Journal of the Korean Data & Information Science Society
2016, **27**(3), 689–700

http://dx.doi.org/10.7465/jkdi.2016.27.3.689
한국데이터정보과학회지

비모수적 코플라를 이용한 반복측정 이변량 자료의 조건부 결합 분포 추정[†]

곽민정[1]

[1]영남대학교 통계학과

접수 2016년 4월 15일, 수정 2016년 5월 9일, 게재확정 2016년 5월 19일

요 약

본 논문에서는 이변량 경시적 자료의 조건부 결합 분포를 추정하기 위하여 회귀 모형과 코플라 모형을 연구하였다. 주변 분포의 추정을 위하여 시변 전환 모형을 고려하였고, 이변량 반응변수 각각에 대한 주변 분포를 경험 분포를 이용한 비모수적 코플라를 이용하여 결합하여 조건부 결합 분포를 추정하였다. 주변 분포 모형의 모수 추정치는 추정방정식의 해로 얻어낼 수 있으며 우리가 제안한 모형은 조건부 평균 모형만으로 자료를 설명하기 어려운 경우에 적용될 수 있다. 시변 전환 모형과 비모수적 코플라 모형을 결합한 본 논문의 방법은 반복 측정된 이변량 경시적 자료에 대한 모형화가 모형에 대한 가정에서 비교적 자유로운 장점이 있다. 우리는 본 논문의 방법을 반복 측정된 이변량 콜레스테롤 자료를 분석하는데 적용하여 보았다.

주요용어: 경험 코플라, 시변 전환 모형, 이변량 경시적 자료, 조건부 결합 분포.

1. 서론

경시적 자료 (longitudinal data)는 각 개인에게서 관측치가 시간에 따라 반복적으로 얻어지는 경우에 발생한다. 연구에 참여한 각각의 개인들에 대해 시간의 흐름에 따라 규칙적으로 혹은 불규칙적으로 관측치가 얻어지며, 동일한 개체에서 관측치가 여러번 얻어지므로 관측치들이 서로 독립이라는 가정이 성립하지 않는다. 이러한 경시적 자료를 분석함에 있어서는 다변량 자료의 특성과 시계열 자료의 특성을 함께 고려하여야 한다. 첫째, 경시적 자료가 다변량 자료와 다른 특징은 관측치들이 시간에 따라 순서가 정해져 있다는 점이고, 둘째로 시계열 자료와 다른 특징은 시계열 자료와는 달리 한 개체에서 얻어지는 측정 시점들의 숫자가 상대적으로 적다는 점이다. 의학 통계에서 경시적 자료의 예로는 임상 시험에 있어서 두 가지의 서로 다른 치료법을 같은 환자에게 처리하여 반응변수의 변화를 관측하게 되는 교차설계에서 얻어지는 비교적 단순한 경시적 자료부터, 정기적으로 병원에 방문하여 동일한 환자에 대하여 각종 임상적 수치를 반복하여 관측, 기록하는 다소 복잡한 경시적 자료에 이르기까지 매우 다양하다.

경시적 자료에 대한 통계학적인 이론들은 Diggle 등 (1994)과 Lindsey (1993)에 의해 잘 정리가 되어 있고, 경시적 자료를 분석하기 위하여 사용된 각종 회귀 분석들에 대한 최근의 요약은 Molenberghs과 Verbeke (2005)에서 얻을 수 있으며, 최근의 경시적 자료 분석 연구 결과로는 Cho와 Dashnyam (2013), Jeon과 Lee (2014) 등이 있다. 시간에 따라 반복적으로 관측치가 얻어지는 경시적 자료

[†] 이 논문은 2014 년도 정부 (미래창조과학부) 의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (NRF-2014R1A1A1002465).

[1] (38541) 경북 경산시 대학로 280, 영남대학교 통계학과, 교수. E-mail: mjkwak@yu.ac.kr

의 분석 목적은 여러 가지가 있으나, 그 중의 대표적인 분석 목적은 회귀 분석을 이용하여 관측된 반응변수와 독립변수간의 상관구조를 설명하거나 반응변수의 조건부 평균에 대한 추론을 하는 것이다.

본 연구의 자료는 미국의 한 연구기관에서 얻어진 아동들의 성장과 건강에 대하여 각종 임상수치를 수집한 자료이다. 총 2500 여명의 백인과 흑인 청소년기 소녀들을 대상으로 1986년부터 1997년까지 2년마다 심혈관계 질환 관련 각종 임상수치를 관측하여 기록한 자료이다 (NGHSRG, 1992; NHBPEP, 2004). 콜레스테롤은 지방의 일종으로 인체의 기능을 정상적으로 유지시키는 데 필수적으로 필요한 구성 성분으로 모두 다섯 종류로 나뉘는데 이중 '좋은 콜레스테롤'이라 불리는 고밀도 지질단백질 (high-density lipoprotein; HDL)과 '나쁜 콜레스테롤'로 불리는 저밀도 지질단백질 (low-density lipoprotein; LDL)이 중요하게 다루어진다. 콜레스테롤은 인체에 필요한 필수 영양소로서 체내에서 일정량 합성되지만, 많으면 건강에 해롭다. 혈액 속의 콜레스테롤 농도가 높으면 동맥경화의 원인이 되거나 협심증, 심근경색증 등의 심장질환과 뇌졸중, 고혈압 등의 뇌혈관 질환이 발생할 가능성이 높아진다고 알려져 있다 (Anderson, 1987). 이런 질병들은 최근 발생하는 주요 사망원인이기도 하기 때문에 무엇보다 이러한 질병에 대한 이해와 예방, 조기 진료를 위해 콜레스테롤 수치의 적절한 측정과 관리는 중요한 문제로 여겨진다.

우리는 이 논문에서 반복 측정된 이변량 콜레스테롤 자료 (HDL, LDL)를 분석하였다. 일변량 경시적 자료에 대한 분석 방법은 많이 알려져 있지만, 서로 연관이 있는 다변량 경시적 자료에 대한 통계적 분석 방법은 많지 않다. Song 등 (2009)는 가우시안 코플라를 이용하여 결합 회귀분석을 실시하였으며, Leon과 Wu (2011)은 역시 가우시안 코플라 함수를 이용하여 이변량 자료를 모형화 하였다. 그러나 그들 모두 고정 시점에서 단면적으로 얻어진 다변량 자료를 분석하는데 그쳤다. 이에 본 논문에서는 각각의 일변량 경시적 자료에 대하여 주변 분포에 대한 적절한 통계적 모형을 세우고, 비모수적 코플라 (copula) 함수를 이용하여 주변 분포들을 결합하여 결합 조건부 분포를 얻고자 한다. 본 논문의 새로운 점은 (1) 특정 한 시점이 아닌 반복 측정된 이변량 자료를 모형화 한다는 점 (2) 비모수적인 코플라 함수를 이용하여 모형에 대한 가정에서 좀 더 자유로운 분석을 할 수 있다는 점을 들 수 있다.

본 논문의 구조는 다음과 같다. 2절에서는 본 연구의 자료 출처와 구조를 설명하고 조건부 분포와 결합 분포를 추정하는 방법에 대하여 각각 설명한다. 3절에서는 2절에서 제안된 방법을 이용하여 연구자료를 분석하고 결과를 해석하였다. 4절에서는 본 논문의 내용을 요약정리하고 본 연구의 장점과 단점, 그리고 향후 연구 문제에 대하여 토의하였다.

2. 대상 및 방법

2.1. 연구 대상

미국 국립보건원 (NGHS)에서는 세 군데의 대학병원 센터들을 중심으로 1986년부터 2379명의 9∼10세의 여아 (백인 49%, 흑인 51%)들을 대상으로 1997년까지 10여년에 걸쳐 해마다 아동의 성장과 심혈관질환의 위험요소를 측정하였다. 지역은 전체 인구중 인종별로 가구당 수입과 부모의 교육수준을 잘 반영할 수 있도록 선택되었다. 연구자들은 이 여아들을 대상으로 표준화된 프로토콜에 따라 부모의 동의를 거쳐 키, 몸무게, 혈압, 생활방식을 묻는 설문지 등을 해마다 측정하여 기록하였다. 추적율은 인종별로 백인 여아 74% 에서 흑인 여아 95%에 이르며 여아들은 10년간 평균적으로 8.8번 센터를 방문하였다 (백인 평균 8.6회/10년, 흑인 평균 9.0회/10년). 특별히 혈액 채취가 필요한 콜레스테롤은 약 3년에 한번씩 (백인 평균 3.1회/10년, 흑인 평균 3.0회/10년) 측정되었는데, 금식 후에 총콜레스테롤, 중성지방, HDL 콜레스테롤, LDL 콜레스테롤을 측정하여 기록하였다.

2.2. 자료 구조

우리는 경시적 자료 분석에서 일반적으로 쓰이는 다음의 자료 구조를 생각한다. 우리는 시간에 따라 반복 측정된 n명의 독립적인 개체들로 이루어진 경시적 자료를 가정하며 i번째 개체에서 n_i개의 관측치가 $t_{ij} \in \mathcal{T}, j = 1, \ldots, n_i$시점에서 얻어진다고 가정한다. 여기서 $\mathcal{T}$는 시간 인덱스를 나타내는 유계 집합으로 관심의 대상이 되는 관찰 기간을 나타낸다. 정해진 시점 $t \in \mathcal{T}$에 대하여, $\mathbf{Y}(t) = (Y^1(t), Y^2(t))$는 시점 t에서의 실수 범위의 이변량 반응변수를 나타내고, $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))^T$는 p차원 실수 벡터로 표현되는 독립변수들의 집합이다. 윗첨자 ^{T}는 벡터나 행렬의 전치를 나타낸다. 경시적 자료 구조 $\{\mathbf{X}(t), \mathbf{Y}(t), t)\}$에 대하여 연구자가 n명의 개체에 대하여 실제로 얻은 관측치는 $\{\mathbf{X}(t_{ij}), \mathbf{Y}(t_{ij}), t_{ij}; j = 1, \ldots, n_i, i = 1, \ldots, n\}$로 표현할 수 있다. 여기서 아래 첨자 i는 i번째 개체를 나타내며, 아래 첨자 j는 j번째 시점을 나타낸다. 여기서 유의할 점은 실제 임상 연구에서 개체를 연속적인 시간에 따라 관측하는 것은 불가능 하므로, n명의 각각의 개체는 서로 다른 관측 시점들의 집합 $\mathbf{t} = (t_1, \ldots, t_J)^T \in \mathcal{T}^J$, $J > 1$의 부분 집합에서 관측치가 얻어진다. 실제 연구에서 이상적인 상황은 미리 정해진 관측 시점에서 모든 개체들이 정확하게 관측되어지는 경우이지만, 임상 연구 진행에 있어 여러 가지 현실적인 문제로 미리 정해진 관측 시점에서만 관측치를 얻기는 거의 불가능한 상태이다. 다시 말하면 모든 개체들이 각각의 시점에서 관측되지는 않는다. 이에 t_j시점에서 관측된 개체들의 수를 n_j로 표기하였다. 이 자료의 구조는 NGHS 자료의 구조와도 일치하는 것이다.

2.3. 주변 조건부 분포 모형

우리는 이변량 반응변수 각각에 대하여 일변량 주변 조건부 분포를 추정한다. 시점 t에서 공변량 $\mathbf{X}(t)$가 $x(t)$로 주어졌을때 일변량 반응변수 $Y^k(t)$, $k = 1, 2$의 주변 조건부 분포를 $F_{kt}(y_k|x) = \Pr[Y^k(t) \leq y_k(t)|X(t) = x(t)], k = 1, 2$로 정의한다. 이때, 우리는 다음의 시변 전환 모형 (time-varying transformation model) 을 가정한다.

$$g_k[1 - F_{kt}(y_k|x)] = h_k(y_k, t) + X^T(t)\beta_k(t), \quad k = 1, 2. \tag{2.1}$$

각각의 일변량 인덱스 $k = 1, 2$에 대하여, $g_k(\cdot)$는 알려져 있는 감소하는 형태의 연결함수이며, $h_k \equiv h_k(y_k, t)$는 알려져 있지 않은 y_k에 대하여 증가하는 형태의 기저함수이다. 회귀 모수는 $\beta_k(t) = (\beta_{k1}(t), \ldots, \beta_{kp}(t))^T$이며 각각의 $\beta_r(t)$, $r = 1, \ldots, p$는 시간 $t \in [l_t, u_t]$의 유계함수이다. 윗첨자 ^{T}는 벡터나 행렬의 전치를 나타낸다. 위의 주변 조건부 분포 모형의 유용한 예로는 비례함수 모형을 위한 연결함수 $g_k[1 - F_{kt}(y_k|x)] = \log[-\log\{1 - F_{kt}(y_k|x)\}]$와 비례오즈 모형 $g_k[1 - F_{kt}(y_k|x)] = -\log[\{1 - F_{kt}(y_k|x)\}/F_{kt}(y_k|x)]$을 고려할 수 있다.

함수 $\phi_k(\cdot) = g_k^{-1}(\cdot)$, $k = 1, 2$를 g_k의 역함수라고 표기하면, 관측시점 t에 대하여 위의 모형은 Cheng 등 (1995)이 제안한 다음의 전환 모형과 동등함을 알 수 있다.

$$h_k(Y^k(t), t) = -X^T(t)\beta_k(t) + \epsilon_k, \quad k = 1, 2.$$

여기서 $\epsilon_k = g_k[1 - F_{kt}(Y^k(t)|X(t))]$는 누적분포함수로 $G_k(\cdot) = 1 - \phi_k(\cdot)$를 가지는 확률오차임을 알 수 있다. 모형에서 알 수 있듯이 우리는 일변량 자료 $Y^1(t)$과 $Y^2(t)$각각에 대하여 서로 다른 회귀 모수와 연결 함수를 이용할 수 있으므로 이변량 자료 $(Y^1(t), Y^2(t))$각각의 구성성분에 대하여 좀 더 유연한 모형을 적합할 수 있게 된다. 예를 들어, $Y^1(t)$은 비례위험 모형을 이용하여 적합하는 동시에 $Y^2(t)$는 비례오즈 모형을 이용하여 적합할 수 있다. 모형 (2.1)에서 시점 t에서 공변량 $\{X(t)\}$가 주어졌을때 $Y^k(t)$의 주변 조건부 분포함수는 $F_{kt}(y_k|x) = 1 - \phi_k\{h_k(y_k, t) + X^T(t)\beta_k(t)\}$로 나타낼 수 있다. 즉, 모형 (2.1) 는 각각의 주변 조건부 분포함수 $F_{kt}(y_k|x)$에 대한 일반화 선형 구조를 가정하는 것

이며 시간에 따라 변하는 회귀 모수를 포함시킴으로써 다양한 자료를 적합하기에 더 유연한 모형이라고 할 수 있다.

구체적으로 추정량을 구하는 방법은, 먼저 Cheng 등 (1995)의 추정방정식 (estimating equation)에 대한 해로써 모형 (2.1) 의 회귀계수 $\beta_k(t)$, $k = 1, 2$에 대한 추정치 $\widehat{\beta}_k(t)$를 구하였다. 그리고 모형간의 관계를 이용하여 기저함수 h_k에 대한 추정치 $\widehat{h}_k$ 다음의 추정방정식에 대한 해로 얻어질 수 있다.

$$\frac{1}{n_j}\sum_i \left[I\{Y_i^k(t_j) > y_k\} - \phi\left(\widehat{h}(y_k, t_j) + X_i^T(t_j)\widehat{\beta}_k(t_j)\right)\right] = 0, \quad k = 1, 2.$$

실제 추정치의 계산에서 각각의 추정치 $\widehat{\beta}_k(t)$와 $\widehat{h}_k$는 뉴튼-랩슨 방법 등을 이용한 반복적인 알고리즘에서 수렴된 값으로서의 추정치로 구하였다. 이렇게 얻어진 추정치 $\widehat{\beta}_k(t)$와 $\widehat{h}_k$을 모형 (2.1)에 대입하여 주변 조건부 분포함수 $F_{kt}(y_k|x)$에 대한 추정량 $\widehat{F}_{kt}(y_k|x)$을 다음과 같이 얻었다.

$$\widehat{F}_{kt}(y_k|x) = 1 - \phi\left(\widehat{h}(y_k, t_j) + X_i^T(t_j)\widehat{\beta}_k(t_j)\right), \quad k = 1, 2.$$

2.4. 결합 조건부 분포 모형

많은 다변량 모형에서 각 변량간의 의존성은 코플라 (copula) 함수를 이용하여 표현될 수 있다. Sklar 정리 (Sklar, 1959)에 의하면, Y^1의 주변 분포를 F_1, Y^2의 주변 분포를 F_2라고 했을때, Y^1과 Y^2의 결합 분포 J는 코플라 함수 C를 이용하여 다음과 같이 표현된다.

$$J(y_1, y_2) = \Pr\{Y^1 \le y_1, Y^2 \le y_2\} = C\{F_1(y_1), F_2(y_2)\}, \quad y_1, y_2 \in R.$$

Genest과 MacKay (1986) 등에 의해 코플라 함수는 광범위하게 연구되어 왔으며, (수축기 혈압, 확장기 혈압), (HDL,LDL) 등의 이변량 자료에 존재하는 변수들간의 상호 연관성을 코플라 함수를 이용하여 모형을 적합할 수 있다. Genest 등 (1995), Oakes (1986), Joe (1993) 등에 의해 연구된 모수적 코플라 함수는 코플라 함수가 변수들 간의 의존성을 설명하기 위하여 실수 혹은 벡터 값을 가지는 모수의 함수적 형태로 표현되며 일반적으로 많이 사용되는 코플라 함수로는 아르키메디안 코플라 함수 (Archimedean copula), 극단값 코플라 함수 (extreme value copula) 등이 있다. 모수적 코플라 함수를 이용한 모형에서 코플라 모수는 유사 로그 우도 함수 (pseudo log-likelihood function) 를 최대로 하는 최대 우도 추정치를 코플라 모수의 추정치로 사용하며 이렇게 얻어진 추정치는 최대 우도 추정치의 점근성질인 일치성 (consistency)과 대표본 정규성 (asymptotic normality)을 지니게 된다. 반면 실제 자료분석에 있어서 특정한 모수의 함수 형태를 가정한다는 점에서 다소 제약적이라고 할 수 있다. 실제로 많은 모수적 코플라 함수들을 자료에 적용하여 자료를 가장 잘 설명하는 코플라 함수를 선택하는 문제에 대하여 논의되고 있다. 본 연구에서는 코플라 함수에 대한 모수적 가정을 하지 않은 비모수적 경험 분포 (empirical distribution)를 이용하여 이변량 반응변수의 결합분포를 추정하고자 한다.

$J_t(y_1, y_2) = \Pr\{Y^1(t) \le y_1(t), Y^2(t) \le y_2(t)\}$를 이변량 반응 변수 $(Y_i^1(t), Y_i^2(t))$의 결합 분포라고 하면 그에 대응되는 코플라는 $C_t(u_1, u_2) = J_t(F_{1t}^{-1}(u_1), F_{2t}^{-1}(u_2))$, $u_1, u_2 \in [0, 1]$로 표현되어 질 수 있으며 이는 다시 $J_t(y_1, y_2) = C_t(F_{1t}(y_1), F_{2t}(y_2))$, $y_1, y_K \in [l_y, u_y]$로 나타내어 진다. 따라서 $J_t(y_1, y_2)$에 대한 추정치 J_{nt}는 비모수적 경험 분포 함수를 이용하여 다음과 같이 나타낼 수 있다.

$$J_{nt}(y_1, y_2) = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i^1(t) \le y_1(t), Y_i^2(t) \le y_2(t)\}, \quad y_1, y_2 \in [l_y, u_y].$$

다음으로 공변량이 $X(t) = x(t)$, 시점이 $t \in [l_t, u_t]$로 주어졌을 때, 이변량 반응 변수 $(Y^1(t), Y^2(t))$의 결합 조건부 분포함수를 다음과 같이 정의한다.

$$J_t(y_1, y_K|x) = \Pr\left[F_{1t}(Y^1(t)|x) \le F_{1t}(y_1(t)|x), F_{2t}(Y^2(t)|x) \le F_{2t}(y_2(t)|x)\right].$$
$$\equiv C_t(u_1, u_2), \quad t \in [l_t, u_t],$$

여기에서 $u_1 = F_{1t}(y_1|x), u_2 = F_{2t}(y_2|x)$를 나타낸다. 위의 모형에서 $Y^1(t)$, $Y^2(t)$각각에 대한 주변 조건부 함수 $F_{1t}(y_1|x)$, $F_{2t}(y_2|x)$의 자리에 2.3절에서 추정된 주변 조건부 분포 함수 $\widehat{F}_{1t}$, $\widehat{F}_{2t}$을 대입하여 다음의 추정된 결합 조건부 분포 함수를 얻는다.

$$\widehat{J}_{nt}(y_1, y_2|x)$$
$$= \frac{1}{n}\sum_{i=1}^{n} I\left\{\widehat{F}_{1t}(Y_i^1(t)|x) \le \widehat{F}_{1t}(y_1(t)|x), \widehat{F}_{2t}(Y_i^2(t)|x) \le \widehat{F}_{2t}(y_2(t)|x)\right\}, \tag{2.2}$$
$$\equiv \widehat{C}_{nt}(u_1, u_2), \quad t \in [l_t, u_t].$$

3. 결과

미 국립보건원에서 실시한 NGHS 연구에서는 1166 명의 백인 여아와 1213 명의 흑인 여아를 대상으로 10여년간 추적하여 각종 임상 자료를 수집하였다. 각 연구대상의 방문 횟수는 1회부터 10회까지 다양하며 대략 평균 8.8회 표준편차 2.2회임을 알 수 있었다. Daniels 등 (1998), Thompson 등 (2007), Obarzanek 등 (2010)은 NGHS 자료를 이용하여 혈압에 대해 단면적 자료에 대한 기초 분석을 실시하였다. 같은 자료가 통계적 방법론의 적용에 쓰인 예로는 최근 Wu와 Tian (2013)은 혈압을 반응변수로 하고 나이, 인종, 키를 공변량으로 하여 일변량 조건부 분포를 추정하였다. 기존의 분석 결과들은 모두 단면적 시점에서의 자료를 분석한 것이라는 한계가 있다. 이에 우리는 이변량 콜레스테롤 자료에 대하여 결합 조건부 분포를 추정하였다.

처음 관측이 시작된 나이는 대략 9살이며 전체 관측시점의 나이 범위를 [9, 19)로 간주하여 같은 간격으로 이루어진 $J = 20$개의 나이 구간들, 즉 $[9.0, 9.5), \ldots, [18.5, 19.0)$을 생각하였다. 시점 t에서 얻어진 이변량 반응변수 (HDL, LDL)를 $(Y^1(t), Y^2(t))$로 나타내었고 $\mathbf{X}(t) = (X_1, X_2(t))$는 공변량을 나타낸다. 여기서 인종을 X_1(1 = 백인, 2 = 흑인), t시점에서의 BMI 백분위수를 $X_2(t)$라고 하였다. 분석에 사용된 총 관측치의 개수는 6697개 이다. 자료의 치우침 정도를 보정하기 위하여 반응 변수에 로그 변환을 취하여 HDL, LDL 대신에 log(HDL), log(LDL)을 사용하였다.

3.1. 조건부 주변확률 추정

우리는 $(Y^1(t), Y^2(t)) = (\text{HDL}, \text{LDL})$각각의 반응변수에 대한 조건부 주변 분포 $F_{kt}, k = 1, 2$를 추정하기 위하여 다음의 비례 오즈 모형을 사용하였다.

$$-\log\left[\frac{1 - F_{kt}(y|X_1, X_2(t))}{F_{kt}(y|X_1, X_2(t))}\right] = h(y, t) + \beta_1(t)X_1 + b_2(t)X_2(t)$$

Figure 3.1은 공변량이 $x_1 = 1, 2$, $x_2(t) = 5, 25, 50, 75, 95$로 주어졌을 각각의 경우에 $\widehat{F}_{1t}(60)$과 $\widehat{F}_{2t}(110)$을 나타낸 것이다. 그림에서 추정된 {HDL < 60}인 조건부 확률은 나이의 증가에 따라 거의 변화하지 않음을 볼 수 있는 반면, 추정된 {LDL < 110}인 조건부 확률은 나이가 14세 일때까지 꾸준히 증가하다가 그 이후로는 감소한 경향을 볼 수 있다. BMI 백분위수가 높은 경우 {HDL < 60}인 조건부 확률이 높아지는 반면, {LDL < 110}인 조건부 확률이 낮아짐을 볼 수 있다. BMI 백분위수가 같은 경우, {HDL < 60}인 조건부 확률은 백인이 흑인에 비해 약간 높은 것을 볼 수 있다. BMI 백분위수가 같은 경우, {LDL < 110}인 조건부 확률은 나이가 10세 전후로는 흑인이 백인에 비해 약간 높은 것을 볼 수 있으나 나이가 증가함에따라 그 차이는 거의 없는 것으로 보인다.

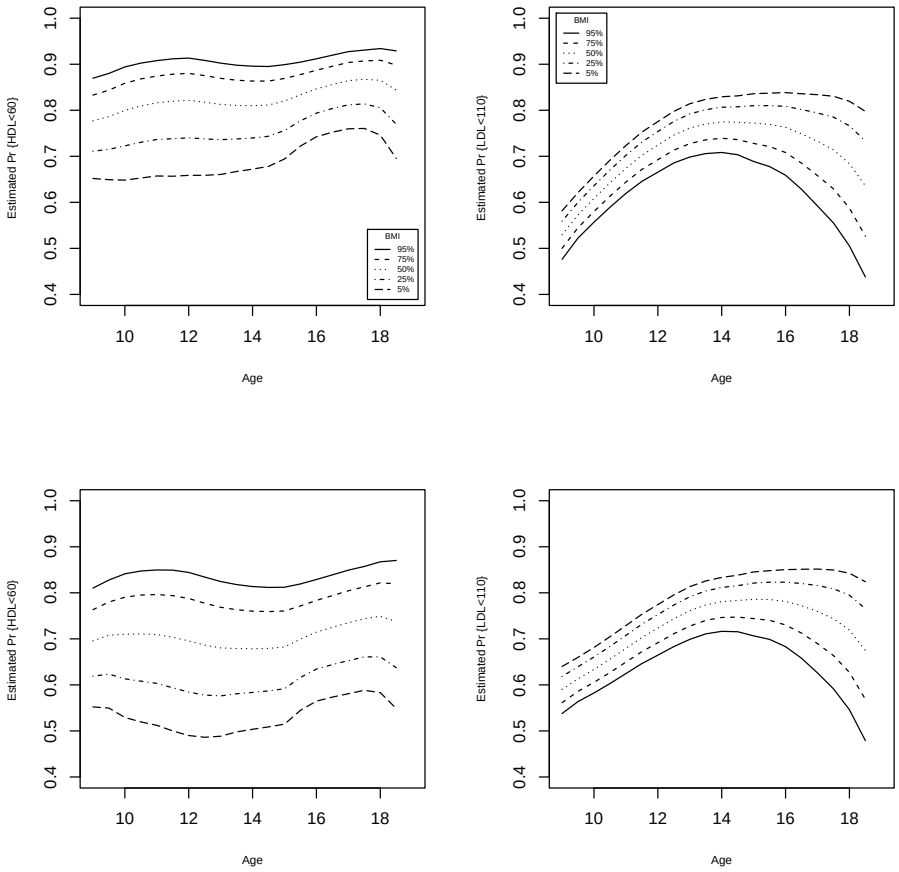

Figure 3.1 Estimated marginal conditional distribution of HDL and LDL of Caucassian and African-American girls for each BMI percentile, respectively. Upper-left panel shows the estimated conditional probability of {HDL < 60} for Caucasian girls, upper-right panel shows the estimated conditional probability of {LDL < 110} for Caucasian girls, lower-left panel shows the estimated conditional probability of {HDL < 60} for African-American girls, lower-right panel shows the estimated conditional probability of {LDL < 110} for African-American girls for each five BMI percentiles, respectively.

3.2. 조건부 결합확률 추정

이변량 반응변수 $(Y^1(t), Y^2(t))$ =(HDL, LDL) 에 대하여 $Y^1(t)$와 $Y^2(t)$의 상관성을 비모수적 코플라 (nonparametric copula) 함수를 이용하여 조건부 결합 분포를 추정하였다. 먼저 두 변수 간의 상관성이 존재하는지를 살펴보기 위하여 분위수 변환된 조건부 주변 분포의 상관계수를 계산하였다. Figure 3.2는 백인 여아와 흑인 여아 각각에 대하여 나이에 따른 $\widehat{F}_{1t}\{y_1|x\}$과 $\widehat{F}_{2t}\{y_2|x\}$의 상관 계수를 그린 그림이다. 나이에 따른 상관 계수의 경향을 파악하기 위하여 평활된 곡선을 함께 나타내었다. 백인의 경우 9세에서 약 0.2인 상관계수가 19세 즈음에 약 0.5로 나이가 증가함에 따라 상관계수가 증가하는 경향을 보이며, 흑인의 경우 9세에서 약 0.15인 상관계수가 19세 즈음에 약 0.4로 역시 나이가 증가함에 따라 상관계수가 증가하는 경향을 보인다.

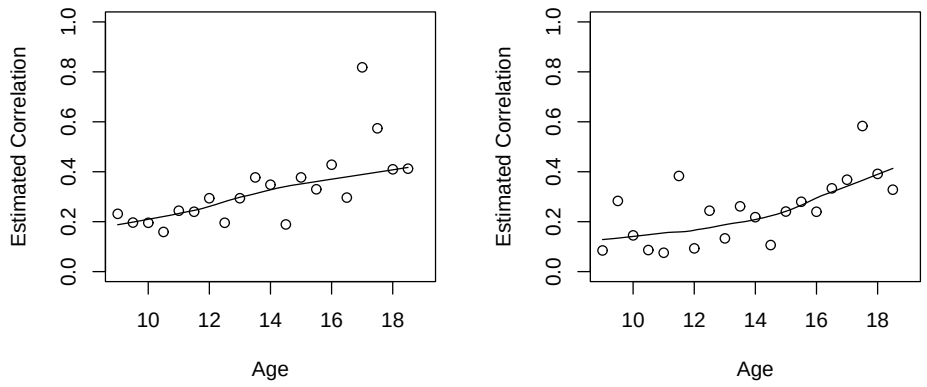

Figure 3.2 Plot of correlation between quantile transformed conditional marginal distribution for $Y^1(t)$ and $Y^2(t)$ over age for Caucasian and African-American girls, respectively. Left panel shows the plot of correlation for Caucasian girls and right panel shows the plot of correlation for African-American girls. Smoothed curves are overlapped in each plot.

다음으로 조건부 결합 확률 $\Pr\{Y^1(t) \le y_1(t), Y^2(t) \le y_2(t)|X = x\}$을 추정하기 위하여 우리는 3.1 절에서 $Y^1(t)$, $Y^2(t)$각각 일변량 반응변수에 대한 조건부 주변 분포 F_{1t}, F_{2t}에 대한 추정량 $\widehat{F}_{1t}$, $\widehat{F}_{2t}$를 얻었다. 이렇게 얻은 $\widehat{F}_{1t}$, $\widehat{F}_{2t}$각각에 대하여 이변량 경험 분포함수를 이용한 비모수적 코플라를 이용하여 추정된 두 개의 조건부 주변 분포를 결합하였다. 이를 식으로 나타내면 다음과 같다.

$$\widehat{J}_{nt}(y_1, y_2|x) = \frac{1}{n}\sum_{i=1}^{n} I\left\{\widehat{F}_{1t}(Y_i^1(t)|x) \le \widehat{F}_{1t}(y_1(t)|x), \widehat{F}_{2t}(Y_i^2(t)|x) \le \widehat{F}_{2t}(y_2(t)|x)\right\}$$

Figure 3.3과 3.4는 두가지의 BMI 백분위 수 $x_2(t) = 50, 75$와 인종별로 비모수적 코플라 함수를 이용하여 추정된 조건부 결합 확률 $\widehat{\Pr}\{\text{HDL}(t) < 60\text{mg/dl}, \text{LDL}(t) < 110\text{mg/dl}\}$을 그린 것이다. 각각의 그림에서 실선은 BMI 백분위수 50% 에 해당하는 추정치와 신뢰구간을 나타내며 점선은 BMI 백분위수 75% 에 해당하는 추정치와 신뢰구간을 나타낸다. 실제로 추정된 확률들의 지그재그 형태의 변동을 부드럽게 나타내기 위하여 가우시안 커널 $K_h(t_j, t) = exp(-(t_j - t)^2/(2h))$을 이용한 평활을 실시하여 부드러운 곡선형태로 표시하였다. 띠너비 (bandwidth) 는 교차 검증법 (cross validation)을 통하여 $h = 1.5$를 사용하였다. Figure 3.3은 백인 여아에 대하여 추정된 조건부 결합 확률 $\widehat{\Pr}\{\text{HDL}(t) < 60\text{mg/dl}, \text{LDL}(t) < 110\text{mg/dl}\}$을 BMI 백분위수 50%와 75%에 대하여 조건부 결합 확률 추정치와 95% 붓스트랩 신뢰구간을 그린 것이다. 붓스트랩 반복 횟수는 200회로 하였다. 백인 여아의 경우 BMI 백분위수가 50%인 경우 조건부 결합 확률이 전체적으로 완만하게 감소하는 경향을 보이며, BMI 백분위수가 75%인 경우 조건부 결합 확률이 성장 초기에는 일정기간 변화가 거의 없다가 13세 이후 완만하게 증가하는 경향을 보인다. Figure 3.4는 흑인 여아에 대하여 추정된 조건부 결합 확률 $\widehat{\Pr}\{\text{HDL}(t) < 60\text{mg/dl}, \text{LDL}(t) < 110\text{mg/dl}\}$을 BMI 백분위수 50% 와 75% 에 대하여 조건부 결합 확률 추정치와 반복횟수 200회를 통한 95% 붓스트랩 신뢰구간을 그린 것이다. 흑인 여아의 경우 BMI 백분위수가 50% 인 경우와 75%인 경우 모두 조건부 결합 확률이 전체적으로 완만하게 증가하는 경향을 보인다. 같은 BMI 백분위수에 대하여 백인과 흑인을 비교하면, 추정된 조건부 결합확률 $\widehat{\Pr}\{\text{HDL}(t) < 60\text{mg/dl}, \text{LDL}(t) < 110\text{mg/dl}\}$은 백인의 경우보다 흑인의 경우에 더 높음을 알 수 있다.

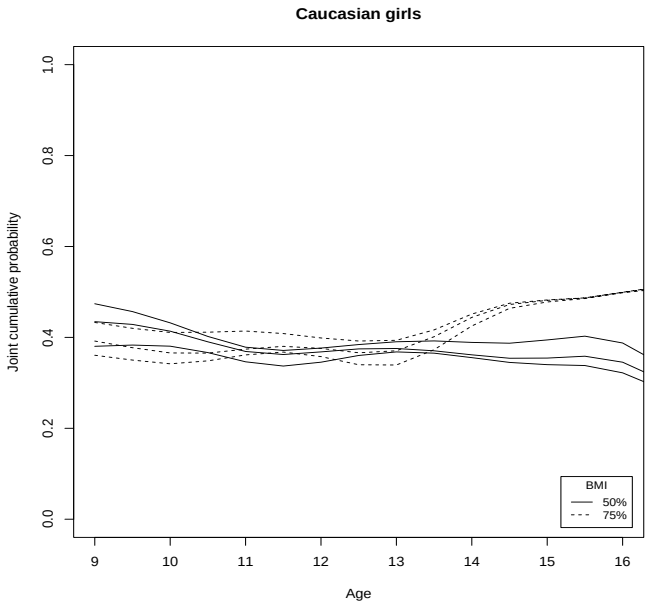

Figure 3.3 The estimated conditional joint probabilities $\widehat{\Pr}\{\text{HDL(t)} < 60\text{mg/dl}, \text{LDL(t)} < 110\text{mg/dl}\}$ and pointwise 95% bootstrap confidence interval for Caucasian ($x_1 = 1$) girls for two selected BMI percentiles.

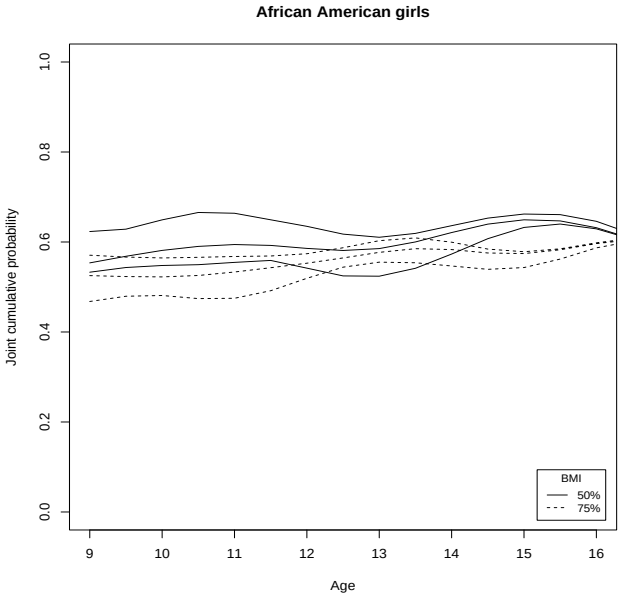

Figure 3.4 The estimated conditional joint probabilities $\widehat{\Pr}\{\text{HDL(t)} < 60\text{mg/dl}, \text{LDL(t)} < 110\text{mg/dl}\}$ and pointwise 95% bootstrap confidence interval for African-American ($x_1 = 2$) girls for two selected BMI percentiles.

3.3. 모의 실험

실제로 우리가 제안한 방법이 참값을 제대로 추정하는지 알아보기 위하여 모의실험을 실시하였다. 우리는 이변량 반응변수 $Y(t) = (Y^1(t), Y^2(t))$에 대하여, 일변량 반응변수 Y^1과 Y^2각각에 대한 주변 분포로 공변량 $\mathbf{X}(t) = (X^1, X^2(t))^T$을 이용한 비례 오즈 모형 $g_k[S_{kt}(y_k|x)] = -\log[S_{kt}(y_k|x)/F_{kt}(y_k|x)]$, $t \in [0,1]$, $k = 1,2$을 고려하였다. 공변량 X^1은 성공 확률이 0.5인 베르누이 분포에서 난수를 생성하였고, 공변량 X^2는 $100X^2(t) \sim U(15,85) + N(0,16)$에서 난수를 생성하였다. 주변 분포의 회귀 모수는 $\beta_1(t) = t^2 - 1.7t - 0.2$과 $\beta^2(t) = -1 - 0.5\cos(1.1\pi t)$로 설정하였고, 기저함수는 $h(y_k,t) = -log\,[\{1 - \Phi(y_k|t)\}/\Phi(y_k|t)], k = 1,2$로 설정하였다. 여기서 $\Phi(y_1|t)$로는 평균인 98+10t이고 표준편차가 5인 정규분포의 누적분포함수를 사용하였으며, $\Phi(y_2|t)$로는 평균인 78+10t이고 표준편차가 5인 정규분포의 누적분포함수를 사용하였다. 개별 모의실험에서 표본의 크기는 n= 1200으로 하였으며 각 개체당 관측횟수는 10번으로 설정하였다.

i번째 개체의 l번째 관측 시점을 t^{il}로 표시하면, 각 i번째 개체에 대하여 관측시점 t^{il}는 균등분포 $U[(l-1)/10, l/10]$ for $l = 1, \ldots, 10$에서 난수를 발생시켜 얻었으며 관측시점 t^{il}에서 공변량 $\mathbf{X}_i(t)$을 생성하였다. 개체 내의 상관성을 고려하기 위하여 오차항에 대한 분포로 $\epsilon(t^{il}) \sim \epsilon_{i0} I(D = 1) + L(0,1)I(D = 0)$을 고려하였다. 여기서 D는 성공 확률이 0.5인 베르누이 분포이고, $\epsilon_{i0} \sim L(0,1)$이며 $L(0,1)$는 표준 로지스틱 분포를 나타낸다. 관측시점 t^{il}에서 반응변수 $Y^1(t^{il})$과 $Y^2(t^{il})$의 상관성을 고려하기 위하여 상관계수가 $\rho(t^{il})$로 주어지는 가우시안 코플라 함수를 이용하였다. 모의 실험에 이용된 이변량 반응 변수 자료 $(Y_i^1(t^{il}), Y_i^2(t^{il}))$는 식 (2.1)에 $\boldsymbol{\beta}(t^{il})$, $\mathbf{X}_i(t^{il})$과 $\epsilon(t^{il})$를 대입하여 얻어내었다.

Figure 3.5는 시간의 변화에 때른 결합 조건부 확률에 대한 참값과 추정값, 그리고 95% 붓스트랩 신뢰구간을 나타낸다. 이 그림은 500번의 모의실험에 기초한 것이다. 참값은 실선으로 나타내었으며 추정값과 그 신뢰구간은 점선으로 표시하였으며 그림에서 볼 수 있듯이 추정값이 참값과 가까움을 확인할 수 있었다.

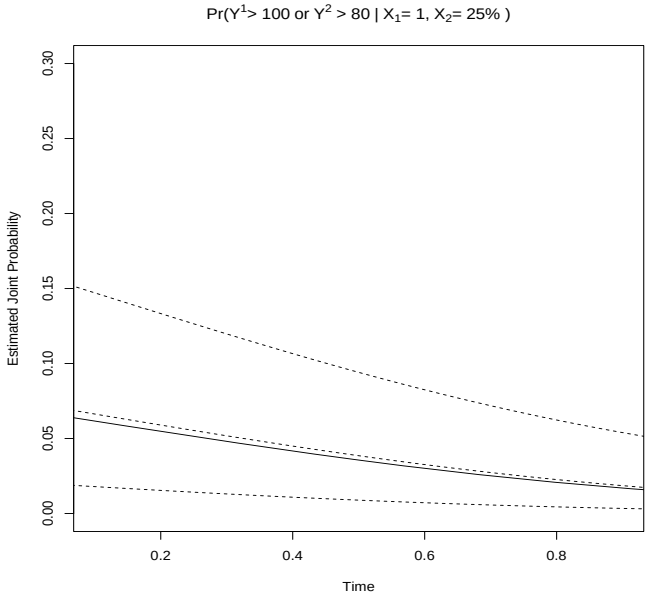

Figure 3.5 The true and estimated joint probabilities $\Pr\{Y^1 > 100 \text{ or } Y^2 > 80 | X_1(t) = 1,\ X_2(t) = 25\%\}$ using Gaussian copula. The solid line represents the true probabilities and the dashed lines represent the mean, the pointwise lower and upper 2.5% percentiles of the estimated probabilities.

4. 결론

본 논문에서는 시간에 따라 반복 측정된 이변량 콜레스테롤 자료를 바탕으로 일변량 각각에 대해서 준모수적인 조건부 주변 분포를 구하고, 이렇게 구한 두 개의 조건부 주변 분포를 비모수적인 이변량 경험함수를 이용하여 자료를 설명하는 적절한 모형을 찾아내어 이변량 자료의 상관성을 고려한 조건부 결합확률을 추정하였다. 각각의 일변량 자료에 대하여 시변 전환 모형을 고려함으로써 주변 분포에 대한 선택의 폭이 넓어진다는 장점이 있으며, 첫번째 일변량 반응변수와 두번째 일변량 반응변수에 서로 다른 주변 회귀 모형을 적용할 수 있다는 장점도 있다. 두개의 일변량 조건부 분포를 결합하기 위하여 경험분포를 이용하였는데, 이는 두 반응변수간의 상관성을 설명하기 위하여 모수적인 함수 형태를 가정하지 않으므로 실제 특정한 모형을 가정하기 힘든 일반적인 상황에서 유용하게 쓰일수 있다는 장점이 있다.

본 논문에서는 반복 측정된 이변량 반응변수의 경우를 생각하였으나, 일반적으로 M-변량 반응변수로 확장할 수 있다. 즉, 조건부 결합 분포를 구하기 위하여 M-차원 경험 분포를 이용할 수 있다. 모수적 코플라의 대표적인 예인 가우시안 코플라의 경우 상관계수 행렬을 통해 다변량 변수들간의 상관성을 쉽게 해석할 수 있는 반면 경험 분포를 이용한 다변량 조건부 결합 분포에서는 변수들간의 상관성을 나타내는 측도를 간편하게 나타낼 수 없다는 단점이 있다. 이는 모형에 대한 어떠한 가정도 하지 않은 비모수적 방법에 대한 일장일단 (trade-off)으로 볼 수 있다. 또한 차원이나 반복 시점의 갯수가 커지면서 각 시점에서 측정되는 자료의 갯수가 적어질 경우에는 안정적인 추정치를 구하기 어렵다는 점을 염두에 두어야 한다. 다음 연구에서는 본 논문의 결과를 바탕으로 추정량의 편향 정도를 유도하고 부드러운 곡선을 구하기 위한 서로 다른 평활방법의 비교 분석등을 고려하고자 한다.

본 논문에서는 각 시점에서 두 개의 반응변수간의 연관성을 코플라 함수를 통하여 설명하고자 하였다. 경시적 자료의 특성상 시간에 따라 반복 측정된 경시적 자료의 특성을 고려할 때 자료의 시점간 연관성을 모형화 하는 것을 다음 연구 주제로 고려할 수 있다.

21세기 인류의 평균수명은 100세를 넘을 것이나 건강수명은 79세 밖에 되지 않아 약 20년을 심혈관계 질환 등과 같은 만성질환으로 고통 받을 것으로 추정되고 있다. 우리나라 통계청이 발표한 2007년 발표한 보고서에 의하면 남성의 평균 수명이 75.1세, 여성이 82.3세로 2002년에 비해 각각 1.7세와 1.9세가 높아졌으며 평균적으로도 1.7세가 연장된 것으로 나타났다. 이에 따른 사망원인도 변화하여 생활습관성 질환의 사망률 비율은 점차 증가 추세로 40대까지는 심장질환으로 인한 사망률이 증가하며, 50대 이후로는 뇌혈관계질환으로 인한 사망률이 지속적인 증가세를 보이는 것으로 나타나 심혈관계 질환의 심각성이 부각되고 있다. 우리나라는 최근 인구의 고령화와 생활양식의 변화로 심혈관 질환이 급격히 증가하고 있는 추세이다. 특히 우리나라 30세 이상에서 고혈압과 고혈압 전기에 해당하는 비율이 58.5%에 달함에도 30~40대 성인 중 고혈압 환자는 대부분 본인이 환자라는 사실도 인지하지 못하고 있으며, 약물치료로 혈압과 콜레스테롤을 적정 수준으로 유지하고 있는 환자 비율은 전체 환자 3명 중 1명꼴에 그치고 있다. 본 연구에서 얻은 결과를 바탕으로 코플라를 이용한 결합 분포의 추정은 10년간 경시적으로 얻어진 이변량 콜레스테롤 심혈관 위험요인 자료를 가지고 다양한 이변량 경시적 자료 분석 모형을 개발하는데 도움이 될 것이라 생각한다.

References

Anderson, K. M., Castelli, W. P. and Levy, D. (1987). Cholesterol and mortality. 30 years of follow-up from the Framingham study. *Journal of American Medical Association*, **257**, 2176-2180.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, **82**, 835-845.

Cho, G. Y. and Dashnyam, O. (2013). Generalized methods of moments in marginal models for longitudinal data with time-dependent covariates. *Journal of the Korean Data & Information Science Society*, **24**, 877-883.

Daniels, S. R., McMahon, R .P., Obarzanek, E., Waclawiw, M. A., Similo, S. L., Biro, F. M., Schreiber, G. B., Kimm, S. Y., Morrison, J. A. and Barton, B. A. (1998). Longitudinal correlates of change in blood pressure in adolescent girls. *Hypertension*, **31**, 97-103.

Diggle, P. J., Liang, K. Y. and Zeger S. L. (1994). *Analysis of longitudinal data*, Oxford University Press, Oxford.

Genest, C., Ghoudi, K. and Rivest, L. P. (1995). A semiparametric estimation procedures of dependence parameters in multivariate families of distributions. *Biometrika*, **82**, 543-552.

Genest, C. and MacKay, J. (1986). A joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, **40**, 280-283.

Jeon J. Y. and Lee K. (2014) Review and discussion of marginalized random effects models. *Journal of the Korean Data & Information Science Society*, **25**, 1263-1272.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis*, **46**, 262-282.

Leon, A.R. and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, **30**, 175-185.

Lindsey, J. K. (1993). *Models for repeated measurements*, Oxford University Press, Oxford.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer, New York.

National Heart, Lung, and Blood Institute Growth and Health Research Group (NGHSRG) (1992). Obesity and cardiovascular disease risk factors in black and white girls: The NHLBI growth and health study. *Americal Journal of Public Health*, **82**, 1613-1620.

National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents (NHBPEP Working Group) (2004). The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. *Pediatrics*, **114**, 555-576.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika*, **73**, 353-361.

Obarzanek, E., Wu, C. O., Cutler, J. A., Kavey, R. W., Pearson, R. W. and Daniels, S. R. (2010). Prevalence and incidence of hypertension in adolescent girls. *Journal of Pediatrics*, **157**, 461-467.

Sklar, A. (1959). Fonctions de répartition àn dimensions et leurs marges. *Publications de L'Institute de Statistique de L'Université de Paris*, **8**, 229-231.

Song, P. X. K., Li, M. and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian Copulas. *Biometrics*, **65**, 60-68.

Thompson, D. R., Obarzanek, E., Franko, D. L., Barton, B. A., Morrison, J., Biro, F. M., Daniels, S. R. and Striegel-Moore, R. H. (2007). Childhood overweight and cardiovascular disease risk factors: The national heart, lung, and blood institute growth and health study. *Journal of Pediatrics*, **150**, 18-25.

Wu, C. O. and Tian, X. (2013). Nonparametric estimation of conditional distribution functions and rank-tracking probabilities with time-varying transformation models in longitudinal studies. *Journal of the American Statistical Association*, **108**, 971-982.

Journal of the Korean Data & Information Science Society
2016, **27**(3), 689–700

http://dx.doi.org/10.7465/jkdi.2016.27.3.689
한국데이터정보과학회지

Estimation of the joint conditional distribution for repeatedly measured bivariate cholesterol data using nonparametric copula[†]

Minjung Kwak[1]

[1]Department of Statistics, Yeungnam University

Received 15 April 2016, revised 9 May 2016, accepted 19 May 2016

Abstract

We study estimation and inference of the joint conditional distributions of bivariate longitudinal outcomes using regression models and copulas. For the estimation of marginal models we consider a class of time-varying transformation models and combine the two marginal models using nonparametric empirical copulas. Regression parameters in the transformation model can be obtained as the solution of estimating equations and our models and estimation method can be applied in many situations where the conditional mean-based models are not good enough. Nonparametric copulas combined with time-varying transformation models may allow quite flexible modeling for the joint conditional distributions for bivariate longitudinal data. We apply our method to an epidemiological study of repeatedly measured bivariate cholesterol data.

Keywords: Bivariate longitudinal data, empirical copula, joint conditional distribution, time-varying transformation models.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2014R1A1A1002465).

[1] Professor, Department of Statistics, Yeungnam University, Gyeongbuk 38541, Korea.
E-mail: mjkwak@yu.ac.kr