패널자료 품질개선 연구(Ⅳ)

www.kli.re.kr

홍민기 · 한치록 · 김재광
신동균 · 김기민 · 이고은

KLI
한국노동연구원

목 차

요 약 ····· i

제1장 서 론 ····· (홍민기) ····· 1

제2장 동태적 패널모형에서 표본이탈 교정
····· (한치록 · 이고은) ····· 3

제1절 서 론 ····· 3

제2절 동태적 패널모형의 추정 ····· 4

제3절 균형패널에서 효율성 제고 방안 ····· 8

1. 도 입 ····· 8

2. FD/IV와 FOD/IV의 결합 ····· 12

3. 모의실험 ····· 13

제4절 표본이탈이 있는 동태적 패널모형의 추정 ····· 15

1. 표본이탈 존재 시 비균형패널이나 균형화된 부분표본의 분석 ····· 15

2. FD/IV 추정시 내생적 표본이탈 편향의 교정 ····· 21

3. FOD/IV ····· 22

4. Full GMM ····· 27

제5절 사업체 패널의 분석 ····· 31

1. 표본이탈을 고려하지 않음 ····· 33

2. 표본이탈을 고려함 ····· 35

제6절 소 결 ····· 38

[부록] Stata 실행 결과 ····· 40

제3장 Paradata를 이용한 무응답 자료 회귀분석
…… (김재광 · 김기민) …… 56
제1절 서 론 …… 56
제2절 제안된 방법론 …… 59
제3절 모의실험 …… 63
제4절 실증 분석 …… 64
제5절 결론 및 향후 과제 …… 68

제4장 사업체 패널조사의 표본 설계 관련 연구 …… (김재광) …… 69
제1절 서 론 …… 69
제2절 사업체 조사와 관련된 기초 연구 …… 70
제3절 사업체 패널조사 설계방안 : 방안 A …… 73
1. 조사 모집단 결정 …… 73
2. 층화 및 표본 추출 …… 75
제4절 사업체 패널조사 설계방안 : 방안 B …… 76
제5절 패널 마모 처리 …… 79
제6절 소 결 …… 80

제5장 한국에서의 자료검증 연구의 필요성에 대하여
…… (신동균) …… 81
제1절 서 론 …… 81
제2절 서베이 자료에 나타난 측정오차 사례 …… 84
제3절 측정오차를 무시했을 때 발생할 추정결과상 편의
예시 …… 95
제4절 자료검증 연구결과 활용 사례 …… 105
제5절 소 결 …… 109

제6장 결론 : 요약 및 시사점 ····· (홍민기) ····· 112
제1절 동태적 패널모형에서 표본이탈 교정 ····· 112
제2절 Paradata를 이용한 무응답 자료 회귀분석 ····· 113
제3절 사업체 패널조사의 표본 설계 관련 연구 ····· 114
제4절 한국에서의 자료검증 연구의 필요성에 대하여 ····· 115

참고문헌 ····· 117

표목차

<표 2- 1> 표본이탈이 없을 때 $n=100$, $T=10$, 1,000회 반복 ····· 14
<표 2- 2> 엄밀히 외생적인 표본이탈 시 전체 자료를 이용한 추정 ····· 17
<표 2- 3> 약하게 외생적인 표본이탈 시 전체 자료를 이용한 추정 ····· 18
<표 2- 4> 내생적인 표본이탈 시 전체 자료를 이용한 추정 ····· 18
<표 2- 5> 외생적인 표본이탈 시 균형화한 부분자료를 이용한 추정 ····· 19
<표 2- 6> 약하게 외생적인 표본이탈 시 균형화한 부분자료를 이용한 추정 ····· 20
<표 2- 7> 내생적인 표본이탈 시 균형화한 부분자료를 이용한 추정 ····· 20
<표 2- 8> 표본이탈이 엄밀하게 외생적인 경우 ····· 23
<표 2- 9> 표본이탈이 약하게 외생적인 경우 ····· 24
<표 2-10> 표본이탈이 내생적(1기 전의 종속변수에 의존)인 경우 ··· 25
<표 2-11> 표본이탈이 엄밀히 외생적인 경우(FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용함) ····· 28
<표 2-12> 표본이탈이 약하게 외생적인 경우(FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용함) ····· 29
<표 2-13> 표본이탈이 내생적인 경우(FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용함) ····· 29
<표 2-14> 표본이탈이 엄밀히 외생적인 경우(Full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용함) ····· 30
<표 2-15> 표본이탈이 약하게 외생적인 경우(Full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용함) ····· 30

<표 2-16> 표본이탈이 내생적인 경우(Full GMM의 도구변수를
표본관측 방정식의 독립변수로 사용함) ····· 31
<표 2-17> 시간에 따른 응답자 수 ····· 32
<표 2-18> 시간에 따른 표본이탈 이유 ····· 33
<표 2-19> 표본이탈을 고려하지 않은 FD/IV 추정결과 ····· 34
<표 2-20> 표본이탈을 고려하지 않은 full GMM 추정결과 ····· 34
<표 2-21> FD/IV의 도구변수를 표본관측 방정식의 독립변수로
사용 ····· 36
<표 2-22> Full GMM의 도구변수를 표본관측 방정식의
독립변수로 사용 ····· 37
<표 2-23> 계수값 비교 ····· 38

<표 3- 1> Monte Carlo bias and standard error of ····· 64
<표 3- 2> 첫 컨택 반응에 따른 연도별 1인당 매출액 응답여부 ····· 65
<표 3- 3> 사업체 패널자료에 네 가지 방법을 적용한 분석결과 ····· 67

<표 4- 1> 산업 구분 및 사업장 규모별 모집단 사업장 수 현황 ····· 74

<표 5- 1> 산업(업종)변수의 측정오차 ····· 90
<표 5- 2> 직종변수의 측정오차 ····· 91

그림목차

[그림 3- 1] 각 방법들의 추정된 결과에 따른 log(1인당 매출액)
분포 ····· 67

요 약

◈ 동태적 패널모형에서 표본이탈 교정

본 연구에서는 홍민기 등(2014)이 고려한 표본이탈의 문제를 동태적 패널모형으로 확장하여 분석한다. 특히 표본이탈이 존재할 때 동태적 패널모형의 효율적인 추정방법에 대해 논의한다. 이를 위해 모의실험을 실시하고 그 결과를 토대로 사업체 패널에 적용한다.

동태적 패널모형이란 종속변수의 과거값을 설명변수로 사용하는 패널자료 모형을 의미한다. 표본이탈이 존재하지 않는 일반적인 상황에서는 고정효과를 1계차분(FD)하는 'Arellano-Bond(1991)의 full GMM' 또는 'FD/IV'를 이용해 모수를 추정한다. Full GMM은 매기마다 사용 가능한 도구변수를 모두 사용하는 반면, FD/IV는 매기마다 동일한 개수의 도구변수를 사용한다. 즉 full GMM이 FD/IV보다 더 많은 도구변수를 사용하게 되는데, 이로 인해 full GMM은 FD/IV보다 더 작은 분산을 갖고 더 큰 편향을 갖는다. 본고의 제2절에서 표본이탈이 없을 때 full GMM과 FD/IV의 차이점에 대해 더 자세히 설명하고 제3절에서는 모의실험을 통해 두 추정방법의 편향과 분산을 비교한다.

한편 동태적 패널모형의 고정효과를 FOD(Forward Orthogonal Deviations, Arellano and Bover, 1995)로 제거하게 되면 오차항이 동분산을 갖고 시간에 걸쳐 비상관되어 차분한 경우보다 더 효율적일 것으로 추측할 수 있다. 즉 'FOD/IV'를 사용하면 FD/IV보다 더 작은 분산을 갖게 될 것으로 예상할 수 있다. 그러나 모의실험 결과에 따르면 FOD/IV의 분산이 FD/IV의 분산보다 더 크게 도출되었으며, 이 결과를 토대로 편향이 작으면서 분산이 작은 'FD-FOD/GMM'을 제3

절에서 제시하고 있다.

자료에 표본이탈이 존재하게 되면 이상의 방법들을 적용할 수 있는지 살펴보아야 한다. 만약 표본이탈이 엄밀히 외생적으로 발생하면 비균형패널 또는 균형화한 부분자료를 이용하여 이상의 방법들을 사용할 수 있다. 특히 비균형패널을 이용할 때 표본이탈이 약하게 외생적으로 발생하면 추정량이 일관적으로 추정되어야 하는데, FOD 변환을 사용한 추정법은 FOD가 미래의 표본이탈 여부에 의존하여 비일관적으로 추정된다. 따라서 표본이탈이 존재할 때 FOD 변환을 사용한 추정법은 사용할 수 없다. 4.1절에서는 비균형패널 또는 균형화한 부분자료를 이용할 수 있는 경우에 대해 살펴보고, 4.3절에서는 FOD 변환의 추정법에 대해 논의한다.

표본이탈이 내생적으로 발생하면 이로 인한 편향을 교정하여 사용해야 한다. Wooldridge(2002; 2010)는 동태적 패널모형에 Heckman (1976)의 2단계 추정법을 적용해 편향교정항(Inverse Mills Ratio : IMR)으로 편향을 교정한다. 본고에서는 'FD/IV with IMR'과 'full GMM with IMR'을 제시하고 있는데, 전자는 FD/IV의 도구변수를 편향교정항의 독립변수로 사용하고, 후자는 full GMM의 도구변수를 편향교정항의 독립변수로 사용한다. 표본이탈이 없을 때와 마찬가지로 내생적 표본이탈이 존재할 때 'full GMM with IMR'의 분산이 'FD/IV with IMR'의 분산보다 더 작다. 편향교정항의 독립변수에 따른 두 추정방법의 비교는 4.4절에 제시되어 있다. 마지막으로 제4절의 내용을 토대로 사업체 패널의 일부에 적용한 결과가 제5절에 제시되어 있다.

◈ Paradata를 이용한 무응답 자료 회귀분석

무응답 자료를 이용한 회귀분석에는 무응답 메커니즘에 대한 이해가 필요하다. 일반적으로 무응답 성향과 관련 있는 변수를 아는 경우에는 그 관련 변수를 회귀분석의 설명변수에 포함시켜서 분석하

는 것이 무응답에 따른 선택 편향을 줄여주는 것으로 알려져 있다. 사업체 패널조사에서는 면접 당시의 '첫 컨택 반응'이라는 변수가 일종의 paradata로서 얻어진다. 이 '첫 컨택 반응'변수는 응답 성향에 대한 좋은 정보를 제공한다. 본 연구에서는 이러한 paradata를 사용하여 무응답 편향을 줄이는 회귀분석 방법에 대하여 다루었다.

제안된 방법은 이 '첫 컨택 반응'변수(Z)를 분석하고자 하는 회귀모형의 설명변수에 추가하여 확장된 회귀모형(augmented outcome regression model)을 통해 응답 자료만을 사용하여 회귀계수를 추정한 후 추가적으로 '첫 컨택 반응'변수를 기존 설명변수(X)로 설명하는 회귀모형을 별도로 세워서 전체 자료를 바탕으로 모형을 적합시켜 Z에 대한 예측값을 X의 함수로 표현한 후 이를 확장된 회귀모형의 Z에 대입해 줌으로써 최종적으로는 Y에 대한 X의 회귀모형을 얻어내는 것이다.

제안된 방법론은 모의실험을 통해서 기존의 방법론과 비교하였고 무응답 메커니즘이 X뿐만 아니라 Z에 의존하는 경우에도 효율적인 추정량을 제공하였다. 또한 기존의 방법론인 확장된 무응답 성향 모형을 사용한 가중치 조정 방법론보다 더 효율적인 추정을 구현하는 것을 확인하였다. 또한 제안된 방법론은 사업체 패널조사에 적용하였다. 고려한 회귀모형은 1인당 매출액을 종속변수로 하고 사업체 규모와 산업을 설명변수로 하는 모형이었는데 그 모형에 추가적으로 '첫 컨택 반응'변수를 넣어서 제안된 방법론을 적용하였다. 적용 결과, 본 연구에서 제안된 방법은 다른 방법들에 비해 추정치 분산의 추정량이 다소 작게 나왔다.

◈ 사업체 패널조사의 표본 설계 관련 연구

본 보고서에서는 기존의 사업체 패널조사에 나타난 몇 가지 문제점을 바탕으로 이를 해결하기 위한 새로운 방안으로서 새로운 패널 표본 설계 또는 기존 패널을 보완하는 표본 재설계를 할 때 유의해

야 할 사항들을 정리하였다. 이를 위하여 사업체 패널을 새롭게 설계하고자 할 때 고려해야 할 사항과 기존 패널에 대한 재설계와 관련된 사항 두 가지로 나누어서 다루었다.

먼저 새로운 패널 설계에서는 조사 모집단의 결정과 관련된 내용과 층화와 관련된 내용을 나누어서 설명하였다. 사업체 패널조사는 많은 조사항목을 가지는 다목적 조사이며, 이 조사를 통해 산업별, 사업장 규모별, 지역별 통계 생산이 가능하도록 하는 것을 기본 원칙으로 설계되었다. 새로운 표본 설계 방안도 원칙면에서는 이와 동일하며 이를 위하여 산업별 분류, 사업장 규모 및 지역을 층화변수로 하는 층화 추출을 사용하고자 하는 것이 동일하다. 층화 추출은 거의 대부분의 사업체 조사에서 실시하고 있는 표본 추출방법이다. 층화 추출에서 표본 배정은 여러 가지 제한조건들을 만족하면서 전체 추정량의 분산을 최소화하는 최적화 문제의 해로써 얻어지는데 이를 위해서는 mathematical programming이 사용된다.

사업체 패널조사 설계의 또 다른 방안으로는 이전의 조사에서 응답해오던 기존 표본은 그대로 두고 나머지 모집단에서 표본을 추가하는 추가 표본 설계를 고려할 수 있다. 패널조사에서 표본 탈락 또는 패널 마모(panel attrition)는 흔히 발생하는 현상으로 적절한 시점에서 계속 표본 추가를 해주어서 적정 표본 수를 유지하고 모집단에 새롭게 진입한 신규 사업체들을 포함함으로써 전체적인 횡단면적 대표성을 제고하는 효과를 가지게 된다. 추가 표본 설계에서는 층화추출을 기본으로 하되 기존의 표본을 포함한 상태에서 새로운 제한조건을 가지는 최적화 문제로 풀어서 표본 배정을 실시하고 그로부터 신규 표본을 추출하면 된다.

◈ 한국에서의 자료검증 연구의 필요성에 대하여

정보 수집의 용이성, 그리고 수집된 정보의 풍부함 등의 이유로 많은 사회과학 분야의 연구들은 조사(survey)를 통하여 획득된 자료

에 근거하여 수행되고 있다. 본 연구의 목적은 그러한 자료들이 얼마나 심각한 측정오차 문제를 안고 있는가, 그리고 이러한 조사자료에 존재하는 측정오차가 분석결과를 얼마나 왜곡시킬 수 있는가를 예시하고 이를 통하여 보다 본격적인 조사자료 검증연구의 필요성을 역설하는 데에 있다. 조사자료의 정확성을 판단할 수 있는 검증자료(validation data)에 대한 접근이 용이하지 않은 상태에서 본 연구에서는 KLIPS 자료에 나타난 변수들의 시계열상 일관성을 기준으로 몇 가지 변수들에 대해 측정오차의 심각성을 가늠해 보았으며, 나아가 측정오차 문제가 가장 심각할 것으로 판단되는 임금변수에 대해서는 KLIPS 자료로 계산한 결과와 임금대장에 기초한 기존의 연구결과를 비교함으로써 그 심각성을 평가하여 보았다. 그 결과 조사자료에 나타난 측정오차의 문제는 외국의 경우처럼 심각한 수준인 것으로 나타났으며, 그 오차는 종종 고전적인 가정을 위해하는 것으로 나타났다. 더구나 KLIPS 자료는 패널자료이기 때문에 자체적으로 시계열상이나 다른 변수들과의 일치성을 기준으로 어느 정도 편집이 가능하다는 점을 고려해 볼 때 일반적으로 횡단면적 조사자료에 나타난 변수들의 측정오차는 본 연구에서 예시한 것보다 더 심각할 수 있음을 짐작할 수 있다. 아울러 본 연구에서는 조사자료 검증결과가 조사자료에 근거하여 도출된 왜곡된 분석 결과를 보정하는 데에 어떻게 사용될 수 있는가에 대해 예시하고 있다. 최종적으로 본 연구에서는 임금 및 각종 소득, 부가급여, 정규 및 초과 근로시간, 근속기간, 고용형태, 연령, 성, 학력, 산업, 직종, 사업체 규모, 노조 등 기본적인 노동시장 변수들에 대해 PSID 자료검증 연구와 유사한 성격의 자료검증 연구를 수행할 것을 제안한다. 이를 통해 조사 결과 획득된 주요 변수들에 내재해 있는 측정오차의 특성과 정도에 대한 정보를 획득할 수 있으며, 이 정보는 향후 조사자료에 근거하여 수행될 모든 연구들의 분석결과들을 보정하는 데에 도움을 줄 것으로 판단된다. 아울러 조사자료에 존재하는 측정오차의 문제가 특정 시점에서의 변수값만이 아니라 해당 변수의 두 시점 사이의 변화에도 영향

을 미칠 수 있기 때문에 자료검증 조사 및 연구를 시점을 달리하여 최소한 두 차례에 걸쳐 수행할 것을 제안한다.

제 1 장
서 론

본 연구는 패널자료 품질개선 연구 시리즈의 네 번째 연구이다. 『패널자료 품질개선 연구(Ⅱ)』에서는 패널자료에서 소득 히핑(heaping) 표본이탈, 측정오차가 어느 정도인지를 분석하였다. 그리고 『패널자료 품질개선 연구(Ⅲ)』에서는 소득 히핑, 표본이탈, 측정오차 문제를 통계적으로 교정하는 방법에 대해 연구하였다.

패널 품질개선 연구 시리즈에서는 패널 표본이탈의 문제를 계속 다루어 왔다. 『패널자료 품질개선 연구(Ⅱ)』에서는 노동 패널자료의 표본이탈 사유를 소멸과 비응답으로 나누어 분석하고, 표본이탈이 임금에 대한 회귀분석과 임금 불평등 분석에 어떠한 영향을 미치는지를 분석하였다. 그리고 『패널자료 품질개선 연구(Ⅲ)』 제4장에서는 패널자료에서 표본이탈을 교정하기 위한 계량경제학적 방법을 소개하였다. 특히 표본이탈로 인하여 추정에 편의(bias)가 존재하는지 여부를 검증하는 방법과 표본이탈의 편향을 교정하는 방법을 개괄하였다. 본 연구의 제2장에서는 동태적 패널(dynamic panel) 모형에서 표본이탈 문제를 다룬다. 표본이탈이 존재하는 자료를 이용하여 동태적 패널모형을 분석할 때 발생하는 문제점을 설명하고, 기존의 해결방법을 검토하며, 효율성을 제고한 새로운 해결책을 알아본다. 또한 모의실험을 통하여 이 추정량의 통계적 특성을 보이며, 이 방법을 사업체 패널에 응용한다. 그리고 표본이탈이 내생적으로 발생하였을 경우 동태적 패널모형에서 표본이탈을 교정하는 방법을 제시한다.

『패널자료 품질개선 연구(Ⅲ)』의 제5장에서는 표본이탈의 경우를 'Non-ignorable missing'의 관점에서 다루었다. 그리고 제6장에서는 조사과정에서 발생하는 모든 상황에 대한 기록을 담은 paradata 정보를 활용하여 가중치를 보정하는 방법을 다루었다. 본 연구의 제3장에서는 이 두 연구의 연장선상에서, paradata에 있는 사업체의 응답 성향에 대한 정보를 이용하여 무응답 편향을 줄이는 새로운 방법을 제시한다.

본 연구 제4장에서는 기존의 사업체 패널조사에 나타난 몇 가지 문제점을 바탕으로 이를 해결하기 위한 새로운 방안으로 새로운 패널 표본설계 또는 기존 패널을 보완하는 표본 재설계를 할 때 유의해야 할 사항들을 정리한다. 특히 표본이탈을 고려하면서 표본을 추출할 때 고려해야 하는 점을 검토한다.

본 연구 제5장에서는 노동 패널자료에 나타난 변수들의 시계열상 일관성을 기준으로 몇 가지 변수들에 대해 측정오차의 심각성을 검토한다. 이러한 검토과정을 통하여 임금, 근로시간, 근속기간, 고용형태, 연령, 성, 학력, 산업, 직종, 사업체 규모, 노조 등 기본적인 노동시장 변수들에 대해 미국의 'Panel Study of Income Dynamics' 자료검증 연구와 비슷한 자료검증 연구를 수행할 것을 제안하고 있다.

제 2 장
동태적 패널모형에서 표본이탈 교정[1)]

제1절 서 론

미시 패널자료에서 표본이탈(attrition)은 흔하게 나타나는 현상이다. 개인이나 기업은 여러 가지 사유로 인하여 추적 조사로부터 벗어날 수 있다. 개인들의 경우 사망, 주소 추적의 실패, 응답 거절 등의 이유가 전형적이며, 사업체들의 경우 사업체가 소멸하거나 응답을 거절하는 경우 표본이탈이 발생한다. 본 장에서는 홍민기 등(2014)이 고려한 표본이탈의 문제를 동태적 패널모형으로 확장하여 연구한다. 표본이탈이 존재하는 자료를 이용하여 동태적 패널모형을 분석할 때 발생하는 문제점을 설명하고, 기존의 해결방법을 검토하며, 효율성을 제고한 새로운 해결책을 알아본다. 또한 모의실험을 통하여 이 추정량의 통계적 특성을 보이며, 이 방법을 사업체 패널에 응용한다.

이하에서는 우선 제2절에서 표본이탈이 없는 동태적 패널모형의 추정을 설명한다. 여기에서는 1계차분(FD)을 한 후 간단한 도구변수를 사용하는 방법과 Arellano and Bond(1991)의 적률법(Generalized Method of Moments : GMM)을 고려한다. 제3절에서는 FD와 FOD를 이용하여 간단한 추정량들을 효율적으로 결합하는 방법을 검토한다. 제4절에서는 표

1) 본고의 작성에 큰 도움을 주신 고려대학교 박상수 교수님께 감사드립니다.

본이탈 존재 시 사용할 방법들을 살펴본다. 4.1절에서는 표본이탈을 무시하고 모든 자료나 균형화한 부분자료를 사용하는 경우에 대한 계량경제학적 이론과 모의실험 결과를 제시한다. 4.2절에서는 FD 이후에 역밀즈 비율(Inverse Mills Ratio)을 도입하여 표본이탈로 인한 편향을 교정하는 방법(Wooldridge, 2002; 2010)을 살펴보고 모의실험을 통하여 편향교정을 확인한다. 4.3절에서는 제3절에서 검토한 방법이 표본이탈 존재 시에도 이용 가능한지 살펴본다. FOD의 방법은 표본이탈이 종속변수와 완전히 무관하게 결정되는 경우(엄밀히 외생적인 표본이탈)에만 일관성을 지니며, 그 밖의 경우에는 편향이 존재할 수 있음을 모의실험과 대수학을 이용하여 설명한다. 4.4절에서는 Arellano와 Bond의 차분 GMM의 편향을 교정하는 추정량을 소개하고, 이것을 기존의 FD에 편향을 교정하는 추정량과 모의실험을 통해 비교한다. 모의실험 결과에 따르면, 완전한 GMM의 경우 표본크기가 작을 때 편향이 더 커지지만 분산은 크게 줄어든다. 제5절에서는 완전한 GMM의 편향을 교정하는 방법을 사업체 패널에 응용한다. 제6절은 본 장을 맺는다.

제2절 동태적 패널모형의 추정

설명변수로서 종속변수의 과거값이 사용되는 모형을 동태적 패널자료 모형이라고 한다. 본 절에서는 우선 표본이탈이 없는 동태적 패널모형의 추정을 설명한다. 수식으로 나타내면

$$y = X'\beta + u = X_{1,}'\beta_1 + X_{2,}'\beta_2 + X_{3,}'\beta_3 + u \qquad (2.1)$$

와 같다. 여기에서 $X_{1,}$는 외생변수, $X_{2,}$는 선결변수, $X_{3,}$는 내생변수이다. 선결변수 $X_{2,}$에는 시차종속변수 y_{-1}이 포함되어 있다. 외생변수들은 고유오차항과 아무런 관련이 없으며(모든 s와 t에 대하여 $E[X_{1,is}v] = 0$), 선결변수들은 $s \leq t$인 경우 $E[X_{2,is}v] = 0$, 그리고 내생변수들은 $s < t$

일 때 $E[X_{3,is}v]=0$을 만족시킨다. 오차항은 $u=\eta_i+v$로서, η_i는 관측불가의 시간불변 요소(고정효과, fixed effects)이며 v는 고유오차항이다. 고유오차항 v는 이분산성을 가질 수 있지만 반드시 시간에 걸친 상관이 없어야 한다. 만일 고유오차항에 시계열상관이 있으면, 이 시계열상관과 설명변수의 영향이 별도로 식별되지 않는다.

표본이탈의 문제가 없을 때, 흔히 차분적률방법(Difference GMM) 또는 시스템적률방법(System GMM)을 사용하여 모수를 추정한다. 차분 GMM은 모형을

$$\Delta y=\Delta X'\beta+\Delta v \tag{2.2}$$

로 차분하여 고정효과(η_i)를 제거한 후 Δv와 무관한 변수들을 도구변수로 사용한다. $X_{1,}$가 외생적, $X_{2,}$가 선결변수, $X_{3,}$가 내생적이라 할 때, 표준적인 가정하에서 $X_{1,i1}$, ..., $X_{1,iT}$, $X_{2,i1}$, ..., $X_{2,-1}$, $X_{3,i1}$, ..., $X_{3,-2}$ 및 y_{i1}, ..., y_{-2}가 Δv와 무관하다. 이 도구변수들을 모두 활용하여 GMM으로 추정하는 것이 Arellano and Bond(1991)의 차분 GMM이다. Stata 소프트웨어에서 이 방법은 xtabond라는 명령어로 구현되어 있다.

시스템 GMM은 이 차분방정식과 관련된 정보 이외에도 수준 방정식으로부터 추가적 정보를 구한다. 만일 차분하고 적절히 시차를 준 변수들이 전체 오차 η_i+v와 무관하다면 이로부터 별도의 적률조건(moment conditions)을 도출할 수 있다. 이 수준방정식과 관련된 적률조건들을 사용한 적률 추정량을 '수준 GMM 추정량'이라고 하며, 시스템 GMM은 차분 GMM에 등장하는 적률조건들과 수준 GMM에 등장하는 적률조건들을 모두 사용하는 GMM이다. 차분 GMM에 비하여 시스템 GMM은 더 많은 적률조건들을 사용하므로 더 효율적이다. 그러나 수준 GMM 부분이 제대로 작동하기 위해서는 차분한 변수들(특히 Δy)이 고정효과인 η_i와 무관하여야 한다. 이 추가적 요구조건은 때로 매우 중요할 수 있다. 만일 개인들이 신규로 노동시장에 진입하거나 사업체가 사업을 시작한 지 얼마 되지 않았다면 고정효과가 Δy에 영향을 미칠 수도 있다. 이 경우 차분 GMM은 사용하여도 좋으나 시스템 GMM은 잘못된 정보를 제공할 수 있다.

이하의 절에서 표본이탈이 존재하는 경우에 대하여 분석하려면 좀더 수학적으로 엄밀하게 GMM 추정절차를 표현할 필요가 있다. 차분 GMM의 절차를 설명하면 다음과 같다. 차분방정식 $\Delta y = \Delta X'\beta + \Delta v$의 오차항과 무관한 변수들의 집합을 z_{it}라 하자. 그러면 적률조건들은

$$E\begin{bmatrix} z_{i3}(\Delta y_{i3} - \Delta X_{i3}'\beta) \\ z_{i4}(\Delta y_{i4} - \Delta X_{i4}'\beta) \\ \vdots \\ z_{iT}(\Delta y_{iT} - \Delta X_{iT}'\beta) \end{bmatrix} = 0 \tag{2.3}$$

이다. 균형패널에서는 이 적률조건들에 포함된 함수들을 일단 모든 i에 걸쳐서 합산하여

$$\sum_{i=1}^{n}\begin{bmatrix} z_{i3}(\Delta y_{i3} - \Delta X_{i3}'\beta) \\ z_{i4}(\Delta y_{i4} - \Delta X_{i4}'\beta) \\ \vdots \\ z_{iT}(\Delta y_{iT} - \Delta X_{iT}'\beta) \end{bmatrix} \tag{2.4}$$

를 구한다(비균형패널에서는 어떤 t에 대해서는 이 항들이 관측되지 않는 i가 존재할 것이다. 이때에는 관측되는 i에 대해서만 합산한다). 추정하려는 모수의 개수에 비하여 적률조건들이 매우 많으므로 이 함수들을 모수의 개수만큼으로 줄여야 한다. 가장 최적으로 함수의 개수를 줄이는 방법은 앞에 "$D'\Omega^{-1}$"를 곱하는 것이다. 여기에서 D는 (2.4)를 모수에 대하여 미분하여 기댓값을 구한 것이고 Ω는 (2.4)를 참값에서 평가하여 구한 분산공분산 행렬이다. 우선 D는

$$\hat{D}' = -\sum_{i=1}^{n}[\Delta X_{i3}z_{i3}' \cdots \Delta X_{iT}z_{iT}']$$

으로 추정한다. $E[\Delta v|z, z_{-1}, ...] = 0$라는 가정하에

$$\Omega = E\sum_{i=1}^{n}\begin{bmatrix} z_{i3}(\Delta v_{i3})^2 z_{i3}' & z_{i3}\Delta v_{i3}\Delta v_{i4}z_{i4}' \cdots & 0 \\ z_{i4}\Delta v_{i4}\Delta v_{i3}z_{i3}' & z_{i4}(\Delta v_{i4})^2 z_{i4}' \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \cdots z_{iT}(\Delta v_{iT})^2 z_{iT}' \end{bmatrix}$$

이다. 한 단계 효율적 추정법(one-step efficient estimation)에서는 v_{it}가 IID라고 가정하여

$$E[z(\Delta v)^2 z'] = 2\sigma_v^2 E[zz'],$$
$$E[z\Delta v \Delta v_{+1} z_{+1}'] = -\sigma_v^2 E[zz_{+1}']$$

라고 축약한 후 Ω에서 σ_v^2을 나누어 $\sigma_v^{-2}\Omega$ 행렬을 자연적으로 추정하고, 2단계 최적 추정에서는 한 단계 추정으로부터 구한 잔차를 차분한 $\Delta\hat{v}$로 치환하여 추정한다.

한 단계 추정이든 두 단계 추정이든 이상의 차분 GMM 방법을 이하에서는 'full GMM'이라고도 부를 것이다.

한편, 패널의 수가 많고 기간이 짧은 자료가 아니면, 이상의 GMM 방법들은 너무 많은 적률조건을 사용하므로 편향이 클 수 있다. 또한 GMM 추정량은 계산이 너무 복잡하다고 생각될 수 있다. 그래서 차분한 모형 (2.2)의 한 설명변수당 하나의 도구변수만을 사용하여 도구변수 추정을 할 수도 있다. 예를 들어 $X_{1,}$가 외생적, $X_{2,}$가 선결변수, $X_{3,}$가 내생적이라면 설명변수 $\Delta X_{1,}$, $\Delta X_{2,}$, $\Delta X_{3,}$에 대하여 1, $\Delta X_{1,}$, $X_{2,-1}$, $X_{3,-2}$를 도구변수로 사용할 수도 있다(여기에서 정보를 잃지 않는다면 $X_{2,-2}$를 도구변수에 추가하기도 한다). 이 도구변수 추정법에서는 도구변수의 개수가 전 기간에 동일하여 분석이 간편하고 너무 많은 도구변수의 문제가 없다.[2] 표본이탈을 고려할 경우 분석이 그 자체로 복잡하므로 이 간편한 방법을 사용하기도 한다. 이 방법을 이하에서는 'FD/IV'라고 하겠다.

2) 이 경우 t에 걸친 적률함수들의 합을 적률함수로 사용하는 것과 같다. 도구변수의 개수가 모든 t에서 동일하면 이러한 변환이 가능하다.

제3절 균형패널에서 효율성 제고 방안

1. 도 입

동태적 모형에서 표준적인 방법은 (i) 모든 s와 t에서 $E[X_{1,\,is}v]=0$, (ii) $s \le t$에 대하여 $E[X_{2,\,is}v]=0$, (iii) $s<t$에 대하여 $E[X_{3,\,is}v]=0$라는 적률조건을 이용하는 것이다. 이는 v_{it}가 시간에 걸쳐 비상관되었고, 고정효과 및 초기조건이 추후의 오차항과 비상관되었다는 가정을 기초로 한다.

이러한 가정들이 함의하는 모든 선형 적률조건들은 차분 GMM에서 사용된다. 그러므로 비선형 적률조건들을 도입하거나 더 많은 가정들(예를 들어 시스템 GMM에서 사용되는 적률조건들)을 추가로 도입하지 않는 한 차분 GMM은 가장 효율적인 방법이다. 따라서 차분 GMM(full GMM)보다 더 효율적인 방법을 고안한다는 것은 추정량을 더 복잡하게 만들거나(예를 들어 2차 함수 형태의 적률조건 활용) 가정들을 더 추가한다는 것(예를 들어 시스템 GMM처럼)을 의미할 뿐이다.

반면 FD/IV 추정법은 차분방정식 $\Delta y=\Delta X'\beta+\Delta v$에 대하여 각 t에서 동일한 수의 도구변수들만을 사용한다. 예를 들어 단순한 모형 $\Delta y=\alpha\Delta y_{-1}+\Delta v$에 대하여 y_{-2}를 도구변수로 사용하는 것이다. 이 방법은 차분 GMM(full GMM)보다 도구변수를 더 적게 사용한다. 차분 GMM(full GMM)이 각 t에서 y_{i1}, ..., y_{-2}를 도구변수로 사용함에 반하여 FD 추정법은 y_{-2}만을 도구변수로 사용한다. 또한 y_{-2}만을 도구변수로 사용한다 할지라도, 최적 GMM의 방법은 $t=3$부터 $t=T$까지 $E[y_{-2}(\Delta y-\alpha\Delta y_{-1})]=0$이라는 $T-2$개의 적률조건들($t=3,\ldots,T$)을 최적으로 결합함에 반하여 FD/IV 추정법은 적률함수들을 단순 합한

$$\sum_{t=3}^{T} y_{-2}(\Delta y-\alpha\Delta y_{-1}) \tag{3.1}$$

의 적률함수를 사용한다.

이렇게 적률조건의 개수를 줄임으로써 GMM 추정량의 편향을 획기적으로 줄일 수 있다. 모의실험을 통하여 모든 도구변수들을 사용한 최적 GMM과 (3.1)을 사용한 도구변수 추정량을 비교해 본 결과, α의 참값이 0.5이고 $n=100$, $T=10$일 때 모든 적률조건들을 사용한 full GMM의 평균이 0.43으로 추정된 반면 (3.1)만을 사용한 도구변수 추정량은 거의 편향되지 않았다. 물론 모든 적률조건들을 다 사용한 최적 GMM의 분산이 작기는 하지만 (3.1)에 기초한 추정량의 상대적으로 작은 편향은 좋은 성질로 간주된다.

그런데, 동태적 패널모형에서 고정효과를 제거하는 데에 차분의 방법만 있는 것은 아니다. 특히 v가 IID라는 가정하에 'Forward Orthogonal Deviations(FOD, Arellano and Bover, 1995)'를 구하면 이렇게 변환된 오차항은 여전히 동분산적이고 시간에 걸쳐 비상관이다. 구체적으로 x_t 변수에 대하여 각 t마다 우선 $x_t - \frac{1}{T-t}\sum_{s=t+1}^{T} x_s$로써 'forward demeaning'을 한다. 원래 변수 x_t가 white-noise일 때 이 'forward deviations'는 시간에 걸쳐 비상관(orthogonal)이지만 이분산적이다. 이 이분산성을 바로잡기 위하여 $C_{T-t}=\left(\frac{T-t}{T-t+1}\right)^{1/2}$을 곱하면

$$\ddot{x}_t = C_{T-t}\left(x_t - \frac{1}{T-t}\sum_{s=t+1}^{T} x_s\right) \tag{3.2}$$

는 시간에 걸쳐 white-noise이다. 이 $\ddot{x}_t$를 $x_1,\ldots,x_T$의 FOD라 한다. FOD를 사용하여 Arellano와 Bond의 full GMM을 나타낼 수 있다. 우선 추정하고자 하는 방정식을 FOD로 나타내면

$$\ddot{y} = \ddot{X}\beta + \ddot{v}$$

가 된다. 각 t에서 도구변수들을 z라고 하자. 그러면 $\ddot{v}$가 t에 걸쳐서 동분산적이고 비상관이므로 각 t마다 $\ddot{X}$를 도구변수들에 대하여 회귀하여 fitted values를 구하고, 모든 자료를 pool하여 $\ddot{y}$를 이 fitted values에 대

하여 OLS 회귀하면 된다.

이 방법은 단지 한 단계 full GMM을 계산하는 좀더 단순한 계산방법일 뿐이다. 그러므로 표본크기가 크지 않을 때, 과다적률조건(many moment conditions)의 문제를 그대로 가지고 있어서 편향이 비교적 크다. 그런데 이 방법에 착안하여 우선 추정방정식을 FOD로 표현하고 여기에 고정된 개수의 도구변수들만을 고려하는 방법을 생각해 볼 수 있다. 예를 들어 단순한 모형에서 $\ddot{y}=\alpha\ddot{x}+\ddot{v}$ 방정식(여기에서 $\ddot{x}$는 우변 시차종속변수의 FOD)에 대하여 y_{-1}(또는 1과 y_{-1})을 도구변수로 사용할 수 있다. 이 경우 추정량은

$$E\left[\sum_{t=2}^{T-1} y_{-1}(\ddot{y}-\alpha\ddot{y}_{-1})\right]=0 \tag{3.3}$$

이라는 적률조건을 사용한다. 이를 'FOD/IV 추정량'이라 하자. (3.3)에는 y_{-1}만을 도구변수로 사용하였으나 더 많은 변수를 사용할 수도 있다. FOD/IV 추정방법과 full GMM의 차이는, full GMM이 각각의 t에서 상이한 수의 도구변수를 사용하였음에 반하여, FOD/IV 추정방법에서는 동일한 수의 도구변수를 활용하여 과다 적률조건의 문제를 완화한다는 것이다.

FOD/IV 추정량과 FD/IV 추정량의 차이는 FD/IV 추정량이 차분으로써 고정효과들을 제거하는 반면 FOD/IV는 FOD를 구함으로써 고정효과를 제거한다는 데에 있다. 오차항 t에 걸쳐 white-noise일 때 오차항의 FOD가 여전히 t에 걸쳐 white-noise이므로, FOD/IV가 FD/IV보다 더 효율적일 것으로 추측할 수 있다. 사실 이러한 추측을 바탕으로 Hayakawa 등은 시계열자료에 대하여 FOD/IV 추정방법을 제안한다.

하지만 놀랍게도 T가 고정된 동태적 패널자료의 경우 이러한 추측은 맞지 않는다. 모의실험 결과에 따르면 α가 작을 때 FD/IV가 오히려 FOD/IV보다 더 효율적인 것으로 보인다. 이는 근본적으로 FD/IV와 FOD/IV가 사용하는 적률조건이 동일하지 않고, FD/IV가 가장 강한 도구변수들을 사용하는 반면 FOD/IV는 큰 관계가 없는 적률조건들도 사용하기 때문이다. 단순한 모형에서 FD/IV와 FOD/IV가 사용하는 적률조

건들은 각각

$$\sum_{t=3}^{T} E[y_{-2}\Delta u] = 0 \quad \text{과} \quad \sum_{t=2}^{T-1} E[y_{-1}\ddot{u}] = 0$$

이다. 그런데

$$\ddot{u} = -\, C_{T-t} \sum_{j=1}^{T-t} \left(1 - \frac{j-1}{T-t}\right) \Delta u_{+j}, \;\; C_m = \sqrt{m/(m+1)}$$

이므로 FOD/IV를 위한 적률조건은

$$-\sum_{t=2}^{T-1} C_{T-t} \sum_{j=1}^{T-t} \left(1 - \frac{j-1}{T-t}\right) E[y_{-1}\Delta u_{+j}] = 0$$

이다. 양변에 -1을 곱한 후 다시 정리하면

$$\sum_{j=2}^{T-1} \left\{ \sum_{t=j+1}^{T} C_{T-t+j-1} \left(1 - \frac{j-2}{T-t+j-1}\right) E[y_{-j}\Delta u] \right\} = 0$$

가 된다. 이로부터, FOD/IV 추정량은 $j=2$뿐 아니라 더 큰 j(시차)와 관련된 적률조건들도 활용함을 알 수 있다. 반면 FD/IV에서는 $j=2$만을 사용하고, $E[y_{-j}\Delta u]$의 계수가 모두 동일하다. FOD/IV에서 더 먼 시차($j>2$)의 적률조건들도 결부됨을 알 수 있다.

FD/IV와 FOD/IV의 상대적 효율성을 분석적으로 도출하는 것은 유용하지 않아 보인다. 모의실험을 통하여 이 둘을 비교하면 단순모형의 경우 α의 값이 0에 가까우면 FD/IV가 더 낫고 α의 값이 1에 가까우면 FOD/IV가 더 나은 경우도 있다. 직관적으로 이는 충격의 지속성이 강할 때 먼 시차의 적률조건들이 도움이 되기 때문일 것으로 추측된다. 충격의 지속성이 약하면 먼 시차들은 별 정보를 주지 못한다.

표본크기(n)가 클 때, 가장 효율적인 방법은 모든 적률조건들을 전부 사용하는 것(full GMM)이다. 하지만 적률조건이 많거나 n이 크지 않을 때, 이 추정량은 편향이 크다. 반면 FD/IV와 FOD/IV는 작은 수의 적률조건만을 사용하므로 편향이 작다(적률조건의 수와 편향과의 관계는 Han

and Phillips, 2006 참조).

이하에서는 일반적인 모형 $y = X'\beta + u$, $u = \eta_i + v$에 대하여 FD/IV와 FOD/IV를 정의하고 이 둘을 결합한 추정량을 제시한다. GMM의 방식으로 최적으로 FD/IV와 FOD/IV를 결합하면 그 결과는 FD/IV나 FOD/IV보다 상대적으로 항상 더 낫다.

2. FD/IV와 FOD/IV의 결합

본 소절에서는 FD/IV, FOD/IV와 이들을 결합한 GMM 추정량을 정의할 것이다. 모형은 $y = X'\beta + u$, $u = \eta_i + v$이다. $X_{1,}$, $X_{2,}$, $X_{3,}$가 각각 외생, 선결, 내생 설명변수라 하자(y_{-1}은 $X_{2,}$에 포함됨). 시간을 맞추는 것이 편리하므로 1계차분을 $\tilde{y} = y - y_{+1}$이라고 정의한다. $\tilde{X}$와 $\tilde{v}$도 동일하게 정의된다. 이 차분들은 $t = 2, \ldots, T-1$에 대하여 정의된다. $\ddot{y}$ 등은 앞에서 정의한 FOD들로서, 마찬가지로 $t = 2, \ldots, T-1$에 대하여 정의된다.

우리가 사용할 FD/IV의 방법은 $\tilde{y} = \tilde{X}'\beta + \tilde{v}$에 대해서 1, $\widetilde{X_{1,}}$, $X_{2,}$ 및 $X_{3,-1}$를 도구변수(Z^a)로 사용한다($t = 2$에서 관측되는 한 $X_{2,-2}$도 사용할 수 있다). 그러면 FD/IV 추정량은 우선 $\tilde{X}$를 Z^a로 OLS 회귀하여 fitted values $\hat{\tilde{X}}$를 구한 후, $\tilde{y}$를 $\hat{\tilde{X}}$로 OLS 회귀하여 구한다. 상수항 1을 도구변수 집합에 포함시키는 것이 좋다.

FOD/IV의 방법은 우선 $\ddot{y} = \ddot{X}'\beta + \ddot{v}$에 대하여 1, $\ddot{X}_{1,}$, $X_{2,}$ 및 $X_{3,-1}$를 도구변수(Z^b)로 사용한다($t = 2$에서 관측되는 한 $X_{2,-2}$도 사용할 수 있다). 그러면 FOD/IV 방법과 동일하게 $\ddot{X}$를 Z^b에 대하여 OLS 회귀하여 fitted values $\hat{\ddot{X}}$를 구한 후, $\ddot{y}$를 $\hat{\ddot{X}}$에 대하여 OLS 회귀하여 구한다.

이 두 추정방법에 상응하는 표본 적률조건들은 각각

$$\sum_{i=1}^{n}\sum_{t=2}^{T-1}\hat{\tilde{X}}(\tilde{y} - \tilde{X}'\beta) = 0$$

$$\sum_{i=1}^{n}\sum_{t=2}^{T-1}\hat{\ddot{X}}(\ddot{y}-\ddot{X}'\beta)=0$$

이다. FD/IV 추정량과 FOD/IV 추정량은 각각 이들의 해이다. 두 방정식을 행렬로 나타내면 각각 $\hat{\tilde{X}}'(\tilde{y}-\tilde{X}\beta)=0$과 $\hat{\ddot{X}}'(\ddot{y}-\ddot{X}\beta)=0$로 쓸 수 있다. FD/IV와 FOD/IV를 최적으로 결합하기 위해서는 1계도함수와 공분산행렬의 추정치가 필요하다. 1계도함수 행렬은

$$\hat{D}'=-[\tilde{X}'\hat{\tilde{X}},\ \ddot{X}'\hat{\ddot{X}}]$$

이다(여기에 -1을 곱하여도 상관없다). 공분산행렬을 추정하기 위해서, 각 추정에서 구한 잔차 $\hat{\tilde{v}}$와 $\hat{\ddot{v}}$를 이용하여 $2k\times 1$ 벡터 $\xi=\begin{pmatrix}\hat{\tilde{X}}\hat{\tilde{v}}\\ \hat{\ddot{X}}\hat{\ddot{v}}\end{pmatrix}$를 계산하자. 각 i에서 이를 t에 걸쳐 합산하여 $\xi_i=\sum_{t=2}^{T-1}\xi$를 구하자. 관측치들이 i에 대해서 독립이라는 가정하에 공분산행렬의 추정값은

$$\hat{\Omega}=\sum_{i=1}^{n}\xi_i\xi_i'$$

이다. 이제 $H'=\hat{D}'\widehat{\Omega^{-1}}$라 하고, $H'=(H_a',H_b')$로 구획하면 FD/IV와 FOD/IV를 최적으로 결합하는 GMM 추정량은

$$H_a'\hat{\tilde{X}}'(\tilde{y}-\tilde{X}\beta)+H_b'\hat{\ddot{X}}'(\ddot{y}-\ddot{X}\beta)=0$$

을 만족시킨다. 즉 최적의 GMM 추정량은

$$\hat{\beta}=(H_a'\hat{\tilde{X}}'\tilde{X}+H_b'\hat{\ddot{X}}'\ddot{X})^{-1}(H_a'\hat{\tilde{X}}'\tilde{y}+H_b'\hat{\ddot{X}}'\ddot{y})$$

이다. 이 추정량을 FD-FOD이라 하자. FD-FOD는 최소한 FD/IV와 FOD/IV만큼 좋다.

3. 모의실험

설명변수가 y_{-1}뿐인 단순한 동태적 패널모형에 대하여 모의실험을 해

보았다. $n=100$, $T=10$에 대하여 1,000회 반복하여 자료를 발생시킨 후 FD/IV, FOD/IV, FD-FOD 및 Arellano와 Bond의 차분 GMM(full GMM; 모든 도구변수들을 사용) 추정을 해본 결과는 <표 2-1>과 같다.

추정 결과 Arellano-Bond full GMM의 편향이 가장 크다(하지만 n이 증가하면서 이 편향은 0으로 줄어든다). FD/IV, FOD/IV, FD-FOD 모두 편향이 거의 없다. Full GMM 추정량의 분산이 가장 작다. 처음 세 추정량 중에서 FD-FOD의 분산이 가장 작다. α의 값이 클 때, FD/IV에 비한 FD-FOD의 효율성 이득은 무시할 수 없을 만큼 크다.

표본이탈이 존재할 때, Wooldridge(2002; 2010)는 FD/IV의 방법과 Heckman의 표본선택 편향 교정방법을 결합한다. 이 방법이 편리한 것은 사실이지만, 앞의 모의실험 결과에 따르면 FD-FOD의 방법은 편향을 증가시키지 않으면서도 상당한 효율성 증대를 가져올 수 있었다. 그러므로 표본이탈이 존재할 때에도, FD/IV와 FOD/IV를 결합시키는 것이 더 나은지 확인할 필요가 있다. 또한 Arellano-Bond의 full GMM은 표본 크기가 작을 때 상당한 편향을 가지지만 분산으로써 측정한 효율성이 매우 높으므로 유용성이 있다. 다음 절에서는 이처럼 FOD를 사용하는 추정량과 full GMM 추정량을 Heckman의 표본선택 편향 교정방법과 결합한 추정방법을 검토할 것이다.

〈표 2-1〉 표본이탈이 없을 때 n=100, T=10, 1,000회 반복

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.200	0.337	0.695	1.324
FOD/IV	0.198	0.406	0.692	1.734
FD-FOD	0.200	0.301	0.697	1.088
full GMM	0.180	0.202	0.625	0.475

제4절 표본이탈이 있는 동태적 패널모형의 추정

표본이탈이 존재할 때 연구자가 취할 수 있는 방법은 두 가지이다. 하나는 표본이탈을 무시하고 전체 비균형패널이나 균형화한 부분패널에 대하여 통상적인 방법을 적용하는 것이다. 또 하나의 방법은 표본이탈을 고려하여 이로 인한 편향을 교정하여 주는 것이다. 4.1절에서는 동태적 패널모형의 경우 우선 표본이탈이 어떠한 성질을 보일 때 기존의 방법을 수정없이 사용할 때에도 편향이 발생하지 않는지 살펴볼 것이다. 본 절의 주요 대상은 표본이탈을 무시한 방법이 작용하지 않는 경우이다. 먼저 표본이탈 편향을 교정한 기존의 방법을 설명하고, 다음으로 FOD/IV 방법을 고려한다. 마지막으로 내생적 표본이탈 존재 시 full GMM을 사용하여 효율성을 제고하는 방법을 설명한다.

1. 표본이탈 존재 시 비균형패널이나 균형화된 부분표본의 분석

표본이탈이 존재한다고 하자. 표본관측 여부를 a변수로 나타내자. 개체 i가 t기에 관측되면 $a=1$이고 패널로부터 이탈하면 $a=0$이다. 균형패널에서는 모든 a가 1이며, 비균형패널에서는 어떤 i와 t에 대해서 $a=0$이다.

IV 또는 GMM에서 사용할 적률조건들을 일반적인 형태로 표현하면 $E[g_i(\beta)]=0$이다. Arellano and Bond(1991) 추정량의 경우 (2.4)를 기초로 추정이 이루어지는데, 표본이탈이 일어나 t마다 이용 가능한 관측치의 수가 다르면 그 대신

$$\sum_{i=1}^{n}\begin{bmatrix} a_{i3}z_{i3}(\Delta y_{i3}-\Delta X_{i3}'\beta) \\ a_{i4}z_{i4}(\Delta y_{i4}-\Delta X_{i4}'\beta) \\ \vdots \\ a_{iT}z_{iT}(\Delta y_{iT}-\Delta X_{iT}'\beta) \end{bmatrix} \tag{4.1}$$

에 기초하여 추정한다. 여기에서 표본이탈이 한번 일어나면 표본으로 복귀하지 않는다고 가정하였다. 즉 $a=1$이면 모든 $s \le t$에 대하여 $a_{is}=1$이다.

이 (4.1)의 함수들은 변형된 적률조건 $E[az(\Delta y - \Delta X'\beta)]=0$에 상응한다. 그러므로 이 변형된 적률조건들이 성립하는 한 비균형패널을 사용한 추정량은 일관성(consistency)을 지닐 것이다. 모수의 참값에서 이 변형된 적률조건은 $E[az\Delta v]=0$임을 의미한다. 그러므로 도구변수가 주어졌을 때 표본이탈 변수 a가 모두 차분한 오차항 Δv과 무관하면 비균형패널을 모두 사용하는 추정량은 일관적이다. 예를 들어, 만일 표본이탈 여부(a_{it})가 $t-2$기 이전의 종속변수값들에 의하여 결정되며, $t-2$기까지의 정보와 고정효과가 주어졌을 때 y_{-1} 및 y와는 무관하다면 비균형패널을 사용한 추정량은 일관적이다. 즉 2기 이전까지의 정보가 주어졌을 때 표본이탈이 현재 및 1기 전의 종속변수값과 무관하게 발생하여야 함을 의미한다. 그러지 않고, 만일 예를 들어 직전 기의 사업성과가 낮아 사업체가 소멸하여 표본이탈이 일어나는 경우, 표본이탈을 제대로 고려해주지 못하면 추정량이 편향될 수 있다.

분석과 계산이 더 쉬운 FD/IV 추정량의 경우, 딱 맞게 식별된 모형의 추정량은

$$\left(\sum_{i=1}^{n}\sum_{t=3}^{T} az\Delta X'\right)^{-1}\sum_{i=1}^{n}\sum_{t=3}^{T} az\Delta y$$

이다. 이 경우에는 $E\left(\sum_{t=3}^{T} az\Delta v\right)=0$이면 추정량이 일관적일 것이다. 다만 어떤 t에서 $E(az\Delta v)\neq 0$이면서 $E\left(\sum_{t=3}^{T} az\Delta v\right)=0$가 되는 경우는 상상하기 어려우므로 이 IV 추정량이 더 일반적인 조건하에서 일관적이라고 논하기에는 무리가 있다.

모의실험을 해보자. 단일 변수 y에 대하여 동태적 모형 $y=\alpha y_{-1}+u$, $u_{it}=\eta_i+v_{it}$를 고려한다. 표본이탈을 결정할 변수 a는 $a_{i1}\equiv 1$이며 $t \ge 2$는 다음에 따라 생성된다:

$$a = a_{-1} \times I(1.5 + c_1 y_{-1} + c_2 y_{-2} + \sqrt{1-c_1^2}\,\varepsilon > 0) \qquad (4.2)$$

표본이탈 이후 복귀하지 않으므로 $a=0$이면 모든 $s>t$에 대하여 $a_{is}=0$이다. 여기에서 ε_{it}는 독립적인 $N(0,1)$ 임의변수이다. $c_1=0$이면 표본이탈은 외생적이라고 할 수 있으며, $c_1 \neq 0$이면 표본이탈은 내생적이라고 할 수 있다(a가 Δv와 연관되므로). $c_1=0$이고 $c_2=0$이면 표본이탈은 종속변수의 값과 완전히 무관하며, 이때 표본이탈은 엄밀히 외생적이다. 만약 $c_1=0$이고 $c_2 \neq 0$이면, 표본이탈이 2기 전의 종속변수값에 의해 결정되지만, Δv와 무관하게 된다. 그러나 종속변수의 과거값에 의해 표본이탈이 발생하므로, 이 경우를 약하게 외생적(또는 선결, pre-determined)인 표본이탈이라고 하자.

이하에서는 FD/IV 추정량과 full GMM 추정량의 분포를 1,000회 반복하여 모의실험한 결과를 보고한다.[3] FOD/IV에 관해서는 다음 장에서 설명할 것이다.

우선 <표 2-2>는 표본이탈이 엄밀히 외생적일 때($c_1=0,\ c_2=0$), 전체 자료(비균형패널)를 사용한 FD/IV 및 full GMM 추정량의 성질을 요약하고 있다.

두 추정량들 모두 일관적인 것으로 보인다. Full GMM 추정량은 여타 추정량에 비하여 더 많은 수의 적률조건을 사용하며, 따라서 편향은 좀 더 크고 분산은 현저하게 작다.

〈표 2-2〉 엄밀히 외생적인 표본이탈 시 전체 자료를 이용한 추정

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.199	0.326	0.698	1.251
full GMM	0.195	0.232	0.681	0.504

주 : $n=500,\ T=10,\ c_1=0,\ c_2=0.$

3) 이상의 실험에서 도구변수 집합에 상수항을 포함시켰다. 이 실험들에서는 고정효과의 평균이 0으로 설정되었으므로 상수항을 포함시키는 것은 그리 중요하지 않으나 실제 자료의 분석에서는 도구변수에 상수항을 포함시키는 것이 중요할 수 있다(Han and Kim, 2014).

〈표 2-3〉 약하게 외생적인 표본이탈 시 전체 자료를 이용한 추정

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.200	0.658	0.699	3.360
full GMM	0.199	0.471	0.674	1.242

주 : $n=500,\ T=10,\ c_1=0,\ c_2=0.5.$

다음으로 약하게 외생적인 표본이탈이 일어날 때($c_1=0,\ c_2=0.5$)의 모의실험 결과이다.

FD/IV와 full GMM에서는 편향의 증가가 관측되지 않는다. 하지만 FOD/IV 추정량은 편향이 있는 것처럼 보이기도 한다. 표본크기(n)를 2,000으로 늘렸을 때, FOD/IV 추정량의 평균은 $\alpha=0.2$에서 0.190으로 추정되었고 $\alpha=0.7$에서 0.671로 추정되어, n이 증가하면서 편향이 감소하는지 분명하지 않다. n=5,000으로 증가시켜 보았는데 α=0.2일 때 FOD/IV 추정량의 평균은 여전히 0.189로 추정되어 편향이 제거되지 않아 보인다. 이에 대해서는 추후에 설명한다.

다음 <표 2-4>에서 보는 것처럼, 식 (4.2)에서 c_1=0.5, 즉 표본이탈이 y_{-1}의 값이 작을수록 더 빈번히 일어난다면 전체 자료를 이용한 추정량들은 크게 편향될 수 있다.

전체 관측치 중 모든 기간에 대하여 관측되는 i만을 대상으로 균형패널(balanced subset)을 만들어 분석하는 방법도 있다. 이를 수식으로 표현하면, $A_i=I\left(\sum_{t=1}^{T}a=T\right)$라는 변수, 즉 모든 t에서 관측이 이루어진 i에 1의 값을 부여하는 변수 A_i를 만든 후 (4.2)의 a_{it}를 A_i로 대체하는

〈표 2-4〉 내생적인 표본이탈 시 전체 자료를 이용한 추정

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.119	0.479	0.419	2.302
full GMM	0.146	0.384	0.580	1.413

주 : $n=500,\ T=10,\ c_1=0.5,\ c_2=0.$

것이다. 그러면 이에 상응하는 적률조건들은 $E[A_i z \Delta v] = 0$이라는 것이다. 이 조건이 성립하기 위해서는, 간략히 말해서 모든 t에서 A_i가 Δv와 무관하다는 것이다. 다시 말해 a_{is}와 Δv가 모든 s와 t에서 서로 무관하다는 것이다. 전체 자료를 사용하는 경우에는 a가 v 및 v_{-1}과 무관하다는 것이 일관성의 조건이었으므로, 균형화된 패널의 분석이 일관적이기 위해서는 전체 자료를 사용하는 경우보다 더 강한 조건을 필요로 한다. 그렇다고 하여, 균형화한 패널의 분석이 반드시 더 큰 편향을 초래한다는 것은 아님에 유의해야 한다.

모의실험을 해보자. 식 (4.2)에 따라 a가 생성된다. 만일 $c_1 = 0$이고 $c_2 \neq 0$이라면 전체 자료(비균형패널)를 이용한 추정량은 여전히 일관적이다. 하지만 $c_1 = 0$이더라도 $c_2 \neq 0$이면 표본이탈이 더 과거의 종속변수값에 의존하므로 균형화한 부분표본($A_i = 1$인 i)을 사용한 추정량은 편향될 수 있다. 다음 모의실험 결과들을 살펴보자.

우선 <표 2-5>는 표본이탈이 엄밀히 외생적인 경우의 결과를 보여준다. Full GMM을 제외한 여타 추정량들의 편향은 무시할 만하다. Full GMM의 경우에는 표본크기가 작을 때 편향이 눈에 띄나, $n = 2{,}000$으로 증가시키면 사라짐을 확인하였다($n = 2{,}000$에서 $\alpha = 0.7$인 경우의 평균은 0.691).

반면 $c_2 = 0.5$로서, 표본이탈이 약하게 외생적이면 다음 <표 2-6>의 결과를 얻는다.

두 추정량 모두 편향이 현저해 보이며, $n = 2{,}000$까지 표본크기를 증가시켜 보았으나 FD/IV에서도 편향이 거의 유지되었다($\alpha = 0.2$일 때 평균은 0.187). 앞에서 이야기한 것처럼, 표본이탈이 약하게라도 종속변수

〈표 2-5〉 외생적인 표본이탈 시 균형화한 부분자료를 이용한 추정

	$\alpha = 0.2$		$\alpha = 0.7$	
	평균	$n \times$분산	평균	$n \times$분산
FD/IV	0.199	0.594	0.699	2.204
full GMM	0.192	0.410	0.668	0.922

주 : $n = 500,\ T = 10,\ c_1 = 0,\ c_2 = 0.$

에 의존할 때, 균형화한 부분자료를 사용하여 추정하면 편향이 발생한다. 물론 $c_1 \neq 0$이어서 표본이탈이 동시대의 차분한 오차항에 의존하는 경우에도 균형화한 부분자료를 이용한 추정량은 편향되어 있다.

만일 $c_1 \neq 0$이면 전체 자료를 이용하든 균형화한 부분자료를 이용하든 모두 편향되어 있을 것이다. $c_1 = 0.5$, $c_2 = 0$일 때, 전체 자료를 이용하면 그 결과는 <표 2-4>와 같고, 균형화한 부분자료를 이용한 결과는 <표 2-7>과 같다. 양자 모두 편향되어 있으나 <표 2-7>에 균형화한 부분자료를 이용한 추정량의 편향이 오히려 더 작다.

본 소절에서는 FOD/IV와 FD-FOD에 대하여 고려하지 않았다. 이는 FOD 변환을 한 값들이 현재 기부터 마지막 기까지의 표본이탈 여부에 의존하게 되어 추가적 편향의 요인이 되기 때문이다. 이에 대해서는 4.3절에서 상세히 설명한다.

한편, 표본이탈이 관측가능값과 무작위적인 요소에 의해 결정될 때, 표본잔류 확률 역수 가중치(inverse probability weighting)의 방법으로 표본이탈 편향을 제거하는 방법이 있으나(Moffitt et al., 1999), 실제 응용연구에서 연구자의 관심은 주로 내생적 표본이탈의 문제를 해결하는 데에 있다.

〈표 2-6〉 약하게 외생적인 표본이탈 시 균형화한 부분자료를 이용한 추정

	$\alpha = 0.2$		$\alpha = 0.7$	
	평균	$n \times$분산	평균	$n \times$분산
FD/IV	0.186	0.695	0.646	1.818
full GMM	0.178	0.471	0.627	0.859

주 : $n = 100$, $T = 10$, $c_1 = 0$, $c_2 = 0.5$.

〈표 2-7〉 내생적인 표본이탈 시 균형화한 부분자료를 이용한 추정

	$\alpha = 0.2$		$\alpha = 0.7$	
	평균	$n \times$분산	평균	$n \times$분산
FD/IV	0.185	0.553	0.642	1.584
full GMM	0.180	0.396	0.631	0.792

주 : $n = 500$, $T = 10$, $c_1 = 0.5$, $c_2 = 0$.

2. FD/IV 추정 시 내생적 표본이탈 편향의 교정

Wooldridge(2002; 2010)는 동태적 패널에서 Heckman의 2단계 추정법을 사용하여 표본이탈 편향을 교정하고자 한 거의 유일한 연구자이다. Wooldridge는 (2.2)처럼 우선 1계차분으로써 고정효과를 제거한다.

$$\Delta y = \Delta X'\beta + \Delta v \tag{4.3}$$

표본관측 여부 a가 Δv와 연관되었을 때 표본이탈은 내생적이다. 이 표본관측 변수가 다음의 모형에 의하여 설명된다고 하자.

$$a = I(w'\delta + \varepsilon > 0) \tag{4.4}$$

여기에서 w는 회귀오차항 Δv와 무관하며 ε는 Δv와 연관될 수 있다. 만일 z가 w를 포함하면, 오차항들이 정규분포를 갖는다는 가정하에

$$E(\Delta v|z,\ a=1) = \rho_t \lambda(w'\delta_t) \tag{4.5}$$

즉

$$E(\Delta y|z,\ a=1) = E(\Delta X|z,\ a=1)'\beta + \rho_t \lambda(w'\delta_t) \tag{4.6}$$

를 얻는다. 여기에서 $\lambda(\cdot)$는 Inverse Mills Ratio(IMR), 즉 $\phi(\cdot)$을 표준정규분포의 확률밀도함수, $\Phi(\cdot)$를 표준정규분포의 누적확률분포함수라 할 때 $\lambda(c) = \phi(c)/\Phi(c)$이고, ρ_t는 미지의 모수이다. 그러므로 만일 δ_t를 알고 있다면, t기에 살아남은 관측치들만을 이용하여 Δy를 ΔX와 $\lambda(w'\delta_t)$에 도구변수 회귀함으로써 편향을 교정한 추정값을 구할 수 있다. 물론 δ_t는 알 수 없으며 식 (4.4)를 추정하여 구한 $\hat{\delta_t}$를 이용하여 $\hat{\lambda} = \lambda(w'\hat{\delta_t})$를 계산함으로써 이를 대신할 수 있다. 기울기 β의 추정을 위해서는 모든 t를 모아서 하나의 회귀를 한다. 이때 편향교정항(Inverse Mills Ratio)의 계수가 각 t마다 다를 수 있으므로 ΔX 및 편향교정항과 시간더미의 상호작용항을 설명변수로 하여 추정한다. 구체적으로, 다음의 단계를 따른다.

(1) 각각의 t에서 회귀식 (4.4)를 프로빗(probit)으로 추정하여 Inverse Mills Ratio $\lambda(w'\delta_t)$를 추정한다. 이를 $\hat{\lambda}$라 하자.

(2) 원래 방정식 (4.3)의 우변에 $\hat{\lambda}$를 추가하여 도구변수 추정을 한다. 이때 도구변수로서 z와 $\hat{\lambda}$를 사용한다.

표본이탈 문제가 없을 때 이상의 추정방법은 1계차분 후 도구변수 추정을 하는(Anderson and Hsiao, 1981) 방법에 해당한다. 이 방법은 두 가지 점에서 정보를 효율적으로 이용하고 있지 못하다. 첫째, 이용 가능한 모든 도구변수들을 사용하지 않는다. 둘째, 차분한 오차항에 시계열상관이 존재하여 통상적인 도구변수 추정량이 효율적이지 못하다. 이하에서는 이 두 가지 비효율성의 문제를 다룬다.

3. FOD/IV

본 소절에서는 표본이탈 존재 시 FOD/IV의 방법을 사용할 수 있는지 검토한다. FOD/IV의 방법은 먼저 고정효과를 제거하기 위하여 자료를 FOD 변환한 후, 가장 연관성 있는 소수의 도구변수만을 사용하는 것이다. 표본이탈이 존재할 때에는 Wooldridge의 FD/IV의 방법과 유사하게 편향교정항(IMR)을 추가시킬 것이다.

이 방법은 1계차분으로써 표현된 식 (4.3)을 FOD들로 치환하고, 최소한의 도구변수를 사용하는 것이다. FOD들은 (3.2)에 정의되어 있다. $\ddot{y}$, $\ddot{X}$, $\ddot{v}$를 이 FOD들이라 할 때, 표본이탈 존재 시 FOD/IV의 방법이 작동하기 위해서는 (4.5)와 유사한 다음의 관계가 충족되어야 한다.

$$E(\ddot{v}|z,\ a=1)=\rho_t\lambda(w'\delta) \tag{4.7}$$

여기에서 z는 w를 포함하도록 정의되었음에 유의해야 한다.

모의실험을 통해 FD/IV와 FOD/IV를 비교해 보자. 앞에서와 마찬가지로 식 (4.2)에 따라서 표본이탈이 발생한다. $c_1=0$이면 표본이탈이 엄밀히 외생적으로 일어나고, $c_1\neq 0$이면 표본이탈이 변환한 오차항에 대하

여 내생적으로 일어난다. 만약 $c_1 = 0$이고 $c_2 \neq 0$이면 표본이탈이 약하게 외생적으로 일어난다.

4.1절에서 살펴본 것처럼, 표본이탈이 외생적으로 일어나거나 약하게 외생적으로 일어나면 교정 없이 전체 자료를 사용하는 추정량은 일관적일 것으로 추측할 수 있다. 표본이탈 여부가 Δv 또는 $\ddot{v}$와 무관하므로, 내생적인 표본이탈로 인한 편향이 없기 때문이다. 그러나 표본이탈이 내생적으로 일어나면 추정량에 편향이 발생하므로, 편향교정항을 추가하면 일관적인 추정량을 얻을 수 있을 것으로 예상할 수 있다. 이러한 예상은 FD/IV의 경우에는 들어맞지만, FOD/IV의 경우에는 맞지 않는다. 그 이유에 관해서는 본 소절의 마지막에 대수적으로 설명하고, 우선은 모의실험을 통하여 확인해 보자.

편향교정항의 독립변수 w가 될 수 있는 조건은 Δv 또는 $\ddot{v}$와 무관해야 한다는 것이다. 따라서 FOD/IV의 경우에도 FD/IV에서 사용하는 도구변수 z를 식 (4.4)의 독립변수로 사용하여 편향교정항을 추정한다.

<표 2-8>은 표본이탈이 엄밀히 외생적으로 일어난 경우이고, <표 2-9>는 표본이탈이 약하게 외생적으로 일어난 경우이다. 'FD/IV with IMR'과 'FOD/IV with IMR'은 편향교정항을 추가해 추정한 것이다. 각각의 모의실험은 1,000회 반복한다.

엄밀히 외생적인 표본이탈의 경우에는 표본이탈 여부가 Δv 또는 $\ddot{v}$와

〈표 2-8〉 표본이탈이 엄밀하게 외생적인 경우

	$\alpha = 0.2$		$\alpha = 0.7$	
	평균	$n \times$분산	평균	$n \times$분산
FD/IV	0.199	0.473	0.699	2.047
FOD/IV	0.199	0.545	0.698	2.133
FD/IV with IMR	0.199	0.509	0.699	2.409
FOD/IV with IMR	0.198	0.567	0.696	2.254

주 : 1. $n = 500$, $T = 10$, $c_1 = 0$, $c_2 = 0$.
2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Inverse Mills Ratio)를 추가하여 편향을 교정한 것임.

전적으로 무관하므로 IMR을 추가하든 그렇지 않든 추정량에 큰 변화가 없을 것이다. 실제로 <표 2-8>의 모든 추정량들은 표본이탈이 엄밀히 외생적으로 일어났기 때문에 일관적으로 추정된 것을 확인할 수 있다. 하지만 표본이탈이 약하게 외생적으로 일어난 <표 2-9>의 경우, FD/IV은 일관적으로 보이는 반면, FOD/IV은 그렇지 않아 보인다. 표본이탈 여부는 2기 전의 독립변수에 의해 결정되고, 이는 Δv와 관련이 없으므로 Wooldridge의 'FD/IV with IMR'은 FD/IV와 같이 일관적으로 추정되어야 하며, 모의실험 결과도 이를 보여준다. 그러나 FOD/IV와 'FOD/IV with IMR'은 편향된 것으로 보이며, 그 이유는 나중에 설명할 것이다.

다음으로 <표 2-10>은 표본이탈이 내생적으로 일어난 경우에 대하여 모의실험한 결과이다. $c_1 = 0.5$, $c_2 = 0$으로 설정하였다. 내생적 표본이탈의 경우 FD/IV와 FOD/IV 모두 편향이 존재하며, 편향교정항을 추가하면 이 추정량들은 일관적일 것으로 기대된다. 실제 'FD/IV with IMR'은 편향이 교정되어 일관적으로 추정되는 것으로 보인다. 반면, 'FOD/IV with IMR'은 편향이 교정되지 않는다. $\alpha = 0.2$인 경우에는 양의 편향이 관측되며, $\alpha = 0.7$인 경우는 편향이 작아 보인다. $\alpha = 0.5$에 대하여 추가로 실험을 해보았는데, 그 평균은 0.518로 참값보다 약간 더 크게 나왔다. 이러한 양의 편향은 본 소절의 마지막에서 이론적으로도 살펴볼 것이다.

〈표 2-9〉 표본이탈이 약하게 외생적인 경우

	$\alpha = 0.2$		$\alpha = 0.7$	
	평균	$n \times$분산	평균	$n \times$분산
FD/IV	0.200	0.658	0.699	3.360
FOD/IV	0.189	0.781	0.667	3.112
FD/IV with IMR	0.199	0.635	0.699	2.815
FOD/IV with IMR	0.224	0.945	0.708	3.266

주 : 1. $n = 500$, $T = 10$, $c_1 = 0$, $c_2 = 0.5$.
2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Inverse Mills Ratio)을 추가하여 편향을 교정한 것임.

〈표 2-10〉 표본이탈이 내생적(1기 전의 종속변수에 의존)인 경우

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.119	0.479	0.419	2.302
FOD/IV	0.091	0.623	0.455	3.093
FD/IV with IMR	0.198	0.556	0.695	2.734
FOD/IV with IMR	0.217	0.794	0.706	3.251

주 : 1. $n=400,\ T=10,\ c_1=0.5,\ c_2=0$
2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Inverse Mills Ratio)을 추가하여 편향을 교정한 것임.

FOD/IV에 IMR을 추가하여 편향의 교정을 시도하는 방법은 일관적인 것처럼 보일 수 있으나 사실은 그렇지 않다. FD/IV의 방법과 달리 FOD/IV의 방법은 IMR을 포함시킴으로써 편향을 교정할 수 없다. 이하에서는 이러한 현상이 발생하는 원인을 간단한 예제를 통하여 대수적으로 살펴본다.

다음의 간단한 동태적 패널모형을 고려하자.

$$y=\eta_i+\alpha x+v,\ \ x=y_{-1}$$

여기에서 y가 $t=0,1,2,3$에 대하여 관측된다. 그리고 표본관측을 나타내는 변수 a_{it}는 다음에 의하여 결정된다.

$$a=a_{-1}\times I\{\delta y_{-2}+e>0\} \tag{4.8}$$

여기에서 오차항 e는 y_{is} 및 v_{is}와 독립이다. $a=1$이면 y가 관측되고 $a=0$이면 y가 관측되지 않는다. 모든 i에 대하여 $a_{i0}=1$이다. a가 관측되면 $\ddot{y}_{-1}$이 정의되고, 이에 해당하는 $\ddot{v}_{-1}$은 y_{-2}와 무관하다. 그런데 a_{it}는 y_{-2}에 의존하므로, 표본이탈은 약하게 외생적이다.

FOD 변환을 위하여 $C_m=\sqrt{m/(m+1)}$이라 하면,

$$\ddot{v}_{i1}\{a_{i2}=1\}=\ddot{v}_{i1a}\{a_{i2}=1,\ a_{i3}=1\}+\ddot{v}_{i1b}\{a_{i2}=1,\ a_{i3}=0\} \tag{4.9}$$

이며, 여기에서

$$\ddot{v}_{i1a} = C_2[v_{i1} - (1/2)(v_{i2} + v_{i3})],\ \ddot{v}_{i1b} = C_1(v_{i1} - v_{i2})$$

이다(여기에 -1을 곱하여도 좋다). 여기에서 중요한 것은 $\ddot{v}_{i1}$이 a_{i3}의 값에 의존한다는 것이다. (4.9)의 양변에 다음의 조건부 평균을 취하자.

$$E(\ddot{v}_{i1}|a_{i2}=1,\ a_{i3}=1, y_{i0},\ y_{i1}) = E(\ddot{v}_{i1a}|a_{i2}=1,\ a_{i3}=1,\ y_{i0},\ y_{i1}) = C_2 v_{i1}$$
$$E(\ddot{v}_{i1}|a_{i2}=1,\ a_{i3}=0, y_{i0},\ y_{i1}) = E(\ddot{v}_{i1b}|a_{i2}=1,\ a_{i3}=0,\ y_{i0},\ y_{i1}) = C_1 v_{i1}$$

(각 식의 두 번째 항등식은 고정효과 또한 주어진 경우에 성립한다. 여기에서는 그렇다고 가정하였다.) 여기에서 a_{i3}을 평균하여 소거하면,

$$\begin{aligned} E(\ddot{v}_{i1}|a_{i2}=1,\ y_{i0},\ y_{i1}) &= C_2 u_{i1} P(a_{i3}=1|a_{i2}=1, y_{i0}, y_{i1}) \\ &\quad + C_1 u_{i1}[1 - P(a_{i3}=1|a_{i2}=1, y_{i0}, y_{i1})] \\ &= C_2 u_{i1}\Phi(\delta y_{i1}) + C_1 u_{i1}[1 - \Phi(\delta y_{i1})] \\ &= C_1 u_{i1} + (C_2 - C_1) u_{i1}\Phi(\delta y_{i1}) \end{aligned}$$

을 얻는다. $t=1$일 때 y_{i0}을 도구변수로 사용한다면, 추가적으로

$$\begin{aligned} E(\ddot{v}_{i1}|a_{i2}=1, y_{i0}) = &\ C_1 E(u_{i1}|a_{i2}=1, y_{i0}) \\ &+ (C_2 - C_1) E[u_{i1}\Phi(\delta y_{i1})|a_{i2}=1, y_{i0}] \end{aligned}$$

을 얻는다. 이 표현식에서 첫 번째 편향 항은 표본관측 여부 변수가 (4.8)에 따라 생성될 때 0이며, 그렇지 않는 경우라도 IMR로써 교정할 수 있다. 하지만 두 번째 줄에 표현된 편향은 교정할 수 없다. 특히 u_{i1}과 $\Phi(y_{i1})$이 일반적으로 양의 상관관계를 가지고, $C_2 > C_1$이므로, 만일 $\delta > 0$이라면 IMR을 포함시켜 편향교정을 시도한 이후에도 양의 편향이 존재할 것이다. <표 2-9>에서 $\alpha = 0.2$일 때 'FOD/IV with IMR'에 보이는 양의 편향은 이러한 점을 반영하는 것으로 보인다. 만일 C_1과 C_2를 동일하게 두고($C_1 = C_2 = 1$) 고정효과를 제거하는 'forward demeaning'을 할 수도 있을 것이다. 이 방법이 $T=3$일 때에는 작동할지도 모르지만 더 큰 T에서는 문제가 발생하므로 해결책이 되지 못한다.

이상에서, 표본이탈이 선결(predetermined)되어 있는 경우에도 FOD를 사용한 교정은 작동하지 않음을 보았다. 간단히 설명하여, 그 이유는

FOD가 미래의 관측여부에 의존하므로 $\ddot{v}$가 a_{+1}뿐 아니라 a_{+2}, a_{+3}, ... 등에도 의존하기 때문이다. 이 경우, 앞에서 살펴본 것처럼 편향은 추정할 수 없는(또는 추정이 매우 어려운) 미지수에 의존하며, FD의 경우처럼 간명하게 표본이탈을 교정할 수는 없다. 이처럼 FOD를 이용한 표본이탈 교정이 작동하지 않으므로, 표본이탈이 없는 경우처럼 FOD와 FD를 결합하여 효율성을 제고하는 방법도 사용할 수 없다.

4. Full GMM

다음으로 표본이탈 존재 시 Arellano-Bond의 full GMM을 사용하는 방법을 설명한다.

제2절에서 설명한 것처럼 full GMM은 식 (2.2)처럼 차분으로써 고정효과를 제거하고, 식 (2.3)에 표현된 모든 적률조건들을 활용하여 GMM을 하는 것이다. 이때 차분된 t기의 방정식에 사용될 도구변수들 z는 Δv, 즉 $v - v_{-1}$과 상관되지 않은 모든 변수들을 포함한다. 기존의 문헌에서는 상수항에 대하여 특별히 언급하고 있지 않으나, 이 상수항은 중요하며 포함시키는 것이 좋다(Han and Kim, 2014).

이 방법을 표본이탈 교정과 결합시키는 방법으로 다음을 고려한다.

(1) 각 t에서 관측여부(a_{it})를 모든 도구변수들—단순 모형이라면 (1, y_{i0}, y_{i1}, ..., y_{-2})—에 대하여 프로빗(probit) 회귀하여 편향교정항(Inverse Mills Ratio)을 추정한다. 이를 $\hat{\lambda}$라 하자.

(2) 차분한 방정식에 $\hat{\lambda}$와 시간더미 변수의 상호작용항을 설명변수 및 도구변수로 포함시키고 GMM 추정을 한다.

FD/IV에 비하여 full GMM은 상당한 효율성 제고를 가져온다. 표본이탈 편향 교정 시에도 이와 유사한 현상이 관측된다. 통상적인 경우 full GMM의 편향이 상대적으로 큰 것처럼 표본이탈 교정 시에도 FD/IV에 비하여 full GMM의 편향이 더 크다.

이하에서 모의실험 결과를 보고한다.

먼저 편향교정항의 설명변수를 FD/IV의 도구변수와 동일하게 사용한 경우를 살펴본다. 다음으로 편향교정항의 독립변수를 full GMM의 도구변수와 동일하게 사용한 경우를 살펴본다. 전자의 경우는 편향교정항의 독립변수가 매기 동일한 것을 의미하고, 후자의 경우는 편향교정항의 독립변수가 시간에 따라 증가하는 것을 의미한다.

각각의 모의실험은 4.3절에서 분석한 것처럼, 표본이탈이 엄밀히 외생적으로 일어나는 경우, 약하게 외생적으로 일어나는 경우, 내생적으로 일어나는 경우로 나누어 분석한다. FD/IV와 full GMM 모두 고정효과를 차분하므로, 외생적인 표본이탈 및 약하게 외생적인 표본이탈에서는 표본이탈로 인한 편향이 존재하지 않는다. 모의실험은 Stata를 이용하며 각각 200회씩 반복한다.

<표 2-11>, <표 2-12>, <표 2-13>은 FD/IV의 도구변수를 식 (4.4)의 독립변수로 사용하여 분석한 결과이다. <표 2-11>은 엄밀히 외생적인 표본이탈이 일어난 경우이고, <표 2-12>는 약하게 외생적인 표본이탈이 일어난 경우이다. 두 경우 모두 내생적인 표본이탈로 인한 편향이 없기 때문에 편향교정항을 넣기 전과 후의 추정량 사이에는 차이가 없다. 다만 full GMM의 경우 FD/IV보다 편향이 더 큰데, 이는 3.1절에서 설명한 것처럼, 많은 수의 적률조건으로 인한 편향으로 추정된다. 모든 경우에서 full GMM의 분산이 FD/IV의 분산보다 더 작다.

〈표 2-11〉 표본이탈이 엄밀히 외생적인 경우(FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용함)

	$\alpha = 0.2$		$\alpha = 0.7$	
	평균	$n \times$분산	평균	$n \times$분산
FD/IV	0.199	0.484	0.698	2.950
full GMM	0.191	0.326	0.669	0.988
FD/IV with IMR	0.198	0.515	0.696	3.342
full GMM with IMR	0.189	0.343	0.665	1.060

주 : 1. $n = 500$, $T = 10$, $c_1 = 0$, $c_2 = 0$.

2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Inverse Mills Ratio)을 추가하여 편향을 교정한 것임.

〈표 2-12〉 표본이탈이 약하게 외생적인 경우(FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용함)

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.200	0.712	0.693	5.313
full GMM	0.190	0.451	0.661	1.237
FD/IV with IMR	0.200	0.666	0.699	3.450
full GMM with IMR	0.189	0.468	0.663	1.268

주 : 1. $n=500,\ T=10,\ c_1=0,\ c_2=0.5$.
2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Inverse Mills Ratio)을 추가하여 편향을 교정한 것임.

〈표 2-13〉 표본이탈이 내생적인 경우(FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용함)

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.119	0.519	0.379	2.668
full GMM	0.166	0.424	0.621	1.256
FD/IV with IMR	0.197	0.601	0.696	3.295
full GMM with IMR	0.191	0.467	0.674	1.173

주 : 1. $n=500,\ T=10,\ c_1=0.5,\ c_2=0$.
2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Inverse Mills Ratio)을 추가하여 편향을 교정한 것임.

<표 2-13>은 표본이탈이 내생적으로 일어난 경우로 편향교정항 없이 추정 시 편향이 존재한다. 그런데 full GMM의 경우, FD/IV보다 편향으로 인한 손실이 적어 보인다. 앞에서와 마찬가지로 full GMM의 분산이 FD/IV의 분산보다 작다.

<표 2-14>, <표 2-15>, <표 2-16>은 full GMM의 도구변수들을 식(4.4)의 독립변수로 사용하여 분석한 결과이다. <표 2-14>는 엄밀히 외생적인 표본이탈이 일어난 경우, <표 2-15>는 약하게 외생적인 표본이탈이 일어난 경우, <표 2-16>은 내생적인 표본이탈이 일어난 경우이다. 모든 일관적인 추정량에 대해서 full GMM의 편향이 더 크고 분산은

〈표 2-14〉 표본이탈이 엄밀히 외생적인 경우(Full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용함)

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.199	0.484	0.698	2.950
full GMM	0.191	0.326	0.669	0.988
FD/IV with IMR	0.198	0.515	0.696	3.342
full GMM with IMR	0.189	0.354	0.662	1.105

주 : 1. $n=500,\ T=10,\ c_1=0,\ c_2=0.$
2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Iinverse Mills Ratio)을 추가하여 편향을 교정한 것임. FD/IV, full GMM, FD/IV with IMR은 <표 2-11>과 동일함.

〈표 2-15〉 표본이탈이 약하게 외생적인 경우(Full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용함)

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.200	0.712	0.693	5.313
full GMM	0.190	0.451	0.661	1.237
FD/IV with IMR	0.200	0.666	0.699	3.450
full GMM with IMR	0.190	0.472	0.661	1.349

주 : 1. $n=500,\ T=10,\ c_1=0,\ c_2=0.5.$
2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Iinverse Mills Ratio)을 추가하여 편향을 교정한 것임. FD/IV, full GMM, FD/IV with IMR은 <표 2-12>와 동일함.

FD/IV의 분산보다 더 작았다.

표본이탈이 내생적인 경우, 표본관측 방정식의 독립변수를 1개만 사용한 경우와 full GMM의 도구변수와 동일하게 사용한 경우를 비교해 보자. 표본이탈이 내생적으로 발생(1기 전의 종속변수에 의존)할 때, 표본관측 방정식의 독립변수를 FD/IV의 도구변수와 동일하게 사용(매기 독립변수를 1개만 사용)한 결과가 <표 2-13>에 제시되어 있고, full GMM의 도구변수를 독립변수로 사용한 경우가 <표 2-16>에 제시되어

〈표 2-16〉 표본이탈이 내생적인 경우(Full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용함)

	$\alpha=0.2$		$\alpha=0.7$	
	평균	$n\times$분산	평균	$n\times$분산
FD/IV	0.119	0.519	0.379	2.668
full GMM	0.166	0.424	0.621	1.256
FD/IV with IMR	0.197	0.601	0.696	3.295
full GMM with IMR	0.188	0.461	0.661	1.204

주 : 1. $n=500$, $T=10$, $c_1=0.5$, $c_2=0$.
2. FD/IV와 FOD/IV는 전체 자료를 이용한 것이며, 'with IMR'이 추가된 것은 편향교정항(Iinverse Mills Ratio)을 추가하여 편향을 교정한 것임. FD/IV, full GMM, FD/IV with IMR은 <표 2-13>과 동일함.

있다.

$\alpha=0.2$일 때, <표 2-13>에서 'full GMM with IMR'의 평균은 0.191, 표본크기 곱하기 분산은 0.467이고, <표 2-16>에서의 평균은 0.188, 표본크기 곱하기 분산은 0.461이다. 즉 표본관측 방정식의 독립변수 개수가 증가할 때 편향은 약간 커지고 분산은 미미하게 줄어들었다. $\alpha=0.7$일 때, <표 2-13>에서 'full GMM with IMR'의 평균은 0.674, 표본크기 곱하기 분산은 1.173이었음에 반하여, <표 2-16>에서 양자는 각각 0.661과 1.204이다. 즉, 표본관측 방정식에서 최소의 설명변수를 사용한 경우가 모든 설명변수를 사용한 경우보다 편향과 분산이 오히려 더 작았다.

제5절 사업체 패널의 분석

본 절에서는 사업체 패널의 일부를 이용하여, 표본이탈 존재 시 동태적 패널모형 추정법(앞 절에서 논의한 FD/IV, full GMM)을 적용해 본다. Full GMM에서 IMR로써 편향을 교정하는 경우에는 표본관측 방정식의 독립변수를 full GMM의 도구변수와 동일하게 사용한 경우와 그렇

지 않은 경우로 나누어 분석해 본다. 분석에 사용한 Stata 명령어와 그 결과는 본 장의 부록에 수록되어 있다.

자료의 응답자는 사업체이며, 2005년부터 2011년까지 격년(총 4기)으로 구성되어 있다. 1기(2005년)에는 1,615개의 사업체가 관측되었으며, 사업체는 2기부터 표본을 이탈한다. 이탈과 복귀를 반복한 100개의 사업체를 제외한 1,515개 사업체의 자료를 이용해 분석을 실시한다. 분석대상인 1,515개의 사업체 중 528개의 사업체가 한번 표본을 이탈하면 다시 복귀하지 않았다. 시간에 따른 응답자의 수는 <표 2-17>에 제시되어 있으며, 자료에서 관측되면 $a=1$이고, 관측되지 않으면 $a=0$이다. 2기에는 261개의 사업체가 표본을 이탈했고, 3기에는 추가로 136개(=397-261), 4기에는 추가로 131개(=528-397)의 사업체가 표본을 이탈했다. 사업체가 표본을 이탈한 이유는 응답거절과 소멸로 나뉘는데, 이는 <표 2-18>에서 알 수 있다. 표본을 이탈한 528개 사업체 중에서 조사응답을 거절한 사업체는 368개, 소멸한 사업체는 160개이다. 분석에서는 표본이탈의 사유를 구분하지 않으나, 홍민기 외(2014)처럼 응답거절과 소멸을 별도로 분석할 수도 있을 것이다.

종속변수는 생산성의 로그값(ltfp)이며, 주요 독립변수는 전체 근로자 중 정규직 근로자 비율의 로그값(lregratio)이다. 통제변수는 시간더미, 시장경쟁정도(comp),[4] 시장수요상태(demand),[5] 노사관계(relation),[6] 외

〈표 2-17〉 시간에 따른 응답자 수

관측여부	1기	2기	3기	4기	전 체
관 측	1,515	1,254	1,118	987	4,874
표본이탈	0	261	397	528	1,186
전 체	1,515	1,515	1,515	1,515	6,060

4) 시장경쟁정도가 '매우 심하다', '심하다'이면 1의 값을 부여하고, '보통', '약하다', '매우 약하다'이면 0의 값을 부여하여 더미변수로 만들었다.
5) 시장수요상태가 '빨리 늘어남', '늘어나는 편'이면 1의 값을 부여하고, '정체', '줄어드는 편', '빨리 줄어듦'이면 0의 값을 부여하여 더미변수로 만들었다.
6) 노사관계가 '매우 좋음', '좋은 편'이면 1의 값을 부여하고, '보통', '나쁜 편', '매우 나쁨'이면 0의 값을 부여하여 더미변수로 만들었다.

〈표 2-18〉 시간에 따른 표본이탈 이유

이탈이유	2기	3기	4기	전 체
응답거절	163	261	368	792
소 멸	98	136	160	394
전 체	261	397	528	1,186

국인지분비율(foreignratio), 성과급 지급여부(performwage), 이직률(ijik)을 사용했다. 분석에는 종속변수의 1기 전 과거값을 독립변수로 사용하는 동태적 패널모형을 고려한다. 종속변수의 1기 전 과거값, 경영체제, 경영상 해고경험은 약하게 외생적이라고 가정하고, 외국인지분비율, 성과급 지급여부, 이직률은 내생적이라고 가정했다. 여타 시장경쟁정도, 시장수요상태, 노사관계, 노조 존재여부는 엄밀히 외생적이라고 간주한다. 5.1절에서는 표본이탈을 고려하지 않을 때, 균형화한 부분자료와 비균형패널을 이용해 분석한다. 5.2절에서는 표본이탈을 고려할 때, FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용한 결과와 full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용하여 분석한 결과를 비교한다.

1. 표본이탈을 고려하지 않음

먼저 표본이탈을 고려하지 않은 동태적 패널모형을 추정해 본다. <표 2-19>에는 FD/IV 추정법의 (1) 균형화한 부분자료, (2) 비균형패널을 사용한 결과가 제시되어 있다. <표 2-17>에서, 4기 모두 자료에 관측된 응답업체 수는 총 987개이다. 균형화한 부분자료는 4기 모두 자료에서 관측된 응답자만 분석대상으로 고려한다. 4.1절에서 설명한 것처럼, $A_i = I\left(\sum_{t=1}^{4} a = 4\right)$를 a대신 사용한다. 부록에 수록된 Stata 결과에서는 이를 'bal(=balanced subset)'이란 변수로 생성하였다.

<표 2-19>에서 종속변수의 1기 전 과거값(ltfp_{t-1})의 계수를 보면, 균형화한 부분자료에서는 0.1856이고, 비균형패널에서는 0.2259로 모두 5% 유의수준 내에서 유의하다. 주요 독립변수인 정규직 비율의 로그값(lre-

〈표 2-19〉 표본이탈을 고려하지 않은 FD/IV 추정결과

ltfp	(1) 균형화한 부분자료		(2) 비균형패널	
$ltfp_{t-1}$	0.1856	(0.1074)	0.2259	(0.0897)
lregratio	0.7176	(0.4858)	0.9305	(0.4790)

주 : 추가적 설명변수(결과를 보고하지 않음)로는 own, dismissal, foreignratio, performwage, ijik, comp, demand, relation, union이 있음. 약하게 외생적이라 간주한 lregratio, own, dismissal, $ltfp_{t-1}$의 도구변수는 1기 전의 값을 사용하고, 내생적이라 간주한 foreignratio, performwage, ijik의 도구변수는 2기 전의 값을 사용함. 각 분석에는 시간더미와 상수항이 포함되어 있음. 클러스터 표준오차를 사용함. 괄호 안은 표준오차임.

gratio)은 균형화한 부분자료에서는 0.7176, 비균형패널에서는 0.9305의 값을 갖고 모두 5% 유의수준 내에서 유의하지 않다. 종속변수의 1기 전 과거값($ltfp_{t-1}$)을 제외한 나머지 변수들은 모두 5% 유의수준 내에서 유의하지 않다.

다음으로 <표 2-20>에는 full GMM 추정법의 (1) 균형화한 부분자료, (2) 비균형패널을 사용한 결과가 제시되어 있다. 종속변수의 1기 전 과거값($ltfp_{t-1}$)의 계수는 균형화한 부분자료, 비균형패널에서 구한 값들 모두 5% 유의수준 내에서 유의하다. <표 2-19>의 FD/IV 추정결과보다 더 작은 값을 가지며, 표본오차의 값도 더 작다. 정규직 비율의 로그값(lregratio)은 비균형패널을 이용했을 때, 그 값이 0.7625로 5% 유의수준 내에서 유의한 결과가 도출되었다.

〈표 2-20〉 표본이탈을 고려하지 않은 full GMM 추정결과

lftp	(1) 균형화한 부분자료		(2) 비균형패널	
$ltfp_{t-1}$	0.1455	(0.0646)	0.1731	(0.0677)
lregratio	0.3838	(0.3570)	0.7625	(0.4187)

주 : 추가적 설명변수(결과를 보고하지 않음)로는 own, dismissal, foreignratio, performwage, ijik, comp, demand, relation, union이 있음. 외생적이라 간주한 lregratio, own, dismissal, $ltfp_{t-1}$의 도구변수는 첫 기부터 현재의 1기 전까지 과거값을 사용하고, 내생적이라 간주한 foreignratio, performwage, ijik의 도구변수는 첫 기부터 현재의 2기 전까지의 과거값을 사용함. 각 분석에는 시간더미와 상수항이 포함되어 있음. 클러스터 표준오차를 사용함. 괄호 안은 표준오차임.

2. 표본이탈을 고려함

다음으로 표본이탈을 고려하여 분석을 진행한다. 먼저, 4.2절에 제시된 Wooldridge(2002; 2010) 방법을 이용하여 편향을 교정한다. 다음으로 4.4절에 제시된 full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용하여 편향을 교정한다. 전자의 경우는 FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용하는 것으로, 매기 독립변수의 개수가 동일하다. 편향교정항을 만들기 위해서는 매기마다 전기의 $a_{it}=1$인 관측치를 대상으로 프로빗(probit) 회귀한다. 이때 표본관측 방정식의 종속변수는 ltfp를 차분한 값의 존재여부를 나타내는 $a1_{it}$이고, 독립변수는 엄밀히 외생적인 변수들(comp, demand, relation, union)의 1기 전 과거값, 약하게 외생적인 변수들(lregratio, own, dismissal, $ltfp_{t-1}$)의 1기 전 과거값, 내생적인 변수들(foreignratio, performwage, ijik)의 2기 전 과거값을 사용한다. 이상의 방법으로 구한 편향교정항을 추가한 FD/IV와 full GMM의 추정결과가 <표 2-21>에 있다.

<표 2-21>에서 종속변수의 1기 전 과거값($ltfp_{t-1}$)의 계수와 정규직 비율의 로그값(lregratio)의 계수를 비교해 보자. 먼저, 종속변수의 1기 전 과거값($ltfp_{t-1}$)은 FD/IV에서는 0.1823, full GMM에서는 0.1719로 추정되었다. 두 값 모두 5% 유의수준 내에서 유의하며, full GMM의 값이 더 작게 추정되었지만 그 값이 대체로 비슷하다. <표 2-19>에서 비균형패널을 사용한 경우 FD/IV의 값은 0.2259이고, <표 2-20>에서 비균형패널을 사용한 경우 full GMM의 값은 0.1731로, full GMM의 경우 표본이탈을 고려하지 않을 때와 값이 비슷하다.

정규직 비율의 로그값(lregratio)은 FD/IV에서는 1.2594, full GMM에서는 0.8040이 도출되었으며, 전자는 5%, 후자는 10% 유의수준 내에서 유의하다. 여기에서도 full GMM의 경우 표본이탈을 고려하지 않을 때와 값이 비슷하다. <표 2-21>에서 다른 통제변수들은 5% 유의수준 내에서 유의하지 않다.

<표 2-21>의 마지막 행에는 imr3와 imr4가 모두 유의하지 않다는 귀

〈표 2-21〉 FD/IV의 도구변수를 표본관측 방정식의 독립변수로 사용

ltfp	(1) FD/IV		(2) full GMM	
$ltfp_{t-1}$	0.1823	(0.0757)	0.1719	(0.0707)
lregratio	1.2594	(0.5193)	0.8040	(0.4222)
imr3	-0.8046	(3.0580)	-0.0074	(2.6839)
imr4	3.8396	(2.9854)	1.4339	(1.9165)
F 검정	1.69	[0.4306]	0.58	[0.7489]

주 : 추가적 설명변수(결과를 보고하지 않음)로는 own, dismissal, foreignratio, performwage, ijik, comp, demand, relation, union이 있음. FD/IV 추정법에서, 약하게 외생적이라 간주한 lregratio, own, dismissal, $ltfp_{t-1}$의 도구변수는 1기 전의 값을 사용하고, 내생적이라 간주한 foreignratio, performwage, ijik의 도구변수는 2기 전의 값을 사용함. Full GMM의 추정법에서, 약하게 외생적이라 간주한 lregratio, own, dismissal의 도구변수는 첫 기부터 현재의 1기 전까지 과거값을 사용하고, 내생적이라 간주한 foreignratio, performwage, ijik의 도구변수는 첫 기부터 현재의 2기 전까지의 과거값을 사용함. 각 분석에는 시간더미와 상수항이 포함되어 있음. 클러스터 표준오차를 사용함. 괄호 안은 표준오차임. F 검정은 imr3와 imr4가 모두 유의하지 않다는 귀무가설을 검정함(괄호 안은 p-value임).

무가설을 검정한 결과가 제시되어 있다. FD/IV의 경우 F 값이 1.69이고, full GMM의 경우 0.58로, imr3와 imr4가 모두 유의하지 않다는 귀무가설을 기각하지 못한다. 즉 자료에는 내생적 표본이탈로 인한 편향이 존재하지 않는다.

<표 2-22>에는 full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용하여 편향을 교정한 방법의 FD/IV와 full GMM 추정법이 제시되어 있다. 편향교정항을 만들기 위해서는 매기마다 전기의 $a_{it}=1$인 관측치를 대상으로 프로빗(probit) 회귀한다. 표본관측 방정식의 종속변수는 앞서 제시한 것처럼, ltfp를 차분한 값의 존재여부를 나타내는 $a1_{it}$이다. 독립변수는 엄밀히 외생적인 변수들(comp, demand, relation, union)과 약하게 외생적인 변수들(lregratio, own, dismissal)의 첫 기부터 1기 전까지의 과거값, 종속변수의 1기 전 과거값($ltfp_{t-1}$)과 내생적인 변수들(foreignratio, performwage, ijik)의 첫 기부터 2기 전까지의 과거값을 사용한다. <표 2-22>에서 종속변수의 1기 전 과거값($ltfp_{t-1}$)의 계수와

〈표 2-22〉 Full GMM의 도구변수를 표본관측 방정식의 독립변수로 사용

ltfp	(1) FD/IV		(2) full GMM	
$ltfp_{t-1}$	0.1615	(0.0775)	0.1479	(0.0659)
lregratio	1.2233	(0.5298)	0.7096	(0.4392)
imr3	-0.4779	(2.9564)	0.3001	(2.6222)
imr4	1.9201	(2.8682)	-0.5901	(1.8390)
F 검정	0.49	[0.7817]	0.11	[0.9461]

주 : 추가적 설명변수(결과를 보고하지 않음)로는 own, dismissal, foreignratio, performwage, ijik, comp, demand, relation, union이 있음. FD/IV 추정법에서는 약하게 외생적이라 간주한 lregratio, own, dismissal, $ltfp_{t-1}$의 도구변수는 1기 전의 값을 사용하고, 내생적이라 간주한 foreignratio, performwage, ijik의 도구변수는 2기 전의 값을 사용함. Full GMM의 추정법에서는 약하게 외생적이라 간주한 lregratio, own, dismissal, $ltfp_{t-1}$의 도구변수는 첫 기부터 현재의 1기 전까지 과거값을 사용하고, 내생적이라 간주한 foreignratio, performwage, ijik의 도구변수는 첫 기부터 현재의 2기 전까지의 과거값을 사용함. 각 분석에는 시간더미와 상수항이 포함되어 있음. 클러스터 표준오차를 사용함. 괄호 안은 표준오차임. F 검정은 imr3와 imr4가 모두 유의하지 않다는 귀무가설을 검정함(괄호 안은 p-value임).

정규직 비율의 로그값(lregratio)의 계수를 비교해 보자. 종속변수의 1기 전 과거값($ltfp_{t-1}$)은 FD/IV에서는 0.1615, full GMM에서는 0.1479로 추정되었으며, 모두 5% 유의수준 내에서 유의하다.

<표 2-21>과 <표 2-22>의 full GMM의 종속변수의 1기 전 과거값($ltfp_{t-1}$)을 비교해 보면, 표본관측 방정식의 독립변수를 full GMM의 도구변수보다 적게 사용한 <표 2-21>에서의 계수값은 0.1719, 표준오차는 0.0707로 도출되었고, 자신의 도구변수를 표본관측 방정식의 독립변수로 사용한 <표 2-22>에서의 계수값은 0.1479, 표준오차는 0.0659로 도출되었다.

정규직 비율의 로그값(lregratio)의 경우, FD/IV에서는 1.2233이 추정되었고 이는 5% 유의수준 내에서 유의하다. 그러나 full GMM의 경우 정규직 비율의 로그값(lregratio)은 5% 유의수준 내에서 유의하지 않다.

<표 2-22>의 마지막 행에는 imr3와 imr4가 모두 유의하지 않다는 귀무가설을 검정한 결과가 제시되어 있는데, 앞에서와 마찬가지로 5% 유

〈표 2-23〉 계수값 비교

	$ltfp_{t-1}$		lregratio	
균형패널 FD/IV	0.1856	(0.1074)	0.7176	(0.4858)
균형패널 full GMM	0.1455	(0.0646)	0.3838	(0.3570)
비균형패널 FD/IV	0.2259	(0.0897)	0.9305	(0.4790)
비균형패널 full GMM	0.1731	(0.0677)	0.7625	(0.4187)
FD/IV with IMR1	0.1823	(0.0757)	1.2594	(0.5193)
full GMM with IMR1	0.1719	(0.0707)	0.8040	(0.4222)
full GMM with IMRfull	0.1479	(0.0659)	0.7096	(0.4392)

주 : 괄호 안은 클러스터 표준오차임. IMR1은 표본관측 방정식의 독립변수를 FD/IV의 도구변수를 사용하여 IMR을 추정한 것이고, IMRfull은 full GMM의 도구변수를 사용하여 IMR을 추정한 것임.

의수준 내에서 귀무가설을 기각하지 못한다.

<표 2-23>에는 종속변수의 1기 전 과거값($ltfp_{t-1}$)과 정규직 비율의 로그값(lregratio)의 계수값이 비교되어 있다. 여기에서 표본관측 방정식의 독립변수를 full GMM의 도구변수로 사용한 'full GMM with IMRfull'의 정규직 비율의 로그값(lregratio)의 계수는 5% 유의수준 내에서 유의하지 않다. 균형패널, 비균형패널 및 동일한 편향교정항을 사용한 분석에서, 대체적으로 full GMM으로 추정한 계수값이 FD/IV의 계수값보다 작고 표준오차도 더 작다.

제6절 소 결

본 장에서는 동태적 패널자료 모형에서 표본이탈로 인한 편향의 교정방법들에 대하여 살펴보았다. 표본이탈이 없을 때 간편한 FD/IV, FOD/IV의 방법 및 이 둘을 GMM으로 결합한 방법(FD-FOD)은 full GMM보다 덜 효율적이나 편향을 크게 줄인다. FD-FOD의 방법은 FD/ IV만큼 편향이 작으며, full GMM보다는 덜 효율적이나 FD/IV에 비하여 상당한

효율성을 지닌다.

표본이탈이 존재할 때, 표본이탈이 엄밀히 외생적인 경우를 제외하면, 균형화한 부분패널을 사용하면 편향이 발생한다. $t+1$기의 표본이탈이 엄밀히 외생적이지는 않지만, 외생변수가 주어질 때 t기 및 $t-1$기의 내생변수값과 무관하면, 모든 패널자료(비균형패널)를 활용한 추정은 일관적이다. 그러나 만일 $t+1$기의 표본이탈 여부가 t기 또는 $t-1$기의 종속변수값에 의존하면 어느 경우에나 편향이 발생하며, 이때 모든 자료(비균형패널)와 균형화된 부분자료(balanced subset) 중 어느 쪽이 작은 편향을 주는지는 일정하지 않다.

$t+1$기의 표본이탈이 $t+1$기, t기, $t-1$기의 종속변수에 의존할 때 편향을 교정하여야 한다. 차분 이후 도구변수를 사용하는 간편한 방법에 대해서는 Wooldridge(2010)가 설명하고 있다. 표본이탈이 없는 경우에 사용할 수 있었던 FOD/IV의 방법은, FOD 변수가 미래의 표본이탈에 의존하므로, 제거할 수 없는 편향을 도입하며, 따라서 이 편향을 제거하는 방법이 발견되지 않는 한 사용이 불가능하다. 따라서 FD와 FOD를 결합하여 효율성을 제고하는 방법도 사용이 불가능하다.

FD 이후 모든 도구변수들을 사용하여 full GMM을 하는 방법은 이 경우에도 사용 가능하며, 모의실험에 따르면 현저한 효율성 제고를 가져오는 것으로 보인다.

표본이탈 편향교정 방법을 사업체 패널의 일부에 응용하여 보았다. Stata 명령어는 부록에 제시되어 있다. 본 분석에서는 표본이탈의 원인(응답거절과 소멸)은 구분하지 않고 분석하였다. 좀더 엄밀한 분석을 위해서는 이 두 원인에 따른 표본이탈을 따로 고려하는 것이 적절해 보인다. 엄밀한 분석은 실제 응용연구의 주제로 남겨 둔다.

〔부록〕 Stata 실행 결과

```
. xtset id tt
       panel variable:  id (unbalanced)
        time variable:  tt, 1 to 4
                delta:  1 unit


. // Generate attrition indicators: a and bal(=balanced subset)
. gen reenter0 = .
(7798 missing values generated)

. replace reenter0 = 1 if response > f.response
(100 real changes made)

. by id: egen reenter = sum(reenter0)

. replace reenter = . if response==.
(1338 real changes made, 1338 to missing)

. drop reenter0

. gen a = .
(7798 missing values generated)

. replace a = response if reenter==0
(6060 real changes made)

. replace a = 0 if a==2 | a==3
(1186 real changes made)

. by id: egen suma = sum(a)

. gen bal = 1 if suma==4
(3850 missing values generated)

. replace bal = 0 if a~=. & suma~=4
(2112 real changes made)

. drop suma
```

```
. tab a tt

           |                    tt
         a |         1         2         3         4 |     Total
-----------+--------------------------------------------+---------
         0 |         0       261       397       528 |     1,186
         1 |     1,515     1,254     1,118       987 |     4,874
-----------+--------------------------------------------+---------
     Total |     1,515     1,515     1,515     1,515 |     6,060


. tab response tt if a==0

           |                tt
  response |         2         3         4 |     Total
-----------+---------------------------------+---------
         2 |       163       261       368 |       792
         3 |        98       136       160 |       394
-----------+---------------------------------+---------
     Total |       261       397       528 |     1,186


. // Generate other variables
. rename tfp_valadd_lp_incB_sm ltfp

. rename valadd_per_imp_ln_L valadd

. rename reg_ratio_ln lregratio

. gen own = .
(7798 missing values generated)

. replace own = 1 if ownsystem1==1 | ownsystem2==1
(3937 real changes made)

. replace own = 0 if ownsystem3==1 | ownsystem4==1 | ownsystem5==1
(2562 real changes made)

. gen comp = .
(7798 missing values generated)

. replace comp = 1 if competition1==1 | competition2==1
```

```
(4701 real changes made)

. replace comp = 0 if competition3==1 | competition4==1 | competition5==1
(1798 real changes made)

. gen demand = .
(7798 missing values generated)

. replace demand = 1 if demand1==1 | demand2==1
(2593 real changes made)

. replace demand = 0 if demand3==1 | demand4==1 | demand5==1
(3906 real changes made)

. gen relation = .
(7798 missing values generated)

. replace relation = 1 if relation4==1 | relation5==1
(4567 real changes made)

. replace relation = 0 if relation1==1 | relation2==1 | relation3==1
(1932 real changes made)

. global y "ltfp"

. global exo "comp demand relation union"

. global pre "lregratio own dismissal"

. global endo "foreignratio performwage ijik"


. // Without IMR
. ivregress 2sls d.$y d.($exo) (d.($pre l.$y $endo) = l.($pre) l2.($y $endo)) i.tt if
bal==1, vce(cl id)
note: 1.tt identifies no observations in the sample
note: 2.tt identifies no observations in the sample
note: 4.tt omitted because of collinearity

Instrumental variables (2SLS) regression          Number of obs =     1972
                                                  Wald chi2(12) =     9.49
                                                  Prob > chi2   =   0.6605
```

R-squared = .
Root MSE = 4.9373

(Std. Err. adjusted for 987 clusters in id)

D.ltfp	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
lregratio						
D1.	.7176316	.4858031	1.48	0.140	-.2345249	1.669788
own						
D1.	.3909385	.4660577	0.84	0.402	-.5225178	1.304395
dismissal						
D1.	-.0007572	.6490916	-0.00	0.999	-1.272953	1.271439
ltfp						
LD.	.1855715	.1074261	1.73	0.084	-.0249797	.3961228
foreignratio						
D1.	.026614	.0436253	0.61	0.542	-.05889	.1121179
performwage						
D1.	5.770848	8.764716	0.66	0.510	-11.40768	22.94938
ijik						
D1.	-2.381989	8.31474	-0.29	0.775	-18.67858	13.9146
comp						
D1.	-.0845175	.1963512	-0.43	0.667	-.4693587	.3003238
demand						
D1.	-.3057655	.5977492	-0.51	0.609	-1.477332	.8658015
relation						
D1.	-.1718772	.5064395	-0.34	0.734	-1.16448	.8207261
union						
D1.	-.636949	.977833	-0.65	0.515	-2.553467	1.279568
tt						

```
           1 |          0  (empty)
           2 |          0  (empty)
           3 |   1.561369    2.2548     0.69   0.489    -2.857958    5.980695
           4 |          0  (omitted)
             |
       _cons |  -.9315941   1.304959    -0.71   0.475    -3.489267    1.626079
------------------------------------------------------------------------------
Instrumented:  D.lregratio D.own D.dismissal LD.ltfp D.foreignratio
               D.performwage D.ijk
Instruments:   D.comp D.demand D.relation D.union 3.tt L.lregratio L.own
               L.dismissal L2.ltfp L2.foreignratio L2.performwage L2.ijk


. ivregress 2sls d.$y d.($exo) (d.($pre 1.$y $endo) = l.($pre) l2.($y $endo)) i.tt,
vce(cl id)
note: 1.tt identifies no observations in the sample
note: 2.tt identifies no observations in the sample
note: 4.tt omitted because of collinearity

Instrumental variables (2SLS) regression        Number of obs =    2317
                                                Wald chi2(12) =   13.54
                                                Prob > chi2   =  0.3311
                                                R-squared     =       .
                                                Root MSE      =  4.9854

                                  (Std. Err. adjusted for 1332 clusters in id)
------------------------------------------------------------------------------
             |               Robust
      D.ltfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  lregratio  |
         D1. |    .930452   .4789827     1.94   0.052    -.0083368    1.869241
             |
         own |
         D1. |   .2484557   .4129464     0.60   0.547    -.5609043    1.057816
             |
  dismissal  |
         D1. |  -.1062962   .5209792    -0.20   0.838    -1.127397    .9148043
             |
       ltfp  |
         LD. |   .2258958   .0896727     2.52   0.012     .0501405    .4016511
             |
```

```
foreignratio |
         D1. |   .0470845   .0514734     0.91   0.360    -.0538016    .1479706
             |
 performwage |
         D1. |   5.847677   5.844325     1.00   0.317    -5.606989    17.30234
             |
        ijik |
         D1. |  -.2996832   12.05784    -0.02   0.980    -23.93261    23.33324
             |
        comp |
         D1. |  -.0296387   .3033503    -0.10   0.922    -.6241944     .564917
             |
      demand |
         D1. |  -.2647548   .3754318    -0.71   0.481    -1.000588    .4710781
             |
    relation |
         D1. |  -.0852416   .4055735    -0.21   0.834    -.8801511    .7096679
             |
       union |
         D1. |  -.7283657   .9962417    -0.73   0.465    -2.680964    1.224232
             |
          tt |
           1 |           0  (empty)
           2 |           0  (empty)
           3 |    1.687039   1.746654     0.97   0.334    -1.736339    5.110417
           4 |           0  (omitted)
             |
       _cons |  -.9298828   .9710807    -0.96   0.338    -2.833166    .9734003
------------------------------------------------------------------------------
Instrumented:  D.lregratio D.own D.dismissal LD.ltfp D.foreignratio
               D.performwage D.ijik
Instruments:   D.comp D.demand D.relation D.union 3.tt L.lregratio L.own
               L.dismissal L2.ltfp L2.foreignratio L2.performwage L2.ijik
```

```
. xtabond $y $exo tdummy1-tdummy4 if bal==1, lags(1) pre($pre) endo($endo) vce(r)
note: tdummy1 dropped from div() because of collinearity
note: tdummy1 dropped because of collinearity
note: tdummy3 dropped because of collinearity
Arellano-Bond dynamic panel-data estimation  Number of obs      =      1972
Group variable: id                           Number of groups   =       987
Time variable: tt
```

```
                                        Obs per group:  min =         1
                                                        avg =  1.997974
                                                        max =         2

Number of instruments =     34          Wald chi2(13)       =     17.87
                                        Prob > chi2         =    0.1625
One-step results
                                (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
        ltfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ltfp |
         L1. |    .145487   .0646337     2.25   0.024     .0188074    .2721666
             |
   lregratio |    .383825   .3570295     1.08   0.282      -.31594     1.08359
         own |   .3137595   .2985698     1.05   0.293    -.2714266    .8989456
   dismissal |    .304619   .3415432     0.89   0.372    -.3647933    .9740313
foreignratio |   .0132546   .0239396     0.55   0.580    -.0336662    .0601754
 performwage |   .7396518   .9221325     0.80   0.422    -1.067695    2.546998
        ijik |  -2.133165   1.737454    -1.23   0.220    -5.538513    1.272182
        comp |  -.0662793   .1104402    -0.60   0.548     -.282738    .1501795
      demand |    .011265   .1815761     0.06   0.951    -.3446176    .3671477
    relation |    .080375   .1740394     0.46   0.644    -.2607359     .421486
       union |  -.3408238    .25263    -1.35   0.177    -.8359695    .1543218
     tdummy2 |  -.0824439    .187042    -0.44   0.659    -.4490394    .2841516
     tdummy4 |  -.1803518   .1963568    -0.92   0.358    -.5652041    .2045004
       _cons |    2.71249    .623912     4.35   0.000     1.489645    3.935335
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).ltfp L(1/.).lregratio L(1/.).own L(1/.).dismissal
L(2/.).foreignratio
                  L(2/.).performwage L(2/.).ijik
        Standard: D.comp D.demand D.relation D.union D.tdummy2 D.tdummy3 D.tdummy4
Instruments for level equation
        Standard: _cons


. xtabond $y $exo tdummy1-tdummy4, lags(1) pre($pre) endo($endo) vce(r)
note: tdummy4 dropped from div() because of collinearity
note: tdummy1 dropped because of collinearity
note: tdummy2 dropped because of collinearity
```

```
Arellano-Bond dynamic panel-data estimation  Number of obs      =      2317
Group variable: id                           Number of groups   =      1332
Time variable: tt
                                             Obs per group: min =         1
                                                            avg =  1.739489
                                                            max =         2

Number of instruments =     34               Wald chi2(13)      =     20.01
                                             Prob > chi2        =    0.0949
One-step results
                                  (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
        ltfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ltfp |
         L1. |   .1730985   .0676675     2.56   0.011     .0404726    .3057244
             |
   lregratio |   .7624961   .4186618     1.82   0.069     -.058066    1.583058
         own |   .2412155   .2835823     0.85   0.395     -.3145956   .7970266
  dismissal  |   .1317381   .2946444     0.45   0.655     -.4457543   .7092306
foreignratio |   .0421513   .0315784     1.33   0.182     -.0197412   .1040438
 performwage |   .6014988   .9277193     0.65   0.517     -1.216798   2.419795
        ijik |  -.7729694   1.946524    -0.40   0.691     -4.588087   3.042148
        comp |  -.0283009   .1197455    -0.24   0.813     -.2629978   .2063959
      demand |   .0305975   .1650764     0.19   0.853     -.2929462   .3541412
    relation |   .1709686   .1659379     1.03   0.303     -.1542637   .4962009
       union |  -.3234634   .2422687    -1.34   0.182     -.7983013   .1513744
    tdummy3  |   .1407281   .1922509     0.73   0.464     -.2360767   .5175328
    tdummy4  |  -.0058008    .130548    -0.04   0.965     -.2616702   .2500686
       _cons |    2.10403   .7938431     2.65   0.008     .5481261    3.659934
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type:  L(2/.).ltfp  L(1/.).lregratio  L(1/.).own  L(1/.).dismissal
L(2/.).foreignratio
                   L(2/.).performwage L(2/.).ijik
        Standard: D.comp D.demand D.relation D.union D.tdummy1 D.tdummy2 D.tdummy3
Instruments for level equation
        Standard: _cons


. // With IMR(using FD/IV)
```

```
. gen a1 = 1

. replace a1 = 0 if d.$y==.
(3732 real changes made)

. forv t=3/4 {
  2.    qui probit a1 l.($exo $pre) l2.($y $endo) if l.a==1 & tt==`t'
  3.    gen a_imr`t' = 0
  4.    capture drop xb
  5.    predict xb if tt==`t', xb
  6.    replace a_imr`t' = (normalden(xb)/normal(xb))*a1 if tt==`t'
  7. }
(6514 missing values generated)
(1886 real changes made, 770 to missing)
(6466 missing values generated)
(2063 real changes made, 862 to missing)

. capture drop xb


. ivregress 2sls d.$y d.($exo) (d.($pre l.$y $endo) = l.($pre) l2.($y $endo)) a_imr*
i.tt, vce(cl id)
note: 1.tt identifies no observations in the sample
note: 2.tt identifies no observations in the sample
note: 4.tt omitted because of collinearity

Instrumental variables (2SLS) regression        Number of obs =    2317
                                                Wald chi2(14) =   17.67
                                                Prob > chi2   =  0.2221
                                                R-squared     =       .
                                                Root MSE      =  4.2802

                                (Std. Err. adjusted for 1332 clusters in id)
------------------------------------------------------------------------------
             |               Robust
     D.ltfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
  lregratio |
        D1. |   1.259417   .5192626     2.43   0.015     .2416812    2.277153
            |
        own |
        D1. |   .4839533   .4263586     1.14   0.256    -.3516941    1.319601
```

dismissal						
D1.	.0716282	.418449	0.17	0.864	-.7485168	.8917732
ltfp						
LD.	.182342	.0757244	2.41	0.016	.0339249	.3307592
foreignratio						
D1.	.0256932	.0447259	0.57	0.566	-.061968	.1133545
performwage						
D1.	2.725833	2.865516	0.95	0.341	-2.890476	8.342142
ijik						
D1.	3.123014	9.193425	0.34	0.734	-14.89577	21.1418
comp						
D1.	.054918	.2514947	0.22	0.827	-.4380024	.5478385
demand						
D1.	-.0946781	.2288658	-0.41	0.679	-.5432469	.3538907
relation						
D1.	.1389667	.2468427	0.56	0.573	-.3448362	.6227695
union						
D1.	-.2578444	.5778644	-0.45	0.655	-1.390438	.874749
a_imr3	-.8045888	3.057996	-0.26	0.792	-6.798151	5.188974
a_imr4	3.839589	2.985375	1.29	0.198	-2.011638	9.690816
tt						
1	0	(empty)				
2	0	(empty)				
3	2.10983	1.709419	1.23	0.217	-1.240568	5.460229
4	0	(omitted)				
_cons	-1.44277	1.146559	-1.26	0.208	-3.689985	.8044451

Instrumented: D.lregratio D.own D.dismissal LD.ltfp D.foreignratio D.performwage D.ijik
Instruments: D.comp D.demand D.relation D.union a_imr3 a_imr4 3.tt

```
              L.lregratio L.own L.dismissal L2.ltfp L2.foreignratio
              L2.performwage L2.ijik


. test a_imr3 a_imr4

 ( 1)  a_imr3 = 0
 ( 2)  a_imr4 = 0

           chi2(  2) =    1.69
         Prob > chi2 =    0.4306


. xtabond $y $exo tdummy1-tdummy4, lags(1) pre($pre) endo($endo) diffvars(a_imr*) vce(r)
nocons
note: tdummy1 dropped from div() because of collinearity
note: tdummy1 dropped because of collinearity

Arellano-Bond dynamic panel-data estimation  Number of obs      =      2317
Group variable: id                           Number of groups   =      1332
Time variable: tt
                                             Obs per group: min =         1
                                                            avg =  1.739489
                                                            max =         2

Number of instruments =     35               Wald chi2(13)      =     17.68
                                             Prob > chi2        =    0.1701
One-step results
                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
        ltfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ltfp |
         L1. |   .1718582   .0706841     2.43   0.015     .0333198    .3103965
             |
  lregratio  |   .8039855   .4222387     1.90   0.057    -.0235872    1.631558
         own |   .2980362   .2978231     1.00   0.317    -.2856864    .8817589
   dismissal |   .1479501   .2957438     0.50   0.617     -.431697    .7275973
foreignratio |   .0352369   .0357975     0.98   0.325    -.0349249    .1053987
 performwage |   .7088859   .9440899     0.75   0.453    -1.141496    2.559268
        ijik |  -1.008906    1.88238    -0.54   0.592    -4.698303    2.680492
```

```
        comp |  -.0261907   .1215205    -0.22   0.829    -.2643665    .2119852
      demand |   .0045513   .1628204     0.03   0.978    -.3145708    .3236734
    relation |   .1663938   .1667759     1.00   0.318    -.1604809    .4932686
       union |   -.349445   .2451091    -1.43   0.154    -.8298501    .1309601
     tdummy2 |   .3356642   .5898725     0.57   0.569    -.8204647    1.491793
     tdummy3 |   .4663034    .467222     1.00   0.318    -.4494348    1.382042
     tdummy4 |            0  (omitted)
      a_imr3 |  -.0074246   2.683923    -0.00   0.998    -5.267817    5.252968
      a_imr4 |   1.433892   1.916497     0.75   0.454    -2.322374    5.190158
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type:  L(2/.).ltfp  L(1/.).lregratio  L(1/.).own  L(1/.).dismissal
L(2/.).foreignratio
                   L(2/.).performwage L(2/.).ijik
        Standard: D.comp D.demand D.relation D.union D.tdummy2 D.tdummy3 D.tdummy4
a_imr3 a_imr4


. test a_imr3 a_imr4

 ( 1)  a_imr3 = 0
 ( 2)  a_imr4 = 0

           chi2(  2) =    0.58
         Prob > chi2 =    0.7489


. // With IMR(using full GMM)
. qui probit a1 l.($exo $pre) l2.($y $endo) if l.a==1 & tt==3

. gen b_imr3 = 0

. predict xb if tt==3, xb
(6514 missing values generated)

. replace b_imr3 = (normalden(xb)/normal(xb))*a1 if tt==3
(1886 real changes made, 770 to missing)

. capture drop xb


. qui probit a1 l.($exo $pre) l2.($exo $pre $y $endo) l3.($y $endo) if l.a==1 & tt==4
```

```
. gen b_imr4 = 0

. predict xb if tt==4, xb
(6682 missing values generated)

. replace b_imr4 = (normalden(xb)/normal(xb))*a1 if tt==4
(2063 real changes made, 1078 to missing)

. capture drop xb


. ivregress 2sls d.$y d.($exo) (d.($pre 1.$y $endo) = l.($pre) l2.($y $endo)) b_imr*
i.tt, vce(cl id)
note: 1.tt identifies no observations in the sample
note: 2.tt identifies no observations in the sample
note: 4.tt omitted because of collinearity

Instrumental variables (2SLS) regression        Number of obs =    2101
                                                Wald chi2(14) =   16.03
                                                Prob > chi2   =  0.3115
                                                R-squared     =       .
                                                Root MSE      =  4.1514

                                  (Std. Err. adjusted for 1116 clusters in id)
------------------------------------------------------------------------------
             |               Robust
     D.ltfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
  lregratio |
        D1. |   1.223348   .5297515     2.31   0.021      .185054    2.261642
            |
        own |
        D1. |   .4748458   .4248676     1.12   0.264    -.3578794    1.307571
            |
  dismissal |
        D1. |   .0900399   .4564106     0.20   0.844    -.8045084    .9845882
            |
      ltfp |
        LD. |   .1614598   .0774832     2.08   0.037     .0095955    .3133241
            |
foreignratio |
        D1. |   .0334019   .0466166     0.72   0.474    -.0579649    .1247687
```

```
             |
performwage  |
         D1. |   3.414525   3.425043     1.00   0.319    -3.298436    10.12749
             |
        ijik |
         D1. |  -1.172768   8.736939    -0.13   0.893    -18.29685    15.95132
             |
        comp |
         D1. |  -.1102863   .2247703    -0.49   0.624     -.550828    .3302553
             |
      demand |
         D1. |  -.1144727   .2772644    -0.41   0.680    -.6579009    .4289555
             |
    relation |
         D1. |  -.0218888   .2849124    -0.08   0.939    -.5803069    .5365293
             |
       union |
         D1. |  -.4159559   .6188205    -0.67   0.501    -1.628822    .7969101
             |
      b_imr3 |  -.4778888   2.956433    -0.16   0.872    -6.272392    5.316614
      b_imr4 |   1.920129    2.86821     0.67   0.503     -3.70146    7.541718
             |
          tt |
           1 |           0  (empty)
           2 |           0  (empty)
           3 |    1.511401   1.661396     0.91   0.363    -1.744876    4.767678
           4 |           0  (omitted)
             |
       _cons |  -1.008477   1.143223    -0.88   0.378    -3.249154    1.232199
------------------------------------------------------------------------------
Instrumented:  D.lregratio D.own D.dismissal LD.ltfp D.foreignratio
               D.performwage D.ijik
Instruments:   D.comp D.demand D.relation D.union b_imr3 b_imr4 3.tt
               L.lregratio L.own L.dismissal L2.ltfp L2.foreignratio
               L2.performwage L2.ijik


. test b_imr3 b_imr4

 ( 1)  b_imr3 = 0
 ( 2)  b_imr4 = 0
```

```
          chi2(  2) =    0.49
        Prob > chi2 =    0.7817

. xtabond $y $exo tdummy1-tdummy4, lags(1) pre($pre) endo($endo) diffvars(b_imr*) vce(r)
nocons
note: tdummy1 dropped from div() because of collinearity
note: tdummy1 dropped because of collinearity

Arellano-Bond dynamic panel-data estimation  Number of obs      =      2101
Group variable: id                           Number of groups   =      1116
Time variable: tt
                                             Obs per group: min =         1
                                                            avg =  1.882616
                                                            max =         2

Number of instruments =     35               Wald chi2(13)      =     14.31
                                             Prob > chi2        =    0.3525
One-step results
                                    (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
        ltfp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ltfp |
         L1. |   .1478972   .0659195     2.24   0.025     .0186974    .2770969
             |
    lregratio |   .7096442   .4391729     1.62   0.106    -.1511188    1.570407
         own |   .2184929   .3081945     0.71   0.478    -.3855573    .8225431
   dismissal |   .1283182   .3228336     0.40   0.691     -.504424    .7610604
foreignratio |    .035724   .0352153     1.01   0.310    -.0332967    .1047447
 performwage |   1.081453   .8948113     1.21   0.227    -.6723446    2.835251
        ijik |  -.9340839    1.80534    -0.52   0.605    -4.472485    2.604317
        comp |  -.0848754   .1177296    -0.72   0.471    -.3156212    .1458704
      demand |   .0152972   .1764819     0.09   0.931    -.3306009    .3611953
    relation |   .0901066   .1791542     0.50   0.615    -.2610292    .4412424
       union |  -.2957054   .2414729    -1.22   0.221    -.7689835    .1775727
     tdummy2 |           0  (omitted)
     tdummy3 |   .1350852     .53392     0.25   0.800    -.9113787    1.181549
     tdummy4 |   .0018101    .603746     0.00   0.998     -1.18151    1.185131
      b_imr3 |   .3001371   2.622226     0.11   0.909    -4.839332    5.439606
      b_imr4 |  -.5901189   1.839042    -0.32   0.748    -4.194575    3.014337
------------------------------------------------------------------------------
```

```
Instruments for differenced equation
        GMM-type: L(2/.).ltfp L(1/.).lregratio L(1/.).own L(1/.).dismissal
L(2/.).foreignratio
                  L(2/.).performwage L(2/.).ijik
        Standard: D.comp D.demand D.relation D.union D.tdummy2 D.tdummy3 D.tdummy4
b_imr3 b_imr4


. test b_imr3 b_imr4

 ( 1)  b_imr3 = 0
 ( 2)  b_imr4 = 0

           chi2(  2) =    0.11
         Prob > chi2 =    0.9461
```

제 3 장
Paradata를 이용한 무응답 자료 회귀분석

제1절 서 론

Paradata는 조사의 목적과는 별도로 조사과정에서 발생한 정보를 수집한 자료로서 조사의 질에 대해 좋은 정보를 제공한다. Paradata를 이용하여 조사의 품질을 평가하고 또 추정을 개선하고자 하는 노력은 최근의 조사 방법론의 주요 연구 주제이기도 하다. Couper and Lyberg (2005)은 paradata와 관련된 주요 개념을 정립하였고, Kreuter et al. (2010)은 paradata를 이용하여 무응답 보정을 하는 방법에 대한 연구를 하였고, Durrant et al.(2011)은 paradata를 이용하여 조사 시기를 결정하는 방법과 관련한 연구를 하였으며, Wagner et al.(2012)는 paradata를 이용하여 실사과정을 모니터하고 이를 이용하여 'responsive design'을 하는 방법론을 제안하였다. 또한, 무응답 가구에 대한 followup 조사를 하는 경우에는 방문횟수가 paradata가 될 수 있다. 이와 관련하여 Wood et al.(2006)과 Kim and Im(2014)은 followup 방문 횟수를 무응답 모형에 사용하여 얻어지는 무응답 보정 방법론에 대한 연구를 하였다.

사업체 패널조사에서는 '첫 컨택 반응'변수와 '응답시간'이 이러한 Paradata로 간주될 수 있다. '첫 컨택 반응'변수는 응답률 자체에 대한 정보를 제공하고 '응답시간'은 응답의 질, 특히 measurement error 여부에 대한 정보를 제공한다고 할 수 있다. 무응답 성향 점수를 사용하여 가중치

조정을 할 때 이러한 paradata를 적절히 사용하여 무응답 성향을 더 정확하게 추정하고 이를 바탕으로 사업체 패널조사를 통해 얻는 통계의 질을 높이는 것이 본 연구의 목표이다. 또한 이와 관련된 이론 연구를 통해서 조사 통계 분야의 추정방법론에 기여하고자 한다.

사업체 패널조사에서 얻는 '첫 컨택 반응'변수는 다음 중 하나의 값을 가진다.

① 우호적 반응 : 조사자료를 달라고 하기, 사전 설문에 응답, 방문 날짜 정함 등
② 재연락 : 지금 당장은 못 하다 다시 연락을 달라고 함
③ 내부 결재 중 : 회사 차원에서 검토하여 하겠다고 함
④ 매우 바쁨 : 바쁘다는 이야기만 함
⑤ 부정적 반응 : 여러 가지 사유로 조사에 응하지 않겠다고 함

매우 바쁨이나 부정적 반응의 경우에는 무응답 비율이 높아지고, 이는 향후 관심 모수의 추정에 어려움을 가져다준다. 본 연구에서는 산업별 기업 규모에 따른 이윤 분포에 대한 추정을 할 때 위의 paradata를 이용하여 보다 향상된 추정을 하는 것을 고려하고자 한다. 이를 위해서 '첫 컨택 반응'변수를 Z라고 하고, '상용 근로자 수'를 X_1, 그리고 '산업분류'를 X_2라고 하자. 또한, '기업 이윤'을 Y라고 하자. Y는 무응답을 가질 수 있으며 이 무응답 여부에 대한 지시변수를 δ라고 정의하자.

관심 모수는 Y를 $X=(X_1, X_2)$에 대한 회귀모형을 적합할 때 생기는 회귀계수라고 하자. 즉,

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+e, \qquad e \sim (0, \sigma^2) \tag{1}$$

의 회귀모형을 적합시킨다고 하자. 이 경우 X는 첫해 연도에서 관측되고 매년 큰 변동이 없으므로 항상 관측되는 설명변수로 간주할 수 있지만 기업의 이윤 Y는 매년 달라지는 것이므로 해당 연도에서 관측되지 않을 수 있다.

이러한 경우 무응답 메커니즘이 Missing at random(MAR)이라 함은 다음의 가정을 의미한다.

$$P(\delta = 1|X,\ Y) = P(\delta = 1|X). \qquad (2)$$

이러한 MAR 하에서 회귀계수 추정을 할 때 사용될 수 있는 방법으로는 CC(Complete-Case) 방법과 응답 확률을 추정하여 그의 역수를 가중치로 넣는 무응답 가중치 조정법이 있다. 가정 (2)에서는

$$f(Y|X, \delta = 1) = f(Y|X)$$

이 성립하므로 $f(Y|X, \delta = 1)$의 모수를 추정하는 CC 방법이 (1)의 회귀계수를 추정하는 데에도 사용될 수 있으며 무응답 가중치 조정법보다 더 통계적으로 효율적인 (즉, 분산이 작은) 추정량을 구현한다.

그러나 (2)의 MAR 가정은 사용하기에 편리함에도 불구하고 비현실적일 수 있다. 일반적으로 MAR 가정이 성립하려면 설명변수 X를 확장시키는 것이 중요하다. 즉, 실제 무응답 메커니즘이 (X_1, X_2)와 어떤 잠재변수 U의 함수라고 하자. 즉,

$$P(\delta = 1|X_1, X_2,\ Y, U) = P(\delta = 1|X_1, X_2,\ U)$$

이 성립한다고 하자. 그런데 U가 X_1과 X_3으로 충분히 설명될 수 있다면

$$\delta \perp Y|(X_{1,}\ X_{2,}\ X_3)$$

이 성립하므로 $(X_{1,}\ X_2)$만을 고려한 모형에서는 MAR이 성립하지 않지만 X_3까지 고려한 확장된 모형에서는 MAR이 성립하게 되는 것이다. 따라서 본 연구에서는 paradata Z를 모형에 추가함으로써 MAR 가정이 더 잘 성립할 수 있게 함과 동시에 그러한 경우 어떠한 추정법을 사용하여 모형 (1)의 회귀계수들을 효율적으로 추정할 수 있을 것인지에 대해 논의하고자 한다.

제2절 제안된 방법론

식 (2)에서 표현되는 MAR 가정은

$$\delta \perp Y|X$$

로도 나타낼 수 있는데 이러한 가정이 성립하지 않을 경우 X를 확장하여 조건을 약화시켜 준다. 즉

$$\delta \perp Y|X \Rightarrow \delta \perp Y|(X, Z)$$

이 성립하므로 (2)의 MAR 조건보다는 다음의 확장된 MAR 조건이 더 약한 조건이다.

$$P(\delta = 1|X, Z, Y) = P(\delta = 1|X, Z). \quad (3)$$

특히, 사업체 패널조사에서처럼 첫 컨택 반응변수를 Z로 사용하는 경우 Z는 무응답 여부에 대한 설명력이 높을 것이므로 (3)의 조건이 상당한 설득력을 가지게 된다. 식 (3)으로 표현되는 무응답 메커니즘을 확장된 MAR(MAR augmented by paradata)라고 부르도록 하자. 이러한 가정이 성립한다고 할 때 흔히 사용되는 방법은 $P(\delta = 1|X, Z)$의 무응답 모형을 설정하고 그 모형의 모수들을 추정한 후 이를 가지고 무응답 조정 가중치로 사용하는 것이다. 김기민(2013)은 사업체 패널조사 자료를 가지고 Z를 포함하는 로지스틱 회귀모형을 고려하여 무응답 조정 가중치를 구현하였다. 무응답 조정 가중치와 관련된 보다 깊이 있는 내용은 Kim and Shao(2013)의 제5장에서 찾아볼 수 있다.

본 연구에서는 이러한 무응답 조정 가중치를 이용한 방법 외에 새로운 방법을 제안하고자 한다. 김기민(2013)이 고려한 방법론이 일종의 '확장된 무응답 모형(Augmented nonresponse model)'을 사용한 것이라고 한다면 본 연구에서 제안하는 방법론은 '확장된 회귀모형(Augmented outcome regression model)'을 사용한 것이라고 할 수 있다. 즉 먼저 다음과

같은 회귀모형을 고려한다.

$$Y = \gamma_0 + \gamma_1 X + \gamma_2 Z + e, \qquad e \sim (0,\ \sigma_e^2) \tag{4}$$

위의 식 (4)의 모형이 $f(Y|X,\ Z)$에 대한 유효한 모형이라고 한다면 (3)의 가정에 의해서

$$f(Y|X,\ Z,\ \delta = 1) = f(Y|X,\ Z)$$

이 성립하게 되고 이러한 경우 $f(Y|X,\ Z,\ \delta = 1)$의 추정에 사용되는 CC 방법은 (4)의 회귀계수에도 최적 추정량을 제공한다. 이렇게 해서 ($\hat{\gamma_0}$, $\hat{\gamma_1}$, $\hat{\gamma_2}$)이 얻어지면 이를 바탕으로 $E(Y|X,\ Z)$의 추정을 할 수 있게 되지만 이는 본래 관심이었던 $E(Y|X)$와는 다른 것이다. 따라서

$$E(Y|X) = E\{E(Y|X,\ Z)|X\}$$

을 이용하여야 하고, 이를 위해서는 $E(Z|X)$를 계산해서 Z 자리에 넣어 줌으로써 식 (1)의 계수 추정치를 얻어낸다. 이를 정리하면 다음과 같다.

1. Specify $E(Y|X,\ Z) = \gamma_0 + \gamma_1 X + \gamma_2 Z$
2. Use the CC method to estimate γ:

$$\sum_{i \in S} \omega_i\, \delta_i\, \{y_i - m(Y|x_i,\ z_i, ;\gamma)\}(1,\ x_i,\ z_i) = 0.$$

3. Use the full data to estimate $E(Z|X) = \alpha_0 + \alpha_1 X$.
4. The final estimate for $E(Y|X)$ is

$$\widehat{E(Y|X)} = \hat{\gamma_0} + \hat{\gamma_1} X + \hat{\gamma_2}(\hat{\alpha_0} + \hat{\alpha_1} X)$$

That is, we have

$$\hat{\beta_0} = \hat{\gamma_0} + \hat{\gamma_2}\,\hat{\alpha_0}$$
$$\hat{\beta_1} = \hat{\gamma_1} + \hat{\gamma_2}\,\hat{\alpha_1}$$

이러한 방법으로 얻어지는 회귀계수의 분산 추정량은 붓스트랩(boo-tstrap)이나 잭나이프(jack-knife) 같은 resampling 방법을 사용하면 될

것이다. 만약 테일러 전개를 이용한 선형 근사법을 사용하고자 한다면

$$\hat{\beta}=\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}=(I_2,\ \hat{\alpha})\hat{\gamma}$$

이고 이때 $\hat{\alpha}=(\hat{\alpha_0},\ \hat{\alpha_1})'$이고 $\hat{\gamma}=(\hat{\gamma_0},\ \hat{\gamma_1},\ \hat{\gamma_2})'$으로 표현된다. 따라서

$$V(\hat{\beta})=(I_2,\ \hat{\alpha})\,V(\hat{\gamma})(I_2,\ \hat{\alpha})'+\hat{\gamma_2^2}\,V(\hat{\alpha})$$

의 형태로 분산추정치를 구하면 된다.

실제 자료에서는 X와 Z가 대부분 범주형(categorical) 변수이므로 이러한 경우 회귀모형을 fully nonparametric하게 구현될 수 있다. 이러한 경우에는 (4)의 모형보다는

$$Y=\gamma_0+\gamma_1\tilde{X}+\gamma_2\tilde{Z}+\gamma_3\tilde{X}\tilde{Z}+e\,,\qquad e\sim(0,\ \sigma_e^2) \tag{5}$$

의 모형을 고려할 수 있는데 여기에서 $\tilde{X}$는 X로부터 유도된 지시변수이다. 즉, X가 {1, 2, 3} 중 하나의 값을 갖는 이산형 확률변수라고 한다면 $\tilde{X}=(\tilde{X_1},\ \tilde{X_2})$이고

$$\tilde{X_1}=\begin{cases}1 & \text{if } X=1\\ 0 & otherwise\end{cases}$$
$$\tilde{X_2}=\begin{cases}1 & \text{if } X=2\\ 0 & otherwise\end{cases}$$

으로 표현된다. 이렇게 해서 $\hat{\gamma}=(\hat{\gamma_0},\ \hat{\gamma_1},\ \hat{\gamma_2},\ \hat{\gamma_3})$가 얻어지면 $E(\tilde{Z}|\tilde{X})$를 구해서 $\tilde{Z}$ 대신에 사용해야 하는데 $E(\tilde{Z}|\tilde{X})$의 추정은 다음과 같이 구현된다.

$$\widehat{E(\tilde{Z}|\tilde{X})}=\frac{\sum_{i\in S}\omega_i I(\tilde{Z}_i=1,\ \tilde{X}_i=1)}{\sum_{i\in S}\omega_i I(\tilde{X}_i=1)} \tag{6}$$

식 (6)을 $E(Y|\tilde{X},\ \tilde{Z})$에 대입하면 $E(Y|\tilde{X})$를 얻어낼 수 있게 된다.

이 방법론은 무응답 대체(imputation)의 형태로도 구현될 수 있다. 조

건부 기댓값의 성질에 의하면

$$E(Y|X)= E\{E(Y|X, Z)|X\}$$

이고 따라서 $\theta = E(Y)$의 추정을 $\hat{\theta}= \sum_{i \in S} w_i y_i$으로 할 때 imputation을 사용하여

$$\hat{\theta_I} = \sum_{i \in S} w_i \{\delta_i y_i + (1-\delta_i) E(Y|x_i, z_i, \delta_i = 1)\}$$

의 형태로 구현할 수 있는데 여기에서 조건부 평균 $E(Y|x_i, z_i, \delta_i = 1)$은 응답자 내에서의 조건부 분포를 이용해서 nonparametric하게 구할 수 있다. 만약 x 중에서 연속형 변수 항목이 있으면 nonparametric kernel regression 방법을 사용하게 될 것이다. 즉 특정 $z_i = z$하에서 조건부 평균 추정은

$$\hat{E}(Y|x_i, z_i = z, \delta_i = 1)= \frac{\sum_{j \in S_z} w_j \delta_j K_h(x_j - x_i) y_j}{\sum_{j \in S_z} w_j \delta_j K_h(x_j - x_i)}$$

의 형태로 구현하게 될 것이다. 여기에서 S_z는 $z_i = z$을 만족하는 표본의 부분집합을 나타내고 $K_h(x)$는 bandwidth가 h인 Kernel function을 지칭한다. 이러한 imputation 방법을 사용한 후 이 자료를 가지고 회귀분석을 적용할 수도 있다. 이렇게 해서 얻어지는 회귀분석 추정량은 앞에서 설명한 two-step으로 얻어진 회귀분석 추정량과 근사적으로 동일함을 보일 수 있다. 만약 조건부 평균 $E(Y|x_i, z_i, \delta_i = 1)$이 x와 z의 일차함수로 표현되는 경우에는 완전히 동일해진다.

제3절 모의실험

제안된 방법론의 타당성을 체크하기 위하여 다음과 같은 간단한 모의실험을 실시하였다. 사용한 모형은 다음과 같다.

• Outcome model

$y_{i=\beta_0} + \beta_1 x_{i1} + \beta_2 x_{i2} + e, \qquad e \sim N(0,1)$
where $(\beta_0, \beta_1, \beta_2) = (1, 0.6, 0.5)$, $x_{i1} \sim Gamma(1,2)$ and
$x_{i2} \sim Bernoulli(0.8)$

• Paradata model

$z_i = 1.2\, y_i + u, \qquad u \sim N(0,1)$

• Response model

$\delta_i \sim Bernoulli(\pi_i)$
where $\log it(\pi_i) = \phi + \phi_1 x_{1i} + \phi_2 z_i$ and $(\phi_{0,} \phi_{1,} \phi_2) = (1, 0.5, -0.4)$.
(Response rate is 66%).

위의 모형으로부터 n=200의 자료를 B=2,000번 반복해서 독립적으로 발생시켰고 각 샘플마다 다음의 네 가지 추정량을 계산하였다.

1. 단순 CC 방법
2. PS1 : MAR을 가정한 성향점수 가중치 조정법
3. PS2 : Paradata를 사용한 성향점수 가중치 조정법
   $P(\delta = 1|X, Z)$를 로지스틱 회귀모형을 적합시켜 구현
4. 제안된 새로운 방법론

이러한 시뮬레이션을 통해 각 추정량의 몬테 카를로(Monte Carlo) 평균과 표준 편차를 계산하였고 이를 <표 3-1>에 기술하였다. <표 3-1>

〈표 3-1〉 Monte Carlo bias and standard error of $\hat{\beta}$

Parameter	Estimator	Bias	S.E.
β_0	CC	0.17	0.209
	PS1	0.17	0.211
	PS2	0.01	0.223
	New	0.00	0.194
β_1	CC	-0.02	0.042
	PS1	-0.02	0.042
	PS2	0.00	0.044
	New	0.00	0.039
β_2	CC	0.02	0.214
	PS1	0.02	0.215
	PS2	0.00	0.235
	New	0.00	0.199

주 : CC는 Complete case estimator, PS1은 Propensity score adjusted estimator (MAR), PS2는 Propensity score adjusted estimator using paradata, New는 proposed method를 의미함.

의 결과를 정리하면 다음과 같다.

1. CC method와 PS1은 Bias를 보여주지만 PS2와 새로 제안된 방법론은 Unbiased한 추정값을 구현하는 것을 보여준다.
2. PS2 방법보다는 새로 제안된 방법론이 더 작은 표준 편차값을 갖는다. 이는 $E(Y|X, Z)$의 추정이 최적 추정으로 구현되는 것에 기인하는 것으로 설명할 수 있을 것이다.

제4절 실증 분석

본 절에서는 제2절에서 제시한 방법들을 실제 자료에 적용하여 비교해 보고자 한다.

여기에서 활용한 자료는 사업체 패널조사(Workplace Panel Survey : WPS) 자료이다. 이 자료는 한국노동연구원에서 1,700여 개 사업체를 대상으로 2005년부터 격년마다 실시하며 고용 및 재무 현황, 인적자원 개발 및 관리, 작업장 혁신, 노사관계 등에 대한 내용을 포함하고 있다.

분석 자료는 사업체 패널조사의 민간부문 사업체에 대해 2차(WPS, 2007), 3차(WPS, 2009)와 4차(WPS, 2011) 데이터를 pooled(N=6,460) 하여 구성하였다. 기본 분석모형은 종속변수(Y)는 1인당 매출액, 독립변수(X)는 사업체 규모와 산업으로 구성하였다. 본 연구는 우리의 관심변수에는 영향을 미치지 않지만 관심변수의 응답여부에는 영향을 미치는 변수를 통해 $E(Y|X)$의 정확도와 효율성을 높이는 추정을 하고자 하는 것이므로 이에 상응하는 변수를 선택해야 한다.

<표 3-2>는 첫 컨택 반응을 다섯 가지로 범주화하여 응답여부에 대해 분석한 결과이다. 첫 컨택 시 우호적인 반응을 보이는 경우 응답률이 68.2%, 부정적인 반응을 보이는 경우 32.5%로 나타나 사업체의 매출에는 영향을 미치지 않지만 응답여부에는 영향을 미칠 것으로 판단되는

〈표 3-2〉 첫 컨택 반응에 따른 연도별 1인당 매출액 응답여부

		우호적 반응	재연락	내부 결재중	매우 바쁨	부정적 반응	전 체
2007	응답	781	85	67	25	109	1,067
	무응답	233	43	33	13	226	548
	응답률(%)	77.0	66.4	67.0	65.8	32.5	66.1
2009	응답	659	82	51	23	115	930
	무응답	355	46	49	15	220	685
	응답률(%)	65.0	64.1	51.0	60.5	34.3	57.6
2011	응답	634	82	52	21	103	892
	무응답	380	46	48	17	232	723
	응답률(%)	62.5	64.1	52.0	55.3	30.7	55.2
Pooled	응답	2,074	249	170	69	327	2,889
	무응답	968	135	130	45	678	1,956
	응답률(%)	68.2	64.8	56.7	60.5	32.5	59.6

Z에 해당하는 변수로 첫 컨택 반응을 선택하였다.

- Y : log(1인당 매출액)
- X : 사업체 산업(12개) · 규모(5개)
- Z : 첫 컨택 반응

첫 번째는 CC 방법으로 MAR을 가정하여 응답한 자료만을 가지고 단순히 $E(Y|X)$를 구하는 것이다.

두 번째는 로지스틱 회귀모형으로 1인당 매출액 무응답 여부에 대해 분석하여 응답확률을 구한 후, 그 응답확률을 Y에 대한 모형에 가중하여 $E(Y|X)$를 구한다. 이때, 1인당 매출액 무응답 여부에 대한 로지스틱 회귀모형은 사업체 산업 · 규모를 가지고 구축하였다.

세 번째는 사업체의 산업 · 규모뿐 아니라 첫 컨택 반응변수(Z)를 포함하여 로지스틱 회귀모형으로 응답확률을 구하고 두 번째 방법과 마찬가지로 Y의 모형에 가중하여 $E(Y|X)$를 구한다.

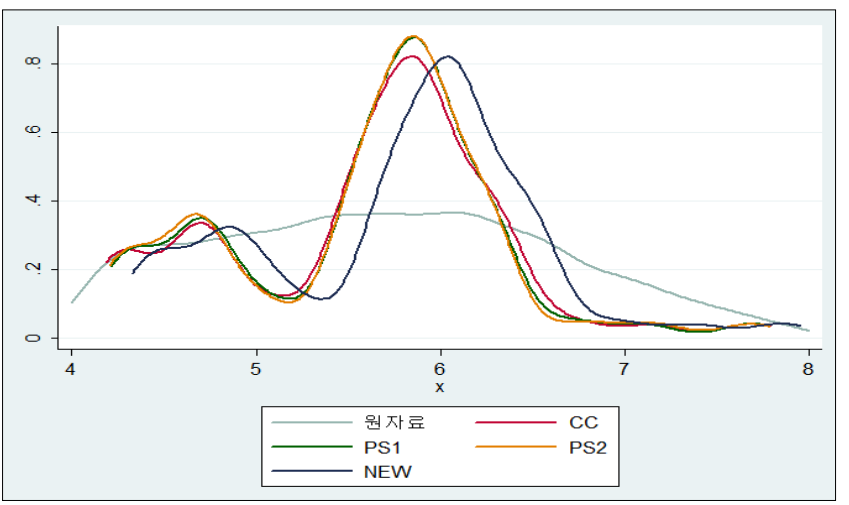
마지막은 기본 모형에 첫 컨택 반응변수(Z)까지 포함하여 응답한 자료에 대한 $E(Y|X, Z)$를 구한 후, 다시 Z에 대해 모형을 설정하여 $E(Z|X)$를 구하여 추정된 계수를 다시 $E(Y|X, Z)$에 넣어 $E(Y|X)$를 구하는 방법이다. 이때, 첫 컨택 반응에 대한 정보인 Z는 매출액에 대해 무응답을 한 사업체에 대해서도 정보가 있으므로 전체 사업체에 대해 추정계수를 구한다. 제3절의 모의실험은 X와 Z 모두 연속형 변수에 대해 보여주었으나, 여기에서는 X와 Z가 범주형 변수(categorical variable)이기 때문에 식 (6)을 사용하여 $\tilde{Z}$ 대신 사용할 $E(\widetilde{Z}|\widetilde{X})$를 구하여 추정하였다. 회귀계수의 분산 추정량은 붓스트랩(boostrap) 방법(N=2,000)을 통해 구하였다.

<표 3-3>은 네 가지 방법을 적용하여 분석한 결과이고, [그림 3-1]은 추정된 결과에 따른 log(1인당 매출액)의 분포이다. 새로 제안된 방법으로 분석한 결과에서는 다른 방법들과 달리 대다수의 변수에서 표준오차가 다소 작게 나타났고, 세 가지 방법(CC, PS1, PS2)의 분포는 큰 차이를 보이지 않았으나 새로 제안한 방법의 분포는 세 가지 방법에 비하여 오른쪽으로 이동한 형태를 보였다.

〈표 3-3〉 사업체 패널자료에 네 가지 방법을 적용한 분석결과

	CC		PS1		PS2		New	
	계수	표준오차	계수	표준오차	계수	표준오차	계수	표준오차
상수항	5.468	0.059	5.487	0.062	5.462	0.062	5.629	0.050
100~299인 이하	0.151	0.048	0.139	0.048	0.162	0.047	0.152	0.049
300~499인 이하	0.458	0.051	0.403	0.052	0.418	0.052	0.440	0.051
500인 이상	0.588	0.056	0.522	0.056	0.510	0.056	0.571	0.061
화학공업	0.380	0.076	0.369	0.083	0.369	0.083	0.421	0.074
금속, 자동차, 운송	0.257	0.072	0.254	0.076	0.259	0.076	0.326	0.058
전기, 전자, 정밀	0.008	0.079	0.015	0.081	0.034	0.081	0.079	0.064
건설업	0.178	0.091	0.157	0.093	0.176	0.094	0.166	0.099
개인서비스업	0.342	0.078	0.359	0.083	0.383	0.084	0.353	0.070
운수업	-1.284	0.080	-1.274	0.085	-1.258	0.084	-1.300	0.082
통신업	0.093	0.141	0.104	0.143	0.159	0.135	0.066	0.151
금융보험업	1.178	0.117	1.178	0.114	1.252	0.113	1.268	0.104
사업서비스업	-1.224	0.078	-1.175	0.079	-1.185	0.078	-1.168	0.084
사회서비스업	-0.881	0.084	-0.853	0.076	-0.874	0.075	-0.889	0.071
전기, 가스, 수도사업	2.169	0.159	2.152	0.178	2.162	0.179	2.171	0.071

〔그림 3-1〕 각 방법들의 추정된 결과에 따른 log(1인당 매출액) 분포



주 : 1) kernel 함수는 gaussian을 사용.

제5절 결론 및 향후 과제

본 연구에서는 paradata를 이용하여 기존의 MAR보다 약한 무응답 메커니즘 가정을 고려하였고 그 가정하에 더 효율적인 추정 방법론에 대해 제안하고 간단한 시뮬레이션을 통해 제안된 방법론의 타당성과 효율성을 보여주었다. 본 연구에서 제안된 방법론은 확장된 회귀모형을 사용하여 그에 대한 최적 추정을 구현한 후 paradata 부분을 조건부 기댓값을 취해 없애주었는데, 이는 기존의 확장된 무응답 모형을 사용하는 방법론보다 더 효율적인 추정을 구현하게 된다. 본 연구에서 제안된 방법론을 한국노동연구원의 사업체 패널조사에 적용하였다. 적용 결과, 본 연구에서 제안된 방법은 다른 방법들에 비하여 추정치의 분산 추정량이 다소 작게 나왔다.

또한 보다 일반적인 회귀모형(예를 들어 로지스틱 회귀모형이나 비모수 회귀모형)으로의 확장도 추가적으로 연구될 계획이다.

제 4 장
사업체 패널조사의 표본 설계 관련 연구

제1절 서 론

한국노동연구원에서 실시하고 있는 사업체 패널조사는 2005년부터 격년으로 실시하는 대규모 사업체 조사로 우리나라 기업의 인적자원관리와 노동수요 및 노사관계 등 다양한 이슈와 관련한 기초 자료를 제공한다. 이 사업체 패널조사는 전국 30인 이상 사업체를 모집단으로 전국의 1,700여 개 표본 사업체를 층화 추출 하여 실시하고 있으며, 패널사업체의 휴폐업, 응답거절 등의 사유로 표본 탈락이 진행되어 왔다. 표본 탈락이 계속적으로 진행되면서 표본의 대표성과 동태적 변화를 정확하게 파악하는 데 여러 문제점이 발생하였고 이에 따라 새로운 패널 표본 설계 또는 기존 패널에 대한 재설계가 필요한 실정이다.

사업체 조사는 일반 사회조사와는 다른 속성을 가지고 있고 따라서 가구방문을 바탕으로 한 일반 사회조사와는 다른 방식으로 표본을 추출하게 된다. 사업체 조사는 사회조사보다 더 왜도(skewness)가 심한 자료를 다루게 되고 따라서 아주 큰 규모의 사업체를 어떻게 뽑느냐에 따라 통계가 많이 달라지게 된다. 사업체의 시계열적 변동을 설계와 가중치 작성에 어떻게 반영할 것인가도 중요한 문제이다. 또 사업체 조사자료는 응답거절이 일어난 경우에도 그 사업체에 대한 기존 정보가 있으므로 이를 반영하여 추정을 개선할 수 있다. 이 경우 흔히 사용될 수 있는 방법

으로는 ratio imputation이 가능할 것이다.

또한 사업체 조사는 일반 사회조사와는 달리 원자료 공개에 좀더 소극적이다. 이는 사업체 조사자료에서 응답업체에 대한 정보 식별이 더 용이해서 confidentiality protection이 더 어려워지기 때문이다. 추정과 가중치 조정법은 비교적 단순한 방법을 사용하는 경향이 있지만 종종 outlier가 발생하기 때문에 이를 처리하는 방법에 대한 연구가 필요하다. 특히 세분화된 단위에서 통계를 생산해야 하는 경우에는 추정이 불안정해지므로 모형을 이용하여 추정을 개선하는 방법론도 고려할 수 있을 것이다.

본 보고서에서는 기존의 사업체 패널조사에 나타난 몇 가지 문제점을 바탕으로 이를 해결하기 위한 새로운 방안으로서 새로운 패널 표본 설계 또는 기존 패널을 보완하는 표본 재설계를 할 때 유의해야 할 사항들을 문헌 조사 및 자료 분석을 통해 정리하고자 한다. 이를 위하여 제2장에서는 사업체 패널조사의 특성에 대한 일반적인 이슈들을 다루고 해외사례 및 기타 참고문헌을 정리하여 소개하였다. 제3장에서는 사업체 패널을 새롭게 설계하고자 할 때 고려해야 할 사항을 다루었고 제4장에서는 기존 패널 재설계에 대한 내용을 다루었다. 제5장에서는 향후 발생하게 될 단위 무응답이나 패널 탈락에 대한 내용을 정리하였다.

제2절 사업체 조사와 관련된 기초 연구

표본 조사를 하기 전에 목표 모집단과 조사 모집단을 정확하게 정의하고 시작하는 것이 필요하다. 목표 모집단(target population)이 조사의 목적을 위한 이상적인 모집단이라고 한다면 조사 모집단(survey population)은 조사를 실시하기 위해 현실적으로 고려하게 되는 모집단이라고 할 수 있다. 예를 들어 목표 모집단은 영세규모(종업원 수 10인 이하)도 포함하지만 현실적으로 그러한 영세규모의 사업체를 모집단에 포함시키면 패널 유지가 힘들게 되므로 조사 모집단의 정의에서 규모를 어느

정도 이상(예 : 종업원 수 30인 이상)으로 국한시키고 있다. 이러한 경우 조사 모집단의 목표 모집단에 대한 커버리지가 어느 정도 되는지에 대한 검토가 선행되어야 할 것이다.

이렇게 조사 모집단이 정해지면 이 조사 모집단에 대한 리스트인 표본 추출틀을 만들어야 한다. 이를 위해서는 사업체 등록 명부를 이용하거나 사업체 총조사 자료를 이용한다. 사업체 등록 명부로는 통계청에서 2000년부터 구축한 사업체 통합관리시스템에서 매년 업데이트되는 명부를 고려할 수 있으나 이러한 자료를 표본 설계를 위해 사용할 수 있을지는 미지수이다. 사업체 총조사 자료를 사용하는 경우에는 조사 연도와 총조사 연도와의 괴리를 해결해야 한다. 예를 들어 조사 모집단이 2014년에 조사 적격 사업체라고 한다면 2010년 사업체 총조사 자료와는 4년의 공백이 있게 된다. 그래서 지난 4년간 적격 사업체로 신규 진입하게 되는 사업체 리스트를 확보하여 이를 사업체 총조사 자료에 추가하여 표본 추출틀을 만들어야 한다. 또한 사업체 총조사 자료에 대한 quality check를 해야 하는데 예를 들어 업종이나 종업원 수가 어느 정도 정확한지에 대한 체크(이를 profiling이라고 한다)가 필요하다. 이러한 profiling은 대규모 업체 위주로 실시하는 것이 일반적이다.

사업체 조사를 위한 표본 추출틀을 결정하는 데 고려해야 할 사항으로는 먼저 사업체 모집단 자료의 분포가 상당히 skewed되는 경향이 있으므로 이러한 대규모 사업체에 대한 전수조사를 실시할 경우 사용 가능해야 한다는 것이다. 또한 산업분류가 일치하는지 확인할 필요가 있다. 또, 조사 단위로 사업체를 사용하는 경우 그것이 분석 단위와 일치하는지 확인해야 한다. 사업체가 본사와 지사로 구분되는 경우 이를 합쳐서 분석을 할 것인지 아니면 별도로 분석할 것인지에 따라 조사 단위가 결정될 수 있을 것이다.

이렇게 표본 추출틀이 결정되면 표본 사업체를 추출해야 하는데 주로 사용되는 표본 추출방법은 층화 추출이다. 이를 위해서는 먼저 표본 추출 단위를 결정해야 하고 그 후 층을 정하고, 층 내 표본 수 배정을 해야 한다. 층의 결정은 통계 공표 단위와 모집단 자료의 분포, 그리고 실사 단계에서의 용의성 등을 고려해서 결정하게 된다. 특히 층 내의 모든 원

소를 다 조사하게 되는 층을 전수층(certainty stratum)이라고 하는데 전수층은 대규모 사업체를 표본에 포함함으로써 추정이 안정적으로 나올 수 있도록 한다는 장점이 있다. 일반적으로는 산업 분류별로 종업원 수를 기준으로 층의 경계를 결정한 후 주요 항목별 층별 예상 표준오차를 계산한 후 전체 비용 및 층별 CV와 관련한 몇 가지 제한조건을 만족하면서 전체 CV를 최소화하는 표본 배정을 찾아냄으로써 최적 표본 설계를 구현하게 되는데, 보다 자세한 내용은 제3장에서 다루기로 한다.

이렇게 해서 얻어진 사업체 샘플은 실제 조사 단계에서 프레임과는 다른 정보를 얻게 되는 경우가 있다. 예를 들어 명부에는 종업원 수가 40명이라고 했는데 방문하여 보니 종업원 수가 20명일 수도 있고 또는 그 반대로 100명일 수도 있을 것이다. 이러한 문제의 사업체 샘플은 전수조사 당시에 얻어진 자료의 문제였는지 아니면 실제 변동에 의한 것인지에 따라 다르게 처리될 수 있을 것이다. 전수조사 당시에 얻어진 자료의 문제라면 이는 표본 추출틀(sampling frame)의 문제이므로 이를 수정해서 그 바뀐 값을 표본 추출틀에 대체한 후 사용한다. 즉, 위의 예에서 종업원 수가 40명이라는 정보가 잘못된 정보였으므로 이를 제외하고 층 내 비슷한 규모의 다른 사업체로 대체하여 조사한다. 만약 전수조사 당시의 문제가 아니라 실제 변동에 의한 것이라면 이는 샘플링 자체의 문제가 아니라 자료의 dynamics에 의한 것이므로 원래 표본 설계 당시의 층에 대한 대표성을 가지는 표본으로 사용한다.

소규모층의 사업체들은 대부분 매우 큰 가중치를 갖게 되는데 이러한 소규모 사업체 샘플 중에서 일부가 아주 큰 규모로 성장해 버리면 전체 추정이 왜곡되어 나타날 수 있다. 이러한 현상을 stratum jumper라고 부른다. stratum jumper는 일종의 outlier 문제로 볼 수 있고 이러한 경우에는 stratum jumper에 해당되는 사업체의 가중치를 줄여주고 동일 층의 다른 표본 사업체의 가중치를 늘려주어서 추정량의 불안정성을 보완할 수 있다.

이렇게 해서 표본을 추출하면 가중치 조정을 추가적으로 실시하는 것이 일반적이다. 가중치 조정은 표본의 횡단면적 대표성을 향상시키기 위해서 실시하는 calibration과 단위 무응답을 보정해주는 무응답 보정의

두 가지로 구분된다. calibration은 모집단에 대해 알려져 있는 정보를 이용하여 표본의 가중치를 조정해서 그 정보에 대한 표본 가중치 추정량이 모집단 참값과 동일해지도록 하는 방법론을 지칭한다. 이러한 calibration과 관련된 내용은 Kim and Park(2010)를 참고하기 바란다.

사업체 조사의 또 다른 특징은 에디팅과 imputation 단계에서 사용할 수 있는 정보가 많이 있다는 것이다. 샘플링 프레임 자체도 많은 정보를 가지고 있을 뿐만 아니라 해당 사업체의 전년도 설문 응답값 등도 해당 연도의 무응답값에 대한 좋은 정보를 제공한다. 이러한 정보는 에디팅과 imputation을 통해서 무응답값에 대한 결측치 처리기법으로 사용될 수 있는데 이러한 방법론은 해당 설문에 대한 이해와 축적된 지식을 통해 통계적 또는 논리적 모형을 사용하여 구현될 수 있을 것이다.

제3절 사업체 패널조사 설계방안 : 방안 A

다음으로는 사업체 패널조사를 새롭게 설계할 때 고려할 만한 사항들을 살펴보고자 한다. 사업체 패널조사는 많은 조사항목을 가지는 다목적 조사이며 이 조사를 통해 산업별, 사업장 규모별, 지역별 통계 생산이 가능하도록 하는 것을 기본 원칙으로 설계되었다. 새로운 표본 설계 방안도 원칙면에서는 이와 동일하며 이를 위하여 산업별 분류, 사업장 규모 및 지역을 층화변수로 하는 층화 추출을 사용하는 것도 동일하다. 층화 추출은 거의 대부분의 사업체 조사에서 실시하고 있는 표본 추출방법이다.

1. 조사 모집단 결정

표본 추출을 위해서는 먼저 조사 모집단이 결정되어야 하는데 이는 목표 모집단과 조금 다를 수 있다. 이때 중요한 것은 조사 모집단이 목표 모집단에 비해서 얼마나 다른가에 대한 정보를 어느 정도 제공해야 한다

는 것이다. 예를 들어 사업체 패널조사는 조사 모집단으로 종업원 수 30인 이상 사업체로 정하는데 이러한 적격 기준을 적용할 때 전체 사업체 중에서 어느 정도를 포함하는지 그리고 이러한 기준이 적정한 것인지에 대한 검토가 필요할 것으로 보인다. <표 4-1>은 모집단 자료에서 산업별로 30인 미만의 사업체를 제외한 분포를 보여주는데 30~99인의 규모에 대부분이 몰려 있으므로 30인 미만의 사업체도 상당히 많을 것임을 알 수 있다.

기존 표본 설계에서 일반 사업장은 30인 이상, 공공 사업장은 20인 이상으로 적격 기준을 정했는데 만약 새롭게 표본 설계를 하게 되면 이 기준에 대한 검토가 필요할 것으로 보인다. 이러한 결정은 통계학적인 검토보다는 실사 측면에서 먼저 살펴보아야 하고 패널 유지 가능성 측면에서도 검토해야 할 것이다. 만약 적격 사업장 규모를 10인 이상으로 확장한다고 할 경우, 패널 유지 측면에서 생성과 소멸을 어떻게 관리할 것인가에 대한 내부 지침을 잘 확보해 두어야 할 것이다.

〈표 4-1〉 산업 구분 및 사업장 규모별 모집단 사업장 수 현황

(단위 : 개소)

산업분류		30~99인	100~299인	300~499인	500인~	전 체
경공업		3,835	779	112	62	4,788
화학공업		2,441	543	84	65	3,133
금속, 자동차, 운송		4,663	883	101	116	5,763
전기, 전자, 정밀공업		2,351	644	89	115	3,199
건설업		1,455	201	19	21	1,696
개인서비스업		2,800	507	59	45	3,411
유통 서비스업	운수업	1,976	1,019	80	46	3,121
	통신업	393	95	12	10	510
사업 서비스업	금융 및 보험업	192	51	34	44	321
	기타	3,453	886	173	134	4,646
사회서비스업		4,474	733	97	152	5,456
전기, 가스 및 수도사업		27	12	0	0	39
전 체		28,060	6,353	860	810	36,083

자료 : 사업체노동실태현황 결과(2004년 12월 말 기준).

2. 층화 및 표본 추출

기존의 사업체 패널조사의 층화는 산업구분, 사업장 규모 및 지역 구분을 이용하였다. 사업장 규모 구분은 사업장의 상용 근로자 수를 기준으로 30~99인, 100~299인, 300~499인, 500인 이상 등으로 구분하였다. 지역 구분은 서울권, 경기·인천권, 강원·충청권, 전라·제주권, 영남권 등 전국을 5개 권역으로 구분하였다. 사업체의 업종은 산업의 특성에 따라 12개 업종으로 구분하였다. 따라서 표본 설계에 사용된 전체 층의 수는 지역(5)×산업분류(12)×규모(4)=240개 층이다.

이러한 층의 결정은 어떠한 통계를 생산하느냐에 따라 달라질 수 있다. 산업 구분별 비교, 사업장 규모별 비교, 지역별 비교가 필요한 경우 그에 따라 비교 집단에 해당하는 층에 어느 정도의 표본 수를 확보하는가가 중요할 것이다. 『사업체 패널자료 품질개선 연구』 보고서(홍민기 외)에 나와 있듯이 30인 근처에서 사업체 패널이 과소표집된 증거가 보이므로 30~99인의 층을 좀 더 세분화하여, 30~49인, 50~99인으로 나누는 방법이 더 나을 것으로 판단된다. 이렇게 하는 경우 전체 층의 수는 지역(5)×산업분류(12)×규모(5)=300개 층이다.

이렇게 층이 결정되면 층 내의 표본 수를 결정해주는 표본 배정을 실시해야 한다. 표본 배정 방법론으로 비례배분이나 제곱근 배분, 네이만 배정과 같은 교과서적인 방법론을 샘플링 교재에서는 다루고 있지만 실무에서는 mathematical programming 방법을 주로 사용한다. mathematical programming은 여러 가지 제한조건을 가지는 상황에서 목적함수를 최소화하고자 할 때 최적화 문제를 얻어내는 프로그래밍 기법으로 엑셀의 solver, SAS의 Proc NLP이나 Proc Optmodel, 또는 R의 alabama package를 사용하여 구현한다. Valliant et al.(2013)의 제5장에서 보다 자세한 내용을 참고할 수 있다.

층화 추출의 경우에 사용하는 제한조건으로는 전체 실사 비용, 모비율 추정에 대한 산업 구분별 오차한계, 그리고 층별 모집단 수 등이 포함될 것이고 목적함수로는 모비율 추정에 대한 전국 통계의 분산을 사용할 수 있을 것이다. 실사 비용은 지역별로 차이가 있을 것이므로 이를 비용함

수로 표현할 수 있고, 산업구분별 오차한계는 8%보다는 5% 이하로 줄이는 것이 적당할 것으로 보인다. 그리고 층별 표본 수는 층별 모집단 수보다 항상 작거나 같아야 한다는 조건이 제한조건에 포함되어야 한다.

제4절 사업체 패널조사 설계방안 : 방안 B

사업체 패널조사 설계의 또 다른 방안으로 이전의 조사에서 응답해오던 기존 표본은 그대로 두고 나머지 모집단에서 표본을 추가하는 추가 표본 설계를 고려할 수 있다. 패널조사에서 표본 탈락 또는 패널 마모(panel attrition)는 흔히 발생하는 현상으로 적절한 시점에서 계속 표본 추가를 해주어서 적정 표본 수를 유지하고 모집단에 새롭게 진입한 신규 사업체들을 포함함으로써 전체적인 횡단면적 대표성을 제고하는 효과를 가지게 된다.

추가 표본 설계에서는 층화 추출을 기본으로 하되 기존의 표본과 새로운 표본을 어떻게 결합하여 가중치를 계산할 것인가를 정하는 것이 핵심이다. 이를 위해서 일단 기존의 표본 자료에서 탈락이나 부적격으로 처리되어 떨어져 나간 사업체를 제외한 나머지 표본 사업체(이를 생존 표본 사업체라고 하자) 리스트를 층별로 작성한다. 이 생존 표본은 추가 표본 설계 시 새 표본에 포함되도록 한다. 다만, 생존 표본 사업체의 규모가 초기와 달라져서 층간 변동이 생기는 경우에는(예를 들어 상용 근로자 수가 40명에서 200명으로 늘었다면) 마지막 시점에서의 해당 층으로 포함시켜야 한다.

층 h에서의 모집단 크기를 N_h라고 하고 생존 표본 크기를 r_h라고 하면 최종 표본 수 n_h는 $r_h \leq n_h \leq N_h$을 만족하여야 할 것이다. 그리고 전체 비용을 C라고 하고 층 h에서의 사업체 하나당 조사하는 데 드는 실사 비용(조사원 인건비 포함)을 C_h, 그리고 초기 비용을 C_0라고 하자. $W_h = N_h/N$이라고 하고 층 h에서의 층 내 분산을 S_h^2라고 할 때, 이러

한 제한조건을 포함하는 optimization problem을 기술하면 다음과 같다.

$$\min_{n_h} \sum_{h=1}^{H} W_h^2 \frac{1}{n_h}\left(1-\frac{n_h}{N_h}\right)S_h^2$$

subject to

$$r_h \le n_h \le N_h,\ h=1,\cdots,H$$
$$\sum_{h=1}^{H} C_h n_h \le C$$
$$N_D^{-2} \sum_{h\in D} N_h^2 \frac{1}{n_h}\left(1-\frac{n_h}{N_h}\right)S_h^2 \le d^2/4$$

for each D, where $N_D = \sum_{h \in D} N_h$(이때, D는 산업 분류를 의미함.)

즉 마지막 constraint는 각 산업분류별 평균값에 대한 오차 한계가 주어진 d값(예 : $d=0.05$)보다 더 작게 나오게 되는 조건을 의미한다. 위의 최적화 문제를 solver나 Proc NLP 등을 사용해서 풀면 최적 n_h^*가 얻어지고 그것으로부터 $m_h^* = n_h^* - r_h$을 계산하면 각 층으로부터 m_h^*만큼의 표본 사업체를 추출하면 된다.

이렇게 해서 추가 표본 사업체를 각 층별로 추출하면 그 이후에는 기존의 생존 표본과 결합하여 표본 추출 확률을 바탕으로 한 가중치를 계산하면 된다. 기존 표본이나 신규 추가 표본 모두 동일한 층 내에서는 동일한 가중치를 갖게 된다.

위의 최적화 문제에서 층 내 분산 S_h^2을 알아야 하는데 단순하게 비율 추정의 문제로 간주하고 $S_h^2 \le 0.25$을 이용하여 단순히 0.25를 사용할 수도 있으나 표본 추가의 경우에는 과거 자료의 정보가 있으므로 S_h^2에 대한 보다 정확한 추정을 할 수 있고 이를 이용하여 최적 배분 문제를 풀어낼 수 있을 것이다.

이렇게 해서 얻어진 표본을 바탕으로 실사를 통해 최종 표본을 확정지을 수 있다. 실사 단계에서는 우편이나 직접 방문을 통해 조사를 하는데 먼저 추출 당시에 사용했던 정보와 일치하는지 확인이 필요하다. 만약

업종이나 지역, 종업원 수에 대한 정보에서 차이가 발생한다면 (다른 층에 속한다면) 이 차이가 프레임 오차에 의한 것인지 아니면 사업체 자체의 변동에 의한 것인지를 확인해야 한다. 프레임 오차에 의한 것이라면 해당 사업체를 해당 층의 다른 사업체로 대체해야 하고 만약 실제 변동에 의한 것이라면 그 정보를 그대로 사용해야 할 것이다.

다음으로는 근로자 가중치의 사용에 대해 알아보고자 한다. 사업체 패널조사는 표본 추출 단위는 사업체이지만 분석 단위는 사업체 단위의 항목도 있고 근로자 단위의 항목도 있을 것이다. 이러한 경우 만약 사업체 단위의 분석법으로 근로자 단위에 대한 모수 추정을 하게 되면 소규모 사업체에 대해 지나치게 많은 가중치를 부여하게 됨으로써 소규모 사업체에 근무하는 근로자 현황이 over-represent되는 경향이 나타나게 될 것이다. 즉, 평균 추정을 할 때 사업체 i에 근로자 수가 M_i라고 하고 사업체 i의 근로자 j에게 얻어지는 특정 항목 관측값을 y_{ij}라고 한다면 모집단 전체의 근로자 단위 평균은

$$\theta_1 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \overline{Y_i}}{\sum_{i=1}^{N} M_i}$$

으로써 사업체 단위 평균인

$$\theta_2 = \frac{1}{N}\sum_{i=1}^{N} \overline{Y_i}$$

과는 완전히 다른 의미를 가진다. θ_1이 근로자 수에 비례하는 가중 평균의 개념이라면 θ_2는 소규모 사업체와 대규모 사업체가 동일한 비중을 갖는 단순 평균의 개념이라고 할 수 있다. 사업체 설문에는 사업체 가중치를 사용해야 하지만 근로자 설문에는 근로자 가중치를 적용하는 것이 더 정확하다. 근로자 가중치는 사업체 가중치에 근로자 수를 곱한 후 표본 가중치 합이 해당 업종의 근로자 총수에 맞추어지도록 가중치를 결정한다.

제5절 패널 마모 처리

패널 표본 마모로 인한 대표성 왜곡 문제는 사업체 조사의 경우에는 폐업, 응답거부 등으로 인해 생존 표본이 더 이상 대표성을 가지지 못하게 되는 경우에 발생한다. 이러한 경우, 패널 표본 추가를 통해 횡단면적 대표성을 보완할 수 있지만 매번 패널 표본 추가를 할 수는 없는 상황이고 빈번한 패널 표본 추가는 표본 사업체의 변화로 인한 통계의 불안정성을 가져온다고 볼 수 있다.

따라서 이러한 패널 마모 처리방법으로는 imputation이나 가중치 조정을 사용하여야 하는데 이에 대한 자세한 기술은 이 보고서의 범위를 넘는 것이므로 본 보고서에서는 간략하게 전체적인 내용만을 기술하고자 한다. 먼저 단순한 응답거절인지, 폐업 등으로 인한 부적격 탈락인지에 따라 처리방법이 달라질 것이다. 폐업 등으로 탈락한 것이라면 이는 패널 표본의 변동으로 볼 수 있는 것이므로 단순한 무응답 처리를 하는 것보다는 하나의 변동 카테고리로 인정하는 것이 좋다(예를 들어 고령자 패널의 경우, 사망과 무응답을 구분해야 하는 것과 동일하다. 만약 어느 표본이 사망 등으로 인하여 소멸한 것이라면 이는 무응답 처리를 해서는 안 된다. 사망은 하나의 응답된 상태로 인정해야 하는 것이다). 응답거절의 경우에는 가중치 조정을 해야 하는데 층화 추출에서 사용한 층을 바탕으로 동질성을 가정하여 처리한다.

가중치 조정법은 가장 단순하게 처리할 수 있는 것으로 층 내에서 응답 사업체 표본값의 평균이 층 내 전체 표본을 대표할 수 있도록 응답 사업체의 가중치를 올려주는 것이다. 즉 층 h에서 처음 r_h 사업체가 응답하였고 나머지 $n_h - r_h$ 사업체가 응답하지 않았다고 한다면 응답한 사업체의 최종 가중치는 다음과 같이 계산된다.

$$w_{hi}^{*} = w_{hi} \times \frac{\sum_{i=1}^{n_h} w_{hi}}{\sum_{i=1}^{r_h} w_{hi}}$$

이때 w_{hi}는 사업체 (hi)의 설계 가중치(사업체 가중치, 근로자 가중치)이다.

이러한 가중치 조정법은 가장 단순하게 적용할 수 있는 방법이나 만약 r_h가 작은 경우(20 이하)에는 유효 표본 수가 적어지므로 불안정한 통계치가 나오게 될 위험이 있다. 이러한 문제점을 극복하기 위해서는 층을 같은 산업분류 내에서 비슷한 규모끼리 병합하거나 아니면 지역 구분을 없애고 병합하는 방법을 고려할 수 있을 것이다.

제6절 소 결

본 연구에서는 사업체 패널조사의 패널 마모 처리방법의 하나로서 표본 추가를 할 때 사용하는 표본 추출방법론에 대한 대략적인 방향을 기술하였다. 표본 추출은 층화 추출을 하되 여러 가지 제한조건을 가지는 일종의 최적화 문제로 간주하여 linear programming 기법을 사용하여 구현한 후 이를 바탕으로 층화 추출을 실시하면 된다. 그 외에도 근로자 가중치 작성이나 무응답 처리 등에 대한 방법론에 대해서도 간단하게 기술하였다.

제 5 장
한국에서의 자료검증 연구의 필요성에 대하여[7)]

제1절 서 론

많은 국내외 실증분석 연구들은 상당 부분 서베이(Survey)를 통하여 획득된 자료에 근거하여 수행되어 왔다. 특히 노동경제학 분야에서 서베이 자료 의존도는 매우 높은 실정이다. 그러나 서베이 조사를 통하여 수집된 변수들은 응답오차나 코딩 에러 등 제요인들에 의해 심한 측정오차를 수반하고 있으며 이러한 측정오차는 경우에 따라 분석 결과를 왜곡시키게 된다. 외국의 경우 측정오차 문제의 심각성과 그 대처방안에 대해서 오래 전부터 연구가 진행되어 왔으나 국내의 경우 측정오차 문제에 대한 연구는 초기 단계에 있다고 할 수 있다.[8)]

한편에서는 이러한 측정오차의 문제를 통계 및 계량경제학적 기법을 이용하여 해소 및 완화시키고자 노력하여 왔으나 대부분의 노력들(예를 들어 도구변수 추정법)은 고전적 가정(classical assumption)이라는 지나치게 엄격한 가정하에 진행되어 왔다. 여기에서 고전적 가정이란 측정오차가 해당 변수의 참값, 분석에 포함된 다른 변수들의 참값, 다른 변수들

7) 중간보고서 심의에 참여하여 소중한 논평을 해주신 심사자분들께 감사의 말씀을 전한다.

8) 예외적으로 소득변수에 존재하는 히핑(heaping) 현상에 대한 연구는 추진되어 왔는데 이에 대해서는 홍민기 외(2014)에 소개된 문헌들을 참조하기 바란다.

의 측정오차, 나아가 교란항(stochastic disturbances)과 독립이라는 가정을 말한다. 따라서 이러한 고전적 가정이 만족되지 않을 경우 분석결과는 편의를 가질 수밖에 없다. 심지어 고전적 가정이 만족된다고 하더라도 연구 주제 및 목적에 따라 측정오차는 모수 추론에 편의를 갖는 결과를 낳게 할 수 있다. 예를 들어 연구자들은 종종 특정 근로자가 두 조사 시점 사이에서 직종을 전환하였는가를 판단하는 방법으로 두 조사 시점에서 획득한 직종 코드를 비교하곤 한다. 그러나 Mellow and Sider(1983), Bound et al.(2001) 등에 따르면 서베이를 통하여 수집된 직종변수의 측정오차는 매우 심각한 수준이며, 이 오차로 말미암아 (직종분류 단위에 따라 다르지만) 직종 전환확률의 추정치가 심히 과대평가되는 경향이 생긴다. 그것은 두 시점 사이에 실제로는 직종전환이 없었는데 측정오차에 의해 두 시점에서 서로 다른 직종 코드를 보고할 확률이, 실제로는 변했는데 측정오차에 의해 같은 코드를 보고할 확률보다 크기 때문이다. 유사한 논리로(제Ⅱ장에서 상세하게 소개하겠지만) 서베이를 통하여 획득된 소득변수로 계산한 소득변동성의 크기는 실제의 변동성을 항상 과대평가하는 방향으로 작용할 것이다. 보다 일반적으로 측정오차는 1차 적률보다는 2차 적률의 추정치에 편의를 가져다주게 된다.

이러한 현실에서 서베이 자료들에 나타난 각종 인구・사회・경제변수들의 측정오차가 고전적인 가정과 부합되는가를 검증하고, 나아가 측정오차의 성격과 그 크기를 연구하는 작업은 모수값 추정과 관련하여 매우 중요한 정보를 제공할 것이다. 비록 서베이 방식의 획기적인 혁명 없이는 연구자들이 항상 측정오차가 수반된 변수를 사용하여 분석할 수밖에 없지만, 사용되는 변수의 측정오차의 성격과 그 정도에 대한 정보를 알 수 있다면 해당 정보를 이용하여 (서베이 변수를 사용하여 도출된) 편의가 있는 결과를 보정할 수 있을 것이다. 바로 이러한 취지하에 미국의 경우 서베이 자료검증 작업이 실시되어 왔다. 대표적으로 PSID VS(PSID Validation Study)를 들 수 있는데, PSID VS는 기본적으로 주요 노동시장변수들에 대해 응답자들로부터 정보를 획득하고, 해당 응답자들이 소속되어 있는 직장으로부터 같은 변수에 대해 정보를 수집한 후 이 두 경로를 통하여 획득한 정보가 얼마나 일치하는가를 검증하는 방식을 취하

고 있다.[9] 그러나 (i) 주요 노동시장변수들에 대한 응답 방식에서 한국인과 미국인 사이에 차이가 존재할 수 있으며 (ii) PSID VS는 조사의 편리성을 위하여 특정 지역의 특정 기업에 한하여 실시함으로써 그 결과를 일반화시키기에는 무리가 따르며 (iii) 그나마 PSID VS 성격의 자료검증 연구는 세계적으로도 드문 실정이다.

이에 본고에서는 한국에서의 서베이 자료검증 연구(이하 KLI VS)의 필요성을 제기하고자 한다. 노동시장 관련 연구들에 한정할 경우에도 많은 연구들은 경제활동인구조사, 가계조사, KLIPS, 고령자패널조사, GOMS, 교육고용패널조사 등 서베이 자료에 근거하여 분석을 수행하여 왔으며 도출된 연구결과들은 직간접적으로 정책수립에 영향을 미치고 있다. 따라서 분석에 사용된 서베이 자료들에 나타난 측정오차에 의해 도출된 실증분석 결과가 어떤 방향으로 얼마나 편의를 가지게 되는가에 대한 연구는 학문적으로나 정책적으로나 매우 중요하다고 판단된다. 비록 서베이 자료검증에 대한 외국(주로 미국)의 연구결과가 존재하나 전술한 바와 같이 서베이 응답방식에서 국가 간 차이가 존재할 가능성이 있다. 한편 서베이 자료의 검증방식도 반드시 PSID VS 방식을 따를 필요는 없으나 한국의 경우 정부가 보유하고 있는 집계자료는 접근이 어렵고, 이용 가능하다고 하더라도 개인 차원에서의 다양한 정보가 존재하지 않는다. 이와는 달리 PSID VS처럼 개인이 응답한 결과와 해당 응답자가 소속된 직장에서 (예를 들어) 인사관리 담당자가 보관 기록에 근거하여 보고한 결과를 대조하는 방식을 따를 경우 성, 연령, 학력, 근속연수, 직무, 근로시간, 임금, 고용형태, 각종 인센티브 등 다양한 개인 특성들 뿐만 아니라 규모, 노조조직 등 다양한 사업체 특성들의 일치성을 검증해 볼 수 있는 장점이 있다. 이처럼 서베이를 통하여 얻는 주요 노동시장 관련 변수들에 나타난 측정오차의 성격과 정도를 연구할 필요가 있다.

제2장에서는 서베이 자료 중의 하나인 KLIPS 자료상에 나타난 측정

9) 물론 서베이를 통해 획득한 정보를 검증하기 위한 검증자료의 원천(validation source)은 다양하다. 대표적으로는 사용자가 소장하고 있는 기록과 행정 자료(administrative data)를 들 수 있으며, 심지어 같은 개인에 대한 재면접(reinterview) 조사와 같은 변수의 시계열상 혹은 다른 변수들과의 일치성 검증 등의 방법도 사용된다.

오차의 실태를 몇 가지 예를 들어 보고하며, 제3장에서는 이러한 측정오차 문제를 무시하고 단순히 서베이 자료를 이용하여 분석결과를 도출하였을 때 발생하는 분석결과의 편의 문제를 예를 들어 소개한다. 제4장에서는 자료검증 연구를 수행할 경우 얻어질 정보의 활용성을 PSID VS 결과를 이용하여 예시하며, 마지막으로 제5장에서는 결론을 내린다.

제2절 서베이 자료에 나타난 측정오차 사례

보다 정확한 자료와의 직접 대조과정 없이 서베이 변수에 나타난 측정오차를 판단하기는 쉽지가 않다. 특히 횡단면 자료의 경우는 더욱 그러하다. 다만 패널자료가 이용·가능한 경우 특정 변수가 시계열상에서 일관성을 유지하고 있는가에 대한 판단은 변수에 따라 어느 정도 가능하다. 예를 들어 성이나 출생연도는 시간의 경과상에서도 변하지 말아야 하며,[10] 교육수준도 두 조사 연도 사이에서 감소하거나 '지나치게' 증가하는 일이 없어야 할 것이다. 본 장에서는 KLIPS 자료를 이용하여 서베이 조사를 통해 나타난 변수에 존재할 수 있는 측정오차에 대해 몇 가지 예시를 들고자 한다. 물론 KLIPS 자료상에 나타난 측정오차의 성격이나 정도가 서베이 자료에 나타나는 측정오차의 성격과 정도를 대표한다고 할 수는 없다. 오히려 KLIPS는 패널자료이기 때문에 특정 변수에 대한 시계열상의 일치성을 검증할 수 있는 관계로 이미 공개된 자료상에 나타난 측정오차의 문제는 일반 서베이 자료상에 나타난 그것보다 덜 심각할 가능성도 있다. 그럼에도 불구하고 직접적인 자료 검증조사 없이 측정오차의 실태를 간접적으로나마 알아볼 수 있는 방법은 KLIPS와 같은 패널자료에서나 가능하다. 한편 여기에서 지칭하는 측정오차란 응답자의 응답오차나 조사기관의 코딩 에러 혹은 관리기관의 편집 에러를 구분하지 않고 최종적으로 공개된 변수에 내재한 오차를 의미한다.

10) 표본 기간 내에 성을 전환했거나 호적상 출생연도를 변경한 경우는 고려하지 않는다.

우선 가장 측정오차가 없을 것이라고 예상되는 변수인 응답자의 '성' 변수를 보자. 1998년부터 2012년까지 15차 조사 기간 동안 최소 두 번(2년) 이상 응답한 응답자들의 총 수는 20,485명이다. 이 중 응답한 성 정보가 조사상에서 완전하게 일치하지 않는 경우는 총 7명이다. <부표 1>에서는 이 7명에 대한 성 변수의 시계열값들을 보고하고 있다. 5명의 경우 시계열상 일관된 성 정보가 보고되고 있다가 어느 특정 연도에 1회에 한하여 다른 성 정보가 보고되고 있어 단순한 응답이나 코딩 에러로 볼 수 있으나 응답번호 212404나 349405처럼 연속적으로 특정 성을 보고하다가 특정 시점 이후 연속적으로 다른 성을 보고하는 경우도 있어서 이러한 변화가 실제의 성 전환을 의미하는지 아니면 코딩상의 문제인지(예를 들어 전 연도의 값을 무조건 입력) 조사가 필요해 보인다.

다음으로 출생연도 변수에 나타난 측정오차 상황을 검토해 보자. 15회차까지 최소한 두 번 이상 생년이 보고된 응답자의 수는 성 변수의 경우와 비슷하게 20,484명이다. 이 중 응답한 모든 연도에 대해 일관성 있게 모두 음력이나 양력으로 답하되 보고된 출생연도가 모든 응답연도들 사이에 완전하게 일치하지 않는 응답자의 수는 47명으로 나타나 명백한 측정오차의 경우로 분류된다. 이 중 개인번호 20605는 출생연도의 변화가 10년을 나타내고 있으며, 개인 번호 108401은 1951년으로 일관성 있게 보고되다가 1955년으로 5년 연속 응답한 후 다시 1951년으로 4년 연속 응답하고 있다. 출생연도가 달리 기록된 경우 중 음력과 양력을 달리하면서 보고한 경우도 적지 않다. 이 경우 보고된 출생월과 일을 이용하여 대조해 본 결과 27명은 출생연월이 일치하지 않는 것으로 나타났다. 이 불일치는 출생연도의 불일치 때문일 수도 있고 출생월이나 일의 불일치 때문일 수도 있다. 물론 성이나 출생연도 변수에 나타난 오차발생의 빈도는 실제 분석결과에 크게 영향을 주지 않겠지만 성이나 출생연도라는 변수의 성격을 고려하면 본 예는 서베이 변수에 존재하는 측정오차의 중요성을 의식하게 하기에 충분하다고 본다.[11]

---

11) 본 연구의 중가보고서 제출 후 한국노동연구원 KLIPS 담당 팀으로부터 2014년 현재(10월 2일) 공개되어 있는 1~15차 자료상에는 본문에서 언급한 바대로의 성이나 출생연도에서 불일치가 전혀 존재하지 않는다는 연락을 받았다. 그러나 본

노동경제학 연구에서 가장 흔하게 사용되는 변수들 중의 하나는 교육투자 수준이다. KLIPS는 응답자들로부터 매 조사시점에서 학력 상태를 학교, 이수여부, 그리고 학년으로 나누어 조사하고 있다. 우선 15차 조사기간 동안 최소 두 번 이상 학력 정보가 보고되어 있는 경우는 앞의 경우들과 유사하게 20,486명이다. 이 중 가장 명백한 측정오차의 사례로서 인접한 두 조사 연도 사이에 보고된 학력이 강등되었거나(예를 들어 대학교에서 고등학교로) 두 단계 이상 증가한 경우(예를 들어 고등학교에서 대학원으로)는 총 717명으로 나타났다. 이 중 강등된 후의 학력이 전문대인 경우는 261명으로 나타났다. 이 261명에 대한 보고된 학력의 저하 현상이 실제의 현상을 나타내는지, 아니면 측정오차인지 판단하기 위해서는 추가적인 검증작업이 필요할 것이나[12] 최소한 그 나머지 456명은 일단 측정오차의 경우로 분류될 수 있다. 나아가 261명 모두가 실제로 상위 학교에서 다시 2년제 대학을 선택한 결과라고 하더라도 연구목적에 따라서는 이들의 학력이 대학이나 대학원으로 기록되어야 할 경우가 많다. 예를 들어 교육투자의 효율/비효율성을 분석할 경우 우리의 관심은 실제의 투자연수(비록 그 투자가 잘못된 판단이었다고 하더라도)에 있어야 하기 때문이다. 이상의 논의는 학력을 학교 수준으로만 평가했을 경우이며 여기에 덧붙여 이수여부와 학년 정보를 추가하여 실제의 교육

연구에서 사용된 1~12차 공개자료와 13~15차 학술대회용 자료를 바탕으로 재확인해 본 결과 본고에서 보고한 바대로의 성이나 출생연도상의 불일치가 확인되었다(요청에 의해 자료 재전송 가능). 또한 10월 초 현재 공개된 1~15차 자료를 이용하여 검토해 본 결과 본문에서 보고한 불일치 문제가 완전히 해소되었음을 확인하였다. 이에 본 연구에서는 중간 보고 후 KLIPS 담당 팀에서 해당 불일치 문제를 수정한 것으로 추측한다. 그 추측의 또 다른 근거로서 분석은 하였으나 보고서의 간결성을 위하여 중간 보고 때 누락시킨 출생 관련 변수들의 불일치 문제들은 10월 2일 현재 공개된 파일들에도 여전히 존재한다는 점이다 (이 변수들에 대해서도 요청에 의해 공개 가능). 사실 변수들에 존재하는 오차를 교정하여 보다 정확한 변수값을 제공하는 것은 가치가 있는 일일 것이다. 문제는 측정오차의 교정 문제가 통상적으로 생각하는 것보다 훨씬 복잡다단한 성격의 것이어서 때로는 차라리 그대로 두는 것이 교정하는 것보다 더 나을 수 있다는 점이다. 그 예를 차후 근속기간의 측정오차 문제를 논할 때 제시하겠다.

12) 총 261명 중 4년제 대학에서 2년제 대학으로 감소한 경우가 253명, 대학원 석사에서 2년제 대학으로 감소한 경우가 7명, 대학원 박사에서 2년제 대학으로 감소한 경우가 1명이다.

연수를 도출하고 이 교육연수가 인접한 두 조사연도 사이에 감소하거나 2년 이상 증가한 경우를 포함하여 교육변수에 존재하는 측정오차의 규모를 측정할 경우 그 규모는 앞서 보여준 수치들이 나타내는 것보다 훨씬 커질 것이다. 이러한 학력변수에 나타난 측정오차는 회귀분석을 통하여 교육투자 수익률을 추정할 때 (만약 해당 오차가 고전적 가정을 만족할 경우) 추정된 수익률의 크기를 과소평가하는 방향으로 작용할 것이다.

가장 측정오차가 심할 것으로 판단되는 변수들의 예로서 임금이나 근로시간을 들 수 있다. 그러나 임금과 근로시간에 관한 한 패널자료상에 나타난 시계열상 일관성을 기준으로 측정오차를 가늠해 볼 수는 없다. 그것은 개개인의 임금이나 근로시간에 미치는 요인들이 너무나 많아서 기계적으로 통제하기가 어렵기 때문이다. 여기에서는 기존의 국내외 연구 결과들을 이용하여 임금 및 근로시간의 측정오차에 대한 정보를 개연적으로나마 소개하고자 한다. 우선 임금변수에 존재하는 측정오차의 성격 및 정도를 임금경직성이라는 기준에서 평가해 보자. 많은 기존 연구들은 경직성의 척도로서 두 시점 사이에서 직업 변동을 경험하지 않은 전체 근속자들 중 명목임금의 동결을 경험한 사람들의 비중을 사용하여 왔다(예를 들어 MacLaughlin, 1994; Card and Hyslop, 1996; Kahn, 1997; Altonji and Devereux, 1999; Smith, 2000; Nickell and Quintini, 2003; Elsby, Shin, and Solon, 2013). 여기에서 만약 측정오차가 고전적 가정을 만족할 경우 서베이 자료로 추정한 임금경직성의 정도는 실제의 경직성을 항상 과소평가하는 방향으로 나타날 것이다. 그것은 실제로 명목임금의 동결을 경험한 근속자가 측정오차에 의해 임금 변동을 경험한 것으로 관찰될 확률이 실제로는 변동을 경험한 근속자가 오차에 의해 동결을 경험한 것으로 관찰될 확률보다 더 크기 때문이다. 그러나 임금변수에 존재하는 측정오차가 고전적 가정을 만족하지 않을 경우 그 결과는 달라질 수 있다. 흔히 서베이를 통하여 추출된 임금변수는 심한 라운딩(rounding) 오차를 안고 있다고 알려져 왔다(Dickens et al., 2006; Elsby et al., 2013). 예를 들어 전년도에 월 95만 원을 받았던 근로자가 다음 연도에 105만 원을 받은 경우 해당 노동자는 임금이 10% 이상 증가했음에도 불구하고 두 연도에 모두 100만 원을 받았다고 보고하여 임금이 동결

되는 것처럼 관찰되는 경향이 있다. 이러한 형태의 측정오차로 말미암아 관찰된 임금변수로 계산한 임금경직성은 실제의 임금경직성 정도를 과대평가하는 경향이 생긴다. Elsby et al.(2013)이 강조하였듯이 서베이 자료에는 두 형태의 측정오차가 모두 존재하며, 어느 형태의 측정오차가 더 지배적인가는 실증적으로 판단될 성격의 것이다. 실제로 Elsby et al.(2013)은 미국의 서베이 자료인 Current Population Survey(CPS) 자료와 영국의 임금대장에 기초한 자료인 New Earnings Survey(NES) 자료로 임금경직성의 정도를 비교 분석하면서 측정오차로부터 비교적 자유로운 NES 자료상에서 임금경직성의 정도가 훨씬 낮게 나타난 것으로 보아 서베이 자료에 나타난 임금경직성의 정도는 라운딩 오차에 의해 과대평가되어 있다고 하였다.

한국의 경우는 어떠한가? 우선 박선영・신동균(2014)은 이러한 기존 연구방법을 따라서 KLIPS 자료에 나타난 명목임금의 경직성을 분석하였다. 동 연구는 KLIPS 자료의 월평균 급여변수를 이용하여 25~59세 사이의 상용직 임금근로자 중 인접한 두 조사연도 사이에 직장변동을 경험하지 않았고 초과근로를 수행한 적이 없으며 성과급을 적용받지 않은 근로자들에 대해 분석을 시도하였다. 그 결과 전체 근속자 중 임금동결을 경험한 근속자들의 비중은 가장 작게는 2004-2003 기간의 14.3%에서 2011-2010의 27%로 나타났다. 여기에서 라운딩 오차 가설을 간접적으로 검토해 보기 위하여 박선영・신동균(2014)은 전년도 기준으로 응답한 월급여의 값이 50만 원의 배수인 경우(50만 원, 100만 원, 150만 원, 200만 원 등)를 표본에서 제거하고 재추정하였다. 그 결과 표본의 규모는 전체적으로 약 40% 정도 감소하였으며, 추정된 명목임금 경직성의 정도는 가장 낮게는 1999-1998의 10.7%에서 가장 높게는 2012-2011의 16.9%로 나타나 라운딩 효과를 제거하지 않았을 때와 비교하여 낮아졌음을 알 수 있다. 물론 라운딩 효과가 50만 원의 배수가 아니라 10만 원의 배수에서 나타날 수 있음을 고려하면 실제 라운딩 오차의 규모는 본 예시에 나타난 것보다 훨씬 더 클 수 있음을 박선영・신동균(2014)는 역설하고 있다. 보다 직접적으로 Park et al.(2014)은 고용노동부의 직종별임금실태 조사자료를 이용하여 임금경직성의 정도를 재추정하였다. 사업체 조사자료

성격상 추출된 임금자료는 임금대장에 기초한 자료이기 때문에 측정오차로부터 상대적으로 자유로울 것으로 판단되었다. 분석결과 측정된 임금경직성은 앞서 사용한 KLIPS 표본의 제약과 가장 유사한 제약하에서 약 2% 정도로 매우 낮게 나타나 Elsby et al.(2013)의 결과를 재확인시켜주고 있다. 요약하면 한국의 경우도 서베이를 통하여 얻어진 임금변수는 고전적 측정오차보다는 라운딩 오차를 강하게 내포하게 되며, 이에 따라 서베이 변수를 이용하여 측정한 임금경직성의 정도는 실제의 경직성을 과대평가하게 된다.

한편 근로시간의 경우 시계열상의 일치성을 기준으로 오류를 가늠해 보기가 더욱 어렵다. 그럼에도 불구하고 박선영・신동균(2014)은 앞에서 소개한 50만 원 단위에서 라운딩 현상을 제거한 표본을 이용하여 두 조사시점 사이의 정규 근로시간 변동의 분포를 도출해 보았다. 동 연구에 제시된 상대돗수 분포를 <부록 2>에 옮겨놓았다. 그림에서 알 수 있듯이 상당히 안정적인 집단(상용직 임금근로자 중 일자리 변동이 없고 초과근로를 수행한 적이 업고 성과급제도 적용되지 않은 직장)에서의 정규 근로시간은 심한 변동을 보이고 있다. 두 조사시점에서 정규 근로시간의 변동이 없었던 응답자의 비중은 40%를 약간 상회할 정도이며 정규 근로시간이 12시간 이상 증가 혹은 감소한 집단의 규모도 12%를 상회한다. 이 변동 중 어느 정도가 실제의 변동이며 어느 정도가 오차에 의한 것인지는 자료검증 연구를 통해 판단되어야 할 것이다.

다음으로는 산업 및 직종 변수의 측정오차 상황에 대해 생각해 보자. 이 변수들 역시 검증자료가 이용 가능하지 않기 때문에 측정오차의 규모에 대한 판단은 어렵다. 시계열상에서의 일관성도 해당 변수들 시간가변 변수이기 때문에 판단의 기준으로 삼기가 어렵다. 다만 두 조사시점 사이에서 직업 변동이 없었던 근로자들의 경우 산업 및 직종의 변동도 발생하지 않을 것이라는 가정하에 보고된 산업 및 직종 변수들의 시계열상 일관성을 검토해 볼 수 있다. 이를 위해서는 박선영・신동균(2014)이 사용한 표본 1과 표본 3을 이용한다. 동 연구는 KLIPS가 매 조사시점에 보고하는 현 일자리의 시작시점에 대한 정보를 이용하여 일자리의 변동이 있었는지를 판단하고 있다. 즉 조사시점에서 일자리의 시작시점이 전

년 조사시점 이전이면 두 조사시점 사이에는 일자리 변동이 없었다고 정의하고 이를 근속자라고 칭하였다. 물론 이 조사시점 변수에도 측정오차가 존재할 수 있다. 그러나 본 연구에서 조사시점 변수로부터 현 일자리에서의 근속기간을 유도하고 이 근속기간 변수의 두 조사시점 사이의 차이와 두 조사시점 사이의 시간 거리를 비교해 본 결과 약 99%의 일치성을 발견하였다. 저자들은 이러한 높은 일치율이 근속연수의 측정오차를 줄이기 위한 KLIPS의 자료 편집작업 때문이라고 판단한다. 그 의미에 대해서는 아래에서 추가적으로 언급하겠다. 한편 박선영・신동균(2014)의 표본 1에서는 표본 제약으로 25~59세, 학생 제외, 월근로시간이 정해져 있는 임금근로자를 부과하였으며, 또한 유도된 임금 분포에서 상하 1%씩 극단값들을 제거하였다. 표본 3에서는 이에 추가하여 초과근로를 수행하였거나 성과급이 적용되는 사업체에서 근로한 응답자들, 나아가 임시・일용직 근로자들도 표본에서 제거하였다.

우선 <표 5-1>에 나타난 산업변수의 측정오차 문제를 살펴보자. 1998년부터 2012년까지의 조사자료를 바탕으로 앞서 도출한 표본(인접한 두 조사시점 사이의 비교가 가능한 경우만)의 규모는 표본 1과 3의 경우 각각 29,363개와 15,984개로 나타났다. 이 중 제3열에 제시된 대분류상 불일치 표본이란 일자리 변동이 없었던 근속자들 중 두 조사시점에서 보고한 산업이 대분류 기준에서 일치하지 않는 경우를 말하며 733개(2.5%)에 이른다. 대분류상에서는 일치하지만 중분류상에서는 불일치하는 표본의

〈표 5-1〉 산업(업종)변수의 측정오차

	대상 표본	정상 표본	측정오차표본			
			대분류상 불일치 표본	대분류상 일치하지만 중분류상 불일치표본	대분류 및 중분류상 일치하지만 세분류상 불일치표본	소계
표본 1개 (비중)	29,363 (100%)	27,538 (93.78%)	733 (2.50%)	587 (2.00%)	505 (1.72%)	1,825 (6.22%)
표본 3개 (비중)	15,984 (100%)	15,067 (94.26%)	374 (2.34%)	251 (1.57%)	292 (1.83%)	917 (5.74%)

수는 587개(2%)이며 따라서 중분류상에서 불일치하는 표본의 총규모는 1,320(4.5%)개이다. 이를 소분류까지 확대하면 총 1,825개(6.22%)의 표본이 세분류상에서 불일치성을 보이고 있다. 물론 세세분류 이상으로 확대하면 불일치하는 표본의 규모는 더욱 증가할 것이며 이에 따라 정상표본의 규모는 줄어들 것이다. 그러나 산업을 보다 자세하게 분류할수록 불일치로 보고되는 경우가 사실 오류가 아니라 실제의 변동일 확률이 증가할 수 있다. 한편 표본 1에서 임시·일용직을 제거할 경우(표본 3) 불일치성을 갖는 표본의 규모는 약간 감소하는 경향을 보인다.

한편 <표 5-2>에서 볼 수 있듯이 직종면에서의 불일치성은 보다 크게 나타나고 있다. 표본 1을 사용할 경우 대분류상에서조차 불일치성을 갖는 표본의 규모가 6.11%로 나타났으며 세분류상에서 오차라고 의심받는 표본의 규모는 총 비교대상 표본의 10.53%로 나타나고 있다. 산업의 경우와 마찬가지로 직종변수도 임시·일용직을 표본에서 제거할 경우 측정오차의 규모는 약간 줄어들지만 여전히 산업변수보다 오차의 규모가 심함을 알 수 있다. 물론 이러한 시계열상의 일치성을 검토하는 방법은 산업 및 직종변수의 측정오차를 간접적으로 검토해 보는 방법이며, 보다 직접적으로는 측정오차로부터 자유로운 정보를 획득하고 서베이 정보를 이와 대조하여야 할 것이다.

측정오차가 산업보다는 직종에서 더 크게 나타나는 현상은 미국의 검증연구 결과와 일치한다. 그러나 Bound et al.(2001)에 소개된 자료검증

〈표 5-2〉 직종변수의 측정오차

	대상 표본	정상 표본	측정오차표본			
			대분류상 불일치 표본	대분류상 일치하지만 중분류상 불일치표본	대분류 및 중분류상 일치하지만 세분류상 불일치표본	소계
표본1 개(비중)	29,184 (100%)	26,112 (89.47%)	1,784 (6.11%)	649 (2.22%)	639 (2.19%)	3,072 (10.53%)
표본3 개(비중)	15,899 (100%)	14,434 (90.79%)	858 (5.40%)	300 (1.89%)	307 (1.93%)	1,465 (9.21%)

연구결과들을 종합해 보면 한국의 경우 앞서 제시한 <표 5-1>과 <표 5-2>의 수치들이 미국의 경우보다 현저하게 낮음을 알 수 있다. 예를 들어 Mellow and Sider(1983)는 두 서베이 자료, 즉 CPS를 통하여 획득한 산업정보와 Employment Opportunity Pilot Project로부터 획득한 정보를 사용자로부터 얻은 산업정보와 비교하여 1단위(대분류)와 3단위(세분류) 상에서 두 자료의 일치성을 보고하였는데, 이 중 일치성이 더 높은 (따라서 측정오차 문제가 덜 심각하게 나타난) CPS 검증자료 결과를 보면 1단위에서는 92.3%, 3단위에서는 84.1%의 일치율을 보이고 있다. 또한 동 연구가 보고한 직종변수의 일치율은 이들보다 현저하게 낮음을 알 수 있다(CPS 자료의 경우 1단위에서 81%, 3단위에서 57.6%). 이 두 연구의 차이는 (i) 박선영・신동균(2014)과 Mellow and Sider(1983) 사이의 검증방법상의 차이 때문일 수도 있으며 (ii) 박선영・신동균(2014)이 사용한 표본은 변동성이 작은 매우 안정적인 집단이기 때문에 이 표본을 이용하여 계산한 산업 및 직종의 불일치성도 비교적 작게 나타난 것일 수 있으며 (iii) KLIPS가 자체적으로 변수의 측정오차를 줄이기 위해 사후적인 편집과정을 거쳤기 때문일 수 있다. 주지하는 바와 같이 CPS는 기본적으로 횡단면 자료이기 때문에 시계열상의 일치성을 검증하기가 어려운 반면 KLIPS는 패널자료이므로 어느 정도까지는 자체 검증작업이 가능하다.

이는 앞서 언급한 KLIPS 자료에 나타난 근속연수의 정확성에 대한 이슈를 떠올리게 한다. 기존의 수많은 외국 연구들이 지적하였듯이 PSID나 NLS(National Longitudinal Survey) 자료상에 나타난 근속연수의 측정오차 문제는 심각한 수준이다(예를 들어 Altonji and Shakotko, 1987; Topel, 1991; Brown and Light, 1992). Topel(1991)의 연구를 인용하면 PSID 백인 남성 자료를 1968~1983년 기간에 대해 분석해 보면 인접한 두 조사연도 사이에 이직을 경험하지 않은 근속자들이 보고한 근속연수의 변화는 -31년에서 7.5년 사이로 광범위하게 나타난다. 이와 비교하면 KLIPS 자료에 나타난 근속연수 정보는 놀라울 정도로 정확하다. 검토결과 특정 조사시점에서 KLIPS가 보고하고 있는 근속개월 수의 두 조사연도 사이의 차이는 모든 경우에서 정확하게 두 조사시점 사이의 거리와

같게 나타나고 있다. PSID뿐만 아니라 많은 외국 패널조사상에서 나타난 결과와 비교해 볼 때, 모든 응답자들이 근속기간, 그것도 개월 수로 보고한 근속기간이 정확하게 두 조사시점 사이의 거리와 같을 확률은 사실상 영(0)에 가깝다고 볼 수 있다. 이처럼 KLIPS 자료상에 나타난 근속개월 수 정보가 정확한 이유는 KLIPS가 응답자들이 보고한 근속기간 정보를 기록한 것이 아니라 조사시점 정보와 여타 정보를 이용하여 자체적으로 근속기간 정보를 제조하였기 때문이라고 추측된다. 물론 자체 편집작업을 통하여 사용자들에게 보다 정확한 변수값을 제공하는 것은 가치있는 일일 것이다. 그러나 많은 경우 서베이를 통하여 수집된 정보는 그 성격이 매우 복잡·다단하여 단순한 논리에 근거하여 수정할 경우 오히려 측정오차의 문제를 더욱 심각하게 만들 수 있다.

근속기간 변수를 이용하여 예를 들어 보자. 만약 추측한 대로 KLIPS가 다양한 정보(예를 들어 인터뷰 시점과 전년 조사시점 이후 이직을 경험하였는지의 여부)를 이용하여 근속자의 경우 두 조사시점 사이의 근속기간 변화를 정확하게 두 시점 사이의 시간거리와 같게 맞추었다고 하자. 문제는 이러한 방식으로는 근속기간 중 표본에 처음으로 관찰되는 연도에서의 근속기간값을 보정하기가 불가능하다는 점이다. 이 경우 첫 연도의 조사시점 기준 근속기간은 (i) 응답자가 보고한 값을 그대로 사용하거나 (현 직장에서 처음으로 일하기 시작한 시점) (ii) 특정 값(1년, 0.5년 혹은 0년)을 대입하는 일일 것이다. Altonji and Shakotko(1987)는 어느 방식을 따라도 추정된 근속기간에 체계적인 편의가 발생할 수 있음을 보이고 있다. 동 연구에 의하면 가장 합리적인 보정방법은 응답자가 매년 보고한 근속기간 정보를 모든 연도에 대해 다 사용하는 것이다. 예를 들어 특정 응답자가 3개 연도에만 연속으로 조사에 응하였다고 가정하자. 첫 연도 응답에서는 현 사용자와 10년 전부터 일을 시작하였다고 보고했으며, 두 번째 및 세 번째 연도에서는 모두 이전 조사시점 이래로 이직을 경험한 바가 없으며 근속기간이 각각 2년 및 3년 되었다고 보고하였다고 가정하자. 이 경우 우리는 “아마도 첫 연도 근속연수가 실제로는 1년인 것이 10년으로 잘못 보고되었다”는 가정하에 세 연도의 근속연수들을 각각 1, 2, 3년으로 기록할 수 있다. 유사한 판단 근거로서 보고된 세 연도

의 근속연수가 (1, 2, 8)인 경우 이를 내부적으로 일치시키기 위하여 (1, 2, 3)으로 기록할 수도 있다. 그러나 Altonji and Shakotko(1987)에 의하면 (비록 확률이 상대적으로 낮지만) 첫 번째 경우 첫 10년의 보고가 정확한 값이며 나머지 두 연도에 보고된 값들이 동시에 잘못될 가능성도 있다. 마찬가지로 두 번째 경우도 첫 두 연도의 값들이 오류를 포함하고 있고 세 번째 연도에 보고한 근속기간만이 정확한 값일 수 있다. 결국 동 연구는 특정 개인이 근속기간 내에서 조사연도별로 보고한 모든 값들을 이용하여 근속기간 변수를 제조해야 하며, 각 개인별로 근속기간에 대한 시계열 자료를 이용하여 가중회귀모형을 추정함으로써 현 직장을 시작한 시점에 대한 가장 효율적인 추정치를 얻을 수 있다고 하였다.[13)]

이러한 기존의 연구결과들을 고려해 볼 때 KLIPS 자료상에 나타난 근속기간 변수는 비록 두 인접한 조사 사이의 변화라는 면에서는 정확한 값을 제공하고 있으나, 근속기간 중 조사에 처음 응답한 시점에서 보고한 근속기간에 오류가 있을 경우 결국 모든 조사연도에서의 근속기간이 체계적으로 과대평가되거나 과소평가되는 현상을 만들어내고 있다. 이는 연구 주제에 따라 분석결과를 심각하게 왜곡시킬 수 있다. 예를 들어 임금결정에 있어서 현물시장 이론이 타당한지 아니면 암묵적 계약이론이 설득력이 있는지를 실증적으로 검증하는 방법으로 많은 기존의 연구들은 (예를 들어 Beaudry and DiNardo, 1991) 특정 시점에서의 임금이 (i) 현 사용자와 고용관계를 맺은 시점에서의 실업률(최초 실업률) (ii) 그 시점 이래로 경기가 가장 좋았을 때의 실업률(최저 실업률), 그리고 (iii) 현재 실업률 중 어느 변수에 의해 보다 잘 설명되는가를 회귀분석을 통하여 판단하였다. 만약 현 근속연수의 시작시점에 측정오차가 존재할 경우 이는 최초 실업률을 잘못 사용하게 되는 결과를 낳게 되며, 최초 실업률 계수의 추정치를 통계적으로 무의미하게 만드는 방향으로 작용하여 고전적인 암묵적 계약이론을 기각하는 방향으로 작용할 것이다.

13) 가중치 사용방법과 해당 알고리즘에 대해서는 동 연구의 부록을 참고하기 바란다.

제3절 측정오차를 무시했을 때 발생할 추정결과상 편의 예시

제2장에서는 노동시장 분석에서 흔히 사용하는 변수들에 나타나는 측정오차의 예를 KLIPS 자료를 이용하여 소개하였다. 여기에서는 이러한 서베이 변수에 존재하는 측정오차를 무시하였을 때 분석결과에 어떤 영향을 미치는가에 대해 논한다. 물론 그 영향의 성격 및 정도는 연구주제 및 변수의 성격에 따라 다르게 나타날 것이므로 이 논의 역시 포괄적이라기보다는 예시의 성격을 지닌다.

우선 노동시장 분석에서 흔히 사용되는, 그리고 측정오차의 문제가 상당히 심각할 수 있는 임금변수를 이용하여 예를 들어 보자. 임금의 참값을 W, 임금변수에 존재하는 측정오차를 V, 그리고 서베이를 통하여 얻어진 임금변수를 W^*로 표기하자. 편의상 모든 임금변수는 로그 변환된 변수라고 하자. 그러면 특정 개인 i에 대해

$$W_i^* = W_i + V_i \tag{1}$$

가 성립된다. 우선 초기 단계에서 논의를 간단하게 하기 위해 측정오차항은 고전적 가정을 만족한다고 하자. 즉 모든 개인 i에 대해 $E(V_i) = 0$이고 $Var(V_i) = \sigma_V^2$, 서로 다른 개인들($i \neq j$)에 대해 $Cov(V_i, V_j) = 0$가 성립되며, 모든 개인들(i 및 j)에 대해 $Cov(W_i, V_j) = 0$가 성립한다고 하자. 한편 $E(W_i) = \mu_W$, 그리고 $Var(W_i) = \sigma_W^2$라고 하자. 명백하게

$$E(W_i^*) = \mu_W, \;\; Var(W_i^*) = \sigma_W^2 + \sigma_V^2 \tag{2}$$

가 성립한다. 이에 따라 모평균과 모분산에 대한 추정치들로 각각 표본평균과 표본분산을 사용할 경우 비록 측정오차를 수반한 변수(W^*)로 계산한 표본평균은 실제의 평균임금에 대한 불평추정량이 되나 같은 변수로 계산한 표본분산은 실제의 임금불평등 정도를 과대평가하게 된다. 즉 다른 조건이 같을 경우 측정오차로부터 보다 자유로운 집계자료로 계산

한 임금(내지 소득) 불평등 정도보다 측정오차를 수반한 서베이 자료로 계산한 임금(내지 소득) 불평등도가 더 크게 나타나는 경향이 있다.

이러한 측정오차의 존재는 임금의 횡단면적 분포에 영향을 줄 뿐만 아니라 임금변동에도 영향을 미친다. 이를 검토해 보기 위하여 식 (1)을 패널자료에 근거한 식으로 바꾸어 보자.

$$W_{it}^{*} = W_{it} + V_{it} \tag{3}$$

역시 논의를 간략하게 하기 위해 측정오차의 고전적 가정이 성립된다고 하자. 즉 모든 i 및 t에 대해 $E(V_{it}) = 0$, $Var(V_{it}) = \sigma_V^2$, $E(W_{it}) = \mu_{Wt}$, 그리고 $Var(W_{it}) = \sigma_V^2$이 성립되며, 모든 개인에 대해 $t \neq s$일 때 Cov, $(V_{it}, V_{is}) = 0$ 그리고 모든 t 및 s에 대해 $i \neq j$일 때 $Cov(W_{it}, V_{js})$ $= 0$가 성립된다. 우선 연구의 관심이 두 시점 사이에 개개인이 경험하는 임금변동의 분포에 있다고 하자. 이는 임금변동성(wage volatility), 임금경직성(wage rigidity) 내지 유연성(wage flexibility), 임금의 경기변동성(wage cyclicality), 소득이동성(income mobility) 및 소득불평등(income inequality) 등 중요한 노동경제학 주제들을 연구하는 데 흔히 사용되는 변수이다. 식 (2)를 $(t-1)$기에 대해 다시 쓰고 이를 식 (2)에서 차감하면

$$\triangle W_{it}^{*} = \triangle W_{it} + \triangle V_{it} \tag{4}$$

로 표시되며, 여기에서 $\triangle$는 $(t-1)$기와 t기 사이의 차분을 의미한다. 우선 임금변동의 평균과 분산을 계산해 보면,

$$E(\triangle W_{it}^{*}) = E(\triangle W_{it}),\ Var(\triangle W_{it}^{*}) = 2\sigma_W^2(1-\rho_1) + 2\sigma_V^2 \tag{5}$$

여기에서 ρ_1는 참값을 나타내는 임금변수의 1차자동상관계수를 나타낸다. 즉 관찰된 임금변수(W^{*})로 정의되는 평균임금증가율은 실제의 임금변수(W)에 근거한 평균임금증가율과 같으나, 관찰된 변수로 계산한 임금변동성은 실제의 임금변동성을 $2\sigma_V^2$만큼 과대평가하게 된다. 이러한 현상은 임금(내지 소득) 변동성에 따라 발생하는 후생경제비용을 과대평

가하는 방향으로 작용한다.

한편 기존의 연구들은 명목임금의 유연성을 연구함에 있어서 그 척도로서 직장변동을 경험하지 않은 근속자들을 대상으로 두 시점 사이 명목임금의 삭감확률을 사용하여 왔다(Smith, 2000; Nickell and Quintini, 2003; Elsby, Shin and Solon, 2013). 한편 박선영・신동균(2014)은 측정오차가 고전적 가정을 따를 경우 이로 인하여 명목임금 삭감확률이 과대평가될 수 있음을 보였다. 동 연구의 논거를 따르면, (논의의 일반성 상실 없이) W와 V가 정규분포를 따르고 $\Phi(\,)$가 표준정규 분포를 따르는 확률변수의 누적분포함수라고 할 때,

$$Prob(W_t^* - W_{t-1}^* < 0) = \Phi\left(\frac{-(\mu_{W_t} - \mu_{W_{t-1}})}{\sqrt{2(\sigma_W^2 + \sigma_V^2)}}\right) > \tag{6}$$
$$Prob(W_t - W_{t-1} < 0) = \Phi\left(\frac{-(\mu_{W_t} - \mu_{W_{t-1}})}{\sqrt{2}\,\sigma_W}\right)$$

가 성립한다. 물론 이 부등식은 $\mu_{W_t} - \mu_{W_{t-1}} > 0$라는 조건하에 성립한다. 이 조건은 근속자의 명목임금이 생애주기상 실질생산성 증가라든가 물가상승의 요인에 의해 증가하는 한 성립된다. 결국 측정오차를 수반한 변수로 계산한 명목임금의 유연성은 실제의 유연성을 과대평가하게 된다. 얼마나 과대평가할 것인가는 참임금변수의 분산과 오차 분산의 상대적 크기, 그리고 생애주기상에서 임금의 연간 성장률 크기에 의해 결정될 것이다. 그렇다면 한국의 경우 서베이 자료로 계산한 명목임금의 유연성은 실제의 유연성을 어느 정도로 과대평가할 것인가? 우선 박선영・신동균(2014)은 KLIPS 자료로 명목임금 삭감확률을 추정하였다. 앞서 소개한 표본, 즉 두 조사연도 사이에서 직장변동을 경험하지 않았고 25~59세 사이이며 초과근로를 수행한 적이 없고 성과급을 적용받지 않았으며 상용직 임금근로자였던 응답자들을 대상으로 명목임금 삭감을 경험한 응답자들의 비율을 추정한 결과 1990~2000년 및 2000~2001년 두 차분연도 평균이 25.8%로 나타났다. 한편 Park et al.(2014)은 측정오차로부터 비교적 자유로운 직종별임금실태 조사자료를 이용하여 가능한 한 박선영・신동균(2014)이 사용한 표본 제약과 유사한 제약하에 재분석한 결과

두 차분연도 평균이 23.5%로 나타나 비록 서베이 자료로 추정한 비율이 약간 더 크게 나타났으나 그 차이가 그다지 크지 않음을 알 수 있었다. 이 결과는 정확히 Elsby et al.(2013)의 결과와 일치한다. 앞서 소개한 명목임금 경직성 추정에 대한 실증분석 결과와 종합하면 서베이 자료로 추정한 명목임금 경직성의 정도는 실제의 경직성 정도를(라운딩 오차로 인하여) 상당히 과대평가하는 경향이 있으나, 서베이 자료로 추정한 명목임금 유연성의 정도는 실제 값과 그다지 다르지 않은 것으로 나타났다.

다음에서는 서베이 변수에 존재하는 측정오차가 회귀분석 결과에 어떠한 영향을 미칠 수 있는가에 대해 몇 가지 경우를 통하여 알아보자. 우선 논의를 간단하게 하기 위해 다음과 같은 단순회귀모형을 이용하자.

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i \tag{7}$$

여기에서 등식 (7)은 모든 고전적 가정을 만족한다고 하자. 우선 변수의 측정오차는 피설명변수(Y)나 설명변수(X) 어디에도 존재할 수 있다. 예를 들어 등식 (7)에서 피설명변수를 임금, 그리고 설명변수를 근속연수라고 할 때 앞서 보았듯이 두 변수 모두 심한 측정오차를 반영하고 있음을 알 수 있다. 교과서적인 논의에 의하면 피설명변수에 존재하는 측정오차는 적어도 해당 측정오차가 고전적 가정을 만족하는 한(예를 들어) β_2에 대한 최소자승 추정량에 편의를 주지 않지만, 설명변수에 존재하는 측정오차는 고전적 가정을 만족한다고 하더라도 β_2에 대한 최소자승 추정량의 절대값을 줄이는 방향으로 작용하게 된다(attenuation inconsistency). 이 불일치성의 성격 및 논거에 대해 간단히 복습해 보자. 이는 등식 (1)에서 기호를 W에서 X로 바꾸어 놓고 V항에 대한 이전과 같은 고전적 가정을 설정해 놓은 경우에 해당한다. 달리 표현하면 $X_i^* = X_i + V_i$이며 $V = (V_1, V_2, ..., V_N)'$라는 확률변수들의 열벡터는 $X = (X_1, X_2, ..., X_N)'$ 및 $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_N)'$과 독립이며, 나아가 $V \sim N(\underline{0}, \sigma_V^2 I_N)$이다. 여기에서 $\underline{0}$은 영벡터, I_N은 $N \times N$의 항등행렬을 나타낸다. 등식 (7)을 실제로 추정하게 되는 식, 즉 측정오차가 수반된 변

수를 이용하여 정의되는 식으로 다시 표현하면 다음과 같다.

$$Y_i = \beta_1 + \beta_2 X_i^* + (\epsilon_i - \beta_2 V_i) \tag{8}$$

여기에서 β_2에 대한 최소자승 추정량이 불일치성을 가지게 됨은 설명변수와 오차항의 상관관계에서 파악될 수 있다. 앞서 제시한 고전적 가정들을 이용하여 검토해 보면 $Cov(X_i^*, \epsilon_i - \beta_2 V_i) = -\beta_2 \sigma_V^2 \neq 0$로서 측정오차가 존재하는 한 β_2에 대한 최소자승 추정량은 생략된 변수(V_i)에 의해 불일치성을 가지게 된다. 역시 교과서적인 논의에 의하면

$$plim\widehat{\beta_2} = \beta_2 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_V^2}\right) \tag{9}$$

로서 측정오차가 존재하는 한 대표본에서도 최소자승 추정치는 실제 β_2의 절대값을 과소평가하게 된다. 얼마나 과소평가하게 되는가는 참값의 분산과 측정오차 분산의 상대적 크기에 의해 결정될 것이다.

흔히 선형회귀모형에서 설명변수의 내생성에 대처하는 방법으로 도구변수를 사용한다. 그러나 설명변수의 내생성이 구조적인 요인들(예를 들어 연립방정식 모형이라든가 동태적 구조 혹은 변수의 생략)에 의해 발생하는 것이 아니라 등식 (8)에서처럼 측정오차에 의해 발생한 것이라면 도구변수접근법은 태생적 한계점을 가진다. 관찰되는 변수(X_i^*)와 교란항이 모두 공통의 측정오차항(V_i)을 포함하고 있기 때문에 측정오차가 포함되어 있는 설명변수와 상관관계가 없으면서 교란항과 상관관계가 없는 도구변수를 찾기는 쉽지가 않다.

이러한 논의를 최근 노동시장 분석에서 흔히 사용되는 패널자료를 이용한 모형으로 확대하여 보자. 교과서적인 논의에 의하면 개인고정효과를 통제할 목적으로 패널자료를 이용하여 차분모형을 추정할 경우 설명변수에 존재하는 측정오차는 추정된 회귀계수의 편의 문제를 더 심각하게 만들 수 있다. 이를 간단하게 예시하기 위해 등식 (7)을 패널자료에 근거한 모형으로 바꾸어 보자.

$$Y_{it} = \beta_1 + \beta_2 X_{it} + \alpha_i + u_{it} \tag{10}$$

한편 $X_{it}^{*} = X_{it} + V_{it}$에서 모든 i 및 t에 대해 $E(V_{it}) = 0$, 그리고 $E(V_{it}V_{js})$는 $i = j$이고 $t = s$인 경우는 σ_V^2, $i = j$이고 $|t-s| = k$인 경우 $\rho_{kV}\sigma_V^2$(ρ_{kV}는 k기 떨어져 있는 V항들의 자동상관계수를 나타냄), 그리고 $i \neq j$인 경우는 0이 된다. 유사하게 모든 i 및 t에 대해 $E(X_{it}) = 0$, 그리고 $E[(X_{it}-\mu_X)(X_{js}-\mu_X)]$는 $i = j$이고 $t = s$인 경우는 σ_X^2, $i = j$이고 $|t-s| = k$인 경우 $\rho_{kX}\sigma_X^2$, 그리고 $i \neq j$인 경우는 0이 된다. 마지막으로

$$Cov(X_{it}u_{is}) = Cov(X_{it}V_{is}) = Cov(\alpha_i V_{it}) = Cov(u_{it}V_{is}) = 0$$

이라고 가정하자. 등식 (10)을 관찰되는 변수(X_{it}^{*})로 다시 표시하면

$$Y_{it} = \beta_1 + \beta_2 X_{it}^{*} + \alpha_i + (u_{it} - \beta V_{it}) \tag{11}$$

이 된다.

여기에서 두 가지 경우를 비교해 보자. 첫째는 개인고정효과를 통제하지 않고 특정 연도 t에서 추출된 횡단면자료를 이용하여 등식 (11)을 단순히 최소자승법으로 추정한 경우를 생각해 보자. 중간 과정을 생략하고 도출된 결과만 보고하면 다음과 같다.

$$plim\widehat{\beta_2} = \beta_2 + \frac{Cov(X_{it}\alpha_i)}{\sigma_X^2 + \sigma_V^2} - \beta_2 \frac{\sigma_V^2}{\sigma_X^2 + \sigma_V^2} \tag{12}$$

여기에서 우변의 두 번째 항은 생략된 변수(개인고정효과)에 의한 추정량의 편의를, 그리고 마지막 항은 측정오차에 의한 추정량의 편의를 나타낸다. 물론 개인고정효과를 통제하지 못함으로써 발생하는 편의의 크기나 방향은 $Cov(X_{it}\alpha_i)$에 의해 결정될 것이다. 측정오차에 의한 편의의 크기는 이전의 논의와 마찬가지로 참값의 분산과 오차의 분산의 상대적 크기에 의해 결정될 것이다.

그러나 패널자료를 이용하는 주된 목적 중의 하나는 관찰할 수 없는 개인고정효과를 통제함으로써 변수의 생략에 의한 추정량의 불일치성을 해소하는 데에 있다. 이에 두 번째 경우로 고정효과모형에서 설명변수의

측정오차가 회귀계수의 추정량에 어떤 영향을 미치는가를 검토해 보자. 간결한 논의를 위하여 $t=1, 2$로 2기의 데이터만 이용한 모델을 생각해 보자. 이 경우 β_2에 대한 고정효과 추정량(covariance estimator)은 등식 (11)을 1기와 2기 사이에 차분한 후 이 차분된 등식에 최소자승법을 적용함으로써 도출될 수 있다. 상수항이 소거됨으로써 추정량의 형태와 그 확률극한은 다음과 같이 도출된다.

$$plim\widehat{\beta_2^{fixed}} = plim\frac{\sum_{i=1}^{N}(X_{i2}^{*}-X_{i1}^{*})(Y_{i2}-Y_{i1})}{\sum_{i=1}^{N}(X_{i2}^{*}-X_{i1}^{*})^2} =$$

$$=\beta_2-\beta_2\frac{\sigma_V^2}{(\frac{1-\rho_{1X}}{1-\rho_{1V}})\sigma_X^2+\sigma_V^2} \qquad (13)$$

식 (12)와 (13)을 비교해 보면 개인고정효과를 통제하기 위하여 차분모형을 사용할 경우 생략된 고정효과에 의한 추정량의 불일치성이 제거되었으나, $\rho_{1X}=\rho_{1V}$인 특수한 경우를 제외하고는 측정오차에 의한 추정량의 편의는 다르게 나타나고 있다. 특히 $\rho_{1X}>\rho_{1V}$인 경우 측정오차에 의한 추정량의 불일치성은 식 (12)의 최소자승 추정량보다는 식 (13)의 고정효과 추정량에서 더 크게 나타난다. 만약 그 차이가 상당히 클 경우 고정효과 추정량의 불일치성은 종합적으로 볼 때 오히려 최소자승 추정량의 불일치성보다 더 클 수 있다. 즉 추정량의 불일치성과 관련하여 아무 조치도 취하지 않는 것이 무엇인가를 하는 것보다 더 나은 경우가 발생할 수 있다.

물론 이상의 모든 논의는 각 모형에서 측정오차가 고전적인 가정을 따른다는 것을 전제로 하고 있으며, 그 고전적 가정이 파괴될 경우 논의는 보다 복잡해질 수 있다. 예를 들어 설명해 보자. 교과서적인 논의에 의하면 피설명변수에 존재하는 측정오차는 고전적 가정을 만족하는 한 전반적인 추정의 정확성을 하락시킬지언정 회귀계수의 추정량에 불일치성을 가져다주지는 않는다. 그러나 피설명변수가 어떤 특수한 형태의 측정오

차 패턴을 따를 때는 회귀계수의 추정량이 불일치성을 가지게 된다. 이를 구체적인 예를 들어 설명해 보자. 오랫동안 거시경제학자들은 시계열 자료에 근거하여 실질임금이 경기변동상 경직적이라고 믿어 왔다(Solon et al.(1994)의 문헌연구 부분 참조). 그러나 대략 1980년대 중반부터 개인임금에 대한 패널자료를 이용한 많은 연구들은 실질임금이 매우 경기순행적(procyclical)이라는 사실을 발견하여 왔다. Solon et al.(1994)은 왜 시계열 자료와 패널자료상의 결과들이 다르게 나타나는가를 구성의 효과로 설명하고 있다. 요약하면 거시 시계열 자료에 나타난 시간당 임금률(예를 들어 미국 노동통계국(Bureau of Labor Statistics) 자료)은 경기하강기와 비교하여 경기상승기에 저임금 및 저숙련 근로자들에 대한 가중치를 더 크게 두는 방식으로 계산되어 있어서 도출된 시간당 임금률은 경기역행적으로 편의를 갖게 되며, 개인 패널자료를 이용하여 이러한 노동력 구성의 변화효과를 통제할 경우 실질임금은 매우 경기순행적이라는 사실을 발견하였다. 이 발견은 거시 및 노동경제학계에서는 혁명적인 사건이었다. 동 연구를 전후로 적어도 50편 정도의 논문이 데이터와 디플레이터 혹은 경기변동지수를 바꾸어 가면서 결과의 강건성을 검증한 결과 모두 실질임금은 매우 경기순행적이라는 사실을 재확인하였다. 이러한 발견은 임금경직성에 근거하여 실업변동을 설명하고자 하는, 혹은 임금경직성 자체를 설명하고자 하는 많은 케인지안 이론들의 생명력을 축소시켰다. 여기에서 그 50여 편의 논문들이 공통적으로 채택한 추정모형을 검토해 보자.

$$\triangle \ln w_{it} = \beta_1 + \beta_2 t + \beta_3 X_{it} + \beta_4 \triangle UR_t + \epsilon_{it} \qquad (14)$$

여기에서 w_{it}는 시간당 실질임금률로서 외국 연구의 경우 흔히 연간 근로소득(E_{it})을 연간 근로시간(h_{it})으로 나누어 정의한다. 표기를 간단하게 하기 위해 디플레이터 표시는 하지 않는다.[14] 한편 t는 추세선을, X_{it}는 잠재경력(연령-교육연수-6)을, UR은 경기변동지수로서의 실업률을, ϵ_{it}는 교란항을 나타낸다. β_4는 실질임금의 경기변동에 대한 반응정

14) 또한 논의의 복잡성을 피하기 위해 등식 (14)를 최소자승 추정법으로 추정하였을 때 발생하는 실업률 계수 추정치의 표준오차 과소평가 문제는 생략한다.

도를 나타내는 모수로서 음, 영, 그리고 양으로 나타날 경우 각각 경기순행적, 경직적, 경기역행적임을 시사한다. 앞서 언급한 50여 편의 연구들은 공통적으로 추정된 β_4가 약 -0.015임을 보고하고 있다. 이는 실업률이 1%포인트 증가할 때 실질임금이 약 1.5% 감소하는 것으로 실질임금이 경기변동상 매우 유연하게 조정되어 감을 의미한다. 이제 측정오차 문제를 제기해 보자. 흔히 근로소득 및 근로시간 변수는 심한 측정오차를 안고 있다. 이에 Shin and Solon(2007)은 이를 명시적으로 모형에 고려하여 등식 (14)를 다음과 같이 현실적인 모형으로 바꾸었다. 우선 측정오차가 고전적 가정을 만족한다는 가정을 해보자. 즉 $\triangle \ln E_{it}^{*} = \triangle \ln E_{it} + V_{Eit}$그리고 $\triangle \ln h_{it}^{*} = \triangle \ln h_{it} + V_{hit}$로서 두 V항들은 앞서 소개한 고전적 가정들을 모두 만족한다고 하자. 이를 식 (14)에 대입하면

$$\triangle \ln w_{it}^{*} = \beta_1 + \beta_2 t + \beta_3 X_{it} + \beta_4 \triangle UR_t + (\epsilon_{it} + V_{Eit} - V_{hit}) \quad (15)$$

가 된다. 이는 측정오차가 고전적 가정을 만족하는 한 등식 (15)에 최소자승법을 적용함으로써 β_4계수를 일치적으로 추정할 수 있음을 의미한다.

그러나 Shin and Solon(2007)에 나타난 보다 일반적인 경우를 생각해 보자. 즉 동 연구는

$$\triangle \ln E_{it}^{*} = \lambda_E \triangle \ln E_{it} + V_{Eit} \quad \text{및} \quad \triangle \ln h_{it}^{*} = \lambda_h \triangle \ln h_{it} + V_{hit}$$

형태의 보다 일반적인 측정오차 모형에서 고전적 가정을 만족하는 오차는 $\lambda_E = \lambda_h = 1$라는 특수한 경우임을 보이고 있다. 나아가 이 두 식을 등식 (15)에 대입하고 정리하여 다음의 식을 도출하였다.

$$\begin{aligned} \triangle \ln w_{it}^{*} = {} & \lambda_E(\beta_1 + \beta_2 t + \beta_3 X_{it} + \beta_4 \triangle UR_t) \\ & + (\lambda_E - \lambda_h) \triangle \ln h_{it} + (\lambda_E \epsilon_{it} + V_{Eit} - V_{hit}) \end{aligned} \quad (16)$$

이 추정식은 기존의 수많은 연구들이 사용하여 온 등식 (15)와 비교하여 두 가지 시사점을 주고 있다. 이 두 시사점들 역시 Shin and Solon(2007)에 제시되어 있으나 여기에서는 해당 논의를 보다 일반화시키고자 한다. 첫째, 기존의 연구들은 실질임금의 경기순행성(β_4)을 추정함에 있어서 $\triangle \ln h_{it}$

을 통제하지 않음으로 인하여 $Cov(\triangle UR_t, \triangle \ln h_{it}) < 0$인 경우 추정된 경기순행성은 전형적인 변수의 생략에 의한 편의 문제를 안게 된다. 이 생략된 변수에 의한 추정량의 불일치성을 평가하기 위하여 우선 Bound et al.(1994) 등의 연구를 따라 임금 및 근로시간의 변동에 존재하는 측정오차가 고전적 가정보다는 평균회귀(mean-reversion)의 가정을 따른다고 하자. 즉 $0 < \lambda_E < 1$ 및 $0 < \lambda_h < 1$이 성립한다. 그러나 이 경우에도 불일치성의 방향을 예측하는 것은 간단한 일이 아니다. 교과서적인 논의에 따르면 불일치성의 방향은 $Cov(\triangle UR_t, \triangle \ln h_{it})$의 부호와 $(\lambda_E - \lambda_h)$ 부호의 곱으로 나타난다. 정형화된 사실에 따라 전자의 부호가 음이라고 할 때, 결국 평균회귀의 경향이 근로시간 변동보다 근로소득 변동에서 더 크게 나타날 경우, 기존의 연구들을 따라서 등식 (15)에 최소자승법을 적용하여 실질임금의 경기변동성을 추정할 경우 도출된 추정치는 실제의 경기변동성을 과소평가하는 방향으로 나타날 것이다. 반대로 평균회귀 경향이 근로시간에서 더 크게 나타날 경우 등식 (15)에 근거하여 추정된 실질임금의 경기순행성 정도는 실제의 경기순행성을 과대평가하는 방향으로 나타날 것이다. 등식 (16)이 주는 두 번째 시사점은 Shin and Solon (2007)에 언급되었듯이 근로소득 변동에 존재하는 평균회귀 성격의 측정오차(λ_E)로 인하여 추정된 경기변동성이 항상 경기역행적인 방향으로 편의를 갖는다는 것이다. 이 두 시사점을 종합하면 다음과 같은 결론이 도출된다. 평균회귀 성향이 근로시간 변동보다 근로소득 변동에서 더 크게 나타날 경우 실제의 경기변동성은 등식 (15)에 최소자승법을 적용함으로써 얻어지는 경기변동성보다 클 것이다. 평균회귀 경향이 근로시간 변동에서 더 크게 나타날 경우 두 효과는 반대 방향으로 작용하여 어느 효과가 더 큰가에 따라 결론이 달라질 것이다. 마지막으로 평균회귀 성향이 두 변수에서 비슷하게 나타날 경우 등식 (15)를 이용하여 추정된 실질임금의 경기순행성은 실제의 크기를 역시 과소평가하게 된다. 결국 우리의 관심은 λ_E와 λ_h의 값이 무엇인가에 집중되는데 이들 역시 자료검증 연구를 통하여 얻어질 성격의 정보이다.

제4절 자료검증 연구결과 활용 사례

본 장에서는 자료검증 연구를 수행할 경우 도출되는 검증결과들이 서베이 자료를 이용하여 분석한 결과를 수정 및 보정하는 데 어떻게 활용될 수 있는지에 대해 기존 연구들을 이용하여 예시하겠다. 우선 바로 앞장 마지막 부분에서 소개된 실질임금의 경기변동성 추정에서 활용된 사례를 소개하겠다. 앞서 요약·소개하였듯이 최근 약 50편의 논문들은 추정식 (15)를 최소자승법으로 추정한 결과 추정된 β_4가 대략 -0.015임을 보였다. 그러나 Bound et al.(1994)은 근로소득 및 근로시간에 존재하는 측정오차가 고전적 가정을 따른다기보다는 평균회귀의 성향을 보이고 있다는 것을 발견하였다. 동 연구는 4년의 간격을 두고 2회에 걸쳐 수행된 PSID VS에서 수집된 개인 단위의 자료를 이용하여

$$\triangle \ln E_{it}^{*} = \lambda_E \triangle \ln E_{it} + V_{Eit}$$

식과

$$\triangle \ln h_{it}^{*} = \lambda_h \triangle \ln h_{it} + V_{hit}$$

식을 최소자승법으로 추정하고, 이로부터 λ_E와 λ_h의 추정치를 도출하였다. 표본제약과 추정방법에 따라 다소 차이는 나지만 Shin and Solon(2007)은 두 추정치가 대략 0.8로 비슷하다는 결론에 도달하였다. 이 결과를 이용하면 수많은 기존 연구들이 등식 (15)에 근거하여 추정한 경기변동성의 크기(-0.015)조차도 실제의 크기를 절대값면에서 약 20% 정도 과소평가하고 있다는 결론에 도달하여 실제 경기변동성의 크기는 약 -0.019 정도에 이름을 알 수 있다.

한국의 경우는 어떠한가? 우선 박선영·신동균(2014)은 한국의 경우 임금의 경직성 내지 유연성을 논함에 있어서 보다 합당한 변수는 시간당 임금률이 아니라 월급여임을 역설하고 있다. 동 연구는 1차부터 15차까지의 KLIPS 자료를 이용하여 대부분의 외국 연구들과 마찬가지로 등식

(15)를 남성을 대상으로 추정해 본 결과 추정된 β_4가 약 -0.0254로 나타나 실질임금이 외국의 경우보다 더 경기순행적임을 보였다. 과연 한국에서도 임금에 대한 서베이 조사에서 응답자들의 반응이 임금변동을 희석시키는 방향으로 나타날 것인가에 대해서는 실제 자료검증 연구를 통해 판단되어야 하지만, 앞서 박선영·신동균(2014)의 연구에서 본 것처럼 특정 시점에서 조사된 임금이 심한 라운딩 현상을 보인다면 아마도 두 시점 사이의 임금변동은 미국의 경우와 마찬가지로 평균회귀의 경향을 보일 것이라는 판단이 선다. 이에 PSID VS에 근거한 Bound et al.(1994)의 연구결과를 한국에 그대로 적용하면($\widehat{\lambda_E} \approx 0.8$) 실제 외환위기 이후 한국 노동시장에서 실질임금의 경기순행성은 대략 -0.032에 이른다.

다음에서는 자료검증 연구결과 활용 사례의 다른 예로서 Shin and Shin(2008)을 인용하겠다. 많은 노동경제학 분야에서 부문 간 노동의 규모는 매우 큰 관심을 끌어 왔다. 예를 들어 다른 조건이 같을 경우 부문 간을 이동한 근로자는 그렇지 않은 근로자들과 비교하여 이전(former) 부문에서 습득한 인적자본의 상실로 인하여 이동 후 적어도 단기적으로는 보다 큰 임금손실을 경험하게 된다. 해당 이동이 비자발적인 경우에는 더욱 그러하다. 한편 노동 및 거시경제학 분야에서 주장하는 부문이동가설에 의하면 실업변동의 상당 부분은 부문 간을 이동하는 근로자들이 보다 장기간의 실업을 경험하는 데에서 비롯된다고 한다. 이처럼 부문 간 노동이동의 규모 및 원인은 많은 학자들의 관심을 끌어 왔다. 두 시점 사이에서 부문 간 노동이동의 규모를 추정함에 있어서 많은 기존 연구들은 두 시점에서 서베이를 통하여 추출한 부문(산업 혹은 직종) 정보를 비교하여 다르면 '부문 간(inter-sectoral)', 같으면 '부문 내(intra-sectoral)' 이동으로 정의하였다.[15] 직관적으로 볼 때 이처럼 서베이를 통하여 얻은 부문 정보를 두 시점 사이에 비교함으로써 얻어지는 부문 간 노동이동의 규모는 항상 실제의 노동이동 규모를 과대평가할 것이다. 그것은 특정 조사시점에서 수집된 산업 혹은 직종변수에 존재하는 측정오차가 고전적 가정을 만족할 경우 오차에 의해 부문 내 이동이 부문 간

15) 여기에서 부문 내의 이동이란 고용주는 변하였으나 두 조사시점 사이에 부문이 같은 경우를 말한다.

이동으로 분류될 확률이 그 반대의 경우보다 더 클 것이기 때문이다. 그렇다면 측정오차가 수반된 변수를 이용하여 계산된 부문 간 이동 규모는 실제의 규모를 얼마나 과대평가할 것인가? Shin and Shin(2008)은 우선 부문을 2단위 산업을 기준으로 분류하고 PSID가 1986년부터 1996년까지 매년 전년도의 1월부터 12월까지 개인별로 보고한 월별 노동력 상태정보를 이용하여, 동 표본기간 중 실업을 경험하면서 직장으로 이동한 근로자들 중 47%가 부문 간 이동하였다고 보고하고 있다. 과연 이 47% 중 몇 %가 실제로 이동한 사람이며 나머지는 측정오차에 의한 것인가? 이를 판단하기 위해 역시 Mellow and Sider(1983)의 검증결과를 활용해 보자. PSID VS와는 달리 동 연구는 두 서베이 자료, 즉 CPS를 통하여 획득한 산업정보와 Employment Op- portunity Pilot Project로부터 획득한 정보를 사용자로부터 얻은 산업정보와 비교하여 1단위(대분류)와 3단위(세분류)상에서 두 자료의 일치성을 보고하였다. 이 중 일치성이 더 높은(따라서 측정오차 문제가 덜 심각하게 나타난) CPS 검증자료 결과를 이용해 보자. 동 연구결과에 의하면 1단위에서는 92.3%, 3단위에서는 84.1%의 일치성을 보이고 있다. 또한 동 연구는 직종을 중심으로 부문을 정의하고 유사분석을 수행한 결과 일치성이 산업을 사용한 경우보다 현저하게 나타났다. 한편 Shin and Shin(2008)이 추정한 47%의 부문 간 이동은 2단위의 산업을 기준으로 하였기 때문에 여기에서는 1단위와 3단위의 추정치를 단순 평균하여 88.2%를 2단위에서의 일치율(agrement rate)로 사용한다.

우선 Shin and Shin(2008)은 가장 간단한 경우로 산업이 두 개인 경우를 상정하였다. 우선 산업 간을 이동하는 근로자들 비중의 참값을 q라고 하자. 특정 근로자가 산업 간 이동을 한 것으로 관찰되는 경우는 (i) 실제로 이동했으면서 산업정보가 두 조사시점에서 모두 정확하게 보고되거나, (ii) 실제로 이동했으면서 두 시점에서의 산업정보가 모두 틀렸거나, (iii) 혹은 이동하지는 않았는데 두 시점 중 한 시점에서만 산업정보에 오류가 발생하는 경우로 나누어 볼 수 있다. 이 세 경우를 모두 고려하면 결과적으로 '산업 간' 이동으로 관찰되는 비율은

$$q(0.882^2+0.118^2)+(1-q)2(0.882)(0.118) \quad (17)$$
$$=0.584q+0.208$$

로 표시된다. 따라서 $0.584q+0.208=0.47$를 풀면 실제로 부문 간을 이동한 비율은 0.449로 추정되며, 이는 0.47의 95.5%로서 편의의 규모는 4.5% 정도에 불과하다.

물론 두 개의 산업을 가정한 것은 현실을 지나치게 단순화시킨 것이며, 편의의 정도는 산업 수에 따라 증가할 것이다. 그러나 일반적으로 N개의 산업으로 확대할 경우 등식 (17)의 유도는 상당히 복잡해진다. 대신 Shin and Shin(2008)은 또 다른 극단적인 경우로 산업을 연속체(continuum)로 보았다. 이 경우 실수로 잘못 보고된 산업들이 두 조사시점에서 같게 나타날 확률은 영(0)이 되며, 실제 산업 간을 이동한 근로자들은 이동 전이나 이동 후의 산업을 잘못 보고한 경우에도 여전히 산업 간을 이동한 것으로 기록될 것이다. 반면 산업 내 이동자의 경우 두 조사연도 중 적어도 한 연도에서 산업을 잘못 보고할 경우(그럴 확률은 $1-0.882^2$) 산업 간 이동자로 기록될 것이다. 결국 서베이를 통하여 관찰되는 산업 간 이동자의 비중은

$$q+(1-q)(1-0.882^2)=0.778q+0.222 \quad (18)$$

로 나타날 것이다. 이를 0.47과 같다고 놓고 q에 대해 정리하면 실제 산업 간 이동자 규모는 31.9%로 추정되며 이는 47%의 68%에 해당한다. 결국 식 (17)과 식 (18)의 두 극단적인 경우들로 도출되는 편의의 크기는 약 4.5~32%로 실제 2단위 산업에서 편의의 크기는 이 사이에 있을 것으로 추정된다. 앞서 언급하였듯이 이 예시에서는 측정오차에 의한 산업간 이동규모의 과대추정 문제가 가장 덜 심각할 경우를 상정하고 있으며 이에 따라 실제 편의의 규모는 훨씬 더 클 것으로 판단된다. 한국의 경우 산업 및 직종별로 각각 1, 2, 3단위에서 서베이를 통하여 수집된 정보와 고용주 기록을 바탕으로 수집된 정보 사이의 일치율이 얼마나 될 것인가에 대해서는 자료검증 연구를 통하여 발견되어야 할 것이다.

제5절 소 결

정보 수집의 용이성, 그리고 수집된 정보의 풍부함 등의 이유로 많은 사회과학 분야의 연구들은 서베이를 통하여 획득된 자료에 근거하여 수행되고 있다. 본고의 목적은 그러한 자료들이 얼마나 심각한 측정오차 문제를 안고 있는가, 그리고 이러한 서베이 자료에 존재하는 측정오차가 분석결과를 얼마나 왜곡시킬 수 있는가를 예시하고 이를 통하여 보다 본격적인 자료검증 연구의 필요성을 역설하는 데에 있다. 서베이 자료에 대한 직접적인 검증자료(validation data)에 대한 접근이 용이하지 않은 상태에서 본고에서는 KLIPS 자료에 나타난 변수들의 시계열상 일관성을 기준으로 몇 가지 변수들에 대해 측정오차의 심각성을 가늠해 보았으며, 나아가 측정오차 문제가 가장 심각할 것으로 판단되는 임금변수에 대해서는 임금대장에 기초한 기존의 연구 결과와의 비교를 통하여 그 심각성을 평가하여 보았다. 그 결과 서베이 자료에 나타난 측정오차의 문제는 외국의 경우처럼 심각한 수준인 것으로 나타났다. 더구나 KLIPS 자료는 패널자료이기 때문에 자체적으로 시계열상이나 다른 변수들과의 일치성을 기준으로 어느 정도 편집이 가능하다는 점을 고려해 볼 때 일반적으로 서베이 자료에 나타난 변수들의 측정오차는 본고에서 예시한 것보다 더 심각할 수 있음을 짐작할 수 있다. 아울러 본고에서는 서베이 자료검증 결과가 서베이 자료에 근거하여 도출된 왜곡된 분석결과를 보정하는 데에 어떻게 사용될 수 있는가에 대해 예시하고 있다. 지면 관계상 검증결과의 유용성에 대해 보다 다양한 소개를 할 수 없었음을 아쉽게 생각하며, 조속한 시일 내에 서베이 자료검증 연구가 실시되기를 기대한다.

일반적으로 서베이 자료검증을 위한 검증자료(validation data)로는 오차로부터 비교적 자유로운 행정자료(administrative data)를 사용할 수 있으며, PSID VS의 경우처럼 서베이 응답자가 소속된 직장으로부터 직접 획득한 정보를 이용할 수도 있다. 행정자료의 사용은 검증비용이 적

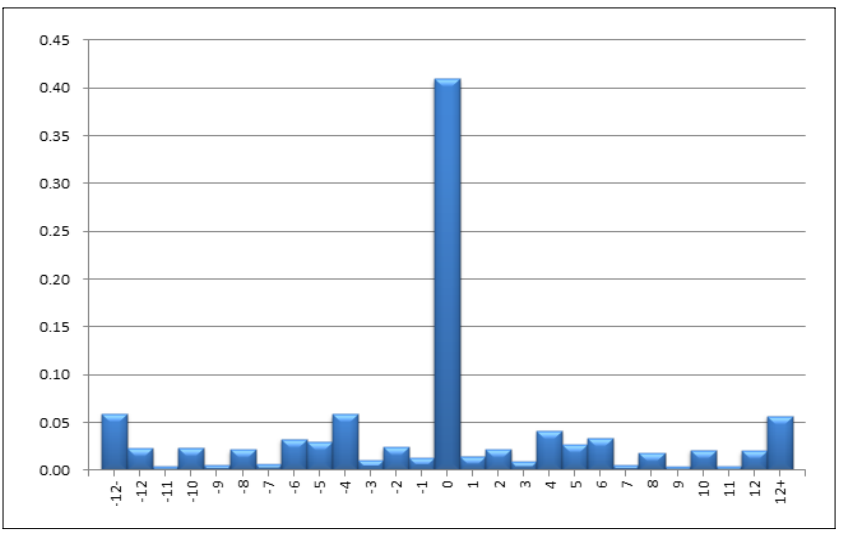
다는 장점이 있지만 (i) 접근이 용이하지 않고 (ii) 이용 가능한 변수가 다양하지 않으며 (iii) 이용 가능한 변수들조차도 많은 경우 서베이를 통하여 획득한 변수들과 정의상 불일치성을 보이고 있다. 세 번째 경우의 예를 들어보자. KLIPS 자료상에 나타난 급여정보가 조사시점에서 응답자가 가지고 있던 주된 일자리에서의 월평균 급여를 나타내는 것과 달리 국세청의 소득정보는 연간 종근무지를 통하여 획득된 총소득을 의미한다. 한편 고용노동부의 다양한 사업체 자료상에 나타난 월급여 정보는 (i) KLIPS 자료와는 달리 세전 월급여를 나타내며 (ii) 역시 KLIPS 자료와는 달리 월평균이 아니라 특정 월에 사용자가 실제로 지불한 액수를 나타낸다. 이와 비교하여 PSID VS와 같이 응답자가 직접 보고한 변수들의 값을 해당 응답자가 소속된 직장을 접촉하여 획득한 변수들의 값과 대조하는 작업은 비록 비용 발생의 문제는 있으나 행정자료와 비교하여 보다 다양한 변수들에 대한 정보를 획득할 수 있으며, 동시에 변수의 정의라든가 정보 발생시점을 일치시킬 수 있는 장점이 있다. 이에 본고에서는 임금 및 각종 소득, 부가급여, 정규 및 초과 근로시간, 근속기간, 고용형태, 연령, 성, 학력, 산업, 직종, 사업체 규모, 노조 등 기본적인 노동시장 변수들에 대해 PSID VS 성격의 자료검증 연구를 수행할 것을 최소한 2회 실시할 것을 제안한다. 첫 연도 검증 연구결과는 서베이 변수에 나타난 측정오차의 특정 시점에서의 성격에 대한 이해를 돕는다. 나아가 동일 응답자에 대해 시차를 두고 2회 이상 검증연구를 수행함으로써 서베이를 통하여 획득된 변수들에 존재하는 측정오차가(예를 들어) 임금 증가율을 어떤 방향으로 얼마나 왜곡시키는가를 이해할 수 있다.

〈부표 1〉 성별 변수의 측정오차 사례

조사회차 \ 개인id	95604	212404	276206	349405	375804	417003	504704
1	-	-	-	1	2	1	-
2	-	2	-	1	2	-	-
3	-	2	-	1	2	-	-
4	-	2	-	1	-	1	-
5	-	2	-	1	-	1	-
6	-	2	-	1	2	1	-
7	-	2	-	1	2	1	-
8	-	2	-	1	2	1	-
9	-	-	-	1	-	1	-
10	2	2	1	1	2	1	1
11	1	2	2	1	1	2	-
12	2	1	2	2	2	1	2
13	-	1	2	2	2	1	-
14	-	-	-	2	2	1	1
15	-	-	2	2	2	1	-

주 : 1=남성, 2=여성.

〈부록 2〉 정규 근로시간 변동의 경험적 분포



자료 : 박선영 · 신동균(2014)에서 차용.

제 6 장
결론 : 요약 및 시사점

본 연구는 패널품질 개선연구 시리즈의 네 번째 연구로서, 패널자료에서 표본이탈을 보정하는 방법, paradata를 이용하여 무응답 편향을 줄이는 방법, 사업체 패널조사의 표본 설계 시 유의해야 할 사항, 패널자료의 자료검증 연구가 필요한 이유에 대해 논의하였다. 본 연구의 결과를 요약하고 시사점을 정리하면 다음과 같다.

제1절 동태적 패널모형에서 표본이탈 교정

이 연구에서는 표본이탈이 존재할 때, 종속변수의 과거값이 설명변수로 사용되는 동태적 패널모형의 효율적인 추정방법에 대해 논의하였다. 표본이탈이 존재하지 않는 일반적인 상황에서는 고정효과를 1계차분(FD)하는 'Arellano-Bond(1991)의 full GMM' 또는 'FD/IV'를 이용해 모수를 추정한다. 한편 동태적 패널모형의 고정효과를 FOD(Forward Orthogonal Deviations, Arellano and Bover, 1995)로 제거하게 되면 오차항이 동분산을 갖고 시간에 걸쳐 비상관되어 차분한 경우보다 더 효율적일 것으로 예상된다. 그러나 모의실험 결과에 따르면 FOD/IV의 분산이 FD/IV의 분산보다 더 크게 도출되었으며, 이 결과를 토대로 편향이 작으면서 분산이 작은 'FD-FOD/GMM'을 제3절에서 제시하였다.

비균형패널을 이용할 때 FOD 변환을 사용한 추정법은 FOD가 미래의 표본이탈 여부에 의존하여 비일관적으로 추정된다. 따라서 표본이탈이 존재할 때 FOD 변환을 사용한 추정법은 사용할 수 없다. 표본이탈이 내생적으로 발생하면 이로 인한 편향을 교정하여 사용해야 한다. Wooldridge (2002; 2010)는 동태적 패널모형에 Heckman(1976)의 2단계 추정법을 적용해 편향교정항(Inverse Mills Ratio : IMR)으로 편향을 교정한다. 본 연구에서는 'FD/IV with IMR'과 'full GMM with IMR'을 제시하고 있는데, 전자는 FD/IV의 도구변수를 편향교정항의 독립변수로 사용하고, 후자는 full GMM의 도구변수를 편향교정항의 독립변수로 사용한다. 표본이탈이 없을 때와 마찬가지로 내생적 표본이탈이 존재할 때 'full GMM with IMR'의 분산이 'FD/IV with IMR'의 분산보다 더 작다. 이러한 내용을 토대로 사업체 패널의 일부에 적용한 결과를 제5절에서 제시하였다.

제2절 Paradata를 이용한 무응답 자료 회귀분석

사업체 패널조사에서는 면접 당시의 '첫 컨택 반응'이라는 변수가 일종의 paradata로서 얻어진다. 이 '첫 컨택 반응'변수는 응답 성향에 대한 좋은 정보를 제공한다. 본 연구에서는 이러한 paradata를 사용하여 무응답 편향을 줄이는 회귀분석 방법에 대하여 다루었다.

제안된 방법은 (1) '첫 컨택 반응'변수(Z)를 분석하고자 하는 회귀모형의 설명변수에 추가하여 확장된 회귀모형(augmented outcome regression model)을 사용하여 응답자료만을 사용하여 회귀계수를 추정하고, (2) 추가적으로 '첫 컨택 반응'변수를 기존 설명변수(X)로 설명하는 회귀모형을 별도로 세워 전체 자료를 바탕으로 모형을 적합시켜 Z에 대한 예측값을 X의 함수로 표현한 후 (3) 이를 확장된 회귀모형의 Z에 대입해 줌으로써 최종적으로는 Y에 대한 X의 회귀모형을 얻어내는 것이다.

본 연구에서는 제안된 방법론을 모의실험을 통해 기존의 방법론과 비교하였고 무응답 메커니즘이 X뿐만 아니라 Z에 의존하는 경우에도 효율

적인 추정량을 제공하였다. 또한 기존의 방법론인 확장된 무응답 성향 모형을 사용한 가중치 조정 방법론보다 더 효율적인 추정을 구현하는 것을 확인하였다. 또한 제안된 방법론을 사업체 패널조사에 적용한 결과 제안한 방법은 다른 방법들에 비하여 추정치 분산의 추정량이 다소 작게 나왔다.

무응답 자료를 이용한 회귀분석에는 무응답 메커니즘에 대한 이해가 필요하다. 일반적으로 무응답 성향과 관련 있는 변수를 아는 경우에는 그 관련 변수를 회귀분석의 설명변수에 포함시켜 분석하는 것이 무응답에 따른 선택 편향을 줄여주는 것으로 알려져 있다. paradata가 무응답 성향에 대한 좋은 정보를 제공하므로, 이를 활용하면 효율적인 추정이 가능함을 보여주었다.

제3절 사업체 패널조사의 표본 설계 관련 연구

이 연구에서는 기존의 사업체 패널조사에 나타난 몇 가지 문제점을 살펴보고, 새로운 패널 표본 설계 또는 기존 패널을 보완하는 표본 재설계를 할 때 유의해야 할 사항들을 정리하였다.

패널 설계에서는 (1) 조사 모집단의 결정과 (2) 층화에 대해 유의하여야 한다. 사업체 패널조사는 많은 조사항목을 가지는 다목적 조사이며 이 조사를 통해 산업별, 사업장 규모별, 지역별 통계 생산이 가능하도록 하는 것을 기본 원칙으로 설계되었다. 새로운 표본 설계 방안도 원칙면에서는 이와 동일하며 이를 위하여 산업별 분류, 사업장 규모 및 지역을 층화변수로 하는 층화 추출을 사용하는 것도 동일하다. 층화 추출에서 표본 배정은 여러 가지 제한조건들을 만족하면서 전체 추정량의 분산을 최소화하는 최적화 문제의 해로서 얻어지는데 이를 위해서 mathematical programming이 사용된다.

사업체 패널조사 설계의 또 다른 방안으로는 이전의 조사에서 응답해 오던 기존 표본은 그대로 두고 나머지 모집단에서 표본을 추가하는 추가

표본 설계를 고려할 수 있다. 패널조사에서 표본 탈락 또는 패널 마모(panel attrition)는 흔히 발생하는 현상으로 적절한 시점에서 계속 표본 추가를 해주어서 적정 표본 수를 유지하고 모집단에 새롭게 진입한 신규 사업체들을 포함함으로써 전체적인 횡단면적 대표성을 제고하는 효과를 가지게 된다. 추가 표본 설계에서는 층화 추출을 기본으로 하되 기존의 표본을 포함한 상태에서 새로운 제한조건을 가지는 최적화 문제로 풀어서 표본 배정을 실시하고 그로부터 신규 표본을 추출하면 된다.

제4절 한국에서의 자료검증 연구의 필요성에 대하여

정보 수집의 용이성, 그리고 수집된 정보의 풍부함 등의 이유로 많은 사회과학 분야의 연구들은 서베이를 통하여 획득된 자료에 근거하여 수행되고 있다. 이 연구에서는 그러한 자료들이 얼마나 심각한 측정오차 문제를 안고 있는가, 그리고 이러한 서베이 자료에 존재하는 측정오차가 분석결과를 얼마나 왜곡시킬 수 있는가를 예시하고 이를 통하여 보다 본격적인 서베이 자료검증 연구의 필요성을 보여주었다.

본 연구에서는 노동패널 자료변수들의 시계열상 일관성을 기준으로 몇 가지 변수들에 대해 측정오차의 심각성을 가늠해 보았으며, 나아가 측정오차 문제가 가장 심각할 것으로 판단되는 임금변수에 대해서는 노동 패널자료로 계산한 결과와 임금대장에 기초한 기존의 연구 결과를 비교하여 보았다.

최종적으로 본 연구에서는 임금, 소득, 부가급여, 근로시간, 근속기간, 고용형태, 연령, 성, 학력, 산업, 직종, 사업체 규모, 노조 등 기본적인 노동시장 변수들에 대해 미국의 PSID 자료검증 연구와 유사한 성격의 자료검증 연구를 수행할 것을 제안하였다. 이처럼 서베이를 통하여 획득된 주요 변수들에 내재해 있는 측정오차의 특성과 정도에 대한 정보를 획득할 수 있으며, 이 정보는 향후 서베이 자료에 근거하여 수행될 모든 연구들의 분석결과들을 보정하는 데에 도움을 줄 것으로 판단된다. 아울러

서베이 자료에 존재하는 측정오차의 문제가 특정 시점에서의 변수값만 아니라 해당 변수의 두 시점 사이의 변화에도 영향을 미칠 수 있기 때문에 자료검증 조사 및 연구를 시점을 달리하여 최소한 두 차례에 걸쳐 수행할 것을 제안하였다.

참고문헌

김기민(2013),『Paradata를 활용한 패널 무응답 보정을 위한 가중치 부여 방안』, 한국노동연구원.

박선영・신동균(2014),「한국의 명목 및 실질임금의 유연성 정도와 성격에 대하여」,『노동경제논집』37(2), pp.1~47.

홍민기・김재광・한치록・김기민(2014),『패널자료 품질개선 연구 III』, 한국노동연구원,

Altonji, J. and R. Shakotko(1987), "Do Wages Rise with Job Seniority?" *Review of Economic Studies* 54, pp.437~59.

Altonji, J. G. and P. J. Devereux(1999), "The Extent and Consequences of Downward Wage Rigidity," Working Paper No.7236, National Bureau of Economic Research.

Anderson, T. W. and C. Hsiao(1981), "Estimation of dynamic models with error components," *Journal of the American Statistical Association* 76, pp.598~606.

Arellano, M. and O. Bover(1995), "Another Look at the Instrumental Variable Estimation of Error-components Models," *Journal of Econometrics* 68, pp.29~51.

Arellano, M. and S. Bond(1991), "Some Tests of Specification for Panel Data : Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies* 58, pp.277~297.

Beaudry, P. and J. DiNardo(1991), "The Effect of Implicit Contracts on the Movement of Wages over the Business Cycle : Evidence from Micro Data," *Journal of Political Economy* 99, pp.665-88.

Bound, J. and A. B. Krueger(1991), "The Extent of Measurement Error

in Longitudinal Earnings Data : Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9, (January 1991), pp.1~24.

Bound, J., C. Brown and N. Mathiowetz(2001), "Measurement error in survey data," *Handbook of Econometrics* Vol.5. J. Heckman and E. Leamer(eds), Elsevier, pp.3705~3843.

Bound, J., C. Brown, G. J. Duncan and W. L. Rodgers(1994), "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data," *Journal of Labor Economics* 12, (July 1994), pp.345~368.

Brown, C. and A. Light(1992), "Interpreting Panel Data on Job Tenure," *Journal of Labor Economics* 10, pp.219~257.

Card, D. and D. Hyslop(1996), "Does Inflation Grease the Wheels of the Labor Market?" Working Paper No. 5538, National Bureau of Economic Research.

Couper, M. P. and L. Lyberg(2005), *The Use of Paradata in Survey Research*, Proceedings of the International Statistical Institute Meetings.

Dickens, W. T., L. Goette, E. L. Groshen, S. Holden, J. Messina, M. E. Schweitzer, J. Trunen and M. E. Ward(2006), "How Wages Change : Micro Evidence from the International Wage Flexibility Project," Federal Reserve Bank of Cleveland Working Paper No.16~20.

Durrant, G. B., J. D'Arrigo and F. Steele(2011), "Using Paradata to Predict Best Times of Contact, Conditioning on Household and Interviewer Influences," *Journal of the Royal Statistical Society : Series A* 174, pp.1029~1049.

Elsby, M. W.L., D. Shin and G. Solon(2013), "Wage Adjustment in the Great Recession," Working Paper No. 19478, National Bureau of Economic Research.

Han, C. and H. Kim(2014), "The Role of Constant Instruments in Dynamic Panel Estimation," *Economics Letters* 124, pp.500~503.

Han, C. and C. Peter, B. Phillips(2006), “GMM with Many Moment Conditions,” *Econometrica* 74(1), pp.147~192.

Heckman, J. J.(1976), “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models,” *Annals of Economic and Social Measurement* 5, pp.475~492.

Kahn, S.(1997), “Evidence of Nominal Wage Stickiness from Microdata,” *American Economic Review* 87(5) pp.993~1008.

Kim, J. K. and Park, M.(2010), “Calibration estimation in survey sampling,” *International Statistical Review* 78, pp.21~39.

Kim, J. K. and J. Shao(2013), “Statistical Methods for Handling Incomeplete Data,” Chapmann & Hall/CRC.

Kim, J. K. and J. Im(2014), “Propensity Score Weighting Adjustment with Several Follow-ups,” *Biometrika* 101, pp.439~448.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T. M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R. M. Groves and T. E. Raghunathan(2010), “Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-response : Examples from Multiple Surveys,” *Journal of the Royal Statistical Society : Series A* 17, pp.389~407.

McLaughlin, K.(1994), “Rigid Wages?” *Journal of Monetary Economics* 34(3), pp.383~414.

Mellow, W. and H. Sider(1983), “Accuracy of Response in Labor Market Surveys : Evidence and Implications,” *Journal of Labor Economics* 1, pp.331~344.

Moffitt, R., J. Fitzgerald and P. Gottschalk(1999), “Sample Attrition in Panel Data : the Role of Selection on Observables,” *Annals of Economics and Statistics* 55/56, pp.129~152.

Nickell, S. and G. Quintini(2003), “Nominal Wage Rigidity and the Rate of Inflation.” *Economic Journal* 113(490), pp.762~781.

Park, S., D. Shin and J. Yoo(2014), “Are Nominal Wages Rigid in Korea,” unpublished manuscript.

Shin, D. and G. Solon(2007), “New Evidence on Real Wage Cyclicality Within Employer-Employee Matches,” *Scottish Journal of Political Economy* 54(5), (November 2007), pp.646~660.

Shin, D. and K. Shin(2008), “Fluctuations of Unemployment and Inter- and Intra-Sectoral Reallocations of Workers,” *International Economic Journal* 22(2), (June 2008), pp.231~251.

Smith, J. C.(2000), “Nominal Wage Rigidity in the United Kingdom.” *Economic Journal* 110(462), pp.C176~195.

Solon, G., R. Barsky and J. A. Parker(1994), “Measuring the Cyclicality of Real Wages : How Important Is Composition Bias?” *Quarterly Journal of Economics* 109(1), pp.1~26.

Topel, R.(1991), “Specific Capital, Mobility, and Wages : Wages Rise with Job Seniority,” *Journal of Political Economy* 99, pp.145~76.

Valliant, R., Dever, J. A. and F. Kreuter(2013), *Practical Tools for Designing and Weighting Survey Samples*, Springer.

Wagner, J., B. T. West, N. Kirgis, J. M. Lepkowski, W. G. Axinn and S. K. Ndiyaye(2012), “Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection,” *Journal of Official Statistics* 28, pp.477~499.

Wood, A. M., I. R. White and M. Hotopf(2006), “Using Number of Failed Contact Attempts to Adjust for Nonignorable Non-response,” *Journal of the Royal Statistical Society : Series A* 169, pp.525~542.

Wooldridge, J. M.(2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press.

_____(2010), *Econometric Analysis of Cross Section and Panel Data,* 2nd Edition, The MIT Press.

◈ 執筆陣

• 홍민기(한국노동연구원 연구위원)
• 한치록(고려대학교 교수)
• 김재광(Iowa State University 교수)
• 신동균(경희대학교 교수)
• 김기민(한국노동연구원 책임연구원)
• 이고은(고려대학교 박사과정)

패널자료 품질개선 연구(Ⅳ)

▪ 발행연월일 | 2014년 12월 24일 인쇄
2014년 12월 30일 발행
▪ 발 행 인 | 이 인 재
▪ 발 행 처 | **한국노동연구원**
150-740 서울특별시 영등포구 은행로 30
☎ 대표 (02) 3775-5514 Fax (02) 3775-0697
▪ 조판·인쇄 | 거목정보산업(주) (02) 2164-3232
▪ 등 록 일 자 | 1988년 9월 13일
▪ 등 록 번 호 | 제13-155호

정가 6,000원

ISBN 978-89-7356-541-2