http://dx.doi.org/10.15813/kmr.2019.20.1.001

R을 이용한 구조방정식모델링: 분석절차 및 방법[1]

Structural Equation Modeling Using R: Analysis Procedure and Method

곽 기 영 (Kee-Young Kwahk) 국민대학교 경영대학/비즈니스IT전문대학원[2]

ABSTRACT

This tutorial introduces procedures and methods for performing structural equation modeling using R. For this, we present the whole process of analyzing the structural equations model from the confirmatory factor analysis to the path diagram generation using the lavaan package, which is relatively well evaluated among the R packages supporting the structural equation modeling, together with the R program codes. Considering that research applying structural equation modeling techniques is the mainstream in a variety of social sciences, including business administration, and that there is growing interest in open source R, this tutorial focuses on researchers who are looking for alternatives to traditional commercial statistical packages and is expected that it will be a useful guidebook for them.

Keywords: Structural equation modeling, R programming, lavaan, LISREL, AMOS

1. 서론

구조방정식모델링(structural equation modeling)은 잠재변수를 다룰 수 있고 여러 변수 간의 직접적·간접적 영향관계를 동시에 분석할 수 있다는 장점으로 인하여 경영학을 비롯한 사회과학 분야의 연구에서 널리 활용되고 있다. 이로 인해 구조방정식모델링을 지원하는 LISREL, AMOS, EQS, Mplus 등과 같은 상용 소프트웨어 패키지들이 높은 가격에도 불구하고 많은 사회과학 분야 연구자들의 필수 도구로 받아들여지고 있다.

그러나 IT 환경의 변화와 오픈소스에 대한 관심이 증대함에 따라 상용 소프트웨어 패키지를 대신할 수 있는 오픈소스 통계 프로그램에 대한 요구가 커지고 있다. 또한 데이터분석에 대한 사회적 관심과 함께 대표적 데이터분석 프로그래밍 언어인 R에 대한 관심도 증대하고 있다. R은 통계처리를 위한 목적으로 통계학자들에 의해 개발되기는 하였지만 사용자 기반이 점차 다양해짐에 따라 R의 응용분야는 통계학 이외의 영역으

1) 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A3A2075114).
논문접수일: 2019년 2월 6일; 게재확정일: 2019년 2월 23일

2) 제 1저자 (kykwahk@kookmin.ac.kr)

로 확장되고 있다(곽기영 2017). R은 현재 데이터분석이나 그래픽과 같은 비통계적 목적으로도 폭넓게 이용되고 있으며, 머신러닝(machine learning), 데이터마이닝(data mining)/텍스트마이닝(text mining), 자연어처리(natural language processing) 등 다양한 학문분야로 응용범위를 넓혀가고 있다.

R 환경에는 구조방정식모델링을 지원하는 다수의 라이브러리(R에서는 이를 패키지(package)라고 부른다)가 있다. 예를 들어, 현재 sem, strum, lava, lavaan, OpenMx 등을 사용할 수 있으며, 구조방정식모델링에 대한 연구자들의 선호가 계속되는 한 이를 지원하는 라이브러리는 지속적으로 증가할 것으로 기대된다. 이 가운데 lavaan 패키지는 상용 소프트웨어 못지않은 기능으로 구조방정식모델링 절차를 충실히 구현하고 있으며 지속적으로 업데이트가 이루어지고 있어 구조방정식모델링을 지원하는 R 패키지 가운데 비교적 좋은 평가를 받고 있다(Rosseel 2012). 여기에서는 lavaan 패키지를 이용하여 구조방정식모델링을 수행하는 절차를 살펴본다. 본 튜토리얼은 다음과 같은 내용으로 구성된다. 우선 다음 두 섹션에서는 구조방정식모델링에 대한 일반적인 개념을 살펴본다. 구체적으로 구조방정식모델링의 구성요소와 분석절차를 다룬다. 이어서 lavaan 패키지를 이용한 구조방정식모델링 수행절차를 살펴본다. lavaan 패키지에 포함된 실제 데이터를 이용하여 R 프로그램 코드와 함께 분석절차를 상세히 기술한다. 끝으로 R을 이용한 구조방정식모델링의 시사점을 토의한다.

2. 구조방정식모델링의 구성요소

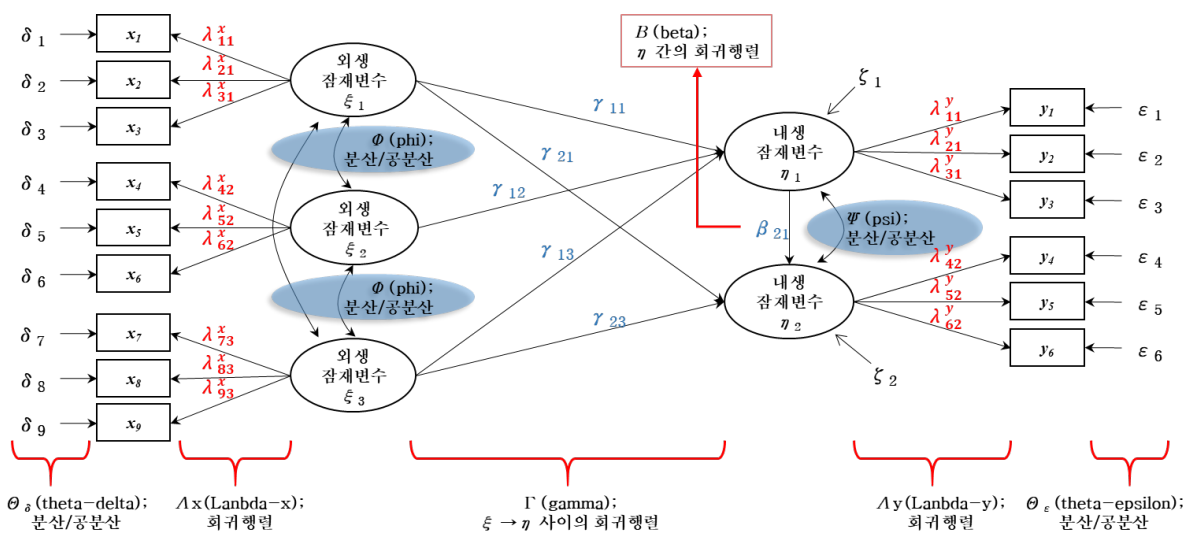
구조방정식모델링은 직접적인 측정이 어려운 잠재변수(latent variable) 간의 영향관계를 분석하기 위한 다변량 통계분석 기법이다. 구조방정식모델링은 전통적인 회귀분석과 달리 잠재변수를 다룰 수 있으며 여러 변수 간의 영향관계를 동시에 분석할 수 있다. 관측변수(observed variable)를 이용하여 잠재요인을 간접적으로 측정한 후 이들 잠재요인 간의 이론적인 영향관계를 분석한다. 따라서 구조방정식모델링은 요인분석과 회귀분석의 특성을 결합한 하이브리드 기법이다.

구조방정식모델(structural equation model)은 잠재변수를 측정하는 측정모델(measurement model)과 측정된 잠재변수 간의 인과관계를 분석하는 구조모델(structural model)로 구성된다. 일반적으로 측정모델에는 확인적 요인분석(confirmatory factor analysis)을 사용하고, 구조모델에는 여러 개의 회귀모델이 결합된 다중회귀분석을 이용한다. 구조방정식모델은 흔히 <그림 1>과 같은 경로도(path diagram)로 표현된다. 경로도는 관측변수와 잠재변수 간의 관계, 그리고 이들 잠재변수 사이의 인과관계를 시각적으로 표현하기 때문에 복잡한 구조를 갖는 구조방정식모델을 이해하는 데 유용하다.

잠재변수는 외생잠재변수(exogenous latent variable)와 내생잠재변수(endogenous latent variable)로 구분된다. 외생잠재변수는 그리스 문자 ξ [ksi]로 나타내며, 내생잠재변수는 η[eta]로 표기한다. 외생잠재변수는 모델 내의 다른 잠재변수에 영향을 미치는 변수로서 모델 내에서 독립변수로서의 역할만 수행한다. 반면에 내생잠재변수는 모델 내의 외생잠재변수에 의해 직접적으로 또는 간접적으로(다른 내생잠재변수를 통해) 영향을 받는 변수로서 모델 내에서 독립변수와 종속변수로서의 역할을 수행한다. 잠재변수는 관측변수에 의해 측정되므로 관측변수 또한 두 가지로 구분된다. 외생잠재변수를 측정하기 위한 관측변수는 외생관측변수(exogenous observed variable)라 하고 알파벳 x로 나타내며, 내생잠재변수를 측정하기 위한 관측변수는 내생관측변수(endogenous observed variable)라 하고 y로 표기한다.

구조방정식모델링에서는 공분산행렬이 분석의 대상이다. 회귀분석은 개별 케이스의 실제 관측값과 회귀식에 의해 추정된 예측값의 차이를 최소화하는 회귀선을 찾는 기법이다. 반면에 구조방정식모델링은 개별 케이스에 관심을 두는 것이 아니라 개별 케이스로부터 얻어진 공분산행렬에 초점을 두어 표본의 공분산행렬과 모델에 의해 예측된 공분산행렬 간의 차이를 가능한 작게 하는 구조방정식모델을 추정한다. 이를 위해 구조방정식모델을 생성하는 데 필요한 다수의 모수(parameter)를 추정한다. 구조방정식모델링에서 추정해야 할 모수로는 구조계수(structural coefficient), 요인적재값(factor loading), 공분산(covariance), 구조오차(structural error), 측정오차(measurement error) 등이 있다.

구조계수는 잠재변수 간의 경로계수를 의미한다. 『외생잠재변수 ξ → 내생잠재변수 η』 간의 회귀계수는 γ [gamma]로 나타내고, 『내생잠재변수 η → 내생잠재변수 η』 간의 회귀계수는 β[beta]로 나타낸다. 요인적재값은 측정변수에서 잠재변수에 이르는 경로계수로서 λ[lambda]로 나타낸다. 외생변수와 내생변수를 구분하기 위하여 『외생잠재변수 ξ → 외생관측변수 x』 간의 회귀계수는 λ^x로 나타내고, 『내생잠재변수 η → 내생관측변수 y』 간의 회귀계수는 λ^y로 나타낸다. 구조방정식모델에는 회귀분석의 회귀모델과 달리 측정오차가 존재한다. 측정오차는 관측변수가 잠재변수를 완전하게 설명하지 못하는 정도를 나타내며 측정오차의 크기를 알고 있으면 통계적 추정에 이를 반영할 수 있고 순수한 구조계수의 크기를 추정할 수 있다. 외생관측변수 x의 측정오차는 δ [delta]로 나타내며 내생관측변수 y의 측정오차는 ε [epsilon]으로 나타낸다. 내생잠재변수에 영향을 미치는 모든 잠재변수(외생잠재변수 및 다른 내생잠재변수)로 설명되지 않고 남아 있는 부분



ξ [ksi] 외생잠재변수
η [eta] 내생잠재변수
x 외생잠재변수의 관측변수
y 내생잠재변수의 관측변수

ζ [zeta] 내생잠재변수 오차
δ [delta] 외생잠재변수의 관측변수 오차
ε [epsilon] 내생잠재변수의 관측변수 오차

Φ [phi] 외생잠재변수 간의 공분산
Ψ [psi] 내생잠재변수 오차 간 공분산
Θ [theta] 관측변수 오차 간 공분산

λ [lambda] 잠재변수에서 관측변수에 이르는 경로계수
γ [gammma] 외생변수에서 내생변수에 이르는 경로계수
β [lambda] 내생변수에서 내생변수에 이르는 경로계수

<그림 1> 구조방정식모델의 구성요소

을 구조오차라고 한다. 구조오차는 구조방정식모델에서 ζ [zeta]로 표시한다. 구조방정식모델링에서는 공분산도 모수로 추정한다. 외생잠재변수 간의 공분산은 ϕ [phi]로 나타내며, 내생잠재변수의 오차(즉 구조오차) 간 공분산은 ψ[psi]로 나타낸다. 관측변수의 측정오차 간 공분산은 외생변수 내생변수 구분 없이 θ[theta]로 나타낸다.

3. 구조방정식모델링 분석절차

구조방정식모델링에 의한 분석은 일반적으로 2단계 접근법에 따라 수행된다(Anderson and Gerbing 1988, 윤지현·곽기영 2014). 첫 번째 단계로서 우선 측정모델을 대상으로 잠재변수 및 관측변수의 단일차원성, 신뢰도, 타당도를 평가한다. 이어지는 두 번째 단계에서 검증된 관측변수들로 구성된 구조모델을 바탕으로 잠재변수 간의 경로분석을 수행하여 이들 잠재변수 간의 영향관계를 검정한다.

3.1 측정모델

측정모델에 대한 평가는 확인적 요인분석을 통해 수행된다. 확인적 요인분석은 모든 잠재변수에 대해 수행되어야 하며 잠재변수별로 개별적으로 평가할 수도 있으나, 일반적으로 전체 잠재변수가 하나의 모델로 구성된 통합 측정모델(pooled measurement model)을 대상으로 한꺼번에 수행한다. 확인적 요인분석을 통해 잠재변수의 단일차원성(unidimensionality), 신뢰도(reliability), 타당도(validity)를 평가할 수 있다. 신뢰도와 타당도를 평가하기 앞서 단일차원성에 대한 평가를 먼저 수행한다.

측정모델을 구성하는 관측변수는 오직 하나의 구성개념(construct, 잠재변수)만을 측정해야 한다. 측정모델의 이러한 특성을 단일차원성이라고 한다. 측정모델의 단일차원성은 각각의 잠재변수가 단일요인모델(single factor model)에 의해 잘 적합되는지로 평가한다. 모델의 전반적인 적합도(fitness)는 다양한 적합도 지표를 통해 평가할 수 있다. 일반적으로 많이 사용되는 적합도 지표를 <표 1>에 정리하였다.[3] 모델

<표 1> 적합도 지표

구분	모델 적합도 지표명		권장 수준	참고문헌
절대적합도	Chisq	Discrepancy Chi Square	p-값 > 0.05	Wheaton et al. (1977)
	RMSEA	Root Mean Square of Error Approximation	RMSEA < 0.08	Browne and Cudeck (1993)
	GFI	Goodness of Fit Index	GFI > 0.9	Joreskog and Sorbom (1984)
증분적합도	AGFI	Adjusted Goodness of Fit Index	AGFI > 0.9	Tanaka and Huba (1985)
	CFI	Comparative Fit Index	CFI > 0.9	Bentler (1990)
	TLI	Tucker-Lewis Index	TLI > 0.9	Bentler and Bonett (1980)
	NFI	Normed Fit Index	NFI > 0.9	Bollen (1989)
간명적합도	Chisq/df	Chi Square/Degrees of Freedom	Chisq/df < 3.0	Marsh and Hocevar (1985)

3) 적합도 지표 가운데 Chisq는 관측된 공분산행렬과 모델에 의해 예측된 공분산행렬이 같다는 귀무가설을 검정한다. 따라서 카이제곱검정(chi-square test) 결과 귀무가설을 기각하지 못하면(p-값 > 0.05) 모델이 데이터를 잘 적합시켜 모델 적합도가 높다고 평가한다. 그러나 카이제곱통계량은 표본크기가 증가함에 따라 함께 커지는 경향이 있기 때문에 표본크기가 클 경우 귀무가설은 쉽게 기각된다. 따라서 표본크기가 클 경우(보통 200 이상) 이 지표는 적합도 평가 항목에서 제외되기도 한다(Hair et al. 2010, Joreskog and Sorbom 1996).

적합도는 크게 절대적합도(absolute fit), 증분적합도(incremental fit), 간명적합도(parsimonious fit)로 나누어 볼 수 있다.[4] 지표 선택에 대한 일치된 견해는 없지만 모델적합도의 각 범주별로 적어도 한 개의 지표를 사용할 것이 권장된다(Hair et al. 2010).

단일차원성은 각 잠재변수에 대해 모든 관측변수가 적정 수준 이상의 요인적재값을 가질 때 충족된다. 단일차원성을 충족하기 위해서는 일반적으로 0.6 이상의 표준화 요인적재값을 가져야 한다. 적합도 지표가 요구되는 수준을 충족하지 못하면 각 관측변수에 대한 요인적재값을 검토하여 낮은 요인적재값(즉 0.6 미만)을 갖는 관측변수를 한 번에 한 개씩 제거한다. 가장 낮은 요인적재값을 갖는 변수부터 시작하여 차례대로 한 번에 한 개씩 제거하며 확인적 요인분석을 반복해서 수행한다. 적합도 지표가 적정 수준에 도달할 때까지 이 과정을 반복한다.

낮은 요인적재값을 갖는 관측변수를 제거한 후에도 모델적합도 수준이 여전히 만족스럽지 못하면 수정 지표(modification indices)를 검토한다. 수정 지표는 추정할 새로운 모수를 모델에 추가하면 적합도가 어떻게 변하는지를 알려준다.[5] 따라서 수정 지표에 따라 모델에 새로운 관계를 설정하여(즉 모수를 추가하여) 적합도를 개선할 수 있다. 그러나 구조방정식모델은 기본적으로 탐색적 성격이 아니기 때문에(즉 모델에 포함될 모수는 이론적 검토를 거쳐 사전에 이미 결정되었기 때문에) 새로운 모수의 추가는 신중할 필요가 있다. 수정 지표에 따라 모델을 수정할 경우 데이터에 대한 과적합(overfitting)으로 인해 분석결과의 일반화 가능성을 약화시킬 위험이 있다. 수정 지표를 이용하는 또 다른 방법은 이를 모델이 잘 적합시키지 못하는 부분에 대한 정보로서 활용하는 것이다. 모수를 추가할 경우 적합도의 큰 개선이 이루어진다는 것은 모델의 해당 부분이 그만큼 데이터를 잘 적합시키지 못한다는 뜻이기도 하다. 이는 큰 수정 지표가 가리키는 부분에 모델의 적합도에 기여하지 못하는 불필요한 변수가 포함되어 있을 수 있다는 것을 의미한다. 따라서 큰 수정 지표값과 관련된 변수 가운데 하나를 제거함으로써 적합도를 개선할 수 있다. 대상 변수 가운데 요인적재값이 작은 변수를 우선적으로 제거한다. 수정 지표와 요인적재값을 고려하여 한 번에 한 개씩 변수를 제거하며 확인적 요인분석을 반복해서 수행한다. 적합도 지표가 적정 수준에 도달할 때까지 이 과정을 반복한다.

측정모델의 단일차원성이 확보되면 이어서 각 잠재변수의 신뢰도를 평가한다. 신뢰도는 측정척도가 측정하려고 의도하는 것을 얼마나 정확하게 오차 없이 측정하고 있는지를 나타낸다. 척도의 신뢰도를 평가하는 지표로서 일반적으로 많이 사용하는 것은 크론바흐 알파계수(Cronbach's coefficient α)이다. 크론바흐 알파계수는 하나의 개념을 여러 측정항목으로 측정할 경우에 나타나는 항목 간의 일관성이나 동질성의 정도를 평가하며, 이를 내적일관성(internal consistency)이라고 한다. 크론바흐 알파계수에 대한 명확한 기준은 없으나 보통 0.7 이상이면 바람직한 수준으로 판단한다(곽기영 2019).

구조방정식모델링에서는 내적일관성 지표와 함께 측정변수의 요인적재값과 측정오차를 함께 고려한 복합신뢰도(composite reliability, CR)의 사용이 권장된다. CR은 다음과 같이 계산하며, 0.7 이상일 경우 잠재변수의 신뢰도가 확보된 것으로 간주한다(Bagozzi and Yi 1988).

4) 증분적합도 지표는 베이스라인(baseline) 모델과 제안 모델 간의 성능 비교를 바탕으로 산출한다. 베이스라인 모델은 관측변수 간의 상관관계가 없다고 가정하며, 이런 이유로 독립 모델(independence model)이라고도 불린다.

5) 모델적합도의 변화는 카이제곱통계량으로 보여준다.

$$CR = \frac{(\sum_1^n \lambda)^2}{(\sum_1^n \lambda)^2 + \sum_1^n \delta(\varepsilon)} = \frac{(\sum_1^n \lambda)^2}{(\sum_1^n \lambda)^2 + \sum_1^n (1-\lambda^2)}$$

여기서, λ 는 표준화 요인적재값, δ 와 ε 은 측정오차, n은 측정변수 개수

신뢰도는 또한 모든 잠재변수에 대해 평균분산추출(average variance extracted, AVE)을 계산하여 평가할 수 있다. AVE는 다음과 같이 계산한다.

$$AVE = \frac{\sum_1^n \lambda^2}{\sum_1^n \lambda^2 + \sum_1^n \delta(\varepsilon)} = \frac{\sum_1^n \lambda^2}{\sum_1^n \lambda^2 + \sum_1^n (1-\lambda^2)} = \frac{\sum_1^n \lambda^2}{n}$$

여기서, λ 는 표준화 요인적재값, δ 와 ε 은 측정오차, n은 측정변수 개수

표준화된 요인적재값의 제곱은 잠재변수에 의해 설명될 수 있는 관측변수 분산의 크기를 나타낸다. 따라서 AVE는 잠재변수에 대한 관측변수의 평균적인 설명력을 의미한다. AVE가 0.5 이상일 경우 신뢰도 요건을 충족한 것으로 판단한다.

타당도는 측정척도가 측정하려고 의도하는 것을 얼마나 충실하게 측정하고 있는지를 나타낸다. 타당도는 크게 집중타당도(convergent validity)와 판별타당도(discriminant validity)의 관점에서 평가한다. 집중타당도는 측정척도가 측정하기로 되어 있는 잠재변수와 관련을 갖는 정도를 나타낸다. 즉 동일한 잠재변수를 측정하는 복수의 측정척도가 서로 어느 정도 일치하는지를 의미한다. 잠재변수의 측정척도는 동일한 잠재변수를 설명하도록 하나로 '집중(수렴)'되어야 한다. 즉 관측변수는 모두 공통적으로 대응되는 잠재변수에 의해 가능한 많은 분산이 설명되어야 한다. 요인적재값이 클수록 집중타당도는 증가한다. 집중타당도가 충족되기 위해서는 측정모델의 모든 관측변수의 요인적재값이 통계적으로 유의해야 한다.

집중타당도는 또한 AVE를 통해서도 평가할 수 있다. 집중타당도가 확보되기 위해서는 AVE는 0.5 이상이어야 한다(Fornell and Larcker 1981). AVE는 요인적재값을 바탕으로 계산되기 때문에 모델 내에 낮은 요인적재값을 갖는 측정항목이 포함되어 있으면 잠재변수의 집중타당도는 낮아진다.

판별타당도는 측정척도가 측정하지 않기로 되어 있는 다른 잠재변수와는 관련을 갖지 않는 정도를 의미하며, 다른 잠재변수에 속한 측정변수 간에는 서로 관련성이 작아야 한다는 것을 나타낸다. 즉 각각 다른 잠재변수를 측정하는 측정척도가 서로 어느 정도 다른지를 나타낸다. 일반적으로 (탐색적) 요인분석을 하는 경우 일정 크기 이상의 교차적재값(cross loading)이 존재하면, 즉 한 측정항목이 두 개 이상의 요인에 높게 적재되면 판별타당도가 충족되지 못한 것으로 간주한다. 그러나 구조방정식모델링은 이론을 바탕으로 잠재요인과 관련 관측변수 간의 측정모델을 수립하기 때문에 이론에 의해 도출되지 않은 교차적재를 직접적으로 평가하지는 않는다.[6] 어떤 잠재변수의 판별타당도에 문제가 있으면 관측변수는 해당 잠재변수의 관측변수들보다 다른 잠재변수의 관측변수들과 더 높은 상관관계를 갖는다. 즉 잠재변수는 자기 자신의 관측변수에 의해서보다는 다른 잠재변수의 관측변수에 의해서 더 잘 설명된다. AVE는 잠재변수에 의해 설명될 수 있는 관측변수의 변동을 나타내므로 이를 잠재변수 간의 상관계수와 비교함으로써 판별타당도를 평가할 수 있다. AVE의 제곱근이 잠재변수 간 상관계수보다 크면 판별타당도를 충족한 것으로 판단한다(Fornell and Larcker 1981).

3.2 구조모델

구조모델에 대한 평가는 잠재변수와 잠재변수 간의 관계에 초점을 맞추어 진행하며, 연구모델에 의해 설정된 이론적 관계가 데이터에 의해 지지되는지 검토한다. 모델의 적합도 평가, 가설과 직접적으로 관련된 경로계

6) 구조방정식모델링에서는 확인적 요인분석 수행 시 잠재변수와 관측변수 간의 경로계수만 자유모수로 설정하여 추정하며, 교차적재 부분은 고정모수로 설정하여 0으로 지정하기 때문에 교차적재는 추정되지 않는다.

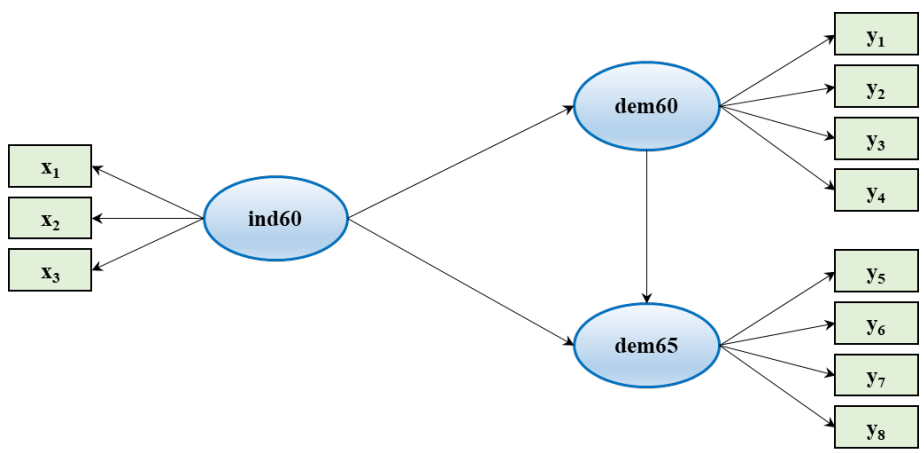
수에 대한 유의성 검정, 독립변수의 설명력을 평가하기 위한 결정계수(R^2)에 대한 검토 등이 주로 수행된다. 가설을 채택하기 위해서는 해당 경로계수의 p-값이 0.05보다 작아야 한다. 또한 경로계수의 부호는 가설화된 관계의 방향과 일치해야 한다.

4. lavaan 패키지를 이용한 구조방정식모델링

여기에서는 lavaan 패키지에 포함되어 있는 PoliticalDemocracy 데이터셋(dataset)을 이용하여 구조방정식모델링 분석절차를 살펴본다(lavaan 2019).[7] 먼저 다음과 같이 lavaan 패키지를 메모리에 적재하고 데이터셋의 구조를 살펴본다.[8] 이 데이터셋은 한 시점에서의 산업화 수준과 두 시점에서의 민주화 수준을 다양한 관점에서 측정한 변수들로 구성되어 있다.

```
> library(lavaan)
> str(PoliticalDemocracy)
'data.frame':    75 obs. of 11 variables:
 $ y1: num  2.5 1.25 7.5 8.9 10 7.5 7.5 7.5 2.5 10 ...
 $ y2: num  0 0 8.8 8.8 3.33 ...
 $ y3: num  3.33 3.33 10 10 10 ...
 $ y4: num  0 0 9.2 9.2 6.67 ...
 $ y5: num  1.25 6.25 8.75 8.91 7.5 ...
 $ y6: num  0 1.1 8.09 8.13 3.33 ...
 $ y7: num  3.73 6.67 10 10 10 ...
 $ y8: num  3.333 0.737 8.212 4.615 6.667 ...
 $ x1: num  4.44 5.38 5.96 6.29 5.86 ...
 $ x2: num  3.64 5.06 6.26 7.57 6.82 ...
 $ x3: num  2.56 3.57 5.22 6.27 4.57 ...
```

PoliticalDemocracy 데이터셋에는 모두 11개의 관측변수가 포함되어 있다. 각 관측변수는 세 개의 잠재변수를 측정하도록 개발되었다. y1, y2, y3, y4 변수는 1960년 시점에서의 민주화 수준(dem60)을 측정하였고, y5, y6, y7, y8 변수는 1965년 시점에서의 민주화 수준(dem65)을 측정하였다. x1, x2, x3 변수는 1960년 시점에서의 산업화 수준(ind60)을 측정하였다. 이들 관측변수와 잠재변수 간 관계, 그리고 연구가설에 해당하는 잠재변수 간의 관계는 <그림 2>와 같다.[9]



<그림 2> 연구모델

7) PoliticalDemocracy 데이터셋과 그 안에 포함된 11개 변수에 대한 설명은 데이터셋에 대한 도움말을 참고한다(?PoliticalDemocracy).
8) lavaan 패키지가 설치되어 있지 않은 경우에는 install.packages() 함수를 이용하여 먼저 다음과 같이 패키지를 설치한다: install.packages("lavaan"). 본 튜토리얼에서 사용하는 패키지는 모두 사전에 설치되어 있다고 가정한다.
9) 본 튜토리얼에서 다루는 연구모델은 설명을 위해 예시로 제시된 것이다. 따라서 이에 대해 이론적 해석이나 실질적 의미를 부여해서는 안된다.

<그림 2>의 연구모델은 개발도상국에서의 산업화와 민주화 간의 영향관계를 설명한다. 현재(예를 들면, 여기에서는 1965년)의 민주화 수준은 과거(예를 들면, 1960년)의 민주화 수준 및 산업화 수준에 의해 결정된다고 가정한다. 산업화 수준은 또한 민주화 수준에 영향을 미친다(예를 들면, 1960년의 산업화 수준은 그 해의 민주화 수준에 영향을 미친다). <그림 2>의 구조방정식모델에서 1960년의 산업화 수준은 외생잠재변수로서 모델 외부의 요인에 의해 설명된다고 가정한다. 1960년과 1965년의 민주화 수준은 내생잠재변수로서 모델 내의 요인(즉 1960년의 산업화 수준과 1960년의 민주화 수준)에 의해 설명된다고 가정한다. 구조방정식모델링을 통해 이들 잠재변수 간의 영향관계를 검정한다.

4.1 확인적 요인분석

먼저 확인적 요인분석을 수행하여 측정모델을 평가해보자. 확인적 요인분석을 수행하기 위해서는 측정모델을 생성해야 한다. 이를 위해 다음과 같이 잠재변수와 관측변수 간 측정모델의 구조를 문자열로 저장한다.

```
> cfa <- "i n d 60 =~ x1 + x2 + x3
+         dem60 =~ y1 + y2 + y3 + y4
+         dem65 =~ y5 + y6 + y7 + y8"
```

각 행은 하나의 잠재변수를 나타낸다. 잠재변수는 대응되는 관측변수들을 바탕으로 다음과 같은 형식으로 정의한다.

잠재변수 =~ 관측변수1 + 관측변수2 + 관측변수3 + … + 관측변수*n*

연산자 =~는 왼쪽의 변수는 오른쪽의 변수에 의해 측정된다는 의미이다. 여기에서 =~ 왼쪽에는 잠재변수가 오고 오른쪽에는 관측변수가 위치한다.

측정모델에 대한 확인적 요인분석은 cfa() 함수를 이용하여 수행한다. cfa() 함수의 첫 번째 인수(model)에 앞서 생성한 측정모델 문자열을 지정하고 data 인수에 관측변수가 포함된 데이터셋을 지정한다. cfa() 함수를 실행하면 다음과 같이 모델추정과 관련된 기본적인 정보가 출력된다.

```
> cfa(model=cfa, data=PoliticalDemocracy)
lavaan 0.6-3 ended normally after 47 iterations

  Optimization method                           NLMINB
  Number of free parameters                         25

  Number of observations                            75

  Estimator                                         ML
  Model Fit Test Statistic                      72.462
  Degrees of freedom                                41
  P-value (Chi-square)                           0.002
```

cfa() 함수는 복잡한 모델의 모수설정 과정 없이 사전에 설정된 기본옵션을 바탕으로 측정모델의 모수를 추정한다. 기본옵션으로는, 예를 들면, 첫째, 각 잠재변수의 첫 번째 관측변수의 요인적재값이 1로 고정된다. 잠재변수는 직접 관찰되지 않으므로 명확히 계량화된 척도가 없다. 하지만 모수추정을 위해서는 잠재변수도 척도를 가져야 한다. 요인적재값을 1로 제약함으로써 잠재변수의 측정단위를 준거변수로 지정된 관측변수(일반적으로 첫 번째 관측변수)와 같도록 하여 잠재변수의 척도를 결정할 수 있다. 둘째, 잔차분산(residual variance)에 대한 추정이 자동으로 추가된다. 셋째, 외생잠재변수는 공분산을 갖도록 설정된다. 넷째, 결측값은 목록별제외(listwise deletion) 방식으로 처리된다. 즉 측정항목 가운데 하나라도 결측값이 있으면 해당 케이스 전체가 분석에서 제외된다.

이들 기본옵션은 인수 지정을 통해 언제든 변경할 수 있다. 예를 들어, 잠재변수의 척도를 결정하는 방법으로서 잠재변수를 평균 0과 분산 1을 갖도록 제약하는 방법을 선택할 수도 있다. 즉 잠재변수의 분산이 단위분산(즉 1)이 되도록 잠재변수의 측정단위를 표준화한

다. 이 경우 요인적재값은 자유모수로서 모두 자유롭게 추정된다. 이렇게 하면 잠재변수 간의 공분산을 좀 더 해석이 용이한 상관계수로 변환할 수 있고 각 잠재변수에 대한 첫 번째 관측변수의 요인적재값을 온전히 추정하고 검정할 수 있다. 잠재변수를 표준화하기 위해서는 std.lv=TRUE를 지정한다. 결측값을 처리하는 방법도 변경할 수 있다. lavaan 패키지는 모수추정 방식으로서 기본적으로 최대우도법(maximum likelihood)을 사용한다(estimator="ML"). 이처럼 최대우도법을 채택하고 있을 경우 결측값 처리 방법으로 FIML(full information maximum likelihood)을 선택할 수 있다. FIML은 다중대체법(multiple imputation)과 유사한 방식으로 결측값을 처리하여 결측값이 포함된 행을 모두 삭제함으로써 생길 수 있는 자료의 손실이나 단일값으로 대체함으로써 생길 수 있는 편향의 위험을 줄일 수 있다. FIML에 의한 결측값 처리를 위해서는 missing="fiml"을 지정한다.

lavaan 패키지의 모델추정 함수에는 모델을 해석하고, 추정하고, 출력하는 방식과 관련된 다양한 옵션이 포함되어 있다. lavOptions() 함수의 도움말을 참고하면 모델추정 시 지정할 수 있는 옵션의 종류를 살펴볼 수 있다. 또는 lavOptions() 함수에 옵션의 이름을 지정함으로써 해당 옵션의 기본설정을 확인할 수 있다.

```
> ?lavOptions
> lavOptions("std.lv")
$`std.lv`
[1] FALSE
```

확인적 요인분석의 결과를 보기 위해서는 다음과 같이 summary() 함수를 이용한다. 인수로서 fit.measures=TRUE와 standardized=TRUE를 지정하면 모델적합도 지표와 표준화 해를 출력한다. 출력결과는 일반적인 상용 구조방정식모델링 소프트웨어와 유사하다.

```
> fit <- cfa(model=cfa, data=PoliticalDemocracy)
> summary(fit, fit.measures=TRUE, standardized=TRUE)
lavaan 0.6-3 ended normally after 47 iterations

  Optimization method                           NLMINB
  Number of free parameters                         25

  Number of observations                            75

  Estimator                                         ML
  Model Fit Test Statistic                      72.462
  Degrees of freedom                                41
  P-value (Chi-square)                           0.002

Model test baseline model:

  Minimum Function Test Statistic              730.654
  Degrees of freedom                                55
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    0.953
  Tucker-Lewis Index (TLI)                       0.938

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -1564.959
  Loglikelihood unrestricted model (H1)      -1528.728

  Number of free parameters                         25
  Akaike (AIC)                                3179.918
  Bayesian (BIC)                              3237.855
  Sample-size adjusted Bayesian (BIC)         3159.062

Root Mean Square Error of Approximation:

  RMSEA                                          0.101
  90 Percent Confidence Interval          0.061  0.139
  P-value RMSEA <= 0.05                          0.021

Standardized Root Mean Square Residual:

  SRMR                                           0.055

Parameter Estimates:

  Information                                 Expected
  Information saturated (h1) model          Structured
  Standard Errors                             Standard
```

Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
ind60 =~						
x1	1.000				0.669	0.920
x2	2.182	0.139	15.714	0.000	1.461	0.973
x3	1.819	0.152	11.956	0.000	1.218	0.872
dem60 =~						
y1	1.000				2.201	0.845
y2	1.354	0.175	7.755	0.000	2.980	0.760
y3	1.044	0.150	6.961	0.000	2.298	0.705
y4	1.300	0.138	9.412	0.000	2.860	0.860
dem65 =~						
y5	1.000				2.084	0.803
y6	1.258	0.164	7.651	0.000	2.623	0.783
y7	1.282	0.158	8.137	0.000	2.673	0.819
y8	1.310	0.154	8.529	0.000	2.730	0.847

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
ind60 ~~						
dem60	0.660	0.206	3.202	0.001	0.448	0.448
dem65	0.774	0.208	3.715	0.000	0.555	0.555
dem60 ~~						
dem65	4.487	0.911	4.924	0.000	0.978	0.978

Variances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.x1	0.082	0.020	4.180	0.000	0.082	0.154
.x2	0.118	0.070	1.689	0.091	0.118	0.053
.x3	0.467	0.090	5.174	0.000	0.467	0.240
.y1	1.942	0.395	4.910	0.000	1.942	0.286
.y2	6.490	1.185	5.479	0.000	6.490	0.422
.y3	5.340	0.943	5.662	0.000	5.340	0.503
.y4	2.887	0.610	4.731	0.000	2.887	0.261
.y5	2.390	0.447	5.351	0.000	2.390	0.355
.y6	4.343	0.796	5.456	0.000	4.343	0.387
.y7	3.510	0.668	5.252	0.000	3.510	0.329
.y8	2.940	0.586	5.019	0.000	2.940	0.283
ind60	0.448	0.087	5.169	0.000	1.000	1.000
dem60	4.845	1.088	4.453	0.000	1.000	1.000
dem65	4.345	1.051	4.134	0.000	1.000	1.000

출력결과는 크게 세 섹션으로 나누어져 있다. 가장 먼저 나타나는 부분은 헤더(header)라고 불리며 다음과 같은 정보를 포함한다: lavaan 패키지 버전, 해가 수렴했는지 여부와 반복 횟수, 관측값 개수, 모델추정 방법, 모델적합도 검정결과(카이제곱 검정통계량, 자유도, p-값). 다음 섹션은 Chisq 적합도 지표 이외의 추가적인 적합도 지표를 보여준다. 이 섹션은 'Model test baseline model'부터 시작하여 'SRMR'로 끝난다. 여기에는 <표 1>에 소개된 적합도 지표를 비롯한 다양한 지표가 제시되어 있다. 마지막 섹션은 모수추정치를 담고 있다. 잠재변수, 공분산, 잔차분산의 순서로 모수추정치를 볼 수 있다.

4.2 모수추정치

parameterEstimates() 함수를 이용하여 모델에 포함된 모수추정치 전체를 다음과 같이 추출할 수 있다. standardized=TRUE를 지정하면 표준화된 추정치가 함께 출력된다. standardizedSolution() 함수 또한 parameterEstimates() 함수와 유사하나, 표준화된 모수추정치만 보여준다.

```
> parameterEstimates(fit, standardized=TRUE)
```

	lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper	std.lv	std.all	std.nox
1	ind60	=~	x1	1.000	0.000	NA	NA	1.000	1.000	0.669	0.920	0.920
2	ind60	=~	x2	2.182	0.139	15.714	0.000	1.910	2.454	1.461	0.973	0.973
3	ind60	=~	x3	1.819	0.152	11.956	0.000	1.521	2.117	1.218	0.872	0.872
4	dem60	=~	y1	1.000	0.000	NA	NA	1.000	1.000	2.201	0.845	0.845
5	dem60	=~	y2	1.354	0.175	7.755	0.000	1.012	1.696	2.980	0.760	0.760
6	dem60	=~	y3	1.044	0.150	6.961	0.000	0.750	1.338	2.298	0.705	0.705
7	dem60	=~	y4	1.300	0.138	9.412	0.000	1.029	1.570	2.860	0.860	0.860
8	dem65	=~	y5	1.000	0.000	NA	NA	1.000	1.000	2.084	0.803	0.803
9	dem65	=~	y6	1.258	0.164	7.651	0.000	0.936	1.581	2.623	0.783	0.783
10	dem65	=~	y7	1.282	0.158	8.137	0.000	0.974	1.591	2.673	0.819	0.819
11	dem65	=~	y8	1.310	0.154	8.529	0.000	1.009	1.611	2.730	0.847	0.847
12	x1	~~	x1	0.082	0.020	4.180	0.000	0.043	0.120	0.082	0.154	0.154
13	x2	~~	x2	0.118	0.070	1.689	0.091	−0.019	0.256	0.118	0.053	0.053
14	x3	~~	x3	0.467	0.090	5.174	0.000	0.290	0.644	0.467	0.240	0.240
15	y1	~~	y1	1.942	0.395	4.910	0.000	1.167	2.717	1.942	0.286	0.286
16	y2	~~	y2	6.490	1.185	5.479	0.000	4.168	8.811	6.490	0.422	0.422
17	y3	~~	y3	5.340	0.943	5.662	0.000	3.491	7.188	5.340	0.503	0.503
18	y4	~~	y4	2.887	0.610	4.731	0.000	1.691	4.083	2.887	0.261	0.261
19	y5	~~	y5	2.390	0.447	5.351	0.000	1.515	3.266	2.390	0.355	0.355
20	y6	~~	y6	4.343	0.796	5.456	0.000	2.783	5.903	4.343	0.387	0.387
21	y7	~~	y7	3.510	0.668	5.252	0.000	2.200	4.819	3.510	0.329	0.329
22	y8	~~	y8	2.940	0.586	5.019	0.000	1.792	4.089	2.940	0.283	0.283
23	ind60	~~	ind60	0.448	0.087	5.169	0.000	0.278	0.618	1.000	1.000	1.000
24	dem60	~~	dem60	4.845	1.088	4.453	0.000	2.713	6.977	1.000	1.000	1.000
25	dem65	~~	dem65	4.345	1.051	4.134	0.000	2.285	6.404	1.000	1.000	1.000
26	ind60	~~	dem60	0.660	0.206	3.202	0.001	0.256	1.065	0.448	0.448	0.448
27	ind60	~~	dem65	0.774	0.208	3.715	0.000	0.366	1.182	0.555	0.555	0.555
28	dem60	~~	dem65	4.487	0.911	4.924	0.000	2.701	6.274	0.978	0.978	0.978

lhs, op, rhs는 모델의 구조를 나타낸다. op 열의 연산자는 lhs 열의 변수와 rhs 열의 변수 간의 관계를 정의한다. =~ 연산자는 변수 간의 측정관계를 나타내며, ~~ 연산자는 분산 또는 공분산 관계를 나타낸다. est 열은 추정치를 나타내며, std.all 열에는 모든 변수(잠재변수와 관측변수)를 표준화하여 산출된 추정치의 표준화값(완전표준화해, completely standardized solution)이 출력된다. std.lv는 잠재변수만을 표준화한 값이고, std.nox는 외생관측변수를 제외한 모든 변수를 표준화한 값이다. =~ 연산자에 의해 정의된 첫 번째 그룹의 모수추정치들은 잠재변수에 대한 관측변수의 요인적재값이다. 예를 들어, ind60에 대한 x1의 요인적재값은 모델설정 시 1로 고정하였으므로 1이 출력되었으며, 표준화된 요인적재값은 0.920이다. ~~ 연산자에 의해 정의된 두 번째 그룹의 모수추정치들은 분산(lhs와 rhs가 같으면) 또는 공분산(lhs와 rhs가 다르면)을 나타낸다. 관측변수의 분산은 잔차분산(residual variance)을 의미한다. 여기에서 잔차분산은 한 개을 제외하고는(x2) 모두 통

계적으로 유의하다(즉 유의하게 0과 다르다). 잔차분산은 모두 0보다 크며, 이는 잠재변수가 관측변수를 완벽하게 예측하지 못한다는 것을 나타낸다. 하지만 어떠한 잠재변수도 관측변수를 완벽하게 설명할 수는 없기 때문에 이는 매우 일반적으로 관찰되는 현상이다.

parameterEstimates() 함수는 모수추정치를 R의 데이터프레임(data frame) 형식으로 출력하기 때문에 이를 이용하면 보다 다양한 방식으로 데이터를 가공할 수 있다. 예를 들어, 다음과 같이 =~ 연산자로 정의된 요인적재값을 추출하여 관련 정보와 함께 하나의 요약 테이블로 만들 수 있다. knitr 패키지의 kable() 함수는 테이블을 좀 더 읽기 쉽도록 만들어준다.

```
> library(knitr)
> options(knitr.kable.NA="")
> library(dplyr)
> parameterEstimates(fit, standardized=TRUE) %>%
+   filter(op=="=~") %>%
+   mutate(stars=ifelse(pvalue < 0.001, "***",
+                       ifelse(pvalue < 0.01, "**",
+                              ifelse(pvalue < 0.05, "*", "")))) %>%
+   select("Latent Factor"=lhs, Indicator=rhs, B=est, SE=se,
+          Z=z, "p-value"=pvalue, Sig.=stars, Beta=std.all) %>%
+   kable(digits=3, format="pandoc", caption="Factor Loadings")
```

Table: Factor Loadings

Latent Factor	Indicator	B	SE	Z	p-value	Sig.	Beta
ind60	x1	1.000	0.000				0.920
ind60	x2	2.182	0.139	15.714	0	***	0.973
ind60	x3	1.819	0.152	11.956	0	***	0.872
dem60	y1	1.000	0.000				0.845
dem60	y2	1.354	0.175	7.755	0	***	0.760
dem60	y3	1.044	0.150	6.961	0	***	0.705
dem60	y4	1.300	0.138	9.412	0	***	0.860
dem65	y5	1.000	0.000				0.803
dem65	y6	1.258	0.164	7.651	0	***	0.783
dem65	y7	1.282	0.158	8.137	0	***	0.819
dem65	y8	1.310	0.154	8.529	0	***	0.847

4.3 상관계수 잔차 행렬

확인적 요인분석의 목표는 관측변수들을 연결하는 잠재된 구조(즉 잠재변수와의 연결관계)를 통해 관측변수 간의 관계를 설명하는 것이다. 예를 들어, 지금의 모델에서 우리는 x1, x2, x3 변수들이 서로 상관관계를 갖는다고 가정한다. 왜냐하면 이 변수들은 모두 동일한 대상, 즉 1960년 시점에서의 산업화 수준(ind60)을 각기 다른 방식으로 측정하고 있기 때문이다. 또한 동일한 잠재변수를 공유하지 않더라도 관측변수 간에 기대되는 상관관계가 존재할 수 있다. 예를 들어, 비록 x1과 y1은 다른 대상(각각 ind60과 dem60)을 측정하고 있지만 둘 간에는 어느 정도 상관관계가 존재할 것으로 기대한다. 왜냐하면 현재의 연구모델에 따르면 1960년 시점에서의 산업화 수준과 민주화 수준은 서로 관련이 있다고 보는 것이 자연스러울 수 있기 때문이다.

이처럼 측정모델은 관측변수 간의 기대된 상관관계

를 표현한 것이기 때문에 모델의 성능을 평가하는 한 가지 방법은 모델이 기대하는 상관계수 행렬(또는 공분산 행렬)과 데이터로부터 얻은 실제 관측된 상관계수 행렬(또는 공분산 행렬) 간의 차이를 살펴보는 것이다. 이 차이가 바로 측정모델의 잔차이다. 어떤 두 변수 간 잔차가 크면 그 변수 간 관계에 대해 모델이 제대로 포착하지 못하는 부분이 존재한다는 것을 의미한다.

상관계수의 잔차 행렬은 다음과 같이 residuals() 함수를 이용하여 구할 수 있다. 인수로서 type="cor"를 지정한다. 잔차 행렬은 출력된 리스트(list) 객체의 cov 원소에 포함되어 있다. 대각선의 불필요한 0은 삭제하고, kable() 함수를 이용하여 테이블을 좀 더 읽기 쉽도록 만든다.

```
> residuals(fit, type="cor")$cov
```

```
      x1     x2     x3     y1     y2     y3     y4     y5     y6     y7     y8
x1  0.000
x2 -0.001  0.000
x3 -0.003  0.002  0.000
y1  0.034 -0.047 -0.083  0.000
y2 -0.099 -0.082 -0.086 -0.038  0.000
y3  0.037  0.005 -0.050  0.083 -0.085  0.000
y4  0.114  0.068  0.054 -0.033  0.066  0.002  0.000
y5  0.155  0.089  0.043  0.075 -0.054  0.022 -0.024  0.000
y6 -0.054 -0.068 -0.047  0.003  0.123 -0.113  0.000 -0.064  0.000
y7 -0.029 -0.040 -0.044 -0.002 -0.028  0.085 -0.008  0.020 -0.032  0.000
y8  0.032 -0.002 -0.039 -0.034 -0.024 -0.054  0.025 -0.050  0.090  0.018  0.000
```

```
> resid.cor <- residuals(fit, type="cor")$cov
> resid.cor[upper.tri(resid.cor, diag=TRUE)] <- NA
> library(knitr)
> kable(resid.cor, digits=2, format="pandoc", caption="Residual Correlations")
```

Table: Residual Correlations

	x1	x2	x3	y1	y2	y3	y4	y5	y6	y7	y8
x1											
x2	0.00										
x3	0.00	0.00									
y1	0.03	-0.05	-0.08								
y2	-0.10	-0.08	-0.09	-0.04							
y3	0.04	0.01	-0.05	0.08	-0.09						
y4	0.11	0.07	0.05	-0.03	0.07	0.00					
y5	0.16	0.09	0.04	0.08	-0.05	0.02	-0.02				
y6	-0.05	-0.07	-0.05	0.00	0.12	-0.11	0.00	-0.06			
y7	-0.03	-0.04	-0.04	0.00	-0.03	0.09	-0.01	0.02	-0.03		
y8	0.03	0.00	-0.04	-0.03	-0.02	-0.05	0.02	-0.05	0.09	0.02	

상관계수 잔차가 0.1보다 큰 변수들이 있는지 살펴보자. 일부 변수 간 관계(x1과 y4, y2와 y6, y3와 y6)를 제외하고 대부분의 잔차는 양호하다. 모델적합도 지표 가운데 RMSEA(root mean square of error approximation), RMR(root mean square residual), SRMR(standardized root mean square residual) 등은 모두 잔차를 바탕으로 계산된다. 상관계수 잔차가 전반적으로 작을수록 이들 지표는 작은 값을 갖게 된다. 따라서 작은 지표값은 모델의 성능이 우수하다는 것을 나타낸다. 왜냐하면 이는 모델에 의한 예측 상관계수 행렬과 관측값에 의한 실제 상관계수 행렬이 서로 유사하다는(즉 잔차가 작다는) 것을 의미하기 때문이다.

4.4 적합도 지표와 모델 개선

fitMeasures() 함수는 lavaan 패키지에 의해 산출되는 적합도 지표를 벡터 형식으로 출력한다.

```
> fitMeasures(fit)
```

npar	fmin	chisq	df
25.000	0.483	72.462	41.000
pvalue	baseline.chisq	baseline.df	baseline.pvalue
0.002	730.654	55.000	0.000
cfi	tli	nnfi	rfi
0.953	0.938	0.938	0.867
nfi	pnfi	ifi	rni
0.901	0.672	0.954	0.953
logl	unrestricted.logl	aic	bic
−1564.959	−1528.728	3179.918	3237.855
ntotal	bic2	rmsea	rmsea.ci.lower
75.000	3159.062	0.101	0.061
rmsea.ci.upper	rmsea.pvalue	rmr	rmr_nomean
0.139	0.021	0.428	0.428
srmr	srmr_bentler	srmr_bentler_nomean	srmr_bollen
0.055	0.055	0.055	0.055
srmr_bollen_nomean	srmr_mplus	srmr_mplus_nomean	cn_05
0.055	0.055	0.055	59.937
cn_01	gfi	agfi	pgfi
68.225	0.854	0.765	0.531
mfi	ecvi		
0.811	1.633		

원하는 지표만을 출력하기 위해서는 다음과 같이 fitMeasures() 함수의 두 번째 인수로 지표명을 지정한다.

```
> fitMeasures(fit, c("chisq", "df", "pvalue", "gfi", "rmsea", "cfi"))
 chisq     df pvalue    gfi  rmsea    cfi
72.462 41.000  0.002  0.854  0.101  0.953
```

간명적합도(Chisq/df < 3)와 CFI(CFI > 0.9)를 제외하고는 적합도 지표들이 모두 권장 수준(p-값 > 0.05, GFI > 0.9, RMSEA < 0.08)을 충족하지 못하고 있다 (<표 1> 참고).

수정 지표를 이용하여 모델을 수정하고 모델 적합도를 개선해보자. 수정 지표는 다음과 같이 summary() 함수에 modindices=TRUE를 지정하거나 modindices() 함수를 직접 실행하여 구할 수 있다.

```
> summary(fit, modindices=TRUE)
> modindices(fit)
```

modindices() 함수는 결과를 데이터프레임 형식으로 출력하기 때문에 원하는 부분만을 추출하거나 지표값의 크기에 따라 소팅할 수 있다. 예를 들어, 요인적재값에 대한 수정 지표는 다음과 같이 테이블 형식으로 출력할 수 있다.

```
> library(dplyr)
> library(knitr)
> modindices(fit) %>%
+     filter(op=="=~") %>%
+     kable(digits=3, format="pandoc",
+           caption="Modification Indices for Factor Loadings")
```

Table: Modification Indices for Factor Loadings

lhs	op	rhs	mi	epc	sepc.lv	sepc.all	sepc.nox
ind60	=~	y1	0.623	-0.269	-0.180	-0.069	-0.069
ind60	=~	y2	2.157	-0.826	-0.553	-0.141	-0.141
ind60	=~	y3	0.004	0.032	0.022	0.007	0.007
…(중략)							
dem65	=~	y2	1.531	-1.520	-3.169	-0.808	-0.808
dem65	=~	y3	0.056	-0.253	-0.526	-0.161	-0.161
dem65	=~	y4	2.563	1.553	3.237	0.973	0.973

mi 열은 모수가 추가될 경우의 카이제곱 변화량을 나타낸다. 이 값은 클수록 모델적합도의 개선효과가 크다는 것을 의미한다. epc 열은 수정된 모델에서의 모수추정치의 변화량(expected parameter change, EPC)을 나타낸다. 수정 지표는 기존의 모델에 포함되지 않은 경로를 수정 가능한 후보 경로로 제안하기 때문에 EPC값은 모수값 0으로부터의 변화량을 나타내며, 따라서 이 값은 결국 수정 모델에서의 모수추정치를 의미한다. 나머지 열은 세 종류의 표준화된 EPC값이다. sepc.lv는 잠재변수만을 표준화한 값이고, sepc.all은 모든 변수를 표준화한 값이며, sepc.nox는 외생관측변수를 제외한 모든 변수를 표준화한 값이다.

우리는 일반적으로 어떤 모수를 추가할 때 적합도가 가장 크게 개선될 수 있는지에 관심을 갖는다. 다음과 같이 수정 지표를 크기 순으로 정렬하면 적합도 개선효과가 큰 모수를 확인할 수 있다. 수정 지표값이 최소 3 이상인 모수만을 출력하였다(minimum.value=3).

예를 들어, 여기에서 첫 번째 행의 y2 ~~ y6 표현식은 두 관측변수 y2와 y6 잔차분산의 상관관계를 허용하는 경우를 나타낸다. 두 관측변수가 잠재변수에 의해 설명되지 못하는 공통의 무언가를 갖고 있다고 판단될 때 이러한 모델설정이 가능하다. 지금의 경우 두 관측변수 y2와 y6는 동일한 측정항목('정치적 반대의 자유')을 연도만 달리하여 측정한다(각각 1960년과 1965년).

```
> modindices(fit, sort.=TRUE, minimum.value=3)
      lhs   op rhs     mi    epc  sepc.lv sepc.all sepc.nox
88     y2  ~~  y6  9.279  2.129    2.129    0.401    0.401
104    y6  ~~  y8  8.668  1.513    1.513    0.423    0.423
81     y1  ~~  y5  8.183  0.884    0.884    0.410    0.410
93     y3  ~~  y6  6.574 -1.590   -1.590   -0.330   -0.330
79     y1  ~~  y3  5.204  1.024    1.024    0.318    0.318
86     y2  ~~  y4  4.911  1.432    1.432    0.331    0.331
94     y3  ~~  y7  4.088  1.152    1.152    0.266    0.266
33 ind60  =~  y5  4.007  0.762    0.510    0.197    0.197
54     x1  ~~  y2  3.785 -0.192   -0.192   -0.263   -0.263
32 ind60  =~  y4  3.568  0.811    0.543    0.163    0.163
85     y2  ~~  y3  3.215 -1.365   -1.365   -0.232   -0.232
100    y5  ~~  y6  3.116 -0.774   -0.774   -0.240   -0.240
102    y5  ~~  y8  3.090 -0.686   -0.686   -0.259   -0.259
```

따라서 이들의 잔차분산 간 상관관계가 존재할 것으로 기대하는 것은 합리적이고 타당하다. 마찬가지 논리로 y1 ~~ y5, y3 ~~ y7, y4 ~~ y8 등의 모수설정도 적용 가능하다. 이러한 이론적 배경과 수정 지표를 참고하여 다음과 같이 모델을 수정하고 확인적 요인분석을 다시 수행한다. 표현식 y2 ~~ y4와 y2 ~~ y6는 합쳐서 y2 ~~ y4 + y6로 나타낼 수 있다.

```
> cfa2 <- "ind60 =~ x1 + x2 + x3
+          dem60 =~ y1 + y2 + y3 + y4
+          dem65 =~ y5 + y6 + y7 + y8
+          y1 ~~ y5
+          y2 ~~ y4 + y6
+          y3 ~~ y7
+          y4 ~~ y8
+          y6 ~~ y8"
> fit2 <- cfa(model=cfa2, data=PoliticalDemocracy)
> fit2
lavaan 0.6-3 ended normally after 78 iterations

  Optimization method                           NLMINB
  Number of free parameters                         31

  Number of observations                            75

  Estimator                                         ML
  Model Fit Test Statistic                      38.125
  Degrees of freedom                                35
  P-value (Chi-square)                           0.329
```

모수를 추가한 모델과 기존 모델 간 비교는 anova() 함수를 이용한다. anova() 함수는 구조방정식모델뿐만 아니라 다양한 유형의 중첩모델(nested model)을 비교할 때 사용할 수 있다.[10]

10) 여기에서 중첩모델이란 추정할 모수의 관점에서 한 모델이 다른 모델의 완전한 부분집합의 형태인 모델을 의미한다. 이는 작은 모델에 포함된 모수들은 모두 큰 모델에도 포함되어 있어야 한다는 뜻이다.

```
> anova(fit, fit2)
Chi Square Difference Test

     Df     AIC     BIC   Chisq Chisq diff Df diff Pr(>Chisq)
fit2 35  3157.6  3229.4  38.125
fit  41  3179.9  3237.9  72.462     34.336       6  5.792e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

anova() 함수는 두 모델에 대한 카이제곱 차이검정을 수행한다. 검정결과에 따르면 기존 모델(fit)과 여섯 개의 모수를 추가한 새로운 모델(fit2) 간의 카이제곱 변화량은 통계적으로 유의한 차이이다($\chi^2(6)=34.336$, p-값 < 0.01). 따라서 관측변수의 잔차분산 간 상관관계를 허용한 새로운 모델이 기존 모델보다 우수하다고 결론 내릴 수 있다. 새로운 모델의 주요 적합도 지표는 다음과 같다.

```
> fitMeasures(fit2, c("chisq", "df", "pvalue", "gfi", "rmsea", "cfi"))
  chisq      df  pvalue     gfi   rmsea     cfi
 38.125  35.000   0.329   0.923   0.035   0.995
```

기존 모델과의 비교를 위해 다음과 같이 여러 모델의 주요 적합도 지표를 한꺼번에 출력할 수 있는 간단한 함수를 생성한다.

```
> library(dplyr)
> library(tibble)
> library(magrittr)
> compareFit <- function(...) {
+   m <- list(...)
+   sapply(m, fitMeasures) %>%
+     set_colnames(paste0("Model", 1:length(m))) %>%
+     as.data.frame() %>%
+     rownames_to_column("Fit_Measures") %>%
+     slice(match(c("chisq", "df", "pvalue",
+                   "gfi", "rmsea", "cfi"), Fit_Measures)) %>%
+     mutate(Fit_Measures=c("Chi-square", "df", "p-value",
+                           "GFI", "RMSEA", "CFI"))}
```

이 함수를 이용하여 기존 모델과 새로운 모델 간의 적합도 지표를 다음과 같이 하나의 테이블로 출력할 수 있다. 여기에서는 테이블 형태의 출력을 위해 kable() 함수 대신에 stargazer() 함수를 이용하였다. stargazer() 함수는 kable() 함수와는 다른 형태의 테이블을 생성할 수 있다.

```
> library(stargazer)
> compareFit(fit, fit2) %>%
+   stargazer(type="text", title="Model Comparison", summary=FALSE,
+             digits=3, digits.extra=0, rownames=FALSE)
```

```
Model Comparison
=======================
Fit_Measures Model1 Model2
-----------------------
Chi-square   72.462  38.125
df              41      35
p-value      0.002   0.329
GFI          0.854   0.923
RMSEA        0.101   0.035
CFI          0.953   0.995
-----------------------
```

새로운 모델의 적합도 지표는 기존 모델에 비해 모두 개선되었다. 모든 적합도 지표가 권장 수준(Chisq/df < 3, p-값 > 0.05, GFI > 0.9, RMSEA < 0.08, CFI > 0.9)을 충족한다(<표 1> 참고). 관측변수의 잔차분산 간 상관관계를 허용하여 여섯 개의 모수를 새로 추정한 모델이 기존의 모델에 비해 데이터를 보다 더 잘 적합시킨다. 수정된 측정모델의 적합도는 전반적으로 양호하며, 따라서 모델의 적합도 관점에서 측정모델의 단일차원성은 확보된 것으로 보인다.

단일차원성이 충족되기 위해서는 또한 각 잠재변수에 대해 모든 관측변수가 적정 수준(일반적으로 0.6) 이상의 표준화 요인적재값을 가지면서 동시에 이들이 통계적으로 유의해야 한다. 표준화 요인적재값과 통계적 유의성은 standardizedsolution() 함수의 실행결과를 바탕으로 다음과 같이 요약 테이블 형식으로 출력할 수 있다.

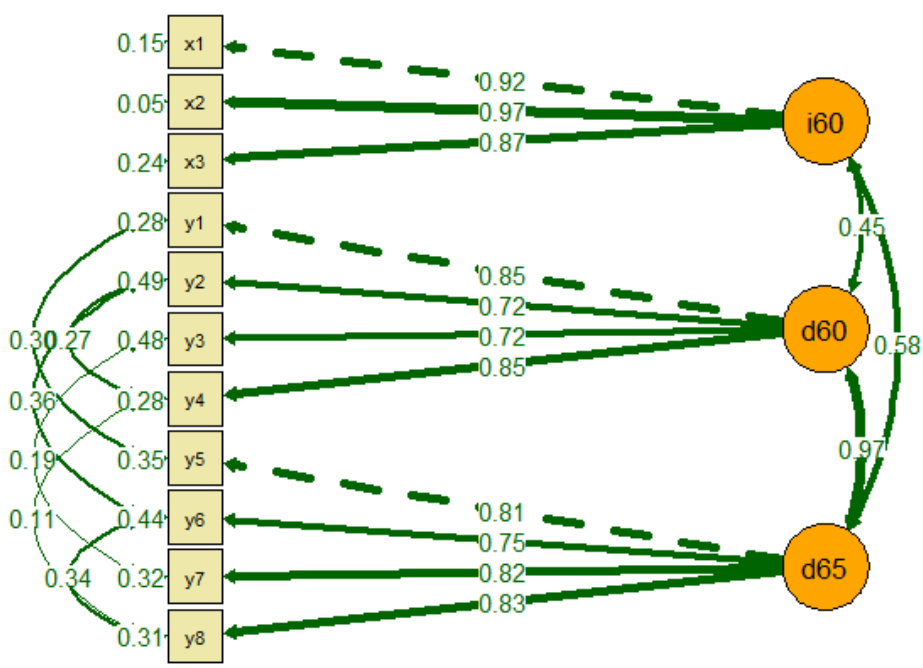
```
> standardizedsolution(fit2)
      lhs op      rhs est.std    se      z pvalue ci.lower ci.upper
1   ind60 =~       x1   0.920 0.023 40.005  0.000    0.875    0.965
2   ind60 =~       x2   0.973 0.016 59.110  0.000    0.941    1.005
3   ind60 =~       x3   0.872 0.031 28.081  0.000    0.811    0.933
… (중략)
32  ind60 ~~ dem60    0.447 0.103  4.323  0.000    0.244    0.649
33  ind60 ~~ dem65    0.578 0.088  6.559  0.000    0.405    0.750
34 dem60 ~~ dem65    0.967 0.029 33.579  0.000    0.910    1.023
> library(dplyr)
> library(stargazer)
> standardizedsolution(fit2) %>%
+   filter(op=="=~") %>%
+   mutate(stars=ifelse(pvalue < 0.001, "***",
+                       ifelse(pvalue < 0.01, "**",
+                              ifelse(pvalue < 0.05, "*", "")))) %>%
+   select(Construct=lhs, Item=rhs, "Factor Loading"=est.std,
+          Z=z, "p-value"=pvalue, Sig.=stars) %>%
+   stargazer(type="text", title="Convergent Validity: Factor Loadings", summary=FALSE,
+             digits=3, digits.extra=0, rownames=FALSE)

Convergent Validity: Factor Loadings
=====================================
Construct Item Factor Loading  Z   p-value Sig.
-------------------------------------
ind60     x1      0.920      40.005   0     ***
ind60     x2      0.973      59.110   0     ***
ind60     x3      0.872      28.081   0     ***
```

```
dem60    y1    0.850    20.200    0    * * *
dem60    y2    0.717    11.157    0    * * *
dem60    y3    0.722    11.439    0    * * *
dem60    y4    0.846    19.462    0    * * *
dem65    y5    0.808    16.847    0    * * *
dem65    y6    0.746    12.924    0    * * *
dem65    y7    0.824    18.392    0    * * *
dem65    y8    0.828    18.477    0    * * *
----------------------------------
```

요인적재값은 모두 0.6 이상이며, 통계적으로 유의하다. 따라서 측정모델의 단일차원성은 확보되었다.

측정모델에 대한 모수추정 결과는 다음과 같이 경로도(path diagram)로 나타낼 수 있다. semPlot 패키지의 semPaths() 함수를 이용한다. semPaths() 함수의 what 인수에는 경로도의 링크에 무엇을 표시할지 지정한다. what="std"는 링크상에 표준화된 모수추정치를 출력한다. 링크의 색상은 edge.color 인수에 지정한다. 노드의 색상은 color 인수에 지정할 수 있으며, 잠재변수(lat)와 관측변수(man) 각각을 리스트 형식으로 별도로 지정할 수 있다. 생성된 그래프는 <그림 3>과 같다.

```
> library(semPlot)
> semPaths(fit2, what="std", layout="tree2", edge.label.cex=1, edge.color="darkgreen",
+          color=list(lat="orange", man="palegoldenrod"), fade=FALSE,
+          style="lisrel", rotation=4, curvature=2)
```



<그림 3> 측정모델

4.5 신뢰도와 타당도

측정모델의 신뢰도는 크론바흐 알파계수, 복합신뢰도, 평균분산추출을 이용하여 평가한다. 이 세 가지 지표는 다음과 같이 semTools 패키지의 reliability() 함수를 이용하여 산출할 수 있다.

```
> library(semTools)
> reliability(fit2)
                ind60     dem60     dem65     total
alpha       0.9023348 0.8587945 0.8827394 0.9149416
omega       0.9437375 0.8375192 0.8556193 0.9134989
omega2      0.9437375 0.8375192 0.8556193 0.9134989
omega3      0.9436899 0.8411795 0.8575539 0.9192055
avevar      0.8588015 0.5996704 0.6408647 0.6320778
```

각 잠재변수별로 크론바흐 알파계수(alpha), 세 종류의 복합신뢰도(omega, omega2, omega3), 평균분산추출(avevar)이 계산된다. 이를 다음과 같이 요약 테이블 형식으로 출력한다(여기에서 복합신뢰도는 첫 번째 지표만 출력).

```
> library(dplyr)
> library(tibble)
> library(stargazer)
> reliability(fit2) %>%
+   t() %>%
+   as.data.frame() %>%
+   rownames_to_column("Construct") %>%
+   slice(-n()) %>%
+   select(Construct, "Composite Reliability"=omega,
+          "Average Variance Extracted"=avevar, "Cronbach's alpha"=alpha) %>%
+   stargazer(type="text", title="Convergent Validity: Reliability",
+             summary=FALSE, digits=3, digits.extra=0, rownames=FALSE)

Convergent Validity and Reliability
=============================================================
Construct  Composite Reliability Average Variance Extracted Cronbach's alpha
-------------------------------------------------------------
ind60              0.944                    0.859                  0.902
dem60              0.838                    0.600                  0.859
dem65              0.856                    0.641                  0.883
-------------------------------------------------------------
```

복합신뢰도와 크론바흐 알파계수 모두 권장 수준인 0.7을 상회하고 있으며, 평균분산추출 또한 권장 수준인 0.5를 초과한다. 따라서 측정모델은 신뢰도를 확보하였다.

타당도는 집중타당도와 판별타당도로 구분하여 평가한다. 집중타당도가 충족되기 위해서는 측정모델의 모든 관측변수의 요인적재값이 통계적으로 유의해야 한다. 앞서 살펴본 바와 같이 요인적재값은 모두 0.7 이상으로 충분히 큰 값을 갖고 있으며 통계적으로 유의하다. 또한 AVE는 0.5 이상이다. 따라서 측정모델은 집중타당도를 충족한 것으로 판단한다.

판별타당도는 잠재변수 간 상관계수와 평균분산추출의 제곱근을 비교함으로써 평가할 수 있다. 잠재변수 간 상관계수는 다음과 같이 lavInspect() 함수를 이용하여 계산한다.

```
> lavInspect(fit2, what="cor.lv")
          ind60  dem60  dem65
ind60   1.000
dem60   0.447   1.000
dem65   0.578   0.967   1.000
```

lavInspect() 함수의 what 인수에 추출하고자 하는 정보를 지정한다. 여기에 지정된 what="cor.lv"는 잠재변수 간의 상관계수를 의미한다.[11] 잠재변수 간 상관계수와 평균분산추출의 제곱근을 비교하기 쉽도록 다음과 같이 요약 테이블을 생성한다. 평균분산추출은 앞서 사용한 semTools 패키지의 reliability() 함수를 이용하여 산출한다.

```
> library(semTools)
> library(dplyr)
> library(tibble)
> library(stargazer)
> lavInspect(fit2, what="cor.lv") %>%
+   as.data.frame() %>%
+   rownames_to_column("Construct") %>%
+   cbind(Square_Root_of_AVE=
+             sqrt(reliability(fit2)["avevar", -ncol(reliability(fit2))])) %>%
+   stargazer(type="text", title="Discriminant Validity: Correlation and AVE",
+             summary=FALSE, digits=3, digits.extra=0, rownames=FALSE)

Discriminant Validity
================================================
Construct   ind60    dem60    dem65   Square_Root_of_AVE
------------------------------------------------
ind60         1      0.447    0.578         0.927
dem60       0.447      1      0.967         0.774
dem65       0.578    0.967      1           0.801
------------------------------------------------
```

측정모델이 판별타당도를 충족하기 위해서는 평균분산추출의 제곱근이 잠재변수 간 상관계수보다 커야 한다. 여기에서는 잠재변수 dem60와 dem65 간의 높은 상관계수(0.967)로 인해 이 요건은 달성하지 못하였다. 하지만 이 둘 간의 상관계수를 제외한 나머지 상관계수는 평균분산추출의 제곱근보다 작으므로 판별타당도는 부분적으로 충족하였다.

4.6 구조모델

확인적 요인분석을 통해 단일차원성, 신뢰도, 타당도에 대한 평가를 마친 측정모델을 바탕으로 구조모델을 생성하고 분석한다. 구조모델에 대한 평가는 잠재변수 간의 관계를 분석하는 경로분석을 통해 수행된다. 이를 위해 먼저 다음과 같이 기존의 측정모델에 잠재변수 간 관계를 추가한 구조모델의 구조를 하나의 문자열로 생성한다.

11) 추출할 수 있는 정보의 목록은 lavInspect() 함수의 도움말(?lavInspect)을 참고한다.

```
> sem <- "# measurement model
+              ind60 =~ x1 + x2 + x3
+              dem60 =~ y1 + y2 + y3 + y4
+              dem65 =~ y5 + y6 + y7 + y8
+            # regressions
+              dem60 ~ ind60
+              dem65 ~ ind60 + dem60
+            # residual correlations
+              y1 ~~ y5
+              y2 ~~ y4 + y6
+              y3 ~~ y7
+              y4 ~~ y8
+              y6 ~~ y8"
```

여기에는 모두 세 개의 연산자가 사용되었다. 연산자 =~는 잠재변수와 측정변수 간의 관계를 정의하며, 연산자 ~~는 공분산(상관계수)을 정의한다. 이 두 가지 연산자는 측정모델을 정의할 때 이미 사용한 바 있다. 구조모델에 추가된 연산자 ~는 회귀모델을 정의한다. R에서 사용하는 일반적인 선형회귀모델을 정의하는 방식과 같다(예를 들면, lm() 함수의 회귀모델 정의 방식). ~ 왼쪽에는 종속변수가 오고 ~ 오른쪽에는 독립변수가 위치한다.

종속변수 ~ 독립변수1 + 독립변수2 + 독립변수3 + … + 독립변수n

구조모델에 대한 평가는 sem() 함수를 이용한다. sem() 함수는 확인적 요인분석을 위해 사용한 cfa() 함수와 유사하다. 첫 번째 인수(model)에 앞서 생성한 구조모델 문자열을 지정하고 data 인수에 관측변수가 포함된 데이터셋을 지정한다.

```
> fit <- sem(model=sem, data=PoliticalDemocracy)
```

구조모델의 평가결과는 확인적 요인분석에서 사용한 결과추출 함수(summary() 함수, parameterEstimates() 함수, standardizedsolution() 함수, fitMeasures() 함수 등)를 동일하게 사용할 수 있다. 예를 들어, fitMeasures() 함수를 이용하여 다음과 같이 주요 적합도 지표를 출력할 수 있다.

```
> fitMeasures(fit, c("chisq", "df", "pvalue", "gfi", "rmsea", "cfi"))
 chisq     df   pvalue      gfi  rmsea    cfi
38.125 35.000   0.329    0.923   0.035  0.995
```

구조모델에 대한 분석은 잠재변수 간 경로에 대한 회귀계수의 도출과 회귀계수에 대한 유의성 평가가 주요 관심사항이다. standardizedsolution() 함수의 출력결과를 이용하여 다음과 같이 회귀계수와 그 유의성을 요약 테이블로 정리할 수 있다.

```
> library(dplyr)
> library(stargazer)
> standardizedsolution(fit) %>%
+   filter(op=="~") %>%
+   mutate(stars=ifelse(pvalue < 0.001, "***",
+                       ifelse(pvalue < 0.01, "**",
+                              ifelse(pvalue < 0.05, "*", "")))) %>%
+   select(Dependent=lhs, Independent=rhs, Coefficient=est.std,
+          Z=z, "p-value"=pvalue, Sig.=stars) %>%
+   stargazer(type="text", title="Regression Coefficients", summary=FALSE,
+             digits=3, digits.extra=0, rownames=FALSE)

Regression Coefficients
=============================================
Dependent Independent Coefficient   Z   p-value Sig.
---------------------------------------------
dem60       ind60       0.447     4.323  0.000  ***
dem65       ind60       0.182     2.587  0.010   **
dem65       dem60       0.885     17.405   0     ***
---------------------------------------------
```

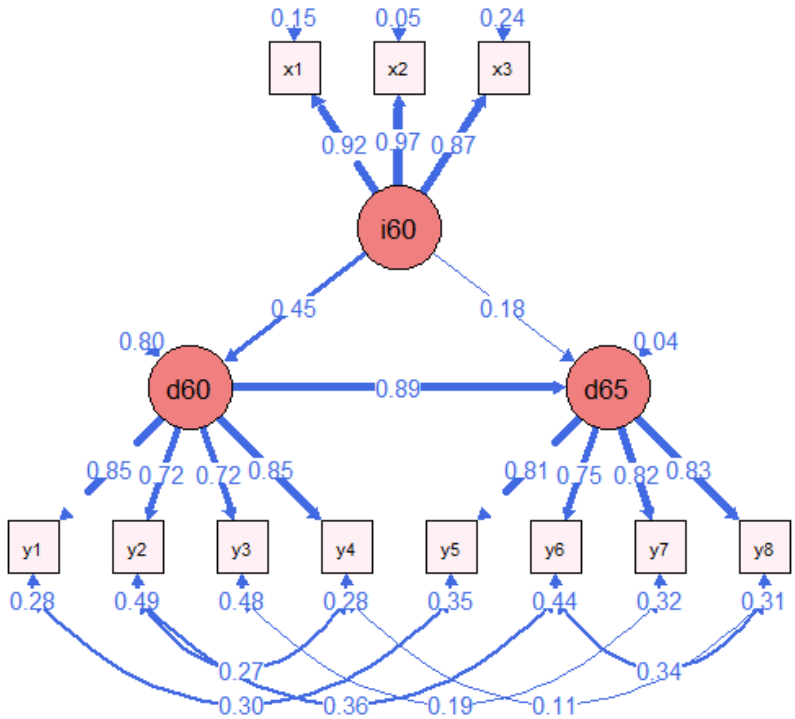
잠재변수 간의 세 경로는 모두 통계적으로 유의하다. 따라서 1960년의 산업화 수준(ind60)은 1960년의 민주화 수준(dem60)에 영향을 미친다는 가설은 채택되었다. 1960년의 산업화 수준과 1960년의 민주화 수준은 1965년의 민주화 수준(dem65)에 영향을 미친다는 가설 또한 채택되었다. 종속변수에 대한 독립변수의 설명력을 나타내는 R^2는 다음과 같이 구할 수 있다.

```
> lavInspect(fit, what="rsquare")
    x1     x2     x3     y1     y2     y3     y4     y5     y6     y7     y8 dem60 dem65
 0.846  0.947  0.761  0.723  0.514  0.522  0.715  0.653  0.557  0.678  0.685  0.200  0.961
```

lavInspect() 함수에 what="rsquare"를 지정한다. 모든 내생변수(내생관측변수와 내생잠재변수)에 대한 R^2가 계산되며, 이 가운데 내생잠재변수(dem60과 dem65)에 대한 R^2가 우리가 관심을 갖는 종속변수에 대한 설명력을 나타내는 R^2이다. 1960년의 민주화 수준(dem60)의 분산은 1960년의 산업화 수준(ind60)에 의해 20.0%가 설명되며, 1965년의 민주화 수준(dem65)의 분산은 두 독립변수 1960년의 산업화 수준과 1960년의 민주화 수준에 의해 96.1%가 설명된다.

구조모델의 분석결과는 semPlot 패키지의 semPaths() 함수를 이용하여 경로도로 나타낼 수 있다. 생성된 경로도는 <그림 4>와 같다.

```
> library(semPlot)
> semPaths(fit, what="std", layout="tree2", edge.label.cex=1, edge.color="royalblue",
+          color=list(lat="lightcoral", man="lavenderblush"), fade=FALSE,
+          style="lisrel", curvature=2)
```



<그림 4> 구조모델

5. 결론

본 튜토리얼은 R을 이용하여 구조방정식모델링을 수행하는 절차와 방법을 실제 데이터를 바탕으로 소개하였다. 구체적으로 R의 lavaan 패키지를 이용하여 확인적 요인분석, 적합도 평가, 모델개선, 신뢰도 및 타당도 평가, 경로계수 추정, 경로도 생성 등의 구조방정식모델을 분석하는 전 과정을 R 프로그램 코드와 함께 제시하였다. 연구자들은 본 튜토리얼에 제시된 절차에 따라 자신의 연구모델을 쉽게 분석할 수 있을 것으로 기대한다. 또한 함께 수록된 프로그램 코드는 분석과정을 이해하고 응용하는 데 있어서 실질적인 도움을 줄 수 있을 것이다.

R은 오픈소스 환경이기 때문에 확장 가능성이 높다. 따라서 기존의 상용 통계패키지에 비해 유연하며 새로운 기능을 신속히 반영할 수 있는 장점을 지닌다. 또한 R은 구조방정식모델링뿐만 아니라 전통적인 통계분석을 위한 다양한 라이브러리를 제공한다. 이러한 특성으로 인해 연구자는 R이라는 하나의 통합환경에서 연구모델을 분석하는 데 필요한 각종 통계처리를 한꺼번에 수행할 수 있는 혜택을 누릴 수 있다.

본 튜토리얼은 대학원의 석사과정 및 박사과정 학생들을 대상으로 개설된 연구방법론 수업이나 단기 워크숍에서 구조방정식모델링 강의를 위한 교재로서 효과적으로 사용할 수 있을 것이다. 구조방정식모델링 기법을 적용한 연구가 경영학을 비롯한 다양한 사회과학 분야에서 주류를 이루고 있고 오픈소스인 R에 대한 관심이 증대하고 있다는 점을 고려할 때 본 튜토리얼은 기존 상용 통계패키지에 대한 대안을 찾고 있는 연구자들에게 유용한 사용지침서가 될 것으로 기대한다.

본 튜토리얼에서는 R을 이용한 기초적인 구조방정식모델링 분석절차 및 방법을 다루었다. 그러나 연구자의 다양한 연구모델을 검정하기 위해서는 기초적인 분석기법을 넘어서는 고급 분석기법이 필요할 수 있다. 매개

효과분석, 조절효과분석, 조절매개효과분석, 다중집단 분석 등은 연구모델의 검정에 종종 등장하는 분석주제로서 R을 이용하여 이들 분석기법의 절차 및 방법을 살펴보는 것은 연구자들의 연구에 실질적인 도움이 될 것으로 생각한다. 후속 튜토리얼에서 이들 주제를 다룰 수 있기를 기대한다.

참고문헌

[국내 문헌]

1. 곽기영 2017. R 기초와 활용, 서울: 도서출판 청람.
2. 곽기영 2019. SPSS를 이용한 통계데이터분석, 서울: 도서출판 청람.
3. 윤지현, 곽기영 2014. "제너러티비티역량: 개념적 정의 및 결정요인," *지식경영연구* (15:3), pp. 95-120.

[국외 문헌]

1. Anderson, J. C., and Gerbing, D. W. 1998. "Structural Equation Modeling in Practice: a Review and Recommended Two-Step Approach," *Psychological Bulletin* (103:3), pp. 411-423.
2. Bagozzi, R. P., and Yi, Y. 1988. "On the Evaluation of Structural Equation Model," *Journal of Academy of Marketing Science* (16:1), pp. 74-94.
3. Bentler, P. M. 1990. "Comparative Fit Indexes in Structural Models," *Psychological Bulletin* (107), pp. 238-246
4. Bentler, P. M. and Bonett, D. G. 1980. "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures," *Psychological Bulletin* (88), pp. 588-606.
5. Bollen, K. A. 1989. *Structural Equations with Latent Variables*, New York: John Wiley and Sons, Inc.
6. Browne, M. W. and Cudeck, R. 1993. "Alternative Ways of Assessing Model Fit," In Bollen, K.A. and Long, J.S. (Eds.), *Testing Structural Equation Models*, CA:Sage, pp.

136-162. Newbury Park
7. Fornell, C., and Larcker, D. F. 1981. "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research* (18), pp. 39-50.
8. Hair Jr., J. F., Black, W. C., Babin, B. J. and Anderson, R. E. 2010. Multivariate Data Analysis: *A Global Perspective* (7th Ed.), Upper Saddle River: Pearson Education.
9. Jöreskog, K. G. and Sörbom, D. 1984. *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods*, Mooresville, Indiana: Scientific Software.
10. Jöreskog, K. G. and Sörbom, D. 1996. *LISREL 8 User's Reference Guide*, Chicago: Scientific Software.
11. lavaan 2019. http://lavaan.ugent.be/tutorial/index.html.
12. Marsh, H. W. and Hocevar, D. 1985. "Application of Confirmatory Factor Analysis to the Study of Self-concept: First- and Higher-order Factor Models and Their Invariance across Groups," *Psychological Bulletin* (97), pp. 562–582.
13. Tanaka, J. S. and Huba, G. J. 1985. "A Fit Index for Covariance Structure Models under Arbitrary GLS Estimation," *British Journal of Mathematical and Statistical Psychology* (38), pp. 197–201.
14. Wheaton, B., Muthen, B., Alwin, D., F., and Summers, G. 1977, "Assessing Reliability and Stability in Panel Models," *Sociological Methodology* (8:1), pp. 84-136.
15. Rosseel, Y. 2012. "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software* (48:2), pp. 1-36.

저 자 소 개

곽 기 영 (Kee-Young Kwahk)

현재 국민대학교 경영대학과 비즈니스IT전문대학원 교수로 재직 중이다. 서울대학교 경영대학을 졸업하고 KAIST 경영과학과와 테크노경영대학원에서 석사 및 박사학위를 취득하였다. 주요 연구관심분야는 Social network analysis and its application, Data analytics, Users' behavior in social media, Knowledge management 등이다.