creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

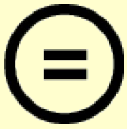
이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

연령-기간-코호트(APC) 모형의 추정 방법론 비교

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

박 세 영

연령-기간-코호트(APC) 모형의 추정 방법론 비교

지도교수 남 정 모

이 논문을 석사 학위논문으로 제출함

2015년 12월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

박 세 영

박세영의 석사 학위논문을 인준함

심사위원 남 정 모 인

심사위원 정 인 경 인

심사위원 박 소 희 인

연세대학교 대학원

2015년 12월 일

차 례

표 차 례

그 림 차 례

국 문 요 약

연령-기간-코호트(APC) 모형의 추정 방법론 비교

역학 분야에서 인구 집단이나 각 개인에게 환경적, 역사적 그리고 사회적 요인들이 복잡하게 미치는 영향을 묘사하기 위해 연령-기간-코호트(Age-Period-Cohort, APC) 분석을 한다. 그러므로 APC 분석은 사회나 인구집단 변화와 특정 질병의 인과 관계 파악에 대한 중요한 연구 문제를 지속적으로 해결하기 위해 널리 사용되어 왔다. 이에 따라 APC 모형은 연령, 기간, 출생 코호트에 따른 질병의 발생률이나 사망률의 경향성을 파악하는 역할을 한다. 그러나 APC 모형을 적합 시킬 때 연령, 기간, 코호트 효과가 결과 변수와 선형적 관계에 있다고 할 수 있다. 이러한 경우 추정에 대한 식별(identification)의 문제가 발생한다. 세 변수가 완전한 선형관계를 가지기 때문에 공선성이 존재하여 연령, 기간, 코호트 효과를 동시에 추정할 수 없는 문제점이 있다.

본 논문에서는 APC 모형을 구축할 때 식별의 문제를 해결할 수 있는 방법들 중에서 역학 분야에서 주로 사용되는 4가지 접근 방법을 이용하여 모의실험을 통해 가상의 상황을 구상하고 상황에 따라 각 방법의 장점과 한계점을 언급하여 APC 모형 구축 시 여러 방법들에 대해 가이드라인을 제시한다. 모의실험의 결과를 통해 제약된 일반화선형모형 보다는 IE 방법과 주성분 분석을 사용하는 것이 각각의 효과가 미치는 영향을 정확하게 볼 수 있다고 판단된다. 질병에 따라 어떤 분석 방법을 적용해야 정확하게 추정할 수 있는지 달라질 수 있지만 본 논문의 모의실험을 바탕으로 APC 분석에 있어서 다음과 같이 간단한 가이드라인을 추천해주고자 한다. 먼저 연령-기간-코호트 분석 이전에 기간과 코호트의 효과가 유의한지 검정한다. 만약 코호트의 효과가 존재하는 경우 IE 방법, 주성분 분석, Median Polish 방법을 만약 코호트의 효과가 존재하지 않는 경우 가장 안정적인 IE 방법과 주성분 분석을 추천한다.

핵심이 되는 말 : APC 모형, 코호트 효과, CGLM, Median Polish, IE, PCA

제 1장 서론

1.1 연구 배경

역학 분야에서 인구 집단이나 각 개인에게 환경적, 역사적 그리고 사회적 요인들이 복잡하게 미치는 영향을 묘사하기 위해 연령-기간-코호트(Age-Period-Cohort, APC) 분석을 한다. 그러므로 APC 분석은 사회나 인구집단 변화와 특정 질병의 인과 관계 파악에 대한 중요한 연구 문제를 지속적으로 해결하기 위해 널리 사용되어 왔다. 이에 따라 APC 모형은 연령, 기간, 출생 코호트에 따른 질병의 발생률이나 사망률의 경향성을 파악하는 역할을 한다. APC 모형 구축의 주요 목적은 질병의 발생률에서 각 요인들의 효과를 평가하는 것이다. 연령 효과는 특정 나이에 관계된 효과로 나이가 들어감에 따라 어떠한 원인에 대한 노출이 축적되거나 생리학적 변화로 인한 효과라고 할 수 있다. 대부분의 경우 동일한 코호트 내에서 시간이 지남에 따라 연령 효과에 의해 질병에 걸리는 비율이 높아지는 경향을 보인다. 기간 효과는 특정 시점에서 널리 퍼져있는 환경의 변화에 따라 발생률이 연령에 관계없이 차이가 나게 되는 효과이다. 예를 들어, 특정 시점에서 병에 걸리는 비율이 증가하는 전염병 같은 경우가 있다. 출생코호트 효과는 태어난 시점에서 따른 효과이다. 같은 출생 코호트에 속한 사람들끼리 시간이 지남에 따라 특정 질병에 대해 유사한 위험을 갖는다고 볼 수 있다. 예를 들어, 특정 시점에서 전 연령에서 질병에 걸리는 분율이 크게 증가하는 전염병이 발생했다면 이를 기간 효과로 보고 특정 연령에서 특히 질병에 더 많이 노출된 경우 이를 코호트 효과로 볼 수 있다. 위의 효과들을 고려해 APC 모형을 적합 시킬 때 연령, 기간, 코호트 효과가 결과 변수와 선형적 관계에 있다고 할 수 있다. 이러한 경우 추

정에 대한 식별(identification)의 문제가 발생한다. 세 변수가 코호트 = 기간 - 연령(또는, 기간 = 연령 + 코호트)라는 완전한 선형관계를 가지기 때문에 공선성이 존재하여 연령, 기간, 코호트 효과를 동시에 추정할 수 없는 문제점이 있다. 이러한 문제점으로 지난 20년간 추정을 위해 여러 가지 해결방안들이 제안되었다(Osmond and Gardner 1982; Clayton and Schifflers 1987; Holford 1992; Tarone and Chu 1992; Robertson and Boyle 1998; Fu 2000; Knight and Fu 2000). 그러나 현재 사용하는 방법들은 여전히 제한점을 가지고 있다고 이의를 제기 받았다. 결과적으로 현존하는 식별문제 해결방법도 완벽한 해결방법이라고 할 수 없는 것이다.

최근 의학통계학(Biostatistics) 분야에서 APC 방법론에 대해 모수를 제약하여 식별문제를 해결하는 것보다 연령, 기간 그리고 출생코호트 효과에 대해 변동 없이 추정할 수 있는 방법론 사용해야한다고 강조하고 있다(Clayton and Schifflers 1987; Holford 1983, 1991, 1992; Robertson, Gandini, and Boyle 1999; Tarone and Chu 1992, 2000).

그러므로 본 논문에서 APC 모형의 식별문제를 해결하기 위한 여러 가지 접근방법들을 간략히 설명하고 모의실험을 통해 상황 별로 적절한 식별 문제 해결방안을 제안하고자 한다.

1.2 연구 목적 및 방법

본 논문에서는 연령-기간-코호트 연구 모형을 구축할 때 식별의 문제를 해결할 수 있는 방법들 중에서 역학 분야에서 주로 사용되는 4가지 접근 방법(Constrained Generalized Model, Median Polish Method, Intrinsic Estimator, Principal Component Analysis)의 방법을 간략하게 소개하고, 모

의실험을 통해 가상의 상황을 구상하여 상황에 따라 각 방법의 장점과 한계점을 언급하여 연령-기간-코호트 모형 구축 시 여러 방법들에 대해 가이드라인을 제시해주고 한다. 또한 추가적으로 실제 자료에 적용하여 각 방법들의 장단점을 알아보고 각각의 추정량을 비교해본다.

1.3 논문의 구성

제 1장에서는 연구의 배경과 목적 및 내용을 소개한다. 2장에서는 APC 모형과 식별문제를 해결할 수 있는 4가지 방법에 대해 이론적 배경과 개념을 간략하게 정리한다. 제 3장에서는 모의실험을 통해 앞에서 소개한 4가지 방법을 비교한다. 제 4장에서는 실제 자료를 이용하여 식별 문제를 해결하고 4가지 방법에 대한 결과들을 비교한다. 마지막으로 5장에서는 결론 및 고찰에 대해 논의한다.

제 2장 이론적 배경

이 장에서는 연령-기간-코호트 모형(Age-Period-Cohort model)의 식별문제(identification problem)를 해결하기 이전에 연령-기간-코호트 분석에 일반적으로 사용되는 일반화선형모형과 식별문제에 대해 간략하게 소개하고 모형의 식별문제를 해결하기 위한 여러 가지 방법들 중 본 논문에서 소개하는 4가지 방법의 개념을 간략하게 정리한다.

2.1 연령-기간-코호트 분석(Age-Period-Cohort Analysis)

연령-기간-코호트(Age-Period-Cohort; APC) 모형은 인구학, 역학 등 여러 분야의 사회 연구를 위해 Mason(1973)에 의해 연계되었다. 특정 사건의 발생률에 대해 APC 모형은 아래와 같이 선형 회귀 형식으로 표현할 수 있다.

$$Y_{ij} = \frac{D_{ij}}{N_{ij}} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ij} \qquad \cdots (1)$$

$$(i = 1,2, \cdots, a;\ j = 1,2, \cdots, p;\ k = 1,2, \cdots, a+p-1)$$

위의 모형에서 Y_{ij} 는 i 번째 연령의 j 번째 기간 그룹에서 관측된 자료의 발생률을 의미한다. D_{ij}는 i 번째 연령의 j 번째 기간 그룹에서 발생한 사건 수, N_{ij}는 i 번째 연령의 j 번째 기간 그룹에서 추정된 인구 집단의 크기를 의미한다. μ 는 절편 또는 평균 발생률, α_i 는 i 번째 연령 그룹의 효과 또는 계수, β_j 는 j 번째 기간 그룹의 효과 또는 계수, γ_k 는 k 번째 출생 코호트 그룹의 효과 또는 계수로 출생 코호트 효과는 연령과 기간의 대각 행렬로 정

의되므로 코호트의 그룹 수는 $a+p-1$로 정의되며 k는 $a-i+j$ 이다. ϵ_{ij} 는 i 번째 연령 효과와 j 번째 기간 효과로 설명되지 않는 임의의 오차항이다.

어떤 형태의 사건이 시간별 혹은 공간별로 또는 크기를 나타내는 다른 지표에 걸쳐 발생할 때에는 사건 발생률에 대한 모형을 설정하는 것이 바람직하다. 즉, APC 분석과 같이 특정 사건의 발생률 자료를 사용하여 모형을 적합하는 경우 (1)번 모형을 아래와 같이 로그 연결함수(log link function)을 사용하여 일반선형회귀모형(Generalized Linear Model; GLM)으로 표현할 수 있다.

$$\log(D_{ij}) = \log(N_{ij}) + \mu + \alpha_i + \beta_j + \gamma_k \quad \cdots(2)$$

(2)번 모형에서 D_{ij} 는 i 번째 연령의 j 번째 기간인 셀 (i,j)의 기대 도수를 의미하며, 각 셀의 기대도수는 포아송(Poisson) 분포를 따른다. 위의 모형에서 $\log(N_{ij})$은 평균의 로그 연결함수 값 $\log(D_{ij})$에 대한 수정항으로 옵셋(offset)이라고 부른다.

위의 회귀모형들은 아래의 수식과 같이 모든 모수들에 대해 중심으로 재매개변수화(reparameterization to center)한 후 모형을 적합 시킨다. 즉, 각 효과 계수들의 합을 0으로 설정한다는 의미이다.

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{p} \beta_j = \sum_{k=1}^{a+p-1} \gamma_k = 0 \quad \cdots(3)$$

위와 같이 재매개변수화(reparameterization) 시킨 후 (1)과 (2)번 모형의 각 효과인 계수들을 계획행렬(design matrix)을 사용하여 아래와 같이 다시

표현할 수 있다.

$$Y = Xb + \epsilon \qquad \cdots(4)$$

행렬로 표현된 일반선형회귀모형에서 Y 는 발생률 또는 로그 변환된 발생률 벡터(vector)를 의미하며 X 는 모형에서 3가지 모수인 연령, 기간, 출생 코호트 효과를 가변수(dummy variable)로 구성하여 각 모수의 효과들을 열벡터(column vector)로 표현한 계획 행렬이다. 그러므로 계획행렬 X 의 차원(dimension)은 $1+(a-1)+(p-1)+(a+p-2)$ 이고 모형의 계획 행렬은 아래와 같이 표현된다.

$$b = (\mu, \alpha_1, \cdots, \alpha_{a-1}, \beta_1, \cdots, \beta_{p-1}, \gamma_1, \cdots, \gamma_{a+p-2})^T \qquad \cdots(5)$$

ϵ 는 임의의 오차에 대한 벡터로 평균은 0이고 대각성분이 1, 비대각원소가 σ^2 로 구성된 상수행렬을 분산으로 가진다.

행렬로 표현된 (4)번 회귀모형에서 최소자승추정량(ordinary least squares estimator)을 구하는 정규방정식은 아래와 같다.

$$\hat{b} = (X^T X)^{-1} X^T Y \qquad \cdots(6)$$

2.2 APC 모형의 식별 문제(Identification Problem)

앞에서 설명한 바와 같이 질병의 발생률이나 인구 집단의 사망률에 대한 경향성을 파악하기 위하여 연령, 기간, 출생 코호트 3가지 요인을 일반화선형모

형으로 적합하여 (6)번 식을 이용해 각 요인의 효과를 추정한다. 그러나 이 효과들은 아래와 같이

코호트 = 기간 - 연령 (또는, 기간 = 연령 + 코호트)

라는 완전한 선형 관계를 가지기 때문에 공선성이 존재하여 모형에서 연령, 기간, 출생 코호트 효과를 동시에 추정할 수 없는 문제가 발생한다. 즉, 모형에서 효과를 추정하기 위해 계획행렬(design matrix) X 의 역행렬이 존재해야 한다. 따라서 X 의 역행렬이 존재하기 위해 X는 정칙행렬(regular matrix)이며 이것은 $|X| \neq 0$을 의미한다. 예를 들어, $n \times p$ 계획 행렬 X의 계수(rank)가 p 라고 가정 한 경우 X 의 계수가 열이나 행의 개수와 같을 때 행렬 X 는 최대 계수(full rank)를 가지는데 이 경우에만 X^TX 도 최대계수를 가지고, 따라서 역행렬이 존재한다. 이와 반대로 연령-기간-코호트 모형과 같이 설명변수들 간에 선형종속 관계가 존재하면 행렬 X 의 계수가 p 보다 작아지게 된다. 이 경우에는 X^TX의 역행렬이 존재하지 않게 되어 정규방정식의 유일한 해를 구할 수 없게 된다.

그러므로 (6)번 모형에서 계획행렬 X 의 계수가 최대 계수가 아닌 비정칙행렬(singular matrix)이기 때문에 $(X^TX)^{-1}$ 가 존재하지 않아 해를 추정할 수 없다. 이러한 문제점을 APC 분석의 모형 식별 문제(identification problem)라고 정의하였다(Kupper et al. 1985).

따라서 APC 모형에서 각 요인의 효과 추정을 위해 추가적인 제한이 필요하다. 다음 장에서 모형 식별 문제를 해결하기 위해 가장 많이 사용되는 4가지 방법의 개념에 대해 간략하게 정리하고 모의실험을 통해 여러 상황을 구성하여 상황별로 적절한 해결방안을 제안하고 가이드라인을 제시해주고자 한다.

2.3 식별 문제에 대한 여러 가지 해결 방안

2.3.1 제약된 일반화선형모형(Constrained Generalized Linear Model; CGLM)

Fienberg and Mason(1985)은 APC 분석 시 연령, 기간, 코호트 효과 중 최소한 가지 이상의 모수를 제약하는 방법을 제안했다. 제약된 일반선형회귀모형은 질병 발생률과 사망률의 경향성을 파악하기 위해 일반적으로 가장 널리 쓰이고 있는 방법으로 연령, 기간, 출생 코호트 효과 중 어떠한 변수를 제약하느냐가 가장 중요한 문제이다.

먼저 3가지 요인에 대해 모든 경우의 수를 고려하면 가능한 모형은 총 6가지이다. 6가지의 모형을 생성하여 제약할 요인을 찾아야 한다. 따라서 각 모형을 분석하여 모형의 이탈도(Deviance)를 구한다. 이 때 이탈도(Deviance) 값이 작을수록 질병 발생률에 영향을 많이 미치며 클수록 미치는 영향이 작으므로 이탈도(Deviance) 값이 큰 요인을 제약하고자 한다. 이 때 발생률이 포아송(Poisson) 분포를 따른다고 가정하고 로그 선형 모형을 통해 각 모형별로 편차 값을 얻을 수 있다.

예를 들어, 아래의 모형에서 제약할 요인을 기간 효과 β_j 라고 정의한다면

$$\log(D_{ij}) = \log(N_{ij}) + \mu + \alpha_i + \beta_j + \gamma_k \qquad \cdots(2)$$

연령별 기간에 대한 발생률 그래프를 통해 p개의 기간 효과들 간의 발생률이 가장 유사해 보이는 효과를 제약 하는 방법이다. 만약 그래프를 통해 첫 번째 기간과 두 번째 기간의 효과가 같다고($\beta_1 = \beta_2$) 정의한다면 기간 효과를 제약하는 일반화선형모형이 되는 것이다. 즉, 모형 적합 시 첫 번째 기간 효과의

계수와 두 번째 기간 효과의 계수가 동일하게 추정하기 때문에 '코호트 = 기간 - 연령' 이라는 선형 종속관계가 성립되지 않으므로 모형에서 계획 행렬 X 의 역행렬이 존재하게 되고 이에 따라 정규방정식을 통해 유일한 해를 구할 수 있게 된다.

이 방법은 코호트 효과가 발생률에 미치는 직접적인 영향을 볼 수 있고 사용이 편리하다는 장점이 있다. 하지만 학계에서 모수의 계수를 제약하는 접근 방식에 대해 많은 문제점을 지적하고 있다(Yang Y 2013; Keyes KM and Utz RL 2010). 먼저 제약할 요인을 찾기 위해서 사전 정보가 필요하지만 이러한 사전 정보를 찾기 쉽지 않고 여러 사건 발생에 대해 다양하지 않다(Mason and Wolfinger 2002). 또한 어떤 모수를 제약하는지에 따라 각 모수의 추정량 패턴이 매우 민감하게 변화하고, 이러한 부분은 여러 논문에서 증명되었다(Glenn 2005; Holford 1991; Kupper et al., 1985).

2.3.2 Median Polish 방법

Selvin(1996)이 제안한 APC 모형에서 Median Polish 방법은 출생 코호트 효과를 연령과 기간의 교호작용이라고 개념화하고 비모수적인 Median Polish 방법(Tukey, 1977)을 확장한 방법이다. 이 방법은 기존의 연령과 기간의 효과로부터 코호트 효과를 독립적으로 추정한다는 점에서 현존하는 다른 방법들과는 차이가 있지만 연령과 기간 효과의 비선형적인 부분을 수량화하여 코호트 효과로 정의한다.

일반적으로 Median Polish 방법은 먼저 연령과 기간 효과만 고려하여 발생률 Y_{ij} 에 대해 이요인 모형을 다음과 같이 구축한다.

$$\log(Y_{ij}) = \mu + \alpha_i + \beta_j + \epsilon_{ij} \ (i = 1,2, \cdots, a;\ j = 1,2, \cdots, p) \qquad \cdots(7)$$

여기서 α_i 는 a개의 연령 효과 중 i 번째 계수, β_j 는 p개의 기간 효과 중 j 번째 계수, ϵ_{ij} 는 i 번째 연령 효과와 j 번째 기간 효과로 설명되지 않는 오차항이다. 앞에서 소개한 연령과 기간 효과만 고려한 이요인 모형을 아래의 모형식과 같이 다시 표현할 수 있다.

$$\log(Y_{ij}) = \mu + \alpha_i + \beta_j + \delta_{ij} \quad \cdots(8)$$

여기서 δ_{ij} 는 연령과 기간효과 사이의 교호작용을 의미한다. 위의 모형에서 교호작용 δ_{ij}의 분산은 두 가지 부분으로 나누어진다. 먼저 체계적인 부분(systematic component)이 코호트 효과로 정의되며 나머지 부분이 임의의 오차로 분류된다. 즉, 오차항 ϵ_{ij} 로부터 교호작용인 코호트 효과가 분리되어 정의되는 것이다.

Median Polish 분석은 먼저 분할표를 이용하여 각각의 행과 열의 중앙값을 반복적으로 제거하여 Median Polish의 잔차를 추정한다. 그 다음 추정된 잔차를 종속변수로 코호트 효과를 독립변수로 둔 선형회귀모형을 적합하여 코호트 효과의 크기를 예측하고 통계적으로 검정한다. 이 때 Median Polish 잔차를 이용한 선형회귀모형은 다음과 같이 표현된다.

$$e_k = \mu_k + \gamma_k + e_{ijk} \quad \cdots(9)$$

여기서 e_k 를 Median Polish의 잔차라고 정의하면 e_k 는 다시 μ 라는 절편, $a+p-1$ 개의 코호트 효과 중 k 번째 효과를 뜻하는 γ_k, 임의의 오차항 e_{ijk} 로 나누어 쓸 수 있다. 따라서 위의 선형회귀모형으로부터 추정된 회귀계수들은 코호트 효과들이 된다.

2.3.3 Intrinsic Estimator(IE) 방법

최근 의학통계학(Biostatistics) 분야에서 임의의 모수를 선택하여 제약하는 APC 방법론 보다 각 모수의 효과를 변화 없이 추정할 수 있는 함수를 사용하는 것을 제안하고 있다. 이러한 추세에 따라 Fu(2000)가 Intrinsic Estimator(IE)를 제안했다. 비정칙치 분해(singular values decomposition; SVD)와 추정함수를 기반으로 추정되는 IE는 연령, 기간 그리고 출생 코호트에 따른 질병 발생 추세에 대해 안정적으로 추정하며 유일한 해를 찾을 수 있다.

Intrinsic Estimator(IE)를 추정하기 위해서 다음과 같이 일반화선형모형을 사용한다.

$$Y = Xb + \epsilon \qquad \cdots(4)$$

위의 선형모형에서 최소자승추정량(ordinary least squares estimator)을 구하는 정규방정식은 아래와 같다.

$$\hat{b} = (X^T X)^{-1} X^T Y \qquad \cdots(6)$$

그러나 앞에서 설명한 바와 같이 계획행렬 X 가 비정칙행렬(singular matrix)이므로 역행렬을 구할 수 없어 유일한 해를 찾을 수 없다. 따라서 $X^T X$ 행렬의 고유치가 0에 대응되는 고유벡터 부분 공간 B_0 와 B_0의 공간에 직교하는 부분 공간으로 분해하고 무어-펜로즈 일반화 역행렬을 이용하여 유일한 해인 $\hat{b}$ 를 구할 수 있다. 즉, 한 모수 공간의 직교 분해를 통해 해를 추정하므로 정규방정식은 다음과 같이 표현된다.

$$\hat{b} = B + tB_0 \quad \cdots(10)$$

여기서 B_0 는 계획행렬의 고유치가 0에 상응하는 고유벡터로 관측된 발생률 Y 와는 관계가 없으며 오직 연령과 기간 그룹의 수에 의해 결정된다. 즉, B_0 는 계획행렬의 함수이다. t 는 B_0 의 방향을 결정하는 임의의 실수이다. 따라서 B_0 는 계획행렬에 의해 고정된 값이지만 임의의 실수 t 를 곱하여 임의의 벡터가 되는 것이다. B 는 t 가 0일 때 기하학적 투사(geometric projection)를 만족하는 값으로 Intrinsic Estimator(IE)가 된다. IE는 다음의 식과 같이 나타낼 수 있다.

$$B = (I - B_0 B_0^T)\hat{b} \quad \cdots(11)$$

IE는 임의의 실수 t 가 0일 때 기하학적 투사가 되면서 모수의 공간은 영공간(null-space)과 영이 아닌 공간(non-null space)으로 분해되어 영이 아닌 공간에서의 값, 즉 B 가 IE가 된다. 여기서 영이 아닌 공간의 B 는 무어-펜로즈 일반화 역행렬을 이용하여 추정할 수 있다. 무어-펜로즈 일반화 역행렬은 통계학에서 역행렬을 정의할 수 없을 때 역행렬을 구하기 위한 자연스러운 방법이다. IE는 기간이 고정되었을 때 불편성을 가지고, 제약된 일반선형회귀모형보다 작은 분산을 가지는 장점이 있다. 또한 요인의 계수를 직접적으로 제약하는 것이 아니므로 제약과 관계없이 동일한 결과를 얻을 수 있다(Yang Yang, 2013). 이러한 IE를 추정하기 위해 주성분 분석 알고리즘을 사용할 수 있으며 알고리즘 순서는 다음과 같다(Yang Y and Wenjiang J. Fu, 2004).

1) X^TX 의 고유치와 고유벡터를 계산하고 이 값들이 주성분이 된다.

2) 주성분의 단위를 표준화(normalizing)시킨다.

3) 고유치가 0에 대응되는 고유벡터 B_0 를 정의한다.

4) 관측된 발생률 Y 와 주성분을 사용하여 회귀모형을 추정한다.

5) IE를 추정하기 위해 모든 고유벡터의 정규직교 행렬(orthonormal matrix)을 이용하여 주성분 회귀 모형의 계수를 변환시킨다.

2.3.4 주성분 분석(Principal Component Analysis; PCA)

주성분 분석(Principal Component Analysis; PCA)은 모형을 적합할 때 APC 모형과 같이 변수 사이에 선형관계가 있거나 변수들 간에 상관관계가 높아 다중공선성이 발생할 때 변수를 삭제하기가 용이하지 않은 경우 통계학에서 주로 사용하는 분석 방법이이다. 주성분이란 새로운 잠재 변수(Latent Variable)로서 기존의 각 변수들을 이용하여 현재 자료를 가장 잘 설명할 수 있는 성분들을 만들어 모형을 적합 시키는 방법으로 차원을 축소하는 개념이다. 주성분은 기존의 각 변수들에 가중치를 부여하고 다음의 식과 같이 선형결합으로 표현할 수 있다.

$$PC_i = \boldsymbol{t_i} = w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{ip}x_p = X\boldsymbol{w_i} \qquad \cdots(12)$$

여기서 w_{ij} 는 i 번째 주성분 PC_i 의 j 번째 변수에 대한 가중치이다. 즉, 주성분 내에서 많은 부분에 대해 설명이 가능한 변수에 가중치를 많이 주는 개념이다. 그러므로 주성분의 분산이 클수록 자료를 가장 많이 설명할 수 있는 주성분이 된다. 만약 자료의 차원을 축소하는 것이 목적이라면 가장 많이 설명하는 제 1 주성분만 선택하여 모형을 추정할 수 있다.

주성분 분석에서 중요한 문제는 주성분의 수를 결정하는 것이다. 이 때 고려하는 것이 고유치의 크기이다. 여기서 고유치란 각 주성분 점수의 분산인 λ 를 의미한다. 즉, 자료의 총 변동 중에서 제 1 주성분 축에 사영되면서 전체 변동에서 제 1축 주성분 점수의 분산(λ_1)만큼 변동이 줄어들고 제 2축 사영에 의하여 제 2축 주성분점수의 분산(λ_2)만큼 변동이 줄어들며, 이런 식으로 아래와 같이 p 개의 주성분 축 사영으로 전체 변동을 완전히 설명할 수 있기 때문이다.

$$\lambda_{total} = \lambda_1 + \lambda_2 + \cdots + \lambda_p \qquad \cdots(13)$$

그러나 각 고유치들은 $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ 와 같은 관계가 있으므로 초기에 정의되는 주성분 축에 의하여 자료 대부분의 변동이 설명될 것으로 기대할 수 있다. 그러므로 주성분 공간의 차원 수를 결정하는 방법은 다음과 같다.

1) 1 이상의 고유치를 갖는 주성분 축만 고려한다.
2) 산비탈 그림(scree plot)을 그려 고유치를 순서대로 놓고 비탈이 꺾이는 점부터의 주성분을 무시한다.

APC 모형 추정의 식별문제 관점에서 주성분 모형의 가장 큰 특징은 최대 주성분 개수를 사용하여 주성분 모형을 적합 시켜 추정한 계수와 무어-펜로즈 일반화 역행렬의 해와 동일하다는 것이다. 여기서 최대 주성분 개수는 X^TX 의 계수(rank)이다. 즉, APC 모형에서 계획행렬 X 가 정칙행렬이므로 X^TX 의 역행렬을 정의할 수 없었고 모형의 모수를 추정할 수 없었다. 그러므로 IE 방법과 동일하게 무어-펜로즈 일반화 역행렬을 사용하여 역행렬을

찾아 해를 구할 수 있다. 그러나 모수의 공간을 분해하여 추정하는 IE 방법과 주성분 분석은 알고리즘이 다르므로 주성분의 개수를 최대로 하지 않고 주성분 분석을 한다면 IE와 다른 해를 추정한다(Yu-Kang Tu and Nicole Kramer, 2012).

제 3장 모의실험

이 장에서는 제 2장에서 소개한 4가지 방법을 통해 추정된 값이 타당한지 확인하고 상황별로 각 방법들을 비교하기 위해 모의실험을 수행하였다.

3.1 모의실험 설계

실제 연령별 기간의 발생률로 구성된 자료에서 실제 효과를 정확하게 추정하는 것은 불가능하다. 그러므로 각 효과의 실제 값을 가정하고 가상의 자료를 생성하여 4가지 방법을 적용하고자 한다. 모의실험에서 연령 효과의 범주 수는 10, 기간의 범주 수는 5로 설정하여 몬테카를로 모의실험을 하여 1000개의 자료를 생성한다. 이 자료는 10 × 5 결과 행렬로 행렬의 각 셀은 아래와 같이 포아송(poisson) 분포를 따른다.

$$Y_{ij} \sim Poisson(\lambda) \qquad \cdots (14)$$

이 때 분포의 λ는 다음의 식과 같다. 여기서 a_i 는 연령의 그룹을 의미한다. 예를 들어, 첫 번째 연령 그룹(a_1)이라면 식에 1을 대입한다. p_j는 기간의 그룹, c_{ij}는 코호트의 그룹을 의미한다. 코호트는 연령별 기간 분할표에서 대각행렬로 구성되므로 만약 첫 번째 연령 그룹(a_1)의 세 번째 기간 그룹(p_3)이라면 '코호트 = 기간 - 연령' 이므로 식에 2를 대입한다.

$$\lambda = \exp(0.3 + 0.1(a_i - 5)^2 + 0.1\sin(p_j) + 0.1\cos(c_{ij}) + 0.1\sin(10 \times c_{ij})) \qquad \cdots (15)$$

위의 식으로부터 실제 연령, 기간, 출생 코호트의 효과는 아래와 같다.

표 1. 가상으로 정의한 실제 연령, 기간, 코호트의 효과

연령 a_i 그룹의 연령 효과	$0.1\times(a_i-5)^2$
기간 p_j 그룹의 기간 효과	$0.1\times\sin(p_j)$
코호트 c_{ij} 그룹의 코호트 효과	$0.1\times\cos(c_{ij})+0.1\times\sin(10c_{ij})$

위에서 언급한 바와 같이 실제 연령별 기간의 발생률로 구성된 자료에서는 현존하는 APC 방법론으로 정확한 효과를 추정할 수 없으므로 각 요인에 대해 가상의 효과를 가정한다. 예를 들어, 연령의 그룹 수가 10이고, 기간의 그룹 수가 5인 경우 각 요인의 효과는 아래와 같다. 여기서 재매개변수화(reparameterization)란 각 요인효과의 평균이 0이 되도록 다시 계산한 값이다. 이후 4가지 방법론을 통해 나온 효과와 재매개변수화를 한 값을 비교하여 각 방법론이 타당한지 살펴보고자 한다.

표 2. 연령의 그룹 수(a)=10, 기간의 그룹 수(p)=5 인 경우 각 요인의 효과

요인	그룹	효과	재매개변수화
연령	a_1	2.025	1.200
	a_2	1.225	0.400
	a_3	0.625	-0.200
	a_4	0.225	-0.600
	a_5	0.025	-0.800
	a_6	0.025	-0.800
	a_7	0.225	-0.600
	a_8	0.625	-0.200
	a_9	1.225	0.400
	a_{10}	2.025	1.200

요인	그룹	효과	재매개변수화
기간	p_1	0.084	0.081
	p_2	0.091	0.087
	p_3	0.014	0.011
	p_4	-0.076	-0.076
	p_5	-0.096	-0.099
출생 코호트	c_1	0.000	-0.006
	c_2	0.050	0.044
	c_3	-0.198	-0.203
	c_4	0.009	0.004
	c_5	0.002	-0.003
	c_6	0.066	0.060
	c_7	0.153	0.147
	c_8	-0.114	-0.120
	c_9	-0.002	-0.007
	c_{10}	-0.135	-0.140
	c_{11}	-0.004	-0.010
	c_{12}	0.142	0.137
	c_{13}	-0.002	-0.008
	c_{14}	0.112	0.106

3.2 모의실험 결과

3.2.1 기간 그룹 수(p) = 5일 때, 3가지 요인의 효과가 모두 존재하는 경우

그림 1은 기간의 그룹 수가 5일 때 3가지 요인인 연령, 기간, 출생 코호트 효과가 모두 존재한다는 가정 하에 생성한 가상의 자료에서 연령, 기간 그리고 출생 코호트 효과를 추정한 결과이다. 각 방법에서 추정된 연령, 기간, 출생 코호트 효과와 가상의 실제 값을 비교하기 위해 1000개의 자료를 분석하여 각 방법에서 추정된 계수와 계수의 평균, 표준편차 그리고 평균오차제곱(Mean Squared Error; MSE)을 계산하였다. 각 방법으로부터 추정된 효과들의 평균과 실제 효과를 비교하여 각 추정량의 불편성(unbiasedness) 정도를 평가하고자 한다. 추정된 효과들의 표준편차는 추정량의 최소 분산(minimum variance) 특징을 보여주기 위한 값으로 작은 값일수록 실제 값을 더욱 정확하게 추정할 수 있다는 것을 의미한다. 평균오차제곱(MSE)은 실제 값과 추정량 차이의 제곱으로 추정량의 편의(bias)와 분산(variance)을 모두 고려할 수 있다. 제약된 일반화선형모형(CGLM) 방법은 하나의 요인을 제약해야한다. 따라서 임의로 첫 번째와 두 번째 연령 그룹을 제약한 모형(CGLM_A1)과 그래프를 통해 다섯 번째와 여섯 번째 연령 그룹을 제약한 모형(CGLM_A2), 첫 번째와 두 번째 기간 그룹을 제약한 모형(CGLM_P), 첫 번째와 두 번째 코호트 그룹을 제약한 모형(CGLM_C)으로 4가지로 설정하여 나머지 방법들과 비교하였다.

그림 1은 각 방법에 따른 추정량의 평균과 MSE를 비교한 것이다. 7가지 방법의 추정량 중 3가지 방법(IE, Median Polish, PCA)이 가상으로 정의한 실제 효과의 패턴과 유사한 것을 확인할 수 있다. 다섯 번째와 여섯 번째 연령과 첫 번째와 두 번째 기간, 코호트 그룹을 제약한 일반화선형모형(CGLM_A2,

CGLM_P, CGLM_C)을 사용한 경우 실제 값과 큰 차이가 있는 것을 확인할 수 있다. 첫 번째와 두 번째 연령기간을 제약한 모형(CGLM_A1)은 다섯 번째와 여섯 번째 연령 기간을 제약한 모형(CGLM_A2)보다 실제 값과 더 큰 차이가 있는 것을 확인할 수 있다. 이는 이전 연구에서 언급한 바와 같이 잘못된 모수 제약은 실제 추정량의 패턴을 명확하게 추정하지 못한다는 결과를 동일하게 재현한 것이다(Glenn 2005; Holford 1991). 그러나 이러한 제약한 일반화선형모형에서 만약 적절한 모수를 선택하여 제약한다면 실제 추정량의 패턴과 유사하게 추정한다(Kupper 1985). 그러므로 제약한 일반화선형모형을 사용할 때 어떤 모수를 제약할지에 대해 주의가 필요하다.

기간 효과를 제약한 일반화선형모형(CGLM_P)의 경우 실제 효과와 거의 비슷한 모양으로 보이나, 그림 1에서 제시된 평균에 의하면 기간 효과를 제약한 일반화선형모형(CGLM_P)의 추정량은 IE와 PCA의 추정량에 비해 실제 값에서 더 많이 벗어나며 MSE도 더 큰 값을 가지는 것을 확인할 수 있다.

그림 1에서 다섯 번째와 여섯 번째 연령 그룹을 제약한 모형(CGLM_A2), 첫 번째와 두 번째 코호트 그룹을 제약한 모형(CGLM_C), Median Polish 방법에 비해 IE와 PCA의 추정량의 표준편차가 0에서 1사이의 값으로 더 작은 값을 가진다. 기간 효과를 제약한 일반화선형모형(CGLM_P)의 추정량도 제약된 방법 중 작은 표준편차를 가지지만, IE와 PCA의 표준편차보다 10배 이상 큰 것을 확인할 수 있다. IE와 PCA의 추정량은 7가지 방법들 중 가장 작은 MSE를 가진다. 연령별 기간의 분할표에서 첫 번째 코호트와 마지막 코호트는 분할표의 첫 번째와 마지막 대각행렬에 위치하여 표본의 크기가 작기 때문에 이들에 대한 MSE가 가장 큰 것을 확인할 수 있다. 따라서 IE와 PCA에 비해 제약된 일반화선형모형이 특히 코호트의 효과를 정확하게 추정하지 못하는 것으로 판단된다.

3.2.2 기간 그룹 수(p) = 5일 때, 3가지 요인 중 1가지 효과가 부재인 경우

앞에서 설명한 3.2.1 경우는 연령, 기간, 출생 코호트의 효과가 모두 존재한다는 가정 하에 1000개의 자료를 생성하여 7가지 방법을 비교하였으나, 이번 장에서는 3가지의 효과 중 1가지의 효과가 부재인 경우 각 방법의 추정량의 타당성을 비교하고자 한다. 예를 들어, 본 논문의 모의실험과 같이 연령과 기간효과만 존재하는 경우 출산율 현상에 응용할 수 있다. 출산율과 같은 현상은 출생 코호트보다 연령과 기간 요인에 더 관계되어있기 때문이다. 따라서 코호트 효과가 없는 경우와 기간 효과가 없는 경우 그리고 연령 효과가 없는 경우 총 3가지의 가정 하에 1000개의 자료를 생성하여 각 방법을 사용해 추정량을 구하고 비교하였다. 3.1에서 설명한 바와 같이 만약 코호트의 효과가 존재하지 않는 경우 1000개의 자료를 생성하고자 한다면 아래와 같이 가상의 효과를 정의할 수 있다.

표 3-1. 가상으로 정의한 실제 연령과 기간의 효과

연령 a_i 그룹의 연령 효과	$0.1\times(a_i-5)^2$
기간 p_j 그룹의 기간 효과	$0.1\times\sin(p_j)$
코호트 c_{ij} 그룹의 코호트 효과	0

연령의 효과가 없는 경우와 기간의 효과가 없는 경우도 각 효과를 0으로 설정하고 1000개의 자료를 생성하여 각 방법을 비교할 수 있다. 그림 2는 코호트의 효과가 존재하지 않는 가정 하에 각 방법에 따라 추정량을 비교한 그림이다. 코호트의 효과가 존재하지 않는 경우에 IE와 PCA의 추정량은 나머지 방법에 비해 거의 0에 가까운 MSE를 가진다. IE는 코호트 효과가 존재하지 않는 자료에서 거의 정확하게 추정하는 것을 확인할 수 있다. 그러나 첫 번째와 두

번째 기간효과를 제약한 일반화선형모형(CGLM_P)과 Median Polish 방법은 대체적으로 출생 코호트의 효과가 변화하는 것으로 보인다. 그림 3과 4는 각각 연령과 기간의 효과가 존재하지 않는 가정 하에 각 방법에 따라 추정량을 비교한 그림이다. 그림 2와 동일하게 IE와 PCA 방법만 실제 효과와 유사하게 추정하는 것을 확인할 수 있다. 그러나 코호트의 효과가 존재하는 경우 Median Polish의 방법의 추정량은 IE와 PCA 방법과 더불어 실제 효과와 유사한 패턴으로 추정하며 코호트의 효과가 존재하지 않는 경우보다 MSE가 감소한 것을 확인할 수 있다.

3.2.3 기간 그룹 수(p) = 5일 때, 3가지 요인 중 2가지 효과가 부재인 경우

마지막으로 3가지의 효과 중 2가지의 효과가 부재인 경우 각 방법의 추정량의 타당성을 비교하고자 한다. 따라서 연령 효과만 있는 경우, 기간 효과만 있는 경우, 코호트 효과만 있는 경우 총 3가지의 가정 하에 1000개의 자료를 생성하여 각 방법을 사용해 추정량을 구하고 비교하였다. 3.1에서 설명한 바와 같이 만약 연령 효과만 존재하는 경우 1000개의 자료를 생성하고자 한다면 아래와 같이 가상의 효과를 정의할 수 있다.

표 3-2. 가상으로 정의한 실제 연령의 효과

연령 a_i 그룹의 연령 효과	$0.1\times(a_i-5)^2$
기간 p_j 그룹의 기간 효과	0
코호트 c_{ij} 그룹의 코호트 효과	0

기간 효과만 있는 경우와 코호트 효과만 있는 경우도 나머지 효과를 0으로 설정하고 1000개의 자료를 생성하여 각 방법을 비교할 수 있다. 그림 5는 연

령 효과만 존재한다는 가정 하에 각 방법에 따라 추정량을 비교한 그림이다. 연령 효과만 존재하는 경우에 IE와 PCA의 추정량은 나머지 방법에 비해 거의 0에 가까운 MSE를 가진다. 그러나 기간과 코호트의 효과가 0임에도 불구하고 IE와 PCA 방법의 추정량이 0 값에 근사하지 못하며 실제 값에서 벗어나는 것을 확인할 수 있다. 이러한 현상은 본 모의실험에서 기간의 수를 5로 고정한 후 각 방법들을 비교하였기 때문이다. 만약 기간의 수를 증가하고 자료를 생성한다면 IE와 PCA 방법에서 연령 효과와 마찬가지로 실제 값인 0에 근사하도록 추정한다(Yang and Fu 2004; Yang Yang 2007).

그림 6과 7은 각각 기간과 코호트의 효과만 존재하는 경우이다. 그림 5와 유사한 결과로 IE와 PCA 방법의 추정량이 나머지 방법에 비해 거의 0에 가까운 MSE를 가지지만 기간 효과만 존재하는 경우 다른 방법에 비해 Median Polish 방법의 추정량 평균이 실제 값과 비슷한 패턴을 보이는 것을 확인할 수 있다.

그림 1-1. p = 5일 때 연령, 기간, 코호트의 효과가 모두 존재하는 경우, 각 방법에 따른 추정량의 평균

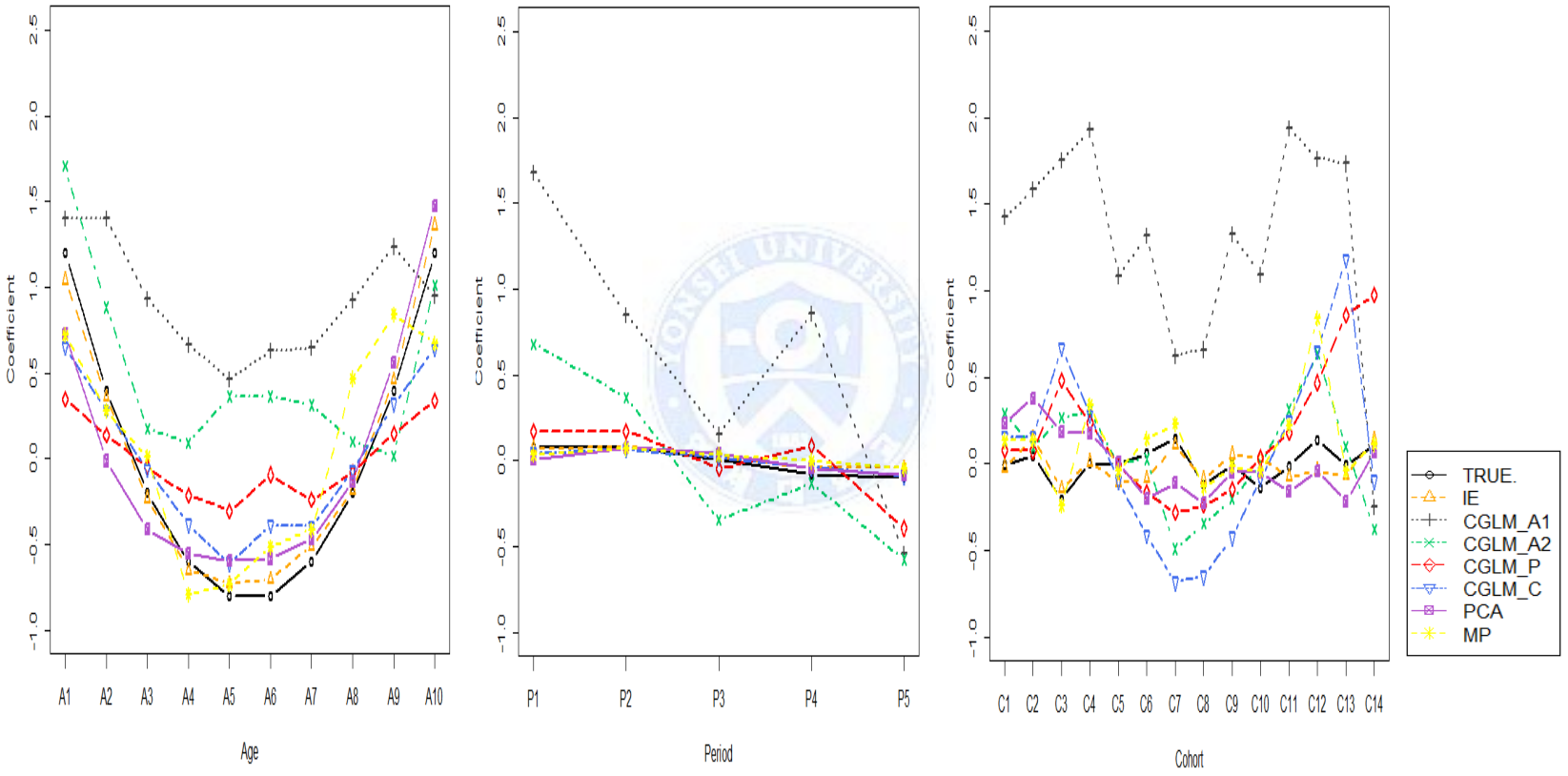


그림 1-2. p = 5일 때 연령, 기간, 코호트의 효과가 모두 존재하는 경우, 각 방법에 따른 추정량의 평균오차제곱

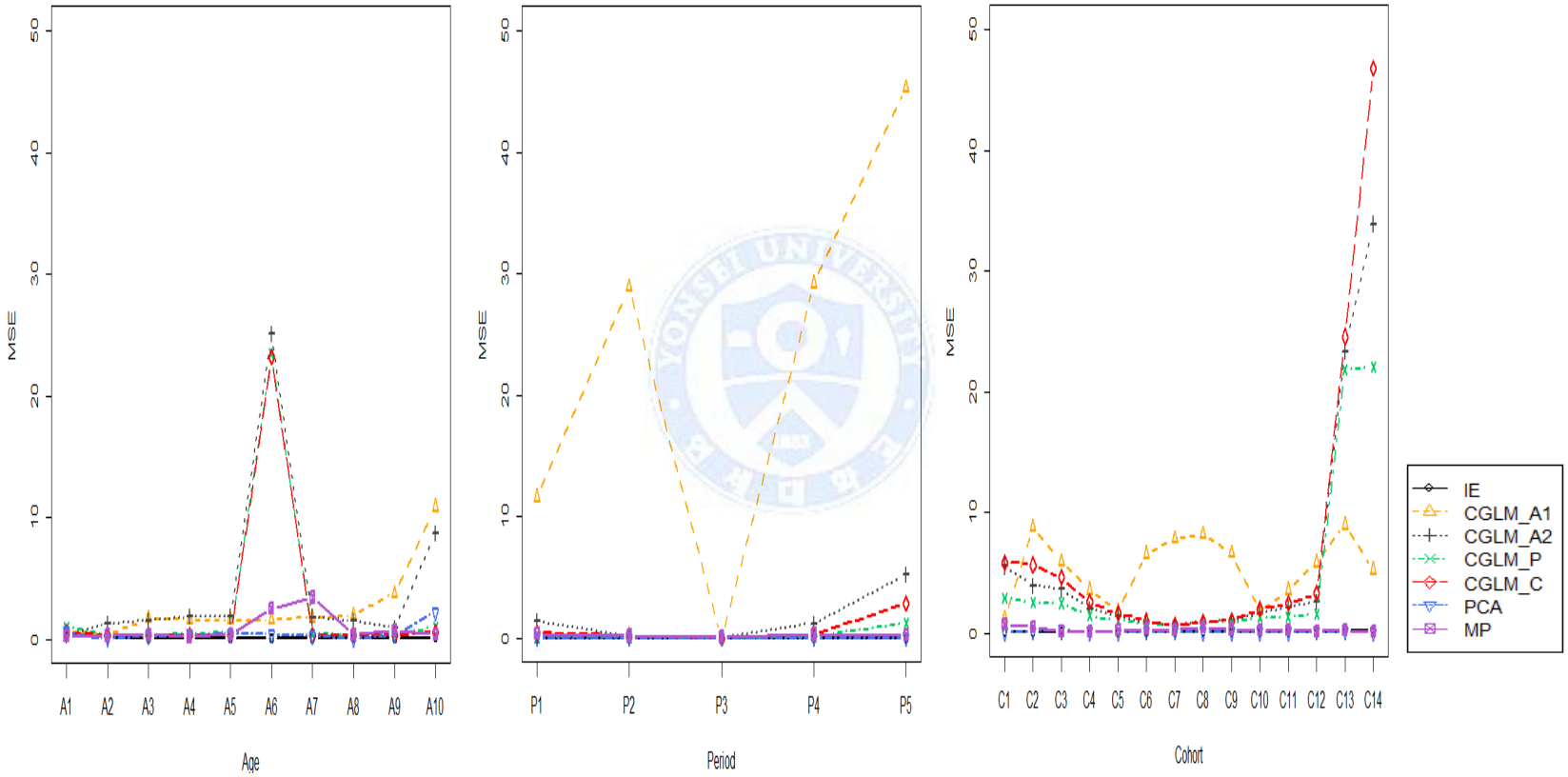


그림 2-1. p = 5일 때 연령과 기간의 효과가 존재하는 경우, 각 방법에 따른 추정량의 평균

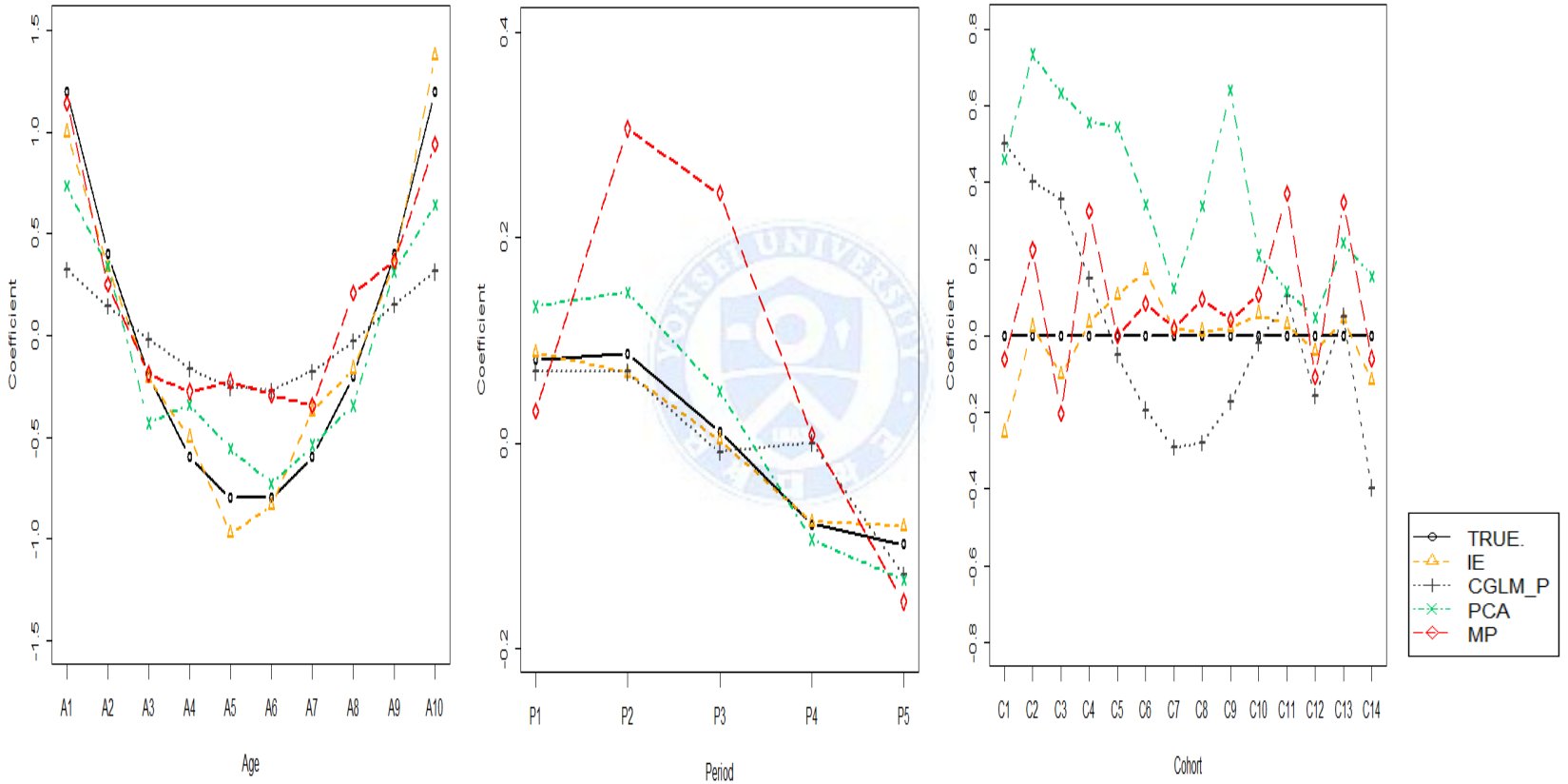


그림 2-2. p = 5일 때 연령과 기간의 효과가 존재하는 경우, 각 방법에 따른 추정량의 평균오차제곱

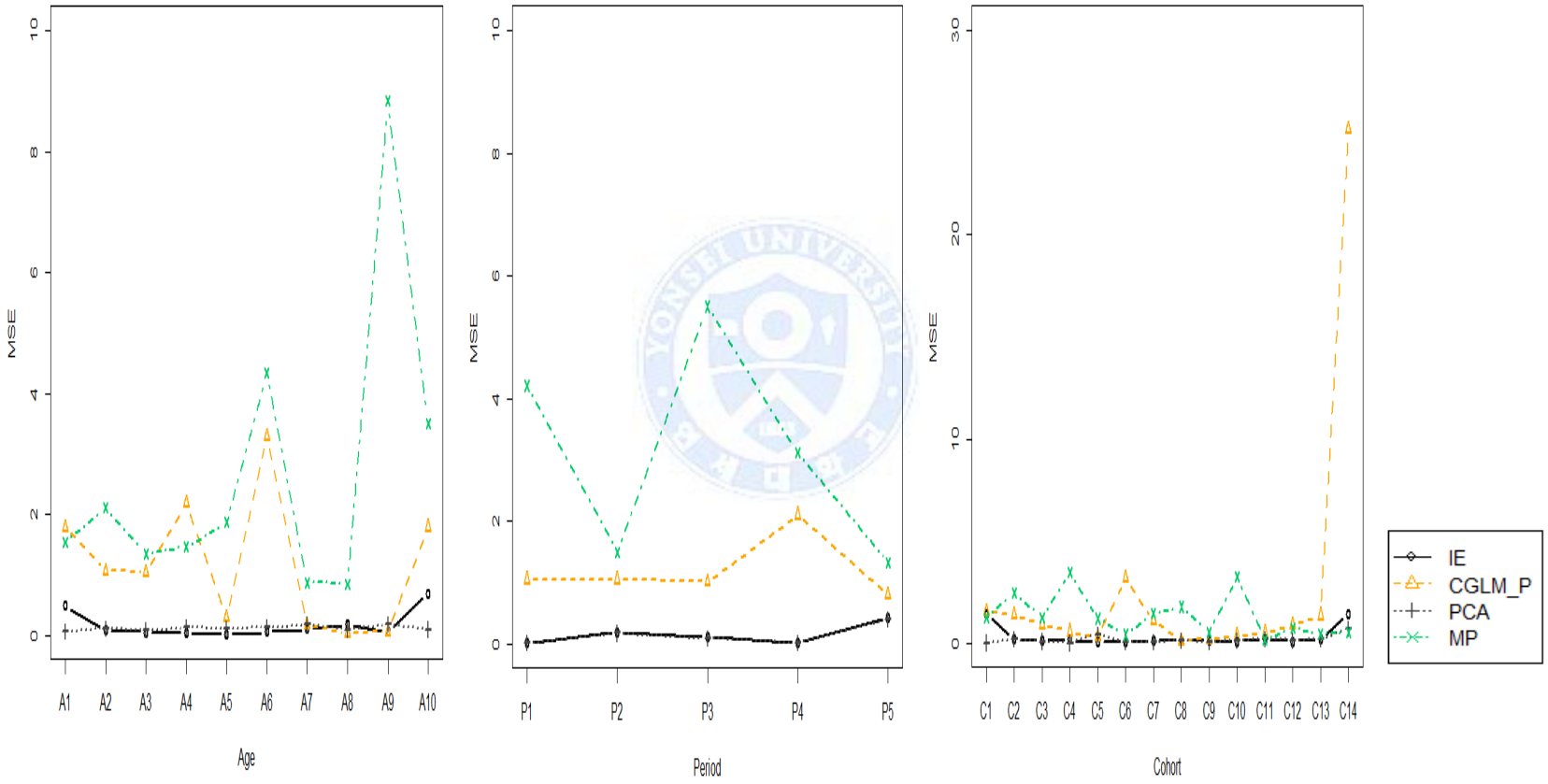


그림 3-1. p = 5일 때 연령과 코호트의 효과가 존재하는 경우, 각 방법에 따른 추정량의 평균

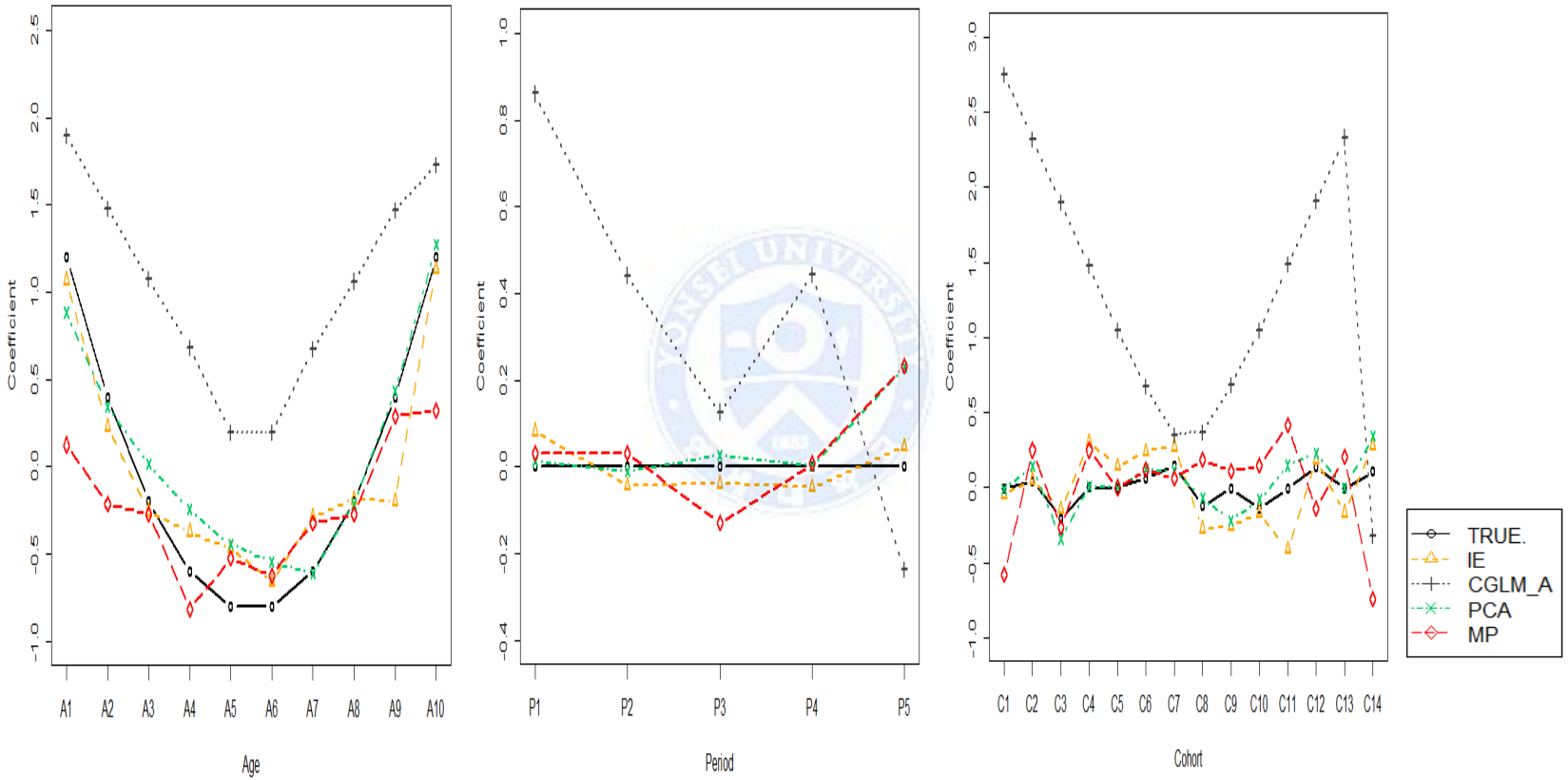


그림 3-2. p = 5일 때 연령과 코호트의 효과가 존재하는 경우, 각 방법에 따른 추정량의 평균오차제곱

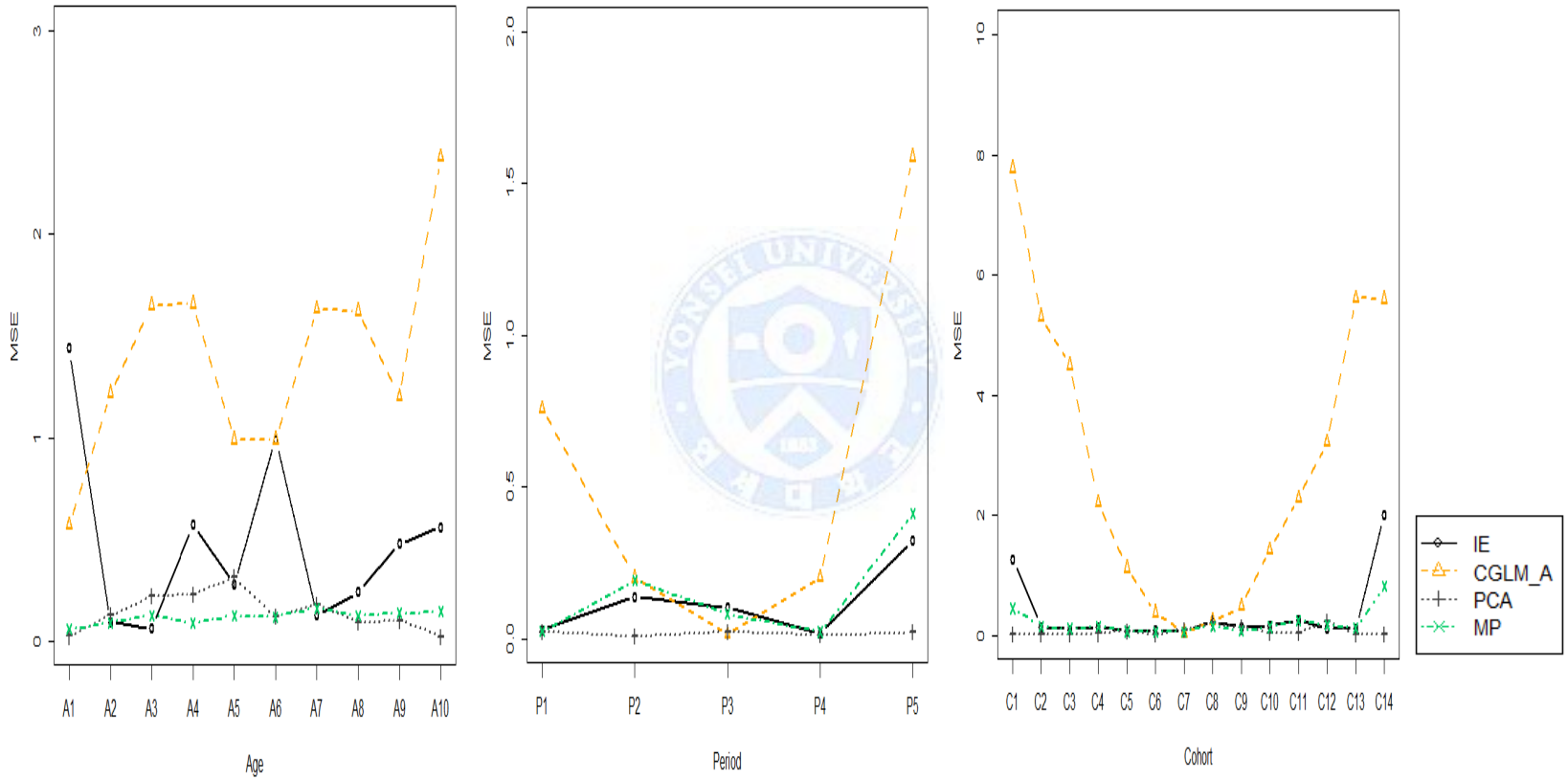


그림 4-1. p = 5일 때 기간과 코호트의 효과가 존재하는 경우, 각 방법에 따른 추정량의 평균

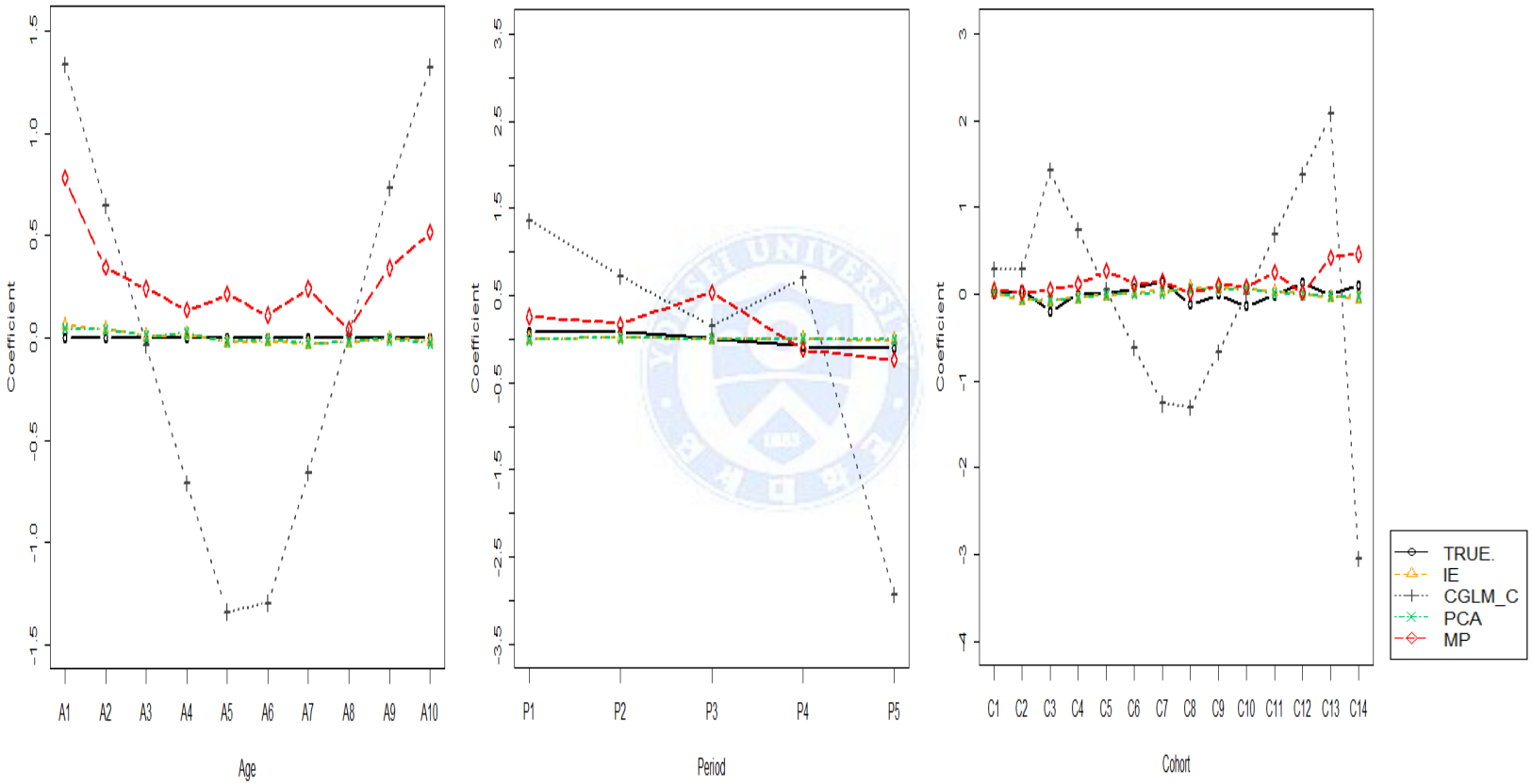


그림 4-2. p = 5일 때 기간과 코호트의 효과가 존재하는 경우, 각 방법에 따른 추정량의 평균오차제곱

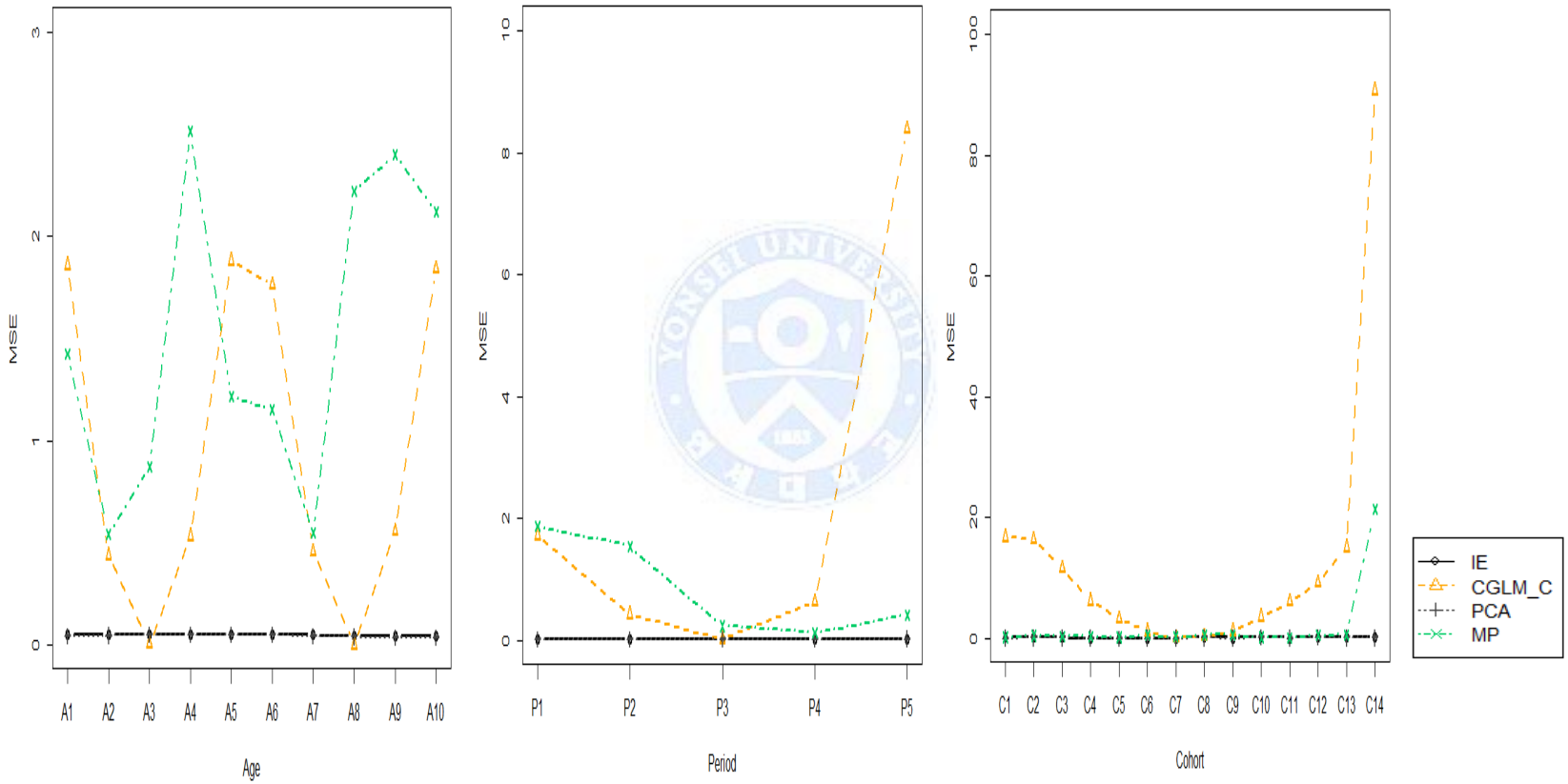


그림 5-1. p = 5일 때 연령 효과만 존재하는 경우, 각 방법에 따른 추정량의 평균

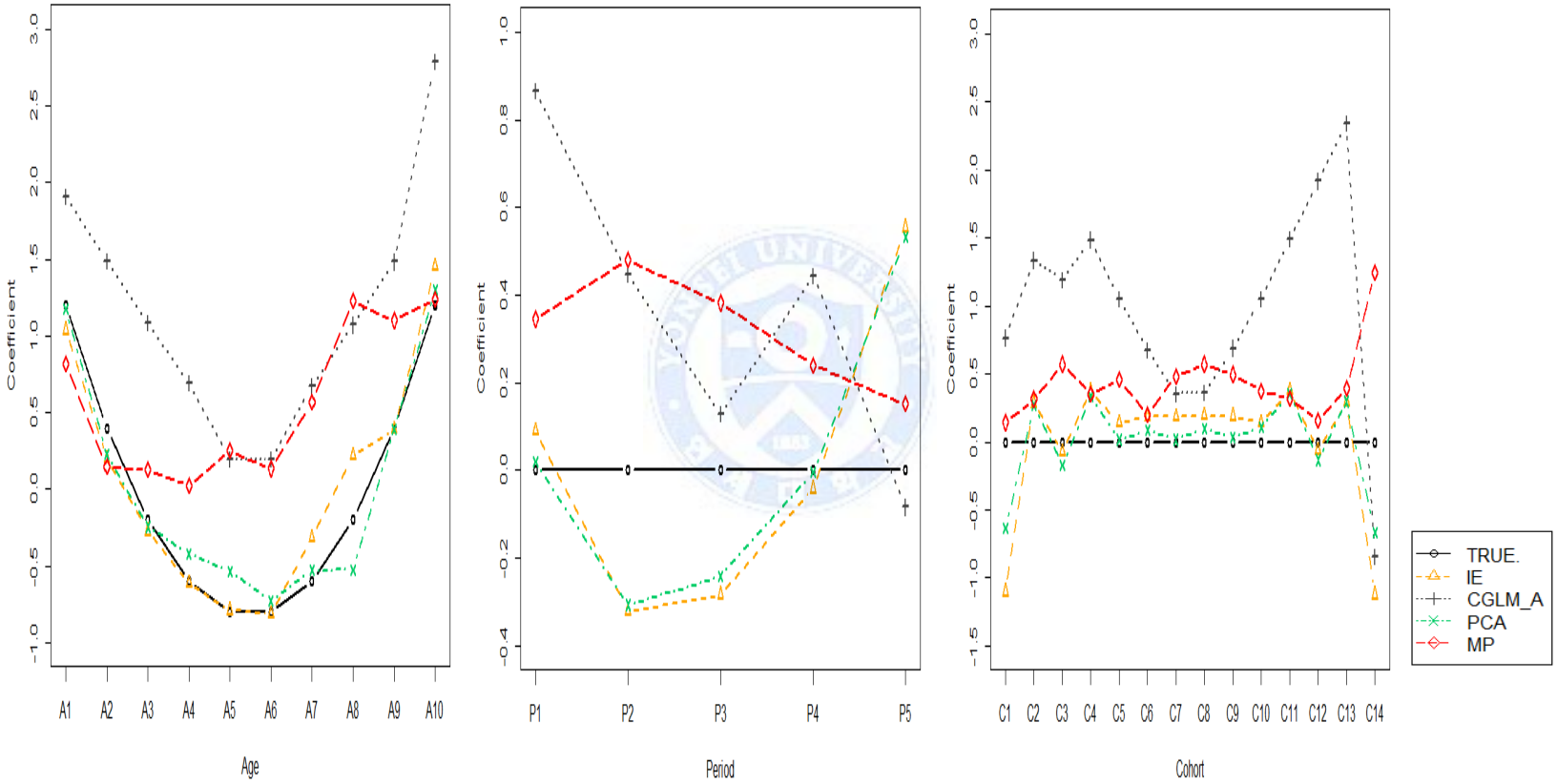


그림 5-2. p = 5일 때 연령 효과만 존재하는 경우, 각 방법에 따른 추정량의 평균오차제곱

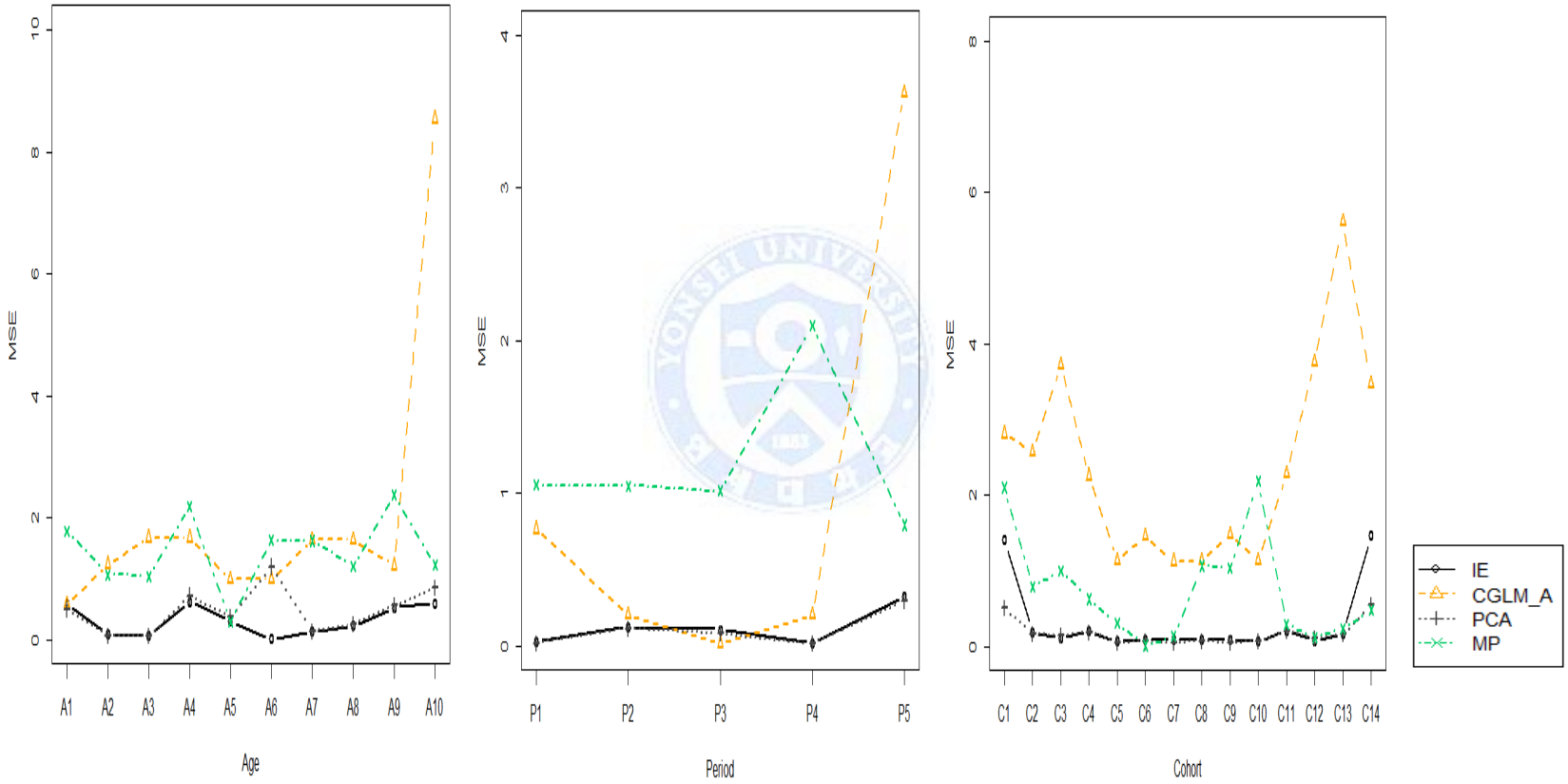


그림 6-1. p = 5일 때 기간 효과만 존재하는 경우, 각 방법에 따른 추정량의 평균

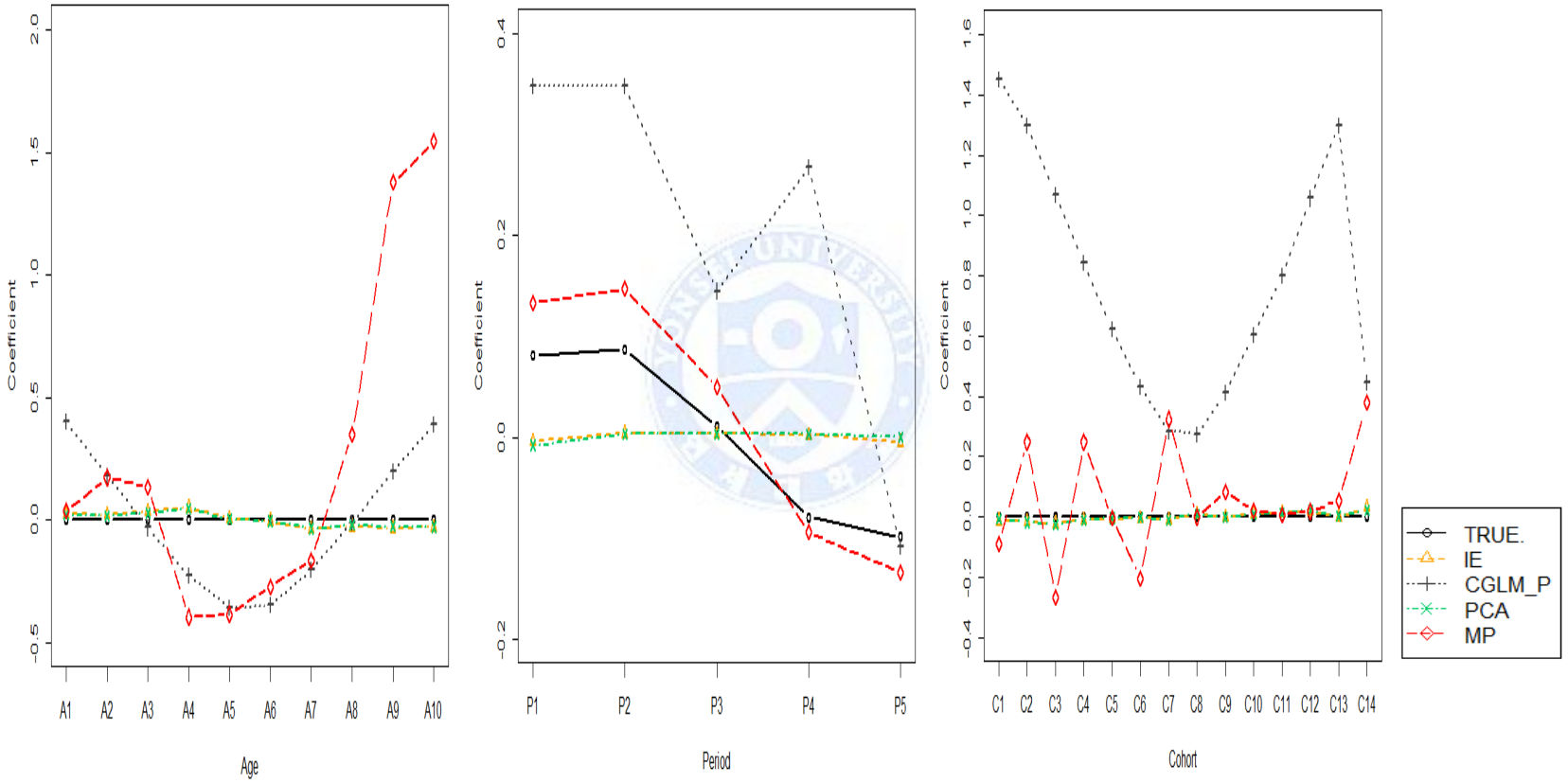


그림 6-2. p = 5일 때 기간 효과만 존재하는 경우, 각 방법에 따른 추정량의 평균오차제곱

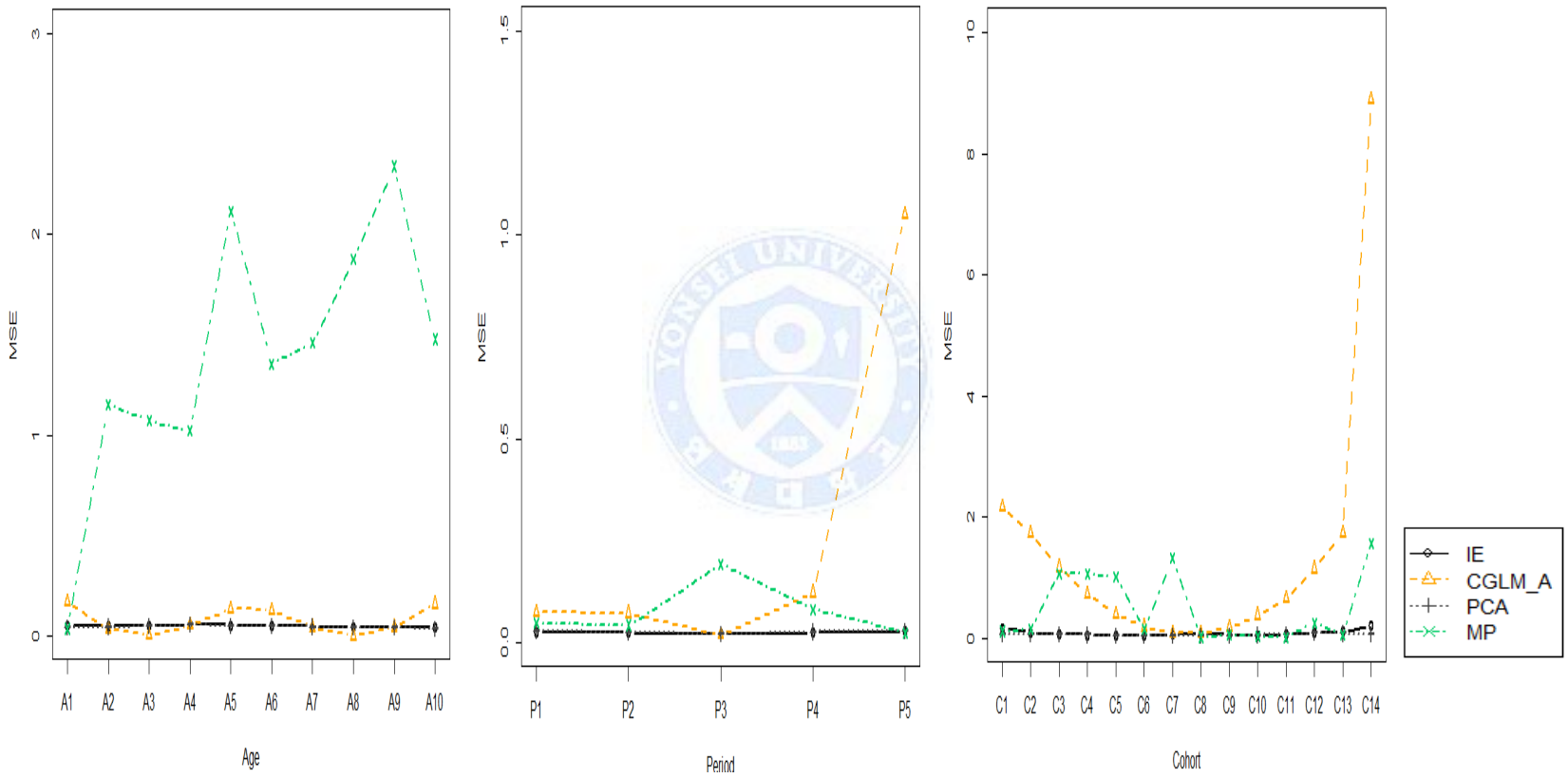


그림 7-1. p = 5일 때 코호트 효과만 존재하는 경우, 각 방법에 따른 추정량의 평균

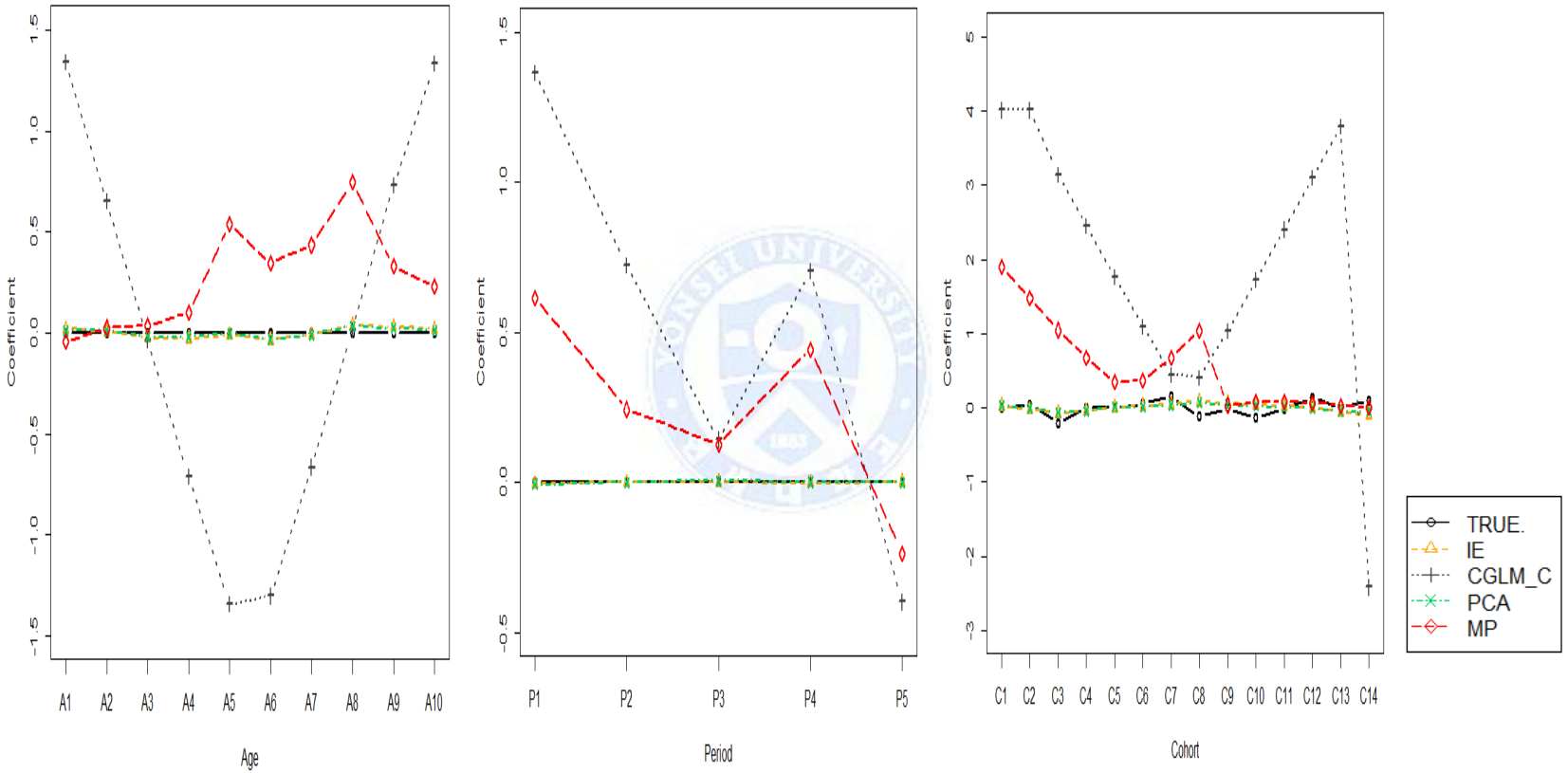
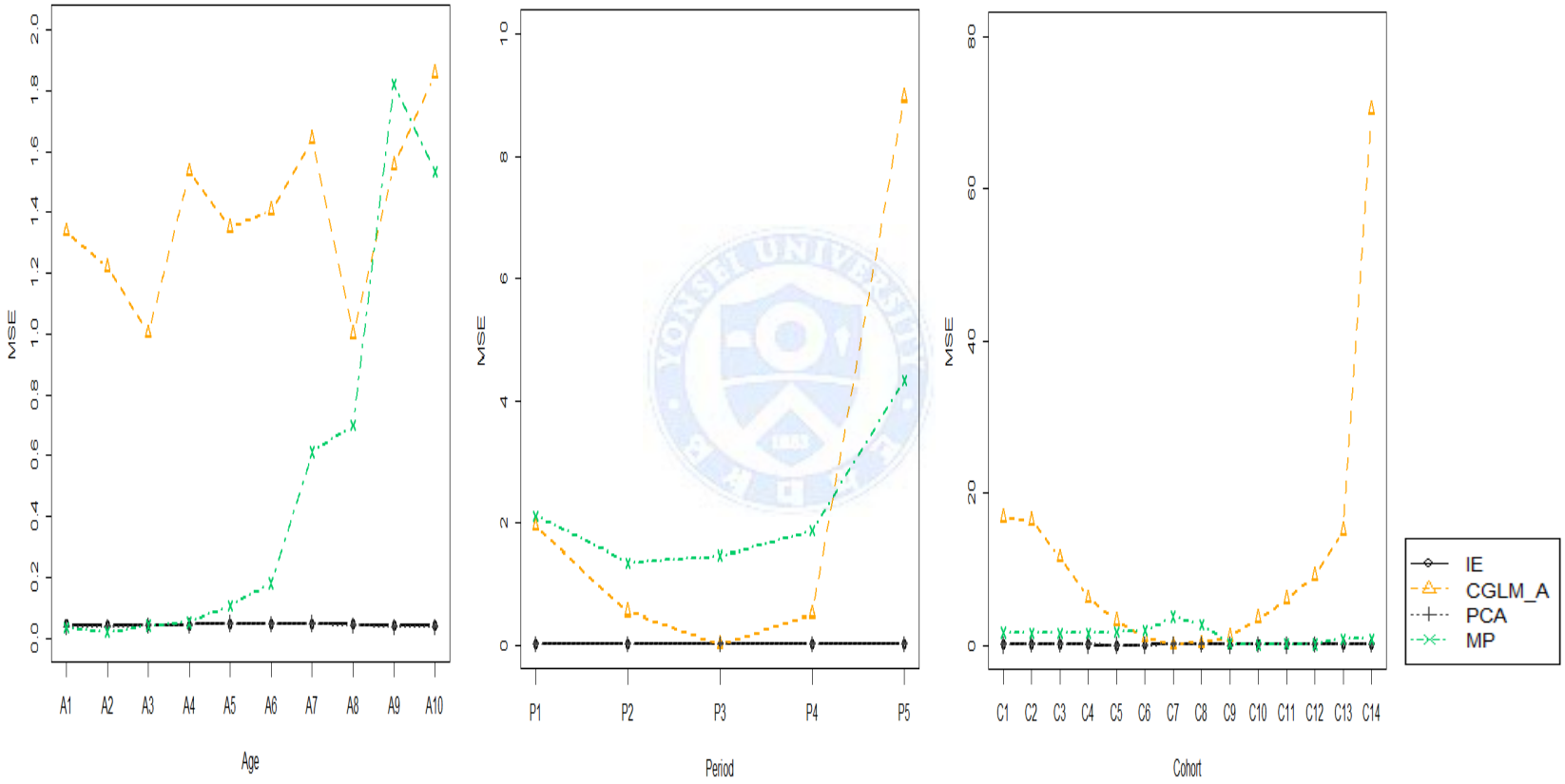


그림 7-2. p = 5일 때 코호트 효과만 존재하는 경우, 각 방법에 따른 추정량의 평균오차제곱



제 4장 실제 자료 분석

이 장에서는 실제 우리나라 5대 암 중 대장암 발생률 자료를 사용하여 제 2장에서 소개한 4가지 방법을 적용하여 각 방법들의 결과를 비교하고자 한다.

4.1 자료 설명

본 논문에서 사용되는 자료는 통계청 국가통계포털의 61개 암종/성/연령(5세)별 암 발생자 수와 발생률 자료이다. 우리나라 30세 이상 성인을 대상으로 1999년부터 2012년까지 조사된 대장암에 대한 발생률 자료로 연령-기간-코호트 분석을 위하여 표 4와 같이 연령은 5세 간격(30-34세 ~ 80-84세)로 정의하고 기간은 5년 간격(1999-2003, 2004-2008, 2009-2013)으로 정의하였다. 연령-기간-코호트 분석을 하기 위해서 연령과 기간을 동일한 간격으로 묶어야 하는데 현재 통계청 국가통계포털에서 2013년의 발생률 자료가 없기 때문에 2009년부터 2012년 자료의 평균치로 2013년 자료를 대체하였다. 또한 30세 미만에서는 30세 이상인 경우보다 암 발생률이 낮으므로 30세 이상부터 84세까지의 자료만 사용하였다. 자료에서 연령의 그룹 수가 11개이고 기간의 그룹 수가 3개이므로 분할표의 대각행렬로 정의되는 코호트는 총 13개로 계산되어 분석에 이용한다. 발생률은 통계청에서 제시한 연령표준화발생률을 사용하였다. 이 때 표준인구는 우리나라 2000년 주민등록연앙인구를 사용하였다. 대장암을 정의하기 위해 국제질병분류법(ICD-10)에 따라 C18(결장의 악성 신생물), C19(직장구불결장 이행부의 악성 신생물), C20(직장의 악성 신생물) 그리고 C21(항문 및 항문관의 악성 신생물)로 정의된다.

표 4. 1999년-2013년의 연령별 대장암 발생률(100,000명 당)

연령	기간		
	1999-2003	2004-2008	2009-2013
30-34	2.18	3.52	3.68
35-39	4.30	5.24	6.94
40-44	8.08	10.56	12.30
45-49	14.04	18.85	21.67
50-54	25.05	34.80	40.42
55-59	38.58	54.57	61.18
60-64	54.35	77.39	91.62
65-69	72.68	100.59	114.69
70-74	88.49	121.83	140.65
75-79	96.65	130.98	156.35
80-84	98.83	130.06	157.14

4.2 자료 분석 결과

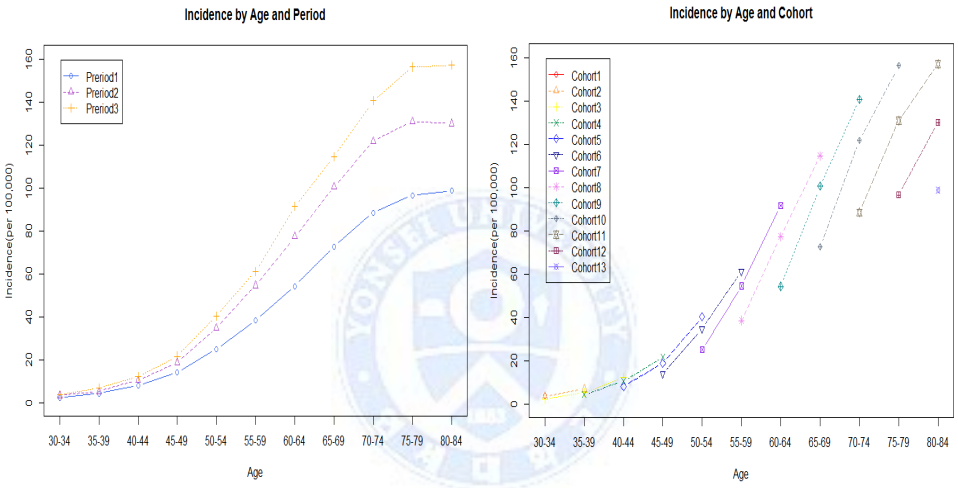
4.2.1 그래프를 이용한 방법(Graphical Approach)

연령-기간-코호트 분석 전 가장 쉽게 접근할 수 있는 방법은 연령에 따른 기간과 코호트의 발생률 그래프를 그려서 각 효과의 패턴을 시각적으로 확인하는 것이다. 연령에 따른 기간별 발생률과 출생코호트별 발생률을 토표로 나타내는 것으로 같은 기간이나 코호트 내에서 연령에 따라 발생률이 달라지는 경우 이를 연령효과로 보고, 모든 연령 그룹에서 기간에 따라 발생률이 달라지는 경우 이를 기간효과로 파악한다. 또한 연령별 발생률이 기간에 따라 평행하지 않은 경우나 발생률이 출생 코호트의 모든 연령대에서 상승되는 경우 이를 코호트 효과로 파악한다. 이 방법을 통해 대략적인 연령, 기간, 코호트

효과의 경향성을 쉽게 확인할 수 있지만 수치적으로 정확한 값이나 유의성 등을 파악하기 어렵다는 단점이 있다(Ho Jin Cho and Woo Hyun Joo 2014).

그림 8은 우리나라 대장암에 대해서 기간과 출생 코호트에 따른 연령별 암 발생률을 도표화한 것이다.

그림 8. 기간과 출생 코호트의 연령별 폐암 발생률(100,000명 당)



위의 그림에서 Period1부터 Period3은 1999년부터 2013년까지 5년 단위로 구간을 나눈 것을 의미하며 Cohort1부터 Cohort13은 연령과 기간의 대각행렬인 13개의 코호트를 의미한다. 그림 8에서 기간에 따른 연령별 대장암 발생률을 보면 기간이 증가할수록 대장암 발생률이 증가하는 것을 볼 수 있다. 또한 1999년부터 2003년 기간 내에서는 연령이 증가함에 따라 발생률이 증가하다가 75세 이상 연령 군에서 감소하는 경향을 보이고 있다. 따라서 약간의 기간 효과가 있는 것으로 판단된다. 출생 코호트에 따른 대장암 발생률을 보면 최근 코호트로 갈수록 발생률이 증가하는 경향성이 있는 것을 확인할 수 있다. 또한 그림 8에서 출생 코호트에 따른 대장암 발생률을 보면 60대 이상 연령 군

에서 기울기가 가파르므로 이 연령대의 발생률이 가장 높다고 볼 수 있다. 그래프를 통해 각 연령, 기간, 코호트의 효과를 통계학적으로 유의한지 판단할 수 없지만 연령-기간-코호트 분석 전에 각 효과들이 각 암 발생률에 대략적으로 어떠한 영향을 미치는지 대략적으로 판단할 수 있다.

4.2.2 제약된 일반화선형모형(Constrained Generalized Linear Model; CGLM)

2장에 언급한 바와 같이 어떤 변수를 제약하느냐에 따라 추정되는 회귀계수의 값과 경향이 매우 큰 차이를 보이기 때문에 본 연구에서는 어떤 모수에 제한을 두는 것이 좋은지 알아보기 위해 아래의 표 5와 같이 5대 암 발생률 자료를 이용하여 연령, 기간, 코호트에 대한 6가지 서로 다른 모형을 구축하고 우도비 검정(Likelihood Ratio Test; LR Test)을 시행하였다.

연령을 포함한 모형의 이탈도(Deviance)가 다른 효과를 포함한 모형의 이탈도보다 매우 작으므로 연령이 각 암 발생과 밀접한 연관이 있다는 것을 알 수 있다. 기간 효과를 포함한 모형의 이탈도 값이 가장 크기 때문에 모형에서 기간 효과를 제한하고 분석하는 것이 좋지만 본 연구에서는 기간 효과가 3개이므로 본 연구에서는 코호트 효과를 제한하기로 한다. 4.2.1절의 각 암의 출생코호트의 연령별 발생률 그래프를 통해 첫 번째 코호트와 두 번째 코호트 효과가 유사한 것을 확인할 수 있었다. 따라서 첫 번째 코호트와 두 번째 코호트 효과를 동일하게 정의하고 연령-코호트-분석을 하였다.

표 6은 대장암 발생률 자료를 이용하여 코호트를 제약하고 분석한 결과이다. 대장암의 경우 연령이 30-34세인 경우에 비해 이후 모든 연령 그룹에서 통계적으로 유의한 증가를 확인할 수 있었다. 80대에서 가장 높은 경향을 보인다. 기간 효과 또한 1999-2003년에 비해 그 이후 기간에서 발생률이 유의하게 증가하였다. 코호트 효과는 첫 번째와 두 번째 코호트를 동일하다고 제한

하였으므로 이들 간의 차이는 존재하지 않는다. 첫 번째 코호트에서 이후 코호트로 갈수록 선형적으로 유의하게 감소하였다. 즉, 나중에 태어난 코호트일수록 대장암 발생률이 낮아지는 경향을 보인다고 할 수 있다. 대장암 발생률 자료에서 연령효과의 범위는 1.62부터 22.95, 기간효과의 범위는 1.78부터 2.42, 코호트효과의 범위는 0.25부터 1.01이다. 따라서 대장암 발생률의 변화는 연령 효과가 가장 큰 영향을 미치는 것을 확인할 수 있다. 따라서 연령효과가 선형적으로 유의하게 증가하는 패턴을 보인다.

표 5. 대장암 발생률 자료를 이용한 우도비 검정 결과

Model	Deviance	DF	LR test	DF	P-value
연령	4747.97	22	4736.50	13	<0.001
기간	71363.29	30	71351.81	21	<0.001
코호트	16760.12	20	16748.65	11	<0.001
연령+기간	274.34	20	262.86	11	<0.001
연령+코호트	218.25	10	206.77	1	<0.001
기간+코호트	678.00	18	666.52	9	<0.001
연령+기간+코호트	11.47	9			

표 6. 대장암 발생률 자료의 코호트 효과를 제약한 일반화선형모형 결과

Variable	Risk Ratio	95% Confidence interval	
		Lower	Upper
Age			
30-34	Reference		
35-39	1.62	1.37	1.92
40-44	2.51	1.94	3.24
45-49	4.62	3.25	6.58
50-54	7.86	5.00	12.35
55-59	9.83	5.65	17.11
60-64	12.33	6.39	23.79
65-69	16.20	7.57	34.68
70-74	20.02	8.43	47.52
75-79	21.26	8.04	56.19
80-84	22.95	7.94	66.37
Period			
1999-2003	Reference		
2004-2008	1.78	1.60	1.97
2009-2013	2.42	1.96	2.99
Cohort			
1919	Reference		
1924	1.00		
1929	1.01	0.87	1.15
1934	1.04	0.82	1.32
1939	1.01	0.71	1.42
1944	0.83	0.53	1.31
1949	0.74	0.42	1.28
1954	0.70	0.36	1.37
1959	0.71	0.33	1.54
1964	0.54	0.23	1.31
1969	0.42	0.15	1.12
1974	0.32	0.09	0.11
1979	0.25	0.07	0.86

4.2.3 Median Polish 방법

표 7은 Median Polish 방법으로 분석한 결과이다. 대장암 발생률 자료에 대해서 연령 효과와 기간 효과는 제약된 일반선형회귀모형의 결과와 동일한 경향을 보였다. 그러나 제약된 일반선형회귀모형의 결과와 달리 코호트 효과가 존재하지 않았다. 아래의 그림 9는 추정된 Median Polish 잔차 그래프이다. 이 그래프를 통해서도 코호트에 따른 특별한 경향성을 보이지 않는 것을 확인할 수 있다.

그림 9. 대장암 발생률 자료의 Median Polish 잔차 그래프

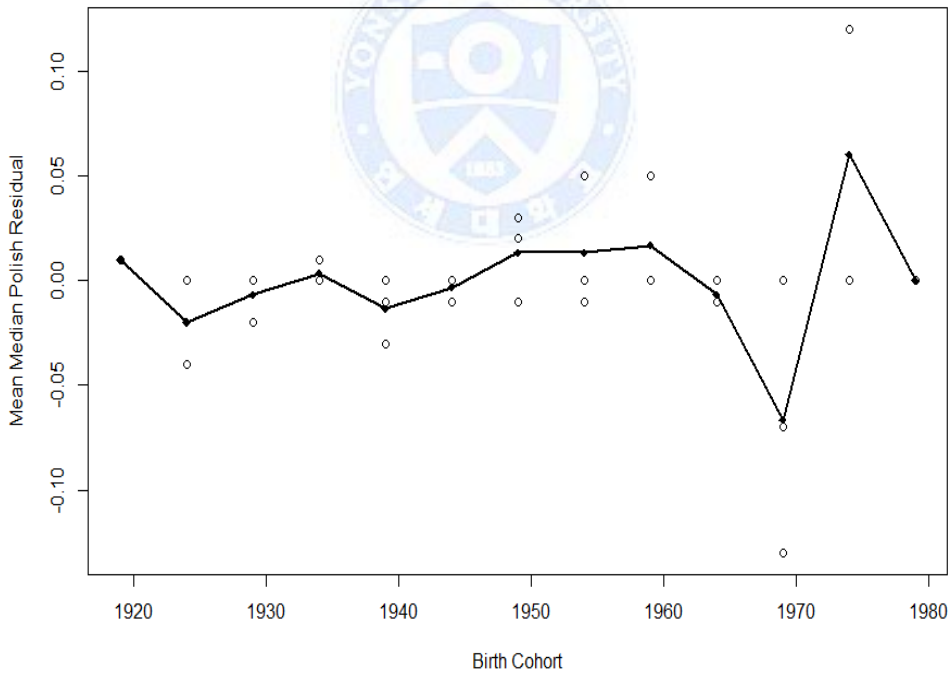


표 7. 대장암 발생률 자료의 Median Polish 방법 결과

Variable	Risk Ratio	95% Confidence interval	
		Lower	Upper
Age			
30-34	Reference		
35-39	2.08	1.89	2.29
40-44	4.18	3.83	4.56
45-49	9.41	8.64	10.23
50-54	18.01	16.59	19.56
55-59	23.32	21.50	25.31
60-64	32.43	29.91	35.16
65-69	48.36	44.60	52.42
70-74	65.23	60.15	70.73
75-79	69.39	63.91	75.33
80-84	73.35	67.35	79.89
Period			
1999-2003	Reference		
2004-2008	1.65	1.62	1.69
2009-2013	2.06	2.01	2.11
Cohort			
1919	Reference		
1924	0.97	0.89	1.05
1929	0.98	0.91	1.06
1934	0.99	0.92	1.07
1939	0.97	0.90	1.05
1944	0.98	0.91	1.06
1949	1.00	0.93	1.08
1954	1.00	0.93	1.08
1959	1.00	0.93	1.08
1964	0.98	0.91	1.06
1969	0.92	0.85	0.99
1974	1.05	0.97	1.13
1979	0.99	0.90	1.08

4.2.4 Intrinsic Estimator(IE) 방법

표 8은 IE 방법을 적용하여 분석한 결과이다. 연령 효과부터 살펴보면 대장암 발생률이 30대부터 유의하게 증가하는 경향을 보였다. 기간 효과도 최근 기간으로 올수록 발생률이 증가하는 경향을 보였다. 그러나 코호트 효과 경우 Median Polish 방법과 달리 통계적으로 유의하였으며 최근 코호트일수록 암 발생률이 낮아지는 경향을 확인할 수 있다.

4.2.5 주성분 분석(Principal Component Analysis; PCA)

표 9는 각 암 발생률 자료를 주성분 분석한 결과이다. 제 2장에서 언급한 바와 같이 주성분 분석에서 중요한 문제는 주성분의 수를 결정하는 것이다. 본 연구에서 사용된 자료를 이용하여 주성분의 분산을 고려한 결과 고유치가 1이상인 경우는 16개이고, 이때 전체 변동의 91%를 설명할 수 있다. 따라서 본 연구에서는 16개의 주성분을 사용하여 암 발생률 자료를 분석한다. 대장암 발생률 자료를 통해 분석한 결과 IE와 동일한 경향성을 보인다. 연령 효과부터 살펴보면 30대부터 유의하게 증가하는 경향을 보이며 기간 효과도 최근 기간으로 올수록 암 발생이 증가하는 것으로 판단된다.

표 8. 대장암 발생률 자료의 Intrinsic Estimator(IE) 방법 결과

Variable	Risk Ratio	95% Confidence interval	
		Lower	Upper
Age			
30-34	0.19	0.17	0.21
35-39	0.29	0.27	0.31
40-44	0.42	0.40	0.44
45-49	0.73	0.69	0.76
50-54	1.16	1.11	1.21
55-59	1.36	1.31	1.41
60-64	1.60	1.55	1.65
65-69	1.97	1.91	2.03
70-74	2.29	2.22	2.35
75-79	2.28	2.21	2.35
80-84	2.31	2.21	2.41
Period			
1999-2003	0.58	0.57	0.59
2004-2008	1.09	1.08	1.11
2009-2013	1.59	1.57	1.61
Cohort			
1919	2.15	1.99	2.32
1924	2.02	1.93	2.12
1929	1.91	1.85	1.97
1934	1.86	1.81	1.92
1939	1.69	1.63	1.75
1944	1.31	1.26	1.36
1949	1.09	1.04	1.13
1954	0.98	0.94	1.02
1959	0.93	0.89	0.98
1964	0.67	0.64	0.70
1969	0.48	0.46	0.51
1974	0.35	0.33	0.38
1979	0.26	0.23	0.30

표 9. 대장암 발생률 자료의 주성분 분석 결과

Variable	Risk Ratio	95% Confidence interval	
		Lower	Upper
Age			
30-34	0.09	0.02	0.44
35-39	0.16	0.05	0.57
40-44	0.27	0.11	0.69
45-49	0.54	0.29	1.01
50-54	0.98	0.71	1.36
55-59	1.35	1.23	1.49
60-64	1.84	1.34	2.54
65-69	2.65	1.41	4.97
70-74	3.55	1.39	9.05
75-79	4.12	1.18	14.34
80-84	4.81	1.02	22.82
Period			
1999-2003	0.67	0.49	0.92
2004-2008	1.08	1.04	1.12
2009-2013	1.38	1.01	1.89
Cohort			
1919	1.46	0.53	4.03
1924	1.59	1.06	2.40
1929	1.74	1.48	2.05
1934	1.96	1.49	2.59
1939	2.06	1.17	3.64
1944	1.86	0.78	4.47
1949	1.79	0.55	5.86
1954	1.89	0.42	8.41
1959	2.07	0.34	12.56
1964	1.72	0.21	14.20
1969	1.43	0.13	16.07
1974	1.16	0.08	17.96
1979	1.12	1.06	15.18

제 5장 결론 및 고찰

본 논문에서 연령-기간-코호트 연구 모형을 구축할 때 식별의 문제를 해결할 수 있는 방법들 중에서 역학 분야에서 주로 사용되는 4가지 접근 방법(Constrained Generalized Model, Median Polish Method, Intrinsic Estimator, Principal Component Analysis)을 비교하기 위해 각 방법에 대해 간략하게 살펴보고 가상의 연령, 기간, 출생 코호트의 효과를 설정하여 모의실험을 하였다. 기존의 여러 논문들에서 제약된 일반화선형의 문제점에 대해 이의를 제기하면서 최근에 제안되고 있는 IE 방법과 비교하였으나 본 논문에서는 현존하는 방법들 중 가장 많이 사용하는 4가지 방법을 모두 비교하였다. 모의실험 설계에서 실제 연령별 기간의 발생률로 구성된 자료에서 현존하는 연령-기간-코호트 분석 방법론으로 실제 효과를 정확하게 추정하는 것은 불가능하므로 각 효과에 대해 가상의 모양으로 설정하였다. 각 방법에서 나온 효과와 가상으로 정의한 효과를 비교하는데 이때 계산되는 효과들은 재매개변수화(Reparameterization)를 통해 각 효과의 합이 0이 되도록 하였다. 모의실험 결과를 통해 연령-기간-코호트 분석 시 가이드라인을 제시해주기 위해 모의실험 설계 시 연령, 기간, 코호트 3가지 요인의 발생에 따라 경우의 수를 고려하여 가상의 상황을 구상하였다.

연령, 기간, 코호트의 효과가 모두 존재하는 경우 IE 방법과 주성분 분석이 가상의 효과 값을 가장 유사하게 추정하였으며, 평균오차제곱(MSE)이 가장 작게 계산되었다. 또한 이전 연구에서 언급한 바와 같이 잘못된 모수 제약은 실제 추정량의 패턴을 유사하게 추정하지 못한다는 결과를 재현하기 위해 임의로 첫 번째 연령과 두 번째 연령 그룹을 제약했다. 본 논문에서도 마찬가지로 임의로 첫 번째 연령과 두 번째 연령 그룹을 제약한 경우 그래프를 통해

비슷한 연령 그룹을 제약한 경우보다 실제 값과 더 큰 차이가 난다는 것을 확인할 수 있었다. 따라서 제약된 일반화선형모형을 사용하는 경우 어떤 모수를 제약할지 결정하기 위하여 여러 참고 문헌과 통계학적 근거의 뒷받침이 필요한 것으로 판단된다.

실제 사회 현상을 고려하면 연령, 기간, 코호트 효과가 모두 존재하는 경우는 드물기 때문에 본 연구에서는 3가지 효과 중 1가지나 2가지가 부재인 경우도 고려하였다. 예를 들어 연령과 기간 효과만 존재하는 경우 출산율 현상에 응용할 수 있으며, 연령과 코호트 효과만 존재하는 경우 우리나라 유방암 발생률에 적용할 수 있다. 코호트 효과가 존재하지 않는 경우 IE 방법과 주성분 분석에서 얻은 추정량은 제약된 일반화선형모형과 Median Polish 방법에 비해 거의 0에 가까운 평균오차제곱(MSE)를 가진다. 또한 IE 방법의 경우 코호트 효과가 존재하지 않는 자료에서 거의 정확하게 추정하는 것을 확인할 수 있다. 그러나 코호트 효과가 존재하는 경우 Median Polish 방법의 추정량은 IE와 주성분 분석의 추정량과 더불어 실제 효과와 유사하게 추정하며 코호트 효과가 존재하지 않는 경우보다 평균오차제곱(MSE)이 감소하는 것을 확인하였다. 실제 자료 분석에서 살펴본 바와 같이 거의 대부분의 질병에는 연령효과가 존재하며 기간과 코호트의 효과 부수적으로 존재한다. 따라서 연령-기간-코호트 분석 전에 기간과 코호트의 효과가 통계적으로 유의한지 검정해보는 것이 중요할 것으로 예상된다.

3가지 효과 중 1가지 효과만 존재하는 경우 전반적으로 IE와 주성분 분석의 추정량이 실제 값과 가장 유사했으며 평균오차제곱(MSE)이 가장 낮은 것을 확인할 수 있었다. 그러나 본 논문에서 기간과 코호트의 효과가 0임에도 불구하고 IE와 주성분 분석 방법의 추정량이 0 값에 근사하지 못하며 실제 값에서 벗어나는 것을 확인할 수 있었다. 이러한 현상은 본 모의실험에서 기간의 수를 5로 고정한 후 각 방법들을 비교하였기 때문이다. 만약 기간의 수를 증가

하고 자료를 생성한다면 연령 효과와 마찬가지로 실제 값인 0에 근사하도록 추정할 것으로 예상된다.

4장에서 실제 자료 분석 결과를 통해 4가지 방법 모두에서 연령과 기간 효과가 유의하게 존재한다는 것을 확인할 수 있었다. 하지만 코호트 효과의 경우 세 모형 간에 차이가 발생하였다. 이처럼 같은 자료를 사용하였지만 어떤 방법을 사용하느냐에 따라 코호트 효과의 존재 여부에 차이가 발생하는 것을 확인해 볼 수 있었다. 하지만 모든 암 발생률 자료에서 각 방법의 효과들이 항상 다르게 나오는 것은 아니므로 제약된 일반화선형방법을 사용하는 것이 좋지 않다고 할 수는 없다. 그러나 위에서 언급한 바와 같이 제약된 일반화선형방법을 사용할 경우 어떠한 변수를 제약해야할지 신중하게 판단해야한다.

본 논문에서는 모의실험의 결과를 통해 제약된 일반화선형모형 보다는 IE 방법과 주성분 분석을 사용하는 것이 각각의 효과가 미치는 영향을 정확하게 볼 수 있다고 판단된다. 질병에 따라 어떤 분석 방법을 적용해야 정확하게 추정할 수 있는지 달라질 수 있지만 본 논문의 모의실험을 바탕으로 연령-기간-코호트 분석에 있어서 다음과 같이 간단한 가이드라인을 추천해주고자 한다. 먼저 연령-기간-코호트 분석 이전에 기간과 코호트의 효과가 유의한지 검정한다. 만약 코호트의 효과가 존재하는 경우 IE 방법, 주성분 분석, Median Polish 방법을 만약 코호트의 효과가 존재하지 않는 경우 가장 안정적인 IE 방법과 주성분 분석을 추천한다.

Theodore R. Holford. 1983. "The Estimation of Age, Period and Cohort Effects for Vital Rates". *Biometrics,* 39: 311-324.

Yang Yang, Sam Schulhofe-Wohl. 2007. "A simulation Study of the Intrinsic Estimator for Age-Period-Cohort Analysis". The Methodology Paper Session at the Annual Meetings of the American Sociological Association at New York.

KATHERINE M. KEYES. 2010. "A Multiphase Method for Estimating Cohort Effects in Age-Period Contingency Table Data". *Ann Epidemiol,* 20: 779-785.

Katherine Margaret Keyes. 2010. "What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971-2006". *Soc Sci Med,* 70(7): 1100-1108.

Sun Ha Jee, Il Soon Kim. 1998. "Projected mortality from lung cancer in South Korea, 1980-2004". *International Journal of Epidemiology,* 27: 365-369.

Yang Yang. 2008. "The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How to Use It". *American Journal of Sociology,* 113(6): 1697-1736.

Yu-Kang Tu, Nicole Kramer. 2012. "Addressing the Identification Problem in Age-Period-Cohort Analysis: A Tutorial on the Use of Partial Least Squares and Principal Components Analysis". *Epidemiology,* 23(4): 1044-3983.

Wenjiang J. Fu. 2011. "On the Intrinsic Estimator and Constrained Estimators in Age-Period-Cohort Models". *Sociological Methods & Research,* 40(3): 453-466.

B.Carstensen. 2007. "Age-Period-Cohort models for the Lexis diagram". *Statist. Med,* 26: 3018-3045.

ABSTRACT

Comparison of Estimation Methods for Age-Period-Cohort(APC) Model

Park, Se Young

Dept. of Biostatistics and Computing

The Graduate School

Yonsei University

Age-Period-Cohort(APC) analysis has played a critical role in studying time-specific phenomena. APC model has been popular tool in epidemiology to identify age, period and cohort trends in disease incidence and/or mortality rates. It is difficult to estimate the true separate effects because of the exact linear dependency. The limitations of existing approaches have been widely acknowledged by statisticians in all disciplines. The aim of the study is to review various methods of APC analysis and to provide a guideline to choose adequate method for evaluating age, period, and cohort effects by study conditions. The constrained generalized linear model(CGLM) approach, median polish approach, intrinsic estimator(IE) approach and principal component analysis(PCA) were used to estimate age, period, and cohort effect using generated the artificial data. While previous studies have used the CGLM method, CGLM depends on arbitrary parameter constraints. Therefore, the significance of period and cohort effects should be determined through APC analysis prior to analysis. Also, if cohort effects are present, the default in IE, PCA or median polish approach can be used. If cohort effects are not present, IE or PCA can be used.

Key words : APC model, Cohort effect, CGLM, Median Polish, IE, PCA