^n$ és $\sum$ egy $n \times n$-es, nemnegatív definit, szimmetrikus mátrix, amelyekre fenáll, hogy $X - \mu$ karakterisztikus függvénye $\varphi X_{-\mu}(t)$ függvénye a $t^T \sum t$ kvadratikus alaknak, azaz $\varphi X_{-\mu}(t) = \phi(t^T \sum t)$, akkor X-nek elliptikus eloszlása van a μ, $\sum$, ϕ paraméterekkel. Jelölése: $X \sim E_n(\mu, \sum, \phi)$*

$n = 1$ esetén az elliptikus eloszlás megegyezik az egydimenziós szimmetrikus eloszlással. A ϕ függvényt karakterisztikus generátornak nevezzük.

6. Definíció (Gauß kopula)**.** *Az n-változós normális eloszlás kopuláját R lineáris korrelációs mátrix esetén a következőképpen definiáljuk:*

$$C_R^{Ga}(u) = \phi_R^n(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_n)),$$

ahol ϕ_R^n jelöli az n-változós normális eloszlásfüggvénynek együttes eloszlásfüggvényét, R korrelációs mátrixszal és ϕ^{-1} jelöli a sztenderd normális eloszlásfüggvény inverzét.

Kétváltozós esetben a kopula felírható a következő alakban is:

$$C_R^{Ga}(u, v) = \int_{-\infty}^{\phi^{-1}(u)} \int_{-\infty}^{\phi^{-1}(v)} \frac{1}{2\pi(1 - R_{12}^2)^{1/2}} \exp\left\{ -\frac{s^2 - 2R_{12}st + t^2}{2(1 - R_{12}^2)} \right\} \mathrm{d}s\mathrm{d}t.$$

Most nézzük hogyan generálhatjuk le a véletlen változót adott C_R^{Ga} Gauß kopulából. Tekintsük kizárólag a szigorúan pozitív definit R mátrixokat. Legyen $R = AA^T$, ahol A $n \times n$-es mátrix, és ha $Z_1, \ldots, Z_n \sim \mathcal{N}(0, 1)$ függetlenek, akkor fennáll, hogy

$$\mu + AZ \sim \mathcal{N}_n(\mu, R).$$

Az A mátrix legyen az R Cholesky felbontása, azaz R-t egy alsó trianguláris mátrix és annak konjugált transzponáltjának szorzatává bontjuk fel. Ezzel könnyen előállíthatjuk a véletlen változót a Gauß kopulából.

Tehát az algoritmus:

- Állítsuk elő R Cholesky felbontását, A-t
- Legyen a $z = (z_1, \ldots, z_n)$ egy n dimenziós, független koordinátájú, normális eloszlású változó
- Legyen $x = Az$
- Legyen $u_i = \phi(x_i)$, $i = 1, \ldots, n$
- Ekkor $(u_1, \ldots, u_n)^T \sim C_R^{Ga}$.

7. Definíció (Student t-kopula)**.** *Az n-változós Student t-eloszlás kopuláját R szimmetrikus pozitív definit mátrix esetén a következőképpen definiáljuk:*

$$C_{\nu,R}^t(u) = t_{\nu,R}^n(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_n)),$$

ahol $t_{\nu,R}^n$ az n-dimenziós, R korrelációs mátrixú és ν szabadságfokú Student-féle t-eloszlás eloszlásfüggvénye, és t_ν^{-1} jelöli az egy dimenziós ν szabadságfokú Student-féle t-eloszlás eloszlásfüggvényének az inverzét.

Kétváltozós esetben a kopula felírható a következő alakban is:

$$C_{\nu,R}^t(u,v) = \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \frac{1}{2\pi(1-R_{12}^2)^{1/2}} \left\{1 + \frac{s^2 - 2R_{12}st + t^2}{\nu(1-R_{12}^2)}\right\}^{-(\nu+2)/2} \mathrm{d}s\mathrm{d}t.$$

Látható, hogy az elliptikus kopulák az elliptikus eloszlású valószínűségi vektorok komponensenkénti transzformáltja. Elliptikus eloszlásokat könnyen szimulálhatunk kopulákkal és vica versa. Ugyanakkor az elliptikus kopuláknak nincs zárt alakja és radiális szimmetriára korlátozottak.

1.4.2. Archimédeszi kopulák

Észszerűnek tűnik, hogy sok pénzügyi és biztosítási alkalmazásban erősebb összefüggés van a nagy veszteségek (pl. részvénypiaci törés/bukás) között, mint a nagy nyereségek között. Ilyen aszimetrikus eseteket nem tudunk elliptikus kopulákkal modellezni. Ebben a fejezetben ismertetem az archimédeszi kopulákat, melyek lehetőséget nyújtanak számos különböző függőségi struktúra modellezésére és ellentétben az elliptikus kopulákkal létezik zárt alakjuk.

8. Definíció (Pszeudoinverz)**.** *Legyen $\varphi : [0,1] \to [0,\infty]$ egy szigorúan monoton csökkenő függvény, melyre teljesül, hogy $\varphi(1) = 0$.*
Ekkor φ pszeudoinverze $\varphi^{[-1]} : [0,\infty] \to [0,1]$ a következő:

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t)\ , & 0 \le t \le \varphi(0), \\ 0\ , & \varphi(0) \le t \le \infty\ . \end{cases}$$

$\varphi^{[-1]}$ monoton csökken $[0,\infty]$-on és szigorúan csökken $[0,\varphi(0)]$-án. Továbbá $\varphi^{[-1]}(\varphi(u)) = u$ a $[0,1]$ intervallumon és fennál, hogy:

$$\varphi(\varphi^{[-1]}(t)) = \begin{cases} t\ , & 0 \le t \le \varphi(0), \\ \varphi(0)\ , & \varphi(0) \le t \le \infty\ . \end{cases}$$

Illetve ha $\varphi(0) = \infty$, akkor $\varphi^{[-1]} = \varphi^{-1}$.

9. Definíció (Archimédeszi kopula)**.** *Legyen $\varphi : [0,1] \to [0,\infty]$ egy szigorúan monoton csökkenő függvény, melyre teljesül, hogy $\varphi(1) = 0$.*
Legyen $C : [0,1]^2 \to [0,1]$:

$$C(u,v) = \varphi^{[-1]}(\varphi(u) + \varphi(v))$$

C függvényt nevezzük archimédeszi kopulának.

A φ függvényt a kopula generátorfüggvényének nevezzük. Ha $\varphi(0) = \infty$, akkor φ szigorú generátorfüggvény. Illetve abban az esetben, ha $\varphi^{[-1]} = \varphi^{-1}$, akkor a

$$C(u,v) = \varphi^{-1}(\varphi(u) + \varphi(v))$$

a szigorú archimédeszi kopula.

5. Tétel. *$C(u,v) = \varphi^{[-1]}(\varphi(u)+\varphi(v))$ akkor és csak akkor kopula, ha φ konvex.*

Legyen $\varphi(t) = (-\ln t)^\theta$, ahol $\theta \geq 1$. Ekkor φ folytonos és $\varphi(1) = 0$.
$\varphi' = -\theta(-\ln t)^{\theta-1}\frac{1}{t}$, tehát $\varphi : [0,1] \to [0,\infty]$ szigorúan monoton csökken. $\varphi''(t) \geq 0$ a $[0,1]$ intervallumon, tehát φ konvex. Ezenfelül $\varphi(0) = \infty$, vagyis φ szigorú generátorfüggvény.

A φ generátorfüggvénnyel előállított C kopula:

$$C_\theta(u,v) = \varphi^{-1}(\varphi(u)+\varphi(v)) = \exp(-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta})$$

a Gumbel-Hougaard kopula.

Legyen $\varphi(t) = (t^{-\theta}-1)/\theta$, ahol $\theta \in [-1,\infty) \setminus \{0\}$. A φ adja a Clayton kopulát:

$$C_\theta(u,v) = \max([u^{-\theta}+v^{-\theta}-1]^{-1/\theta}, 0)$$

Ha $\theta > 0$, akkor a kopula szigorú és leegyszerűsíthető a következő alakra:

$$C_\theta(u,v) = (u^{-\theta}+v^{-\theta}-1)^{-1/\theta}.$$

Legyen $\varphi(t) = -\ln\frac{e^{-\theta t}-1}{e^{-\theta}-1}$, ahol $\theta \in \mathbb{R} \setminus \{0\}$. Az így kapott kapott kopulát:

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$$

Frank kopulának nevezzük.

A Frank kopulák szigorú archimédeszi kopulák. Az egyetlen olyan archimédeszi kopulacsalád, amelyre teljesül a radiális szimmetria.

Legyen $\varphi(t) = 1-t$, $t \in [0,1]$. Ekkor $\varphi^{[-1]}(t) = 1-t$, ha $t \in [0,1]$ és 0, ha $t > 1$, azaz $\varphi^{[-1]}(t) = \max(1-t,0)$. Ekkor $C(u,v) = \max(u+v-1,0) =: W$, tehát a kétváltozós Fréchet által definiált alsó határ is archimédeszi kopula.

A következő tétel lehetővé teszi az archimédeszi kopulák többváltozóssá való kiterjesztését.

6. Tétel. *Legyen C archimédeszi kopula és φ generátorfüggvény. Ekkor:*

- *C szimmetrikus, azaz $C(u,v) = C(v,u)\ \forall u,v \in [0,1]$*
- *C asszociatív, azaz $C(C(u,v),w) = C(u,C(v,w))\ \forall u,v,w \in [0,1]$*

Bizonyítás. A szimmetrikusság az archimédeszi kopula definíciójából következik. Asszociativitás:

$$\begin{aligned} C(C(u,v),w) &= \varphi^{[-1]}(\varphi(\varphi^{[-1]}(\varphi(u)+\varphi(v)))+\varphi(w)) \\ &= \varphi^{[-1]}(\varphi(u)+\varphi(v)+\varphi(w)) \\ &= \varphi^{[-1]}(\varphi(u)+\varphi(\varphi^{[-1]}(\varphi(v)+\varphi(w)))) = C(u,C(v,w)) \end{aligned}$$

□

A fejezet megírásához felhasználtam a [2]-t és az [3], [4] és a [5] cikkeket.

2.

Függési strukturák

2.1. Lineáris korreláció

10. Definíció. *Legyen $(X,Y)^T$ valószínűségi vektorváltozó, ahol a változók szórásnégyzete véges és nem nulla. Ekkor (X,Y) lineáris korrelációs együtthatója:*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}},$$

ahol $Cov(X,Y) = E(X,Y) - E(X)E(Y)$ a vektorváltozó kovarianciája, illetve $Var(X)$ és $Var(Y)$ az X és Y szórásnégyzete.

A lineáris kapcsolat nagyságának, erősségének mérésére a korreláció szolgál. Ha létezik $a \in \mathbb{R} \setminus \{0\}, b \in \mathbb{R}$, hogy $Y = aX + b$, akkor a változók tökéletesen egyenes/fordított arányosak és $|\rho(X,Y)| = 1$. Még fontosabb, hogy a korreláció kölcsönös kapcsolatot jelent. Minden más esetben $-1 < \rho(X,Y) < 1$.

Ha a változók normális eloszlásúak, akkor csak lineáris kapcsolat képzelhető el, vagyis ha nincs közöttük lineáris kapcsolat, akkor függetlenek egymástól, azaz $\rho(X,Y) = 0$.

Probléma viszont, hogy lineáris korrelációból csak a változók közti függőség fokát tudjuk megbecsülni, nem kapunk képet a függőségi struktúráról. Az együtthatóból nem következtethetünk a változók együttes eloszlására.

2.2. Konkordáns korreláció

Legyen $(x, y)^T$ és $(x', y')^T$ két valószínűségi érték a $(X, Y)^T$ vektorváltozóból. $(x, y)^T$ és $(x', y')^T$ konkordáns, ha $(x - x')(y - y') > 0$, illetve diszkordáns, ha $(x - x')(y - y') < 0$.

7. Tétel. *Legyen $(X, Y)^T$ és $(X', Y')^T$ független valószínűségi vektorváltozó, melyek együttes eloszlásfüggvénye rendre H és H'. Legyen F az X és X', illetve G az Y és Y' marginálisa. Jelölje C és C' az $(X, Y)^T$ és $(X', Y')^T$ kopuláit, így $H(x, y) = C(F(x), G(y))$ és $H'(x, y) = C'(F(x), G(y))$. Legyen Q a $(X, Y)^T$ és $(X', Y')^T$ konkordancia és a diszkordancia valószínűségének a különbsége, azaz:*

$$Q = \{P(X - X')(Y - Y') > 0\} - Q - \{P(X - X')(Y - Y') < 0\}$$

Ekkor

$$Q = Q(C, C') = \iint_{[0,1]^2} C'(u, v) \, \mathrm{d}C(u, v) - 1.$$

A tétel teljes bizonyítása a [4]-ben található.

2.3. Kendall-féle tau és Spearman-féle ró

Nem elliptikus eloszlás esetében a lineáris korrelációs együttható gyakran félrevezető. Ilyenkor a Kendall-féle tau vagy a Spearman-féle ró alkalmazása az egyik legjobb alternatíva.

11. Definíció (Kendall-féle korrelációs együttható)**.** *Az $(X, Y)^T$ valószínűségi vektorváltozó Kendall-féle korrelációs együtthatója:*

$$\tau(X, Y) = P\{(X - X')(Y - Y') > 0\} - P\{(X - X')(Y - Y') < 0\},$$

ahol $(X', Y')^T$ $(X, Y)^T$ egy független példánya.

Tehát a Kendall-féle τ a konkordancia valószínűségének és a diszkordancia valószínűségének a különbsége.

8. Tétel. *Legyen $(X,Y)^T$ folytonos valószínűségi vektorváltozó és legyen C egy kopulája. Ekkor $(X,Y)^T$-ra a Kendall-féle tau:*

$$\tau(X,Y) = Q(C,C) = 4\iint_{[0,1]^2} C(u,v)\ \mathrm{d}C(u,v) - 1$$

A tétel teljes bizonyítása a [4]-ben található.

A fenti integrál megegyezik $C(U,V)$ várható értékével, ahol $U,V \sim \mathfrak{U}(0,1)$ és C az együttes eloszlásfüggvény, azaz $\tau(X,Y) = 4E(C(U,V)) - 1$.

12. Definíció (Spearman-féle korrelációs együttható)**.** *Az $(X,Y)^T$ valószínűségi vektorváltozó Spearman-féle korrelációs együtthatója:*

$$\rho_S(X,Y) = 3(P\{(X-X_1)(Y-Y_2) > 0\} - P\{(X-X_1)(Y-Y_2) < 0\}),$$

ahol $(X,Y)^T$, $(X_1,Y_1)^T$ és $(X_2,Y_2)^T$ függetlenek.

9. Tétel. *Legyen $(X,Y)^T$ folytonos valószínűségi vektorváltozó és legyen C egy kopulája. Ekkor $(X,Y)^T$-ra a Spearman-féle ró:*

$$\rho_S(X,Y) = 3\,Q(C,\Pi) = 12\iint_{[0,1]^2} uv\ \mathrm{d}C(u,v) - 3 = 12\iint_{[0,1]^2} C(u,v)\ \mathrm{d}u\mathrm{d}v - 3$$

A tétel teljes bizonyítása a [4]-ben található.

Tehát ha $X \sim F$ és $Y \sim G$, illetve $U = F(X)$ és $V = G(Y)$, akkor

$$\begin{aligned}
\rho_S(X,Y) &= 12\iint_{[0,1]^2} uv\ \mathrm{d}C(u,v) - 3 = 12\,E(U,V) - 3 \\
&= \frac{E(U,V) - 1/4}{1/12} = \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} \\
&= \rho(F(X),G(Y)).
\end{aligned}$$

3.

Regresszió kopula alapon

A leírás alapjai az [6] és a [7] cikkek.

Tegyük fel, hogy rendelkezésre áll az n darab $p+1$ dimenziós $(Y_i, X_{i,1}, ..., X_{i,p})$ változó $(y_i, x_{i,1}, ..., x_{i,p})$, $i = 1, ..., n$ megfigyelt értéke. Feltesszük, hogy minden i-re az Y_i célváltozó és az X_i magyarázóváltozók közt a

$$Y_i = g(X_i, \delta_i, \lambda)$$

regressziós kapcsolat érvényes.

Feltehető, hogy egy standard normális eloszlású ε_i mellett

$$\delta_i = \Phi(\varepsilon_i) \quad \text{és} \quad g(.) \Leftrightarrow F^{-1}_{x_i,\lambda}(.) \qquad \text{vagyis, hogy} \quad Y_i = F^{-1}_{x_i,\lambda}(\Phi(\varepsilon_i))$$

ahol $F_{x_i,\lambda}(.)$ az Y_i eloszlása $X_i = x_i$ esetén.
Tehát az Y_i értékét nem mint

„az X_i egy nem feltétlen lineáris függvénye, plusz egy véletlen hiba”

határozzuk meg, hanem úgy mint

„az Y_i változó X_i értékétől függő eloszlásának

a $\mathcal{U}([0,1])$ eloszlású δ_i szerinti véletlen kvantilise”.

Ugyanis a $\mathcal{N}(0,1)$ standard normális eloszlás Φ eloszlásfüggvénye szerint számolt $\delta_i = \Phi(\varepsilon_i) \sim \mathcal{U}([0,1])$.

A vázolt modell szerint azt tettük fel, hogy az $Y_i|x_i$ eloszlása egy esetleg többdimenziós λ paramétertől függő $F_{x_i,\lambda}(.)$ eloszlás.

A bevezetett modellt tipikusan olyan esetekben szándékozunk alkalmazni, amikor az Y_i változók nem függetlenek.
Ilyen adódhat például ha az adatok „egyed”, „időpont” vagy „hely” referáltak.
Azaz, ha a rendelkezésre álló adatok olyanok, hogy

- egy-egy egyedre vonatkozóan több mérés áll rendelkezésre, vagy
- az adatok egy idősort alkotnak, vagy
- az adatok például egy-egy térbeli ponthoz kötöttek.

Ezekben az esetekben ugyanis a mérések még közelítésként sem tekinthetők függetleneknek. Sőt a mért adatok kovariancia mátrixa remélhetően, az adott kísérleti (mérési) elrendezésnek megfelelően egy jól meghatározott struktúrát mutat.

- Az egyed referált adatok esetén, azaz ha ismételt méréseink vannak, akkor az egy egyedre vonatkozó méréseknek feltehetően van egy, az egyedre jellemző véletlen komponense.

- Az időpont referált adatok esetén feltehető például, hogy az egymásutáni adatok egy ARMA struktúrát követnek.

- A hely referált adatok esetén pedig, ha feltesszük a vizsgált mező stacionárius vagy differencia stacionárius, esetleg izotróp jellegét, akkor a mérések kovariancia mátrixa jól leírhatóvá válik például a mező semivariogram függvényével.

Vagyis a felsorolt 3 példa esetén – és általában is – szükség van arra, hogy a fenti modell szerinti n dimenziós, koordinátáiban standard normális ε vektort ne független koordinátájúnak, hanem egy adott strukturájú kovariancia mátrix szerint függőnek tekintsük.

Ha feltételezük, hogy az ε egy általános n dimenziós normális eloszlást követ, akkor a $\delta \equiv \delta_S$ eloszlása egy S kovariancia mátrix szerinti Gauß-kopula eloszlás lesz. Vagyis a modellünk szerint

$$Y = F_{x,\lambda}^{-1}(\delta_S) \text{ ahol a } \delta_S \sim \text{Gauß-kopula}(S) \ .$$

Ezzel a modellel az Y célváltozó modellezését kétfele bontottuk:

- egyfelől meg kell találni a célváltozó λ-val meghatározott koordinátánkénti modelljét a specifikált modellosztályon belül

- másfelől meg kell találni azt az S kovariancia mátrixot, amely alapján a modell jól közelíti célváltozó értékek közti függést.

A célváltozó értékek koordinátánkénti modellje tipikusan egy általánosított lineáris regresszió modell, binomiális, negatív binomiális, Poisson, gamma stb. célváltozó eloszlást feltételezve.

A kovariancia mátrix típusa tipikusan klaszterezett, ARMA vagy Matern-féle.

Ha a kovariancia mátrix klaszterezett, akkor az az ismételt mérés esetének felel meg. Amikoris egy-egy egyedre vonatkozóan több mérés áll rendelkezésre.

Ha a kovariancia mátrix ARMA típusú, akkor azt feltételezzük, hogy az Y mérések adott p és q fokszám paraméterrel egy `ARMA(p,q)` modell szerinti idősort alkotnak. Az idősor jelleg mint feltételezés azt jelenti, hogy az adatok valamely szempont szerint, (ami nem feltétlen egyezik meg a mérési időponttal) rendezett sorozatot alkotnak. `ARMA(p,q)` típusú modell esetén, ha a mérések sorszáma t, $t = 1, 2, 3, ...$ akkor az adott (p, q)-ra és a becsléssel megállapítható $a_0 = 1, a_1, ..., a_p$ és $b_0 = 1, b_1, ..., b_q$ együtthatókra a célváltozó $y_1, ..., y_t, ..., y_n$ mért értékeire:

$$y_t + a_1 y_{t-1} + ... + a_p y_{t-p} = e_0 + b_1 e_{t-1} + ... + b_q e_{t-q}$$

valamely $e_0, e_1, ..., e_t, ...$ számsorra amely egy független azonos eloszlású 0 várhatóértékű, ismeretlen σ szórású ε_t, $t = 1, ..., n$ hibasorozat megfigyelt értékeinek tekinthető.

A Matern-féle kovariancia mátrixot akkor alkalmazhatjuk, ha a rendelkezésre álló adatok egy stacionárius mező nem feltétlen szabályos rács szerinti helyeken mért értékei. Tehát, ha minden $k = 1, ..., n$ indexű méréshez tartozik egy $u_k \in \mathbb{R}^p$ hely, vagy valamilyen más módon minden $i.$ és $j.$ méréspár $(i, j \in 1, ..., n)$ közt megállapítható egy $|u_i - u_j|$ távolság.

Matérn típusú kovariancia esetén a mérések $n \times n$ méretű S kovariancia mátrixa olyan, hogy a kovariancia $s_{i,j}$ eleme a megfelelő méréspár $h = |u_i - u_j|$ távolsága függvényében a következő semivariogram szerinti:

$$\gamma(h) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{h}{\varphi}\right)^{\kappa} K_{\kappa} \left(\frac{h}{\varphi}\right)$$

Ahol Γ az $n!$ kiterjesztését jelentő gamma függvény a K_κ a módosított másodfajú (más elnevezési szokás szerint III. típusú) κ rangú Bessel függvény. A h a szeparációs, az φ pedig a lecsengési távolság. A Matérn variogram fontos általánosítása több egyszerűbb variogram függvénynek.

Például

ha $\kappa = 1/2$, akkor $\gamma(h) = \exp(-h/\varphi)$ az exponenciális modell

ha $\kappa = 3/2$, akkor $\gamma(h) = (h/\varphi + 1)\exp(-h/\varphi)$

ha $\kappa = 5/2$, akkor $\gamma(h) = ((h/\varphi)^2 + 3(h/\varphi) + 3)\exp(-h/\varphi)$

továbbá $\kappa = 1$-re a Whittle-féle, $\kappa \to \infty$-re a Gauß-féle variogrammot adja.

A $\gamma(h)$ semivariogram alapján a h távolságú mérések $C(h)$ kovarianciáját a

$$C(h) = C(0) - \gamma(h)$$

képlet alapján kaphatjuk meg, ahol tehát $C(0)$ a mérések szórásnégyzete.

Ez a kovariancia fajta akkor alkalmazható, ha feltételezhető, hogy az adatok a hely, illetve a mért objektumok közti távolság függvényében stacionáriusak. A stacionaritás többféleképpen értelmezhető. Ha $\mathbb{R}^k$-beli pontokhoz tartozó mérésekről van szó, akkor a stacionaritás azt jelenti, hogy bármely két $u, v \in \mathbb{R}^k$ pontra és tetszőleges $d \in \mathbb{R}^k$-ra a $cov(y(u), y(v)) = cov(y(u+d), y(v+d))$, vagyis, hogy a mező kovariancia függvénye csak az $u - v$ differenciától függ. Ennél erősebb megkötés, ha a függés csak a $||u-v||$ távolságtól áll fenn, amikoris a mezőt izotrópnak nevezik. A másik fajta stacionaritás értelmezés szerint csak a $D^2(y(u) - y(v))$ differencia variancia függ az $u - v$ differenciától. Az ilyen mezőket differencia stacionárius, vagy belső (vagy lényegileg, angolul intrinsic) stacionárius mezőknek nevezzük. A differencia stacionaritásnak is létezik erősebb változata. Ezek esetén a mezőértékek differenciájának szórása csak az $|u - v|$ távolságtól függ.

4.

Az R-project és a gcmr csomag

4.1. R-projekt

Az R [8], [9] egy nyílt forráskódú, így ingyenes használható, professzionális és folyamatos fejlesztés alatt álló statisztikai szoftvercsomag, amelyben hihetetlen gazdagságban állnak rendelkezésre a már kidolgozott eljárásokat tartalmazó függvények és munkakörnyezetek. Lehetőséget nyújt számos elemző (lineáris és nemlineáris modellezés, klasszikus statisztikai tesztelés, idősor analízis, osztályozás, klaszterezés) és grafikai technika alkalmazására.

Az R sikerét az ingyenes és szabadon használható volta mellett (vagy talán inkább az alapján) elsődlegesen a CRAN (Comprehensive R Archive Network) csomagtárolónak és a felhasználók által megosztható programkódoknak köszönheti. Mára a CRAN több mint 5000 R csomagot számlál, amelyek többnyire lefedik a jelenlegi statisztikai módszerek tárházát.

Az R alapvetően egy interaktív statisztikai/adatelemző környezet, ahol a felhasználók utasításokat adnak ki az R konzolnak a parancsok végrehajtására. Az eredmények szintén itt jelennek meg. De az R legnagyszerűbb tulajdonsága az aktív közösség és a számtalan szabadon elérhető algoritmus.

4.2. A gcmr csomag

A gcmr (Gaussian Copula Marginal Regression) csomagban [7] elérhető regressziós modellek:

általánosított lineáris modell

negatív binomiális regresszió

béta regresszió

Az `gcmr` csomag legfőbb függvénye a `gcmr()`, amely lehetővé teszi Gauß kopula alapú regresszió modellezését. A függvény a minták kiértékelésére a maximum likelihood módszert alkalmazza. Standard argumentumokkal rendelkezik, amelyek segítségével leírhatjuk a általunk használni kívánt módszert. Ezek a következők:

```
gcmr(formula, data, subset, offset, marginal, cormat, start,
     fixed, options = gcmr.options(...), model = TRUE, ...)
```

Ezekből a két legfontosabb a `marginal` és a `cormat` argumentumok, amelyekkel megadhatjuk a modell marginális eloszlását, illetve a kovariancia mátrix típusát.

Előbbi jelenleg a következőket veheti fel:

béta eloszlás

binomiális eloszlás

Gamma eloszlás

normális (Gauß) eloszlás

negatív binomiális eloszlás

Poisson eloszlás

Weibull eloszlás

Mindegyik eloszlás esetén lehetőseg van függvénykapcsolat meghatározására.
A kovariancia mátrix típusa lehet:

független

ARMA(p,q)

klaszterezett

Matern-féle

5.

A kopula alapú regresszió alkalmazása

A kopula alapú regresszió alkalmazását három különböző modell esetén mutatom be:

klaszter

idősor

stacionárius mező

Ehhez az előző fejezetben ismertetett `gcmr` csomagot használom:

```
library("gcmr")
```

5.1. Kopula modell ismételt mérések esetén

A [10] cikkben szereplő adatok egy klinikai kísérlet eredményeit reprezentálják, melyben 59 epilepsziás pacienst figyeltek meg. A kutatás során *progabide* (egyfajta aminosav) hatékonyságát vizsgálták az epilepszia kezelésére. (A 49. sorszámú adatot, mint a többitől jelentősen eltérőt, az elemzésből kihagytuk.)

Ismételt mérések,
`cluster` típusú kovariancia, negatív binomiális koordináta modell,
epilepszia adatok (`epilepsy`).

```
# ----
# 'cluster' tipusu kovariancia

data(epilepsy)
str(epilepsy) # 295 obs. of  6 var
# $ id     : int  1 1 1 1 1 2 2 2 2 2 ...
           : a paciens azonositoja
# $ age    : int  31 31 31 31 31 30 30 30 30 30 ...
           : a paciens eletkora
# $ trt    : int  0 0 0 0 0 0 0 0 0 0 ...
           : jelzi, hogy a paciens 'progabide'-ot (1) vagy
           :                              placebot (0) kapott
# $ counts: int  11 5 3 3 3 11 3 5 3 3 ...
           : az epilepszias rohamok szama
# $ time   : num  8 2 2 2 2 8 2 2 2 2 ...
           : megfigyelesi idoszak (8 jelzi az elso vizsgalatot,
           :                       2 a kesobbieket)
# $ visit : num  0 1 1 1 1 0 1 1 1 1 ...
           : indikator (0 jelzi az alapot, 1 a tovabbiakat)
```

```
table(epilepsy$id)# 1-59 x 5
table(epilepsy$age)/5# 18-57
table(epilepsy$trt) # 0//1: 140 155
table(epilepsy$trt,epilepsy$id) # 0 ha id=1:28; 1 ha id=29:59
table(epilepsy$counts)# 0.. 151
table(epilepsy$time) # 2//8: 236,  59
table(epilepsy$id,epilepsy$time) # 2//8: 59x(4+1)
table(epilepsy$visit) # 0//1: 59, 236
table(epilepsy$id,epilepsy$visit) # 0//1: 59x(1+4)
table(epilepsy$time,epilepsy$visit) # 0//1: 59x(1+4)
```

A `counts` változó tartalmazza az adott időszak alatti rohamok számát. A rohamszámot a magyarázó változók alapján általánosított lineáris regresszió modellel, a célváltozó negatív binomiálisát feltételezve modellezzük. A mérések összefüggésének modellezésére, tekintettel arra, hogy longitudinális adatokról van szó, a `cluster` típusú korreláció mátrixú Gauß-kopulát alkalmazzuk.

```
M <- gcmr(counts ~ offset(log(time)) + visit + trt + visit:trt,
                data = epilepsy, subset = (id != 49),
                marginal = negbin.marg,
                cormat = cluster.cormat(id, type = "ind"))

# a 'counts' a celvaltozo
# paciensenkent 5 adat
# age, trt paciens jellemzo
# minden paciens eseten time 4x2 1x8
# minden paciens eseten vizit 1x0 4x1
# visit=0 <=> time=8 (paciensenkent ez volt 1szer)
# visit=1 <=> time=2 (paciensenkent ez volt 4szer)

summary(M)
M<- gcmr(counts ~ offset(log(time)) + visit:trt,
          data = epilepsy,subset = (id != 49),
          marginal = negbin.marg,
          cormat = cluster.cormat(id, type = "ind"))
```

```
summary(M)
# Coefficients marginal model:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept)  1.40177    0.06969  20.114  < 2e-16 ***
# visit:trt   -0.35154    0.11131  -3.158  0.00159 **
# dispersion   0.73901    0.07169  10.309  < 2e-16 ***
str(M,m=1,give.a=FALSE)
# $ y            : int [1:290, 1] 11 5 3 3 3 11 3 5 3 3 ...
# $ x            : num [1:290, 1:4] 1 1 1 1 1 1 1 1 1 1 ...
# $ offset       :List of 2 mean+$precision
# $ n            : int 290
# $ marginal     :List of 6
# $ cormat       :List of 5
# $ ibeta        : int [1:5] 1 2 3 4 5
# $ igamma       : NULL
# $ nbeta        : num 5
# $ ngamma       : num 0
# $ call         : language gcmr(formula =
# $ estimate     : Named num [1:5] 1.348 0.112 -0.107 -0.302 0.734
# $ lower        : num [1:5] -Inf -Inf -Inf -Inf 1.49e-08
# $ upper        : num [1:5] Inf Inf Inf Inf Inf
# $ fixed        : logi [1:5] NA NA NA NA NA
# $ options      :List of 5
# $ maximum      : num 948
# $ convergence  : int 0
# $ fitted.values: num [1:290] 30.78 8.61 8.61 8.61 8.61 ...
# $ jac          : num [1:290, 1:5] -0.838 -0.493 -0.766 -0.766
# $ hessian      : num [1:5, 1:5] -339.8462 -264.2632 -170.6915
# $ formula      :Class 'formula'  language counts ~
# $ terms        :List of 3
# $ levels       :List of 3
# $ contrasts    :List of 2
# $ model        :'data.frame':  290 obs. of  4 variables:
```

Ugyanezekre az adatokra általánosított lineáris modellt például a

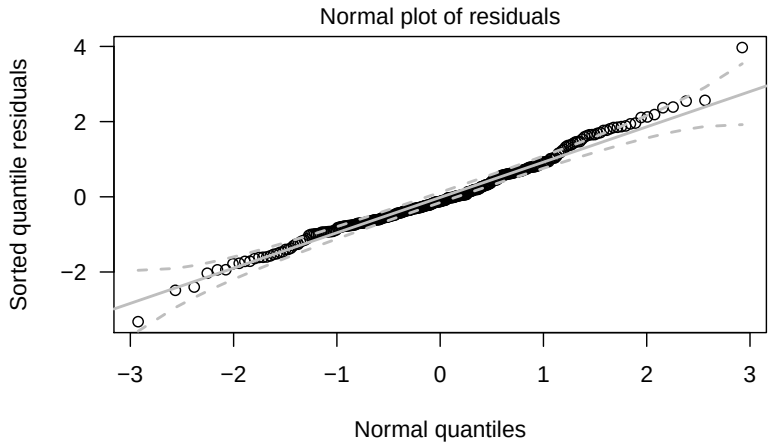
```
W1 <- glm(counts ~ offset(log(time)) + visit + trt + visit:trt,
          data=epilepsy)
```

paranccsal, vagy a célváltozóról Poisson eloszlást feltételezve a

```
W2 <- glm(counts ~ offset(log(time)) + visit + trt + visit:trt,
          data=epilepsy,family=poisson())
```

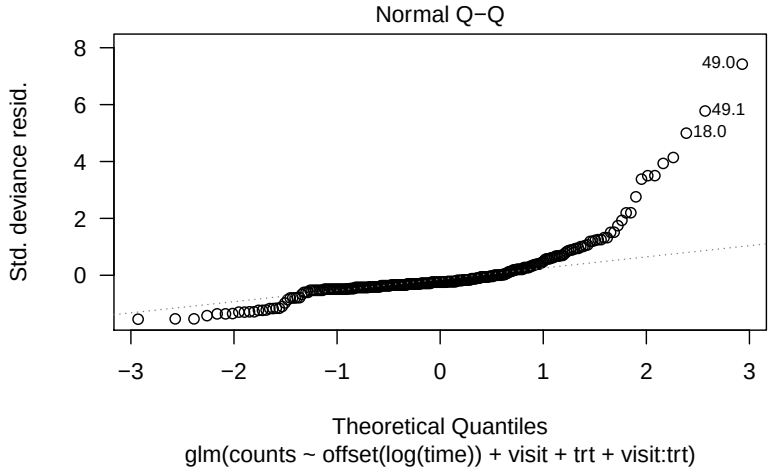
paranccsal illeszthetünk. A kopula modell és e két modell megfelelőségét például az alábbi három ábra alapján ellenőrizhetjük.

```
dev.new()
plot(M,which=3)
dev.new()
plot(W1,which=2)
dev.new()
plot(W2,which=2)
```

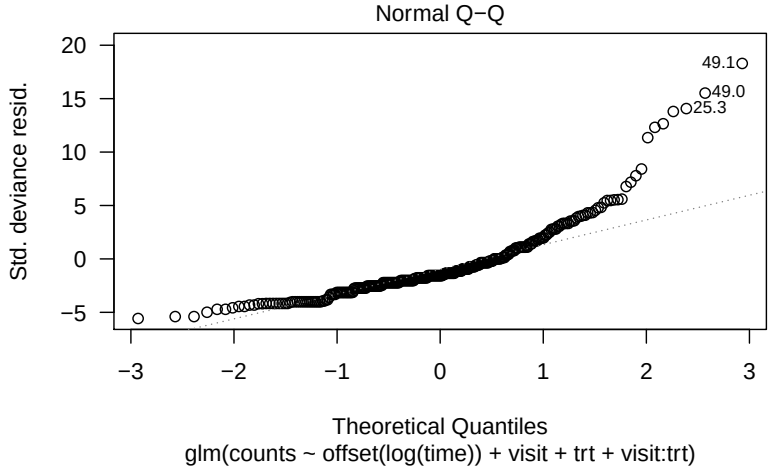


5.1. ábra. Kopula regresszió - epilepszia adatok modellje

Ez a QQ ábra azt mutatja, hogy az illesztett kopula regresszió a maradékok szempontjából nézve jól modellezi az adatokat.



5.2. ábra. GLM regresszió - epilepszia adatok modellje



5.3. ábra. Poisson GLM regresszió - epilepszia adatok modellje

A GLM modellek megfelelősége esetén ennek a két QQ ábrának is lényegében egy egyenes mentén sorakozó pontokat kellene mutatnia. Látható, hogy a két GLM modell a maradékok szempontjából sokkal rosszabb mint a kopula regressziós modell. Az utóbbi két modell lényegében érvényes modellnek sem tekinthető.

5.2. Idősor adatok kopula regresszió modellje

A [11]-ben közölt adatok a Sao Paulo-i "rejtett" munkanélküliséget mérik és az 1991. január és 2005 novembere közti 179 hónapra vonatkoznak. A "rejtett" munkanélküliek körébe tartoznak azok, akik például feketén vannak foglalkoztatva, de azok is, akiket a családon belül fizetség nélkül foglalkoztatnak. A brazil kormány intézete által mért ilyen adat 4% körül ingadozik, minimális értéke mintegy 2%, és nem haladja meg a 6%-ot.

Idősor adatok,
`ARMA` típusú kovariancia, béta koordináta modell,
munkanélküliségi arányok (`HUR`).

```
# ----
# 'arma' tipusu kovariancia

data(HUR)
plot(HUR, ylab = "rate", xlab = "time")
```

A munkanélküliség idősor adatainak modellezésére egyfelől a béta regressziót alkalmazzuk, mint marginális modellt. Másfelől az adatok összefüggésének leírására egy `ARMA(1,3)` modellt alkalmazunk. A "rejtett" munkanélküliségi arányt a `HUR` változó, az idő tényezőt a `trend` skálázott időpont adatok tartalmazzák.

```
trend <- scale(time(HUR))
M <- gcmr(HUR ~ trend | trend, marginal = beta.marg,
                                cormat = arma.cormat(1, 3))
M

summary(M)

str(M,m=1,give.a=FALSE)
```

Az illesztett Gauß-kopula kovariancia mátrixa:

```
M$cormat

# $npar 4
# $start
# function ()
# { tau <- rep(0, p + q)
#     names(tau) <- c(if (p) paste("ar", 1:p, sep = "") else NULL,
#         if (q) paste("ma", 1:q, sep = "") else NULL)
#     tau }
# $chol
# function (tau, not.na)
# { if ((p && any(Mod(polyroot(c(1, -tau[iar]))) < 1.01)) ||
#        (q && any(Mod(polyroot(c(1, tau[ima]))) < 1.01)))
#         return(NULL)
#     n <- length(not.na)
#     rho <- ARMAacf(tau[iar], tau[ima], n - 1)
#     r <- seq(1, n)[not.na]
#     chol(outer(r, r, function(i, j) rho[1 + abs(i - j)]))}
# attr(,"class")
#  "arma.gcmr"   "cormat.gcmr"

par(mfrow = c(2, 2))
plot(M)
```

Látható, hogy a `gcmr` függvény a `stat::ARMAacf` függvény segitségével számolja ki a Gauß-kopula kovariancia paraméteréhez szükséges kovariancia mátrixot.

5.3. Stacionárius mező kopula modellje

A [12] cikk malária adatait használjuk fel. A szerzők 1992-es adatok alapján Gambia 65 településén vizsgálták azt az összefüggést, ami a gyermekek malária fertőzöttsége és az orvosi ellátás, valamint szúnyogok ellen védő hálónak az alkalmazása közt esetlegesen fennáll. A cikkben felhasznált módszer a krigelés. Most ugyanezeken az adatokon a kopula regresszió működési módját mutatjuk be.

Stacionárius mező adatok,
Matérn típusú kovariancia, binomiális koordináta modell,
malária adatok (`malaria`).

```
# ----
# 'matern' tipusu kovariancia

library(sp)
data(malaria)
dim(malaria) # 65 x 10
str(malaria)
# $ x      : hely koordinata x komponens
# $ y      :                 y komponens
# $ cases  : malaria esetszama
# $ size   : megvizsgaltak szama
# $ age    : atlagos kor
# $ netuse : vedohalo alkalmazas gyakorisaga a telepulesen
# $ treated: vedohalo alkalmazas gyakorisaga a vizsgaltak koreben
# $ green  : zöld terület aranya a telepules korzeteben
# $ phc    : van-e egeszsegugyi szolgalat a telepulesen
# $ area   : a telepules jellege
```

A megvizsgált paciensek körében tapasztalt malária esetszámot általánosított lineáris regresszió módszerét felhasználva, a célváltozó-pár binomiális eloszlását feltételezve modellezzük. A két célváltozót a `cases` változó és a `size-cases` differencia (a pozitív és a negatív esetek száma) jelentik. A Gauß-kopula kovariancia mátrixának feltételezett strukturája a Matérn-féle:

```
D <- spDists(cbind(malaria$x, malaria$y)) / 1000
M <- gcmr(cbind(cases, size-cases) ~ netuse + I(green / 100) + phc,
          data = malaria,
          marginal = binomial.marg, cormat = matern.cormat(D),
          seed = 12345)
M
# Marginal model parameters:
# (Intercept)        netuse  I(green/100)           phc
#      -0.8269       -1.1761        2.9476       -0.4051
# Gaussian copula parameters:
#   tau
# 1.511

summary(M)

# Coefficients marginal model:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)   -0.8269     0.4065  -2.034   0.0419 *
# netuse        -1.1761     0.1605  -7.326 2.37e-13 ***
# I(green/100)   2.9476     0.7499   3.930 8.48e-05 ***
# phc           -0.4051     0.1018  -3.978 6.96e-05 ***
#
# Coefficients Gaussian copula:
#     Estimate Std. Error z value Pr(>|z|)
# tau   1.5112     0.3773   4.006 6.18e-05 ***
#
# log likelihood = 252.68,  AIC = 515.36

str(M,m=1,give.a=FALSE)
```

A modellben felhasznált kovariancia mátrix:

```
M$cormat

# $npar #  1
# $start
# function ()
# {   tau <- median(D)
#     names(tau) <- c("tau")
#     attr(tau, "lower") <- sqrt(.Machine$double.eps)
#     tau }
# $chol
# function (tau, not.na)
# {   S <- geoR::matern(D, tau, alpha)
#     q <- try(chol(S[not.na, not.na]), silent = TRUE)
#     if (inherits(q, "try-error"))
#         NULL
#     else q }
# attr(,"class")
# [1] "matern.gcmr" "cormat.gcmr"
```

A `gcmr` függvény a kovariancia függvény kiszámolásához a `geoR` csomag `matern` függvényét használja fel:

```
geoR::matern
# geoR::matern <-
# function (u, phi, kappa)
# {  if (is.vector(u))
#         names(u) <- NULL
#    if (is.matrix(u))
#        dimnames(u) <- list(NULL, NULL)
#    uphi <- u/phi
#    uphi <- ifelse(u > 0, (((2^(-(kappa - 1)))/ifelse(0, Inf,
#        gamma(kappa))) * (uphi^kappa) * besselK(x=uphi,nu=kappa)),1)
#    uphi[u > 600 * phi] <- 0
#     return(uphi)  }
```

Ábrák jegyzéke

Irodalomjegyzék

[1] Abe Sklar. **Fonctions de répartition à n dimensions et leurs marges**. Publ. Inst. Statist. Univ. Paris, **8**:229–231, 1959. (3)

[2] Ludger Rueschendorf. Mathematical risk analysis. Dependence, risk bounds, optimal allocations and portfolios. Springer, 2013. (11)

[3] Thorsten Schmidt. **Coping with copulas**. Indian Statistical Institute, Kolkata, 2007. (11)

[4] Paul Embrechts; Filip Lindskog and Alexander McNeil. **Modelling Dependence With Copulas and Applications to Risk Management**. In S.T Rachev, editor, Handbook of Heavy Tailed Distributions in Finance. Elsevier, 2001. (11), (13), (14)

[5] Nelson R. An introduction to copulas. Springer, New York, 1999. (11)

[6] G. Masarotto and C. Varin. **Gaussian copula marginal regression**. Electronic Journal of Statistics, **6**:1517–1549, 2012. (15)

[7] G. Masarotto and C. Varin. **Gaussian Copula Regression in R**. Journal of Statistical Software, **v077i08**, 2017. (15), (20)

[8] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. (19)

[9] Budapest Users of R Network. r-projekt.hu. (19)

[10] Peter F. Thall and Stephen C. Vail. **Some Covariance Models for Longitudinal Count Data with Overdispersion**. Biometrics, **46**:657–671, 1990. (22)

[11] Andréa V. Rocha and Francisco Cribari-Neto. **Beta autoregressive moving average models**. Test, **18**:529–545, 2009. (28)

[12] Diggle P.; Moyeed R.; Rowlingson B. and Thomson M. **Childhood malaria in The Gambia: a case-study in model-based geostatistics**. Journal of the Royal Statistical Society. Series C, Applied Statistics, **51**:493–506, 2002. (30)