https://doi.org/10.15813/kmr.2019.20.2.001

R을 이용한 구조방정식모델링: 매개효과분석/조절효과분석 및 다중집단분석[1]

Structural Equation Modeling Using R: Mediation/Moderation Effect Analysis and Multiple-Group Analysis

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

ABSTRACT

This tutorial introduces procedures and methods for performing structural equation modeling using R. To do this, we present advanced analysis methods based on structural equation model such as mediation effect analysis, moderation effect analysis, moderated mediation effect analysis, and multiple-group analysis with R program code using R lavaan package that supports structural equation modeling. R is flexible and scalable, unlike traditional commercial statistical packages. Therefore, new analytical techniques are likely to be implemented ahead of any other statistical package. From this point of view, R will be a very appropriate choice for applying new analytical techniques or advanced techniques that researchers need. Considering that various studies in the social sciences are applying structural equations modeling techniques and increasing interest in open source R, this tutorial is expected to be useful for researchers who are looking for alternatives to existing commercial statistical packages.

Keywords: *Structural equation modeling, R programming, lavaan, Mediation effect analysis, Moderation effect analysis, Multiple-group analysis, LISREL, AMOS*

1. 서론

구조방정식모델링(structural equation modeling)은 최근 경영학을 비롯한 다양한 사회과학 분야에서 변수들 간의 영향관계를 분석하는 사실상의 표준적인 방법으로 자리잡고 있다. 이에 따라 구조방정식모델링을 지원하는 LISREL, AMOS, EQS, Mplus 등과 같은 전통적인 상용 소프트웨어 패키지들이 많은 사회과학 분야 연구에서 널리 사용되어 왔다. 그러나 최근 들어 이들 상용 통계패키지를 사용하는 데 따른 높은 비용과 오픈소스 소프트웨어에 대한 증대된 관심에 힘입어 오픈소스 통계 프로그램에 대한 요구 또한 커지고 있다.

1) 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A3A2075114).
논문접수일: 2019년 2월 6일; 1차 수정: 2019년 3월 30일; 게재확정일: 2019년 4월 11일

2) 제 1저자 (kykwahk@kookmin.ac.kr)

R은 뉴질랜드의 오클랜드대학(University of Auckland)에 근무하던 로스 이하카(Ross Ihaka)와 로버트 젠틀맨(Robert Gentleman)에 의해 1990년대 초에 개발되었다. R은 상업용 통계 프로그래밍 언어인 S를 기반으로 만들어졌다. 현재 R은 데이터분석이나 그래픽과 같은 비통계적 목적으로도 폭넓게 이용되고 있지만 전통적으로 통계 분야에 많은 강점을 가지고 있다(곽기영 2017). 따라서 R은 통계적 분석기법을 기반으로 실증분석을 수행하는 사회과학 분야의 연구에서 지속적으로 중요한 역할을 할 수 있을 것으로 기대한다.

'R을 이용한 구조방정식모델링' 튜토리얼은 두 차례에 걸쳐 소개되고 있다. 본 튜토리얼은 두 번째 편으로서 R을 이용하여 구조방정식모델링을 수행하는 전반적인 분석절차 및 방법은 곽기영(2019a)에서 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. R을 이용한 구조방정식모델링에 익숙하지 않은 독자는 곽기영(2019a)을 참고하는 것이 도움이 될 것이다.

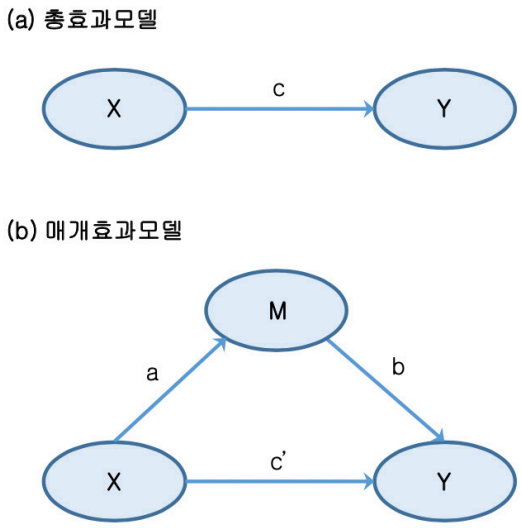
사회과학 분야의 연구모델에서는 흔히 잠재변수 간의 직접적·간접적 영향관계를 동시에 분석하게 되며, 이때 변수 간의 매개효과나 조절효과가 중요한 연구 관심사가 되기도 한다. 또한 변수 간 영향관계에서 집단 간의 차이가 존재하는지도 종종 실증분석을 통해 규명하고자 하는 중요한 연구목표이다. 여기에서는 R 환경에서 개발된 구조방정식모델링 수행 도구인 lavaan 패키지를 이용하여 구조방정식모델에서 이러한 분석을 수행하는 절차 및 방법을 소개한다(lavaan 2019; Rosseel 2012). 구체적으로 본 튜토리얼은 다음과 같은 내용으로 구성된다. 우선 다음 섹션에서는 구조방정식모델에서의 매개효과분석(mediation effect analysis) 및 조절효과분석(moderation effect analysis)과 이 둘을 결합한 조절매개효과분석(moderated mediation effect analysis)을 수행하는 절차를 살펴본다. 이어서 복수의 집단 간의 차이를 비교하는 기법인 다중집단분석(multiple-group analysis)을 수행하는 절차를 소개한다. 끝으로 R을 이용한 구조방정식모델링의 시사점을 토의한다.

2. 매개효과분석과 조절효과분석

2.1 매개효과분석

매개효과분석은 X(독립변수)에서 Y(종속변수)에 이르는 영향관계가 제3의 변수 M(매개변수, mediator)에 의해 매개되는지 검정한다(곽기영 2019b; Baron and Kenny 1986). 매개변수는 독립변수와 종속변수 간의 간접적인 영향관계 메커니즘을 설명하는 역할을 한다. 이런 의미에서 매개효과는 간접효과(indirect effect)라고도 불린다. 매개효과분석을 위한 모델은 <그림 1>과 같다.

<그림 1>의 (a)는 총효과모델로서 회귀계수 c는 Y에 대한 X의 총효과(total effect)를 나타낸다. (b)는 매개효과모델로서 회귀계수 c′은 M을 통제한 상태에서의 Y에 대한 X의 직접효과(direct effect)를 나타낸다. 회귀계수 a와 b의 곱인 ab는 Y에 대한 X의 간접효과(indirect effect)를 나타내며, 매개효과분석은 이러한 간접효과가 존재하는지 여부를 검정한다. 총효과는 직접효과와 간접효과의 합으로 나타낼 수 있다(즉 $c = c' + ab$). 독립변수 X가 매개변수 M에 유의한 영향을 미치고(즉 a가 통계적으로 유의하고), 매개변수 M이 포함된 매개효과모델에서 Y에 대한 X의 직접효과가 매개변수 M이 포함되지 않은 총효과모델에서의 총효과보다 작으면(즉 $c > c'$) 매개효과가 존재한다고 이야기한다. 매개효과가 존재하면 매개변수 M을 통제했을 때 종속변수 Y에 대한 독립변수 X의 영향력(즉 회귀계



<그림 1> 총효과와 매개효과

수)은 감소한다. 매개변수 M이 포함된 매개효과모델에서 Y에 대한 X의 직접효과가 0이면 완전매개(perfect mediation)라고 하며, 0은 아니더라도 직접효과가 총효과에 비해 의미 있는 수준으로 작아지면 부분매개(partial mediation)라고 한다.

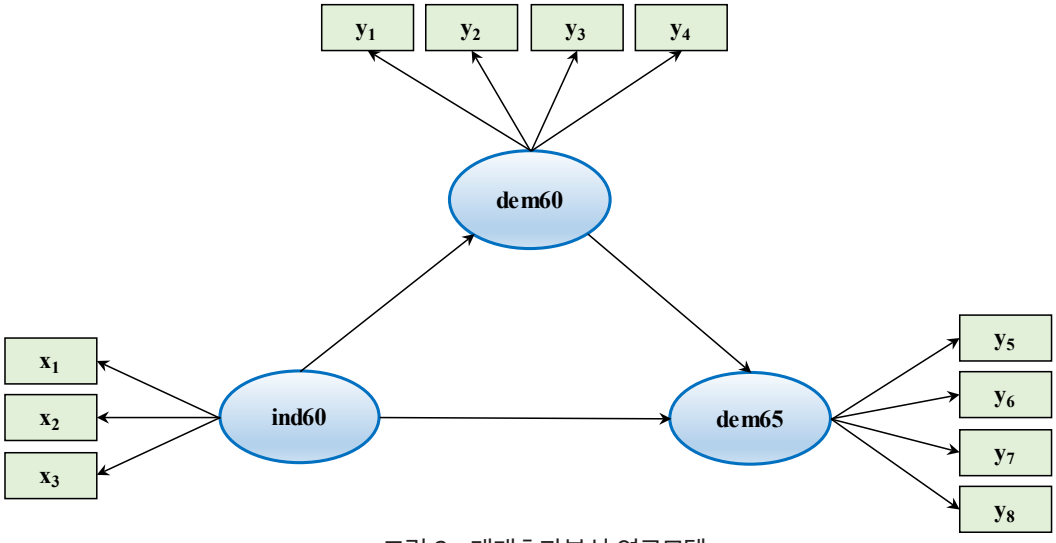
여기에서는 lavaan 패키지에 포함되어 있는 PoliticalDemocracy 데이터셋을 이용하여 구조방정식모델링 분석절차를 살펴본다(lavaan 2019; Michalak 2018a).[3] 먼저 다음과 같이 lavaan 패키지를 메모리에 적재하고 데이터셋(dataset)의 구조를 살펴본다.[4] 이 데이터셋은 한 시점에서의 산업화 수준과 두 시점에서의 민주화 수준을 다양한 관점에서 측정한 변수들로 구성되어 있다.

```
> 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 관계를 나타내는 매개효과분석 연구모델은 <그림 2>와 같다.

<그림 2>의 연구모델은 개발도상국에서의 산업화와 민주화 간의 영향관계를 설명한다. 현재(예를 들면,

3) PoliticalDemocracy 데이터셋과 그 안에 포함된 11개 변수에 대한 설명은 데이터셋에 대한 도움말을 참고한다(?PoliticalDemocracy)
4) 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.



<그림 2> 매개효과분석 연구모델

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.

<그림 2>의 연구모델에서 1960년의 산업화 수준(ind60)이 1960년의 민주화 수준(dem60)을 매개로 1965년의 민주화 수준(dem65)에 영향을 미치는지 분석해보자.[5] 여기에서 ind60는 독립변수이고 dem65는 종속변수이다. dem60는 ind60와 dem65 간의 영향관계를 설명하는 매개변수로의 역할을 수행한다. 매개효과분석을 위한 모델 설정은 다음과 같다.

```
> sem.med <- "# measurement model
+                 ind60 =~ x1 + x2 + x3
+                 dem60 =~ y1 + y2 + y3 + y4
+                 dem65 =~ y5 + y6 + y7 + y8
+               # regressions
+                 dem60 ~ a*ind60
+                 dem65 ~ cp*ind60 + b*dem60
+               # residual correlations
+                 y1 ~~ y5
+                 y2 ~~ y4 + y6
+                 y3 ~~ y7
+                 y4 ~~ y8
+                 y6 ~~ y8
+               # indirect effect: ab
+                 ab := a*b
+               # total effect: c
+                 c := cp + (a*b)"
```

5) 본 튜토리얼에서 다루는 연구모델은 설명을 위해 예시로 제시된 것이다. 따라서 이에 대해 이론적 해석이나 실질적 의미를 부여해서는 안된다.

구조모델(즉 잠재변수 간의 회귀모델)에 포함된 연산자 *는 모수에 대한 레이블(label)을 지정한다.[6] 여기 지정된 a, b, cp는 <그림 1>의 매개효과모델에서 독립변수와 매개변수, 매개변수와 종속변수, 독립변수와 종속변수 간 회귀계수를 나타내는 레이블이다. lavaan 패키지를 이용하여 모수를 추정하면 이들 회귀계수는 지정된 레이블 이름으로 출력된다.

연산자 :=은 새로운 모수를 정의한다. 새로운 모수는 기존 모수의 함수로서 정의되며, 이때 이 함수는 앞서 지정한 모수의 레이블을 이용하여 생성되어야 한다. 여기에서는 ab 모수와 c 모수를 새롭게 정의하였다. ab 모수는 기존의 a 모수와 b 모수의 곱으로 정의되었으며, 이는 매개변수에 의한 간접효과를 나타낸다.[7] c 모수는 직접효과를 나타내는 cp 모수와 앞서 정의한 간접효과(a*b)의 합으로 정의되었으며, 이는 독립변수에서 종속변수에 이르는 총효과를 나타낸다.

sem() 함수를 이용하여 구조모델을 평가하고 매개효과분석을 수행한다. 이때 간접효과의 통계적 유의성 검정을 위해 부트스트래핑을 이용하여 표준오차를 계산한다(se="bootstrap", bootstrap=1000).

```
> library(lavaan)
> set.seed(123)
> fit.med <- sem(model=sem.med, data=PoliticalDemocracy,
+                se="bootstrap", bootstrap=1000)
```

모수추정 결과는 다음과 같이 summary() 함수를 이용하여 출력할 수 있다.

```
> summary(fit.med, standardized=TRUE)
lavaan 0.6-2 ended normally after 68 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

…(중략)

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  dem60 ~
    ind60    (a)     1.483    0.365    4.064    0.000    0.447    0.447
  dem65 ~
    ind60    (cp)    0.572    0.244    2.343    0.019    0.182    0.182
  dem60      (b)     0.837    0.096    8.740    0.000    0.885    0.885
```

6) * 연산자는 모수의 레이블을 지정하는 목적 이외에도 다양한 용도로 사용된다. 모수를 특정 값(예를 들면, 1)으로 고정하거나 초기값을 지정할 때 사용할 수 있다. 또는 모수에 대한 제약조건을 지정할 때도 유용하다. 예를 들어, 다음과 같이 두 관측변수의 요인적재값을 같도록 제약할 수 있다: latent =~ x1 + v*x2 + v*x3. 관측변수 x2, x3에 대해 동일한 한 개의 모수(v)가 추정된다.

7) 함수 정의에 사용된 *는 곱셈을 나타내며, 앞서 살펴본 모수 지정 연산자 *와는 다른 의미를 갖는다.

…(중략)

Defined Parameters:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
ab	1.242	0.364	3.408	0.001	0.395	0.395
c	1.814	0.420	4.320	0.000	0.578	0.578

1960년 산업화 수준(ind60) 한 단위 증가는 1960년 민주화 수준(dem60) 1.483 단위 증가를 가져온다(a=1.483, p-값=0.000). 또한 1960년 산업화 수준을 통제한 상태에서 1960년 민주화 수준(dem60)은 한 단위 증가할 때마다 1965년 민주화 수준(dem65)을 0.837 단위 증가시킨다(b=0.837, p-값=0.000). 즉 1960년 산업화 수준의 증가는 1960년 민주화 수준의 증가를 거쳐 1965년 민주화 수준의 증가로 이어지는 간접적 연관관계를 갖는다. 구체적으로 1960년 산업화 수준과 1960년 민주화 수준 간 경로에서의 1.483 단위의 증가는 1965년 민주화 수준에 있어서 1.242(ab=1.242, p-값=0.001)만큼의 증가를 가져온다. 1960년 민주화 수준을 모델에 포함시키면 1960년 산업화 수준이 1965년 민주화 수준에 미치는 직접효과가 유의수준 0.01에서 통계적으로 유의하지 않으며(cp=0.572, p-값=0.019), 그 크기 또한 총효과에 비해 작아진다(c=1.814 > cp=0.572). 즉 1960년 산업화 수준이 1965년 민주화 수준에 미치는 영향은 1960년 민주화 수준과의 관계를 고려하지 않고 독립적으로 얘기할 수 없다. 따라서 1960년의 민주화 수준은 1960년의 산업화 수준과 1965년의 민주화 수준 간의 관계를 매개한다고 볼 수 있다. 매개효과분석과 관련된 모수만을 추출하여 다음과 같이 요약 테이블로 정리할 수 있다. parameterEstimates() 함수를 이용하여 모델에 포함된 모수추정치를 추출한다.

```
> parameterEstimates(fit.med, standardized=TRUE)
        lhs op        rhs  label    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  0.670   0.920   0.920
 2    ind60 =~         x2          2.180 0.149 14.669  0.000    1.926    2.514  1.460   0.973   0.973
 3    ind60 =~         x3          1.819 0.141 12.903  0.000    1.553    2.083  1.218   0.872   0.872
 …(중략)
 12 dem60  ~     ind60       a 1.483 0.365  4.064  0.000    0.722    2.172  0.447   0.447   0.447
 13 dem65  ~     ind60      cp 0.572 0.244  2.343  0.019    0.095    1.113  0.182   0.182   0.182
 14 dem65  ~     dem60       b 0.837 0.096  8.740  0.000    0.657    1.034  0.885   0.885   0.885
 15       y1 ~~         y5          0.624 0.465  1.341  0.180   -0.282    1.574  0.624   0.296   0.296
 16       y2 ~~         y4          1.313 0.769  1.707  0.088   -0.091    3.058  1.313   0.273   0.273
 17       y2 ~~         y6          2.153 0.877  2.456  0.014    0.470    3.879  2.153   0.356   0.356
 …(중략)
 32   ind60 ~~    ind60          0.448 0.074  6.074  0.000    0.305    0.603  1.000   1.000   1.000
 33 dem60 ~~    dem60          3.956 0.919  4.303  0.000    2.105    5.692  0.800   0.800   0.800
 34 dem65 ~~    dem65          0.172 0.252  0.684  0.494   -0.376    0.648  0.039   0.039   0.039
 35       ab  :=        a*b     ab 1.242 0.364  3.408  0.001    0.578    2.038  0.395   0.395   0.395
 36        c  := cp+(a*b)      c  1.814 0.420  4.320  0.000    0.997    2.672  0.578   0.578   0.578
> library(dplyr)
> library(stargazer)
> parameterEstimates(fit.med, standardized=TRUE) %>%
+   filter(op=="~" | op==":=")  %>%
+   mutate(stars=ifelse(pvalue < 0.001, "***",
+                       ifelse(pvalue < 0.01, "**",
```

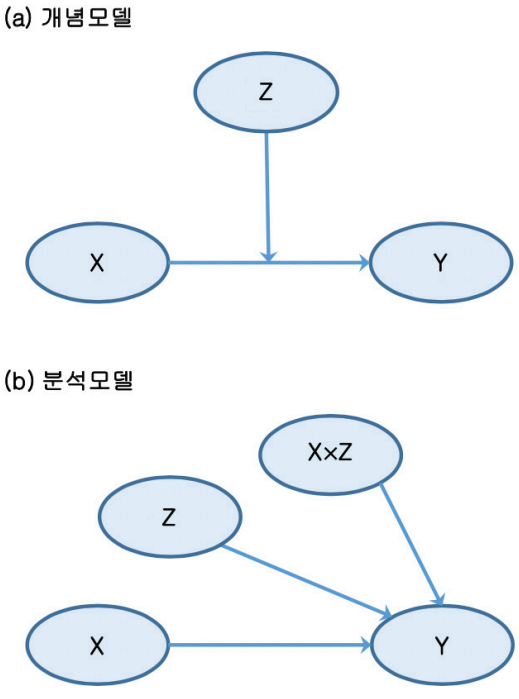
```
+                                  ifelse(pvalue < 0.05, "*", ""))) %>%
+    select(LHS=lhs, RHS=rhs, Label=label, Coefficient=est,
+           Z=z, "p-value"=pvalue, Sig.=stars) %>%
+    stargazer(type="text", title="Regression Coefficients", summary=FALSE,
+              digits=3, digits.extra=0, rownames=FALSE)
```

Regression Coefficients

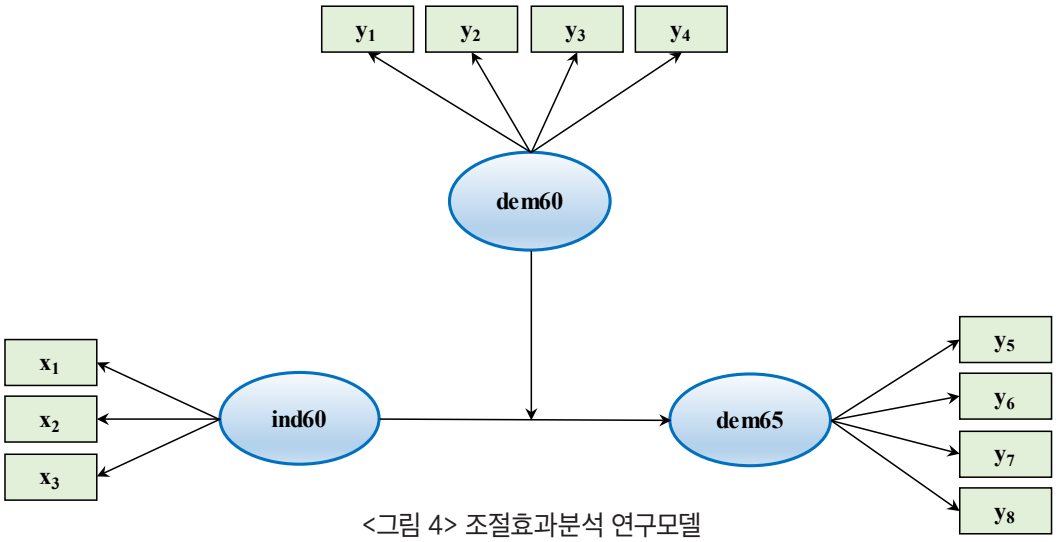
LHS	RHS	Label	Coefficient	Z	p-value	Sig.
dem60	ind60	a	1.483	4.064	0.000	* * *
dem65	ind60	cp	0.572	2.343	0.019	*
dem65	dem60	b	0.837	8.740	0.000	* * *
ab	a* b	ab	1.242	3.408	0.001	* * *
c	cp+(a* b)	c	1.814	4.320	0.000	* * *

2.2 조절효과분석

조절효과분석은 X(독립변수)에서 Y(종속변수)에 이르는 영향관계가 제3의 변수 Z(조절변수, moderator)에 의해 달라지는지 검정한다(곽기영 2019b). 조절효과 모델은 <그림 3>과 같다. 독립변수 X와 조절변수 Z 간의 상호작용(독립변수와 조절변수의 곱으로 표현, 즉 X×Z)이 종속변수 Y에 미치는 영향이 통계적으로 유의한지 검정한다. 조절효과가 존재하면 조절변수의 크기에 따라 독립변수와 종속변수 간의 관계가 강해지거나 약해진다. 또는 영향관계가 반대 방향으로 바뀌기도 한다. 이런 의미에서 조절효과를 상호작용효과(interaction effect)라고도 부른다.



<그림 3> 조절효과



<그림 4> 조절효과분석 연구모델

앞서 살펴봤던 연구모델을 수정하여 1960년의 산업화 수준(ind60)과 1965년의 민주화 수준(dem65) 간의 영향관계가 1960년의 민주화 수준(dem60)에 따라 달라지는지 분석해보자(lavaan 2019; Foldnes 2014; Kearney 2015). 여기에서 ind60는 독립변수이고 dem65는 종속변수이다. dem60는 ind60와 dem65 간의 영향관계를 조절하는 조절변수로의 역할을 수행한다. 조절효과분석 연구모델은 <그림 4>와 같다.

조절효과분석을 위해 먼저 독립변수와 조절변수의 곱으로 표현되는 상호작용항(독립변수x조절변수)을 생성한다. semTools 패키지의 indProd() 함수를 이용하여 독립변수와 조절변수의 모든 가능한 관측변수의 곱을 생성할 수 있다. 여기에서는 독립변수의 관측변수가 3개, 조절변수의 관측변수가 4개 있으므로 모두 12개의 새로운 관측변수가 생성된다(x1.y1 ~ x3.y4). 상호작용항은 모델에 투입되는 기존 변수의 곱으로 만들어지기 때문에 다중공선성의 문제에 취약하다. 따라서 이를 방지하기 위하여 일반적으로 평균중심화(mean centering) 과정을 거친다(meanC=TRUE, doubleMC=TRUE).

```
> library(lavaan)
> library(semTools)
> PoliticalDemocracy.mod <- indProd(PoliticalDemocracy, var1=c("x1", "x2", "x3"),
+                                   var2=c("y1", "y2", "y3", "y4"), match=FALSE,
+                                   meanC=TRUE, residualC=FALSE, doubleMC=TRUE)
> names(PoliticalDemocracy.mod)
 [1] "y1"    "y2"    "y3"    "y4"    "y5"    "y6"    "y7"    "y8"    "x1"    "x2"
[11] "x3"    "x1.y1" "x1.y2" "x1.y3" "x1.y4" "x2.y1" "x2.y2" "x2.y3" "x2.y4" "x3.y1"
[21] "x3.y2" "x3.y3" "x3.y4"
```

조절효과분석을 위한 모델 설정은 다음과 같다.

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

```
+                 # interaction term
+                   ind60dem60 =~ x1.y1 + x1.y2 + x1.y3 + x1.y4 +
+                                 x2.y1 + x2.y2 + x2.y3 + x2.y4 +
+                                 x3.y1 + x3.y2 + x3.y3 + x3.y4
+                 # regressions
+                   dem65 ~ ind60 + dem60 + ind60dem60
+                 # residual correlations
+                   y1 ~~ y5
+                   y2 ~~ y4 + y6
+                   y3 ~~ y7
+                   y4 ~~ y8
+                   y6 ~~ y8"
```

측정모델에 상호작용을 나타내는 새로운 잠재변수(ind60dem60)와 그에 대응되는 관측변수(x1.y1 ~ x3.y4) 간의 관계를 정의한다. 구조모델(즉 잠재변수 간의 회귀모델)에는 독립변수 및 조절변수와 독립변수와 조절변수의 곱으로 생성한 새로운 잠재변수를 포함시킨다.

sem() 함수를 이용하여 조절효과분석을 위한 구조모델을 평가한다. 모수추정 결과는 summary() 함수를 이용하여 출력할 수 있다.

```
> fit.mod <- sem(model=sem.mod, data=PoliticalDemocracy.mod)
> summary(fit.mod, standardized=TRUE)
  lavaan 0.6-3 ended normally after 83 iterations

    Optimization method                           NLMINB
    Number of free parameters                         58

    Number of observations                            75

    Estimator                                         ML
    Model Fit Test Statistic                    1028.376
    Degrees of freedom                               218
    P-value (Chi-square)                           0.000

  …(중략)

  Regressions:
                     Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    dem65 ~
      ind60             0.581    0.220    2.633    0.008    0.184    0.184
      dem60             0.827    0.099    8.341    0.000    0.873    0.873
      ind60dem60        0.066    0.090    0.734    0.463    0.045    0.045

  …(중략)
```

1960년 산업화 수준(ind60)과 1960년 민주화 수준(dem60)은 1965년 민주화 수준(dem65)에 유의한 영향을 미친다(각각 p-값=0.008, 0.000). 그러나 1960년 산업화 수준과 1960년 민주화 수준 간의 상호작용(ind60dem60)은 유의수준 0.05에서 통계적으로 유의하지 않은 것으로 나타났다(p-값=0.463). 따라서 두 예측변수 간에는 상호작

용이 존재하지 않으며, 결과변수와 한 예측변수(독립변수) 간의 관계 패턴은 다른 예측변수(조절변수)의 크기에 따라 달라지지 않는다. 조절효과분석과 관련된 모수만을 추출하여 다음과 같이 요약 테이블로 정리할 수 있다. parameter Estimates() 함수를 이용하여 모델에 포함된 모수추정치를 추출한다.

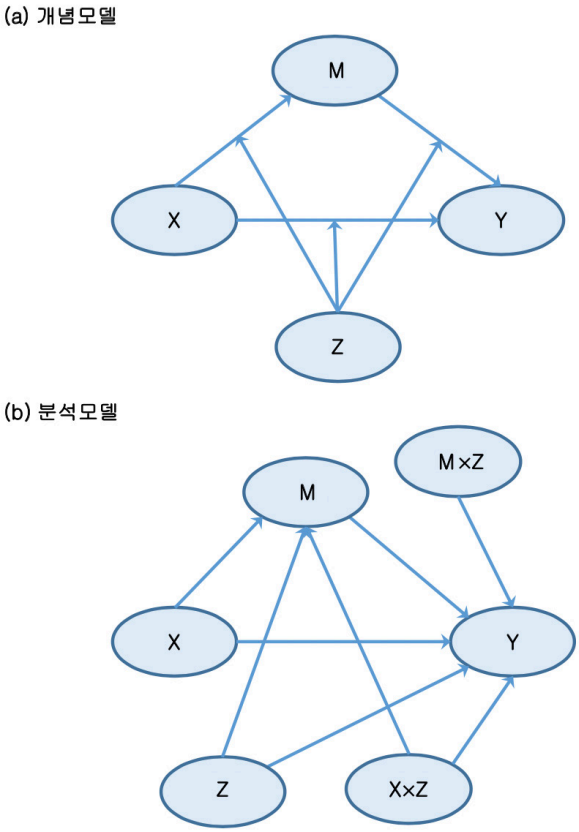
```
> parameterEstimates(fit.mod, standardized=TRUE)
          lhs  op          rhs     est    se      z pvalue ci.lower ci.upper std.lv  std.all std.nox
1       ind60 =~           x1   1.000 0.000     NA     NA    1.000    1.000  0.669   0.918   0.918
2       ind60 =~           x2   2.188 0.139 15.731  0.000    1.915    2.460  1.463   0.975   0.975
3       ind60 =~           x3   1.819 0.153 11.907  0.000    1.520    2.119  1.216   0.871   0.871
…(중략)
24    dem65  ~         ind60   0.581 0.220  2.633  0.008    0.148    1.013  0.184   0.184   0.184
25    dem65  ~         dem60   0.827 0.099  8.341  0.000    0.632    1.021  0.873   0.873   0.873
26    dem65  ~ ind60dem60      0.066 0.090  0.734  0.463   -0.111    0.243  0.045   0.045   0.045
…(중략)
> library(dplyr)
> library(stargazer)
> parameterEstimates(fit.mod, standardized=TRUE) %>%
+   filter(op=="~") %>%
+   mutate(stars=ifelse(pvalue < 0.001, "***",
+                       ifelse(pvalue < 0.01, "**",
+                              ifelse(pvalue < 0.05, "*", "")))) %>%
+  select(LHS=lhs, RHS=rhs, Coefficient=est,
+         Z=z, "p-value"=pvalue, Sig.=stars) %>%
+  stargazer(type="text", title="Regression Coefficients", summary=FALSE,
+            digits=3, digits.extra=0, rownames=FALSE)

Regression Coefficients
=============================================
LHS        RHS       Coefficient    Z    p-value   Sig.
---------------------------------------------
dem65       ind60        0.581     2.633   0.008     **
dem65      dem60         0.827     8.341   0.000    ***
dem65   ind60dem60       0.066     0.734   0.463
---------------------------------------------
```

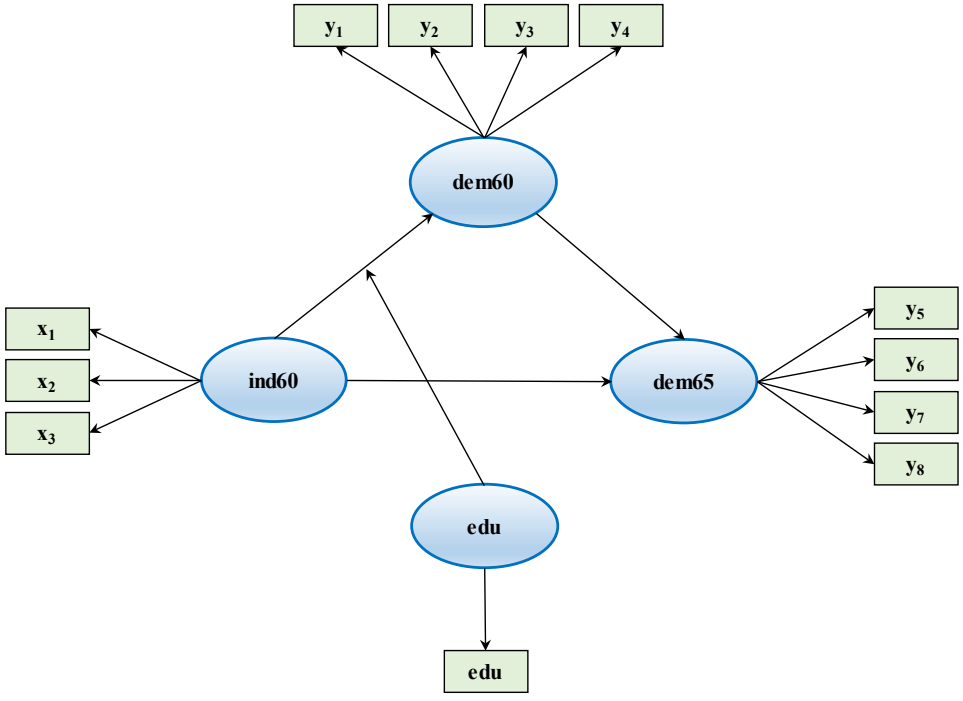
2.3 조절매개효과분석

조절매개효과분석은 매개효과분석과 조절효과분석을 결합한 개념이다. 매개변수에 의해 매개된 두 변수(독립변수와 종속변수) 간 직접적 또는 간접적 영향관계에 제4의 변수(조절변수)가 영향을 미치는지 검정한다. 조절효과가 존재하면 조절변수의 크기에 따라 독립변수와 종속변수 간의 매개관계가 강해지기도 하고 약해지기도 하며, 또는 영향관계가 반대 방향으로 바뀌기도 한다. 조절매개효과모델은 <그림 5>와 같다. 조절효과는 매개효과모델의 모든 경로(즉 X→Y, X→M, M→Y, 또는 이들 세 경로의 조합)에서 발생할 수 있다.

앞서 살펴본 매개효과모델에 조절변수를 도입하여 '1960년 산업화 수준 → 1960년 민주화 수준 → 1965년 민주화 수준'으로 이어지는 간접효과 경로상에서의 조절효과를 분석해보자(lavaan 2019; Michalak 2018b; Washburn 2019). 예를 들어, 교육 수준이 높은 국가와 그렇지 않은 국가는 산업화 수준이 민주화 수준에 미치는 영향에 있어서 차이가 있을 것이라는 가설을 세울 수 있다. 조절매개효과분석을 위한 연구모델은 <그림 6>과 같다.



<그림 5> 조절매개효과



<그림 6> 조절매개효과분석 연구모델

이러한 연구모델을 검정하기 위해 다음과 같이 교육 수준(edu) 변수를 생성하여 조절변수로 활용한다.[8] 교육 수준 변수는 scale() 함수를 이용하여 평균중심화한다.

```
> library(lavaan)
> set.seed(111)
> PoliticalDemocracy$edu <-
+    scale(rnorm(nrow(PoliticalDemocracy), mean=0.5, sd=0.1)*
+              rowMeans(PoliticalDemocracy[, c("x1", "x2", "x3")]),
+            center=TRUE, scale=FALSE)
```

생성한 조절변수와 독립변수(1960년 산업화 수준)를 이용하여 다음과 같이 상호작용항(독립변수x조절변수)을 만든다. 모두 세 개의 새로운 관측변수(x1.edu, x2.edu, x3.edu)가 생성된다.

```
> library(semTools)
> PoliticalDemocracy.modmed <- indProd(PoliticalDemocracy, var1=c("x1", "x2", "x3"),
+                                      var2="edu", match=FALSE, meanC=TRUE,
+                                      residualC=FALSE, doubleMC=TRUE)
> names(PoliticalDemocracy.modmed)
 [1] "y1"     "y2"     "y3"     "y4"     "y5"     "y6"     "y7"     "y8"     "x1"
[10] "x2"     "x3"     "edu"    "x1.edu" "x2.edu" "x3.edu"
```

조절매개효과분석을 위한 모델 설정은 다음과 같다.

```
> sem.modmed <- "# measurement model
+                 ind60 =~ x1 + x2 + x3
+                 dem60 =~ y1 + y2 + y3 + y4
+                 dem65 =~ y5 + y6 + y7 + y8
+               # interaction term
+                 ind60edu =~ x1.edu + x2.edu + x3.edu
+               # regressions
+                 dem60 ~ a1*ind60 + a2*edu + a3*ind60edu
+                 dem65 ~ cp*ind60 + b*dem60
+               # mean and variance of moderator
+                 edu ~ edu.mean*1
+                 edu ~~ edu.var*edu
+               # residual correlations
+                 y1 ~~ y5
+                 y2 ~~ y4 + y6
+                 y3 ~~ y7
+                 y4 ~~ y8
+                 y6 ~~ y8
+               # values of mean and sd of moderator
+                 mean.edu := edu.mean
+                 sd.edu := sqrt(edu.var)
+               # indirect effect conditional on moderator:
```

8) 여기에서 교육 수준은 정규분포로부터 추출한 난수와 산업화 수준을 바탕으로 무작위로 생성하였다. 이는 설명을 위한 목적으로 생성한 임의의 수치이므로 의미를 부여해서는 안된다.

```
+                 # (a1 + a3*ModerationValue)*b
+                   indirect.low := (a1 + a3*(edu.mean-sqrt(edu.var)))*b
+                   indirect.high := (a1 + a3*(edu.mean+sqrt(edu.var)))*b
+                 # direct effect: cp
+                   direct := cp
+                 # total effect: direct effect + indirect effect
+                   total.low := direct + indirect.low
+                   total.high := direct + indirect.high
+                 # index of moderated mediation
+                   mod.med.a3b := a3*b"
```

여기에서는 독립변수(ind60)와 매개변수(dem60) 간 경로에서의 조절효과만 분석한다. 측정모델에 독립변수와 조절변수(edu) 간의 상호작용을 나타내는 새로운 잠재변수(ind60edu)와 그에 대응되는 관측변수(x1.edu, x2.edu, x3.edu) 간의 관계를 정의한다. 구조모델(즉 잠재변수 간의 회귀모델)은 독립변수, 매개변수, 조절변수, 그리고 독립변수와 조절변수의 곱으로 생성한 새로운 잠재변수(상호작용항) 간의 관계를 정의한다.

독립변수와 매개변수 간의 경로에는 세 개의 모수를 추정한다. 구조모델에 레이블로 지정된 a1, a2, a3는 매개경로상의 첫 번째 간접효과 부분을 나타내며 각각 독립변수와 매개변수, 조절변수와 매개변수, 그리고 상호작용항과 매개변수 간의 회귀계수를 의미한다. b는 매개변수와 종속변수 간의 회귀계수를 의미하여, 매개경로의 두 번째 간접효과 부분을 나타낸다. cp는 독립변수와 종속변수 간 회귀계수를 의미하며, 직접효과를 나타낸다.

조절변수가 포함된 모델에서 조절된 매개효과를 파악하기 위해서는, 즉 매개변수에 의한 간접효과가 조절변수의 크기에 따라 어떻게 달라지는지 보기 위해서는 조절변수의 두 가지 수준에 따라 간접효과를 정의할 필요가 있다. 여기에서는 교육 수준의 평균값을 중심으로 한 개의 표준편차만큼 위와 아래에 위치한 값을 이용하여 두 개의 간접효과를 정의한다. 교육 수준이 낮은 경우의 간접효과(indirect.low)는 (a1+a3*(edu.mean-sqrt(edu.var)))*b이며 교육 수준이 높은 경우의 간접효과(indirect.high)는 (a1+a3*(edu.mean+sqrt(edu.var)))*b이다.[9] edu.mean은 교육 수준의 평균을 의미하며 회귀모델 edu ~ edu.mean*1에 의해 계산된다. 회귀모델 y ~ 1은 회귀식의 절편을 추정하며, 이는 곧 y의 평균을 의미한다. edu.var는 교육 수준의 분산을 의미하며, edu ~~ edu.var*edu에 의해 계산된다. 지금처럼 ~~ 연산자의 양쪽에 동일한 변수가 오면 공분산이 아닌 분산을 추정한다. 조절변수의 두 가지 수준에 따른 두 간접효과가 통계적으로 유의한 차이를 보이는지는 a3*b로 정의된 조절매개효과를 검정한다.

sem() 함수를 이용하여 조절매개효과분석을 위한 구조모델에 대한 평가를 수행할 수 있다. 이때 간접효과의 통계적 유의성 검정을 위해 부트스트래핑을 이용하여 표준오차를 계산한다(se="bootstrap", bootstrap=500). 모수 추정 결과는 summary() 함수를 이용하여 출력할 수 있다.

```
> set.seed(111)
> fit.modmed <- sem(model=sem.modmed, data=PoliticalDemocracy.modmed,
+                   se="bootstrap", bootstrap=500)
> summary(fit.modmed, standardized=TRUE)
```

9) 조절변수가 범주형 변수일 경우에는 edu.mean-sqrt(edu.var)과 edu.mean+sqrt(edu.var) 대신에 비교하고자 하는 각 범주의 값을 입력한다.

lavaan 0.6-3 ended normally after 85 iterations

Optimization method	NLMINB
Number of free parameters	56
Number of observations	75
Estimator	ML
Model Fit Test Statistic	183.787
Degrees of freedom	79
P-value (Chi-square)	0.000

…(중략)

Regressions:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
dem60 ~							
ind60	(a1)	1.688	0.538	3.135	0.002	0.493	0.493
edu	(a2)	-0.228	0.428	-0.533	0.594	-0.100	-0.075
ind60edu	(a3)	0.202	0.462	0.437	0.662	0.050	0.050
dem65 ~							
ind60	(cp)	0.574	0.255	2.251	0.024	0.177	0.177
dem60	(b)	0.838	0.095	8.843	0.000	0.883	0.883

…(중략)

Defined Parameters:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
mean.edu	0.000	0.089	0.000	1.000	0.000	0.000
sd.edu	0.751	0.057	13.159	0.000	0.751	1.000
indirect.low	1.287	0.472	2.724	0.006	0.402	0.391
indirect.high	1.541	0.659	2.337	0.019	0.469	0.480
direct	0.574	0.255	2.249	0.024	0.177	0.177
total.low	1.861	0.500	3.720	0.000	0.579	0.568
total.high	2.115	0.719	2.942	0.003	0.645	0.657
mod.med.a3b	0.169	0.389	0.435	0.664	0.044	0.044

교육 수준은 민주화 수준에 영향을 미치지 않는다(a2=-0.228, p-값=0.594). 1960년 산업화 수준(ind60) 한 단위 증가는 1960년 민주화 수준(dem60) 1.688 단위 증가를 가져온다(a1=1.688, p-값=0.002). 그리고 이러한 관계는 교육 수준에 따라서 달라지지 않는다(a3=0.202, p-값=0.662). 또한 1960년 산업화 수준을 통제한 상태에서 1960년 민주화 수준(dem60)은 한 단위 증가할 때마다 1965년 민주화 수준(dem65)을 0.838 단위 증가시킨다(b=0.838, p-값=0.000). 즉 1960년 산업화 수준의 증가는 1960년 민주화 수준의 증가를 거쳐 1965년 민주화 수준의 증가로 이어지는 간접적 연관관계를 갖는다. 그러나 이러한 간접적 영향 관계는 교육 수준에 따라서 다르지 않다(a3*b=0.169, p-값=0.664).

교육 수준이 낮은 국가에 대해 1960년 산업화 수준과 1960년 민주화 수준 간 경로에서의 1.688 단위의 증가는 1965년 민주화 수준에 있어서 1.287(=(a1+a3*(edu.mean-sqrt(edu.var)))

b=(1.688+0.202(0-0.751))*0.838, p-값=0.006)만큼의 증가를 가져온다. 교육 수준이 높은 국가에 대해 1960년 산업화 수준과 1960년 민주화 수준 간 경로에서의 1.688 단위의 증가는 1965년 민주화 수준에 있어서 1.541(=(a1+a3*(edu.mean+sqrt(edu.var)))*b=(1.688+0.202*(0+0.751))*0.838, p-값=0.019)만큼의 증가를 가져온다. 교육 수준이 높은 국가는 그렇지 않은 국가에 비해 1960년 민주화 수준의 매개 역할이 다소 더 강해보이지만 이러한 차이는 앞서 살펴본 대로 통계적으로 유의하지 않다. 즉 교육 수준의 조절매개효과는 존재하지 않는다. 직접효과의 관점에서 보면 1960년 산업화 수준이 1965년 민주화 수준에 미치는 영향은 1960년 민주화 수준과의 관계를 고려할 경우 작아진다(total.low=1.861 > cp=0.574, total.high=2.115 > cp=0.574). 따라서 1960년 민주화 수준은 조절변수를 고려해도 여전히 매개변수로서의 역할을 수행한다. 조절매개효과분석과 관련된 모수만을 추출하여 다음과 같이 요약 테이블로 정리할 수 있다.

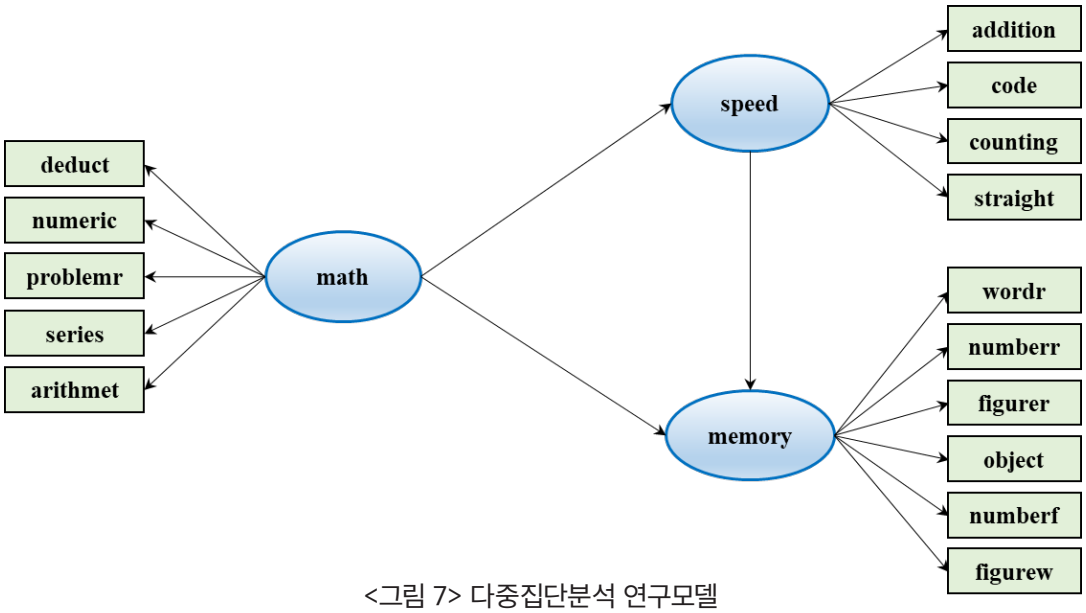
```
> library(dplyr)
> library(stargazer)
> parameterEstimates(fit.modmed, standardized=TRUE) %>%
+   filter(op=="~" | op==":=") %>%
+   mutate(stars=ifelse(pvalue < 0.001, "***",
+                       ifelse(pvalue < 0.01, "**",
+                              ifelse(pvalue < 0.05, "*", "")))) %>%
+   select(LHS=lhs, RHS=rhs, Label=label, Coefficient=est,
+          Z=z, "p-value"=pvalue, Sig.=stars) %>%
+   stargazer(type="text", title="Regression Coefficients", summary=FALSE,
+             digits=3, digits.extra=0, rownames=FALSE)
```

Regression Coefficient

LHS	RHS	Label	Coefficient	Z	p-value	Sig.
dem60	ind60	a1	1.688	3.135	0.002	* *
dem60	edu	a2	−0.228	−0.533	0.594	
dem60	ind60edu	a3	0.202	0.437	0.662	
dem65	ind60	cp	0.574	2.251	0.024	*
dem65	dem60	b	0.838	8.843	0	* * *
mean.edu	edu.mean	mean.edu	0	0	1	
sd.edu	sqrt(edu.var)	sd.edu	0.751	13.159	0	* * *
indirect.low	(a1+a3* (edu.mean−sqrt(edu.var)))* b	indirect.low	1.287	2.724	0.006	* *
indirect.high	(a1+a3* (edu.mean+sqrt(edu.var)))* b	indirect.high	1.541	2.337	0.01 9	*
direct	cp	direct	0.574	2.249	0.024	*
total.low	direct+indirect.low	total.low	1.861	3.720	0.000	* * *
total.high	direct+indirect.high	total.high	2.115	2.942	0.003	* *
mod.med.a3b	a3* b	mod.med.a3b	0.169	0.435	0.664	

3. 다중집단분석

다중집단분석은 집단 간에 모수추정치(예를 들면, 요인적재값, 회귀계수)에 있어서 유의한 차이가 있는지 검정한다. 여기에서는 MBESS 패키지에 포함되어 있는 HS.data 데이터셋을 이용하여 집단 간 차이분석을 수행하는 절차를 살펴본다(lavaan 2019; Bustos 2017). 이 데이터셋은 학생의 공간능력, 어휘능력, 사고속도, 기



<그림 7> 다중집단분석 연구모델

억력, 수리능력 등을 다양한 관점에서 측정한 변수들로 구성되어 있다. 여기에서는 다중집단분석을 통해 성별에 따른 집단 간 차이를 규명한다. 데이터셋에 대한 보다 자세한 내용은 도움말을 참고한다(?HS.data).

```
> library(MBESS)
> data("HS.data")
> names(HS.data)
 [1] "id"       "Gender"   "grade"    "agey"     "agem"     "school"   "visual"
 [8] "cubes"    "paper"    "flags"    "general"  "paragrap" "sentence" "wordc"
[15] "wordm"    "addition" "code"     "counting" "straight" "wordr"    "numberr"
[22] "figurer"  "object"   "numberf"  "figurew"  "deduct"   "numeric"  "problemr"
[29] "series"   "arithmet" "paperrev" "flagssub"
```

분석하고자 하는 연구모델은 <그림 7>과 같다. 수리능력(math)과 사고속도(speed)가 기억력(memory)에 미치는 영향관계를 검정한다. 다중집단분석을 수행하기 전에 먼저 전체 데이터를 대상으로 이들 변수 간의 관계를 검정해보자. 이를 위해 lavaan 패키지의 모델설정 방식에 따라 측정모델과 구조모델의 구조를 다음과 같이 문자열로 생성한다.

```
> sem <- "# measurement model
+             speed =~ addition + code + counting + straight
+             memory =~ wordr + numberr + figurer + object + numberf + figurew
+             math =~  deduct + numeric + problemr + series + arithmet
+           # regressions
+             speed ~ math
+             memory ~ math + speed"
```

sem() 함수를 이용한 전체 모델의 검정결과는 다음과 같다.

```
> library(lavaan)
> fit <- sem(model=sem, data=HS.data)
> summary(fit, fit.measures=TRUE, standardized=TRUE)
  lavaan 0.6-3 ended normally after 165 iterations

    Optimization method                           NLMINB
    Number of free parameters                         33

    Number of observations                           301

    Estimator                                         ML
    Model Fit Test Statistic                     248.862
    Degrees of freedom                                87
    P-value (Chi-square)                           0.000
```

…(중략)

```
 Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   speed ~
     math             0.875    0.140    6.262    0.000    0.619    0.619
   memory ~
     math             0.323    0.071    4.568    0.000    0.529    0.529
     speed            0.087    0.043    2.020    0.043    0.202    0.202
```

…(중략)

회귀분석 부분(즉 구조모델)만을 추출하여 다음과 같이 요약표를 작성할 수 있다.

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

 Regression Coefficient
 ================================================
 Dependent  Independent  Coefficient   Z     p-value  Sig.
 ------------------------------------------------
 speed         math          0.875    6.262   0.000   * * *
 memory        math          0.323    4.568   0.000   * * *
 memory        speed         0.087    2.020   0.043     *
 ------------------------------------------------
```

구조모델의 분석결과를 보면 수리능력은 사고속도에 유의한 영향을 미치는 것으로 나타났으며(p-값=0.000), 또한 수리능력과 사고속도는 기억력에 유의한 영향을 미친다(각각 p-값=0.000, p-값=0.043). 따라서 제시된 연구모델의 가설은 모두 채택되었다(유의수준 0.05).

집단 간 차이분석을 위해서는 sem() 함수의 group 인수에 추가로 집단변수를 지정한다. 동일한 모델에 대해 각 집단별로 검정이 수행된다. 예를 들어, group="Gender"를 지정한 다음 코드는 수리능력, 사고속도, 기억력 간 관계를 정의한 앞서의 구조모델을 남학생 집단과 여학생 집단 각각에 대해 개별적으로 검정한다.

```
> fit1 <- sem(model=sem, data=HS.data, group="Gender")
> summary(fit1, fit.measures=TRUE, standardized=TRUE)
  lavaan 0.6-3 ended normally after 323 iterations

    Optimization method                           NLMINB
    Number of free parameters                         96

    Number of observations per group
    Male                                             146
    Female                                           155

    Estimator                                         ML
    Model Fit Test Statistic                     322.272
    Degrees of freedom                               174
    P-value (Chi-square)                           0.000

…(중략)

 Group 1 [Male]:

…(중략)

 Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   speed ~
     math             0.694    0.145    4.788    0.000    0.635    0.635
   memory ~
     math             0.471    0.111    4.248    0.000    0.780    0.780
     speed           -0.071    0.080   -0.884    0.376   -0.129   -0.129

…(중략)

 Group 2 [Female]:

…(중략)

 Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   speed ~
     math             1.093    0.257    4.252    0.000    0.617    0.617
   memory ~
```

```
    math          0.237    0.088    2.684    0.007    0.415    0.415
    speed         0.123    0.049    2.507    0.012    0.380    0.380

…(중략)
```

집단별 검정결과는 'Group 1 [Male]'과 'Group 2 [Female]'에서 확인할 수 있다. 회귀분석 부분만을 추출하면 다음과 같다.

```
> parameterEstimates(fit1) %>%
+   filter(op=="~") %>%
+   mutate(stars=ifelse(pvalue < 0.001, "***",
+                       ifelse(pvalue < 0.01, "**",
+                              ifelse(pvalue < 0.05, "*", "")))) %>%
+   select(Group=group, Dependent=lhs, Independent=rhs, Coefficient=est,
+          Z=z, "p-value"=pvalue, Sig.=stars) %>%
+   stargazer(type="text", title="Regression Coefficients", summary=FALSE,
+             digits=3, digits.extra=0, rownames=FALSE)

Regression Coefficient
=============================================================
Group   Dependent   Independent   Coefficient     Z     p-value   Sig.
-------------------------------------------------------------
1         speed        math          0.694      4.788    0.000    * * *
1        memory        math          0.471      4.248    0.000    * * *
1        memory        speed        -0.071     -0.884    0.376
2         speed        math          1.093      4.252    0.000    * * *
2        memory        math          0.237      2.684    0.007     * *
2        memory        speed         0.123      2.507    0.012      *
-------------------------------------------------------------
```

구조모델의 분석결과를 보면 두 집단 간에 회귀계수의 크기와 유의성에 있어서 차이가 존재한다. 특히 사고속도가 기억력에 미치는 영향이 남학생 집단(Group 1)에서는 통계적으로 유의하지 않은 반면(p-값=0.376), 여학생 집단(Group 2)에서는 유의수준 0.05에서 통계적으로 유의하다(p-값=0.012). 따라서 사고속도가 기억력에 영향을 미친다는 가설은 남학생 집단에서는 기각되었다.

집단 간의 차이를 확인한 다음에는 이러한 집단 간 차이가 통계적으로 유의한지 검정할 필요가 있다. 집단 간 차이의 유의성 검정은 집단 간에 경로계수가 동일하도록 인위적으로 제약한 모델과 각 집단별로 경로계수를 자유롭게 추정한 모델 간의 적합도를 비교함으로써 수행할 수 있다. 경로계수를 자유롭게 추정한 모델이 제약모델에 비해 유의하게 적합도가 개선되면 집단 간 경로계수의 차이가 통계적으로 유의하다고 결론 내린다.

제약모델을 이용한 검정은 특정 모수가 집단 간에 동일하도록 모델의 모수를 제약한 후 이렇게 생성한 제약모델의 적합도를 평가하는 방식으로 수행한다. 예를 들면, 사고속도와 기억력 간 경로계수가 집단 간에 동일한 값을 갖도록 모델설정 시 해당 경로를 제약하고 모델의 적합도를 평가 및 비교한다. 제약모델의 모수는 다음과 같이 설정할 수 있다.

```
> sem.const <- "# measurement model
+                 speed =~ addition + code + counting + straight
+                 memory =~ wordr + numberr + figurer + object + numberf + figurew
+                 math =~  deduct + numeric + problemr + series + arithmet
+               # regressions
+                 speed ~ math
+                 memory ~ math + c(b, b)*speed"
```

사고속도와 기억력 간 경로계수가 동일하도록 제약한 부분은 집단별 경로계수에 동일한 레이블을 지정한 'memory ~ math + c(b, b)*speed'이다. 여기에서 'c(b, b)*speed'는 사고속도의 경로계수가 두 집단 간에 동일하도록(즉 동일하게 b로) 추정하라는 의미이다. 집단별로 경로계수만을 추출하면 다음과 같다. 사고속도와 기억력 간 경로계수가 남학생 집단과 여학생 집단 모두 동일하게 0.078로 추정된 것을 확인할 수 있다.

```
> fit2 <- sem(model=sem.const, data=HS.data, group="Gender")
> parameterEstimates(fit2) %>%
+   filter(op=="~") %>%
+   mutate(stars=ifelse(pvalue < 0.001, "***",
+                       ifelse(pvalue < 0.01, "**",
+                              ifelse(pvalue < 0.05, "*", "")))) %>%
+   select(Group=group, Dependent=lhs, Independent=rhs, Coefficient=est,
+          Z=z, "p-value"=pvalue, Sig.=stars) %>%
+   stargazer(type="text", title="Regression Coefficients", summary=FALSE,
+             digits=3, digits.extra=0, rownames=FALSE)

Regression Coefficient
=========================================================
Group   Dependent   Independent   Coefficient    Z     p-value  Sig.
---------------------------------------------------------
1         speed         math         0.640     4.684   0.000   ***
1        memory         math         0.350     4.252   0.000   ***
1        memory         speed        0.078     2.067   0.039    *
2         speed         math         1.132     4.312   0.000   ***
2        memory         math         0.279     3.042   0.002    **
2        memory         speed        0.078     2.067   0.039    *
---------------------------------------------------------
```

이제 anova() 함수를 이용하여 다음과 같이 기존의 집단별로 자유롭게 추정한 모델(fit1)과 제약모델(fit2) 간 적합도 비교를 수행한다.

```
> anova(fit1, fit2)
Chi Square Difference Test

     Df   AIC   BIC  Chisq  Chisq diff  Df diff  Pr(>Chisq)
fit1 174 32304 32659 322.27
fit2 175 32306 32658 326.54      4.2703        1    0.03879 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

anova() 함수는 두 모델에 대한 카이제곱 차이검정을 수행한다. 검정결과에 따르면 두 모델 간의 카이제곱 변화량은 통계적으로 유의하다($\chi^2(1)=4.2703$, p-값<0.05). 따라서 집단 간 경로계수를 자유롭게 추정한 모델이 동일하도록 제약한 모델에 비해 데이터를 보다 잘 적합한다고 볼 수 있다. 간단한 함수를 생성하여 두 모델의 주요 적합도 지표를 비교해 보면 다음과 같다. 제시된 적합도 지표 모두 집단 간 경로계수의 차이를 허용한 모델이 더 우수하다는 것을 보여준다.

```
> 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"))}
> library(stargazer)
> compareFit(fit1, fit2) %>%
+    stargazer(type="text", title="Model Comparison", summary=FALSE,
+              digits=3, digits.extra=0, rownames=FALSE)
```

```
Model Comparison
============================
Fit_Measures    Model1   Model2
----------------------------
Chi-square      322.272  326.542
df                174      175
p-value           0.000    0.000
GFI               0.998    0.998
RMSEA             0.075    0.076
CFI               0.877    0.874
----------------------------
```

동일한 레이블을 지정함으로써 일부 모수를 직접 제약하는 방식은 매우 유연한 방법이긴 하지만 모든 모수를 대상으로 제약하고자 할 때는 좀 더 편리한 방법이 있다. 예를 들어, 특정 경로계수가 아닌 모든 경로계수를 대상으로 집단 간에 동일하도록 제약하기 위해서는 group.equal 인수를 이용한다. group.equal 인수에 동일성 제약을 가하고자 하는 모수의 키워드를 지정한다. 경로계수에 대한 집단 간 동일성 제약은 다음과 같이 group.equal=c("regressions")를 지정한다.[10]

```
> fit3 <- sem(model=sem, data=HS.data,
+             group="Gender", group.equal=c("regressions"))
```

10) 집단 간 동일성 제약을 위한 키워드는 이 밖에도 많이 있다. 예를 들면, 요인적재값을 동일하게 하기 위해서는 group.equal=c("loadings")를 지정한다. 좀 더 자세한 내용은 lavaan 패키지 튜토리얼을 참고한다: http://lavaan.ugent.be/tutorial/index.html.

anova() 함수를 이용하여 기존 모델과의 카이제곱 차이검정을 수행하면 다음과 같다.

```
> anova(fit1, fit3)
 Chi Square Difference Test

      Df     AIC     BIC    Chisq  Chisq diff Df diff Pr(>Chisq)
fit1 174   32304   32659   322.27
fit3 177   32306   32651   330.72       8.4518       3    0.03754 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

검정결과에 따르면 두 모델 간의 카이제곱 변화량은 통계적으로 유의하다($\chi^2(3)=8.4518$, p-값<0.05). 따라서 집단 간 경로계수를 자유롭게 추정한 모델이 모든 경로계수를 동일하도록 제약한 모델에 비해 데이터를 보다 잘 적합한다고 볼 수 있다.

세 모델을 다음과 같이 한꺼번에 비교해보면 집단별로 경로계수를 자유롭게 추정한 첫 번째 모델의 적합도가 가장 우수하다. 따라서 남학생 집단과 여학생 집단 간에는 수리능력, 사고속도, 기억력 간 영향관계의 차이가 존재한다.

```
> compareFit(fit1, fit2, fit3) %>%
+   stargazer(type="text", title="Model Comparison", summary=FALSE,
+             digits=3, digits.extra=0, rownames=FALSE)

Model Comparison
=====================================
Fit_Measures    Model1    Model2    Model3
-------------------------------------
Chi-square      322.272   326.542   330.724
df                  174       175       177
p-value           0.000     0.000     0.000
GFI               0.998     0.998     0.998
RMSEA             0.075     0.076     0.076
CFI               0.877     0.874     0.873
-------------------------------------
```

4. 결론

본 튜토리얼은 R을 이용하여 구조방정식모델링을 수행하는 절차와 방법을 실제 데이터를 바탕으로 소개하였다. 구체적으로 R의 lavaan 패키지를 이용하여 매개효과분석, 조절효과분석, 조절매개효과분석, 다중집단분석 등의 구조방정식모델 기반의 고급 분석기법을 R 프로그램 코드와 함께 제시하였다. 연구자들은 본 튜토리얼에 제시된 절차에 따라 필요로 하는 분석기법을 비교적 용이하게 활용할 수 있을 것으로 기대한다. 또한 함께 수록된 R 프로그램 코드는 분석절차를 이해하고 분석방법을 응용하는 데 있어서 실질적인 도움을 제공할 것으로 기대한다.

R은 기존의 상용 통계패키지와 달리 유연하고 확장성이 높다는 특징을 갖고 있다. 따라서 오픈소스의 특성상 새로운 분석기법은 다른 어떤 상용 통계패키지보다도 앞서서 구현될 가능성이 높다. 이러한 관점에서 볼 때 연구자가 필요로 하는 새로운 분석기법이나 고급기법을 적용하는 데 있어서 R은 매우 적절한 선택이 될

것이다. 또한 R은 구조방정식모델링뿐만 아니라 전통적인 통계분석을 위한 다양한 라이브러리를 제공한다. 이러한 특성으로 인해 연구자는 R이라는 하나의 통합환경에서 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. 사회과학 분야에서의 다양한 연구들이 구조방정식모델링 기법을 적용하고 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.

참 고 문 헌

[국내 문헌]

1. 곽기영 2019a. "R을 이용한 구조방정식모델링: 분석절차 및 방법," *지식경영연구* (20:1), pp. 1-26.
2. 곽기영 2019b. SPSS를 이용한 통계데이터분석, 서울: 도서출판 청람.
3. 곽기영 2017. R 기초와 활용, 서울: 도서출판 청람.

[국외 문헌]

1. Baron, R. M., and Kenny, D. A. 1986. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations," *Journal of Personality and Social Psychology* (51:6), pp. 1173-1182.
2. Bustos, C. 2017. http://rpubs.com/clbustos/sem_multigroup_regression.
3. Foldnes, N. 2014. https://rpubs.com/njaalf/workshop2014.
4. Kearney, M. W. 2015. https://rpubs.com/mkearney/103098.
5. lavaan 2019. http://lavaan.ugent.be/tutorial/index.html.
6. Michalak, N. M. 2018a. http://nickmichalak.com/blog_entries/2018/nrg01/nrg01.html.
7. Michalak, N. M. 2018b. http://nickmichalak.com/blog_entries/2018/nrg02/nrg02.html.
8. Rosseel, Y. 2012. "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software* (48:2), pp. 1-36.
9. Washburn, A. N. https://ademos.people.uic.edu/Chapter15.html.

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.