經濟學硏究 제73집 제3호 87-135
DOI : 10.22841/kjes.2025.73.3.003

지역별고용조사 임금자료의 측정오차*

이 지 영**

논문초록 본 연구는 국세청의 지역, 성별, 분위별 근로소득 자료를 활용하여 지역별고용조사 임금자료의 측정오차를 탐구하였다. 본고의 주요 발견은 다음과 같다. 첫째, 지역별고용조사의 임금자료에는 평균회귀 측정오차(mean-reverting measurement error)가 존재한다. 둘째, 평균회귀 측정오차가 있는 자료를 종속변수로 활용하면 계수 값이 희석된다. 셋째, 측정오차의 자기상관이 강하며, 일차차분 자료를 활용하면 측정오차의 영향을 줄일 수 있다. 이러한 발견은 실증분석 시 고전적 측정오차 또는 측정오차가 없다는 가정을 점검할 필요가 있음을 시사한다.

핵심 주제어: 지역별고용조사, 국세청 근로소득자료, 평균회귀측정오차, 희석편의
경제학문헌목록 주제분류: J0, J3

투고 일자: 2024. 9. 4 심사 및 수정 일자: 2025. 5. 6. 게재 확정 일자: 2025. 8. 15.

* 본 논문은 저자가 서울대학교 경제학부 소속일 때 작성한 원고를 수정 및 보완한 연구이다. 또한 본 논문은 한국은행의 공식견해가 아닌 집필자 개인의 견해이므로 본 논문의 내용을 인용하는 경우에는 집필자 명을 반드시 명시해 주시기 바란다. 아울러 본 논문 작성에 많은 도움을 주신 이정민 교수님 그리고 논문을 꼼꼼히 읽고 지적해 주신 익명의 심사위원들과 편집위원께 깊은 사의를 표한다. 본고에 남아 있을 수 있는 오류는 저자의 책임임을 밝힌다.

** 한국은행 강원본부 과장, e-mail: jyhhss@gmail.com

Ⅰ. 서 론

임금 자료는 노동경제학자를 포함한 여러 연구자와 정책입안자 등에게 유용한 정보를 제공한다. 설문을 통해 조사한 임금 자료도 많이 활용되고 있는데, 대부분 측정오차를 포함한다. 예컨대, 본고에서 활용하는 지역별고용조사의 세전 월평균 소득자료 분포는 10만 원, 50만 원, 심지어 100만 원 단위에도 몰려있다(〈Figure A. 1〉).[1] 설문의 질문은 세금 공제 전 상여금을 포함한 최근 3개월 동안의 월평균 임금에 관한 것으로, 이러한 분포는 응답자가 기억력이 좋지 않거나 급여 구성요소에 해당하는 금액을 잘 알지 못해 계산을 잘못한 경우 그려질 수 있다. 예컨대, 원천징수제도로 통장에는 세후 임금만 표기되어, 근로자가 급여 명세서를 꼼꼼히 확인하지 않는 이상 세금, 상여금, 시간외 수당 등 각각에 해당하는 금액을 정확히 알지 못할 수 있다. 또는 가족 구성원의 월급을 대신 응답하는 경우 자신의 월급을 응답할 때보다 오차가 더 클 수 있다(Lee and Lee, 2012; Bound and Krueger, 1991). 이러한 모습은 알려진 사실이지만, 많은 연구자는 측정오차(error-in-variables; measurement errors)가 추정에 미치는 영향을 고려하지 않고 결과를 해석하는 경향이 있다. 연구자들은 실제로 문제가 없다고 믿기보다는 단순히 편의를 위해 고려하지 않았을 것이다.

계량경제학 교과서에서는 참값과 측정오차의 관계가 독립적인 경우 고전적(classical) 측정오차의 특징을 가지며, 이러한 측정오차가 독립변수에 있는 경우 추정치가 0에 가까워지도록 하는 희석편의(attenuation bias)를 발생시킨다고 설명한다(Hansen, 2020, p. 348). 측정오차를 응답값에서 참값을 뺀 값으로 정의할 때, 측정오차가 참값과 음의 관계에 있다면 측정오차는 평균으로 회귀하는(mean-reverting) 특징을 갖는다. 예컨대, 설문조사에서 소득을 보고할 때 고소득자는 실제 소득보다 작게 저소득자는 실제 소득보다 크게 응답하는 경향이 있다. 미국 설문조사 자료와 행정자료를 개인단위에서 연결하고 설문조사 자료의 측정오

1) 〈Figure A. 1〉은 세전 월 평균소득이 300백만 원 이하인 사람의 임금분포로, 150만 원, 200만 원, 250만 원, 300만 원에 자료가 몰려있다. 즉 155만 원을 150만 원 또는 100만 원으로 응답하는 것이다. 고용형태별 근로실태조사의 임금 자료를 통해 고용주가 지급하는 임금의 분포는 〈Figure A. 1〉과 같이 특정 값에 몰려있지 않다. 이러한 형태의 응답오차를 히핑오차(heaping error)라고 한다.

차를 연구한 Duncan and Hill(1985)과 Bound et al. (1994)은 측정오차가 연간 소득 프리미엄을 희석한다고 보고하였고, Lee and Lee(2012)는 측정오차가 시간에 따른 임금격차 변화를 왜곡시킨다고 보고하였다. 국내 연구에서는 응답값으로 소득분배지표를 계산하면 참값으로 가정한 행정자료로 계산한 것에 비해 개선된 것으로 보고하였다(김낙년 · 김종일, 2013; 홍민기, 2017; 이원진 외, 2019). 따라서 설문조사 자료의 측정오차 특징에 따라 발생할 수 있는 문제에 대해 인식하고 해석 시 주의할 필요가 있다.

본 연구는 2015~2018년 국세청 근로소득 자료를 참값으로 가정하고 지역별고용조사 임금자료의 측정오차의 특징과 그 오차가 연구에 미칠 수 있는 영향에 대해 고찰한다. 지역별고용조사는 지역별 임금을 파악할 수 있는 조사로 다른 조사에 비해 국내 모든 지역 정보를 고르게 포함한다는 장점이 있어 많은 연구에서 활용되고 있지만, 자료가 포함하고 있는 측정오차를 고찰한 연구는 확인할 수 없었다. 본고의 주요 발견은 다음과 같다. 첫째, 지역별고용조사의 임금자료에는 평균회귀 측정오차(mean-reverting measurement errors; 비고전적 측정오차)가 존재한다. 둘째, 평균회귀 측정오차가 있는 소득자료를 종속변수로 활용하면 계수 값이 희석된다. 셋째, 측정오차의 자기상관이 강하며, 이 경우 일차 차분 자료를 활용하면 측정오차의 영향을 줄일 수 있다. 나아가 기존 연구에 제시된 측정오차의 종류에 따른 편의의 크기와 모델의 적합성 및 t값의 변화를 정리하였다. 이러한 지역별고용조사의 측정오차에 대한 발견은 실증분석 결과 해석시 고전적 측정오차 또는 측정오차가 없다는 가정을 점검할 필요가 있음을 시사한다.

본고의 구성은 다음과 같다. 제Ⅱ절에서는 기존 연구를 소개하고 제Ⅲ절은 사용한 자료를 서술한다. 제Ⅳ절에서는 측정오차의 이론적 배경을 소개하고 제Ⅴ절에서 지역별고용조사의 측정오차 특징과 그로 인한 영향을 제시한다. 제Ⅵ절에서 결론을 맺는다.

Ⅱ. 선행연구

측정오차를 연구하는 자료검증연구(validation studies)에서는 설문조사 자료를 응답값으로 행정자료를 참값으로 보고 두 자료의 차이를 측정오차로 정의한다. 미국에서는 개인단위에서 응답값과 참값을 연결할 수 있는 자료가 존재하여 자료검증연

구가 활발하다. 공통적으로 설문조사 자료에 포함된 측정오차가 평균으로 회귀하며 자기상관이 존재한다고 보고한다(Duncan and Hill, 1985; Bound and Krueger, 1991; Bound et al., 1994; Bollinger, 1998; Hu and Wansbeek, 2017). 예컨대, 임금자료를 활용한 연구로 Duncan and Hill(1985)은 근로자 개인이 Panel Survey Income Dynamics(PSID)에 보고한 임금과 PSID 응답자가 근무하는 회사의 임금자료를 연결하고 측정오차를 연구하였다. 또한 Bound and Krueger(1991)와 Bound et al. (1994)은 Current Population Survey(CPS)와 국세청 자료를 개인단위에서 결합하여 측정오차의 특징을 연구하였다. 이 연구들은 PSID와 CPS의 측정오차는 평균으로 회귀하는(mean-reverting) 특징이 있으며, 측정오차 간에 자기상관(serial-correlation)이 존재한다고 보고하였다. Pischke(1995)는 측정오차의 이러한 특징이 임시 소득에 의해 결정된다고 하였다.

비모수적인 방법을 통해서도 평균으로 회귀하는 측정오차를 발견한 문헌이 존재한다(An and Hu, 2012; Hu and Sasaki, 2015; Bollinger, 1998). 구체적으로 Bollinger(1998)은 CPS와 Social Security Administration(SSA)의 임금자료를 연결하고 커널 회귀방정식을 추정하였으며 그 과정에서 임금의 참값과 측정오차 사이에 음의 상관관계가 있음을 발견하였다. 다시 말해 저소득층 근로자는 CPS에서 실제 임금보다 큰 값을 보고하고 고소득층 근로자는 실제 임금보다 작은 값을 보고하였다. 임금 이외에 범죄행위 여부(Gibson and Kim, 2008), 노동조합 가입 여부(Bollinger, 2001), 정부 보조금 프로그램 참여 여부(Bollinger and David, 2005; Meyer and Mittag, 2019; Meyer, Mittag, and Goerge, 2022), 학력(Bollinger, 2003) 등과 같이 이분법적인(binary) 응답에 포함된 오차에 관한 연구가 존재하며 이는 정책의 효과 및 임금 격차 정도를 왜곡하는 것으로 보고한다.[2)]

최근 자료검증연구는 설문조사와 행정자료에 모두 오차가 존재한다고 가정하고 행정자료의 오차 특징에 따른 측정오차의 변화를 고찰하고 그러한 오차가 분석 결과에 미친 영향을 연구한다(Kapteyn and Ypma, 2007; Jenkins and Rios-Avila, 2020, 2023; Hyslop and Townsend, 2020; Abowd and Stinson, 2013).[3)]

---

2) Meyer and Mittag(2019)은 정부로부터 빈곤퇴치정책의 일환으로 지원금을 받은 사람이 그렇지 않은 것으로 응답하여 정책의 효과가 희석되었다는 것을 발견하였으며, Feng and Hu(2013)는 개인이 고용된 유형을 잘못 응답하여 실업률이 잘못 계산되었다고 보고하였다. Bollinger(2003)는 학력에 포함된 응답오차가 인종 간의 임금 격차를 왜곡했다고 보고하였다.

국내에서는 응답값과 참값을 개인단위에서 연결할 수 있는 자료가 없어 자료검증 연구는 초기 단계이다. 일부 연구는 국세통계연보와 가구소비실태조사 자료를 활용하여 측정오차가 지니계수와 같은 소득분배지표에 미치는 영향을 연구하였다(김낙년·김종일, 2013; 홍민기, 2017; 이원진 외, 2019). 김낙년·김종일(2013)은 가구소비실태조사의 고소득 근로자의 연간 소득 정보를 국세연보의 자료로 보충하고 지니계수를 계산하였다. 홍민기(2017)는 가구소비실태조사의 고소득 근로자의 가중치를 보정하고 2000년대 후반의 지니계수를 계산하였다. 이들 연구는 측정오차를 보정 한 소득자료로 계산한 소득분배지표가 보정 하지 않은 자료로 계산한 지표에 비해 악화되었다고 보고하였다.

이원진 외(2019)는 2017년 가계금융복지조사의 소득자료와 응답자들이 국세청에 제출한 행정보완자료를 결합한뒤, 동일한 개인의 응답자료와 행정자료로 소득불평등도를 각각 계산하여 비교하였다. 가계금융복지조사는 우선적으로 조사대상자가 국세청에 제출한 소득증빙자료를 통해 조사대상자의 소득자료를 파악하며, 자료를 제출하지 않은 경우 설문을 통해 파악한다. 따라서 가계금융복지조사 자료에서 연구자는 동일한 개인이 응답한 소득 또는 행정보완자료에 기재된 소득 중 한 자료만 관측할 수 있으나, 예외적으로 2017년에는 동일한 개인에 대해 응답값과 행정보완자료를 동시에 확인할 수 있었다. 이원진 외(2019)는 이 점을 활용하여 개인단위에서 응답값과 행정자료를 연결할 수 있었다.[4] 기존 문헌과 동일하게 행정자료로 계산한 소득불평등도는 응답자료를 활용한 경우에 비해 악화된다고 보고하는 한편, 응답값에 내포된 측정오차에 대해 고찰하지 않았다.

---

3) Jenkins and Rios-Avila(2023)은 설문조사 임금에는 응답오차, 조사 기간 기준의 차이에서 발생하는 오차(reference period error)가 포함될 수 있으며, 행정자료 임금에는 설문조사와 연결할 때 발생하는 오차(linkage error)와 소득신고를 하지 않아 발생하는 오차가 포함될 수 있다고 보고한다. 두 자료를 연결할 때 발생할 수 있는 오차 분포를 다양하게 적용하여 추정치 특징의 변화나 모형의 적합도 등을 고찰하였다. Abowd and Stnson(2013)은 Survey of Income and Program Participation(SIPP)와 두 가지 행정자료, SSA와 SSA의 Detaild Earnings Record(DER) 자료를 일자리 수준(job-level)에서 연결하여 측정오차를 고찰한다. 두 행정자료에 가중치를 부여하고 가상의 참값을 생성하였는데, 각 행정자료에 부여하는 가중치에 따라 계산되는 오차의 수준이 달라져 측정오차 고찰 시 참값으로 정의하는 자료의 중요성을 시사한다.

4) 이원진 외(2019)와 같은 자료를 사용한 이혜정(2022)은 조사된 근로소득 자료의 품질을 높이기 위한 통계적인 방법을 제시한다.

다음으로 측정오차의 특징을 탐구하려고 시도한 문헌이 있다. 김준원, 신동균(2016)은 노동패널자료에 응답오차가 존재하며 이는 고전적 가정을 따르지 않는다고 보고하였다. 동일한 개인을 매년 조사하는 패널의 특징을 활용하여 시간의 흐름에 따라 변하지 않아야 하는 변수들이 시간에 따라 다르게 응답하는 정도를 활용해 오차를 파악하였다. 정재현(2020)은 한국조세재정연구원이 수집하는 재정패널을 활용하여 측정오차의 특징을 탐구하였고, 응답자료의 평균소득이 행정보완자료의 평균 소득에 비해 작다는 것을 보고하였다. 그러나 재정패널도 가계금융복지조사와 동일한 방식으로 소득자료를 수집하여 응답값과 행정자료 중 하나의 값만 관측할 수 있다. 따라서 해당 연구에서 비교한 개인은 동일하지 않아 해석에 한계가 있다. 강석훈(2000)은 1996년 국민계정을 행정자료로 보고 이 자료와 가구소비실태조사 자료를 비교하며 거시자료와 미시자료의 일관성을 확인하였다.

본고의 문헌에 대한 기여는 다음과 같다. 첫째, 지역별고용조사 임금자료에 내재된 측정오차의 특징을 파악하고 측정오차의 영향을 고찰한다. 둘째, 측정오차의 종류에 따라 추정치에 발생하는 편의와 추론(inference)에 미치는 영향을 정리하였다.

Ⅲ. 자 료

본 연구에서는 2015~2018년 임금근로자의 근로소득 자료 두 가지를 활용한다. 첫째, 국세청(National Tax Services, NTS) 행정자료이다. 2015~2018년 동안 국세청에 소득을 신고 한 인구 중 횡단면으로 임의추출 된 10%의 자료를 국세청을 통해 받았고, 그중 임금근로자의 자료만 활용하였다. 본 연구는 임금근로자의 근로소득 자료를 연도별, 광역자치단체 수준의 지역별, 성별, 분위 별로 생성된 자료로 재구성하였다.[5] 집단은 연도-지역-성(性)으로 정의하고, 분위는 집단 내에서의 분위를 의미한다.[6] 세종을 제외한 16개 시도지역을 활용하였다. 본 연구에서 활용한

5) 임금근로자의 소득자료만 활용하였기 때문에 국세청 자료의 근로소득은 임금을 의미한다.

6) 본 연구에서 활용하는 자료는 집단을 대표하는 응축된(aggregated) 하나의 값(예, 평균 또는 합)이 아닌 집단 내 분위별 자료이다. 예컨대, 15분위 값의 경우 11분위에서 15분위 임금자료의 평균이 아닌 15분위에 존재하는 값을 추출한 것이다. 따라서 평균값에 비해 개인단위 측정오차 특성을 보다 잘 반영할 것이다. 측정오차가 존재하는 자료의 평균값을 활용하면 그 값에 포함된 측정오차는 개인단위 자료에 포함된 측정오차의 크기에 비해 감소하기 때문이다(Bound et al., 2001, p. 3830). 단, 각 집단 내에 포함된 표본의 숫자가 짝수라면, 중간 분

분위는 10분위, 15분위, …, 90분위, 95분위의 자료이다. 임금근로자의 연간 소득은 세금 공제 전 소득으로 상여금을 포함한다.

둘째, 지역별고용조사(Regional Employment Survey, RES) 자료다. 지역별고용조사는 지역별 고용과 임금 정보를 제공하며 매년 4월과 10월에 조사된다. 본 연구는 4월과 10월 자료(A형)를 모두 활용하였다. 지역별고용조사 자료도 국세청 자료와 동일하게 재구성하였다.[7] 지역별고용조사 자료를 활용하는 이유는 다음과 같다. 지역별고용조사는 다른 설문조사들에 비해 표본 수가 크고 특정 지역에 편향되지 않아 지역별 소득분포를 파악하기에 유리하다. 또한, 지역별고용조사의 임금자료를 활용하는 연구는 많음에도 불구하고 지역별고용조사의 측정오차에 관한 연구와 이를 고려한 연구는 없었다. 최저임금 정책이 개인의 임금에 미친 영향(김태훈, 2019; 김낙년, 2020), 코로나19가 개인의 임금에 미친 영향(정예은·홍지훈, 2023; 김정우 외, 2021), 지역 간 임금격차(김우영, 2012; 서선영·문영만, 2023) 등의 최근 연구에서 지역별고용조사의 임금 자료를 활용하였다. 또한 재정패널이나 가계금융복지조사 임금자료의 측정오차를 탐구한 연구는 존재하지만, 지역별고용조사의 측정오차를 고찰한 연구는 확인되지 않았다.

지역별고용조사의 월 평균 소득자료는 '최근 3개월간 주된 직장(일)에서 받은 월평균 임금 또는 보수는 얼마였습니까? (세금공제 전 월평균 총 수령액 기준, 각종 상여금 및 현물 등 포함)'라는 질문을 통해 조사되었다. 본 연구에서는 월 소득자료에 12개월을 곱하여 연 소득자료로 활용한다. 조사원이 직접 조사하지만 짧은 시간 내에 세금, 상여금 등 여러 가지 요소를 고려하고 계산해야 하므로 정확하지 않을 수 있다. 이러한 오류는 연 소득을 만들기 위해 12를 곱하는 과정에서 과장될 수 있다. 지역별고용조사 임금자료에 측정오차가 발생할 수 있는 원인은 5.3절에서 다루도록 한다.

자료검증연구에서는 개인의 참값과 응답값을 결합하고 측정오차를 계산한다. 그러나 국내의 어떠한 행정자료에도 개인단위에서 설문조사 자료와 연결할 수 있는

위 주변 값들의 평균값을 활용하였다.

7) 지역별 고용조사의 추출 오류(sampling error)를 줄이기 위해 가중치를 활용하여 자료를 재구성하였으나 단위 무응답으로 인한 추출 오류로 인한 문제를 해결하지 못했다. 단, 통계청(2023, 2024)은 단위 무응답이 발생한 경우, 대체 표본을 활용하였고 이에 대해 가중치를 계산하였다고 보고하였다. 따라서, 자료 재구성 시 가중치를 활용하였기 때문에, 단위 무응답으로 발생할 수 있는 오류를 최소화 했을 것이다.

식별변수가 없다. 따라서 본 연구는 국세청과 지역별고용조사의 자료를 연결하기 위한 식별변수를 앞서 정의한 집단으로 활용하였다.[8] 예컨대, 2015년-남성-서울-10분위의 국세청 자료를 2015년-남성-서울-10분위의 지역별고용조사 자료와 연결하였다. 표본 수는 2,304개(=4개년×16개 시도・광역시×2개 성별×18개의 분위)이며, 로그와 소비자물가지수(2015=1)를 활용해 변환한 로그 실질 연 근로소득을 연구에 활용한다. 국세청 임금자료는 참값으로 지역별고용조사의 임금자료는 응답값으로 가정한다.[9]

---

8) 이는 반복되는 횡단면 자료(repeated cross-sectional data)를 집단 수준에서 종단면 자료로 활용하는 pseudo-panel 자료와 유사해 보일 수 있다. 그러나 pseudo-panel 자료의 변수들은 집단별 평균값을 활용하였기 때문에 집단별 분위 값을 활용하는 본고의 자료와 차이가 있다. Pseudo-panel 자료의 이론적 논의를 집단의 평균값이 아닌 집단 내 분위값으로 확장해 볼 수 있을 것이며 이는 후속연구로 남긴다. Deaton(1985), Verbeek(2008) 그리고 Guillerm(2017)은 집단별 평균값으로 생성한 pseudo-panel 자료의 경우 측정오차가 존재하고 개인의 고유한 특성이 설명변수들과 상관관계가 있는 경우 집단 내의 표본 수가 무한으로 증가하면(일반적으로 2,000개 이상) 선형 패널 고정효과 모형을 통해 일치 추정량을 계산할 수 있다고 보고한다. 또한 Moffitt(1993)은 집단별 평균값($\overline{x_c}$)을 개인값(x_{it})의 도구변수로 여기고 추정하는 경우, 선형 패널 고정효과를 통해 얻은 일치 추정량과 동일한 추정량을 계산할 수 있음을 보고한다.

9) 본고에서는 국세청 임금 자료를 지역별고용조사 임금자료의 참값으로 "가정"한다. 본 각주에서는 국세청 임금 자료의 개념과 임금을 보고하는 대상이 지역별고용조사와 다를 가능성과 그것이 결과에 미치는 영향에 대해 간략히 논의한다. 첫째, 비공식 부문 근로자는 지역별고용조사에서는 조사대상에 포함되지만, 국세청 자료에는 포함되지 않을 가능성이 높다. 만일 비공식 부문 근로자의 임금이 국세청 자료에 포착된다면, 국세청 자료의 분산 크기가 커질 것이다. 따라서 본 연구에서 제시한 응답자료의 분산과 국세청 자료의 분산 차이는 과소계산 된 것이다. 또한, 비공식 근로자의 소득이 대부분 저소득이라면 제시한 저소득 근로자의 측정오차는 과소계산 된 것이다. 둘째, 중식비, 교통비, 복리후생비 등과 같은 비과세 소득은 지역별고용조사 자료에 포함될 가능성이 높지만, 국세청 소득자료는 비과세 소득을 차감한 '과세대상 급여금액'이다. 저임금 근로자는 중식비, 교통비 등을, 고임금 근로자는 이에 더해 복리후생비까지 이미 응답자료에 포함했을 가능성이 높다. 만일 국세청 소득자료와 비교가능하게 만들기 위해 지역별 고용조사 자료에서 비과세 임금을 제외하면 본고에서 활용한 응답값보다 작아진다. 즉, 저임금 근로자의 측정오차는 과대, 고임금 근로자는 절대값 기준으로 과소 계산된 것이다. 한편, 선행연구에서도 상술했듯이 최근 연구에서는 참값 자료에 오차가 없다는 가정을 완화하고 측정오차를 고찰한다.

Ⅳ. 측정오차의 이론적 배경

4.1. 추정치의 편의

본 절에서는 추정치에 발생할 수 있는 편의의 크기를 측정오차가 종속변수와 독립변수에 존재하는 경우로 나누어 계산한다. 본고에서는 선형의 최소자승법을 활용한 경우에 대해 고찰한다. 각각의 경우를 최소자승법을 적용한 횡단면 분석과 집단별로 차분 후 최소자승법을 적용한 종단면 분석으로 나누어 계산한다.[10] 첫째, 측정오차가 있는 집단(i)의 임금자료가 종속변수로 활용되는 경우이다.[11] 우선 횡단면 분석시 측정오차로 인해 발생할 수 있는 편의의 크기를 계산한다. 소득 및 임금 프리미엄을 추정할 때, 연구자들은 측정오차(u_1)가 존재하지 않은 임금(w)을 활용하여 식 (1)을 추정하기를 기대한다. 여기서 측정오차는 식 (2)와 같이 응답 값(w^*)에서 참값(w)을 뺀 값이다. 논의를 단순화하기 위해 이변량(bivariate) 회귀분석을 하며, 임의오차(ε)는 다른 변수들과 연관되어 있지 않다고 가정한다.

$$w_i = \beta x_i + \varepsilon_i \tag{1}$$

그러나 많은 경우 식 (2)와 같이 측정오차를 내포하고 있는 응답자료(w^*)를 활용한 식 (3)을 추정하며 임금프리미엄을 계산한다.

$$w_i^* = w_i + u_{1i} \tag{2}$$

$$w_i^* = \beta x_i + \varepsilon_i + u_{1i} \tag{3}$$

측정오차는 가산성 백색잡음(additive white-noise)의 특성이 있는 것으로 가정한다. $E(u_{1i}) = 0$와 $Cov(w_i, u_{1i}) = 0$를 만족하는 고전적 측정오차는 추정치에 편의

10) 자료검증연구 문헌은 고전적 또는 비고전적 측정오차가 선형모델과 비선형모델에 존재하는 경우로 나누어 발달하고 있다. 본 연구에서는 선형모델만을 다루며, 비선형모델에 측정오차가 존재하는 경우는 후속 연구로 남긴다. 비선형모델에 측정오차가 존재하는 경우 발생할 수 있는 문제에 대한 전반적인 이해를 돕기 위해 Chen et al. (2011) 등의 연구를 참고할 수 있다.

11) i는 연도-지역-성(性)으로 구분한 집단을 의미한다.

를 발생시키지 않는다.[12] 본 연구에서는 참값과 음의 관계($Cov(w_i, u_{1i}) < 0$)를 갖는 평균으로 회귀하는 측정오차를 도입한다. 이 관계를 고려하면 식 (2)를 식 (4)와 같이 작성할 수 있으며, δ_1는 -1과 0사이의 값을 갖는다(Bound et al., 1994). 임의오차(η_1)는 측정오차(u_1), 참값(w) 그리고 응답값(w^*)과 관련이 없다고 가정한다. 식 (4)를 식 (2)에 대입하면 식 (5)로 정리할 수 있다. 여기서 $1+\delta_1$은 0과 1사이의 값을 갖는다. 식 (5)를 식 (1)에 대입하여 정리하면 연구자가 활용하는 임금프리미엄 방정식인 식 (6)으로 정리된다.

$$u_{1i} = \delta_1 w_i + \eta_{1i} \tag{4}$$

$$w_i^* = (1+\delta_1) w_i + \eta_{1i} \tag{5}$$

$$w_i^* = (1+\delta_1)\beta x_i + (1+\delta_1)\varepsilon_i + \eta_{1i} \tag{6}$$

정리하면, 종속변수에 활용되는 임금에 내포된 오차가 고전적 측정오차인 경우 식 (3)을 추정하게 되고 평균회귀 측정오차인 경우 식 (6)을 추정하는 것이다. 식 (6)의 δ_1가 0이라면, 고전적 측정오차가 존재하는 식 (3)과 같다. 각각의 경우 횡단면 분석시 β를 계산할 때 발생하는 측정오차의 크기를 〈Table 1〉 (가)의 (1)열과 (3)열에 제시한다. 종속변수에 존재하는 오차가 고전적 측정오차라면 추정치는 참값을 활용하여 추정한 식 (1)의 추정치와 동일하게 계산되고 편의는 없다. 종속변수에 평균회귀 측정오차가 존재하면 추정치는 $\widehat{(1+\delta_1)}\beta$으로 β에 비해 작게 계산된다.[13] 편의의 크기는 $-\delta_1$이며, $1+\delta_1$는 0과 1사이에 존재하여 추정치를 0에 가깝게 만들기 때문에 희석편의(attenuation bias)라고 한다. 많은 실증분석 연구에서 측정오차가 없다고 가정하거나, 존재하여도 심각한 문제를 발생시키지 않는 경우는 고전적 측정오차가 종속변수에 존재할 때이다.

다음으로 종단면 분석시 측정오차로 인해 발생할 수 있는 편의의 크기를 계산한다. 식 (6)에서 수준(level) 변수 대신 차분(△) 형태의 변수를 활용하여 계산하고 이를 다시 쓰면 식 (7)과 같다. 종단면 분석을 활용한 경우의 편의의 크기는

12) 엄밀히 말해 $Cov(x_i, u_{1i}) = 0$와 $Cov(\varepsilon_i, u_{1i}) = 0$도 만족해야 한다.

13) 추정치 계산과정은 부록 1에 제시하며, 추정치는 $(1+\delta_1)\beta$로 확률수렴(convergence in probability)한다.

〈Table 1〉 (가)의 (2)열에 고전적 측정오차일 때의 편의를 (4)열에 평균회귀 측정오차일 때의 편의를 제시한다. 횡단면 분석과 동일하게 고전적 측정오차인 경우 편의가 발생하지 않았고 평균회귀 측정오차인 경우 편의의 크기는 $-\delta_1$로 희석편의가 발생한다. 나아가 종단면 분석의 경우 임의오차(η_1)가 식 (8)과 같이 AR(1)을 따르는 경우를 고려한다. 식 (8)의 ρ_1는 $0 < \rho_1 < 1$의 값을 갖고 e_1는 임의 오차이다. 여전히 희석편의의 크기는 $-\delta_1$이다.

$$\triangle w_i^* = (1+\delta_1)\beta\triangle x_i + (1+\delta_1)\triangle\varepsilon_i + \triangle\eta_{1i} \tag{7}$$

$$\eta_{1it} = \rho_1\eta_{1it-1} + e_{1it} \tag{8}$$

둘째, 독립변수(x)에 활용되는 집단(i)의 근로시간 자료에 측정오차가 포함된 경우이다. 우선 횡단면 분석시 측정오차로 인해 발생할 수 있는 편의의 크기를 계산한다. 노동생산함수를 추정할 때, 연구자는 측정오차(u_2)가 없는 근로시간 자료(x)를 활용하여 식 (1)을 추정하기를 기대한다. 이번 경우 식 (1)에서는 종속변수인 임금(w)에는 측정오차가 존재하지 않고 독립변수인 근로시간에만 측정오차가 존재하는 경우이다. 고전적 측정오차가 종속변수에 존재할 때 편의를 발생시키지 않지만, 독립변수에 존재하면 추정치에 심각한 편의를 발생시킨다는 것은 알려진 사실이다(Hansen, 2020, p. 348; Bound et al., 1994, p. 346; Pischke, 2007, p. 3). 독립변수에 평균으로 회귀하는 측정오차를 도입하면, 측정오차(u_2)와 참값(x)과의 관계는 식 (9)로, 응답값(x^*)과 참값(x)의 관계는 식 (10)으로 작성할 수 있다. 식 (10)은 식 (11)으로 재정리할 수 있으며, δ_2는 δ_1과 같은 특징을 갖는다.

$$u_{2i} = \delta_2 x_i + \eta_{2i} \tag{9}$$

$$x_i^* = (1+\delta_2)x_i + \eta_{2i} \tag{10}$$

$$x_i = \frac{1}{1+\delta_2}x_i^* - \frac{1}{1+\delta_2}\eta_{2i} \tag{11}$$

식 (11)을 식 (1)에 대입하면 식 (12)로 정리할 수 있으며, 이는 평균회귀오차가 포함된 응답자료를 활용하여 노동생산함수를 추정할 때 사용되는 식이다.

$$w_i = \beta\frac{x_i^*}{1+\delta_2} - \beta\frac{\eta_{2i}}{1+\delta_2} + \varepsilon_i \tag{12}$$

횡단면 분석을 활용한 경우, 고전적 측정오차일 때의 편의를 〈Table 1〉 (나)의 (1)열에, 평균회귀오차일 때의 편의를 (3)열에 제시한다. 고전적 측정오차가 포함되었을 때 편의의 크기는 $\frac{\sigma_{u_2}^2}{\sigma_x^2+\sigma_{u_2}^2}$이다. 편의는 0과 1사이 값으로 추정치를 희석한다. 따라서 여러 자료검증연구 문헌에서 이러한 편의를 비례적 희석편의(proportional attenuation bias)라 한다(Bound and Krueger, 1991). 이러한 형태를 일반적으로는 noise to total variance ratio 또는 noise to signal ratio라 한다(Bound et al., 2001). 고전적 측정오차가 독립변수에 존재하는 경우, 희석편의가 발생한다는 것은 계량경제학 교과서와 여러 문헌을 통해 알려져 있다. 평균회귀 측정오차가 포함되었을 때는 편의의 크기는 $\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2} = \lambda_c$에 따라 다르다. 여기서 λ는 전체 분산 대비 참값 분산의 비율과 같은 값으로 자료의 신뢰성(reliability ratio; signal-to-total variance ratio)을 파악할 수 있다(Bound and Krueger, 1991, p. 18). 횡단면 자료에서 계산된 λ는 λ_c로 표기하고 종단면 자료에서의 신뢰성은 λ_p로 표기한다. 편의의 크기가 $0 < \lambda_c < 1$이면, $1 - \frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2} = 1-\lambda_c$이고, $\lambda_c > 1$이면, $\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2} - 1 = \lambda_c - 1$이다. 고전적 측정오차가 포함될 경우와 같이 희석편의가 발생할 수도 있지만, 추정치가 오히려 커질 수도 있다.[14] 각 편의의 크기는 σ_x^2, $\sigma_{u_2}^2$, $\sigma_{\eta_2}^2$의 크기에 따라 결정된다. $\delta_2 = 0$이면, 고전적 측정오차가 포함되었을 때의 편의와 동일하다.

다음으로 종단면 분석시 측정오차로 인해 발생할 수 있는 편의의 크기를 계산한다. 식 (12)에서 수준(level) 변수 대신 차분(△)형태의 변수를 활용하여 식 (13)로 정리한다.

---

14) $\lambda_c = \frac{1}{(1+\delta_2)}\frac{(1+\delta_2)^2\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2}$로 다시 정리할 수 있다. $0 < 1+\delta_2 < 1$이기 때문에 첫 번째 항은 1보다 크지만 두 번째 항은 0과 1사이 값을 갖기 때문에 λ_c 값의 방향은 알 수 없다.

$$\triangle w_i = \beta\frac{\triangle x_i^*}{1+\delta_2} - \beta\frac{\triangle \eta_{2i}}{1+\delta_2} + \triangle \varepsilon_i \tag{13}$$

종단면 분석을 활용한 경우, 고전적 측정오차일 때의 편의를 〈Table 1〉 (나)의 (2)열에, 평균회귀오차일 때의 편의를 (4)열에 제시한다. 고전적 측정오차인 경우의 편의의 크기는 $\frac{(1-\rho_2)\sigma_{u_2}^2}{\sigma_x^2+(1-\rho_2)\sigma_{u_2}^2}$이며 0과 1사이 값을 갖기 때문에 비례적 희석편의이다. $\rho_2=0$라면 횡단면으로 분석할 때 발생하는 편의의 크기와 같고, $\rho_2=1$이면 편의는 사라진다. 평균회귀오차인 경우 추정치에 포함된 편의의 크기는 $\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}=\lambda_p$에 따라 다르다.[15] $0<\lambda_p<1$이면, $1-\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}=1-\lambda_p$을 $\lambda_p>1$이면, $\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}-1=\lambda_p-1$가 편의의 크기이다. 횡단면 분석과 동일하게 평균회귀 측정오차가 포함된 경우에는 희석편의가 발생할 수도 있지만, 추정치가 오히려 커질 수 있다. 각 편의의 크기는 σ_x^2, $\sigma_{u_2}^2$, $\sigma_{\eta_2}^2$의 크기에 따라 결정되며, $\delta_2=0$이면, 고전적 측정오차가 포함되었을 때의 편의와 동일하다. 식 (1)의 종속변수와 독립변수에 측정오차가 동시에 존재하는 경우 추정치에 포함된 편의의 크기를 계산하여 부록 2에 제시한다.

측정오차의 종류와 횡단면, 종단면 분석방법에 따라 편의의 크기를 비교하면서 어떤 방법이 편의를 줄일 수 있는지 파악해본다. 종속변수에 고전적 측정오차가 포함되는 경우는 횡단면이나 종단면 분석방법 모두 편의가 존재하지 않는다. 평균회귀 측정오차가 포함되는 경우에는 편의가 존재하긴 하나 두 방법 모두 동일한 크기의 편의를 발생시킨다. 독립변수에 고전적 측정오차가 포함되는 경우는 횡단면에 비해 종단면 분석이 유리하다. 자기상관이 높을수록 편의는 작아지며, $\rho_2=1$인 경우 편의는 사라진다. 평균회귀 측정오차가 포함되는 경우에는 어느 방법이 유리한지는 상황에 따라 다르다. $0<\lambda_c<1$인 경우 λ_p가 $0<\lambda_p<1$와 $\lambda_p>1$의 값을

15) $\lambda_p=\frac{1}{(1+\delta_2)}\frac{(1+\delta_2)^2\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}$로 다시 정리할 수 있다. $0<1+\delta_2<1$이기 때문에 첫 번째 항은 1보다 크지만 두 번째 항은 0과 1사이 값을 갖기 때문에 λ_p 값의 방향은 알 수 없다.

가질 수 있어 일반화할 수 없다. $\lambda_c > 1$인 경우 종단면 분석의 경우 측정오차의 자기상관이 높을수록 편의가 커지기 때문에 횡단면 분석이 종단면 분석에 비해 유리하다. 다음절에서는 이를 바탕으로 지역별 고용조사에 존재하는 오차의 종류 및 특징을 고찰한다.

〈Table 1〉 Summary of Bias

Classical Measurement error $(\delta_1 = \delta_2 = 0)$		Mean-reverting Measurement error $(-1 < \delta_1, \delta_2 < 0)$	
(1)	(2)	(3)	(4)
Cross-sectional $(\rho_1 = \rho_2 = 0)$	Longitudinal $(0 < \rho_1, \rho_2 \leq 1)$	Cross-sectional & $1-\lambda_c$ or $\lambda_c - 1$ $(\rho_1 = \rho_2 = 0)$	Longitudinal $1-\lambda_p$ or $\lambda_p - 1$ $(0 < \rho_1, \rho_2 \leq 1)$
(가) When measurement errors are in the dependent variables:			
0	0	$-\delta_1$	$-\delta_1$
(나) When measurement errors are in the independent variables:			
$\frac{\sigma_{u_2}^2}{\sigma_x^2 + \sigma_{u_2}^2}$	$\frac{(1-\rho_2)\sigma_{u_2}^2}{\sigma_x^2 + (1-\rho_2)\sigma_{u_2}^2}$	(나-1) $0 < \lambda_c < 1$	(나-1) $0 < \lambda_p < 1$
		$1 - \frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2 + \sigma_{\eta_2}^2}$	$1 - \frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2 + (1-\rho_2)\sigma_{\eta_2}^2}$
		(나-2) $\lambda_c > 1$	(나-2) $\lambda_p > 1$
		$\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2 + \sigma_{\eta_2}^2} - 1$	$\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2 + (1-\rho_2)\sigma_{\eta_2}^2} - 1$

Note: Author summarizes bias as discussed in Chapter 4.

4.2. 추정치의 추론

본 절에서는 앞절에서 계산한 결과를 바탕으로 횡단면 분석과 종단면 분석 시 측정오차가 포함된 회귀분석 모델의 적합성(goodness of fit)과 추정치의 추론(Inference)에 대해 논의한다. 참값으로 회귀 분석한 모델의 적합성과 추정치의 유의성에 비해 측정오차가 포함된 값으로 회귀 분석한 경우 변하는 크기를 측정오차의 종류에 따라 〈Table 2〉에 제시한다. 모델의 적합성과 추정치의 유의성은 각각 잔차 분산(residual variance)과 추정치의 t값을 계산하여 확인하였다. 횡단면 분석과 종단면 분석의 결과는 같다.

〈Table 2〉 Summary of model fitness and changes in T-value

Classical Measurement error ($\delta_1 = \delta_2 = 0$)		Mean-reverting Measurement error ($-1 < \delta_1, \delta_2 < 0$)	
(1)	(2)	(3)	(4)
Cross-sectional ($\rho_1 = \rho_2 = 0$)	Longitudinal ($0 < \rho_1, \rho_2 < 1$)	Cross-sectional ($\rho_1 = \rho_2 = 0$)	Longitudinal ($0 < \rho_1, \rho_2 \leq 1$)
(가) When measurement errors are in the dependent variables:			
1. Model Fitness			
decrease	decrease	unknown	unknown
2. T-value			
decrease	decrease	decrease	decrease
(나) When measurement errors are in the independent variables:			
1. Model Fitness			
decrease	decrease	decrease	decrease
2. T-value			
decrease	decrease	unknown	unknown

Note: Author summarizes the variation in model fit and t-statistics resulting from the use of RES data relative to NTS data. Appendix 3 details calculation.

측정오차의 종류, 측정오차가 포함된 변수와 분석 방법과 무관하게 모델의 적합성과 추정치의 유의성이 대부분 낮아졌다. 이는 참값을 활용한 분석에 비해 측정오차가 포함된 값을 활용하여 분석할 경우, 잔차 분산이 더 커지며 자료의 신뢰성이 떨어질 수 있음을 의미한다. 단, 평균회귀 측정오차가 종속변수에 포함된 경우의 모델 적합성 변화와 독립변수에 포함된 경우 추정치의 유의성 변화는 알 수 없다. 계산과정은 부록 3에 간략히 제시하였다. 또한, 고전적 측정오차가 종속변수에 포함된 경우 추정치에 편의가 발생하지 않았지만, 모델의 적합도와 추정치의 유의성이 낮아진다는 것은 주목할만하다. 이 경우 참값을 활용하여 분석할 때 비해 추정치의 유의성이 낮아져 가설은 보수적으로 검증될 것이다.

V. 결 과

본 절에서는 지역별고용조사 임금자료에 내포된 측정오차 특징을 파악하고 그로 인한 영향을 제시한다. 측정오차는 응답값과 참값의 차, 즉, 지역별고용조사의 임

금자료에서 국세청 근로소득자료를 뺀 값이다. 모든 분석에서 국세청에 소득을 신고한 연도, 성, 시도 별 임금근로자의 수를 가중치로 활용한다.

5.1. 측정오차의 특징

가. 기술적인 증거

지역별고용조사의 측정오차는 평균회귀 측정오차인 것으로 확인되며 기술적인(descriptive) 증거들은 다음과 같다. 측정오차는 응답값에서 참값을 뺀 값으로 계산한다. 첫째로 측정오차의 평균은 0이다($E(u_i)=0$). 〈Table 3〉의 (A)의 (4)열에 제시된 2015년 측정오차의 평균은 0.04이며 표준오차는 0.33로 통계적으로 0과 다르지 않다. 다른 연도에서도 측정오차의 크기는 유사하며 통계적으로 0과 다르지 않다(〈Table A.1〉. (A)의 (4)열, (7)열, (10)열). 두 번째로 지역별고용조사의 측정오차와 참값 간의 관계는 음의 관계를 갖는다($Cov(w_i, u_i)<0$). 〈Table 3〉의 (C)의 (4)열에 분위별 측정오차를 제시한다. 통계적으로 5% 유의수준에서 유의미한 부분은 음영표시를 하였다. 고소득자의 임금 자료는 음의 측정오차를 저소득자의 임금 자료는 양의 측정오차를 포함한다. 즉, 고소득자는 임금을 과소 보고하며 저소득자는 과대 보고한다. 중위소득 근처에 위치한 사람들은(40, 50, 60분위) 실제 소득과 유사하게 보고하여 측정오차가 0과 다르지 않다. 2015년은 〈Table 3〉에 다른 연도는 〈Table A.1〉에 제시한다.

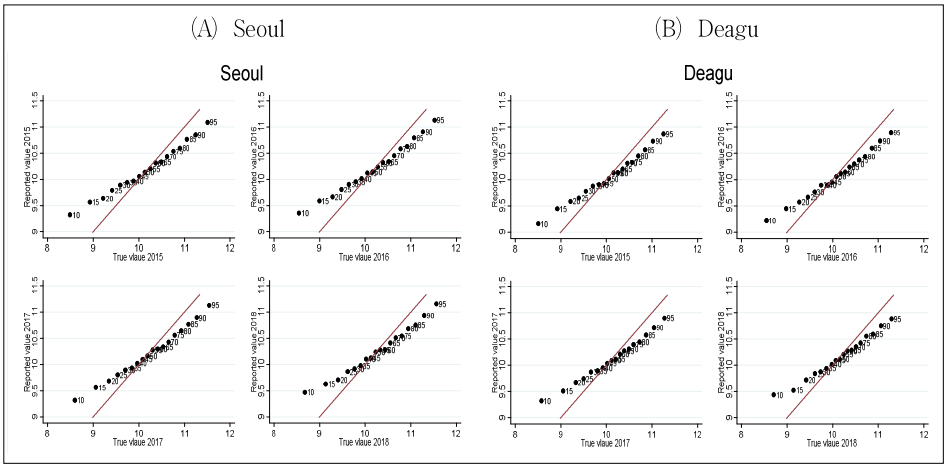
지역과 성별로 나눈 경우에도 평균회귀 측정오차가 확인된다. 지역에 따른 응답값과 참값 간의 관계를 〈Figure 1〉에 제시한다. 대표적으로 서울과 대구를 제시한다. 가로축은 참값 세로축은 응답값을 의미하며, 빨간선은 45도 선으로 두 값이 일치하면 빨간선 위에 점이 놓인다. 서울과 대구 모두 고소득자 응답값은 빨간선 보다 아래에(과소보고) 저소득자의 응답값은 빨간선 보다 위에(과대보고) 위치한다. 이러한 모습은 모든 연도와 모든 지역에서 관찰된다. 지역과 성별에 따른 두 자료의 관계를 〈Figure 2〉에 제시하였고, 소득에 따른 응답 방식은 성별과 무관하였다. 〈Figure 1〉과 동일하게 서울과 대구를 제시한다. 남성과 여성 모두 고소득자는 실제값보다 작게 응답하였고 저소득자는 실제값보다 크게 응답하였다. 〈Table 4〉의 (A)의 (4)열과 (7)열에는 성별과 분위에 따른 측정오차를 제시하며, 고소득자의

〈Table 3〉 Summary Statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Cross-sectional data (2015)			Longitudinal data (2015~2016)		
	obs	RES	NTS	Error	RES	NTS	Error
(A) All							
	576	10.11	10.07	0.04	0.03	0.04	-0.01
		(0.51)	(0.80)	(0.33)	(0.05)	(0.03)	(0.05)
(B) By sex							
Male	288	10.33	10.29	0.03	0.03	0.03	-0.01
		(0.43)	(0.75)	(0.32)	(0.05)	(0.02)	(0.05)
Female	288	9.79	9.75	0.04	0.04	0.05	-0.01
		(0.46)	(0.75)	(0.31)	(0.05)	(0.03)	(0.06)
(C) By percentiles							
90	32	10.75	11.09	**-0.34**	0.03	0.02	0.01
		(0.28)	(0.27)	**(0.06)**	(0.04)	(0.01)	(0.04)
80	32	10.50	10.76	**-0.26**	0.02	0.02	0.00
		(0.31)	(0.32)	**(0.05)**	(0.04)	(0.01)	(0.04)
70	32	10.33	10.50	**-0.17**	0.04	0.02	0.01
		(0.27)	(0.32)	**(0.07)**	(0.05)	(0.01)	(0.06)
60	32	10.21	10.29	-0.08	0.03	0.03	0.00
		(0.29)	(0.30)	(0.05)	(0.05)	(0.01)	(0.04)
50	32	10.08	10.08	-0.01	0.02	0.03	-0.01
		(0.25)	(0.28)	(0.06)	(0.04)	(0.01)	(0.04)
40	32	9.93	9.86	0.07	0.05	0.05	0.00
		(0.24)	(0.25)	(0.06)	(0.06)	(0.01)	(0.06)
30	32	9.83	9.58	**0.25**	0.02	0.06	-0.04
		(0.28)	(0.24)	**(0.08)**	(0.05)	(0.02)	(0.05)
20	32	9.62	9.20	**0.42**	0.03	0.07	-0.04
		(0.26)	(0.26)	**(0.08)**	(0.05)	(0.02)	(0.06)
10	32	9.25	8.47	**0.78**	0.06	0.06	-0.01
		(0.38)	(0.34)	**(0.10)**	(0.07)	(0.04)	(0.07)

Note: Measurement errors are calculated by subtracting the NTS data from the RES data in each percentile. We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data. The standard errors are in parentheses. Measurement errors in the shaded areas are statistically significant at the 5% level.

〈Figure 1〉 Relationship of RES and NTS by region



Notes: X-axis refers to the NTS data and Y-axis refers to the RES data. The red line is 45 degree line. Due to space limitation, we present graphs for two regions. Graphs for other regions are available upon request.

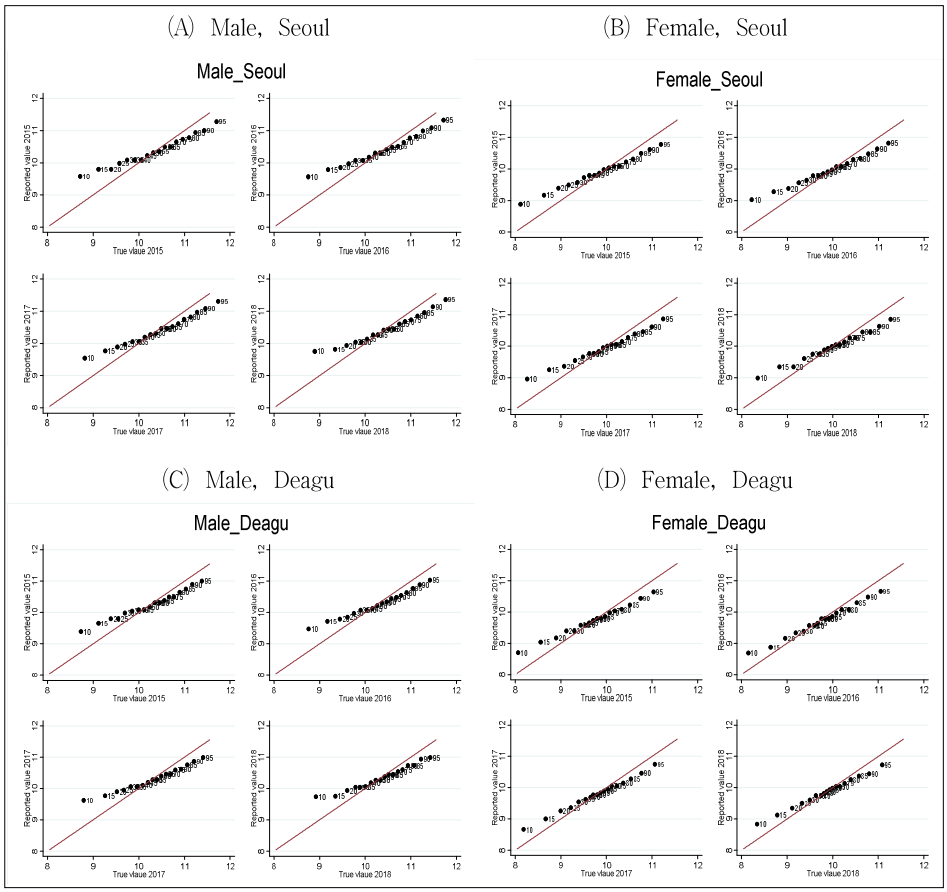
임금자료는 음의 측정오차를 저소득자는 양의 측정오차를 포함하고 중위소득자의 임금자료에 내포된 측정오차의 크기는 0과 다르지 않다. 2015년은 〈Table 4〉에 다른 연도는 〈Table A.3〉에 제시한다.

마지막으로 응답값이 참값에 비해 분산의 크기가 작아 응답값에 평균으로 회귀하는 오차가 존재한다는 것을 알 수 있다. 두 자료의 분산은 〈Figure 3〉에 제시하며, 실선은 참값, 점선은 응답값의 분포이다. 〈Table 3〉의 (A)의 (2)열에 제시한 2015년 응답값의 표준오차는 0.51로 (3)열에 제시한 참값의 표준오차인 0.80에 비해 작다는 것을 통해서도 알 수 있다. 〈Table A.1〉에 제시한 다른 연도 응답값의 표준오차도 참값의 표준오차에 비해 작다.

나아가 측정오차의 자기상관성을 확인하기 위해 연도별 측정오차 간의 상관계수를 〈Table 5〉에 제시한다. 2015년과 2016년 측정오차 간의 상관계수는 0.9880이고 다른 연도 간의 상관계수도 1에 가까운 값으로 연구 기간 지역별고용조사 임금자료의 측정오차는 자기상관성이 매우 높다는 것을 알 수 있다. 측정오차의 자기상관성이 높은 경우 일차차분을 통해서 측정오차의 상당 부분을 없앨 수 있다. 예컨대 2015년과 2016년의 90분위 임금을 일차 차분한 결과 측정오차는 0.01(〈Table 3〉의 (C)의 (7)열)로 계산되었으며 이는 통계적으로 0과 다르지 않았다. 2015년 90분위

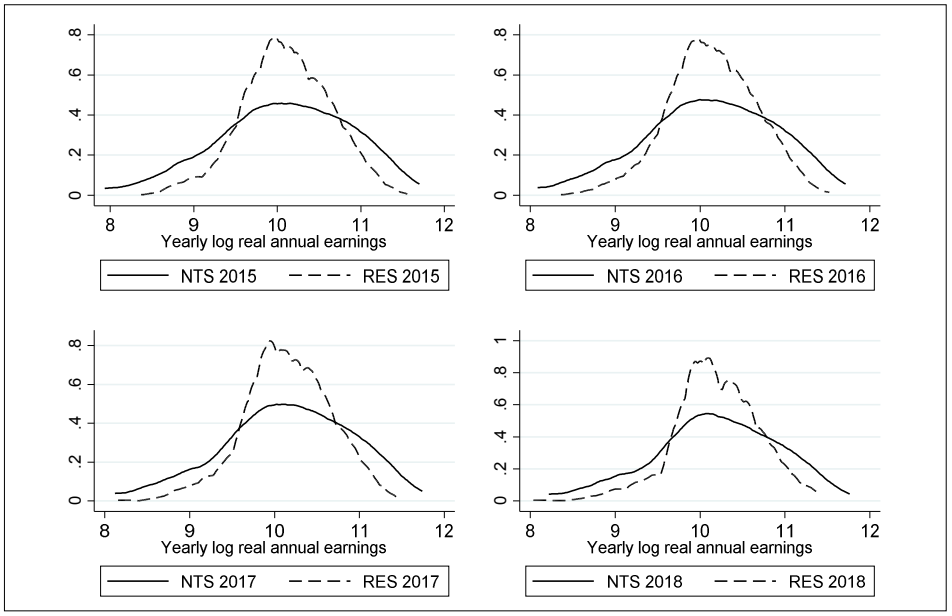
임금에 포함된 측정오차는 -0.34(〈Table 3〉의 (C)의 (4)열)이고 2016년 90분위 임금에 포함된 측정오차는 -0.33(〈Table A.1〉의 (4)열)으로 유사하여 차분을 통해 측정오차를 제거한 것이다. 2015~2016년의 전체 표본을 대상으로 일차 차분한 측정오차를 〈Table 3〉의 (A)의 (7)열에, 남녀로 표본을 나누어 일차 차분한 결과를 〈Table 4〉의 (B)의 (4)열과 (7)열에 제시한다. 모든 분위에서 일차 차분한 측정오차는 통계적으로 0과 다르지 않았다. 다른 연도에 대해 전체 표본을 대상으로 일차 차분한 결과는 〈Table A.2〉에, 남녀 표본을 대상으로 일차 차분한 결과는 〈Table A.4〉에 제시한다. 일차 차분한 측정오차는 모두 0과 다르지 않았다.

〈Figure 2〉 Relationship of RES and NTS by region and sex



Notes: X-axis refers to the NTS data and Y-axis refers to the RES data. The red line is 45 degree line. Due to space limitation, we present graphs for two regions. Graphs for other regions are available upon request.

〈Figure 3〉 Distribution of RES and NTS earning data



Notes: The solid line denotes the NTS data and the dotted line refers to the RES data. The top-left graph represents the year 2015, followed by 2016, 2017, and 2018.

〈Table 4〉 Summary statistics by sex and percentiles

(A) Cross-sectional data (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Male			Female		
p	obs	RES	NTS	Error	RES	NTS	Error
90	16	10.96 (0.11)	11.29 (0.11)	**-0.34** **(0.07)**	10.45 (0.12)	10.79 (0.10)	**-0.34** **(0.06)**
80	16	10.73 (0.10)	11.00 (0.09)	**-0.27** **(0.05)**	10.15 (0.09)	10.40 (0.11)	**-0.25** **(0.03)**
70	16	10.55 (0.08)	10.75 (0.08)	**-0.21** **(0.05)**	10.02 (0.08)	10.14 (0.09)	-0.12 (0.07)
60	16	10.43 (0.09)	10.53 (0.07)	-0.10 (0.05)	9.88 (0.09)	9.94 (0.06)	-0.06 (0.04)
50	16	10.27 (0.08)	10.31 (0.07)	-0.04 (0.06)	9.79 (0.07)	9.76 (0.05)	0.02 (0.04)
40	16	10.11 (0.08)	10.06 (0.08)	0.05 (0.05)	9.67 (0.09)	9.57 (0.03)	0.10 (0.06)
30	16	10.05 (0.09)	9.77 (0.08)	**0.28** **(0.06)**	9.52 (0.08)	9.31 (0.04)	**0.20** **(0.08)**
20	16	9.82 (0.07)	9.40 (0.06)	**0.42** **(0.06)**	9.33 (0.10)	8.90 (0.04)	**0.43** **(0.10)**
10	16	9.55 (0.10)	8.74 (0.07)	**0.81** **(0.07)**	8.81 (0.13)	8.08 (0.06)	**0.74** **(0.12)**

(B) Longitudinal data (2015~2016)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Male			Female		
p	obs	RES	NTS	Error	RES	NTS	Error
90	16	0.03 (0.05)	0.02 (0.01)	0.01 (0.04)	0.03 (0.04)	0.02 (0.01)	0.00 (0.04)
80	16	0.01 (0.03)	0.02 (0.01)	-0.01 (0.03)	0.04 (0.04)	0.02 (0.01)	0.01 (0.04)
70	16	0.04 (0.06)	0.02 (0.01)	0.02 (0.06)	0.03 (0.04)	0.03 (0.01)	0.01 (0.04)
60	16	0.01 (0.04)	0.02 (0.01)	-0.01 (0.04)	0.05 (0.05)	0.03 (0.01)	0.01 (0.05)
50	16	0.02 (0.04)	0.03 (0.01)	-0.01 (0.04)	0.03 (0.04)	0.04 (0.01)	-0.01 (0.04)
40	16	0.05 (0.06)	0.04 (0.01)	0.01 (0.06)	0.04 (0.05)	0.06 (0.01)	-0.02 (0.05)
30	16	0.02 (0.05)	0.06 (0.01)	-0.04 (0.05)	0.03 (0.04)	0.07 (0.02)	-0.04 (0.05)
20	16	0.05 (0.05)	0.07 (0.01)	-0.02 (0.05)	0.01 (0.04)	0.07 (0.02)	-0.06 (0.06)
10	16	0.04 (0.05)	0.05 (0.03)	0.00 (0.05)	0.07 (0.08)	0.09 (0.04)	-0.02 (0.09)

Notes: Measurement errors are calculated by subtracting the NTS data from the RES data in each percentile. We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data. The standard errors are in parentheses. Measurement errors in the shaded areas are statistically significant at the 5% level.

〈Table 5〉 Auto-correlation of measurement errors

	$u_{t=2016}$	$u_{t=2017}$	$u_{t=2018}$
u_{t-1}	0.9880	0.9875	0.9847

Notes: Measurement errors are calculated by subtracting the NTS data from the RES data in each percentile. We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data. All values are statistically significant.

나. 추정을 통한 증거

본 소절에서는 식 (5): $w_i^* = (1+\delta_1)w_i + \eta_{1i}$를 직접 추정하여 $0 < 1+\delta_1 < 1$인지 확인하였다. 횡단면 자료는 최소자승법(OLS), 종단면 자료는 최소자승법의 일

차 차분모형(OLS-FD)을 활용하여 분석하고, 연도 더미를 포함한다. 〈Table 6〉의 (1)열에 횡단면 자료로 분석한 결과를 (4)열에 종단면 자료로 분석한 결과를 제시한다. 〈Table 6〉의 (1)열과 (4)열에 제시된 결과는 $1\widehat{+}\delta_1$이 0과 통계적으로 다르다는 것을 의미하며, (2)열과 (5)열에 제시된 결과는 $1\widehat{+}\delta_1$이 1과 통계적으로 다르다는 것을 의미한다. 따라서 추정치 $1\widehat{+}\delta_1$은 0과 1 사이에 존재하며, 이는 식 (4) : $u_i = \delta_1 w_i + \eta_i$에서 δ_1이 -1과 0 사이에 있음을 의미한다.[16] 즉, 지역별고용조사의 임금자료에 포함된 측정오차와 참값이 음의 상관관계에 있는 것이다.

〈Table 6〉 Relationship between the RES data and the NTS data

Sex	Cross-sectional(OLS)			Longitudinal(OLS-FD)		
	(1)	(2)	(3)	(4)	(5)	(6)
	$1\widehat{+}\delta_1$	$H0: 1+\delta_1 = 1$	obs	$1\widehat{+}\delta_1$	$H0: 1+\delta_1 = 1$	obs
All	0.636*** (0.006)	p=0.000	2,304	0.332*** (0.058)	p=0.000	1,728
Male	0.568*** (0.006)	p=0.000	1,152	0.297*** (0.058)	p=0.000	864
Female	0.624*** (0.012)	p=0.000	1,152	0.283** (0.131)	p=0.000	864

Notes: We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data and by clustering at the regional level. The standard errors are in parentheses. The year fixed effects are controlled.

5.2. 평균회귀 측정오차의 영향

본 절에서는 평균회귀 측정오차의 영향을 살펴보기 위해 응답값과 참값을 각각 활용해 임금프리미엄 추정치와 지니계수를 계산하여 제시한다. 첫째로 임금프리미엄 추정에 미치는 영향을 살펴보기 위해 식 (6) : $w_i^* = (1+\delta_1)\beta x_i + (1+\delta_1)\varepsilon_i + \eta_{1i}$의 종속변수에 응답값($w_i^*$)과 참값($w_i$)을 차례로 넣어 추정한 결과를 비교한다. 이변량(bivariate) 선형 회귀분석방법으로 횡단면 자료는 최소자승법(OLS), 종단면 자

16) $1\widehat{+}\delta_1$는 $1+\delta_1$로 확률수렴($Plim$)하여, $1+\delta_1$이 0과 1사이에 존재한다는 것을 의미한다.

료는 최소자승법 일차 차분모형(OLS-FD)을 활용한다. 국세청 자료의 집단 내 속한 사람 수를 가중치로 활용하여 추정한다. 설명변수는 지역별고용조사 자료를 활용해 임금자료와 유사한 방법으로 연도, 성별, 지역 그리고 20개 백분위 값을 활용하여 생성하였고, 10분위, 15분위, …, 90분위, 95분위 자료를 활용한다. 설명변수는 특정 백분위에 값이 존재하지 않는 경우가 있을 수 있어, 각 분위로부터 2.5분위 낮은 분위에 있는 값부터 2.5분위 높은 분위 사이에 존재하는 값들의 평균값을 활용한다. 생성된 설명변수는 집단별 50세 이상 비중, 전문대학 이상 비중, 미혼 비중, 제조업과 서비스업 고용 비중, 정규직 비중을 포함한다.[17]

응답값을 활용하여 OLS로 추정한 결과는 〈Table 7〉의 (1)열에 OLS-FD로 추정한 결과는 (4)열에 제시하였으며, 참값을 활용하여 OLS로 추정한 결과는 (2)열에 OLS-FD로 추정한 결과는 (5)열에 제시한다. 횡단면 분석 결과 종속변수에 응답값을 활용한 경우, 참값을 활용한 경우에 비해 추정치가 0에 더 가까운 값으로 계산되었다. 이는 희석편의(attenuation bias)로 인한 것으로 보인다. 응답값을 활용하여 계산한 추정치와 참값을 활용하여 계산한 추정치의 차이는 통계적으로 유의미하다(〈Table 7〉의 (3)열). 이러한 모습은 측정오차가 추정치에 미치는 영향을 고찰한 기존 문헌의 결과와 일치한다(Gibson and Kim, 2008).[18] 나아가 응답값을 활용하여

17) 설명변수에 포함된 측정오차의 특징은 알 수 없다. 4.1절과 부록 2에 제시한 것과 같이 측정오차가 설명변수에 존재하는 경우 그 종류에 상관없이 편의를 발생시킬 수 있다. 종속변수에는 평균회귀 측정오차가 존재한 경우, 설명변수에 존재하는 측정오차의 종류는 다음의 세 가지로 나누어 편의의 크기를 계산할 수 있다. 횡단면 분석의 경우를 논의한다. 첫째, 설명변수에 측정오차가 존재하지 않는 경우에는 〈Table 1〉의 (가)의 (3)열에 제시된 크기와 같은 크기의 희석편의가 추정치에 존재한다. 둘째, 설명변수에 고전적 측정오차가 존재하는 경우에는 $1-\frac{(1+\delta_1)\sigma_x^2}{\sigma_x^2+\sigma_{u_2}^2}$ 만큼의 편의를 발생시킨다. 이는 〈Table A.2.1〉의 (3)열에 제시된 편의에 $\delta_2=0$, $\eta_2=u_2$를 대입한 것과 같다. $0<\frac{(1+\delta_1)\sigma_x^2}{\sigma_x^2+\sigma_{u_2}^2}<1$이기 때문에 편의의 크기는 1에서 $\frac{(1+\delta_1)\sigma_x^2}{\sigma_x^2+\sigma_{u_2}^2}$을 빼준 값으로 계산할 수 있다. 이 경우도 추정치에 희석편의가 존재한다. 셋째, 설명변수에 평균회귀 측정오차가 존재하는 경우는 〈Table A.2.1〉의 (3)열에 제시된 크기와 같은 크기의 편의를 발생시킨다. 편의의 크기는 γ_c와 γ_p에 따라 다르다.

18) Gibson and Kim(2008)은 경찰로부터 받은 개인의 범행 여부 자료와 설문조사를 통해 수집한 개인의 범행 여부 응답값을 활용하여 범행여부와 불평등 간의 편상관계수(partial correlation)를 계산하였다. 응답값을 활용하여 추정한 경우 참값을 활용한 경우에 비해 편상관계수의 크기가 0에 더 가깝게 추정되었고, 두 추정치의 차이는 0과 유의미하게 달랐다.

계산한 추정치의 t값의 절대값은 참값을 활용하여 계산한 값에 비해 대체로 작았고 이는 4.2절에서 제시한 결과와 일치하는 결과이다.[19]

〈Table 7〉 The effect of mean-reverting measurement errors on coefficient

	(1)	(2)	(3)	(4)	(5)	(6)
	Cross-sectional			Longitudinal		
	RES	NTS	$H0:\Delta=0$	RES	NTS	$H0:\Delta=0$
Share of aged 50 and above	-2.374*** (0.080) [-29.68]	-4.119*** (0.114) [-36.13]	p=0.000	0.007 (0.011) [0.64]	0.119*** (0.005) [23.80]	p=0.000
Share of junior college and higher	2.253*** (0.029) [77.69]	3.515*** (0.041) [85.73]	p=0.000	-0.026*** (0.006) [-4.33]	-0.100*** (0.003) [-33.33]	p=0.000
Share of the unmarried	2.344*** (0.051) [45.96]	3.542*** (0.078) [45.41]	p=0.000	-0.047*** (0.008) [-5.88]	-0.101*** (0.004) [-25.25]	p=0.000
Share of employment in manufacturing sector	1.971*** (0.071) [27.76]	2.739*** (0.110) [24.90]	p=0.000	-0.046*** (0.009) [-5.11]	-0.078*** (0.005) [-15.60]	p=0.238
Share of employment in service sector	-1.518*** (0.058) [-26.17]	-1.870*** (0.092) [-20.33]	p=0.000	0.052*** (0.007) [7.43]	0.051*** (0.004) [12.75]	p=0.922
Share of regular workers	1.978*** (0.018) [109.89]	3.107*** (0.023) [135.09]	p=0.000	-0.010* (0.005) [-2.00]	-0.082*** (0.002) [-41.00]	p=0.000

Notes: This table reports the estimation result for equation (6) using bi-variate linear regression and the number of individuals in each year-gender-region cell from the NTS data as weights. The standard errors are in parentheses and clustered at regional level. The t-values are in bracket. The year fixed effects are controlled.

종단면 분석 결과 서비스업 고용 비중을 제외한 설명변수의 경우 응답값을 활용하여 계산한 추정치가 참값을 활용하여 계산한 추정치에 비해 0에 더 가까운 값으

19) 응답값으로 계산한 추정치 일부의 t값이 참값으로 계산한 것에 비해 크게 계산된 이유를 두 가지 제시한다. 첫째, 〈Table 2〉에 제시한 결과는 추정치가 확률 수렴한 경우 계산한 결과인데, 본고에서 활용하는 자료의 수가 확률 수렴할 만큼 충분하지 않기 때문일 수 있다. 둘째, 설명변수에 포함된 측정오차의 특징에 따라 t값이 다르게 계산될 수 있기 때문일 수 있다.

로 추정되었고, 두 추정치의 차이는 통계적으로 유의미하였다. 서비스업 고용 비중을 활용한 경우 참값을 활용하여 계산한 추정치가 응답값을 활용하여 계산한 추정치에 비해 0에 더 가깝게 추정되었다. 그러나 두 추정치의 차이는 통계적으로 유의미하지 않았다. 나아가 응답값을 활용하여 계산한 추정치의 t값의 절대값은 참값을 활용하여 계산한 경우에 비해 모두 작았고 이는 4.2절에서 제시한 결과와 일치하는 결과이다.

둘째, 참값과 응답값 각각을 활용해 지니계수를 계산하여 〈Table 8〉에 제시한다. 국세청 자료의 집단 내 속한 사람 수를 가중치로 활용하였다.[20] 응답값을 활용하여 계산한 지니계수는 참값을 활용한 것에 비해 작았다. 즉, 평균회귀 측정오차가 포함된 응답값으로 지니계수를 계산하면 실제 분배 정도보다 개선된다는 것을 확인하였다. 이는 기존 국내 문헌 결과를 지지한다.

〈Table 8〉 Gini Coefficient

Year	RES	NTS
2015	0.279	0.398
2016	0.276	0.391
2017	0.269	0.383
2018	0.264	0.372

Note: We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data.

한편 평균회귀 측정오차가 포함된 자료를 활용할 때 일치 추정량을 계산하는 방법을 기존 문헌을 바탕으로 간략히 소개한다. 고전적 측정오차를 해결하는 방법으로 참값의 자료에서 일부를 층화추출한 자료인 '추가적인(auxiliary) 자료'를 활용하거나 도구변수를 활용하는 방법이 개발되었다.[21] 이 방법들은 비고전적인 측정오차가 있는 경우 그대로 적용할 수 없지만, 여러 문헌에서 이를 그대로 적용하여도 일치 추정치를 계산할 수 있는 방법을 보고한다. Chen, Hong and Tamer(2005)은

20) 지니계수는 ginidesc Stata package를 활용해 계산하였다. 통계청에서 보고하는 지니계수는 가계금융복지조사 자료를 활용하고, 균등화 처분가능 소득을 기준으로 계산한 한편, 본고에서는 세금을 포함한 임금근로소득 자료를 활용하여 계산하여 차이가 존재한다.

21) 가령, Bollinger(1998)은 미국의 SSA 자료에서 social security 번호를 제공한 사람들만의 소득을 참값으로 활용하였다.

비고전적 측정오차인 경우에도 추가적인 자료를 활용하여 최대우도법(Maximum Likelihood Method)의 방법으로 일치 추정량 또는 더 효율적인 추정치로 계산하는 방법을 제시한다. 또한 도구변수를 활용하여 해결하는 방법은 측정오차가 이분법(binary) 설명변수에 포함된 경우(Hu, 2008; Lewbel, 2007 등)와 연속 설명변수에 포함된 경우(Hu and Schennach, 2008; Hahn and Ridder, 2017)로 나누어 개발되었다. 나아가, 선형모델, 비선형모델 등에 따라 가정과 추정 방법을 달리하여 측정오차로 인한 문제를 해결한다(Chen, Hong and Nekipelov, 2011).[22]

5.3. 측정오차의 원인에 대한 논의

지금까지의 논의를 통해 지역별고용조사 임금자료는 평균회귀 측정오차를 포함하는 것으로 확인되었다. 본 절에서는 이러한 측정오차가 발생하는 이유를 논의한다.[23] 첫 번째로 저소득 근로자가 임금을 과대 보고하는 이유를 제시한다. 사회심리학에서는 자기 정체성(self-identity), 사회적으로 바람직함(social desirability), 낙인(stigma) 이론을 제시한다. 저소득자들은 소득과 고용상태를 응답하면서 자신의 정체성을 정의하는 경향이 있기 때문에 사회적으로 바람직하다고 생각되는 수준으로 응답한다는 것이다(Tourangeau et al., 2000, p. 260). 또한 그들은 낮은 임금으로 응답했을 때 받을 수 있는 수치심 또는 사회로부터의 낙인 등을 피하려고 실제 임금보다 높은 수준으로 응답하는 경향이 있다(Mathiowetz et al., 2002).

나아가 저소득자의 측정오차는 절대값 기준으로 고소득자에 비해 크다.[24] 〈Figure 4〉의 가로축은 분위를 세로축은 측정오차를 의미하는데, 모든 연도에서 분위가 낮을수록 그래프의 기울기가 커진다.[25] 이러한 모습이 나타나는 이유를 인지

22) Chen, hong and Nekipelov(2011)은 평균회귀 측정오차가 있는 경우 일치 추정치를 얻을 수 있는 방법을 연구한 문헌을 소개한다. 모수적인 방법, 비모수적(non-parametric)인 방법, 준모수적(semi-parametric) 방법에 따른 추정방법을 소개하고 있으며, 자세한 내용은 각 방법을 제시한 논문을 참고하여 활용하길 바란다.

23) Celhay, Meyer and Mittag(2024)는 미국 사회보장 프로그램 참여 여부에 대한 측정오차 발생 원인을 인지 이론과 사회심리학 이론에서 찾고 제안한 원인이 이론에 부합하는지 검증하였다.

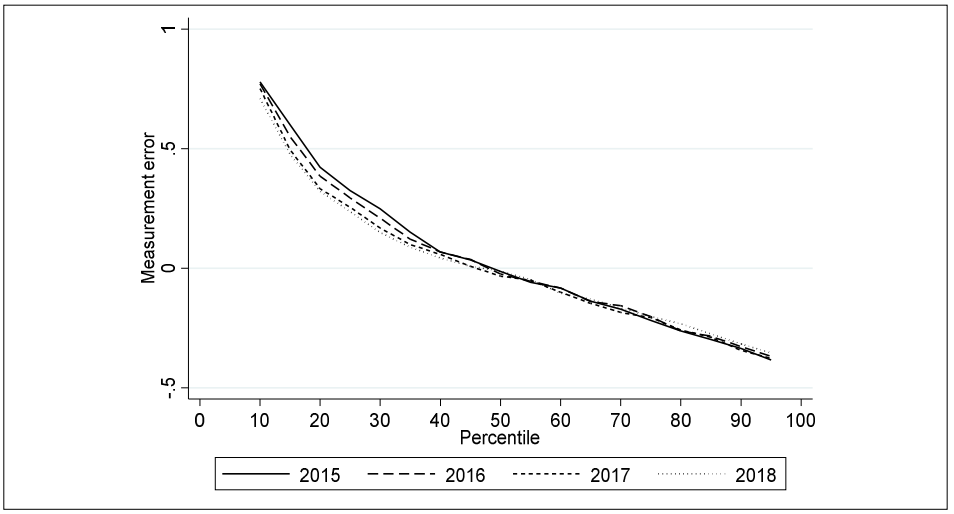
24) 〈Figure 4〉에서 저소득자(특히, 20, 30분위)의 측정오차는 최근으로 오면서 감소한다. 이는 국세청에 소득을 신고하는 비율이 높아져 응답자료와 행정자료 간의 매칭 정확도가 높아지고(〈Table A. 5.〉), 최근 조사대상자들이 이전에 비해 정확한 값을 응답하는 것일 수 있다.

25) 〈Table 3〉의 (C)의 (4)열에 제시한 값을 통해서도 확인할 수 있다. 90분위 임금의 측정오차

이론에서 찾을 수 있다(Moore et al., 2000; Mathiowetz et al., 2002; Bound et al., 2001, p.3743 - p.3748). 이 이론은 저소득자는 고소득자에 비해 예산으로부터 오는 제약이 크기 때문에 소득자료를 응답할 때 추가적인 인지능력이 필요하다고 주장한다. 또한, 고용의 안정성이 낮다면 수입원이 두 개 이상일 수 있는데 이때, 응답자는 특정 기간의 소득을 정확하게 계산하기 어려울 수 있다.

두 번째로 참값을 특정 단위에서 올림, 반올림, 버림하여 응답하기 때문에 오차가 발생할 수 있다. 이러한 응답 방식 때문에 발생한 측정오차를 문헌에서는 히핑오차(heaping error)라 한다(Zinn and Wurbach, 2016).[26] 예컨대, 실제 소득이 1,393,000원인 경우 만 단위에서 버린다면 1,300,000원으로, 만에서 500,000원 단위로 반올림한다면 1,500,000원으로, 만에서 100,000원 단위로 반올림 한다면 1,400,000원으로 응답할 것이다. 이러한 규칙에 따라 생성된 가상의 응답오차는 지역별고용조사의 측정오차를 일부 설명하였다. 자세한 내용은 부록 4에 제시한다.

〈Figure 4〉 Measurement errors by percentiles



Notes: Measurement errors are calculated by subtracting the NTS data from the RES data. We presents the graphs using weights based on the number of individuals in each year-gender-region cell from the NTS data.

는 -0.34이고 10분위 임금의 측정오차는 0.78이다. 다른 연도에서도 유사한 모습이 확인된다.

26) Zinn and Wurbach(2016)은 히핑오차가 추정치에 미치는 영향과 이를 해결하는 방법을 제시한다.

세 번째로 지역별고용조사 임금자료를 재구성하는 과정에서 오차가 과장되었을 수 있다. 지역별고용조사는 최근 3개월 간 받은 평균 임금을 세금과 상여금을 포함해서 응답할 것을 요구한다. 응답자는 상여금 크기를 의도적으로 조절할 수 있다. 특히, 저소득자는 상여금을 받지 않았거나 적게 받았을 가능성이 클 수 있는데, 만일 저소득자가 상여금을 받은 것이 바람직한 모습이라고 생각한다면 이를 받은 것처럼 임금을 과장하여 응답할 수 있다. 본 연구에서 지역별 고용조사의 월 평균 임금을 연 평균 임금으로 만드는 과정에서 이러한 조절이 과장되었을 수 있다.

Ⅵ. 결 론

한국의 자료검증연구는 초기 단계인데, 그 이유는 미국이나 북유럽과 같이 개인 단위에서 참값으로 정의되는 행정자료와 응답값을 연결할 수 없기 때문이다. 본 연구는 국세청의 지역, 성별, 분위별 소득자료를 참값으로 가정하고 지역별고용조사 임금자료의 측정오차를 탐구하였으며 다음의 세 가지 발견을 제시한다. 첫째, 지역별고용조사의 임금자료에는 평균회귀 측정오차(mean-reverting measurement error)가 존재한다. 둘째, 평균회귀 측정오차가 있는 임금자료를 종속변수로 활용하면 계수 값이 희석된다. 셋째, 측정오차의 자기상관이 강하며, 일차 차분 자료를 활용하면 측정오차를 줄일 수 있다. 본 연구의 결과는 실증연구자가 지역별고용조사의 임금 자료의 측정오차에 부여하는 고전적 측정오차 또는 측정오차가 없다는 가정을 점검할 필요가 있으며, 분석 결과 해석 시 측정오차의 영향을 고려하여 해석할 필요가 있음을 시사한다. 본 연구를 시작으로 향후 한국의 여러 설문조사 자료의 측정오차를 개인 수준에서 고찰하는 연구가 가능해지길 바란다.

■ 참 고 문 헌

1. 강석훈, “서베이데이터와 집계데이터의 비교연구: 가구소비실태조사와 국민계정을 중심으로,” 『계량경제학보』, 제11권 제1호, 2000, pp. 41-70.

   (Translated in English) Kang, Seokhoon, “Comparative Study of Survey Data and Aggregate Data: Focusing on the Household Income and Expenditure Survey and the National Accounts,” *Journal of Economic Theory and Econometrics*, Vol. 11, No. 1, 2000, pp. 41-70.

2. 김낙년 · 김종일, “한국 소득분배 지표의 재검토,”『한국경제의 분석』, 제19권 제2호, 2013, pp. 1-64

   (Translated in English) Kim, Nak Nyeon, and Jongil Kim, “Reexamining the Indices of Income Distribution in Korea,” *Journal of Korean Economic Analysis*, Vol. 19, No. 2, 2013, pp1-64.

3. 김낙년, “한국의 최저임금과 고용, 2013-2019,”『한국경제의 분석』, 제26권 제1호, 2020, pp. 145-183.

   (Translated in English) Kim, Nak Nyeon, “Korea’s Minimum Wage and Employment, 2013-2019,” *Journal of Korean Economic Analysis*, Vol. 26, No. 1, 2020, pp. 145-183.

4. 김우영, “한국의 지역간 임금격차: 지역별 고용조사 (RES)를 중심으로,”『노동정책연구』, 제12권 제1호, 2012, pp1-28.

   (Translated in English) Kim, Woo-Yung, “Regional Wage Differentials in Korea : Estimates Based on the Regional Employment Survey,” *Quarterly Journal of Labor Policy*, Vol. 12, No. 1, 2012, pp1-28.

5. 김정우 · 김기민 · 방효훈, “코로나 19 위기가 충남지역 노동시장에 미친 영향: 이중차분법을 활용하여,”『충남연구』, 제5권 제1호, 2021, pp. 27-49.

   (Translated in English) Kim, Jungwoo, Kimin Kim, and Hyohoon Bang, “Impact of the COVID-19 Crisis on the Labor Market in South Chungcheong Province: Using the Difference in Differences Method,” *Chungnam Studies*, Vol. 5, No. 1, 2021, pp. 27-49.

6. 김준원 · 신동균, “서베이 자료에 존재하는 측정오차의 문제점과 자료검증연구의 필요성,”『산업경제연구』, 제29권 제1호, 2016, pp. 249-278.

   (Translated in English) Kim, Joonwon, and Donggyun Shin, “Measurement Errors in Survey-Based Variables: Why Do We Need Data Validation Studies?” *Journal of Industrial Economics and Business*, Vol. 29, No. 1, 2016, pp. 249-278.

7. 김태훈, “최저임금 인상의 고용 및 임금효과,”『노동정책연구』, 제19권 제2호, 2019, pp. 135-174.

   (Translated in English) Kim, Taehoon, “The Effects of the Minimum Wage on Employment and Wages,” *Quarterly Journal of Labor Policy*, Vol. 19, No. 2, 2019, pp. 135-174.

8. 서선영 · 문영만, “부울경 지역의 성별 임금격차 및 해소 방안 연구: 경력단절 변수를 중심으로,”『한국지방정치학회보』, 제13권 제2호, 2023, pp. 49-70.

   (Translated in English) Seo, Seonyoung, and Youngman Moon, “A Study on the Gender

Wage Gap and Solutions in Busan, Ulsan, and Gyeongnam: Focusing on Career Interruption Variables," *The Korean Regional Politics Review*, Vol. 13, No. 2, 2023, pp. 49-70.

9. 이원진 · 정해식 · 전지현, 『소득조사 마이크로데이터 비교 분석-〈가계동향조사〉와 〈가계금융 · 복지조사〉를 중심으로』, 연구보고서(수시) 2019-13, 보건사회연구원, 2019.

(Translated in English) Lee, Wonjin, Haesik Jung, and Jihyun Jun, *An Examination on Income Survey Microdata: 〈Household Income and Expenditure Survey〉 and 〈Survey of Household Finances and Living Conditions〉*, Research Monograph 2019-13, Korea Institute for Health and Social Affairs, 2019.

10. 이혜정, 『조사 자료의 측정오차 보정 및 관리 방안: 근로소득을 중심으로』, 보건복지포럼 2022년 11월 통권 제313호, 한국보건사회연구원, 2022, pp. 59-74.

(Translated in English) Lee, Hyejung, *Measurement Error Correction and Management in Survey Data - Focusing on Earned Income*, Health and Welfare Forum 2022(11)-313, Korea Institute for Health and Social Affairs, 2022, pp. 59-74.

11. 정예은 · 홍지훈, "직무별 재택근무 가능성에 따른 코로나 19의 고용효과 분석," 『시장경제연구』, 제52권 제2호, 2023, pp. 81-112.

(Translated in English) Jeong, Yeeun, and Gihoon Hong, "An Empirical Analysis of the Employment Effects of COVID-19 by Teleworkability," *Journal of Market Economy*, Vol. 52, No. 2, 2023, pp. 81-112.

12. 정재현, 『소득증빙서류와 조사자료 간 교차검증: 재정패널조사를 중심으로』, 조세재정브리프 2020년 12월 통권 제106호, 한국조세재정연구원, 2020.

(Translated in English) Jung, Jaehyun, *Validation Study of Administrative Income and Survey Data: Evidence from the National Survey of Tax and Benefit*, KIPF ISSUE PAPER No. 106, Korea Institute of Public Finance, 2020.

13. 통계청, 『지역별 고용조사 2023년 정기통계품질진단 결과보고서』, 2023, pp. 15-16.

(Translated in English) Statistics Korea, *Regional Employment Survey 2023 Regular Assessment Report*, 2023, pp. 15-16.

14. _____, 『지역별 고용조사 통계정보보고서』, 2024, pp. 61-63.

(Translated in English) Statistics Korea, *Regional Employment Survey Statistical Information Report*, 2024, pp. 61-63.

15. 홍민기, "보정 지니계수," 『경제발전연구』, 제23권 제3호, 2017, pp. 1-22.

(Translated in English) Hong, Minki, "Corrected Gini Coefficient in Korea," *Journal of Korean Economic Development*, Vol. 23, No. 3, 2017, pp. 1-22.

16. Abowd, J. M., and M. H. Stinson, "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," *Review of Economics and Statistics*, Vol. 95, No. 5, 2013, pp. 1451-1467.

17. An, Y., and Y. Hu, "Well-posedness of Measurement Error Models for Self-reported Data," *Journal of Econometrics*, Vol. 168, No. 2, 2012, pp. 259-269.

18. Bollinger, C. R., "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, Vol. 16, No. 3, 1998, pp. 576-594.

19. _____, "Response Error and the Union Wage Differential," *Southern Economic*

Journal, Vol. 68, No. 1, 2001, pp. 60-76.

20. _____, "Measurement Error in Human Capital and the Black-white Wage Gap," *Review of Economics and Statistics*, Vol. 85, No. 3, 2003, pp. 578-585.
21. Bollinger, C. R., and M. H. David, "I didn't Tell, and I won't Tell: Dynamic Response Error in the SIPP," *Journal of Applied Econometrics*, Vol. 20, No. 4, 2005, pp. 563-569.
22. Bound, J., and A. B. Krueger, "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics*, Vol. 9, No. 1, 1991, pp. 1-24.
23. Bound, J., C. Brown, and N. Mathiowetz, *Measurement Error in Survey Data*, In Handbook of Econometrics, Vol. 5, 2001, pp. 3705-3843.
24. Bound, J., C. Brown, G. J. Duncan, and W. L. Rodgers, "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, Vol. 12, No. 3, 1994, pp. 345-368.
25. Celhay, P., B. D. Meyer, and N. Mittag, "What Leads to Measurement Errors? Evidence from Reports of Program Participation in Three Surveys," *Journal of Econometrics*, Vol. 238, No. 2, 2024, 105581.
26. Chen, X., H. Hong, and D. Nekipelov, "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, Vol. 49, No. 4, 2011, pp. 901-937.
27. Chen, X., H. Hong, and E. Tamer, "Measurement Error Models with Auxiliary Data," *The Review of Economic Studies*, Vol. 72, No. 2, 2005, pp. 343-366.
28. Deaton, A., "Panel Data from Time Series of Cross-sections," *Journal of Econometrics*, Vol. 30, No. 1-2, 1985, pp. 109-126.
29. Duncan, G. J., and D. H. Hill, "An Investigation of the Extent and Consequences of Measurement Error in Labor-economic Survey Data," *Journal of Labor Economics*, Vol. 3, No. 4, 1985, pp. 508-532.
30. Feng, S., and Y. Hu, "Misclassification Errors and the Underestimation of the US Unemployment Rate," *American Economic Review*, Vol. 103, No. 2, 2013, pp. 1054-1070.
31. Gibson, J., and B. Kim, "The Effect of Reporting Errors on the Cross-country Relation between Inequality and Crime," *Journal of Development Economics*, Vol. 87, No. 2, 2008, pp. 247-254.
32. Guillerm, M., "Pseudo Panel Methods and an Example of Application to Household Wealth Data," *Economie et Statistique*, Vol. 491, No. 1, 2017, pp. 109-130.
33. Hahn, J., and G. Ridder, "Instrumental Variable Estimation of Nonlinear Models with Nonclassical Measurement Error Using Control Variates," *Journal of Econometrics*, Vol. 200, No. 2, 2017, pp. 238-250.
34. Hansen, B. E., *Econometrics*, Unpublished Manuscript, 2020, p. 348.
35. Hu, Y., "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, Vol. 144, No. 1, 2008, pp. 27-61.
36. Hu, Y., and Y. Sasaki, "Closed-form Estimation of Nonparametric Models with

Non-classical Measurement Errors," *Journal of Econometrics*, Vol. 185, No. 2, 2015, pp. 392-408.

37. Hu, Y., and S. M. Schennach, "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, Vol. 76, No. 1, 2008, pp. 195-216.
38. Hu, Y., and T. Wansbeek, "Measurement Error Models: Editors' Introduction," *Journal of Econometrics*, Vol. 200, No. 2, 2017, pp. 151-153.
39. Hyslop, D. R., and W. Townsend, "Earnings Dynamics and Measurement Error in Matched Survey and Administrative Data," *Journal of Business & Economic Statistics*, Vol. 38, No. 2, 2020, pp. 457-469.
40. Jenkins, S. P., and F. Rios-Avila, "Modelling Errors in Survey and Administrative Data on Employment Earnings: Sensitivity to the Fraction Assumed to have Error-free Earnings," *Economics Letters*, Vol. 192, 2020, 109253.
41. _____, "Reconciling Reports: Modelling Employment Earnings and Measurement Errors Using Linked Survey and Administrative Data," *Journal of the Royal Statistical Society Series A: Statistics in Society*, Vol. 186, No. 1, 2023, pp. 110-136.
42. Kapteyn, A. and J. Y. Ypma, "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data," *Journal of Labor Economics*, Vol. 25, No. 3, 2007, pp. 513-551.
43. Lee, J., and S. Lee, "Does it Matter Who Responded to the Survey? Trends in the US Gender Earnings Gap Revisited," *ILR Review*, Vol. 65, No. 1, 2012, pp. 148-160.
44. Lewbel, A., "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, Vol. 75, No. 2, 2007, pp. 537-551.
45. Mathiowetz, N., C. Brown, and J. Bound, "Measurement Error in Surveys of the Low-income Population," *Studies of Welfare Populations: Data Collection and Research Issues*, 2002, pp. 157-194.
46. Meyer, B. D., and N. Mittag, "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net," *American Economic Journal: Applied Economics*, Vol. 11, No. 2, 2019, pp. 176-204.
47. Meyer, B. D., N. Mittag, and R. M. Goerge, "Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation," *Journal of Human Resources*, Vol. 57, No. 5, 2022, pp. 1605-1644.
48. Moffitt, R., "Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-sections," *Journal of Econometrics*, Vol. 59, No. 1-2, 1993, pp. 99-123.
49. Moore, J. C., L. L. Stinson, and E. J. Welniak, "Income Measurement Error in Surveys: A Review," *Journal of Official Statistics-Stockholm*, Vol. 16, No. 4, 2000, pp. 331-362.
50. Pischke, J. S., "Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study," *Journal of Business & Economic Statistics*, Vol. 13, No. 3, 1995, pp. 305-314.

51. Pischke, S., *Lecture Notes on Measurement Error*, London School of Economics, London. 2007.

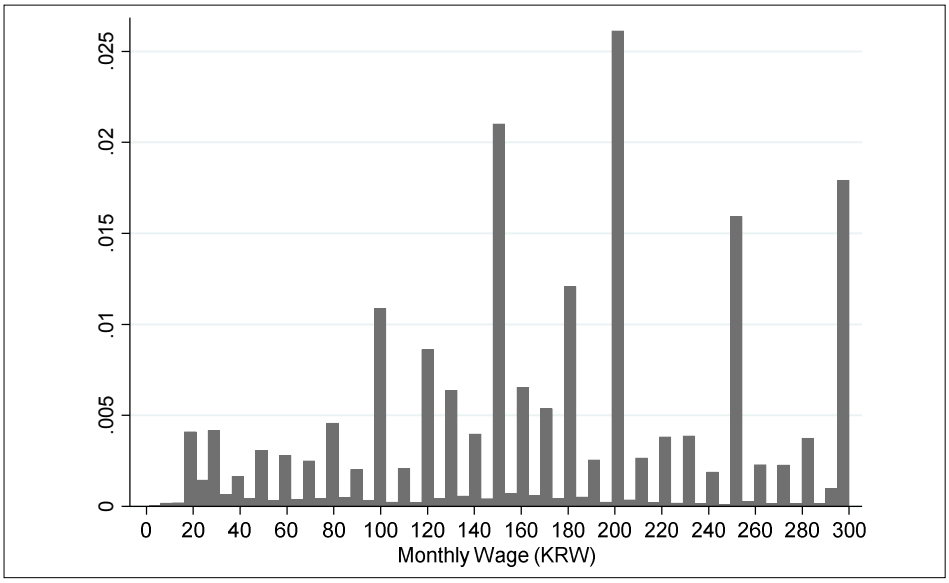
52. Tourangeau, R., L. J. Rips, and K. Rasinski, *The Psychology of Survey Response*, Cambridge University Press, Cambridge. 2000.

53. Verbeek, M., In: Mátyás, L. and P. Sevestre, (eds) The Econom*Pseudo-Panels and Repeated Cross-Sections*, etrics of Panel Data, Advanced Studies in Theoretical and Applied Econometrics, Vol. 46, 2008.

54. Zinn, S., and A. Wurbach, "A Statistical Approach to Address the Problem of Heaping in Self-reported Income Data," *Journal of Applied Statistics*, 2016, Vol. 43, No. 4, p. 682.

〈 부 록: 그림과 표 〉

〈Figure A.1〉 Distribution of RES monthly income (unit: KRW 10,000)



Source: RES (2015~2018).

Note: The distribution illustrates individuals whose monthly income falls below KRW 3 million.

〈Table A.1〉 Summary of cross-sectional statistics (2016, 2017, 2018)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
		2016			2017			2018		
	obs	RES	NTS	Error	RES	NTS	Error	RES	NTS	Error
(A) All										
	576	10.14 (0.51)	10.11 (0.78)	0.03 (0.31)	10.15 (0.50)	10.14 (0.76)	0.01 (0.30)	10.19 (0.49)	10.18 (0.74)	0.01 (0.29)
(B) By sex										
Male	288	10.35 (0.42)	10.33 (0.73)	0.03 (0.32)	10.37 (0.41)	10.37 (0.71)	0.00 (0.32)	10.40 (0.40)	10.40 (0.69)	0.00 (0.30)
Female	288	9.83 (0.46)	9.80 (0.73)	0.04 (0.29)	9.85 (0.46)	9.83 (0.72)	0.02 (0.27)	9.91 (0.46)	9.89 (0.70)	0.02 (0.26)
(C) By percentiles										
90	32	10.78 (0.27)	11.11 (0.27)	**-0.33** **(0.04)**	10.77 (0.26)	11.11 (0.27)	**-0.34** **(0.04)**	10.80 (0.26)	11.12 (0.26)	**-0.32** **(0.04)**
80	32	10.51 (0.29)	10.77 (0.31)	**-0.26** **(0.04)**	10.53 (0.27)	10.79 (0.31)	**-0.26** **(0.06)**	10.56 (0.26)	10.80 (0.30)	**-0.23** **(0.06)**
70	32	10.36 (0.28)	10.52 (0.31)	**-0.16** **(0.05)**	10.35 (0.26)	10.54 (0.31)	**-0.19** **(0.06)**	10.39 (0.25)	10.56 (0.29)	**-0.17** **(0.06)**
60	32	10.23 (0.27)	10.32 (0.30)	-0.08 (0.05)	10.24 (0.25)	10.34 (0.29)	**-0.10** **(0.05)**	10.26 (0.22)	10.37 (0.28)	-0.11 (0.07)
50	32	10.09 (0.25)	10.12 (0.27)	-0.02 (0.04)	10.12 (0.25)	10.15 (0.27)	-0.03 (0.04)	10.18 (0.24)	10.19 (0.25)	-0.01 (0.05)
40	32	9.98 (0.24)	9.91 (0.25)	0.07 (0.05)	10.01 (0.24)	9.95 (0.25)	0.06 (0.04)	10.05 (10.01)	0.22 (0.22)	0.04 (0.04)
30	32	9.85 (0.27)	9.64 (0.23)	**0.21** **(0.08)**	9.87 (0.24)	9.70 (0.24)	**0.17** **(0.06)**	9.91 (0.20)	9.76 (0.22)	**0.15** **(0.07)**
20	32	9.65 (0.28)	9.27 (0.26)	**0.38** **(0.06)**	9.66 (0.30)	9.33 (0.28)	**0.33** **(0.05)**	9.72 (0.31)	9.40 (0.27)	**0.32** **(0.09)**
10	32	9.30 (0.37)	8.53 (0.31)	**0.77** **(0.12)**	9.32 (0.43)	8.57 (0.31)	**0.75** **(0.16)**	9.36 (0.46)	8.65 (0.31)	**0.71** **(0.19)**

Notes: Measurement errors are calculated by subtracting the NTS data from the RES data in each percentile. We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data. The standard errors are in parentheses. Measurement errors in the shaded areas are statistically significant at the 5% level.

〈Table A.2〉 Summary of longitudinal statistics (2016, 2017, 2018)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		2016~2017			2017~2018		
	obs	RES	NTS	Error	RES	NTS	Error
(A) All							
	576	0.02 (0.05)	0.04 (0.03)	-0.02 (0.05)	0.04 (0.05)	0.04 (0.03)	0.00 (0.06)
(B) By sex							
Male	288	0.01 (0.04)	0.04 (0.03)	-0.03 (0.05)	0.03 (0.05)	0.03 (0.03)	0.00 (0.05)
Female	288	0.02 (0.05)	0.04 (0.02)	-0.01 (0.05)	0.06 (0.06)	0.06 (0.03)	0.00 (0.06)
(C) By percentiles							
90	32	0.00 (0.03)	0.01 (0.01)	-0.02 (0.03)	0.04 (0.04)	0.01 (0.01)	0.03 (0.04)
80	32	0.02 (0.04)	0.02 (0.01)	0.00 (0.04)	0.04 (0.04)	0.01 (0.01)	0.03 (0.05)
70	32	-0.01 (0.04)	0.02 (0.01)	-0.03 (0.03)	0.04 (0.05)	0.02 (0.02)	0.02 (0.05)
60	32	0.01 (0.04)	0.03 (0.01)	-0.02 (0.04)	0.03 (0.04)	0.03 (0.02)	0.00 (0.04)
50	32	0.03 (0.04)	0.04 (0.01)	-0.01 (0.03)	0.06 (0.05)	0.04 (0.02)	0.02 (0.04)
40	32	0.04 (0.04)	0.05 (0.01)	-0.01 (0.04)	0.04 (0.05)	0.06 (0.03)	-0.02 (0.05)
30	32	0.02 (0.06)	0.06 (0.02)	-0.04 (0.07)	0.04 (0.05)	0.06 (0.02)	-0.02 (0.05)
20	32	0.02 (0.04)	0.07 (0.03)	-0.05 (0.04)	0.06 (0.06)	0.07 (0.02)	-0.01 (0.06)
10	32	0.02 (0.08)	0.04 (0.03)	-0.02 (0.09)	0.05 (0.08)	0.09 (0.03)	-0.04 (0.09)

Notes: Measurement errors are calculated by subtracting the NTS data from the RES data in each percentile. We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data. The standard errors are in parentheses. Measurement errors in the shaded areas are statistically significant at the 5% level.

〈Table A.3〉 Summary of cross-sectional statistics by sex and percentiles (2016, 2017, 2018)

(A) 2016

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Male			Female		
	obs	RES	NTS	Error	RES	NTS	Error
90	16	10.98 (0.10)	11.31 (0.10)	**-0.33** **(0.04)**	10.48 (0.10)	10.81 (0.10)	**-0.33** **(0.04)**
80	16	10.74 (0.09)	11.02 (0.09)	**-0.28** **(0.03)**	10.19 (0.11)	10.43 (0.10)	**-0.24** **(0.04)**
70	16	10.58 (0.09)	10.77 (0.08)	**-0.19** **(0.04)**	10.05 (0.08)	10.17 (0.08)	**-0.11** **(0.03)**
60	16	10.44 (0.07)	10.55 (0.07)	**-0.11** **(0.03)**	9.93 (0.11)	9.98 (0.06)	-0.05 (0.06)
50	16	10.29 (0.06)	10.34 (0.07)	-0.05 (0.03)	9.81 (0.06)	9.80 (0.05)	0.01 (0.03)
40	16	10.16 (0.08)	10.10 (0.07)	0.06 (0.04)	9.71 (0.08)	9.63 (0.03)	0.08 (0.06)
30	16	10.06 (0.06)	9.82 (0.08)	**0.24** **(0.05)**	9.54 (0.09)	9.38 (0.03)	0.16 (0.09)
20	16	9.87 (0.07)	9.47 (0.06)	**0.40** **(0.03)**	9.34 (0.08)	8.97 (0.04)	**0.37** **(0.08)**
10	16	9.59 (0.10)	8.78 (0.07)	**0.81** **(0.06)**	8.89 (0.18)	8.17 (0.04)	**0.72** **(0.16)**

(B) 2017

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Male			Female		
	obs	RES	NTS	Error	RES	NTS	Error
90	16	10.98 (0.08)	11.32 (0.10)	**-0.34** **(0.04)**	10.48 (0.09)	10.83 (0.10)	**-0.35** **(0.04)**
80	16	10.74 (0.08)	11.03 (0.09)	**-0.29** **(0.03)**	10.23 (0.12)	10.44 (0.10)	**-0.21** **(0.05)**
70	16	10.56 (0.08)	10.79 (0.08)	**-0.23** **(0.04)**	10.06 (0.06)	10.19 (0.08)	**-0.13** **(0.04)**
60	16	10.44 (0.05)	10.58 (0.07)	**-0.13** **(0.03)**	9.95 (0.07)	10.01 (0.06)	-0.06 (0.03)
50	16	10.32 (0.05)	10.37 (0.06)	-0.05 (0.03)	9.84 (0.07)	9.84 (0.05)	0.00 (0.03)
40	16	10.20 (0.09)	10.15 (0.07)	0.05 (0.04)	9.74 (0.05)	9.67 (0.03)	0.07 (0.04)
30	16	10.06 (0.04)	9.89 (0.08)	**0.17** **(0.07)**	9.60 (0.07)	9.43 (0.04)	**0.17** **(0.06)**
20	16	9.91 (0.06)	9.56 (0.07)	**0.35** **(0.03)**	9.33 (0.07)	9.02 (0.04)	**0.31** **(0.07)**
10	16	9.65 (0.12)	8.82 (0.08)	**0.83** **(0.09)**	8.85 (0.21)	8.21 (0.05)	**0.64** **(0.18)**

Notes: Measurement errors are calculated by subtracting the NTS data from the RES data in each percentile. We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data. The standard errors are in parentheses. Measurement errors in the shaded areas are statistically significant at the 5% level.

〈Table A.3〉 Continued

(C) 2018

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Male			Female		
	obs	RES	NTS	Error	RES	NTS	Error
90	16	11.01 (0.09)	11.33 (0.10)	**-0.32** **(0.03)**	10.53 (0.08)	10.84 (0.10)	**-0.32** **(0.06)**
80	16	10.77 (0.08)	11.04 (0.09)	**-0.27** **(0.04)**	10.28 (0.10)	10.47 (0.10)	**-0.19** **(0.04)**
70	16	10.60 (0.08)	10.80 (0.08)	**-0.21** **(0.04)**	10.12 (0.09)	10.23 (0.08)	**-0.12** **(0.03)**
60	16	10.45 (0.04)	10.59 (0.07)	**-0.15** **(0.05)**	10.01 (0.05)	10.06 (0.06)	-0.05 (0.05)
50	16	10.37 (0.08)	10.40 (0.06)	-0.03 (0.06)	9.92 (0.06)	9.91 (0.04)	0.01 (0.03)
40	16	10.23 (0.06)	10.19 (0.07)	0.05 (0.04)	9.80 (0.05)	9.77 (0.03)	0.03 (0.03)
30	16	10.06 (0.05)	9.94 (0.07)	**0.12** **(0.04)**	9.69 (0.09)	9.51 (0.03)	**0.18** **(0.08)**
20	16	9.97 (0.06)	9.63 (0.07)	**0.34** **(0.04)**	9.39 (0.12)	9.09 (0.04)	**0.29** **(0.12)**
10	16	9.72 (0.09)	8.90 (0.07)	**0.81** **(0.07)**	8.87 (0.24)	8.30 (0.05)	**0.57** **(0.22)**

Notes: Measurement errors are calculated by subtracting the NTS data from the RES data in each percentile. We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data. The standard errors are in parentheses. Measurement errors in the shaded areas are statistically significant at the 5% level.

〈Table A.4〉 Summary of longitudinal statistics by sex and percentiles (2016, 2017, 2018)

(A) 2016~2017

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Male			Female		
	obs	RES	NTS	Error	RES	NTS	Error
90	16	-0.01 (0.03)	0.01 (0.01)	-0.01 (0.03)	0.00 (0.03)	0.02 (0.02)	-0.02 (0.04)
80	16	0.00 (0.03)	0.01 (0.01)	-0.02 (0.03)	0.05 (0.04)	0.02 (0.01)	0.03 (0.04)
70	16	-0.02 (0.02)	0.02 (0.01)	-0.04 (0.02)	0.01 (0.04)	0.03 (0.01)	-0.02 (0.04)
60	16	0.00 (0.03)	0.02 (0.01)	-0.02 (0.03)	0.02 (0.05)	0.03 (0.01)	-0.01 (0.05)
50	16	0.03 (0.04)	0.03 (0.01)	-0.01 (0.03)	0.02 (0.03)	0.04 (0.01)	-0.01 (0.04)
40	16	0.04 (0.04)	0.05 (0.01)	-0.01 (0.03)	0.04 (0.05)	0.05 (0.01)	-0.01 (0.05)
30	16	0.00 (0.04)	0.07 (0.02)	-0.08 (0.05)	0.06 (0.06)	0.05 (0.02)	0.01 (0.06)
20	16	0.04 (0.04)	0.09 (0.02)	-0.05 (0.04)	-0.01 (0.04)	0.05 (0.02)	-0.05 (0.04)
10	16	0.06 (0.07)	0.04 (0.04)	0.03 (0.08)	-0.04 (0.07)	0.04 (0.03)	-0.08 (0.05)

(B) 2017~2018

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Male			Female		
	obs	RES	NTS	Error	RES	NTS	Error
90	16	0.03 (0.04)	0.01 (0.01)	0.03 (0.04)	0.04 (0.05)	0.01 (0.01)	0.03 (0.05)
80	16	0.03 (0.04)	0.01 (0.01)	0.03 (0.04)	0.05 (0.05)	0.03 (0.01)	0.02 (0.06)
70	16	0.03 (0.04)	0.01 (0.01)	0.02 (0.04)	0.06 (0.05)	0.04 (0.01)	0.01 (0.05)
60	16	0.00 (0.03)	0.02 (0.01)	-0.02 (0.03)	0.06 (0.04)	0.05 (0.01)	0.01 (0.04)
50	16	0.05 (0.05)	0.02 (0.01)	0.02 (0.05)	0.08 (0.04)	0.07 (0.01)	0.01 (0.04)
40	16	0.03 (0.05)	0.03 (0.01)	0.00 (0.04)	0.06 (0.04)	0.10 (0.01)	-0.04 (0.05)
30	16	0.00 (0.03)	0.05 (0.02)	-0.04 (0.04)	0.09 (0.04)	0.08 (0.02)	0.01 (0.04)
20	16	0.07 (0.04)	0.07 (0.01)	0.00 (0.04)	0.05 (0.09)	0.08 (0.02)	-0.02 (0.08)
10	16	0.06 (0.09)	0.08 (0.03)	-0.02 (0.09)	0.02 (0.07)	0.09 (0.03)	-0.07 (0.07)

Notes: Measurement errors are calculated by subtracting the NTS data from the RES data in each percentile. We computed the standard errors using weights based on the number of individuals in each year-gender-region cell from the NTS data. The standard errors are in parentheses. Measurement errors in the shaded areas are statistically significant at the 5% level.

〈Table A.5〉 Wage workers filing taxes

	(1) Wage workers filing taxes (people)	(2) Wage workers (Economically active population, people)	(3) Ratio of wage workers filing taxes (%, (A))	(4) Ratio of wage workers not filing taxes (%, 1-(A))
2015	17,333,394	19,474,000	89.0	11.0
2016	17,739,258	19,743,000	89.9	10.2
2017	18,004,452	20,006,000	90.0	10.0
2018	18,576,857	20,045,000	92.7	7.3

Sources: Statistical yearbook of National Tax (2015~2018), Economically Active Population Survey (2015~2018).

부 록 1. 편의 계산 과정

종속변수와 독립변수 각각에 평균회귀오차가 포함된 경우 추정치 계산과정을 제시한다. 편의를 위해 {i}를 생략하고 작성한다.

(1) 응답값(w^*)이 종속변수에 사용될 때의 추정치

(1.1) 횡단면 자료: $plim\widehat{(1+\delta_1)}\beta = \frac{cov(x,w^*)}{var(x)} = \frac{(1+\delta_1)var(x)}{var(x)}\beta$

$= (1+\delta_1)\beta$

(1.2) 종단면 자료: $plim\widehat{(1+\delta_1)}\beta = \frac{cov(\triangle x,\triangle w^*)}{var(\triangle x)} = \frac{(1+\delta_1)var(\triangle x)}{var(\triangle x)}\beta$

$= (1+\delta_1)\beta$

(2) 응답값(x^*)이 독립변수에 사용될 때의 추정치

(2.1) 횡단면 자료: $plim\frac{\widehat{\beta}}{1+\delta_2} = \frac{cov(x^*,w)}{var(x^*)} = \frac{cov((1+\delta_2)x+\eta_2, x\beta+\varepsilon)}{var((1+\delta_2)x+\eta_2)}$

(2.2) 종단면 자료: $plim\frac{\widehat{\beta}}{1+\delta_2} = \frac{cov(\triangle x^*,\triangle w)}{var(\triangle x^*)}$

$= \frac{cov((1+\delta_2)\triangle x+\triangle\eta_2,\ \triangle x\beta+\triangle\varepsilon)}{var((1+\delta_2)\triangle x+\triangle\eta_2)}$

$= \frac{(1+\delta_2)var(\triangle x)}{(1+\delta_2)^2var(\triangle x)+(\rho_2-1)^2var(\eta_2)+var(e_2)}\beta$

(3) 측정오차의 자기상관[27]

$cov(u,u_{-1}) = cov(\delta w+\eta, \delta w_{-1}+\eta_{-1}) = \delta var(w)+cov(\eta,\eta_{-1})$

$= \delta\sigma_w^2+\rho\sigma_\eta^2 \neq 0$

27) 측정오차가 존재하는 변수의 종류와 관계없이 동일한 계산으로 하첨자를 생략하였다.

부 록 2. 종속변수와 독립변수에 측정오차가 동시에 존재하는 경우

본 절에서는 식 (1)에서 종속변수와 독립변수에 측정오차가 동시에 존재하는 경우 추정치에 포함된 편의의 크기를 계산한다. 종속변수에 포함된 측정오차(u_1)는 참값(w)과 식 (4)와 같은 관계에 있으며 종속변수의 응답값(w^*)과 참값은 식 (5)와 같은 관계에 있다. 독립변수에 포함된 측정오차(u_2)는 참값(x)과 식 (9)와 같은 관계에 있으며 독립변수의 응답값(x^*)과 참값은 식 (10)과 같은 관계에 있다. 4절에서 제시한 고전적 측정오차와 평균회귀 측정오차에 대한 가정은 동일하다. 우선 횡단면 분석 방법을 활용한 경우 발생할 수 있는 편의의 크기를 계산한다. 종속변수와 독립변수에 평균으로 회귀하는 측정오차가 동시에 존재하게 되면 연구자들은 식 (A. 2. 1)을 추정하게 된다.

$$w_i^* = \frac{(1+\delta_1)}{(1+\delta_2)}\beta x_i^* - \frac{(1+\delta_1)}{(1+\delta_2)}\beta\eta_{2i} + (1+\delta_1)\varepsilon_i + \eta_{1i} \qquad \text{(A. 2. 1)}$$

편의의 크기는 〈Table A. 2. 1〉의 (1)열과 (3)열에 제시한다. 고전적 측정오차가 두 변수에 모두 포함되었을 때 $\frac{\sigma_{u_2}^2}{\sigma_x^2+\sigma_{u_2}^2}$ 만큼의 편의가 발생하며, 이는 〈Table 1〉 (나)의 (1)열에 제시된 독립변수에만 측정오차가 포함된 경우 발생하는 편의의 크기와 동일하다. 고전적 측정오차가 종속변수에만 포함된 경우 발생한 편의의 크기는 0이기 때문일 것이다(〈Table 1〉 (가)의 (1)열). 평균회귀 측정오차가 두 변수에 모두 포함되었을 때 편의의 크기는 $\frac{(1+\delta_2)(1+\delta_1)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2}=\gamma_c$에 따라 다르게 계산된다. $0<\gamma_c<1$이면 $1-\frac{(1+\delta_2)(1+\delta_1)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2}=1-\gamma_c$이며, $\gamma_c>1$이면 $\frac{(1+\delta_2)(1+\delta_1)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2}-1=\gamma_c-1$이다. $\delta_2=0$이고 $\eta_2=0$라면, 평균회귀 측정오차가 종속변수에만 포함되어 있을 때 발생하는 편의의 크기$(-\delta_1)$와 동일하고(〈Table 1〉(가)의 (3)열), $\delta_1=0$이고 $\eta_1=0$라면, 평균회귀 측정오차가 독립변수에만 포함되어 있을 때 발생하는 편의의 크기$(1-\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2}$ 또는 $\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+\sigma_{\eta_2}^2}-1)$

와 동일하게 계산된다(〈Table 1〉 (나)의 (3)열).

둘째로 종단면 분석 방법을 활용한 경우 발생할 수 있는 편의의 크기를 계산한다. 본문에서와 같게 식 (A. 2. 1)에서 수준(level) 변수 대신 차분(△) 형태의 변수를 활용하여 계산하고 이를 다시 쓰면 식 (A. 2. 2)과 같다.

$$\triangle w_i^* = \frac{(1+\delta_1)}{(1+\delta_2)}\beta\triangle x_i^* - \frac{(1+\delta_1)}{(1+\delta_2)}\beta\triangle\eta_{2i} + (1+\delta_1)\triangle\varepsilon_i + \triangle\eta_{1i} \qquad \text{(A. 2. 2)}$$

편의의 크기는 〈Table A. 2. 1.〉의 (2)열과 (4)열에 제시한다. 고전적 측정오차가 두 변수에 모두 포함되었을 때 $\frac{(1-\rho_2)\sigma_{u_2}^2}{\sigma_x^2+(1-\rho_2)\sigma_{u_2}^2}$ 만큼의 편의가 추정되며, 이는 〈Table 1〉의 (나)의 (2)열에 제시된 독립변수에만 측정오차가 포함된 경우 발생하는 편의의 크기와 동일하다. 이 역시 고전적 측정오차가 종속변수에만 포함된 경우 발생한 편의의 크기가 0이기 때문일 것이다(〈Table 1〉 (가)의 (2)열). 평균회귀 측정오차가 두 변수에 모두 포함되었을 때 편의의 크기는 $\frac{(1+\delta_2)(1+\delta_1)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}=\gamma_p$ 에 따라 다르게 계산된다. $0<\gamma_p<1$이면 $1-\frac{(1+\delta_2)(1+\delta_1)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}=1-\gamma_p$이며, $\gamma_c>1$이면 $\frac{(1+\delta_2)(1+\delta_1)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}-1=\gamma_p-1$이다. $\delta_2=0$이고 $\eta_2=0$라면, 평균회귀 측정오차가 종속변수에만 포함되어 있을 때 발생하는 편의의 크기$(-\delta_1)$와 동일하고(〈Table 1〉 (가)의 (4)열), $\delta_1=0$이고 $\eta_1=0$라면, 평균회귀 측정오차가 독립변수에만 포함되어 있을 때 발생하는 편의의 크기$(1-\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}$ 또는 $\frac{(1+\delta_2)\sigma_x^2}{(1+\delta_2)^2\sigma_x^2+(1-\rho_2)\sigma_{\eta_2}^2}-1)$와 동일하게 계산된다(〈Table 1〉 (나)의 (4)열).

〈Table A.2.1〉 Summary bias

Classical Measurement error $(\delta_1=\delta_2=0)$		Mean-reverting Measurement error $(-1<\delta_1,\delta_2<0)$	
(1)	(2)	(3)	(4)
Cross-sectional $(\rho_1=\rho_2=0)$	Longitudinal $(0<\rho_1,\rho_2\leq 1)$	Cross-sectional $(\rho_1=\rho_2=0)$	Longitudinal $(0<\rho_1,\rho_2\leq 1)$
$\dfrac{\sigma^2_{u_2}}{\sigma^2_x+\sigma^2_{u_2}}$	$\dfrac{(1-\rho_2)\sigma^2_{u_2}}{\sigma^2_x+(1-\rho_2)\sigma^2_{u_2}}$	(가) $0<\gamma_c<1$	(가) $0<\gamma_p<1$
		$1-\dfrac{(1+\delta_2)(1+\delta_1)\sigma^2_x}{(1+\delta_2)^2\sigma^2_x+\sigma^2_{\eta_2}}$	$1-\dfrac{(1+\delta_2)(1+\delta_1)\sigma^2_x}{(1+\delta_2)^2\sigma^2_x+(1-\rho_2)\sigma^2_{\eta_2}}$
		(나) $\gamma_c>1$	(나) $\gamma_p>1$
		$\dfrac{(1+\delta_2)(1+\delta_1)\sigma^2_x}{(1+\delta_2)^2\sigma^2_x+\sigma^2_{\eta_2}}-1$	$\dfrac{(1+\delta_2)(1+\delta_1)\sigma^2_x}{(1+\delta_2)^2\sigma^2_x+(1-\rho_2)\sigma^2_{\eta_2}}-1$

Note: Author summarizes the bias as discussed in Appendix 2.

부 록 3. 모델 적합성과 t값 계산과정

4.2절에서 제시한 회귀분석 모델의 적합성과 t값의 계산과정을 제시한다. 경우의 수가 많아 대표적으로 평균회귀 측정오차가 종속변수와 독립변수에 포함되었을 때를 제시한다. 작성의 편의를 위해 하첨자 {i}는 생략한다. 횡단면 분석의 경우를 제시하며, 종단면의 경우는 level 대신 차분(△)을 활용하여 계산하면 〈Table 2〉에 제시한 횡단면 분석 결과와 같은 결론에 도달할 수 있다.

(1) 종속변수에 평균회귀 측정오차가 포함된 경우

(1.1) 모델의 적합도

4.1절에서의 식 (6) : $w^* = (1+\delta_1)\beta x + (1+\delta_1)\varepsilon + \eta_1$ 중 $r := (1+\delta_1)\varepsilon + \eta_1$ 라고 정의한다. 추정한 결과가 $w^* = (1+\delta_1)x\hat{\beta} + \hat{r}$ 일 때, 잔차 분산은 $Plim\,\widehat{\sigma_{\hat{r}}^2} = (1+\delta_1)^2\sigma_\varepsilon^2 + \sigma_{\eta_1}^2 + \delta^2\beta^2(1+\delta_1)^2\sigma_x^2$ 으로 확률 수렴한다. 첫 번째 항은 σ_ε^2 가 $(1+\delta_1)^2$에 비례하여 0과 1사의 값을 지니며, 나머지 두 항은 양수이기 때문에 σ_ε^2과의 비교는 불가능하다. 따라서 회귀분석 모델의 적합도 변화는 알 수 없다.

(1.2) $\hat{\beta}$의 분산과 t값

$\hat{s} := \dfrac{\widehat{\sigma_{\hat{r}}^2}}{\widehat{\sigma_x^2}}$ 와 $s := \dfrac{\sigma_\varepsilon^2}{\sigma^{2_x}}$ 라고 정의하면 $\widehat{\beta_x}$의 분산은 $Plim\,\hat{s} = (1+\delta_1)^2 s + \dfrac{\sigma_{\eta_1}^2}{\sigma_x^2} + \delta_1^2\beta^2(1+\delta_1)^2$ 으로 확률 수렴한다. 첫 번째 항은 s가 $(1+\delta_1)^2$에 비례하여 0과 1사의 값을 지니며, 나머지 두 항은 양수이기 때문에 s와 비교할 수 없다. $\hat{\beta}$의 t값은 $\dfrac{Plim\,t}{\sqrt{n}} = \dfrac{Plim\,\hat{\beta}}{Plim\,\sqrt{\hat{s}}} = \dfrac{\beta}{\sqrt{s + \delta^2\beta^2 + \dfrac{\sigma_{\eta_1}^2}{(1+\delta)^2\sigma_x^2}}}$ 로 확률 수렴하는 것으로 계산되며, 이는 참값을 활용해 계산된 추정치의 t값이 확률 수렴한 값인 $\dfrac{\beta}{\sqrt{s}}$ 보다 작다. 따라서 종속변수에 평균회귀 측정오차가 포함되면 t값은 작아지며 보수적으로 추정하게 된다.

(2) 독립변수에 평균회귀 측정오차가 포함된 경우

(2.1) 모델의 적합도

4.1절에서의 식 (12) : $w=\beta\frac{x^{*}}{1+\delta_2}-\beta\frac{\eta_2}{1+\delta_2}+\varepsilon$중 $r:=-\beta\frac{\eta_2}{1+\delta_2}+\varepsilon$라고 정의한다. 추정한 결과가 $w=\hat{\beta}\frac{x^{*}}{1+\delta_2}+\hat{r}$ 일 때, 잔차 분산은 $Plim\,\widehat{\sigma_{\hat{r}}^2}=(1-\lambda_c)^2\beta^2\sigma_x^2+\frac{\lambda^{2_c}}{(1+\delta_2)^2}\beta^2\sigma_{\eta_2}^2+\sigma_\varepsilon^2$으로 확률 수렴한다. 이는 σ_ε^2보다 큰 값으로 회귀분석 모델의 적합도는 감소한다.

(2.2) $\hat{\beta}$의 분산과 t값

$\hat{s}:=\frac{\widehat{\sigma_{\hat{r}}^2}}{\widehat{\sigma_{x^*}^2}}$, $s:=\frac{\sigma_\varepsilon^2}{\sigma^{2_x}}$와 $c:=1/(1+\delta_2)^2+\frac{\sigma_{\eta_2}^2}{\sigma_x^2}$로 정의하면 $\hat{\beta}$의 분산은 $Plim\,\widehat{\sigma_{\hat{r}}^2}=c((1-\lambda_c)^2\beta^2+\frac{\lambda_c^2}{(1+\delta_2)^2}\beta^2\frac{\sigma_{\eta_2}^2}{\sigma_x^2}+s)$으로 확률 수렴하는 것으로 계산된다. c가 σ_x^2와 $\sigma_{\eta_2}^2$의 값에 따라 0과 1사이의 값을 갖거나 1보다 큰 값일 수 있어, s와 비교할 수 없다. $\hat{\beta}$의 t값은 $\frac{Plim\,t}{\sqrt{n}}=\frac{Plim\,\hat{\beta}}{Plim\,\sqrt{\hat{s}}}=\frac{\lambda_c\beta}{\sqrt{c((1-\lambda_c)^2\beta^2+\frac{\lambda_c^2}{(1+\delta)^2}\beta^2\frac{\sigma_{\eta_2}^2}{\sigma_x^2}+s}}$로 확률 수렴하는 것으로 계산된다. λ_c와 c의 크기가 정해지지 않아 참값을 활용해 계산된 추정치의 t값이 확률 수렴한 값인 $\frac{\beta}{\sqrt{s}}$와 크기를 비교할 수 없다.

부 록 4. 응답 방식과 측정오차

본 절에서는 히핑오차(heaping error)가 지역별고용조사의 측정오차를 얼마나 설명할 수 있는지 확인한다. 기존에 참값이었던 국세청의 연 근로소득 자료를 12로 나누어 월 평균근로소득 자료로 만들었다. 저자는 가상(pseudo)의 응답값을 생성하였는데 이는 국세청의 월 평균근로소득 자료를 다음의 세 가지 규칙에 따라 만들었다. 첫 번째 규칙은 만에서 버림, 두 번째 규칙은 5만 원, 5십만 원 또는 5백만 원 단위로 반올림, 세 번째 규칙은 만 원, 십만 원, 백만 원 단위로 반올림하는 것이다. 예컨대, 월평균 소득의 참값이 1,393,000원인 경우 저자는 첫 번째 규칙에 따라 1,300,000원으로, 두 번째 규칙에 따라 1,500,000원으로, 세 번째 규칙에 따라 1,400,000원으로 응답한 것이다. 가상의 응답오차는 참값인 국세청 월 평균 근로소득자료에서 세 가지 규칙에 따라 만든 가상의 응답값을 뺀 값이며, 이것이 가상의 히핑오차이다.

가상의 응답오차와의 비교를 위해 지역별고용조사의 측정오차도 12로 나누어 월 소득 수준으로 단위를 조정하였다. 세 가지 가상의 응답오차와 측정오차의 분포는 〈Figure A.4.1〉에 제시한다. 지역별고용조사의 측정오차는 0에 집중되어 있으며 대부분의 가상 응답오차는 넓게 분포한다.

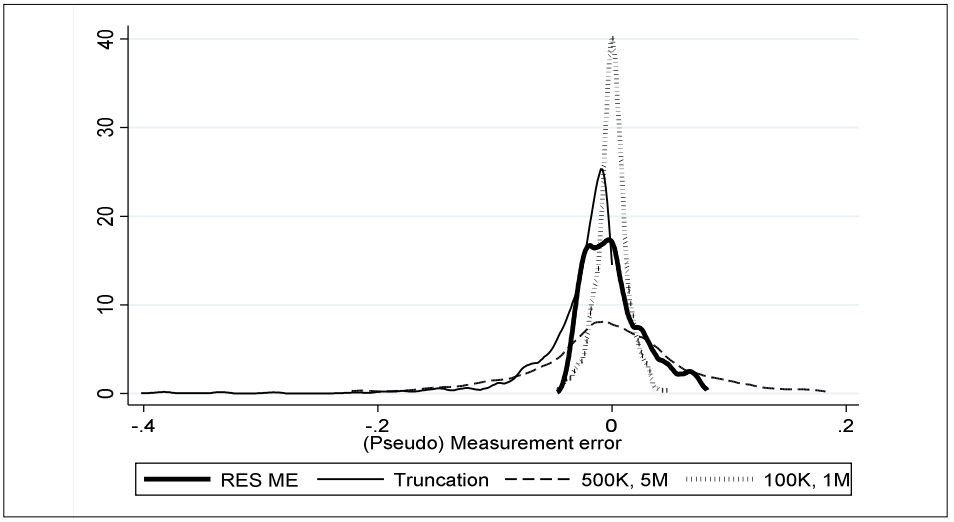
가상의 히핑오차가 지역별고용조사의 측정오차를 얼마나 설명하는지 확인하기 위해 지역별고용조사의 측정오차를 가상의 히핑오차에 회귀분석 하였다. 히핑오차가 지역별고용조사의 측정오차를 모두 설명하면 계수값은 1과 가까운 값이 추정될 것이고 일부만 설명하면 0과 1사이의 값을, 설명하지 않는다면 0일 것이다. 연도더미를 통제하였고 국세청 자료의 집단 내에 속한 사람 수를 가중치로 활용하였다.

계수는 세 가지 가상 응답오차 순서대로 각각 -0.326, -0.015, 0.016로 추정되었다. 각각의 계수 값이 1과 같다는 가설과 0과 같다는 가설을 기각하였다.[28] 다시 말해, 세 가지 규칙으로 응답하는 방식 모두가 지역별고용조사의 측정오차를 일부 설명할 수 있다.

---

28) 각각의 계수 값이 1과 같다는 가설의 세 추정치의 p-value는 모두 0으로 계산되었다.

〈Figure A.4.1〉 Distribution of pseudo-response error and measurement error



Notes: The thick solid line (RES ME) refers to the monthly-adjusted RES measurement error, obtained by dividing the annual error by 12. The thin solid line refers to the response errors generated following the first rule, which truncates amounts at the ten-thousand-won (Truncation). The dotted line indicates the ones following the second rule, which rounds values to KRW 50,000, 500,000, or 5 million (500K, 5M). The thin dotted line represents the ones following the third rule, which rounds values to KRW 10,000, 100,000, or 10 million (100K, 1M). The pseudo-heaping errors are calculated by subtracting the pseudo-response values generated under each of the three rules from the NTS monthly income data.

Measurement Errors of Income Data in the Regional Employment Survey*

Jiyeong Lee**

Abstract

This study explores the measurement errors embedded in the income data of the Regional Employment Survey (RES) using earnings data of National Tax Services (NTS) by region, sex and percentiles. This study mainly presents three findings. First, the surveyed income data have mean-reverting measurement errors. Second, the errors embedded in the dependent variable of the surveyed income data attenuate the coefficients in bi-variate regression model. Third, they are serially auto-correlated. The errors are mostly eliminated by a first-difference ordinary least square model. The findings suggest the need to confirm a conventional assumption of no error or classical measurement error made in empirical studies.

Key Words: Regional Employment Survey, National Tax Services, Mean-reverting measurement error, Attenuation bias
JEL Classification: J0, J3

Received: Sept. 4, 2024. Revised: May 6, 2025. Accepted: Aug. 15, 2025.

* This paper is a revised and supplemented version of a manuscript originally written during the author's affiliation with the Department of Economics at Seoul National University. The views expressed in this paper are solely those of the author and do not represent the official stance of the Bank of Korea. When citing this paper, please be sure to attribute it to the author. I extend sincere gratitude to two anonymous reviewers, editor and particularly professor Jungmin Lee for his helpful comments and advice. All remaining errors are mine.

** Senior Economist, Bank of Korea Gangwon Branch, 31, Jungang-ro, Chuncheon-si, Gangwon-do 24346, Korea, Phone: +82-33-258-3285, e-mail: jyhhss@gmail.com