주 의

1. 이 보고서는 통계개발원에서 시행한 정책연구용역 사업의 연구결과 보고서입니다.

2. 이 보고서 내용을 발표 또는 인용할 때에는 반드시 통계개발원에서 시행한 정책연구용역사업의 결과임을 밝혀야 합니다.

3. 이 보고서에 대한 저작권 일체와 2차적 저작물 또는 편집저작물의 작성권은 통계개발원이 소유하며, 통계개발원은 정책상 필요시 보고서의 내용을 보완 또는 수정할 수 있습니다.

통계개발원 STATISTICAL RESEARCH INSTITUTE

통계개발원 정책연구용역

혼합 조사에서의 추정 방법 개발 최종결과보고서 Estimation with mixed-mode survey data

2012. 12.

통계개발원

통계개발원
정책연구용역

혼합 조사에서의 추정 방법 개발
최종결과보고서

Estimation with mixed-mode survey data

2012. 12.

통계개발원 STATISTICAL RESEARCH INSTITUTE

제 출 문

통 계 개 발 원 장 귀 하

본 보고서를 “혼합조사에서의 추정 방법 개발 연구용역” 과제의 최종연구 결과보고서로 제출합니다.

2012년 12월 12일

연세대학교 통계연구소장 박 상 언

연 구 진

연구책임자 박 상 언(연세대학교 응용통계학과 교수 겸 통계연구소장)

공동연구자 김 재 광(아이오와주립대학교 교수)

김 서 영(통계개발원 사무관)

연구보조원 박 라 나(통계개발원 주무관)

박 승 환(서울대학교 통계학과 박사과정)

이 주 연(연세대학교 응용통계학과 석사과정)

연구결과보고서 요약문

연구과제명	혼합 조사에서의 추정방법 개발		
중심단어	베이즈정리, EM 알고리즘, 회귀모형, 선택편향 보정, Tobit 모형		
연구기관	연세대학교 통계연구소	연구책임자	박 상 언
연구기간	2012 . 6 . 15 ~ 2012 . 12 . 12		

본 보고서는 "혼합 조사에서의 추정 방법 개발"의 최종결과보고서로 2012년 6월부터 2012년 12월까지 6개월여의 기간 동안 이루어진 추정방법 개발에 대한 내용을 담고 있다. 통계청에서 2011년 6월과 10월 두 차례에 걸쳐 전국 초중고 1,081개 학교의 학부모 약 45,501명(2011. 6월 기준)을 상대로 실시한 사교육비 조사 자료를 이용하여 혼합조사에서 발생할 수 있는 여러 문제들을 분석하였다.

본 연구에서는 혼합조사에서 발생하는 모드 효과를 random effect로 간주하고 이를 반영하여 보정하는 방법을 제안하고 사교육비 조사 자료에 적용하였다. 이를 위해 구조 모형과 측정 모형을 세우고 그 모형에서 모수를 추정한 후 베이즈 정리를 이용하여 예측을 하였는데, 모수 추정을 위해서는 몬테 카를로 계산을 사용하여 EM 알고리즘을 적용하였다. 사교육비 자료의 특성을 고려하여, 사교육 시간에 대한 구조 모형은 Tobit 모형을 사용하였고, 사교육비에 대해서는 절편이 없는 회귀 모형을 사용하였다. 측정 모형은 두 모드 효과를 나타내는 잠재 변수에 대해 정규분포를 따르는 것을 가정하였다. 이 모형을 바탕으로 잠재 변인에 대한 예측 모형은 베이즈 정리를 통해서 구현되는데 이로부터 imputation 값을 발생시키는 것은 Kim(2011)이 제안한 parametric fractional imputation 방법을 이용하였다. 2012년 조사의 경우 조사 모드가 랜덤으로 배정되지 않고 응답자에게 선택권이 주어졌다. 이런 경우에는 조사 모드 외에도 선택 편향이 발생하게 되는데 본 연구에서 제안한 방법을 사용하면 편향을 상당히 줄일 수 있는 것을 모의실험을 통해 확인하였다.

본 연구에서 제안한 방법론은 통계청의 다른 혼합 조사에서도 사용할 수 있다. 이를 위해서는 관심 변수에 대해 올바른 구조 모형을 세워야하는데 이를 위해서는 각 조사마다 해당 조사의 관심 변수에 대한 관련 지식을 반영하여 모형을 세워야 할 것이다.

Project Summary

Title of Project	Estimation with mixed-mode survey data		
Key Words	Bayes theorem, EM algorithm, Regression model, Selection bias, Tobit model		
Institute	Yonsei Institute of Statistical Science	**Project Leader**	Sang Un Park
Project Period	2012 . 6 . 15 - 2012 . 12 . 12		

This is the final report for the project "Estimation with multi-mode survey data" which lasted for 6 months beginning June to December, 2012. In this project, we analyze some issues that can occur in mixed-mode surveys using the survey data on private education expenses conducted by National Statistical Office in 2011.

We regard the mode effect existing in mixed-mode surveys as random effects and suggest a methodology that calibrates the effect. In order to do this, we design a structural error model and a measurement error model, and use Monte Carlo computation and apply the EM algorithm to estimate the parameters. We consider the Tobit model as the structural error model for the private education time and expense. Also, we assume that the measurement error model follows the normal distribution regarding the latent variables that measures the mode effects. To obtain the prediction values from the model, we use the parametric fractional imputation method suggested by Kim(2011). Unlike the survey conducted in 2011, the survey mode wasn't randomly allocated in the survey conducted in 2012. In such a case, a selection bias may exist as well as the mode effect. However, we confirm that the bias can be significantly reduced by using the suggested model.

Our methodology can be further applied to other mixed-mode surveys, but we need to design a correct structural model regarding the variables we're interested in. In order to do this, the background knowledge of the variables has to be considered when designing the statistical model.

<목 차>

<표 목차>

<그림 목차>

제1장. 서론

1절. 연구의 필요성

혼합조사(mixed mode survey)는 응답률을 높이고 응답자들의 편의를 위해 다양한 창구의 응답을 허용하는 조사로써 실제 서베이에서 많이 사용되고 있다. 특히 인터넷 사용 인구의 증가로 인하여 방문 조사의 대안으로 일부 국가통계에서도 인터넷 조사가 부분적으로 사용되고 있는 실정이므로 이러한 혼합조사에 대한 보다 체계적인 연구는 매우 중요하나 이에 대한 보다 깊이 있는 통계학적 연구는 부족한 실정이다. 이러한 혼합 조사에서는 종종 각각의 조사 모드에 따라 다른 결과를 얻게 되어서 이에 대한 분석이 어렵게 된다. (de Leeuw, 1992; Dillman, *et al.*, 1996). Dillman (2000)에 의하면 자기 기입식 조사와 조사원에 의해 조사되는 서베이가 다른 결과를 가져오는 이유로써 첫째로 조사원의 간섭으로 인한 응답 차이, 둘째로는 자기 응답은 비주얼에 의존하는 반면 조사원 조사는 듣는 것에 의존하게 되는 것에 의한 응답 차이, 셋째로는 조사 인센티브의 차이로 인한 무응답의 차이로써 자기 기입식은 일반적으로 더 높은 항목 무응답을 가지게 된다. 이러한 차이는 결국 social desirability의 차이, 설문 문항에 대한 이해도의 차이, 사지선다 문항에서 답을 고르는 경향의 차이 등을 가져오게 된다.

이러한 혼합 조사를 통계학적으로 분석하는 문제는 결국 두 조사의 보정(calibration) 문제로 생각해 볼 수 있다. 이러한 보정은 흔히 측정 오차 모형(measurement error model)에서 발생하는 연구 주제로써 두 개의 다른 조사가 있을 때 어떻게 이를 보정하여 일관된 값을 계산해 내느냐의 문제로 접근하게 된다. 본 연구에서는 이러한 문제에 대해 일종의 무응답 대체(imputation) 방법을 적용하는 방법을 제안하고자 한다.

2절. 연구의 내용 및 방법

인터넷 조사가 점차 활발히 진행되고 있는 상황에서 기존에 주로 사용되었던 종이 조사와 인터넷 조사를 사용하여 조사를 진행한 사교육비 조사는 대표적인 혼합조사의 예이다. 실제 혼합조사에서 발생할 수 있는 여러 문제들을 사교육비 조사 자료를 이용하여 연구, 분석하고자 한다.

이를 위하여 통계청에서 실시하는 사교육비 조사 자료를 이용하여 사례 분석하고자 한다.

사교육비 조사는 우리나라의 초중고 학생들의 사교육비 실태를 체계적으로 조사하여 공신력 있는 통계를 정기적으로 작성, 제공하는 것을 목적으로 하는 조사이다. 전국 초중고 1,081개 학교 (1,411학급)의 학부모 약 45,501명(2011.6월 기준)을 상대로 2011년 6월과 10월 두 차례에 걸쳐 조사가 진행되었다. 표본추출은 지역과 학교를 층화변수로 사용하여 층화집락추출 방법을 사용하였다. 사교육비 조사는 응답자 기입방식 조사이며 종이 조사와 인터넷 조사 두 가지 방식을 함께 사용한 혼합조사이다. 종이 조사는 표본학급의 담임교사가 학생 편으로 종이 조사표를 학부모에게 전달, 학부모는 조사표를 직접 작성하고 학생 편으로 제출하는 방식이다(조사대상 학교별로 담당 공무원이 조사표 배부 및 회수). 인터넷 조사는 표본학급의 담임교사가 학생 편으로 인터넷 조사 안내장을 학부모에게 전달, 학부모가 인터넷에 직접 응답하는 방식이다(조사대상 학급의 50%에 대해 인터넷 조사 실시).

제2장. 문헌연구

1절. 문헌연구 결과

1. Emanuela Sala and Peter Lynn (2007). The potential of a multi-mode data collection design to reduce non response bias. The case of survey of employers, Springer Science+Business Media B.V.

본 논문은 단순한 postal survey와 two-phase multi-mode design의 두 가지 방법으로 고용주들을 대상으로 실시된 설문조사를 바탕으로 응답률(response rate)를 비교하였다. Postal survey에서는 상당히 큰 bias가 발생한 반면 multi-mode 방법은 postal sample의 bias를 대체적으로 줄이는 것으로 나타났다. 또한 두 방법 중 어떤 방법이 bias를 줄이는 데 최소의 비용을 발생 시키는지에 대하여 비교하였다.

2. James Wagner, Jennifer Arrieta, Heidi Guyer, Mary Beth Ofstedal (2011). Does Sequence Matter in Multi-Mode Surveys: Results from an Experiment. University of Michigan, Survey Research Center

Multiple mode 방법은 설문의 응답률을 유지하면서 비용을 줄인다는 점에서 많은 관심을 받고 있다. 따라서 어떤 data collection mode를 사용해야 하는지와 mode를 순서대로 또는 동시에 사용해야 하는지에 대한 연구가 많이 이루어져왔다. 하지만, mode의 순서가 결과에 어떤 영향을 끼치는지에 대한 연구는 많이 이루어지지 않았다. 본 연구에서는 미국의 50대 이상을 대상으로 하는 설문에서 screening interview의 실행 순서를 달리하여 실험을 진행하였다. 한 집단에게는 우편 또는 face-to-face 인터뷰를 진행했으면 다른 집단에게는 그 반대의 순서로 진행한 후, 각 실

험의 비용과 응답률 등을 비교하였다.

3. Joseph W. Sakshug and Ting Yan and Roger Tourangeau (2010). Nonresponse error, measurement error, and mode of data collection: tradeoffs in a multi-mode survey of sensitive and non-sensitive items, Public Opinion Quarterly, Vol. 74, No. 5, 907-933

무응답(non-response)과 측정 오류(measurement error) 사이에 trade-off가 존재할 것이라는 가설은 존재하지만 이 관계를 증명한 연구는 많이 이루어지지 않았다. 본 연구에서는 동창회 설문자료를 바탕으로 이 변수들 간 관계를 알아보고자 한다. 설문은 응답자의 학력에 대해 묻는 민감한 질문이 포함되어 있으며 세 가지의 data collection 방법(CATI, IVR, internet survey)을 비교하였다. 이 연구에서 두 mode는 모두 컴퓨터로 진행되는 자기-기입식 설문인데, 이 방법이 측정 오류는 낮추지만 응답 오류를 높일 것이라는 가설을 검증해보았다.

4. Edith D. de Leeuw (2005), To Mix or Not to Mix Data Collection Modes in Surveys, Journal of Official Statistics, Vol. 21, No.2, 233-255

설문을 진행하는 방법이 다양해짐에 따라 어떤 조사 방법이 가장 효과적인지에 대한 관심이 많아졌다. 최근에는 data collection에서 multiple mode와 mixed-mode가 가장 많이 사용되는데, 본 논문에서는 mixed-mode survey 설계의 장단점을 비교 및 정리하여 소개한다.

5. Maria Signore and Giovanna Brancato (2008), Multi-mode Data Collection: What Can Still Be Expected?, Proceedings of Statistics Canada Symposium

본 연구에서는 전체적인 무응답에 대해 data collection mode와 다른 survey characteristic 간의 상호관계를 분석하여 어떤 multi-mode collection 방법이 응답률을 높이는지에 대해 알아본다.

6. Stepehen Ansolabehere and Brian F. Schaffner, Does Survey Mode Still Matter?: Findings From a 2010 Multi-Mode Comparison.
(http://people.umass.edu/schaffne/ansolabehere_schaffner_mode2.pdf)

본 연구에서는 2010년에 이루어진 three-mode survey comparison study의 자료를 바탕으로 설문에 사용된 각 mode들의 결과를 비교했다. 설문은 전국적으로 이루어졌으며 인터넷(opt-in Internet panel), 전화와 우편의 방법으로 동시에 진행됐다. 각 mode를 비교한 결과, 인터넷 방법으로 전화 설문만큼 정확하게 추정 가능한 것으로 나타났다.

7. Annette Jackle and Peter Lynn (2007), Respondent Incentives in a Multi-Mode Panel Survey: Cumulative Effects on Nonresponse and Bias, ISER Working Paper 2007-01

본 연구 논문은 지속적인 incentive 제공이 attrition, bias, item nonresponse에 어떤 영향을 끼치는지에 대해 연구하였다. 영국의 젊은층을 대상으로 한 panel survey 자료를 바탕으로 incentive의 영향력을 분석해보았다.

제3장. 연구내용

1절. 연구내용

제안된 방법론은 다음과 같은 순서를 따라 진행된다.

(1) 모형 설정

(2) 설정된 모형의 모수 추정

(3) 예측치 발생 (prediction)

(4) 보정된 자료를 이용한 통계치 계산

첫 번째로 모형은 크게 측정 모형과 구조 모형을 세우는 것으로 구분되어진다. 논의를 간단하게 하기 위하여 어느 조사에서 두 가지 다른 방법 (방법 A, 방법 B)의 설문을 기입하도록 허용하였다고 하자. 여기서 민감한 관심 항목을 Y라고 하고 두 조사 방법에 의해 관측되는 값은 각각 y_a와 y_b라고 부르기로 하자. 이러한 경우 단순하게 두 방법을 그대로 집계하면 체계적인 편향이 존재하게 된다. 이때 $u_a = y_a - y$는 조사 방법 A에 기인하는 측정 오차가 되고 $u_b = y_b - y$는 조사 방법 B에 기인하는 측정오차가 된다.

한편 참값 y에 대해 다른 보조 변수를 이용한 회귀 모형을 세울 수 있을 것이다.

$$y_i = \beta' x_i + e_i \tag{1}$$

또는 보다 일반적으로는 (1)의 모형을 $f(y_i|x_i;\theta)$으로 표현할 수 있을 것이다. 식 (1)의 모형은 실제로는 관측되지 않는 참값 (y)에 대한 보조 변수와의 구

조적 관계를 나타내는 것으로써 흔히 구조 모형(structural error model)이라고 불린다. 한편 이러한 구조 모형 외에 관측값과 참값의 관계를 나타내는 측정 오차에 대한 모형이 필요하다. 이러한 모형은 측정 모형으로 $g_a(y_a|y)$ 또는 $g_b(y_b|y)$으로 표현될 수 있을 것이다.(예를 들어 $y_{ai}=y_i+u_{ai}$일 때 $u_{ai}\sim N(0,\sigma_a^2)$으로 가정하는 것은 조사방법 A의 측정 모형이 될 것이다.)

조사 방법 A로 자료를 통일하기 위해서는 $\sigma_a^2=0$으로 놓는 것이 필요하다. (그렇지 않으면 자료 구조상 identifiability 문제가 발생하게 된다. 즉, $\sigma_a^2=0$을 가정하지 않으면 전체 모형의 모수 추정이 불가능해진다.) 이러한 경우 (1)의 모형 (또는 더 일반적으로 $f(y_i|x_i;\theta)$으로 표현되는 모형)은 조사 방법 A로 얻어진 자료를 바탕으로 모수 추정이 가능해진다.

이제 모형의 모수 추정을 위해서 구조 모형과 측정 모형을 결합하면 참값에 대한 사후적 모형을 베이즈 정리를 사용하여 얻어낼 수 있게 된다. 조사 방법 B를 사용한 경우에 참값에 대한 사후적 모형은

$$f_b(y|y_b,x)=\frac{f(y|x)g_b(y_b|y)}{\int f(y|x)g_b(y_b|y)dy} \tag{2}$$

으로 표현된다. 여기서 $g_b(y_b|x,y)=g_b(y_b|y)$가 사용되었는데 이것은 y_b를 설명하는데 y가 주어지면 x는 필요하지 않다는 것을 의미한다. 이러한 사후 모형식 (2)의 계산은 정규분포처럼 간단한 경우에는 직접 계산이 가능하나 일반적인 경우에는 MCMC 방법 또는 Kim(2011)이 제안한 Parametric fractional imputation 테크닉을 사용하게 된다.

만약 두 모형(구조 모형과 측정 모형)이 모두 정규분포를 따른다면 (2) 의

모형은 평균이 $\hat{y_i} = \alpha(\hat{\beta_0} + \hat{\beta_1} x_i) + (1-\alpha) y_{ib}$이고 분산이 $\alpha\sigma_e^2$인 정규분포를 따르는데 여기서 $\alpha = \dfrac{\sigma_u^2}{\sigma_e^2 + \sigma_u^2}$으로 표현된다. 한편 σ_u^2의 추정은

$$y_{bi} = \beta_0 + x_i\beta_1 + u_i + e_i$$

임을 이용하여 일종의 Method of moments 방법으로 계산할 수 있다. 즉,

$$\hat{\sigma_u^2} = \frac{1}{n_b} \sum_{i \in S_b} \left\{ \frac{n}{n-p} \left(y_{ib} - \hat{\beta_0} - x_i \hat{\beta_1} \right)^2 - \hat{\sigma_e^2} \right\} \tag{3}$$

으로 계산하면 될 것이다.

만약 정규분포가 아닌 일반적인 모수적 분포를 따르는 경우에는 Monte Carlo EM 방법을 사용해야 할 것이다. 이때 E-step은 (2)의 사후 모형으로부터 참값 y를 몬테 카를로 방법을 통해서 발생(이를 imputation으로 이해할 수 있음)시키는 작업이 되고 M-step은 이렇게 해서 얻어진 imputation 값을 실제값인 것처럼 간주하고 $f(y_i|x_i;\theta)$의 모수와 $g_a(y_a|y)$, $g_b(y_b|y)$의 모수를 최대우도 추정법을 사용하여 추정하는 것이 된다.

본 자료는 상당한 부분(약 15-20%)에서 Y가 0의 값을 갖는다. 따라서 (1)의 회귀 모형을 사용하면 설명력이 떨어지므로 다음과 같은 구조 모형을 고려한다.

$$z_i = \beta' x_i + e_i, \quad e_i \sim N(0, \sigma_e^2) \tag{4}$$

$$y_i = z_i I(z_i > 0)$$

위 모형은 계량 경제학 분야에서 Tobit 모형으로 알려져 있는데, 잠재 변인 z_i를 가정하고 그 잠재 변인에 대한 회귀 모형을 설정한 후 만약 잠재 변인이 0 보다 크면 관측되고 그렇지 않으면 0으로 관측되는 모형이다. 따라서 이러한 모형의 모수 추정을 위해서는 잠재 변인을 먼저 발생시키고 그로부터 모수를 추정하게 된다. 실제로는 자료 A와 자료 B가 별도로 있으므로 자료 A의 관측치에 대해서는 (4)의 모형을 가정하고 자료 B의 관측치에 대해서는

$$
\begin{aligned}
z_{bi} &= z_{ai} + u_i, \quad u_i \sim N(0, \sigma_u^2) \\
z_{ai} &= \beta' x_i + e_i, \quad e_i \sim N(0, \sigma_e^2) \\
y_{bi} &= z_{bi} I(z_{bi} > 0)
\end{aligned}
\tag{5}
$$

으로 모형을 세울 수 있다. 식 (5)의 첫 번째 모형은 측정 모형으로써 두 모드별 잠재 변인의 차이에 대한 모형으로 표현된다. 따라서 각 자료로부터 두 개의 잠재 변인 (z_a, z_b)를 발생시킨 후 그로부터 모수를 추정하는 작업이 필요하게 된다.

이러한 경우 모수 추정을 좀 더 자세하게 설명하려면 다음의 성질을 이용하게 된다.

$$
\begin{aligned}
f(z_a | x, z_b, y_b) &= f(z_a | x, z_b) \\
f(z_b | x, z_a, y_a) &= f(z_b | z_a)
\end{aligned}
\tag{6}
$$

즉, z가 주어진 이상 $y = zI(z > 0)$의 값을 아는 것은 아무 도움이 되지 않는다는 것이다. 또한 z_b를 설명하는 것은 z_a로 충분하므로 두 번째 식이 성립하게 된다.

이제 모수 추정을 위해서는 EM 알고리즘을 사용하게 되는데 이러한 EM 알고리즘은 E-step과 M-step으로 나눠진다. E-step은 주어진 자료에서 잠재 변인을 여러 개 발생시켜 그를 바탕으로 한 몬테 카를로 방법을 사용하는 것이고 M-step은 발생된 잠재 변인값을 바탕으로 스코어식을 풀어서 모수 추정치를 업데이트 하는 것이다.

먼저 E-step에서 자료 A에서는 관측치 (x_i, y_{ai})로부터 잠재 변인값 (z_{ai}, z_{bi})를 발생시키고 자료 B에서는 관측치 (x_i, y_{bi})로부터 잠재 변인값 (z_{ai}, z_{bi})를 발생시키는 작업을 하게 된다. 자료 A에서는 관측값 (x_i, y_{ai})로부터 z_{ai}를 먼저 발생시키고 그로부터 z_{bi}를 발생시키나 자료 B에서는, y_{bi}로부터 z_{bi}를 먼저 발생시키고 그 값과 x_i를 이용하여 z_{ai}를 발생시키게 된다.

먼저 자료 A에서 z_{ai}는 만약 $y_{ai} > 0$이면 $z_{ai} = y_{ai}$가 되고 만약 $y_{ai} = 0$이면 $f(z_a|x, z_a < 0) \propto f(z_a|x) I(z_a < 0)$이므로 주어진 x 값을 이용하여 $f(z_a|x;\hat{\theta})$으로부터 z_a를 발생시킨 후 만약 $z_a > 0$이면 그 값을 버리고 다시 발생시키는 것을 반복함으로써 z_a를 발생시킨다. 그 이후 z_b는 발생된 z_a 값을 바탕으로 측정모형 $z_{bi} = z_{ai} + u_i$, $u_i \sim N(0, \sigma_u^2)$을 이용하여 z_b 값을 발생시킨다. 이를 독립적으로 m번 반복하여 몬테 카를로 샘플을 얻으면 되는데 이 경우 fractional weight는 $w_{ij}^* = 1/m$을 사용하게 된다. 본 자료 분석에서는 $m = 50$을 사용하였다.

자료 B에서는 y_{bi}가 관측되므로 만약 $y_{bi} > 0$이면 $z_{bi} = y_{bi}$가 되고 z_{ai}는 $f(z_a|x, z_b) \propto f(z_a|x) g(z_b|z_a)$이므로 이를 위해서 먼저 $f(z_a|x;\hat{\theta})$으로부터 z_{ai}를 발생시킨 후 이렇게 발생된 m개의 (z_{ai}, z_{bi}) 값에 fractional weight를 부여하는데

이때 사용되는 fractional weight는 $g(z_{bi}|z_{ai})$에 비례하도록 계산한다. 즉, j 번째 몬테 카를로 샘플로 발생된 $(z_{ai}^{*(j)}, z_{bi}^{*(j)})$에 부여되는 fractional weight는

$$w_{ij}^{*} = \frac{g\left(z_{bi}^{*(j)}|z_{ai}^{*(j)}\right)}{\sum_{k=1}^{m} g\left(z_{bi}^{*(k)}|z_{ai}^{*(k)}\right)}$$

으로 계산된다. 또한 만약 $y_{bi}=0$이면 $f(z_a, z_b|x, z_b<0)$으로부터 발생시켜야 하는데 이를 위해서 먼저 단순히 $f(z_a|x;\hat{\theta})$으로부터 z_{ai}를 발생시킨 후 이로부터 $g(z_{bi}|z_{ai})$을 사용하여 z_{bi}를 발생시킨 후 만약 $z_{bi}<0$를 만족하지 않으면 이를 버리고 다시 발생시키는 것을 반복하여 몬테 카를로 샘플을 얻어내면 된다. 이 경우 fractional weight는 $w_{ij}^{*}=1/m$을 사용하게 된다.

위의 방법으로 몬테 카를로 샘플(결국은 fractionally imputed data가 된다)이 얻어지면 이를 바탕으로 최대우도 추정법을 적용하여 모수 추정값을 업데이트 할 수 있다. 만약 가상의 (z_{ai}, z_{bi}) 값이 실제로 관측되었다면 식 (4)의 모수 추정은

$$\sum_{i \in S}\left(z_{ai} - \hat{\beta}' x_i\right)x_i = 0$$

$$\hat{\sigma}_e^2 = \frac{1}{n}\left\{\sum_{i \in S}\left(z_{ai} - \hat{\beta}' x_i\right)^2\right\}$$

으로 표현될 수 있고 식 (5)의 모수 추정은

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{i \in S}\left(z_{bi} - z_{ai}\right)^2$$

으로 계산될 수 있을 것이다. 실제로는 (z_{ai}, z_{bi}) 값이 관측되는 것이 아니라 그것들의 fractional imputation 값을 사용하므로 fractionally imputed data를 이용한 모수 추정은

$$\sum_{i \in S}\sum_{j=1}^{m} w_{ij}^{*}\left(z_{ai}^{*(j)} - \hat{\beta}' x_i\right)x_i = 0$$

$$\hat{\sigma}_e^2 = \frac{1}{n}\left\{\sum_{i \in S}\sum_{j=1}^{m} w_{ij}^{*}\left(z_{ai}^{*(j)} - \hat{\beta}' x_i\right)^2\right\}$$

과

$$\hat{\sigma}_u^2 = \frac{1}{n}\left\{\sum_{i \in S}\sum_{j=1}^{m} w_{ij}^{*}\left(z_{ai}^{*(j)} - z_{bi}^{*(j)}\right)^2\right\}$$

으로 계산된다. 이렇게 해서 얻어진 모수 추정치를 바탕으로 E-step을 다시 실시하는 것을 추정치가 수렴할 때까지 반복하면 EM 알고리즘이 완성된다.

이렇게 해서 잠재 변인값들이 모두 발생되고 이에 대한 fractional weight가 완성되면 자료를 다시 원자료 형태인 $y_{ai}^{*(j)} = z_{ai}^{*(j)} I(z_{ai}^{*(j)} > 0)$으로 변환하여 최종 통계를 계산한다. 즉, 자료 B의 데이터에서 $y_{ai}^{*(j)}$가 얻어지면 이를 이용하여 전체 평균 $\theta = E(Y_a)$의 추정치는 다음과 같이 얻어질 수 있을 것이다.

$$\hat{\theta} = \left\{\sum_{i \in S} w_i\right\}^{-1}\left\{\sum_{i \in S_A} w_i y_{ai} + \sum_{i \in S_B} w_i \sum_{j=1}^{m} w_{ij}^{*} y_{ai}^{*(j)}\right\}.$$

위의 제안된 방법론은 측정 오차 모형 (measurement error model)을 이용한 예측 기법을 바탕으로 제안되었고 실제 자료에서는 y가 사교육 시간과 사교육비의 두 가지 항목을 갖는 벡터가 된다. 이러한 자료의 확장에 대한 자세한 기술은 [부록 4]의 “Bivariate data extension”를 참조하기 바란다.

제4장. 연구결과

주요 관심 변수는 사교육 시간과 사교육비이며 이에 대하여 향후 분석을 진행한다. 사교육 시간은 1주일 동안의 사교육 시간을 의미하며 사교육비는 한 달 동안의 사교육 전체 비용을 의미한다.

1절. 기초자료 분석

실제 분석에 앞서 현행 자료의 분포를 살펴보았다. 1차 조사는 2011년 6월 1일 기준으로 조사되었고 2차 조사는 2011년 10월 1일 기준으로 조사되었는데 각각의 조사에서 대략 50%의 응답을 인터넷 조사로부터 얻어내었다. 각각의 조사에서 인터넷 조사로의 모드 선택은 자발적 선택이 아닌 랜덤 배정이 되었다.

<표 Ⅳ-1> 조사 모드 분포 - 차수별 분포

조사방법별 분포	종이 조사	인터넷 조사
1차 조사	53.30%	46.70%
2차 조사	52.60%	47.40%

<표 Ⅳ-2> 조사 모드별 분포 - 차수별 및 학교별 분포

학교별 분포		초등학교	중학교	일반고	특성화고
1차 조사	종이 조사	25.8%	25.4%	40.8%	8.0%
	인터넷 조사	21.9%	25.5%	43.6%	8.9%
2차 조사	종이 조사	25.4%	25.3%	40.9%	8.4%
	인터넷 조사	22.7%	25.8%	43.5%	8.0%

<표 Ⅳ-3> 조사 모드별 분포 - 차수별 및 지역별 분포

학교별 분포		강남	강북	광역시	중소도시	읍면
1차 조사	종이 조사	4.3%	7.8%	30.0%	43.4%	14.5%
	인터넷 조사	4.1%	8.1%	31.9%	38.2%	17.7%
2차 조사	종이 조사	4.4%	8.0%	30.2%	43.0%	14.4%
	인터넷 조사	4.0%	7.9%	31.6%	38.9%	17.6%

종이 조사와 인터넷 조사의 비율은 약 50% 정도로 두 조사 방식의 각 조사시기별 비율이 서로 비슷하다. 우선, 학교별 조사 비율을 살펴보면, 초등학교의 경우 종이 조사 비율이 인터넷 조사 비율보다 다소 높으며 일반고 조사 비율은 인터넷 조사 비율이 종이 조사보다 다소 높다. 지역별 분포를 살펴보면, 중소도시에서는 종이 조사 비율이, 읍면지역에서는 인터넷 조사 비율이 높으며 다른 지역에서는 비슷한 조사 비율을 보여주고 있다.

주요 변수의 무응답 비율은 다음과 같다.

<표 Ⅳ-4> 항목 별 무응답 분포

무응답 비율		전체	종이 조사	인터넷 조사
1차 조사	교육정도(부)	4.22	4.21	4.23
	교육정도(모)	3.31	3.06	3.52
	경제활동	0.5	0.49	0.51
	방과후 교육비	9.08	9.26	8.93
2차 조사	교육정도(부)	4.43	4.19	4.64
	교육정도(모)	3.44	2.96	3.88
	경제활동	0.52	0.36	0.67
	방과후 교육비	8.43	7.43	9.34

인터넷 조사의 무응답비율이 종이 조사에 비하여 다소 높은 경향을 보이나 유의한 차이는 없는 것으로 나타났다. 주요 관심변수인 사교육 시간에 대한 무응답은 존재하지 않았으나 사교육비에 대한 무응답은 상당히 높았다(1차

조사: 34.67%, 2차 조사: 38.77%). 이러한 무응답자의 대부분은 사교육 시간의 값이 0이기 때문에 이러한 경우에는 사교육비 값을 0으로 대체하였다.

주요변수에 대한 outlier를 확인하기 위하여 각 변수에 대한 4분위수를 살펴보았다.

<표 Ⅳ-5> 사교육 시간 응답값 분포

사교육 시간		1st Q	Median	3rd Q	Max.
1차 조사	종이	0.00	5.00	10.00	49.00
	인터넷	0.00	4.00	9.00	48.00
2차 조사	종이	0.00	4.00	10.00	62.00
	인터넷	0.00	4.00	9.00	46.00

<표 Ⅳ-6> 사교육비 응답값 분포

사교육비		1st Q	Median	3rd Q	Max.
1차 조사	종이	57.00	90.00	135.00	900.00
	인터넷	58.00	90.00	140.00	1170.00
2차 조사	종이	54.00	90.00	135.00	1080.00
	인터넷	57.00	90.00	141.00	1955.00

사교육 시간은 최대값이 50시간 내외로 기입상의 실수가 있거나 극단값이라고 여겨지지 않는다. 사교육비의 값은 사교육 시간에 대하여 비례하여 증가하기 때문에 단순히 사교육비 자체의 값이 크다고 outlier로 분류하기에는 다소 무리가 있기 때문에 사교육 시간당 비용을 통하여 outlier를 찾고자 하였다.

<표 Ⅳ-7> 사교육 시간당 교육비 응답값 분포 Ⅰ

시간당 교육비		1st Q	Median	3rd Q	Max.
1차 조사	종이	1.80	2.89	5.00	45.00
	인터넷	1.80	3.00	5.10	90.00
2차 조사	종이	1.75	2.81	4.92	60.00
	인터넷	1.80	3.00	5.00	56.50

<표 Ⅳ-8> 사교육 시간당 교육비 응답값 분포 Ⅱ

시간당 교육비		5~10 만원	10~15 만원	15~20 만원	20~30 만원	30~40 만원	40~50 만원	50만원 이상
1차 조사	종이	3187	504	123	44	19	3	0
	인터넷	2978	525	162	71	14	4	7
2차 조사	종이	2752	461	133	47	15	3	4
	인터넷	2812	468	165	79	13	9	2

시간당 사교육비가 최대 90만원인 자료가 있기는 하나 발생 가능한 값으로 판단하여 자료에 대하여 논리적 오류는 없다고 결론을 내렸다.

조사 방법에 따른 관심 변수의 평균에 대한 차이가 있는지 알아보기 위하여 계산한 기초통계는 다음과 같다. 기초통계와 조사 방법 간에 관심 변수의 평균이 차이가 있는가에 대한 t-test를 시행하였다.

<표 Ⅳ-9> 조사 방법에 따른 사교육 시간의 차이 검정

사교육 시간		평균	표준편차	t-value	p-value
1차 조사	종이	5.96	6.11	8.917	0.000
	인터넷	5.44	6.21		
2차 조사	종이	5.43	6.13	2.904	0.004
	인터넷	5.27	6.09		

<표 Ⅳ-10> 조사 방법에 따른 사교육비 차이 검정

사교육비		평균	표준편차	t-value	p-value
1차 조사	종이	71.20	77.80	3.808	0.000
	인터넷	68.32	82.46		
2차 조사	종이	64.80	79.36	-1.414	0.157
	인터넷	65.88	82.33		

2차 조사의 사교육비를 제외한 주요 관심 변수들은 조사 방식에 따라서 유의한 평균 차이가 존재한다. 1차 조사에서는 종이 조사가 인터넷 조사보다 더 큰 사교육 시간과 사교육비 값을 보여주었고 2차 조사에서도 사교육 시간에 대해서만큼은 종이 조사가 인터넷 조사보다 더 큰 값을 보여주었다. 아래 표에서 보여지듯이 인터넷 조사가 종이 조사보다 무응답 비율이 더 높았고 이는 인터넷 조사의 조사 신뢰도가 더 낮음을 의미할 수 있는 것으로 볼 수 있다. 또한 <표 Ⅳ-10>의 표준편차 값도 인터넷 조사가 더 크게 나오는 것을 확인할 수 있으므로 종이 조사가 더 신뢰할 수 있는 것으로 판단할 수 있을 것이다.

<표 Ⅳ-11> 조사 모드에 따른 사교육비 항목 무응답 비율

1차 조사	종이 조사	31.95%
	인터넷 조사	37.78%
2차 조사	종이 조사	37.91%
	인터넷 조사	39.73%

다음은 지역별, 학교별 주요 관심변수에 대한 기초통계와 조사 방식 간의 차이에 대한 t-test 결과이다. 일반적으로 말해서 종이 조사가 인터넷 조사보다 표준편차가 더 작은 경향이 있으며 종이 조사가 더 큰 사교육 시간 값을 보여주고 있다.

<표 Ⅳ-12> 조사 모드에 따른 사교육 시간 차이 검정 : 지역별

사교육 시간		조사방법	평균	표준편차	t-value	p-value
1차 조사	강남	종이	8.5	7.1	1.905	0.057
		인터넷	7.9	7.0		
	강북	종이	6.8	6.3	0.024	0.981
		인터넷	6.8	6.8		
	광역시	종이	6.2	6.0	3.747	0.000
		인터넷	5.8	6.1		
	중소도시	종이	5.9	6.1	7.912	0.000
		인터넷	5.2	6.2		
	읍면	종이	4.5	5.6	2.193	0.028
		인터넷	4.2	5.7		
2차 조사	강남	종이	8.0	6.9	0.803	0.422
		인터넷	7.7	6.7		
	강북	종이	6.3	6.4	0.679	0.497
		인터넷	6.1	6.6		
	광역시	종이	5.8	5.9	1.846	0.065
		인터넷	5.6	6.0		
	중소도시	종이	5.3	6.1	2.828	0.005
		인터넷	5.0	6.0		
	읍면	종이	3.8	5.6	-2.663	0.008
		인터넷	4.2	5.8		

<표 Ⅳ-13> 조사 모드에 따른 사교육 시간 차이 검정 : 학교별

사교육 시간		조사방법	평균	표준편차	t-value	p-value
1차 조사	초등학교	종이	7.7	5.9	1.446	0.148
		인터넷	7.5	6.4		
	중학교	종이	7.7	6.7	6.014	0.000
		인터넷	7.0	6.7		
	일반고	종이	4.6	5.4	5.457	0.000
		인터넷	4.2	5.5		
	특성화고	종이	1.6	3.9	-3.208	0.001
		인터넷	2.1	4.8		
2차 조사	초등학교	종이	7.5	6.0	-1.559	0.119
		인터넷	7.6	6.0		
	중학교	종이	7.1	6.7	3.140	0.002
		인터넷	6.7	6.6		
	일반고	종이	4.0	5.4	2.659	0.008
		인터넷	3.8	5.2		
	특성화고	종이	1.3	3.5	-3.911	0.000
		인터넷	1.9	5.1		

<표 Ⅳ-14> 조사 방법에 따른 사교육비 차이 검정: 지역별

사교육비		조사방법	평균	표준편차	t-value	p-value
1차 조사	강남	종이	145.0	126.4	0.254	0.799
		인터넷	143.4	134.9		
	강북	종이	97.3	98.5	-1.236	0.216
		인터넷	101.5	105.2		
	광역시	종이	70.5	70.4	0.972	0.331
		인터넷	69.3	77.9		
	중소도시	종이	69.3	72.3	3.558	0.000
		인터넷	65.4	75.5		
	읍면	종이	42.3	55.1	1.418	0.156
		인터넷	40.4	57.4		
2차 조사	강남	종이	144.8	137.4	-0.014	0.989
		인터넷	144.9	153.5		
	강북	종이	95.1	105.9	0.958	0.338
		인터넷	91.7	105.1		
	광역시	종이	64.7	70.0	-2.311	0.021
		인터넷	67.6	77.0		
	중소도시	종이	60.9	71.8	-1.272	0.204
		인터넷	62.3	73.6		
	읍면	종이	35.3	51.9	-4.582	0.000
		인터넷	41.2	56.9		

<표 Ⅳ-15> 조사 방법에 따른 사교육비 차이 검정: 학교별

사교육비		조사방법	평균	표준편차	t-value	p-value
1차 조사	초등학교	종이	69.9	60.3	2.770	0.006
		인터넷	66.7	59.4		
	중학교	종이	79.5	71.3	-0.432	0.666
		인터넷	80.1	80.9		
	일반고	종이	77.5	91.3	3.944	0.000
		인터넷	72.2	94.6		
	특성화고	종이	17.4	44.9	-1.733	0.083
		인터넷	20.0	48.1		
2차 조사	초등학교	종이	66.5	59.5	-2.345	0.019
		인터넷	69.2	59.7		
	중학교	종이	73.9	72.1	-3.088	0.002
		인터넷	78.4	81.1		
	일반고	종이	68.6	95.2	2.232	0.026
		인터넷	65.6	94.2		
	특성화고	종이	13.9	38.7	-2.721	0.007
		인터넷	17.8	48.2		

2절. 모형 설정

모형 설정을 위하여 학생과 부모의 인적사항 변수들을 보조 변수 x로 이용하였다. 분석의 편의를 위하여 보조 변수 x에 결측값이 있는 경우는 자료에서 제외한 후 분석에 사용하였다. 사용한 보조 변수들은 지역 구분, 교육수준, 성별, 학생 성적, 나이, 소득이다. 성별은 남자는 0, 여자는 1의 값을 갖는다. 나이는 부모 각각 나이의 평균값을 통하여 만들어진 변수이다. 교육수준은 고졸 이하와 대졸 이상을 기준으로 부모 모두 고졸 이하이면 교육수준 1, 부모 중 한 명이 대졸 이상이면 교육 수준 2, 모두 대졸 이상이면 교육 수준 3으로 분류하였다. 분석에 사용된 모형은 모형 (5)이며 분석은 조사시기별 학교별로 진행 되었다. 1차 조사에 대한 모형 설정 결과는 본문에 기재하고 2차 조사에 대한 모형 설정 결과는 [부록 2]에 기재하도록 한다. 다음은 1차 조사 자료를 바탕으로 분석한 사교육 시간과 사교육비에 대한 각각의 모형 설정 결과 및 모수 추정 결과이다.

1. 사교육 시간

사교육 시간에 대한 모형 (5)에 대한 모수 추정의 결과는 다음과 같다. 각 학교별로 모수 추정 결과를 기술하도록 한다. $\widehat{\beta_{(0)}}$는 초기 추정치를 의미하며, $\widehat{\beta_{EM}}$는 EM-algorithm을 통하여 얻어진 최종 모수 추정치이다.

<표 Ⅳ-16> 초등학교 사교육 시간에 대한 모형 선택 결과

사교육 시간	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	0.609	0.648	-0.114
강남	0.000	-	0.000
강북	0.033	0.515	0.301
광역시	0.696	0.417	0.306
중소도시	0.644	0.413	0.698
읍면	-0.275	0.448	-0.256
교육 수준 1	0.000	-	0.000
교육 수준 2	0.678	0.232	0.721
교육 수준 3	0.649	0.211	0.842
성별	-0.222	0.167	-0.157
학생성적	0.864	0.078	0.876
나이	-0.252	0.087	-0.222
소득	0.930	0.050	0.928
$\widehat{\sigma_e}$	6.250	-	6.040
$\widehat{\sigma_u}$	3.215	-	3.877

<표 Ⅳ-17> 중학교 사교육 시간에 대한 모형 선택 결과

사교육 시간	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	2.102	0.985	0.584
강남	0.000	-	0.000
강북	-1.700	0.663	-1.929
광역시	-1.539	0.609	-1.892
중소도시	-1.167	0.601	-1.688
읍면	-2.626	0.659	-3.731
교육 수준 1	0.000	-	0.000
교육 수준 2	0.902	0.285	1.014
교육 수준 3	0.942	0.269	0.962
성별	-0.674	0.217	-0.847
학생성적	1.274	0.091	1.458
나이	-0.404	0.135	-0.269
소득	0.963	0.062	0.955
$\widehat{\sigma_e}$	7.793	-	7.726
$\widehat{\sigma_u}$	2.604	-	3.373

<표 Ⅳ-18> 고등학교 사교육 시간에 대한 모형 선택 결과

사교육 시간	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	0.553	0.776	-0.442
강남	0.000	-	0.000
강북	-1.606	0.451	-0.996
광역시	-3.773	0.388	-3.413
중소도시	-4.157	0.377	-3.979
읍면	-6.332	0.446	-6.988
교육 수준 1	0.000	-	0.000
교육 수준 2	0.998	0.213	1.021
교육 수준 3	1.511	0.205	1.647
성별	0.361	0.163	0.716
학생성적	0.424	0.071	0.452
나이	-0.031	0.105	-0.111
소득	0.911	0.047	0.924
$\widehat{\sigma_e}$	7.309	-	7.367
$\widehat{\sigma_u}$	3.552	-	3.787

<표 Ⅳ-19> 특성화 고등학교 사교육 시간에 대한 모형 선택 결과

사교육 시간	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	-7.9	3.047	-8.826
강남	0.000	-	0.000
강북	-6.55	2.406	-4.955
광역시	-2.929	1.756	-3.599
중소도시	-2.821	1.677	-4.166
읍면	-7.562	1.768	-7.221
교육 수준 1	0.000	-	0.000
교육 수준 2	1.26	0.897	1.413
교육 수준 3	4.342	1.024	3.217
성별	0.833	0.724	1.825
학생성적	0.88	0.297	0.848
나이	-0.407	0.414	-0.392
소득	1.029	0.2	1.148
$\widehat{\sigma_e}$	10.617	-	10.894
$\widehat{\sigma_u}$	6.144	-	5.495

사교육 시간에 대한 모형 설정에 사용된 변수 중에서 지역 구분, 교육 수준, 성별 변수는 범주형 변수이고 학생 성적, 나이, 소득은 연속형 변수이다. 연속형 변수에 대한 모형 설정 결과에 대한 해석은 다른 모든 변수들이 동일할 때 연속형 변수가 한 단위 증가하게 되면 해당 변수의 β만큼 사교육 시간이 증가한다고 해석할 수 있다.

고등학교 사교육 시간에 대한 모형 선택 결과를 예로 들어보도록 하자. 다른 모든 변수들이 동일하다고 했을 때 소득이 한 단위 증가하면 사교육 시간은 0.924만큼 증가한다고 해석할 수 있다. 성별과 학생 성적 변수에 대하

여도 동일한 방법으로 해석할 수 있다.

범주형 변수에 대한 모형 설정 결과 해석은 다음과 같이 할 수 있다. 먼저 한 범주형 변수가 갖는 여러 수준들 중 특정한 수준을 기준으로 정한 후 기준 수준이 사교육 시간에 미치는 효과를 0이라고 한다. 예를 들면 지역구분에서는 강남, 교육 수준에서는 교육 수준 1이 기준이 되는 수준이다. 범주형 변수의 다른 수준들의 β의 의미는 다른 모든 변수들의 값이 동일하다고 할 때 기준 수준이 아닌 다른 수준을 값으로 가질 때 사교육 시간이 변하는 정도라고 할 수 있다.

고등학교 사교육 시간에 대한 모형 선택 결과를 예로 들어보도록 하자. 다른 모든 변수들이 동일하다고 할 때 교육 수준이 2인 사람은 교육 수준이 1인 사람에 비하여 사교육 시간이 1.021만큼 증가 하며 마찬가지로 교육 수준이 3인 사람은 교육 수준이 1인 사람에 비하여 사교육 시간이 1.647 만큼 증가한다고 해석할 수 있다. 지역 구분 역시 강남을 기준으로 하여 다른 모든 변수의 값이 동일하다면 강남이 아닌 다른 지역들은 각각 강북은 0.996, 광역시는 3.413, 중소도시는 3.979, 읍면은 6.988만큼 강남보다 사교육 시간이 줄어들게 된다고 해석할 수 있다.

2. 사교육비

사교육 시간의 잠재 변인을 z_1, 사교육비에 대한 잠재 변인을 z_2라 한다면 사교육비에 대한 모형은 다음과 같은 비율모형을 고려할 수 있다. 여기서 R_i는 시간당 사교육비를 의미한다.

$$z_{2i} = R_i z_{1i,}$$

$$R_i = \gamma' x_i + \eta_{i,} \quad \eta_i \sim N(0, \sigma_\eta^2)$$

위의 모형과 모형 (5)의 측정모형,

$$z_{bi} = z_{ai} + u_i, \quad u_i \sim N(0, \sigma_u^2)$$
$$y_{bi} = z_{bi} I(z_{bi} > 0)$$

을 함께 고려한다면 imputation에 필요한 사교육비에 대한 모형을 설정할 수 있다. 따라서 사교육비에 대한 모형 설정을 위해 추정해야할 모수는 시간당 사교육비 모형의 모수 γ, σ_η^2과 사교육비에 대한 측정모형의 분산 σ_u^2이다. 사교육비 모형에 대한 모수에 대한 추정의 결과는 다음과 같다. 각 학교별로 모수 추정 결과를 기술하도록 한다. 사용한 변수는 지역 구분, 교육 수준, 소득이다.

<표 Ⅳ-20> 초등학교 사교육비에 대한 모형 선택 결과

사교육비	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	11.861	0.601	12.054
강남	0.000	-	0.000
강북	-0.653	0.644	-1.153
광역시	-4.023	0.525	-4.011
중소도시	-4.224	0.519	-3.828
읍면	-5.017	0.567	-4.705
교육 수준 1	0.000	-	0.000
교육 수준 2	0.122	0.299	-0.013
교육 수준 3	0.805	0.267	0.620
소득	0.525	0.065	0.446
$\widehat{\sigma_e}$	7.499	-	6.562
$\widehat{\sigma_u}$	48.109	-	27.029

<표 Ⅳ-21> 중학교 사교육비에 대한 모형 선택 결과

사교육비	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	10.993	0.904	13.100
강남	0.000	-	0.000
강북	-1.002	0.900	-1.627
광역시	-2.316	0.819	-2.818
중소도시	-1.543	0.807	-2.191
읍면	-2.815	0.897	-3.159
교육 수준 1	0.000	-	0.000
교육 수준 2	1.536	0.400	0.813
교육 수준 3	1.802	0.369	1.409
소득	0.638	0.087	0.439
$\widehat{\sigma_e}$	9.811	-	8.410
$\widehat{\sigma_u}$	75.992	-	43.680

<표 Ⅳ-22> 특성화 고등학교 사교육비에 대한 모형 선택 결과

사교육비	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	14.821	2.621	17.582
강남	0.000	-	0.000
강북	2.948	4.013	2.965
광역시	-4.352	2.604	-3.046
중소도시	-2.78	2.461	-1.868
읍면	-2.672	2.67	-0.766
교육 수준 1	0.000	-	0.000
교육 수준 2	2.532	1.433	0.476
교육 수준 3	3.641	1.474	1.511
소득	0.186	0.318	-0.04
$\widehat{\sigma_e}$	11.238	-	8.735
$\widehat{\sigma_u}$	28.536	-	38.267

사교육비에 대한 모형 설정 결과표에서 소득 변수의 β의 의미는 다른 모든 변수들이 동일할 때 소득이 한 단위 증가하게 되면 소득의 β만큼 사교육 시간이 증가한다고 해석할 수 있다. 예를 들면 중학교 사교육비에 대한 모형 선택 결과에서 다른 모든 변수들이 동일하다고 할 때 소득이 한 단위 증가하면 시간당 사교육비는 0.439만큼 증가한다고 해석할 수 있다.

지역 구분과 교육 수준의 β의 의미는 다음과 같이 해석할 수 있다. 지역구분에서는 강남, 교육 수준에서는 교육 수준 1이 기준이 되는 수준이며 그 β 값은 0이다. 범주형 변수의 다른 수준들의 β의 의미는 다른 모든 변수들의 값이 동일하다고 할 때 기준 수준이 아닌 다른 수준을 값으로 가질 때 사교육 시간이 변하는 정도라고 할 수 있다. 예를 들면, 중학교 사교육 시간에

대한 모형 선택 결과에서 다른 모든 변수들이 동일하다고 할 때 교육 수준이 2인 사람은 교육 수준이 1인 사람에 비하여 시간당 사교육비가 0.813만큼 증가하며 마찬가지로 교육 수준이 3인 사람은 교육 수준이 1인 사람에 비하여 시간당 사교육비가 1.409만큼 증가한다고 해석할 수 있다. 지역 구분 역시 강남을 기준으로 하여 다른 모든 변수의 값이 동일하다면 강남이 아닌 다른 지역들은 각각 강북은 1.627, 광역시는 2.818, 중소도시는 2.191, 읍면은 3.159만큼 강남보다 시간당 사교육비가 줄어들게 된다고 해석할 수 있다.

3절. Imputation 결과 분석

Imputation 결과 분석을 위하여 사교육 시간과 사교육비에 대하여 평균, 0 초과 자료들의 평균, 자료의 값이 0인 자료들의 비율과 시간당 사교육비의 평균을 관심 모수로 고려하였다. 각 관심 모수별로 imputation 후 추정 결과를 기술하였다. 종이 조사 열은 종이 조사의 자료들로만 계산된 모수의 추정치, 인터넷 조사 열은 인터넷 조사 자료들로만 계산된 추정치이고 전체 열은 종이 조사와 인터넷 조사를 모두 사용하여 계산된 추정치이다. 전체 열의 모수들의 계산식은 다음과 같다.

$$\hat{\theta}=\left\{\sum_{i\in S}w_i\right\}^{-1}\left\{\sum_{i\in S_A}w_i y_{ai}+\sum_{i\in S_B}w_i y_{bi}\right\}.$$

Imputation 열은 imputation을 완료한 후 다음의 식을 통하여 계산된 각 모수의 추정치를 나타낸다.

$$\hat{\theta}=\left\{\sum_{i\in S}w_i\right\}^{-1}\left\{\sum_{i\in S_A}w_i y_{ai}+\sum_{i\in S_B}w_i\sum_{j=1}^{m}w_{ij}^{*}y_{ai}^{*(j)}\right\}.$$

1. 사교육 시간

<표 Ⅳ-23> 사교육 시간 평균에 대한 추정 결과

평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.868	7.657	7.777	7.460
중학교	7.946	7.159	7.576	7.345
고등학교	4.757	4.331	4.551	4.255
특성화 고	1.792	2.198	1.995	1.641

<표 Ⅳ-24> 사교육 시간 0 초과 평균에 대한 추정 결과

0 초과 평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	9.142	9.185	9.160	8.703
중학교	10.346	9.980	10.180	9.772
고등학교	7.466	7.637	7.544	7.022
특성화 고	6.844	7.841	7.359	6.304

<표 Ⅳ-25> 사교육 시간 0 비율에 대한 추정 결과

0 비율	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.139	0.166	0.151	0.143
중학교	0.232	0.283	0.256	0.248
고등학교	0.363	0.433	0.397	0.394
특성화 고	0.738	0.720	0.729	0.740

2. 사교육비

<표 Ⅳ-26> 사교육비 평균에 대한 추정 결과

평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	72.071	68.479	70.529	70.190
중학교	82.891	83.395	83.128	82.232
고등학교	79.980	74.726	77.444	75.925
특성화 고	19.377	21.766	20.571	19.917

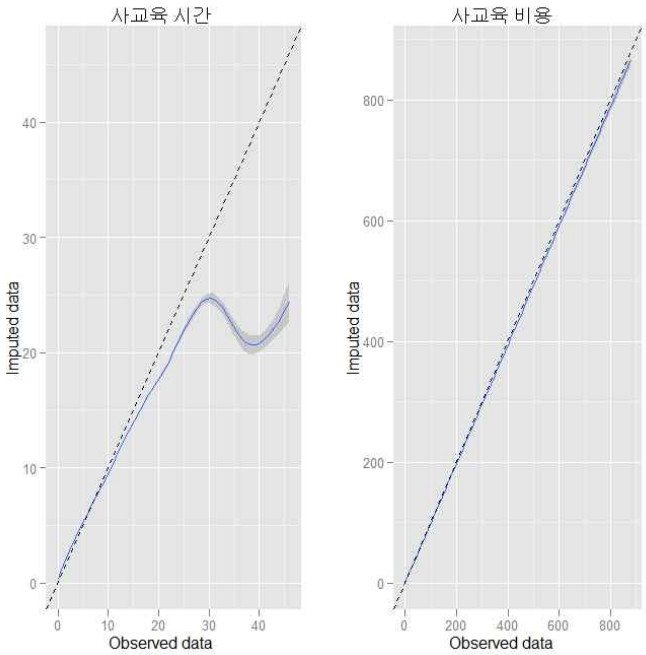
<표 Ⅳ-27> 사교육비 0 초과 평균에 대한 추정 결과

0 초과 평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	83.742	82.141	83.067	82.178
중학교	107.925	116.261	111.707	109.694
고등학교	125.528	131.775	128.362	125.290
특성화 고	74.023	77.645	75.895	75.720

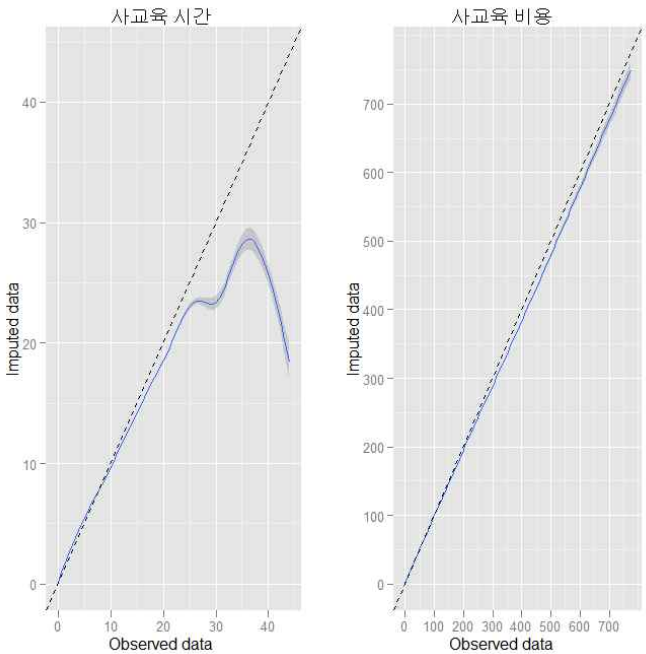
<표 Ⅳ-28> 사교육비 0 비율에 대한 추정 결과

0 비율	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.139	0.166	0.151	0.146
중학교	0.232	0.283	0.256	0.250
고등학교	0.363	0.433	0.397	0.394
특성화 고	0.738	0.720	0.729	0.737

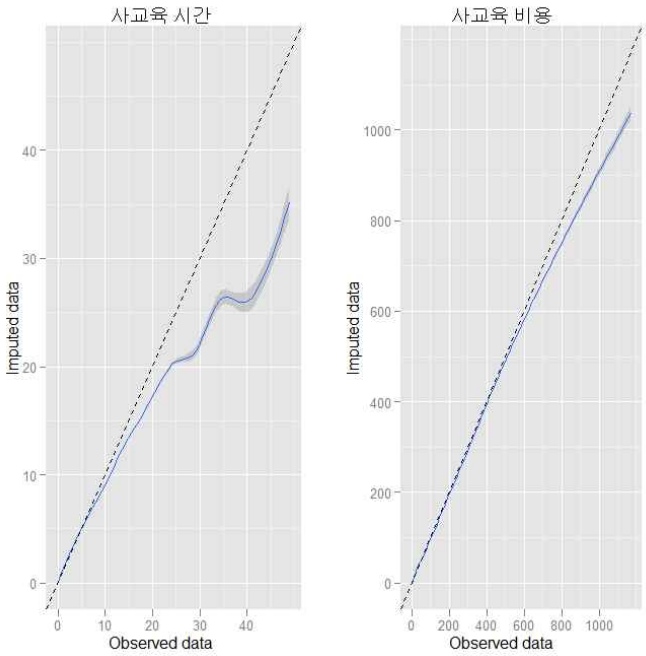
3. 관측 자료와 대체 자료 비교



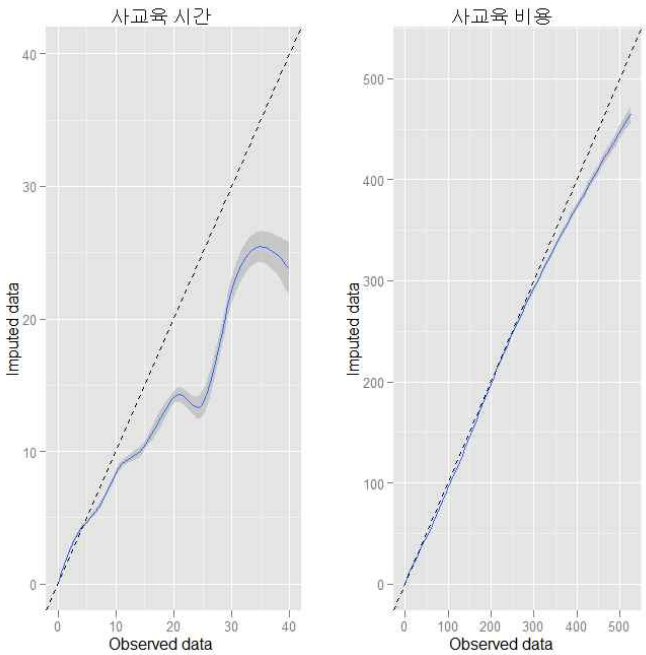
(그림 Ⅳ-1) 초등학교 사교육 시간과 비용에 대한 관측 자료와 대체 자료 비교



(그림 Ⅳ-2) 중학교 사교육 시간과 비용에 대한 관측 자료와 대체 자료 비교



(그림 Ⅳ-3) 고등학교 사교육 시간과 비용에 대한 관측 자료와 대체 자료 비교



(그림 Ⅳ-4) 특성화 고등학교 사교육 시간과 비용에 대한 관측 자료와 대체 자료 비교

4. 시간 당 사교육비

<표 Ⅳ-29> 시간 당 사교육비 평균에 대한 추정 결과

평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	2.691	2.784	2.730	2.647
중학교	3.301	3.809	3.532	3.313
고등학교	5.198	5.464	5.319	5.238
특성화 고	3.511	3.698	3.608	3.670

사교육 시간, 사교육비 변수의 세 가지 관심 모수 모두에 대하여 Imputation 열의 값들이 인터넷 조사와 전체 열의 값들보다 작게 추정되었음을 알 수 있다. 시간당 비용의 평균 또한 Imputation 열의 값들이 인터넷 조사와 전체 열의 값들보다 작게 추정되었음을 위의 표를 통하여 확인할 수 있다. 또한 사교육 시간과 비용에 대한 관측 자료와 대체 자료의 비교 그림에서 볼 수 있듯이 관측 자료 중 극단적으로 큰 값들은 대체 자료에서는 그 크기를 줄여주어 imputation을 통해 생성된 대체 자료에는 극단적으로 큰 값들의 비율이 관측 자료보다 적어짐을 알 수 있다.

이는 고려한 모형 하에서는 이론적으로 인터넷 조사의 평균이 종이 조사의 평균보다 더 크게 나타나도록 되어 있으므로 제안된 방법론은 이를 보정하기 위해 인터넷 조사의 값을 줄여주게 되기 때문이다. 하지만 현재의 자료에서는 선택 편향이 존재하지 않기 때문에 0 초과 평균에서는 종이 조사와 인터넷 조사가 비슷한 평균값을 갖는 반면 0의 비율이 인터넷 조사에서 종이 조사보다 크기 때문에 인터넷 조사의 평균값들이 작아지게 되는 현상을 보인다. 다음 장에서 (5)에서 가정된 모형 하에서 인공적으로 만든 자료와 선택 편향이 존재하는 자료에 대한 모의실험을 통하여 imputation 효과를 더 명확히 확인하고자 한다.

<표 Ⅳ-24>에서 <표 Ⅳ-30>의 분석은 종이 조사 결과를 기준으로 인터넷 조사 결과를 종이 조사 결과로 대체하여 얻어진 자료를 바탕으로 분석한 결과이다. [부록 3]에서는 거꾸로 인터넷 조사결과를 기준으로 종이 조사 결과를 대체하여 얻어진 자료를 분석한 결과를 기술하였다.

[부록 3]의 분석 결과도 역시 대체값의 평균이 본 조사값의 평균보다 작아지는 것을 볼 수 있다. 그 이유는 측정 오차가 있는 자료는 실제 참값의 자료보다 분산이 더 큰 분포를 따르므로 이를 보정한 예측값은 실제 관측치보다 더 평균 쪽으로 가까운 방향으로 얻어지게 될 것이다. 그러나 본 자료의 경우에는 0을 기준으로 그보다 더 작은 값은 0 으로 절삭이 되므로 평균보다 큰 관측값의 보정치는 그대로 줄어들지만 평균보다 작은 관측값의 보정치는 그 값이 커지는 효과가 절삭 때문에 상쇄되는 경향이 있어서 전체적으로는 예측치의 평균이 더 줄어들게 된다. 예를 들어 실제 참값의 평균이 2라고 할 때 관측치가 5라고 하면 예측치는 그보다 작은 값으로(예: 4) 보정되지만 관측치가 -2인 경우 예측치가 -1로 보정되더라도 둘 다 자료상에서는 0으로 나타나기에 보정의 효과가 나타나지 않는다. 따라서 아주 큰 값들은 줄어들지만 아주 작은 값들은 어차피 0에서 머물러 있게 되므로 결과적으로 대체값의 평균은 관측값의 평균보다 더 작아지게 된다.

제5장. 모의 실험 연구

3장에서 제안한 방법론에 대한 검증을 위하여 참값을 알고 있는 인공적인 자료를 생성하여 제안한 방법론을 적용하여 그 결과를 분석해보고자 한다. 또한 기존 자료는 선택 편향이 없는 자료이지만 선택 편향이 있는 경우의 분석을 위하여 기존 자료를 모집단으로 하고 지역, 교육 수준, 소득에 따라 추출확률을 달리하는 PPS sampling 방법을 통하여 선택 편향이 있는 인공적인 자료를 생성한 후 분석을 진행한다.

1절. 추정 방법론 검증

추정 방법론 검증을 위하여 두 가지 모집단을 고려하였다. 첫 번째 모집단은 정규분포 하에서 (5)에서 제안된 모형을 따르는 자료이다. 두 번째 모집단은 관심 변수 y가 정규분포가 아닌 감마분포를 따른다고 가정했을 때 (5)에서 제안된 모형을 따르는 경우이다. 두 모집단 모두 전체 $N=2{,}000$개의 자료를 생성하였고 이를 랜덤하게 A자료와 B자료로 각각 $n_a=n_b=1{,}000$개씩 나누었다. 각 몬테 카를로 샘플은 $B=500$번 반복해서 생성하였다. Fractional imputation 개수는 $M=50$이다.

1. 정규분포가정

$x_1 \sim N(1,1), x_2 \sim N(3,1)$을 따르는 보조 변수 x를 생성한다. 자료 A의 잠재변인 z_a와 관측치 y_a는 다음의 식을 이용하여 생성한다.

$$z_{ai} = -1.7 + x_{1i} + 0.5x_{2i} + e_i \quad e_i \sim N(0,1)$$

$$y_{ai} = z_{ai} I(z_{ai} > 0)$$

자료 B의 잠재 변인 z_b와 관측치 y_b는 다음의 식을 이용하여 생성한다.

$$z_{bi} = z_{ai} + u_i, \quad u_i \sim N(0, 0.8)$$
$$y_{bi} = z_{bi} I(z_{bi} > 0)$$

다음은 3장에서 제안된 방법론을 이용하여 분석한 분석 결과이다.

<표 Ⅴ-1> 정규분포가정 모형 모수 추정 결과

	β_0	β_1	β_2	σ_e^2	σ_u^2
참값	-1.700	1.000	0.500	1.000	0.800
추정값	-1.704	0.999	0.501	0.979	0.858

<표 Ⅴ-2> 정규분포가정 관심 모수 추정 결과

	종이 조사	인터넷 조사	전체	Imputation
평균	1.080	1.174	1.127	1.078
0 초과 평균	1.541	1.734	1.634	1.534
0 비율	0.297	0.323	0.310	0.297

모형의 모수는 측정 모형의 분산을 다소 크게 추정하지만 구조 모형의 모수에 대해서는 참값을 큰 오차 없이 상당히 정확하게 추정하고 있다. 세 가지 관심 모수에 대하여 모두 imputation의 결과가 기존의 전체 자료만 가지고 추정했을 때보다 작게 추정되며 이는 참값(종이 조사 결과)과 가까이 추정되었음을 알 수 있다.

2. 감마분포 가정

이 모의 실험에서는 잠재 변인과 관측치를 구분하지 않도록 한다. 그 이유는 감마분포를 가정하기 때문에 생성된 변수가 항상 0보다 크기 때문이다. $x_1 \sim N(5,0.8), x_2 \sim N(3,0.8)$을 따르는 보조 변수 x를 생성한다. 자료 A의 y_a는 다음의 식을 이용하여 생성한다.

$$y_{ai} \sim \Gamma(\alpha = 2, \frac{1}{2\mu_i})$$
$$\mu_i = \frac{1}{E(y_{ai})} = 2 + x_{1i} + 0.5x_{2i}$$

자료 B의 y_b는 다음의 식을 이용하여 생성하도록 한다.

$$y_{bi} = y_{ai} + u_i, \quad u_i \sim N(0, 0.8).$$

<표 V-3> 감마분포가정 모형 모수 추정 결과

	β_0	β_1	β_2	σ_e^2	σ_u^2
참값	2.000	1.000	0.500	2.000	0.800
추정값	2.076	0.994	0.489	2.005	0.808

<표 V-4> 정규분포가정 관심 모수 추정 결과

	종이 조사	인터넷 조사	전체	Imputation
평균	0.119	1.119	0.119	0.119

감마분포 가정의 경우 잠재 변인 모형을 고려하지 않았기 때문에 이론적 모형 하에서는 종이 조사, 인터넷 조사, 전체 열의 평균이 모두 같아야 한다. Imputation열의 값도 세 열의 평균값과 동일하므로 감마분포 가정 하에서도 3장에서 제안된 방법이 잘 적용됨을 알 수 있다.

2절. 선택 편향 보정 검증

선택 편향이 있는 자료에서 선택 편향에 따른 오차를 보정하기 위한 방법론을 3장에서 제안하였다. 이를 검증하기 위하여 현재의 자료를 모집단으로 하고 인터넷 자료의 추출확률을 다음의 세 가지 변수에 따라 달리 적용하여 PPS sampling 방법을 통해 선택 편향이 있는 자료를 생성하도록 한다.

<표 Ⅴ-5> 추출확률에 사용된 변수들

지역	0 (중소도시, 읍면지역), 1 (강남, 강북, 광역시)
교육	0 (고졸-고졸), 1 (고졸-대졸 이상), 2(대졸 이상-대졸 이상)
소득	0 (0~400만원), 1 (400만원 이상)

인터넷 추출확률은 대도시에 살수록, 부모가 고등교육을 받았을수록, 부모의 소득이 높을수록 높게 가정하였다. 인터넷 자료의 추출확률에 영향을 미치는 변수에 따라 다음의 세 가지 경우에 따른 모의 실험을 고려하도록 한다. 모형의 모수는 구조 모형과 측정 모형의 분산에 대해서만 추정결과를 기술하기로 한다. 관심 모수는 앞에서와 같이 평균, 0 초과 평균, 0의 비율이며

이에 대한 추정결과를 기술한다. 각각의 추출확률과 학교별 모두 PPS sampling의 크기는 $n_a = n_b = 1000$개이며 각 몬테 카를로 샘플은 $B = 100$개씩 반복해서 생성하였다.

서로 다른 추출확률에 따른 PPS sampling을 통한 세 모의 실험 모두 사교육 시간과 사교육비 변수에 대한 인터넷 조사 자료의 평균과 0 초과 평균값이 종이 조사 자료의 값보다 크게 됨을 알 수 있다. 따라서 전체자료를 통하여 관심 모수에 대한 추정을 시행하게 되면 유의한 오차가 발생함을 알 수 있다. Imputation을 통하여 인터넷 조사 자료의 값들을 줄여줌으로써 관심 모수 추정에 있어 오차를 줄일 수 있게 됨을 아래의 표를 통하여 확인 할 수 있다. 다음은 각 모의 실험에 대한 결과표이다.

1. Case 1 (지역-교육)

<표 Ⅴ-6> 지역과 교육에 따른 인터넷 자료의 추출확률

지역	교육	추출확률 (인터넷)
0	0	0.1
0	1	0.4
0	2	0.7
1	0	0.3
1	1	0.6
1	2	0.9

지역: 0 (중소도시, 읍면지역), 1 (강남, 강북, 광역시)
교육: 0 (고졸-고졸), 1 (고졸-대졸 이상), 2 (대졸 이상-대졸 이상)

<표 Ⅴ-7> Case 1, 학교별 사교육 시간 모형 모수 추정

	σ_e	σ_{u_1}
초등학교	5.855	4.472
중학교	7.443	4.332
고등학교	7.379	4.158
특성화 고	14.043	8.087

<표 Ⅴ-8> Case 1, 학교별 사교육 시간 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.276	8.043	7.660	7.304
중학교	7.423	7.795	7.609	7.346
고등학교	4.132	4.952	4.542	4.233
특성화 고	1.616	2.493	2.055	1.652

<표 Ⅴ-9> Case 1, 학교별 사교육 시간 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	8.852	9.278	9.065	8.537
중학교	10.228	10.082	10.155	9.662
고등학교	7.120	7.855	7.488	6.955
특성화 고	6.740	8.059	7.400	6.662

<표 Ⅴ-10> Case 1, 학교별 사교육 시간 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.178	0.133	0.156	0.144
중학교	0.274	0.227	0.251	0.240
고등학교	0.419	0.370	0.395	0.391
특성화 고	0.760	0.691	0.726	0.752

<표 Ⅴ-11> Case 1, 학교별 사교육비 모형 모수 추정

	σ_η	σ_{u_2}
초등학교	6.291	31.394
중학교	8.179	47.759
고등학교	12.211	62.079
특성화 고	8.645	78.525

<표 Ⅴ-12> Case 1, 학교별 사교육비 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.276	8.043	7.660	7.304
중학교	7.423	7.795	7.609	7.346
고등학교	4.132	4.952	4.542	4.233
특성화 고	1.616	2.493	2.055	1.652

<표 Ⅴ-13> Case 1, 학교별 사교육비 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	8.852	9.278	9.065	8.537
중학교	10.228	10.082	10.155	9.662
고등학교	7.120	7.855	7.488	6.955
특성화 고	6.740	8.059	7.400	6.662

<표 Ⅴ-14> Case 1, 학교별 사교육비 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.178	0.133	0.156	0.144
중학교	0.274	0.227	0.251	0.240
고등학교	0.419	0.370	0.395	0.391
특성화 고	0.760	0.691	0.726	0.752

<표 Ⅴ-15> Case 1, 학교별 시간 당 사교육비 평균

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.151	0.199	0.112	0.155
중학교	0.275	0.339	0.223	0.281
고등학교	0.469	0.534	0.402	0.468
특성화 고	0.779	0.813	0.711	0.762

2. Case 2 (지역-소득)

<표 Ⅴ-16> 지역과 소득에 따른 인터넷 자료의 추출확률

지역	소득	추출확률 (인터넷)
0	0	0.2
0	1	0.6
1	0	0.4
1	1	0.8

지역: 0 (중소도시, 읍면지역), 1 (강남, 강북, 광역시)
소득: 0 (0~400만원), 1 (400만원 이상)

<표 Ⅴ-17> Case 2, 학교별 사교육 시간 모형 모수 추정

	σ_e	σ_{u_1}
초등학교	5.788	4.526
중학교	7.489	4.485
고등학교	7.658	4.160
특성화 고	15.137	8.823

<표 Ⅴ-18> Case 2, 학교별 사교육 시간 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.382	8.174	7.778	7.442
중학교	7.494	7.814	7.654	7.399
고등학교	4.209	4.909	4.559	4.263
특성화 고	1.676	2.538	2.107	1.663

<표 Ⅴ-19> Case 2, 학교별 사교육 시간 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.161	0.139	0.150	0.135
중학교	0.263	0.237	0.250	0.237
고등학교	0.413	0.377	0.395	0.392
특성화 고	0.755	0.695	0.725	0.756

<표 Ⅴ-20> Case 2, 학교별 사교육 시간 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.178	0.133	0.156	0.144
중학교	0.274	0.227	0.251	0.240
고등학교	0.419	0.370	0.395	0.391
특성화 고	0.760	0.691	0.726	0.752

<표 Ⅴ-21> Case 2, 학교별 사교육비 모형 모수 추정

	σ_{η}	σ_{u_2}
초등학교	6.179	32.668
중학교	7.924	48.537
고등학교	12.315	62.259
특성화 고	8.856	84.402

<표 Ⅴ-22> Case 2, 학교별 사교육비 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	63.142	77.422	70.282	69.671
중학교	73.819	97.188	85.504	83.902
고등학교	66.043	89.809	77.926	75.928
특성화 고	17.13	26.382	21.756	20.162

<표 Ⅴ-23> Case 2, 학교별 사교육비 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	75.252	89.921	82.587	81.114
중학교	100.221	127.431	113.826	110.448
고등학교	112.539	144.223	128.381	125.176
특성화 고	69.77-	86.423	78.097	81.276

<표 Ⅴ-24> Case 2, 학교별 사교육비 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.161	0.139	0.150	0.141
중학교	0.263	0.237	0.250	0.240
고등학교	0.413	0.377	0.395	0.393
특성화 고	0.755	0.695	0.725	0.752

<표 Ⅴ-25> Case 2, 학교별 시간 당 사교육비 평균

	종이 조사	인터넷 조사	전체	Imputation
초등학교	2.526	2.821	2.676	2.589
중학교	3.084	4.111	3.603	3.229
고등학교	4.925	6.011	5.49	5.175
특성화 고	3.469	3.795	3.66	3.677

3. Case 3 (소득-교육)

<표 Ⅴ-26> 소득과 교육에 따른 인터넷 자료의 추출확률

소득	교육	추출확률 (인터넷)
0	0	0.1
0	1	0.3
0	2	0.4
1	0	0.4
1	1	0.7
1	2	0.9

소득: 0 (0~400만원), 1 (400만원 이상)
교육: 0 (고졸-고졸), 1 (고졸-대졸 이상), 2 (대졸 이상-대졸 이상)

<표 Ⅴ-27> Case 3, 학교별 사교육 시간 모형 모수 추정

	σ_e	σ_{u_1}
초등학교	5.744	4.482
중학교	7.404	4.345
고등학교	7.263	4.111
특성화 고	15.024	8.667

<표 Ⅴ-28> Case 3, 학교별 사교육 시간 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.06	8.442	7.751	7.443
중학교	7.152	8.093	7.623	7.410
고등학교	4.008	5.079	4.544	4.254
특성화 고	1.605	2.662	2.134	1.711

<표 Ⅴ-29> Case 3, 학교별 사교육 시간 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	8.62	9.562	9.091	8.61
중학교	10.091	10.244	10.168	9.709
고등학교	7.071	7.876	7.473	6.969
특성화 고	6.712	8.306	7.509	6.831

<표 Ⅴ-30> Case 3, 학교별 사교육 시간 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.181	0.117	0.149	0.136
중학교	0.291	0.210	0.251	0.237
고등학교	0.433	0.355	0.394	0.390
특성화 고	0.761	0.68	0.721	0.749

<표 Ⅴ-31> Case 3, 학교별 사교육비 모형 모수 추정

	σ_{η}	σ_{u_2}
초등학교	6.199	32.829
중학교	7.963	49.795
고등학교	11.787	62.818
특성화 고	8.485	87.659

<표 Ⅴ-32> Case 3, 학교별 사교육비 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	58.265	82.037	70.151	69.451
중학교	66.456	103.498	84.977	83.168
고등학교	59.826	95.212	77.519	75.418
특성화 고	15.759	29.484	22.622	21.083

<표 Ⅴ-33> Case 3, 학교별 사교육비 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	71.14	92.925	82.0325	80.914
중학교	93.764	130.999	112.382	109.491
고등학교	105.54	147.643	126.592	123.816
특성화 고	65.938	91.985	78.962	83.151

<표 Ⅴ-34> Case 3, 학교별 사교육비 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.181	0.117	0.149	0.142
중학교	0.291	0.210	0.251	0.240
고등학교	0.433	0.355	0.394	0.391
특성화 고	0.761	0.68	0.721	0.746

<표 Ⅴ-35> Case 3, 학교별 시간 당 사교육비 평균

	종이 조사	인터넷 조사	전체	Imputation
초등학교	2.514	3.028	2.778	2.612
중학교	3.004	4.148	3.617	3.269
고등학교	4.534	5.813	5.221	4.977
특성화 고	3.154	4.672	4.042	3.581

제6장. 결론

본 연구는 혼합 조사에서 발생하는 모드 효과 (Mode effect)를 일종의 측정 오차를 나타내는 random effect로 간주하고 이를 반영하여 보정하는 방법을 제안하고 사교육비 조사 자료에 적용하였다. 이 접근법은 또한 측정 오차 모형에서의 예측 방법론으로 이해될 수 있으며 이를 위해서는 구조 모형과 측정 모형을 세우고 그 모형에서의 모수를 추정한 후 베이즈 정리를 이용하여 예측하는 형태를 가지게 된다. 이때 모수 추정을 위해서는 EM 알고리즘이 사용되고 이를 위해서는 몬테 카를로 계산을 사용하게 된다. 모수 추정에서는 주어진 표본에 대한 sample model에 대한 모수 추정이므로 표본 설계 가중치를 사용하지 않는 추정을 하게 된다.

사교육비 자료의 경우 자료의 특성상 사교육 시간에 대한 구조 모형은 Tobit 모형을 사용하였고 사교육비에 대한 구조 모형은 사교육비를 바탕으로 한 일종의 ratio 모형 (절편 없는 회귀 모형)을 사용하였다. 측정 모형은 두 모드 효과를 나타내는 잠재 변수에 대해 정규분포를 따르는 것을 가정하였다. 이러한 모형을 바탕으로 잠재 변인에 대한 예측 모형은 베이즈 정리를 통해서 구현되는데 이로부터 imputation 값의 발생은 Kim (2011)이 제안한 Parametric fractional imputation을 사용하였다.

분석에 사용한 자료는 2011년 사교육비 조사 자료로써 1차 조사와 2차 조사에 각각 본 방법론을 적용하였다. 2011년도 조사에서는 조사 모드가 랜덤하게 배정되었지만 2012년도 조사부터는 응답자가 선택할 수 있도록 조사되었다. 이러한 경우에는 조사 모드 외에도 선택 편향 때문에 두 모드 간 관측치 평균의 차이가 더 커질 수 있는데 본 연구에서 제안된 방법을 사용하면 이러한 선택 편향이 상당히 줄어드는 보정을 하게 된다. 즉, 조사 모드 선택 경향이 예측 모형에 사용된 보조 변수에 의존하는 경우 이러한 선택 경향에 대한 메카니즘은 Rubin (1976)의 Missing at random으로 볼 수 있기 때문에 이를 예측 모형 설정 단계에서 무시할 수 있게 된다. 이는 5장 2절의

모의 실험을 통해 확인한 결과와도 일치한다.

본 연구에서 제안한 방법론은 통계청의 다른 혼합 조사에서도 사용할 수 있다. 이를 위해서는 관심 변수에 대한 올바른 구조 모형을 세우는 것이 중요한데 얼마나 좋은 모형을 사용하느냐가 얼마나 유용한 보정 효과를 가져오는지를 결정한다. 이를 위해서는 각 조사마다 해당 조사의 관심 변수에 대한 관련 지식을 반영하여 모형을 세워야 할 것이다.

예를 들어서 범주형 자료의 경우에는 일단 연속형 잠재 변수를 가정한 후에 연속형 잠재 변수는 구조 모형은 일반적인 회귀 모형(정규분포)을 따르고 측정 오차 모형은 정규분포를 따르는 것으로 한 후에 범주형 자료 관측은 그 잠재 변수 값이 가장 큰 항목이 관측되는 모형을 생각할 수 있을 것이다. 예를 들어 {1,2,3}의 값을 가지는 범주형 변수가 있다고 하면 이에 대한 3개의 잠재 변수 (Z_1, Z_2, Z_3)을 가정한 후에 이 잠재 변수에 대해 측정 오차 모형과 구조 모형을 세운 후에 이를 반영하여 잠재 변수에 대한 예측을 한 후에 최종 대체값은 예측된 잠재 변수 (Z_1^*, Z_2^*, Z_3^*) 값을 바탕으로 결정하는데 만약 Z_1^*가 다른 두 값보다 크면 $Y^* = 1$의 값을 갖고 만약 Z_2^*가 다른 두 값보다 크면 $Y^* = 2$의 값을 갖는 식으로 구현하면 될 것이다. 이러한 범주형 자료의 경우의 모형 설정 및 보정은 실제 자료를 바탕으로 향후에 더 연구될 것이다.

참고문헌

de Leeuw, E.D. (1992). *Data Quality in Mail, Telephone, and Face-to-face Surveys.* Amsterdam: TT-Publicaties.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* 2nd ed. New York, NY: John Wiley & Sons

Dillman, D. A., Sangster, R. L., Tarnai, J., & Rockwood, T. H. (1996). *Understanding differences in people answers to telephone and mail surveys. In: M. T. Braverman & J. K. Slater (Eds.) Advances in Survey Research.* New Directions for Evaluation, Number 70, Summer. San Francisco: Jossey-Bass.

Kim, J.K. (2011). *Parametric fractional imputation for missing data analysis.* Biometrika, 98, 119-132.

Rubin, D.B. (1976). *Inference and missing data.* Biometrika, 63, 581-592.

[부록 1] 2차 조사 자료 분석 결과

제7장.

1절. 사교육 시간

<표 부록Ⅰ-1> 초등학교 사교육 시간에 대한 모형 선택 결과

사교육 시간	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	0.933	0.696	0.736
강남	0.000	-	0.000
강북	0.360	0.537	-0.110
광역시	0.243	0.436	0.073
중소도시	0.334	0.434	0.049
읍면	-1.186	0.471	-1.040
교육 수준 1	0.000	-	0.000
교육 수준 2	0.148	0.244	0.404
교육 수준 3	0.523	0.223	0.533
성별	-0.174	0.177	-0.050
학생성적	0.978	0.088	0.899
나이	-0.324	0.092	-0.281
소득	0.880	0.054	0.904
$\widehat{\sigma_e^2}$	6.480	-	2.512
$\widehat{\sigma_u^2}$	1.143	-	1.410

<표 부록Ⅰ-2> 중학교 사교육 시간에 대한 모형 선택 결과

사교육 시간	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	-1.427	1.057	-1.863
강남	0.000	-	0.000
강북	-0.956	0.704	-1.540
광역시	-0.079	0.648	-0.716
중소도시	-0.116	0.639	-0.780
읍면	-2.039	0.711	-3.103
교육 수준 1	0.000	-	0.000
교육 수준 2	1.039	0.305	1.009
교육 수준 3	1.015	0.289	1.253
성별	-0.773	0.231	-0.812
학생성적	1.680	0.100	1.778
나이	-0.417	0.144	-0.324
소득	1.018	0.069	0.978
$\widehat{\sigma_e^2}$	8.117	-	2.854
$\widehat{\sigma_u^2}$	1.119	-	1.322

<표 부록Ⅰ-3> 고등학교 사교육 시간에 대한 모형 선택 결과

사교육 시간	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	-1.220	0.903	-1.028
강남	0.000	-	0.000
강북	-1.915	0.510	-1.667
광역시	-4.529	0.441	-4.048
중소도시	-5.043	0.428	-4.755
읍면	-7.789	0.515	-7.136
교육 수준 1	0.000	-	0.000
교육 수준 2	1.446	0.250	1.239
교육 수준 3	2.154	0.237	2.112
성별	0.559	0.190	1.039
학생성적	0.519	0.086	0.376
나이	-0.049	0.123	-0.227
소득	0.943	0.056	0.976
$\widehat{\sigma_e^2}$	8.204	-	2.840
$\widehat{\sigma_u^2}$	1.894	-	1.676

<표 부록Ⅰ-4> 특성화 고등학교 사교육 시간에 대한 모형 선택 결과

사교육 시간	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	-7.063	3.301	-10.449
강남	0.000	-	0.000
강북	-6.962	2.475	-3.478
광역시	-6.992	1.928	-6.116
중소도시	-6.051	1.748	-5.000
읍면	-7.920	1.809	-7.476
교육 수준 1	0.000	-	0.000
교육 수준 2	1.551	0.990	1.075
교육 수준 3	4.822	1.075	4.113
성별	3.245	0.795	3.488
학생성적	0.711	0.321	0.986
나이	-0.418	0.454	-0.555
소득	0.594	0.223	0.974
$\widehat{\sigma_e^2}$	11.077	-	3.387
$\widehat{\sigma_u^2}$	8.586	-	2.664

<표 부록Ⅰ-5> 사교육 시간 평균에 대한 추정 결과

평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.546	7.678	7.767	7.718
중학교	7.092	7.397	6.993	7.202
고등학교	3.726	4.076	3.881	3.981
특성화 고	1.268	1.394	1.987	1.669

<표 부록Ⅰ-6> 사교육 시간 0 초과 평균에 대한 추정 결과

0 초과 평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	8.875	9.079	9.151	9.112
중학교	9.906	10.203	10.000	10.107
고등학교	6.994	7.469	7.470	7.469
특성화 고	6.243	6.739	8.508	7.610

<표 부록Ⅰ-7> 사교육 시간 0 비율에 대한 추정 결과

0 비율	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.150	0.154	0.151	0.153
중학교	0.284	0.275	0.301	0.287
고등학교	0.467	0.454	0.480	0.467
특성화 고	0.797	0.793	0.766	0.781

2절. 사교육비

<표 부록Ⅰ-8> 초등학교 사교육비에 대한 모형 선택 결과

사교육비	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	11.049	0.592	11.242
강남	0.000	-	0.000
강북	-1.029	0.629	-1.111
광역시	-3.846	0.512	-3.763
중소도시	-3.945	0.511	-3.677
읍면	-3.258	0.560	-3.466
교육 수준 1	0.000	-	0.000
교육 수준 2	0.355	0.298	0.064
교육 수준 3	0.813	0.265	0.616
소득	0.546	0.065	0.482
$\widehat{\sigma_{\eta}}$	7.311	-	6.260
$\widehat{\sigma_{u}}$	49.927	-	31.951

<표 부록Ⅰ-9> 중학교 사교육비에 대한 모형 선택 결과

사교육비	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	12.922	0.972	14.700
강남	0.000	-	0.000
강북	-0.976	0.949	-1.741
광역시	-5.036	0.868	-4.812
중소도시	-3.429	0.856	-3.859
읍면	-4.507	0.973	-4.491
교육 수준 1	0.000	-	0.000
교육 수준 2	0.828	0.425	0.321
교육 수준 3	1.640	0.394	1.224
소득	0.674	0.097	0.458
$\widehat{\sigma_{\eta}}$	9.908	-	8.440
$\widehat{\sigma_{u}}$	78.997	-	55.921

<표 부록 Ⅰ-10> 고등학교 사교육비에 대한 모형 선택 결과

사교육비	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	18.550	1.210	22.968
강남	0.000	-	0.000
강북	0.684	1.155	-0.187
광역시	-3.675	0.998	-3.284
중소도시	-5.892	0.965	-4.152
읍면	-7.321	1.284	-3.603
교육 수준 1	0.000	-	0.000
교육 수준 2	1.172	0.650	0.035
교육 수준 3	3.428	0.598	1.723
소득	1.123	0.141	0.531
$\widehat{\sigma_{\eta}}$	16.737	-	13.917
$\widehat{\sigma_{u}}$	109.018	-	63.629

<표 부록Ⅰ-11> 특성화 고등학교 사교육비에 대한 모형 선택 결과

사교육비	$\widehat{\beta_{(0)}}$	표준오차	$\widehat{\beta_{EM}}$
절편	13.472	2.928	18.277
강남	0.000	-	0.000
강북	15.903	4.217	8.935
광역시	-4.014	3.055	-2.182
중소도시	-3.123	2.587	-2.025
읍면	-4.865	2.754	-2.175
교육 수준 1	0.000	-	0.000
교육 수준 2	1.081	1.826	0.523
교육 수준 3	3.428	1.742	0.943
소득	0.557	0.413	0.096
$\widehat{\sigma_{\eta}}$	12.294	-	9.284
$\widehat{\sigma_{u}}$	74.286	-	50.969

<표 부록Ⅰ-12> 사교육비 평균에 대한 추정 결과

평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	69.609	69.005	70.676	69.764
중학교	78.894	77.694	81.831	79.696
고등학교	68.655	70.808	68.013	69.443
특성화 고	16.229	14.642	19.708	16.984

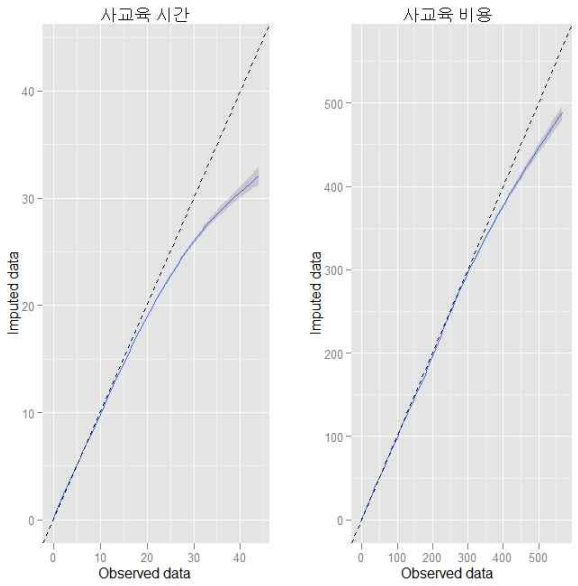
<표 부록 I -13> 사교육비 0 초과 평균에 대한 추정 결과

0 초과 평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	82.218	81.594	83.273	82.358
중학교	110.809	107.166	117.020	111.846
고등학교	129.103	129.744	130.887	130.288
특성화 고	78.621	70.760	84.371	77.464

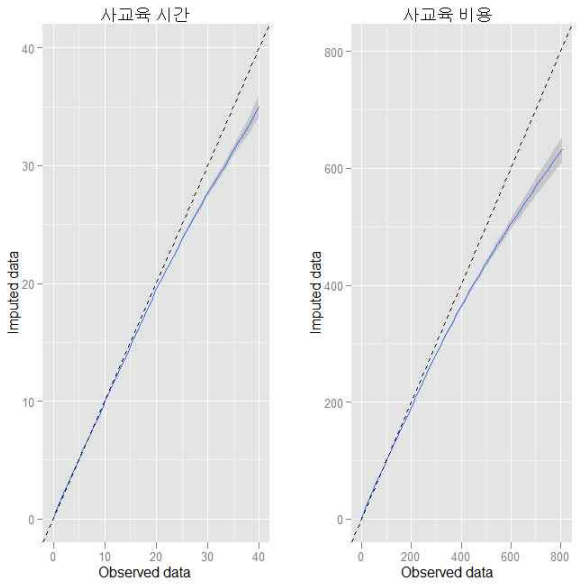
<표 부록 I -14> 사교육비 0 비율에 대한 추정 결과

0 비율	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.153	0.154	0.151	0.153
중학교	0.288	0.275	0.301	0.287
고등학교	0.468	0.454	0.480	0.467
특성화 고	0.794	0.793	0.766	0.781

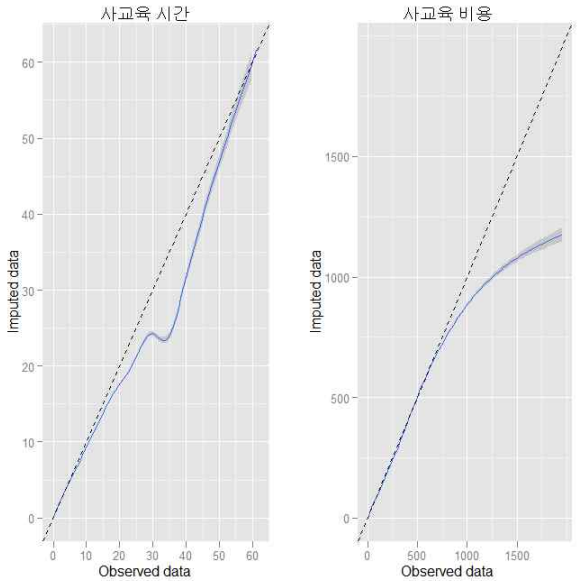
3절. 관측 자료와 대체 자료 비교



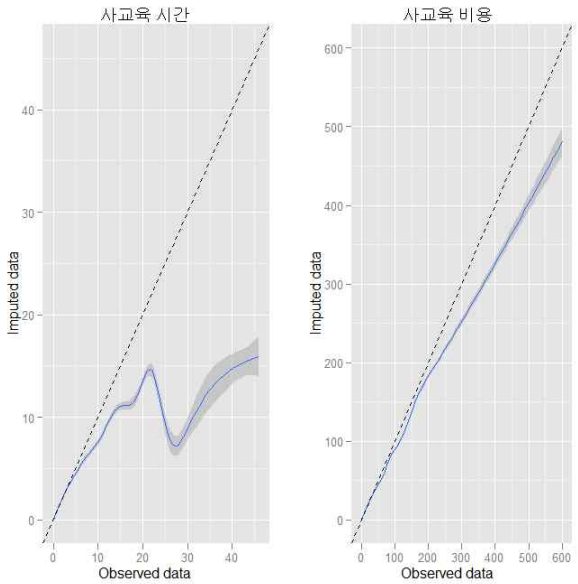
(그림 부록Ⅰ-1) 초등학교 사교육 시간과 비용에 대한 관측 자료와 대체 자료 비교



(그림 부록Ⅰ-2) 중학교 사교육 시간과 비용에 대한 관측 자료와 대체 자료 비교



(그림 부록Ⅰ-3) 고등학교 사교육 시간과 비용에 대한 관측 자료와 대체 자료 비교



(그림 부록Ⅰ-4) 특성화 고등학교 사교육 시간과 비용에 대한 관측 자료와 대체 자료 비교

4절. 시간 당 비용

<표 부록Ⅰ-15> 시간 당 사교육비 평균에 대한 추정 결과

평균	종이 조사	인터넷 조사	전체	Imputation
초등학교	2.635	2.707	2.668	2.585
중학교	3.320	3.821	3.558	3.303
고등학교	5.432	5.501	5.465	5.479
특성화 고	3.489	3.639	3.563	3.855

[부록 2] 2차 조사 자료를 통한 선택 편향 검증

1절. Case 1 (지역-교육)

<표 부록Ⅱ-1> 지역과 교육에 따른 인터넷 자료의 추출확률

지역	교육	추출확률 (인터넷)
0	0	0.1
0	1	0.4
0	2	0.7
1	0	0.3
1	1	0.6
1	2	0.9

지역: 0 (중소도시, 읍면지역), 1 (강남, 강북, 광역시)

교육: 0 (고졸-고졸), 1 (고졸-대졸 이상), 2 (대졸 이상-대졸 이상)

<표 부록Ⅱ-2> Case 1, 학교별 사교육 시간 모형 모수 추정

	σ_e	σ_{u_1}
초등학교	6.149	3.789
중학교	7.918	4.296
고등학교	7.992	4.074
특성화 고	11.804	7.770

<표 부록Ⅱ-3> Case 1, 학교별 사교육 시간 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.339	7.139	8.184	7.661
중학교	6.974	6.790	7.788	7.289
고등학교	3.652	3.446	4.440	3.943
특성화 고	1.327	1.263	2.403	1.833

<표 부록Ⅱ-4> Case 1, 학교별 사교육 시간 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	8.668	8.872	9.345	9.118
중학교	9.644	10.061	10.316	10.195
고등학교	6.872	7.173	7.601	7.408
특성화 고	6.487	6.756	9.127	8.142

<표 부록Ⅱ-5> Case 1, 학교별 사교육 시간 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.153	0.195	0.124	0.160
중학교	0.277	0.325	0.245	0.285
고등학교	0.469	0.520	0.416	0.468
특성화 고	0.796	0.813	0.737	0.775

<표 부록Ⅱ-6> Case 1, 학교별 사교육비 모형 모수 추정

	σ_{η}	σ_{u_2}
초등학교	6.140	29.417
중학교	8.245	47.250
고등학교	11.938	65.361
특성화 고	8.624	80.581

<표 부록Ⅱ-7> Case 1, 학교별 사교육비 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	67.637	58.559	77.969	68.264
중학교	79.809	65.328	97.279	81.303
고등학교	66.685	53.525	84.163	68.844
특성화 고	17.479	12.027	25.832	18.930

<표 부록Ⅱ-8> Case 1, 학교별 사교육비 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	80.295	72.779	89.028	81.248
중학교	110.753	96.801	128.860	113.728
고등학교	125.592	111.396	144.080	129.334
특성화 고	84.298	64.331	98.147	84.112

<표 부록Ⅱ-9> Case 1, 학교별 사교육비 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.158	0.195	0.124	0.160
중학교	0.279	0.325	0.245	0.285
고등학교	0.469	0.520	0.416	0.468
특성화 고	0.793	0.813	0.737	0.775

<표 부록Ⅱ-10> Case 1, 학교별 시간 당 사교육비 평균

	종이 조사	인터넷 조사	전체	Imputation
초등학교	2.487	2.640	2.567	2.550
중학교	2.923	4.028	3.509	3.143
고등학교	5.038	5.845	5.476	5.243
특성화 고	3.376	3.842	3.650	3.807

2절. Case 2 (지역-소득)

<표 부록Ⅱ-11> 지역과 소득에 따른 인터넷 자료의 추출확률

지역	소득	추출확률 (인터넷)
0	0	0.2
0	1	0.6
1	1	0.4
1	2	0.8

지역: 0 (중소도시, 읍면지역), 1 (강남, 강북, 광역시)
소득: 0 (0-400만원), 1(400만원 이상)

<표 부록Ⅱ-12> Case 2, 학교별 사교육 시간 모형 모수 추정

	σ_e	σ_{u_1}
초등학교	6.046	3.879
중학교	7.859	4.353
고등학교	8.040	4.122
특성화 고	11.694	7.702

<표 부록Ⅱ-13> Case 2, 학교별 사교육 시간 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.434	7.161	8.331	7.746
중학교	6.967	6.875	7.657	7.266
고등학교	3.669	3.556	4.405	3.981
특성화 고	1.362	1.337	2.400	1.869

<표 부록Ⅱ-14> Case 2, 학교별 사교육 시간 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	8.704	8.760	9.547	9.167
중학교	9.623	10.016	10.308	10.168
고등학교	6.891	7.271	7.639	7.470
특성화 고	6.462	6.792	8.959	8.042

<표 부록Ⅱ-15> Case 2, 학교별 사교육 시간 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.146	0.183	0.127	0.155
중학교	0.276	0.314	0.257	0.285
고등학교	0.468	0.511	0.423	0.467
특성화 고	0.789	0.803	0.732	0.768

<표 부록Ⅱ-16> Case 2, 학교별 사교육비 모형 모수 추정

	σ_{η}	σ_{u_2}
초등학교	6.160	30.465
중학교	8.276	47.291
고등학교	12.695	64.564
특성화 고	9.221	76.737

<표 부록Ⅱ-17> Case 2, 학교별 사교육비 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	69.716	60.289	80.309	70.299
중학교	80.089	67.718	95.447	81.582
고등학교	67.481	56.514	82.543	69.528
특성화 고	17.972	13.224	25.468	19.346

<표 부록Ⅱ-18> Case 2, 학교별 사교육비 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	82.063	73.749	92.039	83.192
중학교	110.998	98.659	128.491	114.164
고등학교	126.858	115.565	143.135	130.479
특성화 고	83.976	67.191	95.077	83.262

<표 부록Ⅱ-19> Case 2, 학교별 사교육비 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.150	0.183	0.127	0.155
중학교	0.278	0.314	0.257	0.285
고등학교	0.468	0.511	0.423	0.467
특성화 고	0.786	0.803	0.732	0.768

<표 부록Ⅱ-20> Case 2, 학교별 시간 당 사교육비 평균

	종이 조사	인터넷 조사	전체	Imputation
초등학교	2.513	2.767	2.645	2.565
중학교	3.320	3.993	3.667	3.457
고등학교	5.048	5.964	5.542	5.341
특성화 고	3.430	3.312	3.361	3.869

3절. Case 3 (소득-교육)

<표 부록Ⅱ-21> 지역과 교육에 따른 인터넷 자료의 추출확률

지역	교육	추출확률 (인터넷)
0	0	0.1
0	1	0.4
0	2	0.7
1	0	0.3
1	1	0.6
1	2	0.9

교육: 0 (고졸-고졸), 1 (고졸-대졸 이상), 2 (대졸 이상-대졸 이상)
소득: 0 (0-400만원), 1 (400만원 이상)

<표 부록Ⅱ-22> Case 3, 학교별 사교육 시간 모형 모수 추정

	σ_e	σ_{u_1}
초등학교	6.046	3.811
중학교	7.935	4.194
고등학교	7.995	4.059
특성화 고	11.730	7.230

<표 부록Ⅱ-23> Case 3, 학교별 사교육 시간 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	7.420	6.915	8.480	7.698
중학교	7.071	6.573	8.091	7.332
고등학교	3.654	3.321	4.567	3.944
특성화 고	1.428	1.261	2.539	1.900

<표 부록Ⅱ-24> Case 3, 학교별 사교육 시간 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	8.697	8.635	9.545	9.114
중학교	9.723	9.952	10.411	10.200
고등학교	6.874	7.133	7.631	7.413
특성화 고	6.549	6.758	8.791	7.994

<표 부록Ⅱ-25> Case 3, 학교별 사교육 시간 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.147	0.199	0.112	0.155
중학교	0.273	0.339	0.223	0.281
고등학교	0.469	0.534	0.402	0.468
특성화 고	0.782	0.813	0.711	0.762

<표 부록Ⅱ-26> Case 3, 학교별 사교육 시간 모형 모수 추정

	σ_{η}	σ_{u_2}
초등학교	6.033	29.751
중학교	8.090	47.474
고등학교	12.302	66.353
특성화 고	9.042	77.343

<표 부록Ⅱ-27> Case 3, 학교별 사교육비 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	68.472	55.533	82.841	69.187
중학교	80.189	60.736	102.940	81.838
고등학교	67.013	50.861	87.698	69.280
특성화 고	19.436	12.256	29.865	21.060

<표 부록Ⅱ-28> Case 3, 학교별 사교육비 0 초과 평균 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	80.638	69.345	93.237	81.911
중학교	110.644	91.953	132.465	113.852
고등학교	126.239	109.244	146.530	130.224
특성화 고	88.064	65.677	103.407	88.604

<표 부록Ⅱ-29> Case 3, 학교별 사교육비 0 비율 추정

	종이 조사	인터넷 조사	전체	Imputation
초등학교	0.151	0.199	0.112	0.155
중학교	0.275	0.339	0.223	0.281
고등학교	0.469	0.534	0.402	0.468
특성화 고	0.779	0.813	0.711	0.762

<표 부록Ⅱ-30> Case 3, 학교별 시간 당 사교육비 평균

	종이 조사	인터넷 조사	전체	Imputation
초등학교	2.490	2.889	2.702	2.604
중학교	3.043	4.300	3.720	3.304
고등학교	4.778	5.846	5.381	5.074
특성화 고	3.392	3.671	3.556	3.904

[부록 3] 종이 조사를 인터넷 조사로 대체 하는 경우의 imputation 결과

제9장.

1절. 사교육 시간

<표 부록Ⅲ-1> 학교별 사교육 시간 모형 모수 추정

	σ_e		σ_{u_1}	
	1차 조사	2차 조사	1차 조사	2차 조사
초등학교	6.188	6.048	3.624	3.831
중학교	7.371	7.556	5.016	5.134
고등학교	7.282	7.984	4.089	4.349
특성화 고	11.773	13.360	5.382	4.952

<표 부록Ⅲ-2> 사교육 시간 평균에 대한 추정 결과

	종이 조사		인터넷 조사		전체		Imputation	
	1차	2차	1차	2차	1차	2차	1차	2차
초등학교	7.868	7.678	7.657	7.767	7.777	7.718	7.314	7.295
중학교	7.946	7.397	7.159	6.993	7.576	7.202	6.809	6.489
고등학교	4.757	4.076	4.331	3.881	4.551	3.981	4.116	3.602
특성화 고	1.792	1.394	2.198	1.987	1.995	1.669	1.721	1.473

<표 부록Ⅲ-3> 사교육 시간 0 초과 평균에 대한 추정 결과

	종이 조사		인터넷 조사		전체		Imputation	
	1차	2차	1차	2차	1차	2차	1차	2차
초등학교	9.142	9.079	9.185	9.151	9.160	9.112	8.584	8.546
중학교	10.346	10.203	9.980	10.000	10.180	10.107	9.101	9.046
고등학교	7.466	7.469	7.637	7.470	7.544	7.469	6.866	6.822
특성화 고	6.844	6.739	7.841	8.508	7.359	7.610	6.730	7.080

<표 부록Ⅲ-4> 사교육 시간 0 비율에 대한 추정 결과

	종이 조사		인터넷 조사		전체		Imputation	
	1차	2차	1차	2차	1차	2차	1차	2차
초등학교	0.139	0.154	0.166	0.151	0.151	0.153	0.148	0.146
중학교	0.232	0.275	0.283	0.301	0.256	0.287	0.252	0.283
고등학교	0.363	0.454	0.433	0.480	0.397	0.467	0.401	0.472
특성화 고	0.738	0.793	0.720	0.766	0.729	0.781	0.744	0.792

2절. 사교육비

<표 부록Ⅲ- 5> 학교별 사교육비 모형 모수 추정

	σ_e		σ_{u_1}	
	1차 조사	2차 조사	1차 조사	2차 조사
초등학교	7.232	6.541	28.709	26.329
중학교	10.859	11.230	42.329	44.154
고등학교	14.007	13.497	53.116	57.512
특성화 고	11.571	9.676	49.862	47.357

<표 부록Ⅲ-6> 사교육비 평균에 대한 추정 결과

	종이 조사		인터넷 조사		전체		Imputation	
	1차	2차	1차	2차	1차	2차	1차	2차
초등학교	72.071	69.005	68.479	70.676	70.529	69.764	69.784	69.207
중학교	82.891	77.694	83.395	81.831	83.128	79.696	82.121	78.839
고등학교	79.980	70.808	74.726	68.013	77.444	69.443	75.724	67.942
특성화 고	19.377	14.642	21.766	19.708	20.571	16.984	20.113	16.863

<표 부록Ⅲ-7> 사교육비 0 초과 평균에 대한 추정 결과

	종이 조사		인터넷 조사		전체		Imputation	
	1차	2차	1차	2차	1차	2차	1차	2차
초등학교	83.742	81.594	82.141	83.273	83.067	82.358	82.489	81.523
중학교	107.925	107.166	116.261	117.020	111.707	111.846	110.488	110.674
고등학교	125.528	129.744	131.775	130.887	128.362	130.288	126.727	128.696
특성화 고	74.023	70.760	77.645	84.371	75.895	77.464	77.900	80.292

<표 부록Ⅲ-8> 사교육비 0 비율에 대한 추정 결과

	종이 조사		인터넷 조사		전체		Imputation	
	1차	2차	1차	2차	1차	2차	1차	2차
초등학교	0.139	0.154	0.166	0.151	0.151	0.153	0.154	0.151
중학교	0.232	0.275	0.283	0.301	0.256	0.287	0.257	0.288
고등학교	0.363	0.454	0.433	0.480	0.397	0.467	0.402	0.472
특성화 고	0.738	0.793	0.720	0.766	0.729	0.781	0.742	0.790

3절. 시간 당 비용

<표 부록Ⅲ-9> 시간 당 사교육비 평균에 대한 추정 결과

	종이 조사		인터넷 조사		전체		Imputation	
	1차	2차	1차	2차	1차	2차	1차	2차
초등학교	2.691	2.635	2.784	2.707	2.730	2.668	2.724	2.648
중학교	3.301	3.320	3.809	3.821	3.532	3.558	3.676	3.727
고등학교	5.198	5.432	5.464	5.501	5.319	5.465	5.472	5.558
특성화 고	3.511	3.489	3.698	3.639	3.608	3.563	3.974	3.847

[부록 4] Bivariate data extension

제10장.

- Notation:

1. $Y_a = (Y_{a1}, Y_{a2})$: study variables (Time, Cost) for mode A
2. $Y_b = (Y_{b1}, Y_{b2})$: study variables (Time, Cost) for mode B
3. X : auxiliary variable

- Model:

1. Model for Mode A observation:

(a) Partition the sample into school types (Elementary, Middle, High, etc).

(b) Model for Y_{a1} on x: zero-inflated regression model(ex: Tobit model).

$$y_{a1} = \begin{cases} z_{a1} & \text{if } z_{a1} > 0 \\ 0 & otherwise \end{cases}$$

where

$$z_{a1} = x'_i\beta + e_i, \quad e_i \sim N(0, \sigma_e^2).$$

Or, more generally, we have $f(z_{a1}|x) = f(z_{a1}|x;\theta_1)$

(c) Model of Y_{2a} on x and y_{1a}: Ratio model

$$z_{a2i} = R_i z_{a1i}$$

where

$$R_i = x'_i \beta + \eta_i .$$

Or, more generally, we have

$f(z_{a2}|x, z_{a1}) = f(z_{a2}|x, z_{a1}; \theta_2)$.

(Such models can be verified from the regression diagnostics in sample A data.

We may use the residual plots.)

2. Mode for mode B observation

Model of Y_{b1} on y_{a1}: measurement error model

$$z_{b1,i} = z_{a1,i} + u_i$$

where $u_i \sim N(0, \sigma_{u1}^2)$ and

$$y_{b1,i} = \begin{cases} z_{b1,i} & \text{if } z_{b1,i} > 0 \\ 0 & otherwise. \end{cases}$$

Also, Model of Y_{b2} on y_{a2} is

$$z_{b2,i} = z_{a2,i} + u_i$$

where $u_i \sim N(0, \sigma_{u2}^2)$ and

$$y_{b2,i} = \begin{cases} z_{b2,i} & \text{if } z_{b2,i} > 0 \\ 0 & otherwise. \end{cases}$$

The measurement error models can be written as

$$g(z_{b1}|z_{a1};\sigma_{u1}^2) \text{ and } g(z_{b2}|z_{a2};\sigma_{u2}^2).$$

Conditional independence assumption is im-plicitly used in the measurement error model. That is,

$$g(z_{b1}|z_{a1}, z_{a2}, x) = g_1(z_{b1}|z_{a1})$$

and

$$g(z_{b1}|z_{a1}, z_{a2}, x) = g_2(z_{b2}|z_{a2}).$$

- **Remark:**

1. The transformation from (z_1,z_2) to (y_1, y_2) is deterministic. That is $y_k = z_k I(z_k > 0)$. Thus, conditional on z, additional conditioning on y does not add any information:

$$f(z_a|x, z_b, y_b) = f(z_a|x, z_b). \tag{1}$$

2. Also,

$$f(z|y,x) = \begin{cases} I(z=y) & \text{if } y > 0 \\ f(z|x, z<0) & \text{if } y = 0 \end{cases} \tag{2}$$

3. Measurement error model is not based on the sample partition, while the original data model (Model for Mode A data) is based on the sample partition.

- Parameter estimation: EM algorithm using parametric fractional imputation.

1. Imputation for Data A: generate $(z^{*}_{a1,i}, z_{a2,i})$ from the conditional distribution of (z_{a1}, z_{a2}) given the observed value of $x_i, y_{a1,i}, y_{a2,i}$ using the estimated parameters.

   (a) Generation of $z^{*}_{a1,i}$:

$$f(z_{a1}|x_i, y_{a1,i}) = \begin{cases} I(z_{a1} = y_{a1,i}) & \text{if } y_{a1,i} > 0 \\ f(z_{a1}|x_i, z_{a1} < 0) & \text{if } y_{a1,i} = 0 \end{cases}$$

   where

$$f(z_{a1}|x_i, z_{a1} < 0) \propto f(z_{a1}|x_i) I(z_{a1} < 0).$$

   (b) Generation of $z^{*}_{a2,i}$:

$$f(z_{a2}|z^{*}_{a1}, y_{a2,i}) = \begin{cases} I(z_{a2} = y_{a2,i}) & \text{if } y_{a2,i} > 0 \\ f(z_{a2}|x_i, z^{*}_{a1}, z_{a2} < 0) & \text{if } y_{a2,i} = 0 \end{cases}$$

   where

$$f(z_{a2}|x_i, z^{*}_{a1}, z_{a2} < 0) \propto f(z_{a2}|x_i, z^{*}_{a1}) I(z_{a2} < 0)$$

   Thus, the imputed values are easy to be generated from.

2. Imputation model for Data B: Generate $(z^*_{a1,i}, z^*_{a2,i})$ from the conditional distribution of (z_{a1}, z_{a2}) given the observed value of $x_i, y_{b1,i}, y_{b2,i}$ using the estimated parameters.

(a) Generation of $(z^*_{a1,i}, z^*_{b1,i})$:

$$f(z_{a1}, z_{b1}|x_i, y_{b1,i}) = \begin{cases} f(z_{a1}, z_{b1}|\, x_i, z_{b1} = y_{b1,i}) & \text{if } y_{b1,i} > 0 \\ f(z_{a1}, z_{b1}|x_i, z_{b1} < 0) & \text{if } y_{b1,i} = 0 \end{cases}$$

[Case 1] For $y_{b1,i} > 0$,

$$\begin{aligned} f(z_{a1}, z_{b1}|x_i, z_{b1} = y_{b1,i}) &= f(z_{a1}|x_i, z_{b1} = y_{b1,i})\, I(z_{b1} = y_{b1,i}) \\ &= f(z_{a1}|x_i) g_1(y_{b1,i}|z_{a1})\, I(z_{b1} = y_{b1,i}) \end{aligned}$$

where $f(z_{a1}|x_i)$ is the density for the original data model for sample A and $g_1(\cdot)$ is the density for the measurement error model of z_{b1}.

[Case 2]

If $y_{b1,i} = 0$ then, by (2) again,

$$\begin{aligned} f(z_{a1}, z_{b1}|x_i, z_{b1} < 0) &= f(z_{a1}|x_i) f(z_{b1}|\, z_{a1}, x, z_{b1} < 0) \\ &= f(z_{a1}|x_i) f(z_{b1}|\, z_{a1}, z_{b1} < 0)\, I\,(z_{b1} < 0) \\ &= f(z_{a1}|x_i) g_1(z_{b1}|z_{a1})\, I\,(z_{b1} < 0) \end{aligned}$$

(b) Generation of $(z^*_{a2,i}, z^*_{b2,i})$

$$f(z_{a2}, z_{b2}|x_i, z^*_{a1}, y_{b1,i}) = \begin{cases} f(z_{a2}, z_{b2}|\, x_i, z^*_{a1}, z_{b2} = y_{b2,i}) & \text{if } y_{b2,i} > 0 \\ f(z_{a2}, z_{b2}|x_i, z^*_{a1}, z_{b2} < 0) & \text{if } y_{b2,i} = 0 \end{cases}$$

[Case 1]

If $y_{b2,i} > 0$ then

$$f(z_{a2}, z_{b2} | x_i, z^*_{a1}, z_{b2} = y_{b2,1}) = f(z_{a2} | x_i, z^*_{a1}) g_2(y_{b2,i} | z_{a2})\, I(z_{b2} = y_{b2,i})$$

[Case 2]

If $y_{b2,i} = 0$ then

$$f(z_{a2}, z_{b2} | x_i, z^*_{a1}, z_{b2} < 0) \propto f(z_{a2} | x_i, z^*_{a1}) g_2(z_{b2} | z_{a2})\, I(z_{b2} < 0).$$

3. Fractional Imputation

(a) Imputation for Data A: m imputed values are generated

$$z^{*(j)}_{a1,i} \sim f(z_{a1} | x_i, z_{a1} < 0) \quad \text{if } y_{a1,i} = 0$$

$$z^{*(j)}_{a2,i} \sim f(z_{a2} | x_i, z_{a2} < 0) \quad \text{if } y_{a2,i} = 0.$$

Otherwise, we use $z^{*(j)}_{a1,i} = y_{a1,i}$ and $z^{*(j)}_{a2,i} = y_{a2,i}$.

The fractional weights are given by

$$w^*_{ij(t)} \propto \begin{cases} \dfrac{f(z^{*(j)}_{a1,i} | x_i, z^{*(j)}_{a1,i} \le 0; \hat{\theta}_{1(t)})}{f(z^{*(j)}_{a1,i} | x_i, z^{*(j)}_{a1,i} \le 0; \hat{\theta}_{1(0)})} \dfrac{1}{f(z^{*(j)}_{a2,i} | x_i, z^{*(j)}_{a1,i}, z^{*(j)}_{a2,i} \le 0; \hat{\theta}_{2(t)})} & y_{a1,i} > 0, y_{a2,i} > 0 \\ \dfrac{1}{f(z^{*(j)}_{a2,i} | x_i, z^{*(j)}_{a1,i}, z^{*(j)}_{a2,i} \le 0; \hat{\theta}_{2(0)})} & y_{a1,i} = 0, y_{a2,i} = 0 \end{cases}$$

(b) Imputation for Data B: m imputated values

are generated from

$$z^{*(j)}_{a1,i} \sim f(z_{a1} | x_i; \hat{\theta}_{1(0)})$$

$$z^{*(j)}_{a2,i} \sim f(z_{a2} | x_i, z^{*(j)}_{a1,i}; \hat{\theta}_{2(0)}).$$

Also,

$$z_{b1,i}^{*(j)} \sim g_1(z_{b1} \mid z_{a1,i}^{*(j)}, z_{b1} < 0; \hat{\sigma}_{u1(0)}^2) \quad \text{if } y_{b1,i} = 0$$

$$z_{b1,i}^{*(j)} = y_{b1,i} \quad \text{if } y_{b1,i} > 0$$

and

$$z_{b2,i}^{*(j)} \sim g_2(z_{b2} \mid z_{a2,i}^{*(j)}, z_{b2} < 0; \hat{\sigma}_{u2(0)}^2) \quad \text{if } y_{b2,i} = 0$$

$$z_{b2,i}^{*(j)} = y_{b2,i} \quad \text{if } y_{b2,i} > 0$$

The fractional weights are given by for $i, y_{b1,i} > 0, y_{b2,i} > 0$,

$$w_{ij(t)}^{*} \propto g_1(z_{b1,i} \mid z_{a1,i}^{*(j)}; \hat{\sigma}_{u1(t)}^2) g_2(z_{b2,i} \mid z_{a2,i}^{*(j)}; \hat{\sigma}_{u2(t)}^2)$$

$$\times \frac{f(z_{a1,i}^{*(j)} \mid x_i; \hat{\theta}_{1(t)})}{f(z_{a1,i}^{*(j)} \mid x_i; \hat{\theta}_{1(0)})} \frac{f(z_{a2,i}^{*(j)} \mid x_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(t)})}{f(z_{a2,i}^{*(j)} \mid x_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(0)})}$$

and for $i, y_{b1,i} = 0, y_{b2,i} = 0$,

$$w_{ij(t)}^{*} \propto \frac{g_1(z_{b1,i}^{*(j)} \mid z_{a1,i}^{*(j)}, z_{b1,i}^{*(j)} \le 0; \hat{\sigma}_{u1(t)}^2)}{g_1(z_{b1,i}^{*(j)} \mid z_{a1,i}^{*(j)}, z_{b1,i}^{*(j)} \le 0; \hat{\sigma}_{u1(0)}^2)} \frac{g_2(z_{b2,i}^{*(j)} \mid z_{a2,i}^{*(j)}, z_{b2,i}^{*(j)} \le 0; \hat{\sigma}_{u2(t)}^2)}{g_2(z_{b2,i}^{*(j)} \mid z_{a2,i}^{*(j)}, z_{b2,i}^{*(j)} \le 0; \hat{\sigma}_{u2(0)}^2)}$$

$$\times \frac{f(z_{a1,i}^{*(j)} \mid x_i; \hat{\theta}_{1(t)})}{f(z_{a1,i}^{*(j)} \mid x_i; \hat{\theta}_{1(0)})} \frac{f(z_{a2,i}^{*(j)} \mid x_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(t)})}{f(z_{a2,i}^{*(j)} \mid x_i, z_{a1,i}^{*(j)}; \hat{\theta}_{2(0)})} .$$

4. M-step: Parameters are updated by solving the imputed score equations:

$$\sum_{i \in A_a} w_i \sum_{j=1}^{m} w_{ij}^{*} S_1(\theta_1; x_i, z_{a1,i}^{*(j)}) + \sum_{i \in A_b} w_i \sum_{j=1}^{m} w_{ij}^{*} S_1(\theta_1; x_i, z_{a1,i}^{*(j)}) = 0$$

$$\sum_{i \in A_a} w_i \sum_{j=1}^{m} w_{ij}^{*} S_2(\theta_2; x_i, z_{a1,i}^{*(j)}, z_{a2,i}^{*(j)}) + \sum_{i \in A_b} w_i \sum_{j=1}^{m} w_{ij}^{*} S_2(\theta_2; x_i, z_{a1,i}^{*(j)}, z_{a2,i}^{*(j)}) = 0$$

where S_1 and S_2 are the score functions of θ_1 and θ_2, respectively.

The parameters in the measurement error models are updated by

$$\hat{\sigma}_{u1}^2 = \frac{\sum_{i \in A_b} w_i \sum_{j=1}^{m} w_{ij}^* (z_{b1,i} - z_{a1,i}^{*(j)})^2}{\sum_{i \in A_b} w_i}$$

and

$$\hat{\sigma}_{u2}^2 = \frac{\sum_{i \in A_b} w_i \sum_{j=1}^{m} w_{ij}^* (z_{b2,i} - z_{a2,i}^{*(j)})^2}{\sum_{i \in A_b} w_i}$$