조 영 일 교수 지도

석사학위 청구논문

표준화 잔차(Standardized Residuals)를 이용한 확인적 요인분석에서의 극단치 식별

2015

성신여자대학교 대학원

심 리 학 과

신 진 이

표준화 잔차(Standardized Residuals)를 이용한 확인적 요인분석에서의 극단치 식별

조 영 일 교수 지도

이 논문을 석사학위논문으로 제출함

2014년 11월

성신여자대학교 대학원

심 리 학 과

신 진 이

인 준 서

신진이의 석사학위 논문으로 인준함.

2014년 11월

심사위원장 _____ (인)

심 사 위 원 _____ (인)

심 사 위 원 _____ (인)

성신여자대학교 대학원

논문개요

본 연구에서는 구조방정식 모형에서 측정모형에 해당하는 확인적 요인모형에서의 극단치 식별에 관한 경험적 증거를 찾고자 했다. 먼저 요인모형에서 표준화 잔차의 이론적 분포가 χ^2분포와 다른지 검증하였고, 표준화 잔차의 경험적 분포의 평균 및 표준편차와 이론적 분포의 평균 및 표준편차의 비교를 통해서 표준화 잔차가 χ^2분포를 가정하는지 살펴보았다. 연구 결과, 생성된 자료로부터 산출된 χ^2값은 χ^2분포를 따르지 않았다. 따라서 극단치 식별을 위해 χ^2분포가 아닌 생성된 자료의 경험적 분포를 활용하여 백분위수(percentile) 99%, 95%, 90%에 따른 임계값(critical value)를 구하였다. 이 임계값을 이용하여 극단치를 포함한 표본에서 분석모형, 표본크기, 극단치 비율의 다양한 조건 하에서 표준화 잔차의 극단치 탐지율을 비교함으로써 표준화 잔차의 유용성을 검증하였다. 주요 연구결과를 요약하면 다음과 같다.

첫째, 분석모형이 단순하고 표본크기가 크며 극단치 비율이 높은 조건에서 극단치 탐지율이 높게 나타났다. 둘째, 모형과 표본크기의 상호작용효과를 분석한 결과 통계적으로 유의한 것으로 나타났다. 따라서 단순모형과 복잡모형으로 나누어 단순주효과 분석을 실시한 결과 표본크기가 300일 때 극단치 탐지율이 가장 높게 나타났으며, 다음으로 표본크기 50, 100 순으로 극단치 탐지율이 높게 나타났다. 셋째, 표본크기와 극단치 비율의 상호작용효과 역시 통계적으로 유의하게 나타났다. 보다 구체적으로 표본크기가 50과 100인 조건에서는 극단치 비율이 높을수록 극단치 탐지율이 높은 것으로 나타났으며 표본크기가 300인 조건에서는 5%일 때 극단치 탐지율이 가장 높았으며, 다음으로 20%, 15%, 10%순으로 극단치 탐지율이 높은 것으로 나타

났다.

끝으로, 본 연구에 대해 논의하고 제한점 및 추후 연구방향을 제시하였다.

주요어: 극단치, 이상치, 시뮬레이션, 표준화 잔차, outlier

목 차

그 림 목 차

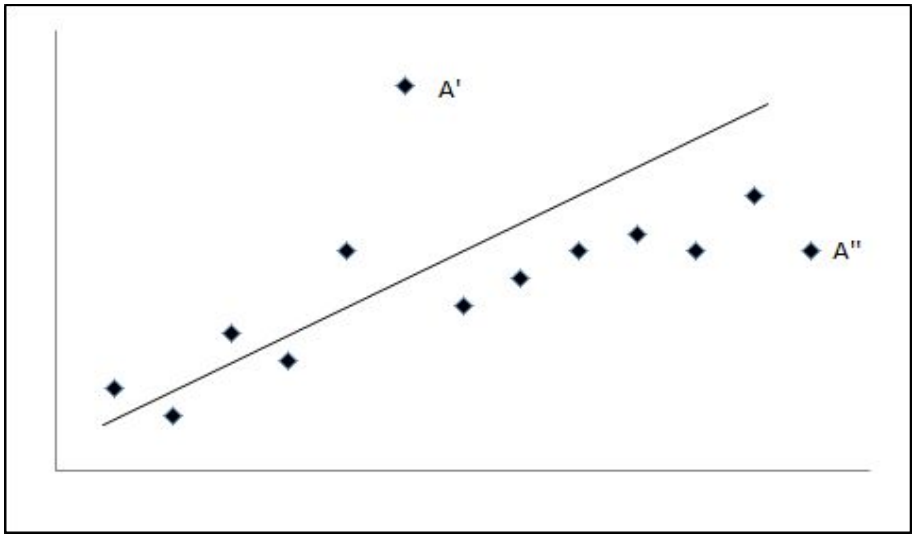
표 목 차

I. 서 론

1. 연구의 필요성 및 목적

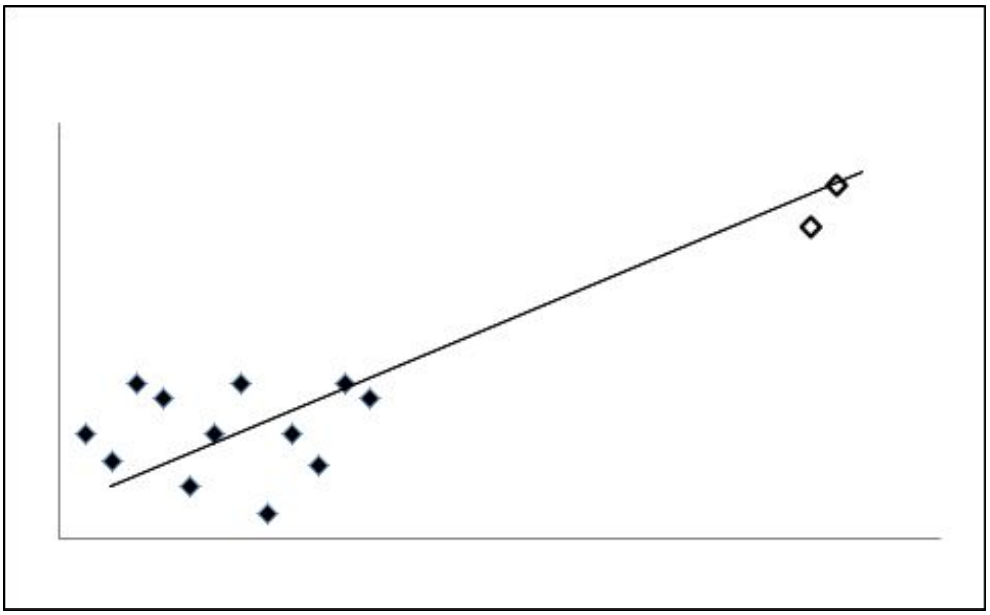
사회과학은 사회에서 일어나는 현상을 대상으로, 표본을 활용하여 다양한 변수들을 관찰하고 이를 요약하여 모집단에 적용 가능한 보다 간명한 이론 및 모형으로 설명하기 위한 과학의 한 분야이다(성태제, 2011). 연구자는 관찰, 면접, 사회조사, 실험연구 등의 연구방법을 사용하여 모집단의 속성을 추정하고자 한다. 모집단의 속성을 추정하는 과정에서 모집단의 속성은 신뢰롭고 타당한 측정도구(척도)를 통해 얻은 자료에 기초하여, 관련 변수들 간의 구조를 탐색하고 이를 활용하여 사회현상을 체계적으로 설명한다(성태제, 2011; Nicola, Kemp, & Snelgar, 2012/2013). 따라서 연구자에게 연구를 수행함에 있어 수집된 자료는 연구문제를 검증하기 위한 중요한 자료가 된다. 자료의 수집은 시간, 비용과 같은 문제로 모집단을 대상으로 할 수 없기 때문에 표집(Sampling)을 통해 구성된 표본(Sample)으로부터 자료를 수집하게 된다. 이렇게 수집된 자료에 대한 정보는 모집단의 정보를 추정하는데 사용한다. 따라서 표본은 모집단을 대표할 수 있는 집단으로 구성되어야 하지만 수집된 자료들이 모두 모집단을 대표할 수 있는 자료가 되지 않을 수도 있다. 구체적으로, 표본은 모집단의 일부분을 표집한 것이기 때문에 표집 시 연구자가 일반화하고자 하는 모집단의 속성을 정확하게 반영하는 관찰치들과 함께 분석결과에 큰 영향을 미치는 극단치(outlier)와 영향력 있는 관찰치(influential case)들도 포함되어 있다. 따라서 표본을 활용하여 모집단의 속성을 정확하게 추정하기 위해서는 극단치와 영향력 있는 관찰치를 식별하는 것은 연구자들에게 매우 중요한 일이다(문수백, 2009; 이희연, 노승철, 2013; Chatterjee & Yilmaz, 1992).

극단치(outlier)는 대부분의 유사한 양상을 보이는 자료들과 비교해서 다른 메커니즘에 의해 생성되었다고 보일 정도로 유사하지 않거나 또는 상반된 특성을 보이는 관찰치로 이상치라고도 한다(Hawkins, 1980). 다시 말해 대부분의 자료들과 다르게 비정상적으로 극단적인 값을 보이는 것을 극단치라고 한다(신형원, 손소영; 2001; Chatterjee & Yilmaz, 1992). 반면에 대다수의 점들이 위치한 분포지역에서 크게 벗어났지만 비슷한 패턴을 보이는 점들로 추정치에 영향을 미치는 관찰치를 영향력 있는 관찰치(influential case)라고 한다(이희연, 노승철, 2013; Chatterjee & Yilmaz, 1992). 극단치와 영향력 있는 관찰치를 구분하는 것은 쉽지 않다(Chatterjee & Yilmaz, 1992). 관찰치 중 일부는 극단치와 동시에 영향력 있는 관찰치 이거나 또는 극단치는 아니지만 영향력 있는 관찰치일 수 있다. 예를 들어 <그림1>에서 A′과 A″을 비교해보면, A′은 다른 관찰치들과 다르게 추정된 회귀선에서 많이 벗어나 있음을 볼 수 있다. 따라서 A′은 극단치로 볼 수 있는 반면에 A″은 추정된 회귀선에서 벗어났지만 다른 관찰치들과 유사한 패턴을 보이고 있기 때문에 영향력 있는 관찰치로 볼 수 있다.

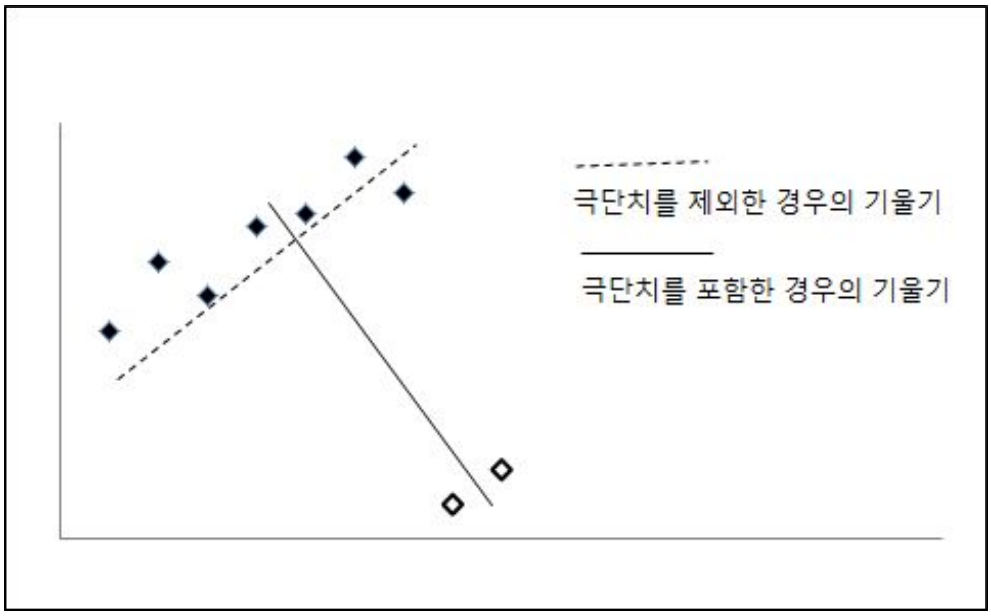


<그림 1> 극단치와 영향력 있는 관찰치의 차이

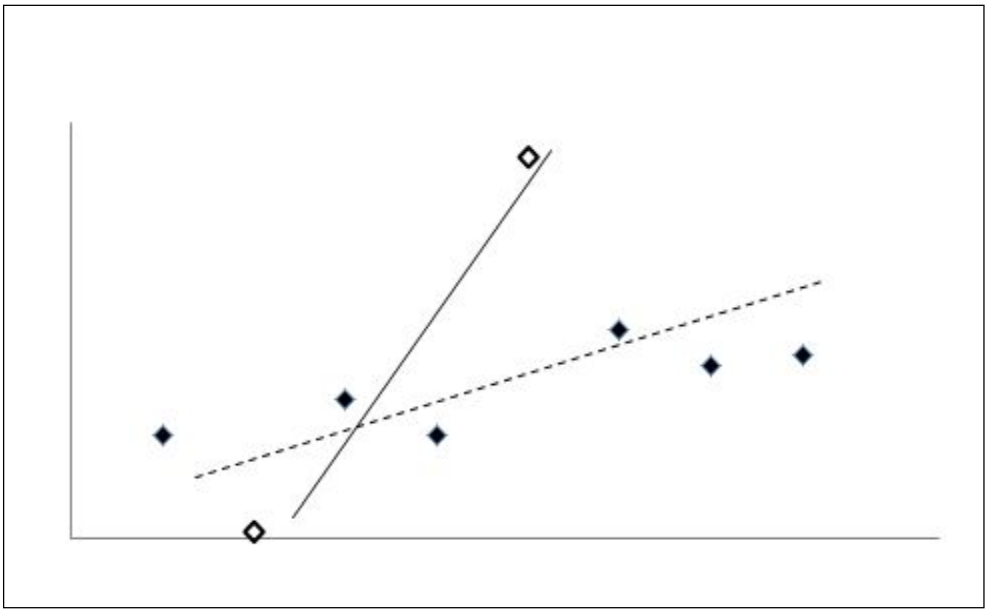
영향력 있는 관찰치와 극단치는 모집단의 속성을 추정하는 모형의 모수 추정에 편향(bias)을 발생 시킬 수 있다. 모형 및 모수의 추정에 편향이 발생하는 경우는 크게 세 가지 경우로 나타날 수 있다. 첫째, 극단치 또는 영향력 있는 관찰치로 인해 회귀선이 추정되는 경우이다. <그림 2>의 회귀선은 오른쪽 끝에 있는 두 개의 관찰치를 제외하면 나머지 관찰치에서는 자료를 적절하게 요약하는 회귀모형을 추정하기가 어렵다. 따라서 오른쪽 끝에 있는 두 개의 관찰치로 인해 회귀선이 추정되었다고 볼 수 있다. 둘째, 회귀선의 방향이 바뀌는 경우이다. <그림 3>의 경우는 오른쪽 하단에 보이는 두 개의 관찰치로 인해 원래 추정되어야할 회귀선인 점선의 방향이 실선의 방향으로 완전히 달라지는 것을 볼 수 있다. 마지막은 회귀선의 기울기에 영향을 미치는 경우로 <그림 4>와 같다. 위아래로 있는 관찰치를 제외한 경우의 회귀선인 점선과 관찰치를 포함한 회귀선인 실선은 기울기의 차이를 보인다. 극단치 또는 영향력 있는 관찰치로 인해 회귀선의 기울기가 상대적으로 보다 급해진 것을 알 수 있다.



<그림 2> 극단치와 영향력 있는 관찰치가 회귀선에 미치는 영향(1)



<그림 3> 극단치와 영향력 있는 관찰치가 회귀선에 미치는 영향(2)



<그림 4> 극단치와 영향력 있는 관찰치가 회귀선에 미치는 영향(3)

세 가지 경우에서 보이듯이, 극단치와 영향력 있는 관찰치의 위치가 추정된 회귀식에 영향을 미친다(이희연, 노승철, 2013). 즉, 극단치와 영향력 있는 관찰치가 연구 결과를 왜곡시킬 수 있다. 따라서 자료를 분석하기 전에 자료를 진단하는 것은 정확한 연구 결과를 산출하기 위한 중요한 작업이다(문수백, 2009; 이희연, 노승철, 2013; Bollen & Arminger, 1991; Chatterjee & Yilmaz, 1992; Rousseeuw & Leroy, 2005).

극단치를 식별하는 방법으로는 모형에 기반을 두지 않는 방법(model-free method)과 모형기반방법(model-based method)이 있다(Bollen & Arminger, 1991).

모형에 기반을 두지 않는 방법은 통계적 모형에 관계없이 모든 모형에 적용 가능한 방법을 말한다. 모형에 기반을 두지 않는 방법으로는 히스토그램(histogram), 산포도(scatter plot), 상자도표(box plot)등이 있다.

모형에 기반을 두지 않는 방법의 장점은 그림이나 도표를 활용하여 극단치를 식별할 수 있기 때문에 연구자가 보다 쉽게 접근할 수 있다는 것이다. 반면에 연구자의 주관적인 판단에 의해서 극단치를 결정해야 하기 때문에 객관성이 결여된다는 단점이 있다(이희연, 노승철 2013; Rousseeuw & Leroy, 2005). 또한 단변량(univariate)의 경우 변수간의 관련성을 쉽게 그림으로 나타낼 수 있지만 다변량(multivariate)의 경우 변수들 간의 관련성을 그림으로 표현하기 어렵다는 한계를 가진다(성태제, 2011; Rousseeuw & Leroy, 2005).

반면에 모형기반방법을 통한 극단치 식별에는 거리를 이용한 Cook's distance, 마할라노비스 거리(Mahalanobis distance)와 잔차의 범위를 이용한 방법인 표준화 잔차(Standardized Residuals)등이 있다(이희연, 노승철, 2013; Pang-Ning, Steinbach & Kumar 2006). 모형기반방법들은 다변량에서도 활용 가능한 방법들로 계산이 어렵긴 하지만 기준 통계치와 자료를 통

해 계산한 통계치를 비교하여 극단치를 식별할 수 있다. 따라서 모형기반방법은 모형에 기반을 두지 않는 방법보다 객관적인 기준을 가지고 극단치를 식별할 수 있다는 장점이 있다(이희연, 노승철, 2013). 앞에서 기술된 모형기반방법은 회귀모형에 기초한 극단치 식별 방법이다.

그런데 최근에는 회귀모형보다 잠재변수를 활용하여 측정변수의 측정오차를 통제하는 요인분석모형(Factor Analysis Model)과 보다 확장된 구조방정식모형(Structural Equation Modeling)이 심리학을 포함한 사회과학의 다양한 영역에서 보다 빈번하게 사용되고 있다(김주환, 김민규, 홍세희, 2009; 홍세희, 2000). 이러한 배경에는 위에서 기술된 모형들이 측정변수들 간의 관련성을 요약하는 잠재변수를 포함시킴으로써 측정오차(measurement error)를 통제하고, 이론적으로 관심이 있는 모형 평가에서 다양한 합치도 지수(fit measures)가 활용 가능하며, 매개효과의 추정 및 유의도 검증의 간명성 등이 있다(문수백 2009; 홍세희, 2000; 홍세희, 2003).

이러한 요인분석모형과 구조방정식모형은 회귀모형과는 다르게 개별 자료로부터 얻어진 공분산행렬 또는 상관행렬에 기초하여 자료를 분석하는 모형으로(배병렬, 2009; 이희연, 노승철, 2013) 개별 자료에 초점을 두지 않았을 뿐만 아니라 잠재변수를 가진 모형의 계산이 더 복잡하기 때문에 극단치 식별이 상대적으로 보다 어렵다(Bollen & Arminger, 1991). 그럼에도 불구하고 많은 연구자들이 잠재변수를 포함한 모형의 극단치 식별을 위해 다양한 방법들을 제안하였다(Aguinis, Gottfredson & Joo, 2013).

잠재변수를 포함한 모형의 극단치 식별을 위한 다양한 연구들 중에서 마할라노비스 거리(Mahalanobis distance)를 극단치 식별에 사용할 수 있다고 많은 연구자들이 주장하고 있다(문수백, 2009; 배병렬,2009; 이희연,노승철, 2013; Kline, 2011). 마할라노비스 거리는 관찰변수들의 평균과 개별 관찰변수와의 거리를 통해 극단치를 식별한다. 그러나 이러한 과정에서 모형에서

예측된 값들이 아닌 표본에서 얻은 관찰치의 평균과 공분산을 이용하여 값을 산출하기 때문에 산출된 값은 관찰치의 평균과 공분산이 모형의 변화에 민감하게 반응한다. 따라서, 동일한 결과를 산출할 수 없다는 단점이 있다. 반면에 잔차를 이용한 극단치의 식별에서는 결과가 모형의 변화에 영향을 받지 않기 때문에 더 유용하게 사용될 수 있다(안병진, 서한손, 2011).

따라서 본 연구에서는 확인적 요인분석모델에서 모형에서 재생산된 값과 관찰된 값을 비교하여 극단치를 식별할 수 있는 표준화 잔차 점수의 유용성을 검증하고자 한다. 연구목적을 달성하기 위해서 시뮬레이션 자료를 활용하여 표준화 잔차 점수의 분포의 가정을 검증하고, 1종 오류(Type-I error)와 탐지율(Detection rates)을 검증하고자 한다.

Ⅱ. 이론적 배경

1. 극단치의 식별에 대한 전통적인 접근방식

극단치를 탐색하는 방법에는 크게 두 가지 방법으로 모형에 기반을 두지 않는 방법과 모형기반방법이 있다(Bollen & Arminger, 1991).

1) 모형에 기반을 두지 않는 방법

모형에 기반을 두지 않는 방법은 모집단의 특성을 요약해주는 모형의 종류와 무관하게 적용 가능한 방법이다. 이 방법은 히스토그램(histogram), 산포도(scatter plot), 상자도표(box plot)와 같이 그림으로 표현하는 방식으로 극단치를 식별할 수 있다. 모형에 기반을 두지 않는 방법을 활용하여 극단치를 식별하는 예시에는 청소년 패널 데이터[1] 중에서 중학교 2학년 929명의 키와 몸무게 자료가 사용되었다. 청소년 패널 데이터에서 키의 평균과 표준편차는 각각 162.81과 6.94이며 몸무게의 평균과 표준편차는 각각 53.22와 8.65이다.

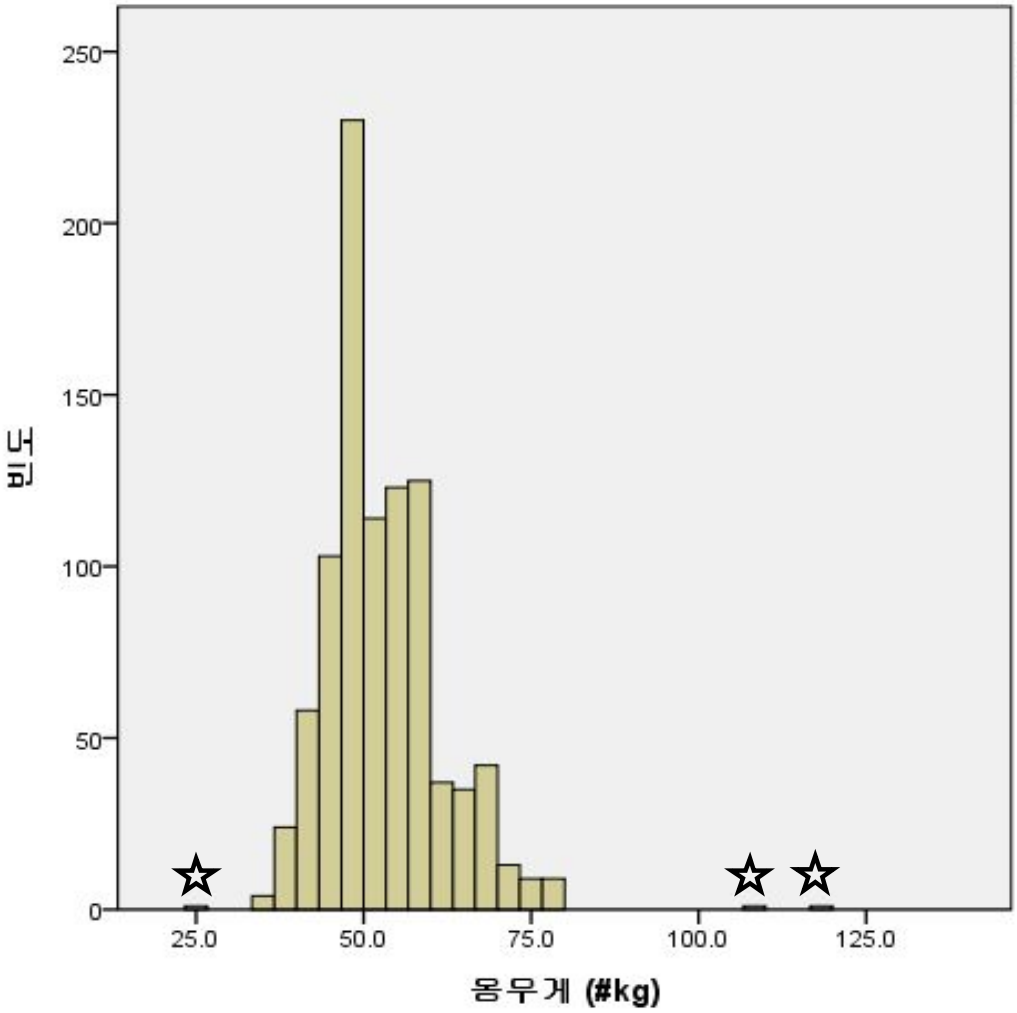
(1) 히스토그램(histogram)

히스토그램은 막대로 주어진 현상의 특성을 나타내는 그래프로 선 그래프가 양적 비연속변수일 때 가진 모든 자료의 수를 이용하여 그려야 하는 취약점을 해결하기 위해 사용될 수 있는 그래프이다(성태제, 2011; 이희연, 노

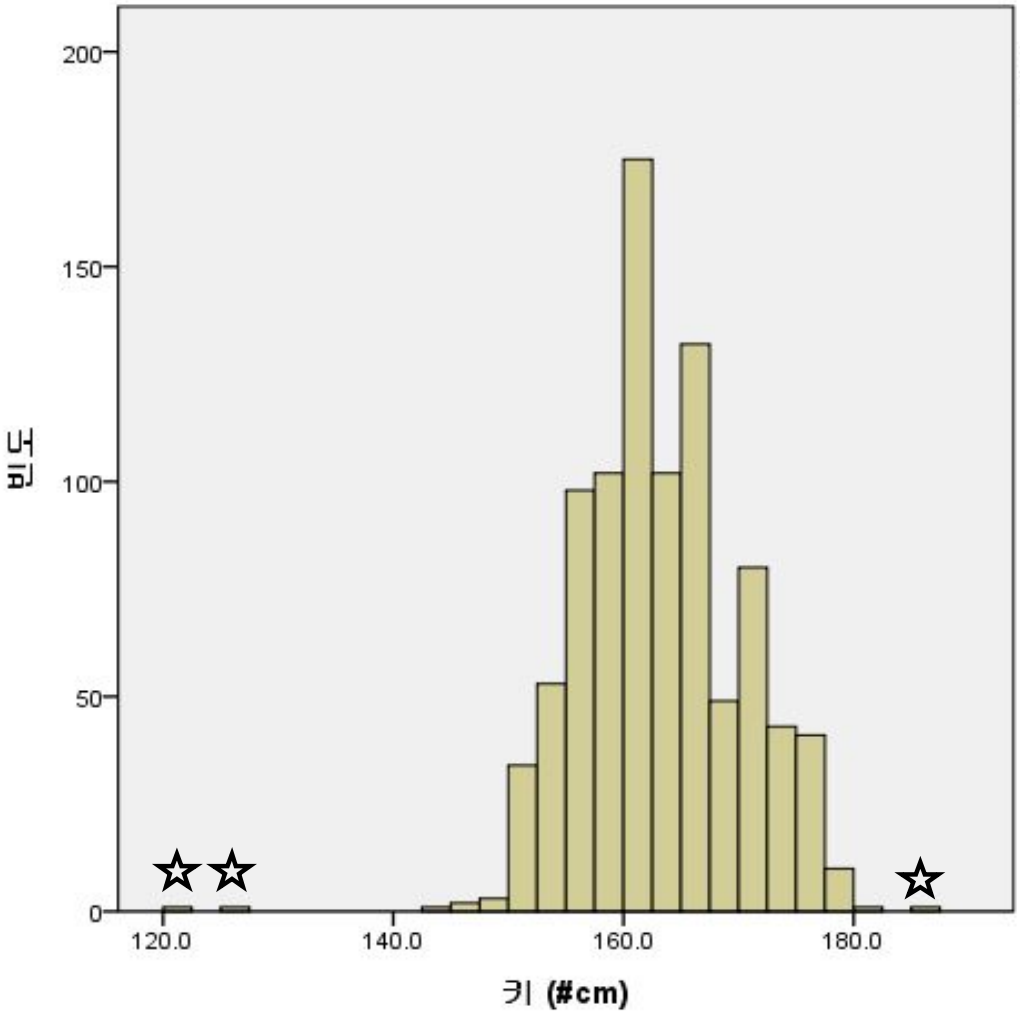
1) 예시에 사용된 자료는 한국청소년정책연구원에서 중학생을 대상으로 실시한 종단 연구인 『한국아동청소년패널조사(KCYPS)』의 중학교 1학년 패널데이터 2차년도 자료 중 일부를 사용하였으며, 보다 정확한 이해를 돕기 위해 5개의 극단치 사례를 임의로 설정하였다.

승철, 2013). 따라서 히스토그램은 이론적으로 관심이 있는 양적 연속 변수일 때 사용할 수 있는 도표로 가로축은 계급구간을 나타내는 급간으로 표시하고 세로축은 빈도를 나타내며 <그림 5>, <그림 6>과 같다. 히스토그램에서 극단치의 식별은 자료가 집중되어 있는 범위를 벗어난 자료들로 볼 수 있다. 따라서 중학교 2학년의 몸무게를 나타낸 <그림 5>에서 양 끝의 별표로 표시된 자료들이 집중되어 있는 구간을 벗어나 있기 때문에 극단치로 볼 수 있으며, 중학교 2학년의 키를 나타낸 <그림 6>에서도 별표로 표시된 양 끝의 자료를 극단치로 볼 수 있다.

보다 구체적으로, 몸무게와 키의 극단치 값을 평균과 비교해보면 <그림 5>의 몸무게의 경우 극단치로 보이는 관찰치 값들은 25kg, 110kg, 120kg로 몸무게 평균인 53.22kg과 두배가 넘는 차이가 나는 것을 알 수 있다. <그림 6>의 키 또한 극단치로 보이는 관찰값들이 120cm, 125cm, 185cm로 평균인 162.81cm와 20cm가 넘게 차이가 난다. 이는 극단치로 보인는 관찰치들이 다른 집중되어 있는 자료들과 다르게 평균에서 멀리 떨어져 있으며 그 빈도 또한 낮아 잘 관찰되지 않는 사례를 의미한다. 즉, 이러한 값들은 상대적으로 전체 자료를 대표할 수 없다는 것을 의미한다(백문수, 2009).



<그림 5> 중학교 2학년의 몸무게 히스토그램



<그림 6> 중학교 2학년의 키 히스토그램

(2) 상자도표(box plot)

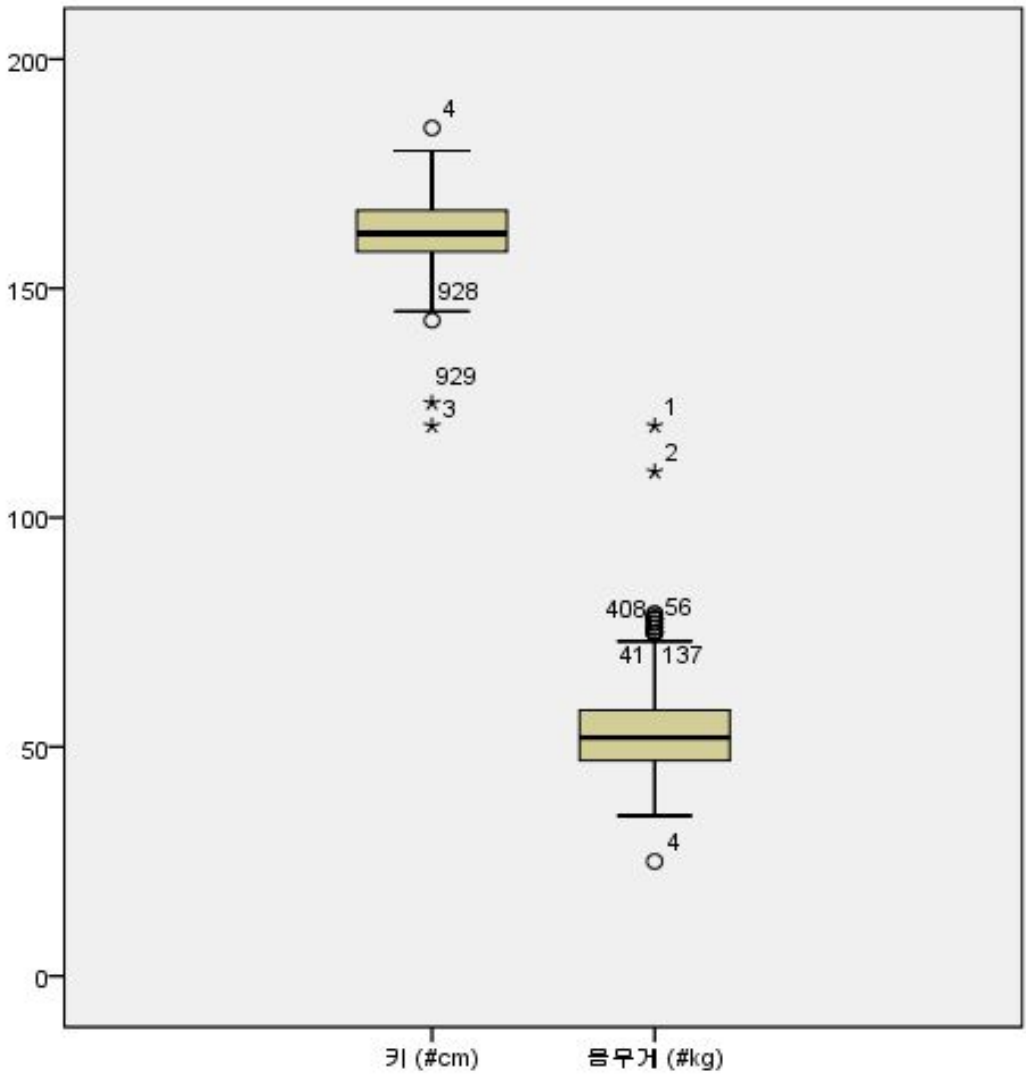
상자도표는 극단치를 제외한 자료를 5개의 지표인 최소값, 1사분위수(Q1), 중앙값(Q2), 3사분위수(Q3), 최대값으로 요약한 도표로 <그림 7>과 같다(문수백, 2009; 성태제, 2011; 이희연, 노승철, 2013; Aguinis, Gottfredson, & Joo, 2013). 중앙값은 자료를 순서대로 정리했을 때 가운데 오는 값으로 50백분위수에 해당하는 값이며, 1사분위수와 3사분위수는 각각 25백분위수와 75백분위수에 해당하는 값이다. 최소값과 최대값은 3사분위수에서 1사분위수를 뺀 값의 3배 사이에 있는 하한값과 상한값이다[2].

상자도표는 자료의 중앙값과 분산을 시각적으로 표현하여 여러 집단 간 또는 변수 간의 분포 특성을 한 눈에 비교할 수 있으며, 데이터의 분산 대칭여부와 양과 음의 왜곡된 정도, 관찰치의 범위, 극단치를 파악할 수 있다(이희연, 노승철, 2013; Nicola, Kemp, & Snelgar 2012/2013). 보다 구체적으로 중앙값과 사분위수 사이의 거리로 비대칭도를 파악할 수 있으며, 3사분위수에서 1사분위수를 뺀 값인 범위(상자의 길이)를 1.5배 이상 벗어난 자료가 그래프에 점 또는 별모양으로 나타나기 때문에 극단치를 식별하기 쉽다.

보다 쉽게 설명하기 위해 중학교 2학년의 키와 몸무게를 비교한 상자도표인 <그림 7>을 보자. 키와 몸무게를 나타낸 각 상자도표에서 상자의 중앙값을 표현한 선이 정중앙에 있으며 각각의 값은 162cm와 52kg을 나타낸다. 중앙값을 나타내는 선을 기준으로 키와 몸무게 모두 상자의 모양이 거의 비슷하다. 이는 두 변수의 분산이 대칭이라는 것을 의미한다. 상자의 길이를 나타내는 범위는 3사분위수와 1사분위수의 차이로 키에서는 167과 158의 차로 그 값이 9이고, 몸무게에서는 58과 47의 차로 11이다. 키와 몸무게의 최소값과 최대값을 계산하면 키는 153.5와 180.5를 몸무게는 31.5와 74.5의 값

2) 최대값: (Q3+ 1.5(Q3-Q1))내의 최대값
최소값: (Q1-1.5(Q3-Q1))내의 최소값

이 나온다. 따라서 극단치들은 최소값과 최대값의 범위를 벗어난 관찰치로 점과 별로 표시한다. 점은 1사분위수와 3사분위수로부터 상자 길이의 1.5배 이상 벗어난, 즉 최소값과 최대값을 벗어난 관찰치를 나타내며, 별은 1사분위수와 3사분위수로부터 상자길이의 3배 이상 벗어난 관찰치 이다. 따라서 <그림 7>의 키에서 3번과 929번 자료는 관찰치가 125와 120으로 1사분위수로부터 상자길이의 3배인 131보다 작으며, 928번 자료의 관찰치는 143으로 최소값인 153.5보다 작다. 1사분위수로부터 4번 자료의 관찰치는 185로 최대값인 180.5보다 큰 관찰치 임을 알 수 있다. 몸무게를 나타낸 상자도표에서도 역시 1번과 2번 자료의 관찰치가 각각 120과 110으로 3사분위수로부터 상자길이의 3배인 91보다 크고, 4번 자료와 408, 56, 41, 137번 자료의 값이 각각 25, 74.9, 75, 75로 최소값과 최대값인 153.5와 180.5의 범위를 벗어난 관찰치라는 것을 확인할 수 있다. 따라서 연구자는 이를 참고하여 위에 기술된 번호의 자료들을 극단치로 간주 할 수 있다.

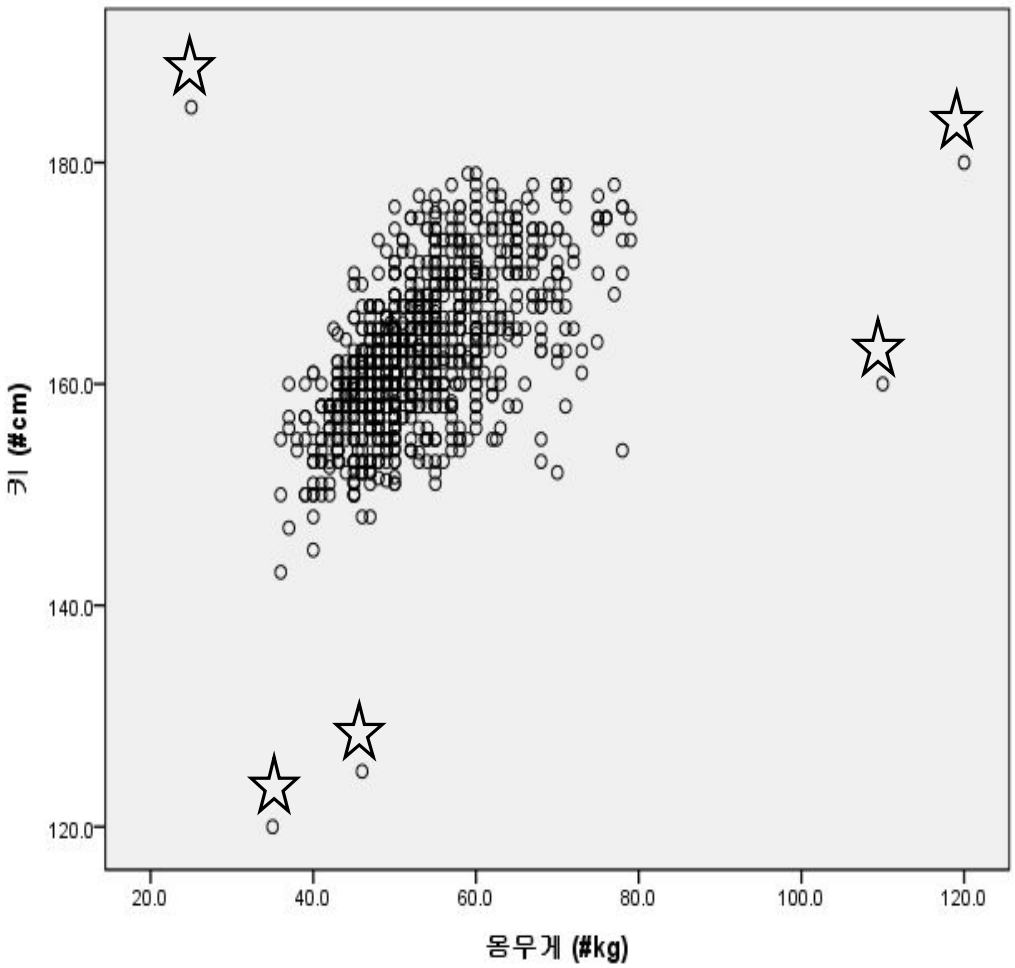


<그림 7> 중학 교 2학년의 키와 몸무게 상자도표

(3) 산포도(scateer plot)

산포도는 <그림 8>과 같이 두 변수의 값을 도표에 나타낸 것으로 두 변수 간의 관계를 알아보기 위해 사용한다. 산포도를 통해 극단치의 식별, 두 변수 간 관련성의 유무, 관련성의 방향성과 같은 두 변수간의 관계에 대한 대략적인 정보를 얻을 수 있지만 관련성의 정도에 대한 정확한 정보는 얻지 못한다. 보다 구체적으로 동일한 산포도로부터 극단치의 식별, 관련성의 유무, 관련성의 정도가 사람마다 다를 수 있다. <그림 8>은 중학교 2학년의 키와 몸무게의 관계를 산포도로 나타낸 것이다. 앞서 기술한 바와 같이 극단치는 대부분의 자료들과 다른 양상을 보이는 값을 말한다. 따라서 이 그림에서 대부분의 밀집되어 있는 자료와 떨어져 별표로 표시된 자료를 극단치로 판단 할 수 있다.

위에 기술한 바와 같이 모형에 기반을 두지 않는 방법은 그림을 통해 극단치를 식별할 수 있기 때문에 누구나 쉽게 접근할 수 있다. 하지만 동일한 자료를 분석함에도 불구하고 극단치로 식별되는 관찰치가 분석에 사용되는 방법에 따라 상이할 수 있다. 예를 들어 위의 그래프들은 모두 동일한 자료인 중학교 2학년의 키와 몸무게 자료를 분석한 결과임에도 불구하고 히스토그램과 산포도는 동일한 자료들을 극단치로 식별하였지만 상자도표의 경우 다른 그래프들과 다르게 더 많은 극단치를 보고하고 있다. 이는 어떤 그래프로 분석하느냐에 따라 극단치에 대한 평가 및 판단이 달라질 수 있다는 것을 의미한다. 이와 더불어 연구자에 따라 극단치를 결정하는데 있어 기준이 다를 수 있다는 점을 감안하면 모형에 기반을 두지 않는 방법을 통한 극단치의 식별은 주관적일 수밖에 없다. 따라서 극단치를 객관적으로 식별할 수 있는 기준이 필요하며, 극단치에 대한 통계적 유의성 검증을 통해 객관적으로 판단할 수 있는 방법으로 모형기반방법이 있다.



<그림 8> 중학교 2학년의 키와 몸무게의 산포도

2) 모형기반방법

모형기반방법은 모형에 기반을 두지 않는 방법과 다르게 모형 구조를 고려한 방법으로 연구자가 설정한 통계적 모형으로부터 벗어난 관찰변수를 극단치로 식별한다(Maimon, & Rokach, 2005; Williams, Baxter, Hawkins, & Gu, 2002; Bollen & Arminger, 1991). 이러한 방법에는 거리를 이용한 Cook's distance 및 마할라노비스 거리(Mahalanobis distance)와 잔차의 범위를 이용한 방법인 표준화 잔차(Standardized Residuals)등이 있다(배병렬, 2009; 이희연, 노승철, 2013).

(1) Cook's distance

Cook's distance는 특정 자료가 전체모형 또는 예측치에 영향을 미치는지의 여부를 판단하는 지수로 식(1.2.1)과 같다.

$$C_i = \frac{\sum[\hat{Y}_i - \hat{Y}_i(i)]^2}{\hat{\sigma}_x^2(p+1)} \qquad (1.2.1)$$

식(1.2.1)에서 p는 상수항을 포함한 관찰변수의 개수를, $\hat{\sigma}_x^2$은 모집단 분산의 예측치 이다. $\hat{Y}_i(i)$는 극단치로 판단되는 i번째 사례의 관찰값을 제거한 경우에 추정된 예측치를 나타낸다. 보다 구체적으로 식에서 보듯이 i번째 사례의 관찰값을 포함한 예측치와 i번째 사례의 관찰값을 포함하지 않은 예측치의 차이를 통해 i번째 사례의 관찰값의 영향력을 판단하는 방법이다. 일반적으로 C가 1보다 클 경우 영향력 있는 사례라고 볼 수 있으며, 이는 모수의 추정치에 미치는 영향력이 크다는 것을 의미한다(이희연, 노승철, 2013).

(2) 마할라노비스 거리(Mahalanobis distance)

마할라노비스 거리는 개별 자료의 점수와 표본평균과의 다변량 거리를 의미하며, 식은 (1.2.2)와 같다.

$$MD_i^2 = (V_i - \overline{V})C^{-1}(V_i - \overline{V})^t \tag{1.2.2}$$

보다 구체적으로 관찰변수 x가 p개 있고, 사례수가 i개 있을 때, 관찰변수의 행렬을 가로만 있는 행렬인 벡터로 나타낼 수 있으며 이것을 식(1.2.3)과 같이 V_i로 나타낸다. 표본평균은 모든 관찰변수의 합을 관찰변수의 개수로 나눈 것으로 식(1.2.4)와 같이 나타낼 수 있다.

$$V_i = (x_{i,1}\ x_{i,2}\ x_{i,3}\ x_{i,4} \cdots\ x_{i,p-1}) \tag{1.2.3}$$

$$\overline{V} = \frac{1}{n}\sum_{i=1}^{n} V_i \tag{1.2.4}$$

그리고 V_i의 공분산인 C를 구하면 식(1.2.5)와 같다. $(V_i - \overline{V})^t$는 $(V_i - \overline{V})$의 전치행렬(Transposed matrix)이다.

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \overline{V})^t(V_i - \overline{V}) \tag{1.2.5}$$

위의 식에서 보는 것과 같이 마할라노비스 거리는 i번째 사례의 관찰변수들의 값인 V_i와 관찰변수들의 평균인 $\overline{V}$와의 거리를 나타내는 것을 알 수

있다. 따라서 만일 관찰변수에 대한 마할라노비스 거리가 0이라면, 응답자의 모든 관찰점수는 그들 각각의 평균과 같으며, 수치가 클수록 평균에서 멀리 떨어져 있음을 나타낸다. 마할라노비스 거리제곱(MD_i^2)은 관찰변수의 수와 동일한 자유도를 갖는 카이스퀘어 통계량으로 극단치를 식별할 수 있다. 예를 들어, 5개 변수의 자료에서 특정 자료의 마할라노비스 거리제곱이 25.82일 경우, 유의수준 .001에서의 $\chi^2(5)$의 임계값(critical value)은 15.09이다. 따라서 이 자료는 유의수준 .001에서 나머지 자료와 유의하게 다르다고 할 수 있다(배병렬, 2009; 이희연, 노승철, 2013; Rousseeuw & Leroy, 2005; Kline, 2011).

(3) 표준화 잔차(Standardized Residuals)

통계분석에서 잔차는 연구자가 설정한 모형이 자료에 부합되는지의 여부를 판단할 수 있기 때문에 중요한 역할을 한다. 잔차란 자료에서 얻은 관찰치(y_i)와 추정된 모형에 의해 예측된 값($\hat{y_i}$)의 차이를 말하며 식(1.3.1)과 같이 나타낼 수 있다.

$$e_i = y_i - \hat{y_i} \tag{1.3.1}$$

모형이 자료에 합치한다면 자료로부터 얻은 관찰치와 추정된 모형에 의해 예측된 값이 비슷하다는 것을 의미하기 때문에 잔차의 값이 작게 나타나며, 반면에 모형이 자료에 합치하지 않는다면 잔차의 값이 크게 나타난다. 잔차의 크기는 척도의 크기에 따라 변수들마다 다르게 산출된다. 따라서 척도가 다른 변수들 사이에서 잔차의 값은 비교가 불가능하기 때문에 비교를 위해 단위를 통일시킬 필요성이 있다. 단위를 통일 시키는 과정을 표준화라고 하며표준화된 잔차를 표준화 잔차(Standardized Residuals)라고 한다. 식(1.3.2)

는 잔차 e_i를 잔차의 표준편차로 나눠준 표준화 잔차의 식을 나타낸다. σ는 모집단의 표준편차이며, h_{ii}는 예측행렬의 대각선에 있는 원소의 값을 나타낸다.

$$z_i = \frac{e_i}{\sigma\sqrt{1-h_{ii}}} \tag{1.3.2}$$

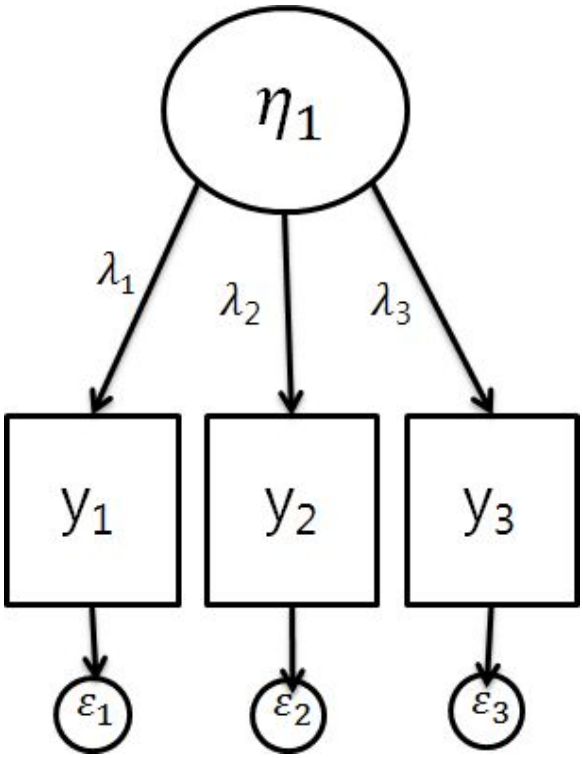
회귀모형에서 잔차는 관찰치가 회귀선에서 수직으로 떨어져 있는 정도를 나타내는 것으로 표준잔차가 ±2보다 클 때 측정치와 예측치가 5% 유의수준에서 다르다고 볼 수 있다(이희연, 노승철, 2013). 일반적으로 표준화 잔차가 ±3보다 클 경우 극단치로 식별한다(문수백, 2009; 이희연, 노승철, 2013).

모형기반방법을 통한 극단치의 식별은 앞서 기술한 모형에 기반을 두지 않는 방법과 달리 비교 가능한 기준 통계치가 있기 때문에 객관적으로 극단치를 식별할 수 있다. 모형기반방법은 회귀모형 뿐만 아니라 잠재변수를 포함하고 있는 요인분석모형과 구조방정식모형에서도 계산이 복잡하지만 적용 가능할 것으로 보인다. Bollen과 Arminger(1998)의 연구에 따르면 요인분석에서 극단치를 포함한 모형의 표준화 잔차 값을 비교한 결과 극단치 사례의 표준화 잔차 값이 다른 정상적인 사례에 비하여 큰 값을 나타내는 것을 발견하였다. 따라서 이를 종합하여 본 연구에서는 모형기반방법 중 회귀모형에 기초한 표준화 잔차를 구조방정식의 측정모형인 요인분석모형으로 확장시켜 다변량 자료의 극단치를 식별하고자 한다.

2. 극단치의 식별에 대한 새로운 접근방식

1) 요인모형(Factor Model)

요인분석은 대량의 자료를 이론적으로 의미 있는 소수의 변수를 추출하는 통계방법으로 공통요인분석이라 불리기도 한다(이순묵, 2000). <그림 9>는 1요인모형으로 한 개의 잠재변수를 3개의 측정변수로 측정하는 모형이다. 예를 들어, '사회성'이라는 잠재변수를 측정한다고 하면 구성개념인 η_1이 '사회성'이 되고 사회성을 측정하는 문항인 y_1, y_2, y_3가 '나는 친구가 많다고 생각한다', '나는 사람들과 많은 모임에 참석하는 것을 좋아한다', '나는 사람들과 있으면 즐겁다'로 측정된 점수가 된다(이수연, 2010). ϵ_i는 사회성을 설명하고 난 나머지 부분으로 측정변수만의 고유요인 또는 잔차이다(이순묵, 2000). λ_i는 잠재변수와 측정변수간의 선형적인 관련성의 정도를 나타내 주는 계수로, λ_1은 η_1과 y_1의 선형적인 관련성을 λ_2는 η_1과 y_2의 선형적인 관련성을 λ_3은 η_1과 y_3의 선형적인 관련성의 정도를 나타낸다.



<그림 9> 1요인모형

위 모형을 식으로 정리하면 식(2.1.1)과 같으며 이는 측정변수와 잠재변수 간의 선형적 관련성을 보여준다.

$$y_1 = \lambda_1\eta_1 + \epsilon_1$$
$$y_2 = \lambda_2\eta_1 + \epsilon_2$$
$$y_3 = \lambda_3\eta_1 + \epsilon_3 \qquad (2.1.1)$$

식(2.1.1)은 수학적으로 회귀모형과 비슷하다. 그러나 회귀모형에서는 예측변수가 모두 직접 측정 가능한 변수인 반면에 요인모형에서의 예측변수인 η는 직접측정 되지 않는다.

요인모형식인 식(2.1.1)을 행렬로 표현하면 식 (2.1.2)와 같다.

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i \tag{2.1.2}$$

식(2.1.2)에서 $\boldsymbol{y}_i$는 관찰변수로 p x 1 벡터이며, $\boldsymbol{\eta}_i$는 서로 독립적이며 동일한 분포(Independently Identically Distributed)를 따르는 무작위 잠재변수 즉, 요인으로 m x 1 벡터이다. $\boldsymbol{\epsilon}_i$역시 독립적이며 동일한 분포를 따르는 무작위 오차변수의 p x 1 벡터로 요인모형에서는 잔차라고도 한다. $\boldsymbol{\Lambda}$는 잠재변수인 $\boldsymbol{\eta}_i$와 측정변수인 $\boldsymbol{y}_i$의 관련성을 포함하고 있는 p x m 행렬로 정규분포를 따르며, 가중치 혹은 요인계수(Factor loading)라고 한다.

요인분석모형은 측정변수에서 모든 공통적인 잠재변수들이 추출된다면 변수들 간에는 더 이상의 상관관계가 존재하지 않을 것이라고 해석할 수 있다. 따라서 공통적인 잠재변수를 설명하고 남은 잔차 간에도 상관이 존재하지 않음을 가정하며 이를 지역독립성(Local independent)이라고 한다. 위의 가정은 아래와 같이 정리할 수 있다.

$$E(\boldsymbol{\eta}_i) = \mathbf{0} \tag{2.1.3}$$

$$E(\boldsymbol{\epsilon}_i) = \mathbf{0} \tag{2.1.4}$$

$$E(\boldsymbol{\eta}_i, \boldsymbol{\epsilon}_i') = \mathbf{0} \tag{2.1.5}$$

$$E(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_i') = \boldsymbol{\Theta} \tag{2.1.6}$$

식(2.1.3)과 식(2.1.4)는 잠재변수 $\boldsymbol{\eta}_i$와 오차변수 $\boldsymbol{\epsilon}_i$의 평균이 각각 0임을 가정하다. 식 (2.1.5)은 잠재변수와 잔차 간 공분산이 0임을 나타내는 식으로 다르게 표현하면 $COV(\boldsymbol{\eta}_i, \boldsymbol{\epsilon}_i') = \mathbf{0}$과 같다. 식(2.1.6)은 잔차 간 공분산인

$COV(\epsilon_i, \epsilon_i') = \Theta$를 나타내는 대각행렬(Diagonal Matrix)이다.

2) 잔차(Residual)의 산출

잔차는 이전에 기술한 바와 같이 자료에서 얻은 관찰치(y_i)와 추정된 모형에 의해 예측된 값($\hat{y_i}$)의 차이를 비교한 값으로 각 사례별로 잔차의 값을 산출 할 수 있다. 요인모형에서는 각 사례별 관찰치로부터 얻은 잔차의 값과 모형에 기반하여 산출된 잔차의 값을 비교하여 두 값의 유사정도에 따라 극단치를 식별 할 수 있다. 이 모형은 다변량분석이기 때문에 여러 변수의 차이를 동시에 고려하여 잔차를 산출하게 되며, 모형에 기초하여 각 사례별로 잔차를 산출하는 방법은 아래와 같다.

(1)비표준화 잔차

요인모형 식인 식(2.1.2)를 바탕으로 오차에 대한 식(2.2.1)로 전환할 수 있다.

$$\epsilon_i = y_i - \Lambda\eta_i \qquad (2.2.1)$$

식(2.2.1)에서 ϵ_i는 잔차, y_i는 관찰변수, Λ는 가중치, η_i는 잠재변수를 나타낸다. 식(2.2.1)에서 우리는 관찰변수인 y_i의 값만 알고 있기 때문에 식을 풀어낼 수 없다. 따라서 잔차 ϵ_i를 추정하기 위하여 η_i에 대한 값을 추정해야 한다. i번째 관찰요인점수의 추정치를 $\hat{\eta}_i$을 사용하여 나타내면 식(2.2.2)와 같이 나타낼 수 있다.

$$\hat{\eta}_i = W y_i \qquad (2.2.2)$$

식(2.2.2)에서 W는 관찰치인 y_i로부터 요인점수의 추정치인 $\hat{\eta}_i$을 추정하기 위해 사용되는 가중치 m x p 행렬이며, ^은 추정치를 나타낸다. 요인점수추정을 위한 가중치추정방법에는 알파추정법(Alpha factoring), 이미지 추정법(Image factoring), 최대우도법(ML: Maximum Likelihood estimation), 최소제곱추정법(Least Squares method)이 있다(이순묵, 2000; 이영준, 2002; 이희연, 노승철, 2013).

보다 구체적으로 알파추정법은 변수들을 요인 별로 분석하였을 때 각 요인에 해당되는 변수들의 신뢰계수인 Cronbach's a가 최대가 될 수 있도록 요인을 추출하는 방법이다. 이미지추정법은 관찰된 자료를 영상(Image)과 잔영(Anti-image)으로 구분하여, 잔영공분산행렬(Anti-image covariance matrix)로부터 영상공분산행렬(Imagecovariance matrix)을 계산해내는 방법이다. 여기서 영상은 한 변수를 종속변수로 다른 모든 변수를 독립변수로 하였을 때 설명은 분산을 잔영은 설명되지 않은 분산을 말한다.

최대우도법은 표본으로부터 계산된 상관행렬이 주어졌을 때, 이로부터 모집단에서의 모수를 추정하는 방법이다. 즉 모집단에서의 요인계수가 어떤 값들이어야 연구자가 표본으로부터 측정한 값들이 나올 확률이 가장 높은지 알아내는 방법이다. 최소제곱추정법은 표본으로부터의 상관행렬과 재생산된 상관행렬과의 차의 제곱합 즉, 잔차를 최소화하는 방법이다.

식(2.2.2)의 $\hat{\eta}_i$을 식(2.2.1)의 η_i에 대입하면 식(2.2.3)을 얻을 수 있다.

$$\begin{aligned} \hat{\epsilon}_i &= y_i - \Lambda \hat{\eta}_i \\ &= y_i - \Lambda W y_i \\ &= (I - \Lambda W) y_i \end{aligned} \qquad (2.2.3)$$

식(2.2.3)에서 $\hat{\epsilon}_i$은 자료에서 산출된 ϵ_i의 추정치에 대한 비표준화된 잔차이다. ϵ_i는 식(2.2.1)을 보면 알 수 있듯이 관찰치인 요인의 진점수로부터 추정된 오차이며, 식(2.2.3)의 $\hat{\epsilon}_i$은 요인의 진점수에서 추정된 요인점수를 뺀 값으로 요인의 추정치에 기반하여 추정된 잔차이다.

가중치를 추정할 수 있는 여러 가지 방법 중 가장 널리 알려진 방법은 최소제곱추정법이다. 최소제곱추정법에서는 모수 추정시 계산이 쉽고 통계프로그램에서 쉽게 실행되기 때문에 많이 사용되고 있다(Rousseeuw & Leroy, 2005). 최소제곱추정법은 앞서 기술한 바와 같이 잔차를 최소화하는 방법으로 식(2.2.4)와 같이 나타낼 수 있다.

$$\sum_{i=1}^{N}(\eta_i - \hat{\eta}_i)'(\eta_i - \hat{\eta}_i) \tag{2.2.4}$$

식(2.2.4)를 정리하면 식(2.2.5)와 같이 나타낼 수 있다.

$$W_r = \sum_{\eta\eta} \Lambda' \sum_{yy}^{-1} \tag{2.2.5}$$

식(2.2.5)에서 $\sum_{yy}^{-1}$는 관찰변수y_i의 모집단 공분산 행렬의 역행렬이고 $\sum_{\eta\eta}$는 잠재변수η_i의 모집단 공분산 행렬이다. 위의 식(2.2.3)에 가중치인 식(2.2.5)를 대입시키면 식(2.2.6)과 같이 최소제곱추정법에 기초한 비표준화된 잔차식을 얻을 수 있다.

$$\hat{\epsilon}_i = (I - \Lambda \sum_{\eta\eta} \Lambda' \sum_{yy}^{-1}) y_i \quad (2.2.6)$$

앞서 기술한 바와 같이 각 사례별로 산출된 오차 ϵ_i와 모형에 기반하여 산출된 잔차 $\hat{\epsilon}_i$의 유사성여부를 설명하기 위해, 식(2.2.3)에 y_i대신 식(2.1.2)를 대입시켜 정리하면 식(2.2.7)을 얻을 수 있다.

$$\begin{aligned} \hat{\epsilon}_i &= y_i - \Lambda \hat{\eta}_i \\ &= (\Lambda \eta_i + \epsilon_i) - \Lambda \hat{\eta}_i \\ &= \Lambda(\eta_i - \hat{\eta}_i) + \epsilon_i \end{aligned} \quad (2.2.7)$$

식(2.2.7)은 사례에서 산출된 잔차ϵ_i뿐만 아니라 $\hat{\epsilon}_i$이 잠재변수인 η_i로부터 추정된 요인점수인 $\hat{\eta}_i$의 차이 즉, 추정의 오차$(\eta_i - \hat{\eta}_i)$를 반영한 다는 것을 알 수 있다. 추정의 오류가 작을수록 두 잔차사이의 차이값이 작아짐을 보여준다.

극단치를 식별하기 위해 자료에서 산출된 오차 ϵ_i와 연구모형에 기반하여 추정된 잔차 $\hat{\epsilon}_i$의 유사여부를 확인할 수 있는 방법으로는 $\hat{\epsilon}_i$와 ϵ_i의 상관이 있다. 상관은 두 집단의 공분산을 각각의 표준편차로 나눈다는 정의에 의해 식(2.2.8)과 같이 나타낼 수 있다.

$$CORR(\hat{\epsilon}_i \epsilon_i') = D_{\hat{\epsilon}}^{-1} COV(\hat{\epsilon}_i \epsilon_i') D_{\epsilon}^{-1} \quad (2.2.8)$$

$D_{\hat{\epsilon}}^{-1}$와 D_{ϵ}^{-1}는 각각 $\hat{\epsilon}_i$과 ϵ_i의 모집단 표준편차의 대각 행렬이며,

$COV(\hat{\epsilon}_i\epsilon_i')$는 $\hat{\epsilon}_i$과 ϵ_i의 모집단 공분산 행렬로 식(2.2.9)와 같이 정리할 수 있다.

$$\begin{aligned}\hat{\epsilon}_i\epsilon_i' &= [(I-\Lambda W)y_i]\,\epsilon_i' \\ &= (I-\Lambda W)y_i\epsilon_i' \\ &= (I-\Lambda W)(\Lambda\eta_i+\epsilon_i)\epsilon_i' \\ &= (I-\Lambda W)(\Lambda\eta_i\epsilon_i'+\epsilon_i\epsilon_i')\end{aligned} \tag{2.2.9}$$

$$\begin{aligned}COV(\hat{\epsilon}_i\epsilon_i') &= COV[(I-\Lambda W)y_i\,\epsilon_i'] \\ &= (I-\Lambda W)[\Lambda(\eta_i\epsilon_i')+(\epsilon_i\epsilon_i') \\ &= (I-\Lambda W)(\Lambda 0+\Theta) \\ &= (I-\Lambda W)\Theta\end{aligned}$$

정리된 식(2.2.9)에 식(2.2.5)의 잔차 $\hat{\epsilon}_i$과 가중치 $W=W_r$을 대입한 식(2.2.10)은 다음과 같이 정리된다.

$$\begin{aligned}COV(\hat{\epsilon}_i\epsilon_i') &= (I-\Lambda W_{(r)})\Theta \\ &= \Theta\sum_{yy}^{-1}\Theta\end{aligned} \tag{2.2.10}$$

분산은 개념적으로 표준편차의 제곱이기 때문에 $\hat{\epsilon}_i$의 분산을 통해 표준편차인 $D_{\hat{\epsilon}}^{-1}$를 찾을 수 있으며, $\hat{\epsilon}_i$의 분산은 식(2.2.3)을 사용하여 식(2.2.11)으로 나타낼 수 있다.

$$VAR(\hat{\epsilon}_i) = (I-\Lambda W)\sum_{yy}(I-\Lambda W)' \tag{2.2.11}$$

식(2.2.11)에서 VAR는 모집단 분산을 의미하며 식(2.2.11)에 최소제곱추정

법으로 추정한 잔차를 대입하면 식(2.2.12)와 같이 정리할 수 있다. 이는 위의 식(2.2.10)인 $\hat{\boldsymbol{\epsilon}}_i$과 $\boldsymbol{\epsilon}_i$,의 공분산과 똑같이 표현된다는 것을 알 수 있다.

$$VAR(\hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{\Theta}\sum_{yy}^{-1}\boldsymbol{\Theta} = COV(\hat{\boldsymbol{\epsilon}}_i, \boldsymbol{\epsilon}_i') \tag{2.2.12}$$

위의 식들은 두 잔차 간의 상관을 계산하기 위한 표준편차를 찾기 위해 계산되었다. 이는 또한 다음에 다룰 표준화된 잔차를 계산하기 위하여 사용한다. 표준화와 비표준화의 차이는 척도에 대한 독립성의 여부이다. 비표준화의 경우 표준화와 달리 단위가 같으면 단위 그대로 비교 가능하지만 단위가 다를 경우 비교가 불가능하다(문수백, 2009). 반면에 표준화의 경우 통계적 유의성을 검증이 가중하며, 단위에 영향을 받지 않고 비교가능하기 때문에 잔차의 비교에서도 단위의 영향을 받지 않는 표준화 잔차를 사용하는 것이 더 유용하다.

(2)표준화 잔차

개념적으로 표준화는 어떤 값을 그 값의 표준편차로 나눠주는 것을 말한다. 잔차를 표준화하는 과정에서도 역시 잔차를 표준편차로 나눠주는데 이 때 표준편차는 오차의 분산에 루트를 씌운 것과 같은 개념이기 때문에 오차 분산에 루트를 씌운 값으로 나눠준다.

식(2.2.13)은 잔차의 표준편차를 고려한 naive standardization으로 케이스마다 다른 값을 가지는 e_{ji}는 j번째 변수와 i번째 관찰치의 비표준화된 잔차이며, $[\boldsymbol{\Theta}]_{jj}$는 j번째 변수에 대한 잔차 분산을 나타낸다. 식(2.2.13)은 변수에 대한 개별사례를 변수의 잔차 표준편차로 나눠서 표준화 잔차를 구하는 것을 나타낸다.

$$e_{ji} / \sqrt{[\Theta]_{jj}} \tag{2.2.13}$$

식(2.2.13)에서 $1/\sqrt{[\Theta]_{jj}}$은 자료로부터 산출된 잔차인 ϵ_{ji}의 표준편차를 포함하며, e_{ji}의 추정방법에 따라 분산이 달라진다. 본 연구에서는 추정방법으로 최소제곱추정법을 사용한다. 식(2.2.14)는 일반적인 오차분산인 식(2.2.11)을 대입한 식을 나타내며 $[VAR(\hat{\boldsymbol{\epsilon}}_i)]_{jj}$는 $VAR(\hat{\boldsymbol{\epsilon}}_i)$의 j번째 대각 요소이다.

$$\hat{e}_{ji} / \sqrt{VAR[\hat{\boldsymbol{\epsilon}}_i]_{jj}} \tag{2.2.14}$$

식(2.2.15)는 최소제곱추정법을 이용한 잔차식인 식(2.2.12)를 대입한 식이다.

$$e_{ji} / \sqrt{VAR[\hat{\boldsymbol{\epsilon}}_i]_{jj}} \tag{2.2.15}$$

식(2.2.14), (2.2.15)를 적용하면, 표준화가 되었기 때문에 표준편차가 1이고 평균이 0인 랜덤 변수를 얻는다.

3) 통계적 유의성 검증

$\boldsymbol{\epsilon}_i$와 $\boldsymbol{\eta}_i$가 정규분포를 따른다고 가정하면, $\boldsymbol{y}_i$는 정규분포를 따르고 식(2.2.3)에 의해서 $\hat{\boldsymbol{\epsilon}}_i$도 정규분포를 따른다. 따라서 앞서 기술한 표준화 잔차도 정규분포를 따르기 때문에 통계적 유의성 검증을 할 수 있다. 여러변수에 대한 통계적 검증에는 χ^2분포를 사용하며 산출 과정은 아래와 같다.

식(2.2.13)을 제곱하면 식(2.2.16)과 같다. 식(2.2.16)에서 q_i는 χ^2값을, Θ^{-1}은 ϵ_i분산의 전치행렬을 나타내며, 이 식은 자유도 p에 대한 χ^2값이다.

$$q_i = \hat{\epsilon}_i' \Theta^{-1} \hat{\epsilon}_i \tag{2.2.16}$$

식(2.2.16)을 Θ^{-1}대신에 $\hat{\epsilon}_i$를 대입하여 식을 다시 풀면 식(2.2.17)과 같고, 이 식은 자유도 p에 대한 χ^2분포를 따른다.

$$Q_i = \hat{\epsilon}_i' [VAR(\hat{\epsilon}_i)]^{-1} \hat{\epsilon}_i \tag{2.2.17}$$

따라서 표준화 잔차를 이용한 극단치의 식별시 χ^2분포를 사용하여 통계적 유의성 검증을 할 수 있다.

Ⅲ. 연구문제 및 가설

본 연구의 연구문제는 다음과 같다.

[연구문제 1] 표준화 잔차 점수의 분포는 χ^2분포를 가정하는가?

가설 1-1. 표준화 잔차 점수의 평균은 χ^2의 기댓값과 같을 것이다.

가설 1-2. 표준화 잔차 점수의 분산은 χ^2의 분산과 같을 것이다.

[연구문제 2] 분석모형의 복잡성은 극단치 탐지율과 1종 오류 확률에 영향을 미치는가?

가설 2-1. 복잡한 모형보다 단순한 모형에서 극단치 탐지율이 높을 것이다.

가설 2-2. 단순한 모형에서 1종 오류 확률은 기댓값과 동일할 것이다.

가설 2-3. 복잡한 모형에서 1종 오류 확률은 기댓값과 다를 것이다.

[연구문제 3] 표본크기는 극단치 탐지율과 1종 오류 확률에 영향을 미치는가?

가설 3-1. 표본크기가 큰 모형에서 극단치의 탐지율이 높을 것이다.

가설 3-2. 표본크기가 큰 모형에서 1종 오류 확률은 기댓값과 동일할 것이다.

가설 3-3. 표본크기가 작은 모형에서 1종 오류 확률은 기댓값과 다를 것이다.

[연구문제 4] 극단치의 비율은 극단치 탐지율과 1종 오류 확률에 영향을 미치는가?

가설 4-1. 극단치의 비율이 낮은 조건에서 극단치의 탐지율이 높을 것이다.

가설 4-2. 극단치의 비율이 낮은 조건에서 1종 오류 확률은 기댓값과 동일할 것이다.

가설 4-3. 극단치의 비율이 높은 조건에서 1종 오류 확률은 기댓값과 다를 것이다.

Ⅳ. 연구 방법

1. 시뮬레이션 연구 설계

본 연구에서는 몬테카를로(Monte Carlo) 시뮬레이션 기법을 사용하였다. 이 방법은 수학적 이론을 바탕으로 다양한 조건에서 경험적인 기준을 찾을 때 많이 사용되는 방법론적 연구 방법이다(홍세희, 2005; Muthén & Muthén, 2002, 2005). 먼저 연구자가 모형의 모수를 결정하고 그 자료의 전집으로부터 많은 표본자료를 생성한 후 수학적 이론이 적용되는지를 확인한다.

본 연구의 목적은 표준화 잔차를 이용하여 극단치 탐지율을 알아보는 것이다. 시뮬레이션 연구 설계의 조건에는 문항이 하나의 요인에만 영향을 받는 단순한 모형, 두 개의 문항이 각각의 요인에 영향을 받는 복잡한 모형과 50, 100, 300의 표본크기, 5%, 10%, 15%, 20%의 극단치 비율로 2x3x4의 요인 설계를 구성하였다. 따라서 <표 1>과 같이 총 24개의 조건을 가진 데이터를 생성하고, 각 조건은 1000번씩 반복되었다.

1) 독립변수

(1) 분석모형

분석모형은 2요인모형으로 관찰변수 모두가 한 개의 잠재변수에 의해서 영향을 받는 단순한 모형인 <그림 10>과 변수들 중 일부가 두 개 의 잠재변수 모두에 의해서 영향을 받는 다소 복잡한 모형인 <그림 11>로 설정하였다.

(2) 표본크기(Sample Size)

Kline(2011)은 구조방정식모형을 이용한 연구에서 표본의 크기가 100이하일 때는 어떠한 구조방정식모형을 활용한 분석에서도 합리적인 결과를 도출하기 힘들며, 최소 200개를 기준으로 보고 있다. 하지만 표본크기는 연구모형, 모형의 크기, 신뢰도, 변수들 사이의 관련성, 결측치의 양 등에 영향을 받기 때문에(Muthén & Muthén, 2002), 어떤 절대적인 기준을 정할 수 없으며, 표본의 크기가 100이하인 자료를 이용해 구조방정식 연구를 수행하기도 한다(MacCallum & Austin, 2000). 따라서 본 연구에서는 작은 표본크기로 50, 중간표본 크기로 100, 큰 표본크기를 300으로 설정하였다.

(3) 극단치의 비율

극단치의 비율에 따른 식별율과 1종 오류를 알아보기 위하여 극단치의 비율을 5%, 10%, 15%, 20%로 설정하였다.

2) 종속변수

(1) 1종 오류(Type-I error)

1종 오류는 영가설이 사실임에도 불구하고 통계검정 결과에 따라 영가설을 기각하고 연구가설을 채택하는 오류이다. 본 연구에서는 극단치 사례가 아님에도 불구하고 극단치 사례로 식별하는 경우의 확률이다.

(2) 극단치 탐지율(Detection rates)

극단치 탐지율은 실제 표본이 가지고 있는 극단치의 사례 중 몇 개의 극단치 사례를 식별했는지를 보는 비율로 본 연구의 종속변수로 설정하였다.

<표 1> 시뮬레이션 연구 설계

모형	표본크기	극단치 비율% (극단치 사례수)
단순 모형	50	5(3)
		10(5)
		15(8)
		20(10)
	100	5(5)
		10(10)
		15(15)
		20(20)
	300	5(15)
		10(30)
		15(45)
		20(60)
복잡 모형	50	5(3)
		10(5)
		15(8)
		20(10)
	100	5(5)
		10(10)
		15(15)
		20(20)
	300	5(15)
		10(30)
		15(45)
		20(60)

2. 자료생성 및 분석방법

1) 자료생성

본 시뮬레이션 연구에서는, Mplus 6.12 프로그램에서 Monte Carlo 기능을 사용하여 자료를 생성하였다. 먼저 각 조건마다 다른 씨드(seed) 값을 설정하여 극단치를 포함하지 않은 1000개의 자료를 생성하였다. 앞서 생성된 자료에 극단치를 포함한 자료를 만들기 위해 조건에 맞는 극단치 비율 사례를 SAS 9.2를 사용하여 생성하였다. 극단치 사례는 마지막에 위치한 사례를 순서로 선정하였으며, 극단치 사례들의 마지막 관찰변수인 y_6에 5를 더함으로써 극단치를 포함한 자료를 생성하였다. 자료 생성시 필요한 측정오차의 공분산은 식 (2.1.1)을 사용하여 계산하였다. 각 모형에서 자료 생성시 설정한 모수치는 아래와 같다.

$$\Theta = \Sigma - \Lambda\Phi\Lambda' \qquad (2.1.1)$$

$\Sigma =$ 문항 간 분산·공분산 행렬

$\Lambda =$ 측정변수의 요인계수(factor loading) 행렬

$\Phi =$ 요인간 분산·공분산 행렬

$\Theta =$ 측정오차의 분산·공분산 행렬

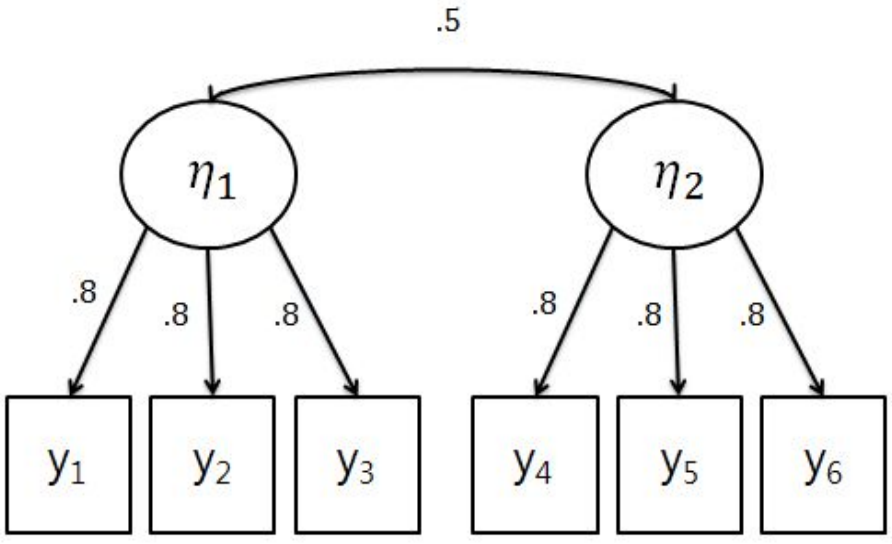
(1) 첫 번째 모형의 모수치

첫 번째 모형인 단순한 모형에 대한 모수치 행렬은 식(2.1.1.1), 식(2.1.1.2)와 같으며, 모형은 <그림 10>과 같다. 단순한 모형은 하나의 관찰치가 하나

의 요인에만 영향을 받는 모형으로 요인계수 값은 각각 .8로 설정했으며 두 요인간 상관은 .5로 설정하였다.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} .8 & 0 \\ .8 & 0 \\ .8 & 0 \\ 0 & .8 \\ 0 & .8 \\ 0 & .8 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} .36 \\ .36 \\ .36 \\ .36 \\ .36 \\ .36 \end{bmatrix} \qquad (2.1.1.1)$$

$$\sum_{\eta\eta} = \begin{bmatrix} 1 \\ .5 & 1 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad (2.1.1.2)$$



<그림 10> 첫 번째 모형의 모수치

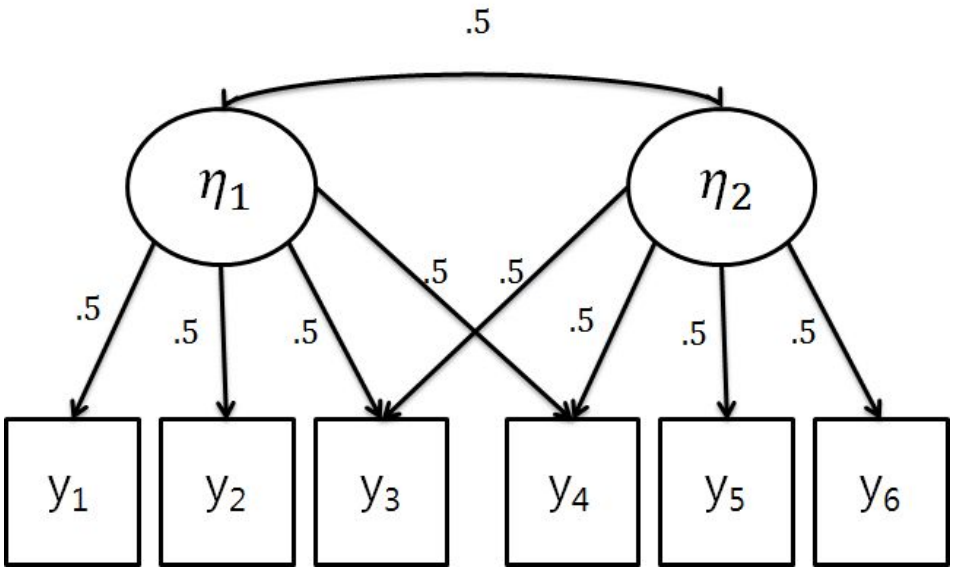
(2) 두 번째 모형의 모수치[3)]

두 번째 다소 복잡한 모형에 대한 모수치 행렬은 식(2.1.2.1), 식(2.1.2.2)와 같으며, <그림 11>과 같다. 복잡한 모형은 단순한 모형과 달리 관찰변수 y_3와y_4 모두 두 개의 요인에 의해 영향을 받는 모델로 설정되었다. 복잡한 모형에서 요인계수 값은 각각 .5로 설정되었으며 두 요인간 상관도 .5로 설정하였다.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} .5 & 0 \\ .5 & 0 \\ .5 & .5 \\ .5 & .5 \\ 0 & .5 \\ 0 & .5 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} .75 \\ .75 \\ .25 \\ .25 \\ .75 \\ .75 \end{bmatrix} \qquad (2.1.2.1)$$

$$\sum_{\eta\eta} = \begin{bmatrix} 1 & \\ .5 & 1 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad (2.1.2.2)$$

3) 두 번째 모형의 경우 오차공분산 계산시 오차공분산이 음수가 되지 않게 하는 계수로 모수치를 설정하기 위하여 요인계수를 .5로 설정하였다.



<그림 11> 두 번째 모형의 모수치

2) 분석방법

분석모형, 표본크기, 극단치의 비율에 따라 생성된 자료를 SAS IML을 이용한 행렬계산을 통해 각 사례별로 표준화된 잔차를 산출하였다. 산출된 표준화 잔차를 이용하여 χ^2값을 계산하였고, χ^2값은 각각 생성된 자료에서의 평균과 표준편차를 통해 카이스퀘어 분포를 가정하는지 검증하였다. 또한 생성된 자료의 분포를 이용하여 백분위수에 따른 임계값을 구하였다. 이 임계값을 이용하여 극단치를 식별하였으며, 각 조건에서 극단치 탐지율의 평균과 표준편차를 구하였다.

Ⅴ. 연구 결과

1. 극단치를 포함하지 않은 조건의 특성

각 조건에서 생성된 1000개 자료의 χ^2값은 <표 2>와 같다. <표 2>에서 평균은 각 조건에서 1000개 자료의 평균과 표준편차를 나타내며, 표준편차는 1000개 자료의 표준편차의 평균과 표준편차를 나타낸다. 자료의 분포가 χ^2분포를 따르는지 검증하기 위하여 χ^2의 기댓값과 분산을 비교하였다. 결과적으로 생성된 자료의 χ^2값은 χ^2의 분포를 다르지 않았다. 보다 구체적으로 χ^2분포에서 χ^2의 기댓값은 자유도와 같으며, 분산은 자유도의 두 배로, 생성된 자료에서 기대되는 χ^2의 평균과 분산, 표준편차는 각각 12와 6, 2.45이다. 하지만 결과적으로 <표 2>에서 보이는 것과 같이 생성된 자료에서 χ^2의 평균의 범위는 13.81에서 28.83으로 χ^2의 기댓값인 6보다 크게 나타났으며, 표준편차 역시 범위가 11.29에서 32.08로 χ^2의 표준편차인 2.45보다 크게 나타났다.

이는 앞서 산출했던 수학적 접근과 경험적 자료의 분포가 다르다는 것을 의미한다. 따라서 자료의 분포가 기존의 알려진 분포의 형태를 가정하지 않기 때문에 분포의 형태를 가정하지 않는 경험적 분포를 통해 얻은 임계값으로 극단치를 식별였다(Owen, 1988).

<표 2> 극단치를 포함하지 않는 조건에서 표준화 잔차의 특성

모형	표본 크기		평균(표준편차)	최소	최대
단순 모형	50	평균	16.76(4.64)	6.10	36.45
		표준편차	14.39(5.57)	4.07	37.69
	100	평균	15.06(3.07)	6.89	31.75
		표준편차	12.61(3.69)	4.43	28.45
	300	평균	13.81(1.62)	9.71	20.26
		표준편차	11.29(1.94)	6.22	19.75
복잡 모형	50	평균	28.83(12.33)	5.21	120.15
		표준편차	32.08(17.56)	3.45	132.27
	100	평균	27.39(10.17)	7.65	73.97
		표준편차	30.61(14.32)	4.75	97.84
	300	평균	26.24(6.54)	7.05	94.32
		표준편차	29.54(9.35)	4.37	128.33

2. 모형, 표본크기, 극단치의 비율, 유의 수준에 따른 극단치 탐지율

앞의 결과에서 생성된 자료가 χ^2분포를 가정하지 않음을 확인하였다. 따라서 극단치가 있는 조건에서 극단치를 식별하기 위하여 생성된 자료의 경험적 분포를 통해 99%, 95%, 90% 백분위수(percentile)에서의 임계값을 알아보았다. 이 임계값을 통하여 본 연구에서는 극단치를 포함하고 있는 사례에서 백분위수에 따른 극단치 탐지율을 연구하고자 하였다.

1) 경험적 분포에 의한 임계값

생성된 자료의 분포를 통해 얻은 99%, 95%, 90% 백분위수에 대한 임계값의 평균과 표준편차는 <표 3>과 같다. 백분위수와 표본크기가 클수록 표준편차를 통하여 값이 안정적인 것을 확인 할 수 있었다. 모형에서는 단순한 모형보다 복잡한 모형에서 더 큰 임계값을 나타냈다. 단순모형에서 백분위수가 99%, 95%, 90%일 때 임계값의 범위는 54.16에서 62.31, 35.77에서 43.55, 27.62에서 34.35로 나타났으며, 복잡모형에서는 단순모형의 임계값 보다 큰 값인 135.91에서 139.13, 84.54에서 89.65, 61.52에서 66.46으로 나타났다.

<표 3> 경험적 분포에 의한 임계값

모형	표본크기	평균(표준편차)		
		99%	95%	90%
단순 모형	50	62.31 (24.92)	43.55 (15.57)	34.35 (11.64)
	100	56.97 (17.97)	39.08 (10.65)	30.49 (7.47)
	300	54.16 (10.24)	35.77 (5.68)	27.62 (3.89)
복잡 모형	50	135.91 (76.47)	89.65 (46.50)	66.46 (32.91)
	100	134.69 (64.64)	87.16 (39.78)	63.86 (27.43)
	300	139.13 (45.87)	84.54 (25.54)	61.52 (17.41)

2) 극단치 사례를 포함한 조건에서 표준화 잔차의 특성

극단치 사례를 포함한 단순모형과 복잡모형의 특성은 <표 4>, <표 5>와 같다. 두 모형 모두 앞서 기술한 극단치를 포함하지 않은 조건의 표준화 잔차의 평균범위인 13.81에서 28.83보다 더 큰 값의 범위인 32.90에서 273.99의 값을 가지는 것을 확인 할 수 있었다. 또한 단순모형에서 극단치 비율이 5%에서 20%로 늘어날수록 표준화 잔차의 평균이 50, 100, 300의 표본크기 조건에서 50.52에서 273.99, 39.17에서 221.40, 34.99에서 216.51로 값이 커지는 경향을 보였다. 복잡모형에서 또한 극단치 비율이 5%에서 20%로 늘어날수록 표본크기의 조건에 따라 40.59에서 83.71, 35.73에서 79.95, 32.90에서 80.27로 표준화 잔차 값이 커지는 경향을 보였으나 단순모형 보다 증가하는

값의 범위는 작았다.

결과적으로 두 모형 모두 극단치 비율이 늘어날수록 표준화 잔차의 평균이 더 큰 값을 가지는 것으로 나타났으며, 모형에 따라서는 단순한 모형이 복잡한 모형보다 더 큰 표준화 잔차 값을 가지는 것으로 나타났다.

<표 4> 단순모형에서 극단치를 포함한 표본의 표준화 잔차의 특성

표본 크기	극단치 비율(%)		평균(표준편차)	최소	최대
50	5	평균	50.52(13.40)	25.73	188.35
		표준편차	108.69(36.70)	33.29	278.04
	10	평균	89.43(20.60)	44.19	211.78
		표준편차	192.52(52.32)	78.56	435.59
	15	평균	168.98(54.02)	88.02	1224.01
		표준편차	324.20(88.44)	156.10	1701.42
	20	평균	273.99(1024.80)	123.07	30151.83
		표준편차	469.10(1471.00)	213.58	44708.97
100	5	평균	39.17(6.97)	22.15	64.42
		표준편차	84.29(21.13)	38.60	176.74
	10	평균	84.69(13.57)	54.11	139.03
		표준편차	186.82(35.48)	107.05	325.80
	15	평균	145.50(20.80)	92.58	221.32
		표준편차	291.39(46.10)	184.59	455.05
	20	평균	221.40(29.42)	143.34	332.49
		표준편차	398.52(56.38)	250.35	621.11
300	5	평균	34.99(3.81)	24.02	48.67
		표준편차	76.22(11.48)	45.01	121.16
	10	평균	81.54(7.38)	58.05	106.60
		표준편차	182.03(19.12)	116.38	249.93
	15	평균	142.36(12.12)	106.45	184.63
		표준편차	287.65(26.91)	206.53	381.16
	20	평균	216.51(16.46)	166.64	276.61
		표준편차	391.77(31.63)	293.93	505.84

<표 5> 복잡모형에서 극단치를 포함한 표본의 표준화 잔차의 특성

표본 크기	극단치 비율(%)		평균(표준편차)	최소	최대
50	5	평균	40.59(28.13)	14.25	612.03
		표준편차	49.05(35.71)	10.14	778.29
	10	평균	50.41(33.33)	21.35	685.38
		표준편차	65.08(45.12)	23.71	889.85
	15	평균	68.52(30.68)	30.42	742.56
		표준편차	92.12(39.92)	43.39	973.30
	20	평균	83.71(27.63)	39.33	527.53
		표준편차	111.86(34.50)	39.33	553.25
100	5	평균	35.73(32.23)	12.73	901.07
		표준편차	42.29(46.96)	16.06	1372.23
	10	평균	45.66(15.52)	21.50	304.29
		표준편차	59.45(18.42)	27.58	322.27
	15	평균	61.31(18.25)	31.26	459.74
		표준편차	82.74(22.55)	44.61	557.53
	20	평균	79.95(25.24)	43.93	462.93
		표준편차	107.06(31.49)	63.99	614.60
300	5	평균	32.90(8.93)	13.00	195.80
		표준편차	38.27(11.47)	16.94	264.85
	10	평균	44.69(11.38)	24.30	246.54
		표준편차	57.40(13.62)	33.76	319.84
	15	평균	62.22(8.98)	39.44	113.05
		표준편차	92.23(11.32)	56.75	150.64
	20	평균	80.27(10.24)	55.12	186.47
		표준편차	116.57(12.84)	72.39	232.67

3) 단순모형에서 극단치 탐지율

단순모형에서 표본크기, 극단치의 비율, 백분위수에 따른 극단치 탐지율은 <표 6>과 같다. 단순 모형의 모든 조건에서 극단치 탐지율은 100%로 나타났다. 보다 구체적으로 생성된 자료의 분포로 얻은 임계값을 통해 극단치를 식별한 결과 표본크기, 극단치 비율, 백분위수에 상관없이 모든 조건에서 극단치 사례는 임계값보다 큰 표준화 잔차 값을 가지고 있음을 알 수 있었다.

<표 6> 단순모형에서 극단치 탐지율의 평균과 표준편차

표본크기	극단치 비율(%)		극단치 탐지율(%)		
			99%	95%	90%
50	5	평균	100	100	100
		표준편차	(0)	(0)	(0)
	10	평균	100	100	100
		표준편차	(0)	(0)	(0)
	15	평균	100	100	100
		표준편차	(0)	(0)	(0)
	20	평균	100	100	100
		표준편차	(0)	(0)	(0)
100	5	평균	100	100	100
		표준편차	(0)	(0)	(0)
	10	평균	100	100	100
		표준편차	(0)	(0)	(0)
	15	평균	100	100	100
		표준편차	(0)	(0)	(0)
	20	평균	100	100	100
		표준편차	(0)	(0)	(0)
300	5	평균	100	100	100
		표준편차	(0)	(0)	(0)
	10	평균	100	100	100
		표준편차	(0)	(0)	(0)
	15	평균	100	100	100
		표준편차	(0)	(0)	(0)
	20	평균	100	100	100
		표준편차	(0)	(0)	(0)

4) 복잡모형에서 극단치 탐지율

복잡모형에서 표본크기, 극단치의 비율, 백분위수에 따른 극단치 탐지율은 <표 7>과 같다. 복잡한 모형에서 극단치 탐지율의 범위는 35.96%에서 99.40%로 단순모형에서 극단치 탐지율인 100% 보다 낮게 나타났다. 복잡한 모형의 모든 조건에서 극단치 탐지율은 백분위수가 낮아질수록 극단치 탐지율은 높아지는 것으로 나타났다. 다시 말해, 백분위수가 99%에서 90%로 낮아질수록 10%비율의 극단치를 포함하고 있는 50, 100, 300의 각 표본크기에서 극단치 탐지율의 범위는 46.37%에서 90.30%, 35.96%에서 89.70%, 98.10%에서 98.70%로 높아졌다. 또한 백분위수가 90%일 때 극단치 탐지율이 89.70%에서 99.1%로 가장 높게 나타났다. 극단치 비율에 따라서는 극단치 비율이 20%일 때 가장 높은 탐지율을 나타냈다. 백분위수 95%에서 표본크기가 50, 100, 300일 때 극단치 비율이 5%부터 20%까지의 범위에서 극단치 탐지율이 77.13%에서 98.12%, 72.72에서 93.94%, 98.90%에서 97.55%로 높아지는 경향이 나타났다. 즉, 복잡한 모형에서는 극단치 비율이 높을수록 극단치 탐지율이 높게 나타났다.

<표 7> 복잡모형에서 극단치 탐지율의 평균과 표준편차

표본크기	극단치 비율(%)		극단치 탐지율(%)		
			99%	95%	90%
50	5	평균	46.37	77.13	90.30
		표준편차	(32.78)	(27.97)	(19.28)
	10	평균	66.58	90.14	96.58
		표준편차	(26.46)	(15.50)	(8.48)
	15	평균	84.79	84.79	99.15
		표준편차	(15.30)	(15.30)	(3.20)
	20	평균	89.99	98.12	99.40
		표준편차	(11.88)	(5.03)	(2.84)
100	5	평균	35.96	72.72	89.70
		표준편차	(25.53)	(24.27)	(16.30)
	10	평균	66.79	91.43	97.75
		표준편차	(19.48)	(10.06)	(4.78)
	15	평균	78.78	91.11	93.51
		표준편차	(11.40)	(5.13)	(3.14)
	20	평균	86.68	93.94	95.31
		표준편차	(7.68)	(3.11)	(2.15)
300	5	평균	98.10	98.09	98.07
		표준편차	(5.19)	(5.19)	(5.29)
	10	평균	61.72	90.20	95.48
		표준편차	(11.70)	(5.13)	(2.63)
	15	평균	82.99	95.92	97.59
		표준편차	(6.71)	(2.30)	(1.22)
	20	평균	90.08	97.55	98.36
		표준편차	(4.30)	(1.29)	(0.81)

5) 모형에 따른 표본크기에서 극단치 탐지율의 평균차이

모형에 따라서 표본크기가 극단치 탐지율에 가지는 효과가 차이가 있는지 알아보기 위하여 이원분산분석을 실시하였다. 표본크기와 모형에 따른 사례수와 극단치 탐지율의 평균과 표준편차는 <표8>에 보고되었다. 단순모형(M=100.00, SD=0.00)이 복잡모형(M=86.70, SD=25.09)보다 극단치 탐지율이 높은 것으로 나타났다. 표본크기에서는 표본크기가 300인 조건에서 극단치 탐지율(M=96.01, SD=6.47)이 가장 높은 것으로 나타났다.

모형과 표본크기가 극단치 탐지율에 미치는 효과에 대하여 2(단순모형 vs. 복잡모형) x 3(50 vs. 100 vs. 300)분산분석을 실시하였다. 그 결과는 <표 9>에 보고되었다. 표본크기와 모형에 따라 극단치 탐지율에 영향을 미치는 상호작용효과는 통계적으로 유의하였다(F(2, 23994)=147.77, p<.001). 따라서 각 변인들의 주효과를 검증하는 대신에 단순주효과(simple main effects)를 검증하였다. 보다 구체적으로 모형의 종류인 단순모형과 복잡모형으로 두 개의 집단을 구분하여 각각의 집단에서 표본크기가 극단치 탐지율에 미치는 효과를 검증하였다.

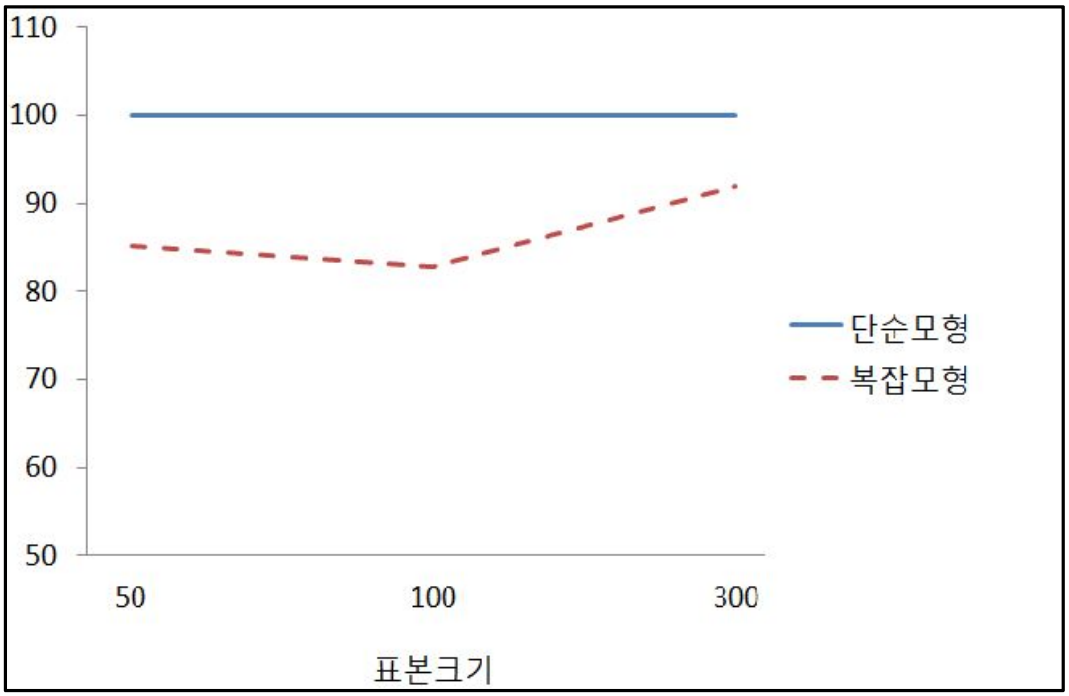
단순모형의 경우 모든 조건에서 극단치 탐지율이 100%이기 때문에 검증이 불가능 했다. 복잡한 모형에서 표본크기에 따른 극단치 탐지율의 차이는 통계적으로 유의하였다(F(2, 11997)=147.768, p=.024). 집단간의 차이를 보다 구체적으로 살펴보기 위해서 Scheffe 사후검정을 실시하였다. 그 결과, 표본크기 300이 극단치 탐지율이 가장 높았고, 다음으로 표본크기 50, 100 순으로 극단치 탐지율이 높게 나타났다. 이는 표본의 크기가 클수록 추정치의 변동이 적기 때문에 안정적인 값을 나타내게 되며 그 결과 적절한 해결책(proper solution)을 제시한다는 Chen, Bollen, Paxton, Curran과 Kirby(2001)의 연구결과와 일치한다.

<표 8> 모형과 표본크기에 따른 극단치 탐지율의 평균과 표준편차

			표본크기			합계
			50	100	300	
모형	단순모형	평균	100.00	100.00	100.00	100.00
		표준편차	0.00	0.00	0.00	0.00
		사례수	4000	4000	4000	12000
	복잡모형	평균	85.28	82.81	92.01	86.70
		표준편차	39.65	14.82	7.20	25.09
		사례수	4000	4000	4000	12000
	합계	평균	92.64	91.40	96.01	93.35
		표준편차	28.98	13.56	6.47	18.95
		사례수	8000	8000	8000	24000

<표 9> 모형과 표본크기에 따른 극단치 탐지율의 이원분산분석 결과

	제곱합	자유도	평균제곱	F	p
표본크기	1061566.54	1	1061566.54	3455.27	<.001
모형	90797.99	2	45398.99	147.77	<.001
표본크기x모형	90797.99	2	45398.99	147.77	<.001
오차	7371709.11	23994	307.23		
합계	217753031.71	24000			



<그림 12> 모형에 따른 표본크기에 대한 상호작용

6) 표본크기에 따른 극단치 비율에서 극단치 탐지율의 평균차이

표본크기에 따라서 극단치 비율이 극단치 탐지율에 가지는 효과가 차이가 있는지 알아보기 위하여 이원분산분석을 실시하였다. 극단치 비율과 표본크기에 따른 사례수와 극단치 탐지율의 평균과 표준편차는 <표 10>에 보고되었다.

표본크기가 300(M=92.64, SD=28.98)인 조건이 극단치 탐지율이 가장 높게 나타났으며, 표본크기가 50(M=91.40, SD=13.56)인 조건과 100(M=96.01, SD=6.47)인 조건 순으로 극단치 탐지율이 높게 나타났다. 극단치 비율에서는 극단치 비율이 높을수록 극단치 탐지율이 높은 것으로 나타났다. 보다 구체적으로 극단치 비율이 20%(M=97.19, SD=4.14)인 조건에서 가장 높은 극단치 탐지율을 나타냈으며 15%(M=94.92, sSD=7.22), 10%(M=92.04,

SD=31.14), 5%(*M*=89.25, *SD*=19.00)조건 순으로 극단치 탐지율이 높은 것으로 나타났다.

표본크기와 극단치 비율이 극단치 탐지율에 미치는 효과에 대하여 3(50 vs. 100 vs. 300) x 4(5 vs. 10 vs. 15 vs. 20)분산분석을 실시하였다. 그 결과는 <표 11>에 보고되었다. 극단치 비율과 표본크기에 따라 극단치 탐지율에 영향을 미치는 상호작용효과는 통계적으로 유의하였다(F(6, 23988)=106.24, p<.001). 따라서 각 변인들의 주효과를 검증하는 대신에 단순주효과(simple main effects)를 검증하였다. 보다 구체적으로 표본크기의 종류인 50, 100, 300으로 세 개의 집단을 구분하여 각각의 집단에서 극단치 비율이 극단치 탐지율에 미치는 효과를 검증하였다.

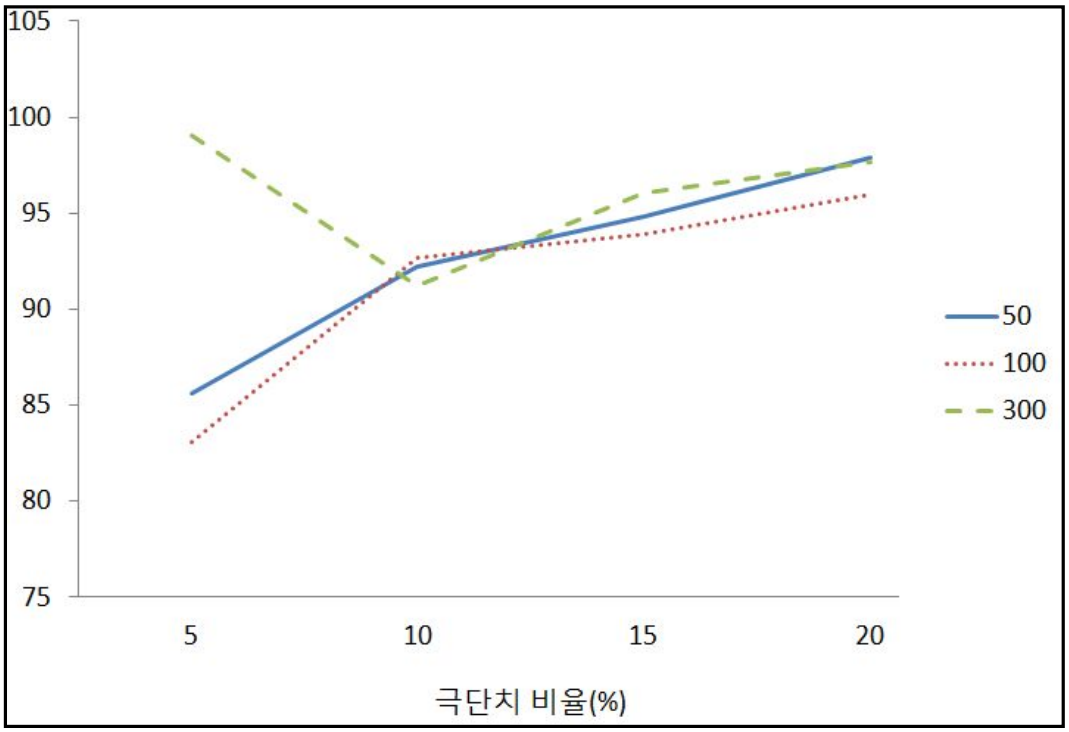
집단간의 차이를 보다 구체적으로 살펴보기 위해서 Scheffe 사후검정을 실시하였다. 그 결과, 표본크기가 50과 100인 조건에서 모든 극단치 비율의 조건에서 극단치 탐지율에 차이가 있는 것으로 나타났으며, 극단치 비율이 높을수록 극단치 탐지율이 높은 것으로 나타났다. 반면에 표본크기가 300인 조건에서만 극단치 비율이 5%일 때 극단치 탐지율이 가장 높은 것으로 나타났으며, 다음으로 20%, 15%, 10% 순으로 극단치 탐지율이 높은 것으로 나타났다. 이는 사례수가 300인 조건에서 극단치 비율이 낮을수록 추정치가 안정된 값을 나타내기 때문인 것으로 해석된다(Chen et al., 2001)

<표 10> 모형과 표본크기에 따른 극단치 탐지율의 평균과 표준편차

			극단치 비율(%)				합계
			5	10	15	20	
표본크기	50	평균	85.63	92.22	94.79	97.92	92.64
		표준편차	21.49	52.11	9.09	4.47	28.98
		사례수	2000	2000	2000	2000	8000
	100	평균	83.06	92.66	93.90	95.99	91.40
		표준편차	21.46	10.12	7.21	4.71	13.56
		사례수	2000	2000	2000	2000	8000
	300	평균	99.04	91.23	96.08	97.67	96.01
		표준편차	3.80	9.59	4.40	2.65	6.47
		사례수	2000	2000	2000	2000	8000
	합계	평균	89.25	92.04	94.92	97.19	93.35
		표준편차	19.00	31.14	7.22	4.14	18.95
		사례수	6000	6000	6000	6000	24000

<표 11> 표본크기와 극단치 비율에 따른 극단치 탐지율의 이원분산분석 결과

	제곱합	자유도	평균제곱	F	p
극단치 비율	90797.99	2	45398.99	134.54	<.001
표본크기	214719.49	3	71573.16	212.11	<.001
극단치 비율x표본크기	215095.62	6	35849.27	106.24	<.001
오차	8094258.53	23988	337.43		
합계	217753031.71	24000			



<그림 13> 극단치 비율에 따른 표본크기에 대한 결과

Ⅵ. 논의 및 제언

1. 연구 결과에 대한 논의

본 연구는 구조방정식 모형에서 측정모형에 해당하는 확인적 요인모형에서 극단치 식별의 영향에 대하여 살펴보았다. 먼저 극단치 식별을 위하여 요인모형에서 극단치 식별에 사용할 수 있는 표준화 잔차의 이론적 및 경험적 분포를 검증하였다. 이러한 목적을 달성하기 위해서 요인모형에서 표준화 잔차의 이론적 분포가 χ^2분포와 다른지 증명하였고, 표준화 잔차의 경험적 분포의 평균 및 표준편차와 이론적 분포의 평균 및 표준편차의 비교를 통해서 표준화 잔차가 χ^2분포를 가정하는지 확인하였다.

연구 결과 생성된 자료로부터 산출된 표준화 잔차의 경험적분포는 χ^2분포를 따르지 않았다. 따라서 극단치 식별을 위해 χ^2분포가 아닌 생성된 자료의 경험적 분포를 통하여 백분위수에 따른 임계값을 구하였다. 이 임계값을 이용하여 극단치를 포함한 표본에서 분석모형, 표본크기, 극단치 비율의 다양한 조건하에서 표준화 잔차의 극단치 탐지율을 비교함으로써 표준화 잔차의 유용성을 검증하였다. 주요 연구결과를 요약하면 다음과 같다.

첫째, 생성된 자료는 χ^2분포를 가정하지 않는 것으로 나타났다. 즉, 연구문제 1에 대한 가설은 지지되지 않았다. 이는 표준화 잔차를 이용한 χ^2값을 산출할 때 예측치가 아닌 모수치를 사용했기 때문에 계산되어야 하는 값보다 과소 추정되거나 과대 추정된 것으로 보인다. 따라서 생성된 자료를 통해 유의수준에 따른 임계값을 구하여 극단치 식별에 이용하였다.

둘째, 표준화 잔차의 평균은 극단치를 포함하지 않은 조건보다 극단치를 포함한 조건에서 더 크게 나타났다. 또한 각 사례들에서 극단치를 포함한

사례들의 표준화 잔차는 극단치를 포함하지 않은 사례들 보다 큰 값을 나타냈다. 이는 극단치가 평균에 영향을 주게 되고 따라서 극단치 비율이 클수록 표준화 잔차 역시 커지는 것을 확인할 수 있었다.

셋째, 극단치 탐지율은 복잡모형보다 단순모형에서 더 높게 나타났다. 보다 구체적으로 단순모형에서 탐지율은 모든 조건에서 100%를 나타낸 반면 복잡모형은 극단치 비율이 높을수록, 표본크기가 클수록 극단치 탐지율이 높은 것으로 나타났다. 따라서 가설 2-1과 3-1, 4-1이 지지되었다.

넷째, 모형과 표본크기의 상호작용효과를 분석한 결과 통계적으로 유의한 것으로 나타났다. 보다 구체적으로 모형에 따른 표본크기가 극단치 탐지율에 영향을 미치는 것으로 나타났다. 복잡모형에서 표본크기가 300인 조건에서 극단치 탐지율이 가장 높게 나타났으며 표본크기가 100인 조건에서 극단치 탐지율이 가장 낮은 것으로 나타났다.

다섯째, 표본크기와 극단치 비율의 상호작용효과 역시 통계적으로 유의하게 나타났다. 표본크기에 따른 극단치 비율에서는 극단치 비율이 클수록 극단치 탐지율이 높게 나타났다.

본 연구에서는 표준화 잔차를 이용하여 극단치를 식별하고자 하였다. 결과적으로 모형이 간명하고 충분한 표본크기가 극단치를 식별하는데 도움을 주는 것으로 나타났다. 분석모형이 복잡한 경우 표본이 극단치를 포함하게 되면 오차 공분산이 음수로 추정되게 되고 이는 요인점수가 수렴할 수 없게 되거나 잘못된 추정결과를 나타내게 된다. 따라서 복잡한 모형과 같은 다차원적 척도보다는 단순한 모형과 같은 단일차원의 척도를 사용하는 것을 권한다(Clark & Watson, 1995). 또한 표본크기가 클수록 모집단을 더 잘 대표하기 때문에 추정치가 모수치에 더 가까운 값을 추정하게 되며 분산이 작아지게 된다. 따라서 모수치와 다른 값인 극단치의 탐지율이 높게 나타나게 된다.

2. 연구의 제한점 및 후속 연구를 위한 제언

본 연구의 제한점은 다음과 같다. 첫째, 본 연구는 오차 분산의 모수치를 알고 실시된 시뮬레이션연구이다. 따라서 모형의 표준화 잔차를 계산하는데 모수치의 오차 공분산을 사용하였다. 하지만 실제 연구에서는 오차공분산의 모수치를 알 수 없기 때문에 이 연구를 일반화하기에는 어렵다. 따라서 표준화 잔차를 계산하는데 있어 모수치의 오차분산이 아닌 추정된 오차분산을 이용한 연구가 추가적으로 필요하다.

둘째, 확인적 요인 모형은 구조방정식의 측정모형으로 구조방정식 모형보다 단순한 모형이다. 위의 결과에서도 언급했듯이 모형이 복잡해지면 오차분산이 음수로 추정되게 되며 이는 요인점수를 추정할 수 없거나 잘못된 값을 추정하게 된다. 또한 모형이 복잡해질수록 극단치의 탐지율이 낮아진다. 따라서 보다 복잡한 구조방정식 모형을 활용하여 표본크기, 다양한 구조방정식 모형의 조건에서 극단치 식별을 위한 적절한 유의수준 연구가 추가적으로 필요하다.

본 연구에서는 표준화 잔차를 이용하여 극단치를 식별하고자 하였다. 앞서 기술한 바와 같이 극단치는 모형의 모수추정에 편향을 발생 시켜 결과를 왜곡 시킬 수 있다. 따라서 모형에 따른 적절한 극단치의 식별 방법이 필요하며, 이는 더 분명한 연구결과를 얻는데 도움을 줄 것이다.

참 고 문 헌

김계수 (2010). Amos 18.0 구조방정식모형 분석. 서울: 한나래.

김주환, 김민규, 홍세희 (2009). 구조방정식모형으로 논문쓰기. 서울: 커뮤니케이션북스.

문수백 (2009). 구조방정식모형링의 이해와 적용. 서울: 학지사.

배병렬 (2009). Amos 17.0 구조방정식모형링: 원리와 실제. 서울: 청람.

성태제 (2011). 현대 기초 통계학의 이해와 적용. 학지사.

신형원, 손소영 (2001). 군집기반 유전자 알고리즘을 이용한 다중 이상치 검출 연구. 대한산업공학회 추계학술대회 논문집, 135-139.

안병진, 서한손 (2011). 동적 그림을 이용한 이상치 검색. 응용통계연구, 24(5), 979-986.

이수연. (2010). 대학생의 행복 척도 개발 및 구인타당도 검증. 한국심리학회지: 학교, 7(2), 107-122.

이순묵 (2000). 요인분석의 기초. 서울: 교육과학사.

이영준 (2002). 요인분석의 이해. 서울: 도서출판 석정.

이충헌 (2008). 비정규성 자료에서의 이상치 검출을 위한 통계적인 방법. 공주대학교 대학원 석사학위논문.

이희연, 노승철 (2013). 고급통계분석론-이론과 실습-제2판. 서울: 문우사.

홍세희 (2000). 특별기고: 구조 방정식 모형의 적합도 지수 선정기준과 그 근거. 한국심리학회지: 임상, 19(1), 161-177.

홍세희 (2003). 구조 방정식 모형의 원리와 응용. 이화여자대학교 경영연구소 추계학술심포지움 발표 논문, 서울: 이화여자대학교.

홍세희 (2005). Wijsman 의 알고리즘을 이용한 표본 상관/공변량 행렬 생

성방법. 사회과학연구논총, 14(단일호), 111-119.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 00(0)*, 1-32.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological methodology, 21*, 235-262.

Chatterjee, S., & Yilmaz, M. (1992). A review of regression diagnostics for behavioral research. *Applied Psychological Measurement, 16(3)*, 209-227.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper Solutions in Structural Equation Models Causes, Consequences, and Strategies. *Sociological Methods &Research, 29(4)*, 468-508.

Clark, L. A., &Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological assessment, 7(3)*, 309.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* Guilford Press.

Maimon, O. Z., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook (Vol. 1).* New York: Springer.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9(4)*, 599-620.

Muthén, L. K., &Muthén, B. O. (2005). *Mplus: statistical analysis with latent variables; user's guide;[Version 3].* Muthén & Muthén.

Nicola, B., Kemp, R., & Snelgar, R. (2013). SPSS를 활용한 심리연구 분석 [SPSS for Psychologists]. (이주일, 문혜진, 정현선, 조영일, 최윤영, 한태영 역). 서울: 시그마프레스. (원전은 2012에 출판).

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika, 75(2),* 237-249.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2007). 데이터 마이닝 [Introduction to data mining]. (용환승, 나연묵, 박종수, 승현우, 이민수, 이상준, 최린 역). 서울: 인피니트북스. (원전은 2006에 출판).

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589). John Wiley & Sons.

Williams, G., Baxter, R., He, H., Hawkins, S., & Gu, L. (2002, December). A comparative study of RNN for outlier detection in data mining. In *2013 IEEE 13th International Conference on Data Mining* (pp. 709-709).

ABSTRACT

Outlier Detection Using Standardized Residuals in Confirmatory Factor Analysis

Jin-Yi Shin
Department of Psychology
The Graduate school of
Sungshin University

This study was to find empirical evidence for outlier detection of the confirmatory factor model for the measurement model in structural equation model. First, in the factor model, it was proved that the theoretical distribution of the standardized residuals is different from chi-square distribution. The standardized residuals confirmed the chi-square distribution assumption through mean and standard deviation. In result the chi-square value calculated from the generating data did not assume the distribution. Thus, to detect outlier, I obtained the critical values from distribution of generating data according to 99%, 95%, 90% percentiles. Using these critical values in a sample which contains outlier, the sample was tested under various conditions like analysis model, sample size, and outlier rate. By comparing the detection rate, usefulness of standardized residuals was proved. The results are as follows:

First, the chi-square value calculated from the generating data did not assume the distribution. Thus, I was able to obtain the critical values from the distribution of generating normal data. This critical value was then used to detect the outlier. Second, the simple analysis model, the data showed that having a larger sample size and high outlier rate produced a higher detection rate. Third, the analysis model and the sample size caused an interaction effect and the effect was statistically significant. Also, an interaction effect between the sample size and the outlier ration was statistically significant.

Finally, we discuss the limitations of this study and suggested future study.

Key Words: Outliers, Simulation, Outlier detection, Standardized residuals.

부 록

<부록 1> 자료 생성을 위한 MPLUS 명령문(ver. 6.12)

<부록 2> 요인점수 계산과 저장을 위한 MPLUS 명령문 (ver.6.12)

<부록 3> 극단치 생성을 위한 SAS 명령문(ver. 9.2)

<부록 4> x^2 계산을 위한 SAS 명령문(ver. 9.2)

<부록 5> x^2값의 평균과 표준편차를 구하기 위한 SAS 명령문(ver 9.2)

<부록 1> 자료 생성을 위한 MPLUS 명령문(ver. 6.12)

<간단한 모형>

```
TITLE: simple model;

MONTECARLO:
            NAMES ARE Y1-Y6;      ! 예측변수
            NOBSERVATIONS = 50;   ! 표본크기
            NREPS = 1000;         ! 반복 횟수
            SEED = 24336;

        REPSAVE=ALL;
        SAVE=DATA(50_5)*.DAT;     ! 저장파일명
        RESULTS=TEMP.SAV;

 MODEL POPULATION:                ! 모형의 모수치 설정
          F1 BY Y1@.8             ! 요인계수 값
                Y2@.8
                Y3@.8;
          F2 BY Y4@.8
                Y5@.8
                Y6@.8;

                Y1@.36;           ! 오차공분산 값
                Y2@.36;
                Y3@.36;
                Y4@.36;
                Y5@.36;
                Y6@.36;

          F1 WITH F2 @.5;         ! 요인간 상관

          F1@1;                   ! 요인의 표준편차
          F2@1;
```

```
        [F1@0];             ! 요인의 평균
        [F2@0];

MODEL:                      ! 설정된 모형
    F1 BY Y1 Y2 Y3;
    F2 BY Y4 Y5 Y6;
    F1 WITH F2;

OUTPUT: TECH9;              ! 표본생성시 잘못된 표본에 대한 설명을 나타냄.
```

<복잡한 모형>

```
TITLE: complex model;

MONTECARLO:
          NAMES ARE Y1-Y6;
          NOBSERVATIONS = 50;
          NREPS = 1000;
          SEED = 11412;

       REPSAVE=ALL;
       SAVE=DATA(50_5)*.DAT;
       RESULTS=TEMP.SAV;

 MODEL POPULATION:
          F1 BY Y1@.5
                Y2@.5
                Y3@.5
                Y4@.5;
          F2 BY Y3@.5
                Y4@.5
                Y5@.5
                Y6@.5;

          Y1@.75;
          Y2@.75;
```

```
        Y3@.25;
        Y4@.25;
        Y5@.75;
        Y6@.75;

        F1 WITH F2 @.5;
        F1@1;
        F2@1;
        [F1@0];
        [F2@0];

MODEL:
    F1 BY Y1 Y2 Y3 Y4;
    F2 BY y3 Y4 Y5 Y6;
    F1 WITH F2;

OUTPUT: TECH9 ;
```

<부록2> 요인점수 계산과 저장을 위한 MPLUS 명령문(ver.6.12)

```
TITLE: 요인점수 저장;

DATA: FILE IS temp_data.dat;                    ! 파일명

VARIABLE: NAMES ARE Y1-Y6;                      ! 변수명

MODEL: F1 BY Y1-Y3;                             ! 설정된 모형
       F2 BY Y4-Y6;
       F1 with F2;

OUTPUT: STANDARDIZED SAMPSTAT;

SAVEDATA: FILE IS temp_result.sav;              ! 저장할 파일명
SAVE = FSCORES;                                 ! 추정된 요인점수만 저장
```

<부록3> 극단치 생성을 위한 SAS 명령문(ver. 9.2)

```
%macro data;
%do i=1012 %to 1045;
data sim_50_5_&i;
infile "D:\data\300\20\data(50_5)&i..dat";
input y1-y6;                                    /*변수명*/
id=_n_;                                         /*아이디 생성*/

if id=48 then y6=y6+5;                          /*극단치 생성*/
if id=49 then y6=y6+5;
if id=50 then y6=y6+5;

drop id;                                        /*아이디 지우기*/
run;

DATA _NULL_;
SET sim_50_5_&i;
FILE  "D:\data\300\20\data(50_5)&i..dat" DSD DROPOVER DLM = ',';
PUT (_all_) (~);
RUN;

%end;
%mend;
%data;
quit;
```

<부록4> x^2 계산을 위한 SAS 명령문(ver. 9.2)

```
options pagesize=60 linesize=80 pageno=1 nodate;
dm log 'clear' output;

%macro data;
%do i=1 %to 1000;
data sim_100_5_&i;
infile "D:\data\300\20\data(50_5)&i..dat"  DSD  DLM = ',';
input y1-y6;
run;

 /*공분산 계산*/
proc corr data=sim_100_5_&i cov out=cov_1 noprint;
var y1-y6;
DATA DCOV; SET COV_1;
IF _TYPE_ = 'COV';
PUT (Y1-Y6) (7. 3);
RUN;

data fac_sc_100_5_&i;
infile "D:\data\300\20\data(50_5)&i...sav";
input fy1-fy6 f1 f1_re f2 f2_re;
run;

 /*잔차 계산*/
proc iml;
/*residual_hat */
use  fac_sc_100_5_&i;
read all var {"f1" "f2"} into eta;

eta_t = t(eta);

use sim_100_5_&i;
```

```
read all var  {"y1", "y2", "y3", "y4", "y5", "y6"} into obs;

lam = {.8 0, .8 0, .8 0, 0 .8, 0 .8, 0 .8};

temp = lam*eta_t;
temp_t = t(temp);
resi_hat = obs - temp_t;

use dcov;
read all into cov;

covi=inv(cov);

vals = {.36, .36, .36, .36, .36, .36};
theta = diag (vals);

var=theta*covi*theta;
vari=inv(var)
chi_total=j(50,1);

do i=1 to 50;

res_ind= resi_hat[i,];

/*chi-square */
resi_t=t(res_ind);
chi=res_ind*vari*resi_t;

chi_total[i,1]=chi[1,1];

end;
nNames = "c1";

create chi_total&i from chi_total[colname=nNames];
append from chi_total;
```

```
run;

proc append base = temp data = chi_total&i;
run;

%end;
%mend;
%data;
quit;

proc print data = temp; run;
```

<부록5> x^2값의 평균과 표준편차를 구하기 위한 SAS 명령문(ver 9.2)

```
data temp1;
  set temp;
  group = int((_n_ - 1)/50)+1;
run;

proc sort data=temp1;
by group;
run;

/*평균과 표준편차계산*/
proc means data=temp1;
class group;
var c1;
OUTPUT OUT=sim100 MEAN=totalMEAN STDDEV=totalSD;
run;

data sim101;
set sim100;

if  group='.' then delete;
run;
proc print data=sim101;
run;
proc means data=sim101;
var totalMEAN totalSD;
run;
quit;
```