2018년도 통계청 연구용역

행정자료 활용에 따른 가계금융복지조사 시계열 조정 등 방법론 연구

최종결과보고서

2018. 10. 30

통계청
Statistics Korea

제 출 문

제 출 문

통계청장 귀하

본 보고서를 **행정자료 활용에 따른 가계금융복지조사 시계열 조정 등 방법론 연구** 최종보고서로 제출합니다.

2018년 10월 30일

한국통계학회 ㊞

연구진

책 임 연 구 원	김재광, KAIST 수리과학과 교수
연 구 원	임종호, 연세대학교 응용통계학과 조교수
연 구 보 조 원	이단향, 아이오와 주립대학교 통계학과 박사과정생

최종결과보고서

차 례

<표 목차>

1. 서론

현재 가계금융복지조사는 2011년 기준 소득부터 공표 중이나 행정자료로 보완 가능한 시간은 2014-2016년으로 제한되어 있다. 행정자료를 활용하여 자료를 보완하는 것은 가계 소득을 보다 정확하게 집계한다는 취지에는 원칙적으로 바람직하지만 이를 바탕으로 통계청에서 발표하기에는 몇 가지 문제점이 발생한다.

첫 번째 문제점은 가계금융복지조사의 표본가구와 그 가구원들에게 100% 행정자료가 얻어지는 것이 아니라 실제로는 80-90%의 매칭 성공률로 인하여 매칭된 집단과 매칭되지 않은 집단의 소득값에 체계적인 차이가 발생할 수 있다는 것이다. 이러한 경우 매칭되지 않은 집단의 조사 소득값에 아무런 보정을 하지 않고 그대로 발표하게 되면 매칭 편향의 위험성을 제거하지 않은 것이 되고 또한 매칭되지 않는 집단의 소득이 과소하게 나타나는 경향이 있으므로 전체 소득이 과소 집계되는 문제점이 발생할수 있다.

두 번째 문제점은 행정자료를 활용하여 보완된 2014-2016년 가금복 조사 자료는 이를 활용하지 않은 2011-2013년 가금복 조사 자료와는 시계열 단절이 발생한다는 것이다. 이러한 시계열 단절이 발생하는 가장 근본적인 이유는 “금융소득”에 대한 행정자료가 2014년부터 활용 가능하기 때문이다. 따라서 이러한 시계열적 단절은 구조적인 문제가 아니라 집계 변화의 방법의 변화에 기인한 것이므로 이러한 문제점을 해결해야할 필요성이 발생한다.

또한 가계동향조사 자료의 과거 자료를 이용하여 가금복 자료의 2011년 이전에 대하여 전체 소득과 분배 지표에 대한 시계열을 구현이 가능한지에 대한 검토도 통계청의 실무진으로부터 제기되었다.

이러한 문제점 들을 해결하고 가능한 방법론을 탐색하기 위해 본 연구진은 6개월간 연구를 진행했는데 첫 번째 매칭 편향의 문제를 해결하기 위하여 매칭이 되지 않은 경우를 일종의 결측(missing) 문제로 환원하여 결측 자료의 대체 방법을 적용하여 이 문제를 해결하였다. 사용된 대체 방법론은 조사 소득값이 측정되는 경우와 측정되지 않는 경우에 따라 달라지는데 자료를 동질적인 그룹으로 나눈 후에 그 그룹 내에서 비대체(ratio imputation)을 하거나 회귀 대체(regression imputation)을 하는 방식으로 접근하였다. 또한 두 번째 문제의 경우 2011-2013년 가금복 조사로부터 동일한 형태의 소득분배지표를 얻기 위해서는 금융소득 부분에 대한 추정을 해야 하는데, 이는 2014-2016년도의 가금복 자료를 분석하여 금융소득에 대한 통계적 모형을 세우고 이를 바탕으로 금융소득을 가구별로 추정하여 집계하는 방법으로 달성될 수 있다. 이는 금융소득 변수가 일종의 무응답된 것으로 간주하고 그에 대한 통계적 예측값을 넣어주는 일종의 종단면 대량 대체(longitudinal mass

imputation) 방법론을 개발하는 것으로 이해할 수 있다.

본 최종보고서에서는 이러한 연구에 대한 분석 과정 및 결과를 정리하였는데 2장에서는 매칭 편향 보정에 대한 내용을 다루었고 3장에서는 금융소득에 대한 시계열 보완 방법론에 대한 내용을 다루었다. 4장에서는 가계동향조사 자료를 이용한 가금복 자료의 시계열 대체 분석에 대한 내용을 다루었다.

2. 매칭 편향 분석

2.1 연구 내용

본 연구에서는 향후 가계금융복지조사에서 일부 조사항목을 폐지하게 될 경우를 고려하여, 아래와 같이 두 가지 시나리오로 대체 방법론을 제안한다.

- 조사값이 있는 경우: 혼합 비(ratio) 대체
- 조사값이 없는 경우: 혼합 회귀(regression) 대체

특정 항목의 조사+행정값이 관심 변수라고 할 때, 해당 항목의 조사값을 사용할 수 있는 경우에는 조사값과 조사+행정값이 강한 양의 상관관계를 가지므로 조사값은 조사+행정값의 핵심 예측변수라 할 수 있다.

조사에서 소득값이 얻어지는 경우에는 조사 소득값이 조사+행정 소득값에 대한 가장 설명력 높은 변수이므로 이를 이용하여 비 대체(ratio imputation)을 실시하되 전체 표본을 몇 개의 그룹(또는 imputation cell)으로 나누어서 그 cell 내에서 비 대체를 실시하는 방식으로 구현하였다. 이러한 cell을 사용한 비 대체 (cell ratio imputation)을 위해서는 cell을 어떻게 만들 것인지가 핵심인데 이를 위하여 통계학의 혼합 모형(mixture model)을 사용하였다.

또한 조사 소득값이 얻어지지 않는 경우에 대한 대체법으로는 확장된 혼합 모형을 사용한 후 설명력 높은 설명 변수를 찾아서 이를 그룹별로 회귀 대체를 실시하는 방식으로 구현하였다.

2.1.1 혼합 모형을 이용한 비 대체법

2.1.1.1 혼합 비 대체 모형 1 (기본 모형)

1) 모형

관심 변수인 특정 항목의 조사+행정값을 y 라고 하고, 조사값을 $\tilde{y}$ 라고 하자. 즉, $\tilde{y}$ 는 y 의 핵심 예측변수이다. $\tilde{y}$ 가 주어졌을 때, y 에 대해 G 개의 혼합 성분을 가지는 일반 혼합 모형을 가정하면 다음과 같다.

$$f(y|\tilde{y}) = \sum_{g=1}^{G} P(z_g = 1|\tilde{y}) f_2(y|\tilde{y}, z_g = 1),$$

단, z_g는 그룹 g 의 지시변수로, 만약 y 가 그룹 g 에 속하면 $z_g = 1$, 그렇지 않으면 $z_g = 0$ 이다. ($g = 1,...,G$) 여기서,

$$P(z_g = 1|\tilde{y}) = \frac{f_1(\tilde{y}|z_g = 1)\pi_g}{\sum_{g=1}^{G} f_{1(}\tilde{y}|z_g = 1)\pi_g}$$

이고, $\pi_g = P(z_g = 1)$ 이다. 즉, $\pi_g = P(z_g = 1)$ 는 y 가 그룹 g 에 속할 사전확률을, $P(z_g = 1|\tilde{y})$ 는 조사값 $\tilde{y}$ 가 주어졌을 때, y 가 그룹 g 에 속할 사후확률을 나타낸다.

본 연구 과제에서는 구체적으로 다음과 같은 G 개의 혼합 성분을 가지는 모수적 혼합 비 대체 모형을 가정한다.

$$\begin{aligned} &\boldsymbol{z} \sim Multinomial(1, \boldsymbol{\pi}), \qquad (1)\\ &T(\tilde{y})|z_g = 1 \sim N(\mu_g, \sigma_g^2),\\ &y|(\tilde{y}, z_g = 1) \sim N(\tilde{y}\beta_g, \tilde{y}\sigma_{e,g}^2), \quad g = 1,...,G, \end{aligned}$$

단, $\boldsymbol{z} = (z_1,...,z_G)'$ 이고, $\boldsymbol{\pi} = (\pi_1,...,\pi_G)'$ 이다. 여기서, $T(\tilde{y})$ 는 그룹 내 정규분포를 따르도록 하는 $\tilde{y}$의 변환 변수로 $\log(\tilde{y})$ 또는 $\sqrt{\tilde{y}}$ 등이 있다.

비대체(ratio imputation)을 구현하기 위해서 모형 (1)에서 y 가 그룹 g 에 속할 때, y 의 평균이 $\tilde{y}\beta_g$ 이고 분산이 $\tilde{y}\sigma_{e,g}^2$ 인 정규분포를 가정했다. 이는 조사값이 커짐에 따라 y 의 분산 또한 커짐을 가정한다.

2) 모수 추정

일반적인 혼합 모형에서 그룹 변수 $\boldsymbol{z}=(z_1,...,z_G)$는 관측되지 않는 잠재 변수이므로, 이러한 경우 최대우도 추정방법(Maximum likelihood estimation method)을 구현하는 기댓값 최대화 알고리즘(Expectation-maximization algorithm; EM 알고리즘)을 이용하여 혼합 모형의 모수를 추정할 수 있다. 마찬가지로 본 연구과제에서도 모형 (1)의 모수를 추정하기 위해 EM 알고리즘을 이용한다.

모형 (1)의 모수를 $\theta=(\theta_1,\theta_2,\boldsymbol{\pi})$를 라고 표기하자. 단, $\theta_1=(\mu_g,\sigma_g^2)$, $\theta_2=(\beta_g,\sigma_{e,g}^2)$, 그리고 $\boldsymbol{\pi}=(\pi_1,....\pi_G)$이다. t번째 반복에서 얻은 모수 값을 $\theta^{(t)}$라고 할 때, EM 알고리즘의 기댓값 단계와 최대화 단계는 다음과 같다.

[기댓값 단계] $\theta^{(t)}$를 이용하여 다음과 같은 확률을 계산한다.

$$p_{ig}^{(t)}=P(z_{ig}=1|\,\widetilde{y_i},y_i;\theta^{(t)})$$

$$=\frac{\pi_g^{(t)}f_1(\widetilde{y_i}|z_{ig}=1;\theta_1^{(t)})f_2(y_i|\widetilde{y_i},z_{ig}=1;\theta_2^{(t)})}{\sum_{g=1}^{G}\pi_g^{(t)}f_1(\widetilde{y_i}|z_{ig}=1;\theta_1^{(t)})f_2(y_i|\widetilde{y_i},z_{ig}=1;\theta_2^{(t)})}$$

[최대화 단계] 모형 (1)의 우도함수를 최대로 하는 모수값을 계산하면 다음과 같다.

$$\mu_g^{(t+1)}=\frac{\sum_{i\in A}p_{ig}^{(t)}\delta_i T(\widetilde{y_i})}{\sum_{i\in A}p_{ig}^{(t)}\delta_i},$$

$$\sigma_g^{2(t+1)}=\frac{\sum_{i\in A}p_{ig}^{(t)}\delta_i T(\widetilde{y_i})^2}{\sum_{i\in A}p_{ig}^{(t)}\delta_i}-(\mu_g^{(t+1)})^2,$$

$$\beta_g^{(t+1)} = \frac{\sum_{i\in A} p_{ig}^{(t)} \delta_i y_i}{\sum_{i\in A} p_{ig}^{(t)} \delta_i \tilde{y_i}},$$

$$\sigma_{e,g}^{2(t+1)} = \frac{\sum_{i\in A} p_{ig}^{(t)} \delta_i \tilde{y_i}^{-1} (y_i - \beta_g^{(t+1)} \tilde{y_i})^2}{\sum_{i\in A} p_{ig}^{(t)} \delta_i},$$

그리고 $\pi_g^{(t+1)} = \left(\sum_{i\in A} \delta_i\right)^{-1} \left(\sum_{i\in A} p_{ig}^{(t)} \delta_i\right)$ 이다. 여기서 A 는 조사 가구원 인덱스 집합을 나타낸다. 알고리즘이 수렴할 때까지 [기댓값 단계]와 [최대화 단계]를 충분히 반복하여 최종 모수 추정값을 얻을 수 있다.

3) 대체 방법

EM 알고리즘을 이용하여 얻은 모수 추정값을 $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\pi})$ 라고 표기하자. 모수 추정값을 이용하여 다음과 같은 두 가지 대체 방법을 고려할 수 있다.

▫ 결정적 대체 (Deterministic imputation): 비매칭 집단의 j 번째 가구원의 조사값 $\tilde{y_j}$ 가 주어졌을 때 각 그룹 $g = 1,\ldots G$ 에 속할 확률을 다음과 같이 계산할 수 있다.

$$\hat{p}_{jg} = P(z_{jg} = 1 | \tilde{y_j}; \hat{\theta}) = \frac{\hat{\pi}_g f_1(\tilde{y_j} | z_{jg} = 1; \hat{\theta_1})}{\sum_{g=1}^{G} \hat{\pi}_g f_1(\tilde{y_j} | z_{jg} = 1; \hat{\theta}_1)},$$

그 다음, $\hat{p}_{j1},\ldots,\hat{p}_{jG}$ 를 각 그룹의 가중치로 하는 가중 평균 비 대체값을 다음과 같이 계산한다.

$$\hat{y}_j = \left(\sum_{g=1}^{G} \hat{p}_{jg} \hat{\beta}_g\right) \tilde{y_j}.$$

▫ 확률적 대체 (Stochastic imputation): 모수 추정값을 모형 (1)에 대입한 다음, $f(y_j | \tilde{y_j}; \hat{\theta}) = \sum_{g=1}^{G} \hat{p}_{jg} f(y_j | \tilde{y}_j, z_{jg} = 1; \hat{\theta})$로부터 랜덤 난수를 추출하여 대체값을 생성한다.

2.1.1.2 조건부 혼합 비 대체 모형 2 (확장된 모형)

1) 모형

관심 변수인 특정 항목의 조사+행정값을 y 라고 하고, 조사값을 $\tilde{y}$ 라고 하자. 혼합 비 대체 모형 1을 바탕으로 보조 변수 x의 정보를 통합하는 확장된 모형을 제안한다. 여기서, 보조 변수 x는 연령대, 교육정도 등의 인구학적 범주형 변수를 고려할 수 있다. x와 $\tilde{y}$ 가 주어졌을 때, y에 대해 G개의 혼합 성분을 가지는 일반 조건부 혼합 모형을 가정하면 다음과 같다.

$$f(y|\,x,\tilde{y}) = \sum_{g=1}^{G} P(z_g = 1|\,x,\tilde{y}) f_2(y|\tilde{y}, z_g = 1)$$

단, z_g는 그룹 g의 지시변수로, 만약 y가 그룹 g에 속하면 $z_g = 1$, 그렇지 않으면 $z_g = 0$ 이다. ($g = 1,...,G$) 여기서,

$$P(z_g = 1|\,x,\tilde{y}) = \frac{f_1(\tilde{y}|z_g = 1)\pi_g(x)}{\sum_{g=1}^{G} f_1(\tilde{y}|z_g = 1)\pi_g(x)}$$

이고, $\pi_g(x) = P(z_g = 1|\,x)$ 이다. 즉, $\pi_g(x)$는 x의 값에 따라 y가 그룹 g에 속할 확률을 다르게 가정한다. 만약, x를 3개의 범주를 가지는 보조 변수라고 하면, $\pi_g(x)$는 아래와 같은 조건부 확률 분포를 가진다.

g	$x=1$	$x=2$	$x=3$
1	$\pi_1(1)$	$\pi_1(2)$	$\pi_1(3)$
2	$\pi_2(1)$	$\pi_2(2)$	$\pi_2(3)$
⋮	⋮	⋮	⋮
G	$\pi_G(1)$	$\pi_G(2)$	$\pi_G(3)$
열 총합	1	1	1

본 연구 과제에서는 구체적으로 다음과 같은 G개의 혼합 성분을 가지는 모수적 조건부 혼합 비 대체 모형을 다음과 같이 가정한다.

$$\boldsymbol{z}|x \sim Multinomial(1, \boldsymbol{\pi}(\boldsymbol{x})), \qquad (2)$$

$$T(\tilde{y})|z_g = 1 \sim N(\mu_g, \sigma_g^2), \quad ,$$

$$y|(\tilde{y}, z_g = 1) \sim N(\tilde{y}\beta_g, \tilde{y}\sigma_{e,g}^2), \quad g = 1,...,G,$$

단, $\boldsymbol{z} = (z_1,...,z_G)'$ 이고, $\boldsymbol{\pi}(\boldsymbol{x}) = (\pi_1(x),...,\pi_G(x))'$ 이다.

2) 모수 추정

혼합 비 대체 모형 1에서와 마찬가지로 EM 알고리즘을 이용하여 모형 (2)의 모수를 추정한다. 모형 (2)의 모수를 $\theta = (\theta_1, \theta_2, \boldsymbol{\pi})$ 라고 표기하자. 단, 보조 변수 x 의 범주 개수가 K라고 할 때, $\theta_1 = (\mu_g, \sigma_g^2)$, $\theta_2 = (\beta_g, \sigma_{e,g}^2)$, 그리고 $\boldsymbol{\pi} = (\pi_1(1),....\pi_G(1),...\pi_1(K),...\pi_G(K))$이다. t번째 반복에서 얻은 모수 값을 $\theta^{(t)}$라고 할 때, EM 알고리즘의 기댓값 단계와 최대화 단계는 다음과 같다.

[기댓값 단계] $\theta^{(t)}$를 이용하여 다음과 같이 확률을 계산한다.

$$p_{ig}^{(t)} = P(z_g = 1|\, x_i, \tilde{y_i}, y_i; \theta^{(t)})$$

$$= \frac{\pi_g^{(t)}(x_i) f_1(\tilde{y_i}|\, z_{ig} = 1; \theta_1^{(t)}) f_2(y_i|\tilde{y_i}, z_{ig} = 1; \theta_2^{(t)})}{\sum_{g=1}^{G} \pi_g^{(t)}(x_i) f_1(\tilde{y_i}|\, z_{ig} = 1; \theta_1^{(t)}) f_2(y_i|\tilde{y_i}, z_{ig} = 1; \theta_2^{(t)})}$$

[최대화 단계] $\theta_1 = (\mu_g, \sigma_g^2)$과 $\theta_2 = (\beta_g, \sigma_{e,g}^2)$는 혼합 비 대체 모형 1 에서와 동일하게 계산할 수 있다. 혼합 확률은 다음과 같이 계산한다.

$$\pi_g^{(t+1)}(x) = \left(\sum_{i \in A} \delta_i I(x_i = x)\right)^{-1} \left(\sum_{i \in A} p_{ig}^{(t)} \delta_i I(x_i = x)\right),$$

여기서 A는 조사 가구원 인덱스 집합을 나타낸다. 알고리즘이 수렴할 때까지 [기댓값 단계]와 [최대화 단계]를 충분히 반복하여 최종 모수 추정값을 얻을 수 있다.

3) 대체 방법

EM 알고리즘을 이용하여 얻은 모수 추정값을 $\hat{\theta}=(\hat{\theta}_1, \hat{\theta}_2, \hat{\pi})$라고 표기하자. 모수 추정값을 이용하여 다음과 같은 두 가지 대체 방법을 고려할 수 있다.

- 결정적 대체 (Deterministic imputation): 비매칭 집단의 j번째 가구원의 보조변수 x_j와 조사값 $\tilde{y}_j$가 주어졌을 때, 각 그룹 $g=1,\dots G$에 속할 확률을 다음과 같이 계산한다.

$$\hat{p}_{jg}=P(z_{jg}=1|x_j,\tilde{y}_j;\hat{\theta})=\frac{\hat{\pi}_g(x_j)f_1(\tilde{y}_j|z_{jg}=1;\hat{\theta}_1)}{\sum_{g=1}^{G}\hat{\pi}_g(x_j)f_1(\tilde{y}_j|z_{jg}=1;\hat{\theta}_1)}$$

그 다음, $\hat{p}_{j1},\dots,\hat{p}_{jG}$를 각 그룹의 가중치로 하는 가중 평균 비 대체값을 다음과 같이 계산한다.

$$\hat{y}_j=\left(\sum_{g=1}^{G}\hat{p}_{jg}\hat{\beta}_g\right)\tilde{y}_j.$$

- 확률적 대체 (Stochastic imputation): 모수 추정값을 모형 (2)에 대입한 다음, $f(y_j|\tilde{y}_j;\hat{\theta})=\sum_{g=1}^{G}\hat{p}_{jg}f(y_j|\tilde{y}_j,z_{jg}=1;\hat{\theta})$로부터 랜덤 난수를 추출하여 대체값을 생성한다.

2.1.2. 혼합 모형을 이용한 회귀 대체법

2.1.2.1. 혼합 회귀 대체 모형 1 (기본 모형)

1) 모형

관심 변수인 특정 항목의 조사+행정값을 y라고 하고, $\tilde{x}$를 p개의 예측 변수(열 벡터)라고 하자. 핵심 예측 변수인 조사값 ($\tilde{y}$)이 없는 경우에는, 그룹(셀) 내에서 비 대체 대신 회귀 대체를 적용할 수 있다. $\tilde{x}$가 주어졌을 때, y에 대해 G개의 혼합 성분을 가지는 일반 혼합 모형을 가정하면, 다음과 같다.

$$f(y|\tilde{x})=\sum_{g=1}^{G}P(z_g=1|\tilde{x})f_2(y|\tilde{x},z_g=1)$$

단, $P(z_g=1|\tilde{x})=\dfrac{f_1(\tilde{x}|z_g=1)\pi_g}{\sum_{g=1}^{G} f_1(\tilde{x}|z_g=1)\pi_g}$ 이고, $\pi_g=P(z_g=1)$ 이다. 마찬가지로, z_g는 그룹 g의 지시변수로, 만약 y가 그룹 g에 속하면 $z_g=1$, 그렇지 않으면 $z_g=0$ 이다.

본 연구 과제에서는 구체적으로 다음과 같은 모수적 혼합 회귀 대체 모형을 가정한다.

$$\begin{aligned} &\boldsymbol{z} \sim Multinomial(1, \boldsymbol{\pi}), \\ &T(\tilde{x})|z_g=1 \sim N(\mu_g, \Sigma_g), \\ &y|(\tilde{x}, z_g=1) \sim N((1,\tilde{x}')\beta_g, \sigma^2_{e,g}) \end{aligned} \tag{3}$$

단, $\boldsymbol{z}=(z_1,...,z_G)'$ 이고, $\boldsymbol{\pi}=(\pi_1,...,\pi_G)'$ 이다. 여기서 $T(\tilde{x})$는 그룹 내 다변량 정규 분포를 따르도록 하는 $\tilde{x}$의 변환 변수로 $\log(\tilde{x})$ 또는 $\sqrt{\tilde{x}}$ 등이 있다.

2) 모수 추정

혼합 비 대체 모형 1과 2에서와 마찬가지로 그룹 변수 $\boldsymbol{z}=(z_1,...,z_G)$가 잠재 변수이므로, EM 알고리즘을 이용하여 모수를 추정할 수 있다. 모형 (3)의 모수를 $\theta=(\theta_1, \theta_2, \boldsymbol{\pi})$라고 표기하자. 단, 모형 (3)에서는 $\theta_1=(\mu_g,\Sigma_g)$, $\theta_2=(\beta_g,\sigma^2_{e,g})$, 그리고 $\boldsymbol{\pi}=(\pi_1,....\pi_G)$ 이다. t번째 반복에서 얻은 모수 값을 $\theta^{(t)}$라고 할 때, EM 알고리즘의 기댓값 단계와 최대화 단계는 다음과 같다.

[기댓값 단계] $\theta^{(t)}$를 이용하여 다음과 같은 확률을 계산한다.

$$\begin{aligned} p_{ig}^{(t)} &= P(z_{ig}=1|\,\tilde{x_i}, y_i; \theta^{(t)}) \\ &= \frac{\pi_g^{(t)} f_1(\tilde{x_i}|z_{ig}=1;\theta_1^{(t)}) f_2(y_i|\tilde{x_i}, z_{ig}=1;\theta_2^{(t)})}{\sum_{g=1}^{G}\pi_g^{(t)} f_1(\tilde{x_i}|z_{ig}=1;\theta_1^{(t)}) f_2(y_i|\tilde{x_i}, z_{ig}=1;\theta_2^{(t)})} \end{aligned}$$

[최대화 단계] 가정한 혼합 모형의 우도함수를 최대로 하는 모수값을 계산한다. 모형 (3)에서는 다음과 같이 계산할 수 있다.

$$\mu_g^{(t+1)} = \frac{\sum_{i\in A} p_{ig}^{(t)} \delta_i T(\widetilde{x_i})}{\sum_{i\in A} p_{ig}^{(t)} \delta_i},$$

$$\Sigma_g^{(t+1)} = \frac{1}{\sum_{i\in A} p_{ig}^{(t)} \delta_i} \sum_{i\in A} p_{ig}^{(t)} \delta_i (T(\widetilde{x_i}) - \mu_g)(T(\widetilde{x_i}) - \mu_g)',$$

$$\beta_g^{(t+1)} = \left(\sum_{i\in A} p_{ig}^{(t)} \delta_i (1, \widetilde{x_i}')'(1, \widetilde{x_i}')\right)^{-1} \left(\sum_{i\in A} p_{ig}^{(t)} \delta_i (1, \widetilde{x_i}')' y_i\right),$$

$$\sigma_{e,g}^{2(t+1)} = \frac{\sum_{i\in A} p_{ig}^{(t)} \delta_i (y_i - (1, \widetilde{x_i}')\beta_g^{(t+1)})^2}{\sum_{i\in A} p_{ig}^{(t)} \delta_i},$$

그리고 $\pi_g^{(t+1)} = \left(\sum_{i\in A} \delta_i\right)^{-1} \left(\sum_{i\in A} p_{ig}^{(t)} \delta_i\right)$ 이다. 여기서 A 는 관측된 조사 가구원의 인덱스 집합을 나타낸다. 알고리즘이 수렴할 때까지 [기댓값 단계]와 [최대화 단계]를 충분히 반복하여 최종 모수 추정값을 얻을 수 있다.

3) 대체 방법

EM 알고리즘을 이용하여 얻은 모수 추정값을 $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\boldsymbol{\pi}})$ 라고 표기하자. 모수 추정값을 이용하여 다음과 같은 두 가지 대체 방법을 고려할 수 있다.

▫ 결정적 대체 (Deterministic imputation): 비매칭 집단의 j 번째 가구원의 조사값 $\widetilde{y_j}$ 가 주어졌을 때 각 그룹 $g = 1, \ldots G$ 에 속할 확률을 계산한 후, 이를 가중치로 하는 가중 평균 회귀 대체값을 다음과 같이 계산한다.

$$\hat{p}_{jg} = P(z_{jg} = 1 | \widetilde{x_j}; \hat{\theta}) = \frac{\hat{\pi}_g f_1(\widetilde{x_j} | z_{jg} = 1; \hat{\theta}_1)}{\sum_{g=1}^{G} \hat{\pi}_g f_1(\widetilde{x_j} | z_{jg} = 1; \hat{\theta}_1)},$$

$$\hat{y}_j = \sum_{g=1}^{G} \hat{p}_{jg}(1, \tilde{x}_j')\hat{\beta}_g.$$

▫ 확률적 대체 (Stochastic imputation): 모수 추정값을 모형 (3)에 대입한 다음, $f(y_j|\tilde{y}_j;\hat{\theta}) = \sum_{g=1}^{G} \hat{p}_{jg} f(y_j|\tilde{y}_j, z_{jg}=1;\hat{\theta})$로부터 랜덤 난수를 추출하여 대체값을 생성한다.

본 최종보고서에서는 결정적 대체 방법을 이용하여 대체한 결과를 제시한다.

2.1.2.2. 조건부 혼합 회귀 대체 모형 2 (확장된 모형)

1) 모형

관심 변수인 특정 항목의 조사+행정값을 y라고 하고, $\tilde{x}$를 p개의 예측 변수(열 벡터)라고 하자. 혼합 회귀 대체 모형 1을 바탕으로 보조변수 x의 정보를 통합하는 확장된 모형을 제안한다. 여기서, 보조 변수 x는 연령대, 교육정도 등의 인구학적 범주형 변수를 고려할 수 있다. x와 $\tilde{x}$가 주어졌을 때, y에 대해 G개의 혼합 성분을 가지는 일반 조건부 혼합 모형을 가정하면, 다음과 같다.

$$f(y|x,\tilde{x}) = \sum_{g=1}^{G} P(z_g=1|x,\tilde{x}) f_2(y|\tilde{x}, z_g=1)$$

단, $P(z_g=1|x,\tilde{x}) = \dfrac{f_1(\tilde{x}|z_g=1)\pi_g(x)}{\sum_{g=1}^{G} f_1(\tilde{x}|z_g=1)\pi_g(x)}$이고, $\pi_g(x) = P(z_g=1|x)$이다. 마찬가지로, z_g는 그룹 g의 지시변수로, 만약 y가 그룹 g에 속하면 $z_g=1$, 그렇지 않으면 $z_g=0$이다. ($g=1,...,G$)

본 연구 과제에서는 구체적으로 G개의 혼합 성분을 가지는 모수적 조건부 혼합 회귀 대체 모형을 다음과 같이 가정한다.

$$\boldsymbol{z}|x \sim Multinomial(1, \boldsymbol{\pi}(\boldsymbol{x})), \qquad (4)$$

$$T(\tilde{x})|z_g = 1 \sim N(\mu_g, \Sigma_g),$$

$$y|(\tilde{x}, z_g = 1) \sim N((1,\tilde{x}')\beta_g, \sigma^2_{e,g})$$

단, $\boldsymbol{z} = (z_1,...,z_G)'$ 이고, $\boldsymbol{\pi}(\boldsymbol{x}) = (\pi_1(x),...,\pi_G(x))'$ 이다.

2) 모수 추정

혼합 회귀 대체 모형 1에서와 마찬가지로 EM 알고리즘을 이용하여 모형 (4)의 모수를 추정한다. 모형 (4)의 모수를 $\theta = (\theta_1, \theta_2, \boldsymbol{\pi})$ 라고 표기하자. 단, 보조 변수 x 의 범주 개수가 K 라고 할 때, $\theta_1 = (\mu_g, \Sigma_g)$, $\theta_2 = (\beta_g, \sigma^2_{e,g})$, 그리고 $\boldsymbol{\pi} = (\pi_1(1),....\pi_G(1),...\pi_1(K),...\pi_G(K))$ 이다. t 번째 반복에서 얻은 모수 값을 $\theta^{(t)}$ 라고 할 때, EM 알고리즘의 기댓값 단계와 최대화 단계는 다음과 같다.

[기댓값 단계] $\theta^{(t)}$ 를 이용하여 다음과 같이 확률을 계산한다.

$$p_{ig}^{(t)} = P(z_g = 1|\, x_i, \tilde{x_i}, y_i; \theta^{(t)})$$

$$= \frac{\pi_g^{(t)}(x_i) f_1(\tilde{x_i}|\, z_{ig} = 1; \theta_1^{(t)}) f_2(y_i|\tilde{x_i}, z_{ig} = 1; \theta_2^{(t)})}{\sum_{g=1}^{G} \pi_g^{(t)}(x_i) f_1(\tilde{x_i}|\, z_{ig} = 1; \theta_1^{(t)}) f_2(y_i|\tilde{x_i}, z_{ig} = 1; \theta_2^{(t)})}$$

[최대화 단계] $\theta_1 = (\mu_g, \Sigma_g)$ 과 $\theta_2 = (\beta_g, \sigma^2_{e,g})$ 는 혼합 회귀 대체 모형 1 에서와 동일하게 계산할 수 있다. 혼합 확률은 다음과 같이 계산한다.

$$\pi_g^{(t+1)}(x) = \left(\sum_{i \in A} \delta_i I(x_i = x)\right)^{-1} \left(\sum_{i \in A} p_{ig}^{(t)} \delta_i I(x_i = x)\right),$$

여기서 A 는 조사 가구원 인덱스 집합을 나타낸다. 알고리즘이 수렴할 때까지 [기댓값 단계]와 [최대화 단계]를 충분히 반복하여 최종 모수 추정값을 얻을 수 있다.

3) 대체 방법

EM 알고리즘을 이용하여 얻은 모수 추정값을 $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\boldsymbol{\pi}})$ 라고 표기하자. 모수

추정값을 이용하여 다음과 같은 두 가지 대체 방법을 고려할 수 있다.

▫ 결정적 대체 (Deterministic imputation): 비매칭 집단의 j번째 가구원의 보조변수 x_j와 조사값 $\tilde{y}_j$가 주어졌을 때, 각 그룹 $g=1,\dots G$에 속할 확률을 계산한 후 이를 가중치로 하는 가중 평균 회귀 대체값을 다음과 같이 계산한다.

$$\hat{p}_{jg}=P(z_{jg}=1|x_j,\tilde{x}_j;\hat{\theta})=\frac{\hat{\pi}_g(x_j)f_1(\tilde{x}_j|z_{jg}=1;\hat{\theta}_1)}{\sum_{g=1}^{G}\hat{\pi}_g(x_j)f_1(\tilde{x}_j|z_{jg}=1;\hat{\theta}_1)},$$

$$\hat{y}_j=\sum_{g=1}^{G}\hat{p}_{jg}(1,\tilde{x}_j')\hat{\beta}_g.$$

▫ 확률적 대체 (Stochastic imputation): 모수 추정값을 모형 (4)에 대입한 다음, $f(y_j|\tilde{y}_j;\hat{\theta})=\sum_{g=1}^{G}\hat{p}_{jg}f(y_j|\tilde{y}_j,z_{jg}=1;\hat{\theta})$로부터 랜덤 난수를 추출하여 대체값을 생성한다.

본 최종보고서에서는 결정적 대체 방법을 이용하여 대체한 결과를 제시한다.

2.1.3. 모형 선택

1) G 선택

혼합 대체 모형에서 혼합 성분 개수 G는 혼합 모형의 모형 복잡성(model complexity)를 반영하는 모수이다. 즉, G가 클수록 편의(bias)는 감소하나 분산(variance)가 증가하는 반면, G가 작을수록 편의는 증가하나 분산이 감소하게 된다. 이러한 G를 선택하는 방법으로 k-fold 교차 검증법(k-fold cross-validation)을 이용하여 평균 제곱근 오차 (Root Mean Squared Error)를 최소로 하는 G를 선택한다. k-fold 교차 검증법 절차는 다음과 같다.

[단계 1] 관측한 자료를 크기가 비슷한 k개의 집단으로 랜덤으로 분할한다. (k-fold)

[단계 2] 첫 번째 집단을 검증 데이터 (test data)로, 나머지 $k-1$개의 집단을 학습 데이터(training data)로 간주하여 혼합 모형의 모수를 추정한 후 검증 데이터의 관심 변수를 예측한 다음 평균 제곱근 오차(Root Mean Squared

Error; RMSE)를 계산한다.

[단계 3] 검증 데이터를 달리 하면서 [단계 2]를 k번 반복한다. 즉, k개의 집단이 정확하게 한 번씩 검증 데이터가 되도록 하면서 [단계 2]를 반복한다.

[단계 4] k개의 평균 제곱근 오차(RMSE)의 평균을 계산한다. (본 최종보고서에서는 이 값을 평균 RMSE라 칭하기로 한다.)

각 $G=1,...,G_0$에 따라 k-fold 교차 검증법 절차를 통해 평균 RMSE를 계산하여 G_0의 평균 RMSE 중 가장 작은 값을 가지는 G를 선택한다. 본 연구 과제에서는 $k=10$으로, $G_0=5$로 두고 자료 분석을 진행하였다.

2) 임의 결측 (Missing at random; MAR) 검정

본 연구 과제에서는 매칭 집단과 비매칭 집단 간 편향을 일으키는 잠재적인 인구학적 변수 x_b를 이용하여, x_b가 주어졌을 때 매칭 집단과 비매칭 집단 간의 조사+행정값 y의 조건부 분포가 동일한지를 확인하는 절차를 수행한다.

비매칭 집단에서는 조사+행정값이 사용할 수 없으므로 이를 결측이라 간주할 수 있다. 그러면, 임의 결측 (MAR) 가정은 주요 인구학적 변수 x_b가 주어졌을 때, 매칭(응답) 집단과 비매칭(무응답) 집단 간의 y의 조건부 분포가 거의 동일함을 의미한다.

혼합 비/회귀 대체 모형 1(기본 모형)과 조건부 혼합 비/회귀 대체 모형 2(확장된 모형)에서 얻어진 대체값이 MAR 가정을 심각하게 위반하는지 판단하기 위해 다음과 같은 통계량을 비교할 수 있다.

▫ X^2- 통계량

[단계 1] 매칭 집단에서 조사+행정값 y와 비매칭 집단에서 대체값을 범주형 변수(y_c)로 변환한다. (예: 조사 및 조사+행정 근로소득을 이용하여 표본 사분위수를 구한 후, 이를 기준으로 4구간 근로소득을 범주형 변수로 정의한다.)

[단계 2] 범주형 변수 $x_b=k$일 때 ($k=1,...,K$), 매칭 여부와 y_c의 독립성을 검정하는 카이제곱 통계량을 계산한 후 다음과 같은 가중 합을 계산한다.

$$X^2=\sum_{x=1}^{K} p_x(k)\chi^2(k)$$

단, $p_x(k) = \sum_{i \in A} I(x_i = k)/n$이고, n은 전체 조사 가구원 수이다.

2.2 연구 결과

본 최종보고서는 2017년 조사 가구원을 분석 대상으로 선정하여 분석 과정 및 결과를 기술하기로 한다. 다른 조사 연도에서도 아래와 같은 자료 분석 절차를 그대로 적용할 수 있다.

2.2.1. 기본 내용

<표 2-1>은 2017년 가계금융복지조사의 가구원 단위 매칭 · 비매칭 빈도표이다. 조사 자료와 행정 자료의 매칭 성공률이 88%이므로, 매칭에 성공한 가구원의 경우 조사+행정 자료를 사용할 수 있지만, 매칭에 실패한 가구원의 경우 조사 자료만 사용 가능하다. 만약 매칭된 가구원은 조사+행정 자료를 사용하고 비매칭된 가구원은 조사 자료를 사용한다면, 두 집단 간 체계적 차이가 발생할 수 있다. 2.1절에서는 이러한 매칭으로 인한 편향을 보정하여 비매칭 가구원의 조사+행정 자료의 대체값을 생성하는 대체 방법론을 개발하였고, 이 절에서는 다음과 같은 소득과 부채 항목에 대하여 개발한 대체 방법론을 적용하고자 한다.

◦ 경상 소득 = 근로 + 사업 + 재산 + 공적이전소득 + 사적이전소득
◦ 금융 부채 = 담보 + 신용 + 카드 + 외상 및 할부 + 사인 간 거래

<표 2-1. 2017년 조사 가구원 단위 매칭 · 비매칭 빈도표>

	가구원 수	백분율 (%)
매칭	43813	88
비매칭	5705	12
전체	49518	100

2.2.2. 자료 현황

2017년 가계금융복지조사의 관심 항목인 소득과 부채의 자료 현황을 먼저 살펴보기로 한다.

2.2.2.1. 소득 세부 항목별 기초 자료 분석

1) 근로소득

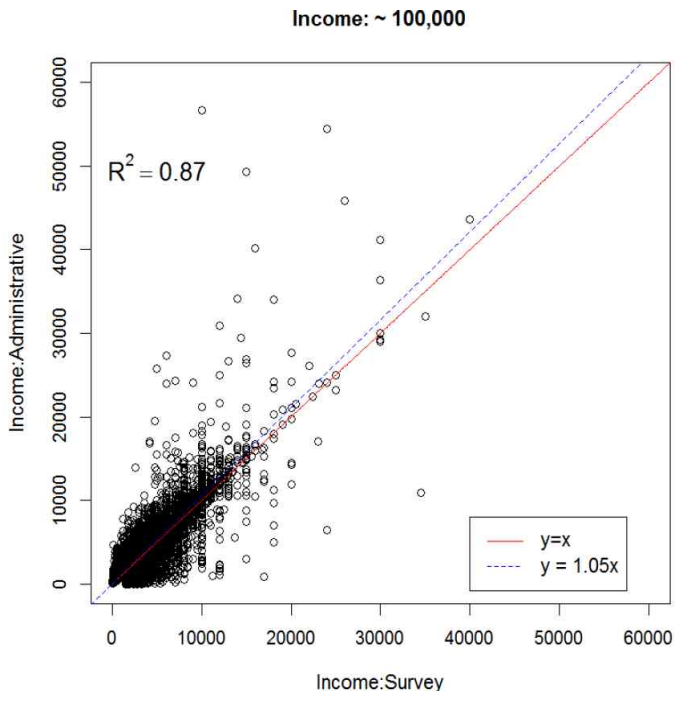
본 연구에서는 조사 근로소득이 0보다 큰 가구원을 근로소득이 있는 가구원으로 정의하고 자료 분석 대상으로 간주한다. <표 2-2>는 2017년 조사 근로소득 유·무와 매칭 집단별 이원 빈도표를 나타낸다. 전체 조사 가구원 49,518명 중 조사 근로소득이 0보다 큰 가구원은 17,711명이며 이를 대상으로 조사 근로소득과 조사+행정 근로소득의 자료를 살펴보면 다음과 같다.

<표 2-2. 근로소득 유·무와 매칭 여부의 이원 빈도표>

		매칭	비매칭	행 합
근로소득	유	15664	2047	17711
	무	28149	3658	31807
열 합		43813	5705	49518

<그림 2-1>은 2017년 조사된 근로소득이 있는 가구원 중 일부 (근로소득이 100,000만원 미만인 가구원)를 대상으로 그린 조사 근로소득과 조사+행정 근로소득의 산점도이다. 조사 근로소득은 조사+행정 근로소득과 강한 양의 선형 관계를 보이지만, 자료가 y=x 선을 기준으로 다소 퍼져있고 조사 근로소득이 커짐에 따라 자료가 퍼져있는 정도, 즉, 분산 또한 커지는 경향을 보인다.

<그림 2-1. 조사 vs. 조사+행정 근로소득 산점도>



<표 2-3>은 2017년 근로소득 대상자 중 매칭 집단 별 조사 근로소득 및 조사+행정 근로소득 요약 통계량을 보여준다. 이미 <그림 2-1>에서 확인했듯이, 매칭 집단의 조사 근로소득과 조사+행정 근로소득의 요약 통계량이 다소 차이가 있음을 알 수 있다. 조사+행정 근로소득에서 조사 근로소득에 비해 평균 및 3사분위수가 높은 반면, 1사분위수와 2사분위수는 오히려 낮다.

<표 2-3. 근로소득 대상자의 매칭 집단별 조사 및 조사+행정 근로소득 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	15664	1440	2400	4000	3145
	조사+행정		1200	2228	4220	3199
비매칭	조사	2047	1500	2400	3710	2929
	조사+행정		-	-	-	-

2) 금융소득

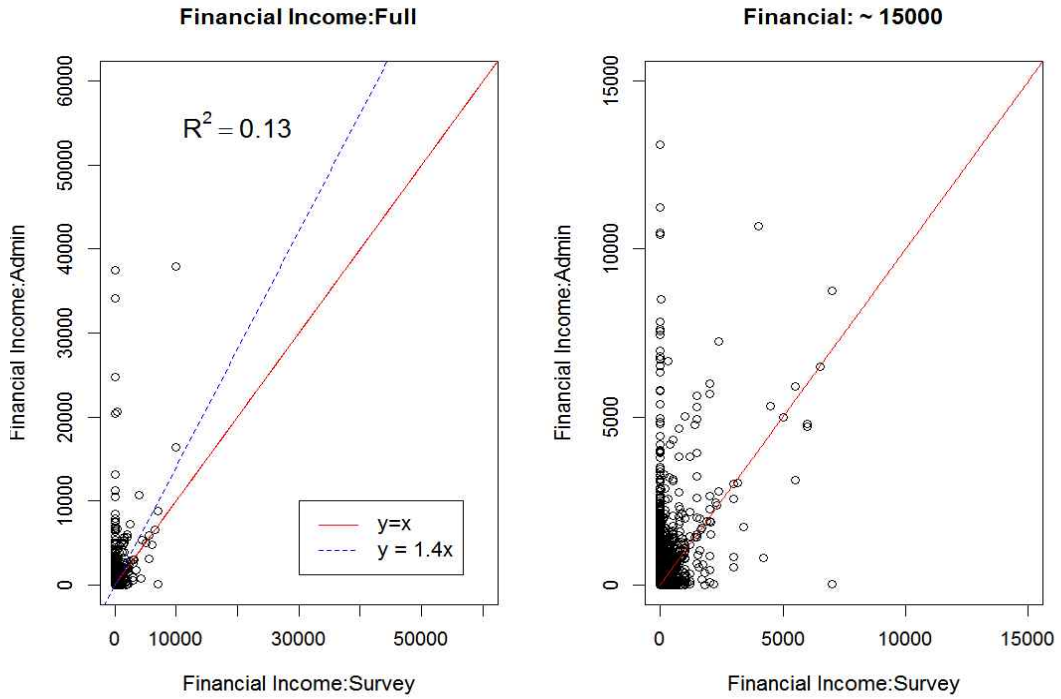
금융소득은 근로소득과 달리, 금융소득이 있는 대상자를 정의하는 기준이 모호하다. 따라서 2017년 조사에 응한 모든 가구원을 대상으로 아래와 같이 금융소득 자료를 정리 및 요약한다. 매칭 집단 별로 조사 금융소득이 0인 가구원 집단과 0보다 큰 가구원 집단으로 이분화하여 정리하면 <표 2-4>와 같다. 매칭 집단에서 조사 금융소득이 0인 가구원은 매칭된 가구원의 93.3%에 해당하고, 그 중 조사+행정 금융소득 또한 0인 가구원은 41.8%, 조사+행정 금융소득이 0보다 큰 가구원은 51.5%이다. 즉, 실제 조사+행정 금융소득이 있으나 조사에서 금융소득이 없다고 응답한 가구원은 51.5%로 비중이 다소 높은 편이다. 비매칭 집단에서는 조사 금융소득이 0이라고 응답한 가구원이 96.2%에 해당한다.

<표 2-4. 매칭 집단별 조사 및 조사+행정 금융소득 0 값 빈도>

			조사 금융소득	
			=0	>0
매칭	조사+행정 금융소득	=0	18325 (41.8%)	123 (0.3%)
		>0	22551 (51.5%)	2814 (6.4%)
	열 총합		40876 (93.3%)	2937 (6.7%)
비매칭	열 총합		5490 (96.2%)	215 (3.8%)

<그림 2-2>는 조사 금융소득과 조사+행정 금융소득의 산점도를 전체 조사 가구원 대상과 일부 가구원 대상 (금융소득이 15,000만원 미만인 가구원)으로 보여준다. 근로소득과는 달리 뚜렷한 선형관계를 보이지는 않으며, 조사 금융소득이 0이고, 조사+행정 금융소득이 0보다 큰 가구원의 분포를 그림에서 확인할 수 있다.

<그림 2-2. 조사 vs. 조사+행정 금융소득 산점도: 전체, 부분>



<표 2-5. 매칭 집단별 조사 및 조사+행정 금융소득 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	11.9
	조사+행정		0	1	28.0	65.3
비매칭	조사	5705	0	0	0	6.1
	조사+행정		–	–	–	–

<표 2-5>는 2017년 모든 조사 가구원을 대상으로 매칭 집단 별 조사 금융소득 및 조사+행정 금융소득 요약 통계량을 제시한다. 매칭 집단을 보면 조사 금융소득과 조사+행정 금융소득의 요약 통계량이 3사분위수와 평균에서 차이 나는 것을 확인할 수 있다.

3) 임대소득

금융소득과 마찬가지로 2017년 조사에 응한 모든 가구원을 대상으로 임대소득 자

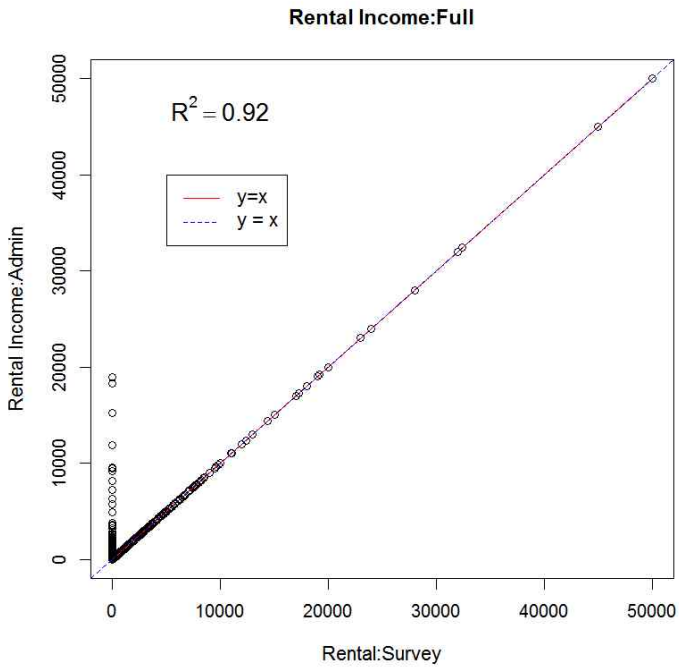
료를 정리하면 다음과 같다. <표 2-6>을 보면 알 수 있듯이, 매칭된 조사 가구원 중 조사 임대소득이 0인 가구원이 94.2%이고, 그 중 조사+행정 임대소득이 0인 가구원이 93.8%로 상당히 높은 비중을 차지한다. 비매칭 집단에서 조사 임대소득이 0인 가구원은 96%이다.

<그림 2-3>을 보면 조사 임대소득과 조사+행정 임대소득이 거의 일치함을 확인할 수 있으며, 조사 임대소득이 0이면서 조사+행정 임대소득이 0보다 큰 가구원의 분포 또한 그림에서 확인할 수 있다.

<표 2-6. 매칭 집단별 조사 및 조사+행정 임대소득 0값 빈도>

			조사 임대소득	
			=0	>0
매칭	조사+행정 임대소득	=0	41107 (93.8%)	0 (0.0%)
		>0	190 (0.4%)	2516 (5.7%)
	열 총합		41297 (94.2%)	2516 (5.7%)
비매칭	열 총합		5478 (96%)	227 (4%)

<그림 2-3. 조사 vs. 조사+행정 임대소득 산점도>



<표 2-7>은 2017년 모든 조사 가구원을 대상으로 매칭 집단 별 조사 임대소득 및 조사+행정 임대소득 요약 통계량을 보여준다. 임대소득 자료 특성상 0인 가구원이 많음으로 인해 조사 및 조사+행정 임대소득의 1,2, 그리고 3사분위 값이 모두 0이다. 매칭 집단에서 조사+행정 임대소득의 평균이 조사 임대소득보다 높지만 요약 통계량의 차이가 크지 않음을 알 수 있다.

<표 2-7. 매칭 집단별 조사 및 조사+행정 임대소득 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	70.0
	조사+행정		0	0	0	76.1
비매칭	조사	5705	0	0	0	45.8
	조사+행정		–	–	–	–

4) 공적연금

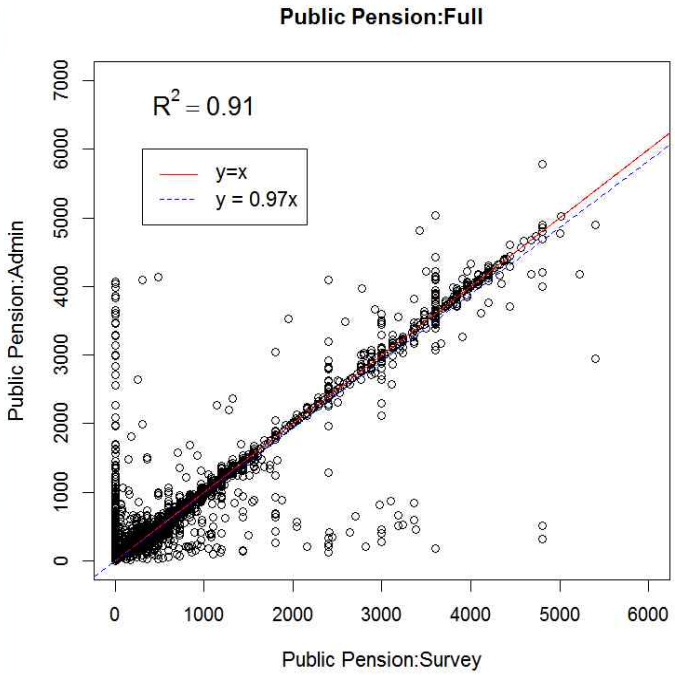
<표 2-8>을 보면 알 수 있듯이, 매칭된 조사 가구원 중 조사 공적연금이 0인 가구원이 90.6%이고, 그 중 조사+행정 공적연금이 0인 가구원이 88.3%로 꽤 높은 비중을 차지한다. 비매칭 집단에서 조사 공적연금이 0인 가구원은 94.4%이다.

<표 2-8. 매칭 집단별 조사 및 조사+행정 공적연금 0 값 빈도>

			조사 공적연금	
			=0	>0
매칭	조사+행정 공적연금	=0	38674 (88.3%)	0 (0.0%)
		>0	1037 (2.4%)	4102 (9.4%)
	열 총합		39711 (90.6%)	4102 (9.4%)
비매칭	열 총합		5386 (94.4%)	319 (5.6%)

<그림 2-4>를 보면 조사 공적연금과 조사+행정 공적연금의 상관관계가 높은 편이나 y=x 선으로부터 거리가 먼 자료들의 분포 및 조사 공적연금이 0이나 조사+행정 공적연금이 0보다 큰 가구원의 분포를 그림에서 확인할 수 있다.

<그림 2-4. 조사 vs. 조사+행정 공적연금 산점도>



<표 2-9>에서는 2017년 모든 조사 가구원을 대상으로 매칭 집단 별 조사 공적연금 및 조사+행정 공적연금 요약 통계량을 보여준다. 공적연금의 자료 특성상 0인 가구원이 많음으로 인해 조사 공적연금과 조사+행정 공적연금의 1,2, 그리고 3사분위 값이 모두 0이나, 매칭 집단에서 조사+행정 공적연금의 평균이 조사 공적연금의 평균보다 높다.

<표 2-9. 매칭 집단별 조사 및 조사+행정 공적연금 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	76.9
	조사+행정		0	0	0	84.9
비매칭	조사	5705	0	0	0	41.3
	조사+행정		–	–	–	–

5) 기초연금

<표 2-10>은 매칭 집단별 조사 및 조사+행정 기초연금의 0값 빈도를 보여준다. 매칭된 조사 가구원 중 조사 기초연금이 0인 가구원이 88.4%이고, 그 중 조사+행정 기초연금이 0인 가구원이 87.8%로 상당히 높은 비중을 차지한다. 비매칭 집단에서 조사 기초연금이 0인 가구원은 91.1%이다.

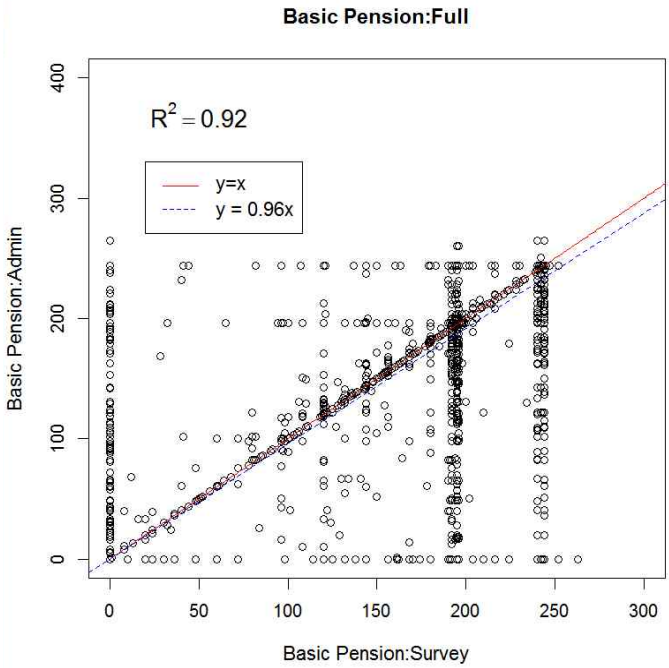
<그림 2-5>를 보면 조사 기초연금과 조사+행정 기초연금의 상관관계가 높은 편이긴 하나 y=x 선을 기준으로 퍼져있는 형태가 비선형이며, 특정 조사 기초연금 값

(200만원 혹은 250만원 근처)에서는 조사+행정 기초연금이 무작위로 퍼져있는 형태 또한 볼 수 있다.

<표 2-10. 매칭 집단별 조사 및 조사+행정 기초연금 0 값 빈도>

			조사 기초연금	
			=0	>0
매칭	조사+행정 기초연금	=0	38477 (87.8%)	159 (0.4%)
		>0	272 (0.6%)	4905 (11.2%)
	열 총합		38749 (88.4%)	5064 (11.6%)
비매칭	열 총합		5198 (91.1%)	507 (8.9%)

<그림 2-5. 조사 vs. 조사+행정 기초연금 산점도>



임대소득, 공적연금과 마찬가지로 자료 특성 상 0값이 많음으로 인해 매칭된 가구원 중 조사 기초연금과 조사+행정 기초연금의 1,2, 그리고 3사분위수가 모두 0임을 <표 2-11>에서 확인할 수 있다. 조사 기초연금의 평균과 조사+행정 기초연금의 평균 또한 거의 유사하다.

<표 2-11. 매칭 집단별 조사 및 조사+행정 기초연금 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	24.44
	조사+행정		0	0	0	24.39
비매칭	조사	5705	0	0	0	19.43
	조사+행정		–	–	–	–

6) 양육수당

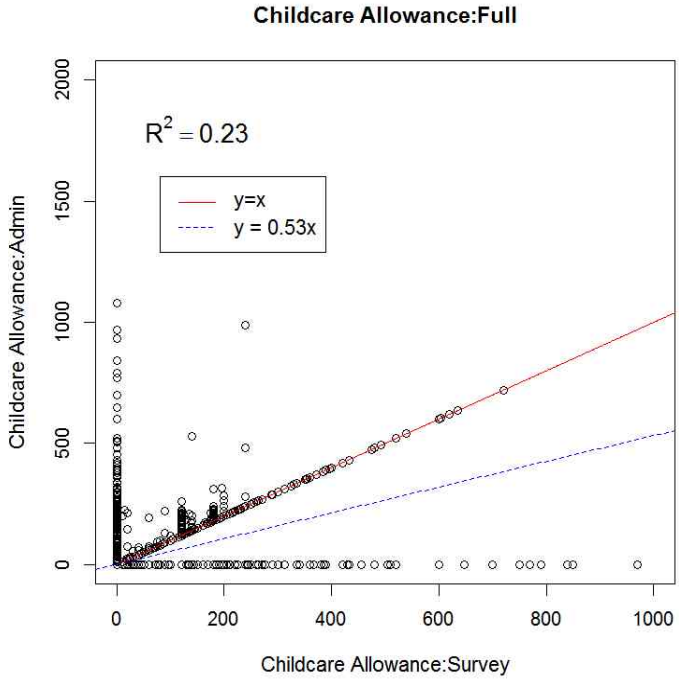
<표 2-12>에서 알 수 있듯이, 매칭된 조사 가구원 중 조사 양육수당이 0인 가구원이 98.3%이고, 그 중 조사+행정 양육수당이 0인 가구원이 96.8%로 상당히 높은 비중을 차지한다. 비매칭 집단의 조사 양육수당이 0인 가구원 또한 98.3%이다.

<표 2-12. 매칭 집단별 조사 및 조사+행정 양육수당 0 값 빈도>

			조사 양육수당	
			=0	>0
매칭	조사+행정 양육수당	=0	42412 (96.8%)	274 (0.6%)
		>0	677 (1.5%)	450 (1.0%)
	열 총합		43089 (98.3%)	724 (1.6%)
비매칭	열 총합		5607 (98.3%)	98 (1.7%)

<그림 2-6>을 보면 2017년 조사 가구원의 양육수당 자료의 특성을 다음과 같이 크게 세 가지 형태로 요약할 수 있다. 조사 양육수당과 조사+행정 양육수당이 일치하는 형태, 조사 양육수당은 0이나 조사+행정 양육수당이 0보다 큰 값으로 퍼져있는 형태, 조사 양육수당은 0보다 큰 양의 값이나 조사+행정 양육수당이 0인 형태이다.

<그림 2-6. 조사 vs. 조사+행정 양육수당 산점도>



<표 2-13>과 같이 매칭 집단 별 요약 통계량을 통해 양육수당의 자료 형태를 더 살펴보면, 다른 기타 소득항목과 마찬가지로 매칭된 가구원 중 조사 및 조사+행정 양육수당의 1,2, 그리고 3사분위수 모두 0이고, 평균은 조사+행정 양육수당이 더 크나, 그 차이가 크지 않다.

<표 2-13. 매칭 집단별 조사 및 조사+행정 양육수당 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	3.1
	조사+행정		0	0	0	4.1
비매칭	조사	5705	0	0	0	2.9
	조사+행정		–	–	–	–

7) 장애수당

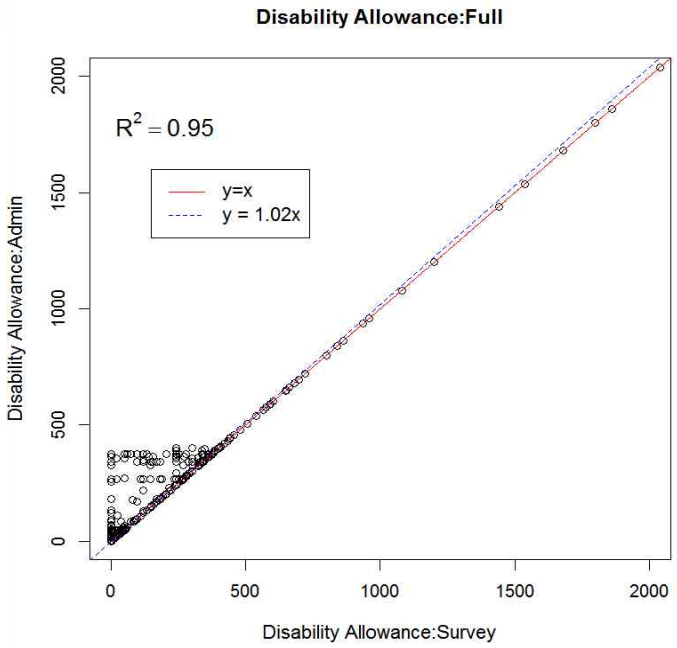
<표 2-14>에서 알 수 있듯이, 매칭된 조사 가구원 중 조사 장애수당이 0인 가구원은 98.4%이고, 그 중 조사+행정 장애수당이 0인 가구원이 98.1%로 상당히 높은 비중을 차지한다. 비매칭 집단의 조사 장애수당이 0인 가구원은 99.1%이다.

<표 2-14. 매칭 집단별 조사 및 조사+행정 장애수당 0 값 빈도>

			조사 장애수당	
			=0	>0
매칭	조사+행정 장애수당	=0	42976 (98.1%)	0 (0.0%)
		>0	130 (0.3%)	707 (1.6%)
	열 총합		43106 (98.4%)	707 (1.6%)
비매칭	열 총합		5655 (99.1%)	50 (0.9%)

<그림 2-7>을 보면, 조사 장애수당이 조사+행정 장애수당과 일치하거나, 조사 장애수당이 500만원보다 작은 가구원의 경우 조사+행정 장애수당이 조사 장애수당보다 큰 자료 분포 또한 확인할 수 있다.

<그림 2-7. 조사 vs. 조사+행정 장애수당 산점도>



<표 2-15>는 매칭 집단별 조사 및 조사+행정 장애수당 요약 통계량을 보여준다. 다른 기타 소득항목과 마찬가지로, 매칭된 가구원 중 조사 및 조사+행정 장애수당의 1,2, 그리고 3사분위수 모두 0이고, 평균은 조사+행정 장애수당이 더 크나, 그 차이가 크지 않다.

<표 2-15. 매칭 집단별 조사 및 조사+행정 장애수당 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	3.1
	조사+행정		0	0	0	3.6
비매칭	조사	5705	0	0	0	1.8
	조사+행정		–	–	–	–

8) 맞춤형 기초생활 보장 지원금

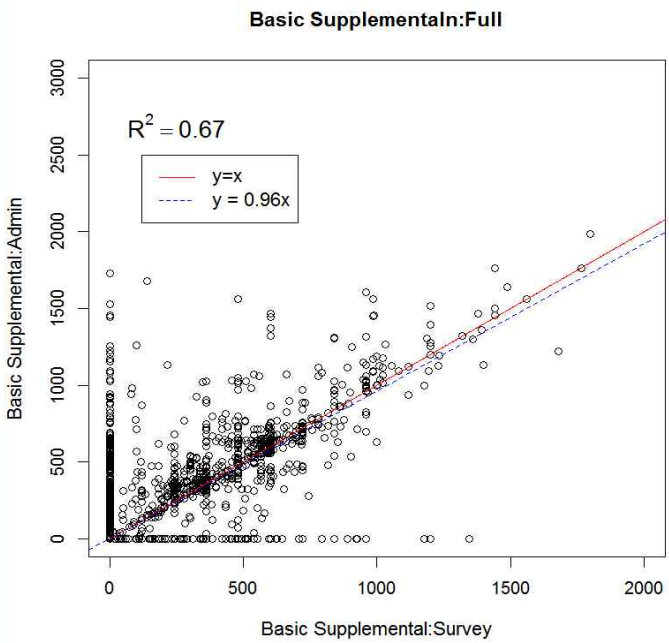
<표 2-16>은 매칭 집단별 조사 및 조사+행정 맞춤형 기초생활 보장 지원금의 0 값 빈도를 보여준다. 매칭된 조사 가구원 중 조사 맞춤형 기초생활 보장 지원금이 0인 가구원이 98.3%이고, 그 중 조사+행정 맞춤형 기초생활 보장 지원금이 0인 가구원이 97.8%로 상당히 높은 비중을 차지한다. 비매칭 집단에서 조사 맞춤형 기초생활 보장 지원금이 0인 가구원은 98.7%이다.

<표 2-16. 매칭 집단별 조사 및 조사+행정 맞춤형 기초생활 보장 지원금 0 값 빈도>

			조사 지원금	
			=0	>0
매칭	조사+행정 지원금	=0	42841 (97.8%)	115 (0.3%)
		>0	221 (0.5%)	636 (1.5%)
	열 총합		43062 (98.3%)	751 (1.7%)
비매칭	열 총합		5628 (98.7%)	77 (1.3%)

<그림 2-8>을 보면 조사 맞춤형 기초생활 보장 지원금과 조사+행정 맞춤형 기초생활 보장 지원금의 산점도 또한 크게 세 가지 유형으로 구분할 수 있다. 조사 맞춤형 기초생활 보장 지원금이 0이나 조사+행정 맞춤형 기초생활 보장 지원금이 0보다 큰 값으로 퍼져있는 유형, 그와 반대로 조사 맞춤형 기초생활 보장 지원금은 0보다 큰 값으로 퍼져있으나 조사+행정 맞춤형 기초생활 보장 지원금이 0인 유형, 조사 및 조사+행정 조사 맞춤형 기초생활 보장 지원금이 y=x를 기준으로 다소 퍼져있으나 뚜렷한 양의 선형 관계를 보이는 유형이다.

<그림 2-8. 조사 vs. 조사+행정 맞춤형 기초생활 보장 지원금 산점도>



<표 2-17>과 같이 매칭 집단 별 요약 통계량을 통해 맞춤형 기초생활 보장 지원금 자료를 더 살펴보면 매칭된 가구원 중 조사 및 조사+행정 맞춤형 기초생활 보장 지원금의 1,2, 그리고 3사분위수는 모두 0이고, 평균은 조사+행정 맞춤형 기초생활 보장 지원금이 더 큰 것을 알 수 있다.

<표 2-17. 매칭 집단별 조사 및 조사+행정 맞춤형기초생활 보장 지원금 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	8.3
	조사+행정		0	0	0	10.4
비매칭	조사	5705	0	0	0	6.1
	조사+행정		–	–	–	–

9) 소득세

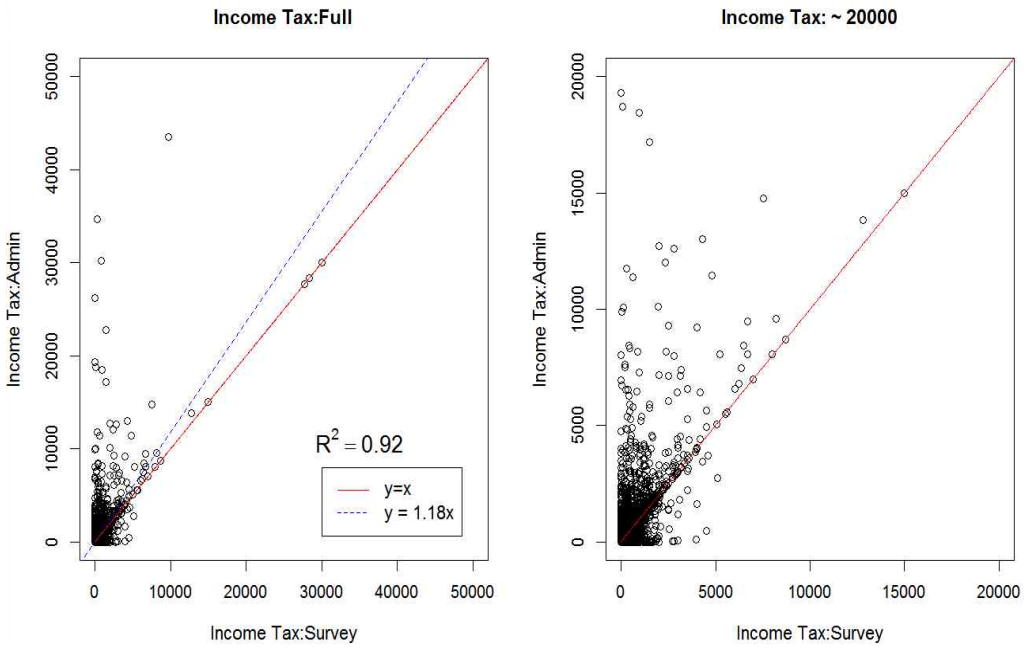
<표 2-18>에서 알 수 있듯이, 매칭된 조사 가구원 중 조사 소득세가 0인 가구원은 68.7%이고, 그 중 조사+행정 소득세가 0인 가구원은 48.4%, 조사+행정 소득세는 20.3%이다. 비매칭 집단의 조사 소득세가 0인 가구원은 69.3%이다.

<표 2-18. 매칭 집단별 조사 및 조사+행정 소득세 0 값 빈도>

			조사 소득세	
			=0	>0
매칭	조사+행정 소득세	=0	21192 (48.4%)	3361 (7.7%)
		>0	8899 (20.3%)	10361 (23.6%)
	열 총합		30091 (68.7%)	13722 (31.3%)
비매칭	열 총합		3955 (69.3%)	1750 (30.7%)

<그림 2-9>는 조사 소득세와 조사+행정 소득세의 산점도를 조사 가구원 전체와 일부(소득세 15,000 만원 미만인 가구원)를 대상으로 보여준다. 조사 소득세와 조사+행정 소득세가 양의 선형 관계를 보이기는 하나, 조사 소득세가 10,000만원 미만인 가구원 중 조사 소득세보다 조사+행정 소득세가 큰 가구원들이 눈에 띄게 분포되어 있는 것을 확인할 수 있다.

<그림 2-9. 조사 vs. 조사+행정 소득세 산점도>



<표 2-19>와 같이 매칭 집단 별 요약 통계량을 통해 조사 및 조사+행정 소득세를 더 살펴보면 매칭된 가구원 중 조사 소득세가 조사+행정 소득세보다 3사분위수

는 더 높으나 평균은 오히려 낮은 것을 알 수 있다.

<표 2-19. 매칭 집단별 조사 및 조사+행정 소득세 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	15.0	61.2
	조사+행정		0	0	12.0	84.7
비매칭	조사	5705	0	0	14.0	51.1
	조사+행정		–	–	–	–

2.2.2.2. 부채 (가구원 단위) 기초 자료 분석

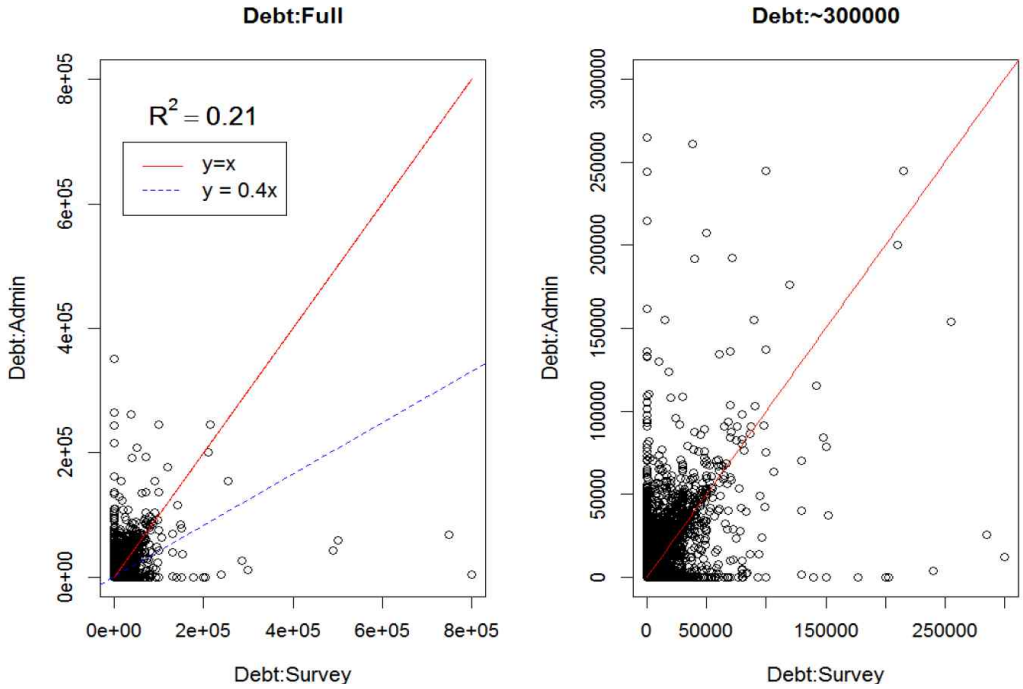
본 연구에서는 가구원 단위의 부채 자료를 살펴보기로 한다. <표 2-20>은 매칭 집단별 조사 및 조사+행정 부채의 0값 빈도를 보여준다. 매칭된 조사 가구원 중 조사 부채가 0인 가구원은 76.8%이고, 그 중 조사+행정 부채가 0인 가구원이 60.8%로 높은 비중을 차지한다. 비매칭 집단에서 조사 부채가 0인 가구원은 81.1%이다.

<표 2-20. 매칭 집단별 조사 및 조사+행정 부채 0 값 빈도>

			조사 부채	
			=0	>0
매칭	조사+행정 부채	=0	26640 (60.8%)	917 (2.1%)
		>0	7001 (16.0%)	9255 (21.1%)
	열 총합		33641 (76.8%)	10172 (23.2%)
비매칭	열 총합		4629 (81.1%)	1076 (18.9%)

<그림 2-10>은 가구원 단위 조사 부채와 조사+행정 부채의 산점도를 전체 조사 가구원 대상과 일부 가구원 대상 (부채가 300,000 만원 미만인 가구원)으로 보여준다. 조사 부채와 조사+행정 부채가 약한 양의 선형 관계를 보이고, 조사 부채가 낮은 가구원에서도 자료의 분포가 꽤 퍼져있는 것을 확인할 수 있다.

<그림 2-10. 가구원 단위 조사 vs. 조사+행정 부채 산점도>



<표 2-21>은 매칭 집단별 조사 및 조사+행정 부채의 요약 통계량을 보여준다. 매칭된 가구원 중 조사 및 조사+행정 부채의 1사분위수와 2사분위수는 모두 0이다. 그러나 조사 부채의 3사분위수는 0인 반면 조사+행정 부채의 3사분위는 1000만원이고, 조사 부채의 평균이 1605만원인 반면 조사+행정 부채의 평균은 2465만원으로 다소 큰 차이가 있다.

<표 2-21. 매칭 집단별 조사 및 조사+행정 부채 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	1605
	조사+행정		0	0	1000	2465
비매칭	조사	5705	0	0	0	1300
	조사+행정		–	–	–	–

2.2.3. 분석 결과: 소득

2.2.3 절에는 2017년 가계금융복지조사의 가구원 단위 자료의 소득 항목에 대하여 2.1절에서 소개한 대체 방법론의 적용 과정 및 결과를 제시한다. 이 절에서는 조사 소득값이 있는 경우에서의 분석 결과만을 제시하고 조사 소득값이 없는 경우에서의 분석 결과는 부록으로 남겨둔다.

2.2.3.1. 근로소득

1) 분석 절차

근로소득은 조사값이 0보다 큰 가구원만 근로소득 대상자로 제한하여 분석하므로, 혼합 비 대체 모형을 이용한 대체 방법론을 바로 적용할 수 있다.

- 후보 모형: 혼합 비 대체 모형1, 조건부 혼합 비 대체 모형 2
- G 선택: 각 후보 모형 별 10-fold 교차 검증법
- 대체: 결정적 대체 방법
- MAR 검정 통계량(X^2) 비교 후 모형 1과 모형 2 중 최종 모형 선택
  - 범주형 변수 x_b: 4개의 범주를 가지는 연령 (30대 미만, 30대, 40대, 50대, 50대 이상)
  - 범주형 변수 y_c: 조사 근로소득 및 조사+행정 근로소득의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

2) 분석 결과

<표 2-22>는 2017년 조사 가구원 중 근로소득이 있는 가구원을 분석 대상자로 제한하여 분석한 결과를 보여준다. 대체값으로 조사값을 그대로 사용할 경우와 혼합 비 대체 모형 - 모형 1(기본 모형)과 조건부 모형 2(확장된 모형)의 결정적 대체 결과를 비교한다.

10-fold 교차 검증법을 통해 혼합 비 대체 모형 1에서 최대 혼합 성분 개수를 5로 두고, $G=1,2,...,5$를 혼합 성분으로 가지는 총 다섯 개의 모형을 평균 RMSE를 기준으로 비교한다. 그 결과, 가장 낮은 평균 RMSE는 $G=4$일 때이고, 그 때의 평균 RMSE는 1539.7이다. 마찬가지로 조건부 혼합 비 대체 모형 2에서 최대 혼합 성분 개수를 5로 두고, $G=1,2,...,5$를 혼합 성분으로 가지는 총 다섯 개의 모형을 평균 RMSE 기준으로 비교한 결과, 가장 낮은 평균 RMSE는 $G=5$일 때이고, 그 때의 평균 RMSE는 1621.2이다.

이제 대체값 후보로, 조사값과 혼합 비 대체 모형 1 ($G=4$), 그리고 조건부 혼합 비 대체 모형 2 ($G=5$)를 비교하여 최종 대체 방법을 선택한다.

평균 RMSE를 기준으로 다시 비교해보면, 대체값으로 조사값을 사용할 경우 평균 RMSE는 1672.4이고, 혼합 비 대체 모형을 사용하여 대체하는 경우 모형 1은 1539.7, 모형 2는 1621.2로 모형 1을 이용한 대체값이 RMSE 기준 가장 좋았다.

MAR 검정 통계량 또한 조사값을 그대로 사용할 경우(X^2 = 41.3)보다 혼합 비 대체 모형을 이용한 대체값이 모형 1에서 X^2=28.2, 모형 2에서 X^2=22.7로 값이 더 낮은 것을 확인할 수 있다.

따라서 비매칭 집단에서는 조사값을 그대로 사용하지 않고, 혼합 비 모형을 사용한 대체값을 사용하되, 모형 2 대비 모형 1의 X^2값 증가량보다 평균 RMSE가 뚜렷하게 감소했으므로 최종 대체 모형으로 모형 1을 선택한다.

<표 2-22. 조사 근로소득 및 대체 모형 별 평균 RMSE 및 X^2 비교>

대체 방법		G	평균 RMSE	X^2
조사값 사용		-	1672.4	41.3
혼합 비 대체	모형 1	4	1539.7	28.2
	모형 2-조건부: 연령 (10-20/30/40/50대이상)	5	1621.2	22.7

◦ 최종 대체 방법: 혼합 비 대체 모형 1 ($G=4$)를 이용한 결정적 대체

<표 2-23>은 최종 대체 모형인 혼합 비 대체 모형 1 ($G=4$)의 모수 추정 결과를 보여준다. 4개의 혼합 성분을 가지는 모형으로, 조사+행정 근로소득이 그룹 4에 속하면 조사 근로소득과 일치하는 값을, 그룹 3에 속하면 조사 근로소득보다 2.16배 높은 값을 평균으로 하는 분포를, 그룹 1 또는 2에 속하면 조사 근로소득보다 0.92배 혹은 0.99배인 값을 평균으로 하는 분포를 따르는 것으로 추정되었다. 추정된 혼합 확률을 보면 조사+행정 근로소득이 그룹 2에 속할 확률이 0.38로 가장 높았고, 그룹 3에 속할 확률이 0.07로 가장 낮았다.

<표 2-23. 최종 대체 모형 모수 추정 결과>

그룹	μ_g	σ_g^2	β_g	$\sigma_{e,g}^2$	π_g
1	7.67	0.12	0.92	265.75	0.31
2	8.42	0.25	0.99	492.25	0.38
3	6.73	1.14	2.16	1837.90	0.07
4	6.93	0.80	1.00	0.00	0.24

<표 2-23>의 모수 추정값을 이용하여 결정적 대체값을 구한 후, 매칭 집단 별 조사 근로소득, 조사+행정 또는 대체 근로소득의 요약통계량을 정리하면 <표 2-24>와 같다. 매칭 집단을 먼저 살펴보면, 조사 근로소득에 비해 조사+행정 근로소득의 1, 2사분위수는 낮은 반면, 3사분위수 및 평균은 더 높다. 비매칭 집단에서 조사 근로소득과 대체 근로소득의 요약통계량을 살펴보면, 조사 근로소득에 비해 대체 근로소득의 1사분위수는 높았으나 2사분위수는 낮았고, 3사분위수 및 평균은 더 높다. 전체 (매칭, 비매칭) 집단에서도 유사한 패턴을 확인할 수 있다.

<표 2-24. 매칭 집단별 근로소득 (조사·조사+행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	15664	1440	2400	4000	3145
	조사+행정		1200	2228	4220	3199
비매칭	조사	2047	1500	2400	3710	2929
	대체		1535	2356	3715	2977
전체 (매칭,비매칭)	(조사, 조사)	17711	1445	2400	4000	3120
	(조사+행정, 조사)		1203	2280	4140	3167
	(조사+행정, 대체)		1240	2270	4140	3173

2.2.3.2 기타 소득 항목

2.2.2절 자료 현황에서 살펴보았듯이, 기타 소득 항목 – 금융소득, 임대소득, 공적연금, 기초연금, 양육수당, 장애수당, 맞춤형 기초생활 보장 지원금,소득세 – 에서는 조사값이 0인 가구원이 많으므로 조사값이 0인 집단과 0보다 큰 집단으로 나누어 독립적으로 대체 모형을 적합한다.

▫ 조사값이 0인 집단

I) 모형

관심 변수인 특정 항목의 조사+행정값을 y 라고 하고, 조사값을 $\tilde{y}$ 라고 하자. 즉, 조사값이 0인 집단의 $\tilde{y}$ 값은 모두 0이다. 2.2.2절 자료 현황에서 기타 소득항목의 기초 분석 결과(0값 빈도표 및 산점도)를 보면 조사값이 0일 때 조사+행정값이 0인 가구원도 있지만, 조사+행정값이 0보다 큰 가구원도 존재한다. 이러한 자료 특성을 반영하여 본 연구 과제에서는 인구학적 범주형 변수 x와 조사값 $\tilde{y}=0$ 가 주어졌을 때, y 의 분포를 다음과 같은 2개의 혼합 성분을 가지는 혼합 모형을 따른다고 가정한다.

$$p_1(x)f_1(y|\tilde{y}=0,x)+p_2(x)f_2(y|\tilde{y}=0,x),$$

단, $p_1(x)=P(y=0\,|\,\tilde{y}=0,\,x)$이고, $p_2(x)=P(y>0\,|\,\tilde{y}=0,x)$이다. 여기서, 보조변수 x는 연령대, 교육정도 등의 인구학적 범주형 변수를 고려할 수 있다. 2.2.2절 자료 현황에서 관찰한 자료의 특성를 반영하기 위해, 구체적으로 다음과 같은 모수적 혼합 모형을 가정한다.

$$f_1(y|\tilde{y}=0,x)=\begin{cases}1 & \text{if} \quad y=0\\ 0 & o.w.\end{cases}, \qquad (5)$$

$$f_2(y|\tilde{y}=0,x)=\frac{1}{\sqrt{2\pi\sigma_0^2(x)}}exp\left\{-\frac{1}{2}(T(y)-\mu_0(x))^2\right\},$$

단, $T(y)$는 y의 변환 함수로, 본 최종보고서에서는 $T(y)=\log(y)$를 사용하였다. 모형 (5)는 조사값이 0인 가구원이 그룹 1에 속하면 조사+행정 값 또한 0을 가정하고, 그룹 2에 속하면 $T(y)\sim N(\mu_0(x),\sigma_0^2(x))$를 가정한다.

ii) 대체 방법

모수를 추정한 후, 다음과 같은 두 가지 대체 방법을 고려할 수 있다.

◦ 결정적 대체(Deterministic imputation): 비매칭 집단의 j번째 가구원의 보조변수 x_j와 조사값 $\tilde{y_j}=0$가 주어졌을 때, 각 그룹 $g=1,2$에 속할 확률을 비교한 후, 만약 $\hat{p_1}(x_j)>\hat{p_2}(x_j)$이면 대체값을 $\hat{y}_j=0$으로, 그렇지 않으면 $\hat{y}_j=\exp(\hat{\mu_0})$로 계산한다.

※만약 2.1절에서 소개한 대체 방법론의 결정적 대체 방법처럼 가중 평균으로 대체값을 대체할 경우, 조사값이 0인 가구원임에도 불구하고 대체값이 0보다 큰 값으로 대체될 가구원의 수가 너무 많아질 수 있으므로 기타 소득 항목의 특성 대비 과대 추정될 가능성이 높다. 따라서 조사값이 0인 집단에서는 결정적 대체 방법으로 가중 평균을 고려하지 않는다.

◦ 부분 확률적 대체(Stochastic imputation): 비매칭 집단의 j번째 가구원의 보조변수 x_j와 조사값 $\tilde{y_j}=0$가 주어졌을 때, 각 그룹 $g=1,2$에 속할 확률로부터 그룹 g를 랜덤하게 선택한다. 만약 $g=1$이 뽑혔다면 대체값을 대체값을 $\hat{y}_j=0$으로, $g=2$이면 $\hat{y}_j=\exp(\hat{\mu_0})$로 계산한다.

▫ 조사값이 0보다 큰 집단

2.2.2절 자료 현황에서 기타 소득항목의 조사값과 조사+행정값의 산점도를 다시 살펴보자. 금융소득을 예로 들면, 조사값을 사용하는 경우라도 조사값과 조사+행정

값의 선형 관계가 강하지 않고, 조사값이 낮은 경우에도 분산의 정도가 꽤 퍼져있는 것을 확인할 수 있다. 이러한 경우, 2.1절에서 소개한 혼합 대체 비 모형 1 또는 조건부 혼합 대체 비 모형 2를 가정하는 것이 자료에 적합하지 않을 수 있다. 따라서 본 연구에서는 혼합 대체 비 모형 1, 2 뿐만 아니라 다음과 같은 모형 또한 후보 모형으로 고려한 후 최종 모형을 선택하기로 한다.

$$\begin{aligned} &\boldsymbol{z} \sim Multinomial(1, \boldsymbol{\pi}), \qquad (6)\\ &T(\tilde{y})|z_g = 1 \sim N(\mu_g, \sigma_g^2),\\ &y|(\tilde{y}, z_g = 1) \sim N(\tilde{y}\beta_g, \sigma_{e,g}^2), \quad g = 1,...,G, \end{aligned}$$

단, $\boldsymbol{z} = (z_1,...,z_G)'$ 이고, $\boldsymbol{\pi} = (\pi_1,...,\pi_G)'$ 이다. 여기서, $T(\tilde{y})$ 는 그룹 내 정규분포를 따르도록 하는 $\tilde{y}$의 변환 변수로 $\log(\tilde{y})$ 또는 $\sqrt{\tilde{y}}$ 등이 있다.
모형 (1)과 모형 (6)의 차이점은 모형 (1)에서는 y가 그룹 g에 속할 때, y의 평균이 $\tilde{y}\beta_g$이고 분산이 $\tilde{y}\sigma_{e,g}^2$인 정규분포를 가정한 반면, 모형 (6)은 평균은 동일하나, 분산이 $\sigma_{e,g}^2$인 정규분포를 가정함으로써, 각 그룹 내에서 조사값의 크기와 관계없이 분산이 동질(homogeneous)하도록 분포를 가정하였다.

모형 (6)은 결국 예측변수가 조사값 하나이고 절편이 없는 회귀 모형이다. 마찬가지로 절편이 없는 조건부 혼합 회귀 대체 모형으로 확장된 모형 또한 가정할 수 있다. 모수 추정 방법 및 대체 방법은 2.1.2절 혼합 회귀 대체를 참고하기 바란다.

1) 금융소득

1-1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 연령 (30대 미만, 30대-50대, 60대 이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: 혼합 비 대체 모형1, 조건부 혼합 비 대체 모형 2, (절편이 없는) 혼합 회귀 대체 모형1, (절편이 없는) 조건부 혼합 회귀 대체 모형 2

- 조건부 혼합 비 / 회귀 대체 모형 2의 보조 변수 x : 연령(30대 미만, 30대, 40대, 50대, 50대 이상)

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석

▫ 대체: 결정적 대체 방법

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b: 연령 (30대 미만, 30대, 40대, 50대, 50대 이상)

◦ 범주형 변수 y_c: 조사 금융소득 및 조사+행정 금융소득의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

1-2) 분석 결과

2017년 조사 가구원 전체를 대상으로 금융소득 항목에 관하여 분석한 결과는 다음과 같다. 먼저 조사값이 0인 집단의 분석 결과를 살펴보자. <표 2-25>는 조사값이 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 연령이 30대 미만, 30~50대, 그리고 60대 이상인 경우 그룹 1에 속할 확률이 각각 0.73, 0.32, 0.29로 30대 미만은 그룹 1에 속할 확률이, 30대 이상은 그룹 2에 속할 확률이 높았다.

<표 2-25. 조사값이 0인 집단의 금융소득 모형 모수 추정 결과>

그룹	x:연령	30대 미만	30대-50대	60대 이상
1	$p_1(x)$	0.73	0.32	0.29
2	$p_2(x)$	0.27	0.68	0.71
	μ_0	1.66	2.53	3.17
	σ_0^2	2.52	3.87	3.80

<표 2-26>는 조사값이 0인 경우와 0보다 큰 경우에 관한 10-fold 교차 검증법 결과를 보여준다. 조사값이 0인 경우에 모형을 사용한 대체 방법이 더 좋았으나, 그 차이가 크지는 않다. 본 최종보고서에는 결정적 대체 방법으로 최종 결과를 기술한다.

조사값이 0보다 큰 경우, 10-fold 교차 검증 결과를 통해 대체값으로 조사값을 그대로 사용할 경우와 혼합 비 대체 모형 - 모형 1(기본 모형)과 조건부 모형 2(확장된 모형) 및 절편이 없는 혼합 회귀 모형 - 모형 1 (기본 모형)과 조건부 모형 2 (확장된 모형)의 결정적 대체 결과를 비교할 수 있다. 각 모형 종류별로 최대 혼합 성분 개수를 5로 두고, $G = 1, 2, ..., 5$를 혼합 성분으로 가지는 다섯 개의 모형을 평균 RMSE를 기준으로 비교한다. 그 결과, 혼합 비 대체 모형보다 절편이 없는 혼합 회귀 모형이 평균 RMSE 기준으로 더 좋았고, 혼합 회귀 대체 모형1 에서 가장 낮은 가장 낮은 평균 RMSE는 $G = 5$일 때이고, 그 때의 평균 RMSE는 1039.5이

다. 조건부 혼합 회귀 대체 모형 2에서 가장 낮은 평균 RMSE는 $G=3$일 때이고, 그 때의 평균 RMSE는 1054.2이다.

이제 대체값 후보로, 조사값과 혼합 회귀 대체 모형 1 ($G=5$), 그리고 조건부 혼합 비 대체 모형 2 ($G=3$)를 비교하여 최종 대체 방법을 선택한다.

평균 RMSE를 기준으로 다시 비교해보면, 대체값으로 조사값을 사용할 경우 평균 RMSE는 1021.4이고, 혼합 회귀 대체 모형을 사용하여 대체하는 경우 모형 1은 1039.5, 모형 2는 1054.2로, 조사값을 이용할 때 평균 RMSE 값이 가장 좋았다.

<표 2-26. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

<table>
<tr><th colspan="2"></th><th colspan="2">대체 방법</th><th>평균 RMSE</th><th>G</th></tr>
<tr><td rowspan="3">조사값 = 0</td><td colspan="3">조사값 대체</td><td>350.5</td><td rowspan="3">-</td></tr>
<tr><td colspan="3">모형 대체 - 결정적</td><td>348.4</td></tr>
<tr><td colspan="3">모형 대체 - 확률적</td><td>349.0</td></tr>
<tr><td rowspan="5">조사값 > 0</td><td colspan="3">조사값 대체</td><td>1021.4</td><td>-</td></tr>
<tr><td rowspan="2">혼합 비</td><td colspan="2">모형 1</td><td>1101.8</td><td>1</td></tr>
<tr><td colspan="2">모형 2 - 조건부: 연령 (10-20/30/40/50대이상)</td><td>1101.8</td><td>1</td></tr>
<tr><td rowspan="2">혼합 회귀</td><td colspan="2">모형 1</td><td>1039.5</td><td>5</td></tr>
<tr><td colspan="2">모형 2 - 조건부: 연령 (10-20/30/40/50대이상)</td><td>1054.2</td><td>3</td></tr>
</table>

<표 2-27>에서는 조사값과 모형 대체값에 대한 MAR 검정 결과를 보여준다. 조사값을 그대로 사용할 경우 X^2은 1894.4이고, 혼합 회귀 대체 모형을 이용한 대체값은 모형 1($G=5$)에서 X^2=755.9, 모형 2($G=3$)에서 X^2=758.2로 값이 현저하게 감소한 것을 확인할 수 있다. 조사값 대비 혼합 회귀 대체 모형 1의 평균 RMSE 값의 증가량보다 X^2 값이 현저하게 감소했으므로, 최종 대체 모형으로 조사값이 0일 때는 모형 대체를, 조사값이 0보다 큰 집단에서는 모형 1을 선택한다.

<표 2-27. 금융소득 대체 후 MAR 검정 결과>

<table>
<tr><th colspan="2"></th><th>X^2</th></tr>
<tr><td colspan="2">조사값 대체</td><td>1894.4</td></tr>
<tr><td rowspan="2">조사값 = 0:
결정적 모형 대체</td><td>조사값>0:
혼합 회귀 모형 1</td><td>755.9</td></tr>
<tr><td>조사값>0:
혼합 회귀 모형 2 - 조건부: 연령</td><td>758.2</td></tr>
</table>

◦ 최종 대체 방법: 조사값이 0인 경우 모형 대체와 조사값이 0보다 큰 경우 혼합 회귀 대체 모형 1($G=5$)를 이용한 결정적 대체

<표 2-28. 조사값 > 0 인 집단의 금융소득 최종 모형 모수 추정 결과>

그룹	μ_g	σ_g^2	β_g	$\sigma_{e,g}^2$	π_g
1	3.41	0.97	3.16	9370	0.31
2	4.68	0.47	0.92	7447	0.31
3	5.36	1.91	0.87	413359	0.19
4	2.96	1.13	1.00	212	0.18
5	6.82	1.84	2.01	152407990	0.01

<표 2-28>에서와 같이 최종 대체 모형의 모수를 추정한 후 결정적 대체값을 구한 다음, 매칭 집단 별 조사 금융소득, 조사+행정 또는 대체 금융소득의 요약통계량을 정리하면 <표 2-29>와 같다. 매칭 집단을 먼저 살펴보면, 조사 금융소득은 1,2, 그리고 3사분위수가 모두 0이고, 평균이 12.0인 반면, 조사+행정 금융소득의 3사분위수는 28.0이고 평균은 65.3으로 조사 금융소득과의 차이를 보여준다. 비매칭 집단에서 조사 금융소득과 대체 금융소득의 요약통계량을 살펴보면, 조사 금융소득이 1,2, 그리고 3사분위수가 모두 0이고 평균이 6.1인 반면, 모형을 이용하여 대체한 금융소득의 2사분위수와 3사분위수가 12.5이고, 평균이 16.6이다. 전체 (매칭, 비매칭) 집단에서도 유사한 패턴을 확인할 수 있으며, (조사+행정, 조사) 금융소득의 3사분위수와 평균보다 (조사+행정, 대체) 금융소득의 3사분위수와 평균이 조금 더 높지만 그 차이가 크지 않다.

<표 2-29. 매칭 집단별 금융소득 (조사·행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	12.0
	조사+행정		0	1	28.0	65.3
비매칭	조사	5705	0	0	0	6.1
	대체		0	12.5	12.5	16.6
전체 (매칭, 비매칭)	(조사, 조사)	49518	0	0	0	11.3
	(조사+행정, 조사)		0	1	20.0	58.5
	(조사+행정, 대체)		0	1	23.8	59.7

2) 임대소득

2-1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 연령 (30대 미만, 30대-50대, 60대 이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: 혼합 비 대체 모형1, (절편이 없는) 혼합 회귀 대체 모형1

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 2$ 로 선택하여 분석 (2.2절 자료 현황의 임대소득 산점도 참고)

▫ 대체: 결정적 대체 방법

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b : 연령 (30대 미만, 30대-50대, 60대 이상)

◦ 범주형 변수 y_c : 조사 임대소득 및 조사+행정 임대소득의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

2-2) 분석 결과

2017년 조사 가구원 전체를 대상으로 임대소득 항목에 관하여 분석한 결과는 다음과 같다. 먼저 조사값이 0인 집단의 분석 결과를 살펴보자. <표 2-30>은 조사값이 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 연령이 30대 미만, 30~50대, 그리고 60대 이상인 경우 그룹 1에 속할 확률이 각각 1, 0.99, 0.99로 모두 그룹 1에 속할 확률에 1에 가깝다. (교육정도를 보조 변수로 사용해도 유사한 결과를 나타냄)

<표 2-30. 조사값이 0인 집단의 임대소득 모형 모수 추정 결과>

그룹	x:연령	30대 미만	30대-50대	60대 이상
1	$p_1(x)$	1	0.99	0.99
2	$p_2(x)$	0	0.01	0.01
	μ_0	6.95	5.89	6.53
	σ_0^2	1.24	3.43	1.66

<표 2-31>은 조사값이 0인 집단과 0보다 큰 집단에 관한 10-fold 교차 검증법 결과를 보여준다. 조사값이 0인 경우에 모형을 사용한 대체 방법(결정적)이 조사값을 사용한 경우와 동일한 평균 RMSE를 가진다. 이는 임대소득 자료 특성상, 조사값이 0이면 조사+행정 자료값이 0인 경우가 매칭 집단에 압도적이었으므로, 대체 모형 또한 그룹 1에 속할 확률이 모든 연령 범주에서 1에 가까웠고, 따라서 결정적

대체 방법을 이용한 대체값 또한 모두 0으로 계산되었기 때문이다.

조사값이 0보다 큰 경우, 10-fold 교차 검증 결과를 통해 대체값으로 조사값을 그대로 사용할 경우와 혼합 비 대체 모형 - 모형 1(기본 모형)과 절편이 없는 혼합 회귀 모형 - 모형 1 (기본 모형)의 결정적 대체 결과를 비교할 수 있다. 각 모형 종류별로 최대 혼합 성분 개수를 2로 두고, $G=1,2$를 혼합 성분으로 가지는 두 개의 모형을 평균 RMSE를 기준으로 비교한다. 그 결과, 모든 모형에서 $G=1$에서 평균 RMSE 값이 0이다. 따라서 비매칭 집단에서 모형을 이용한 대체값을 사용하지 않고, 조사값을 그대로 사용하기로 한다.

<표 2-31. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

	대체 방법		평균 RMSE	G
조사값 = 0	조사값 대체		180.9	-
	모형 대체 - 결정적		180.9	
	모형 대체 - 확률적		185.9	
조사값 > 0	조사값 대체		0	-
	혼합 비	모형 1	0	1
		모형 2 - 조건부: 연령	-	-
	혼합 회귀	모형 1	0	1
		모형 2 - 조건부: 연령	-	-

◦ 최종 대체 방법: 조사값 사용

<표 2-32>는 매칭 집단 별 조사 임대소득, 조사+행정 또는 대체 임대소득의 요약통계량을 정리한 표이다. 여기서, 대체값은 조사값과 동일한 값이다.

<표 2-32. 매칭 집단별 임대소득 (조사·조사+행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	70.0
	조사 + 행정		0	0	0	76.1
비매칭	조사	5705	0	0	0	45.8
	대체(=조사)		0	0	0	45.8
전체 (매칭,비매칭)	(조사, 조사)	49518	0	0	0	67.3
	(조사+행정, 조사)		0	0	0	72.6
	(조사+행정, 대체)		0	0	0	72.6

3) 공적연금

3-1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 연령 (60대 미만, 60대 이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: 혼합 비 대체 모형1, 조건부 혼합 비 대체 모형 2, (절편이 없는) 혼합 회귀 대체 모형1, (절편이 없는) 조건부 혼합 회귀 대체 모형 2

- 조건부 혼합 비 / 회귀 대체 모형 2의 보조 변수 x : 교육정도 (초졸 미만, 중·고졸, 대졸 이상)

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석

▫ 대체: 결정적 대체 방법

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b : 교육정도 (초졸 미만, 중·고졸, 대졸 이상)

◦ 범주형 변수 y_c : 조사 공적연금 및 조사+행정 공적연금의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

3-2) 분석 결과

2017년 조사 가구원 전체를 대상으로 공적연금 항목에 관한 분석 결과는 다음과 같다. <표 2-33>은 조사값이 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 연령이 60대 미만과 60대 이상인 경우 그룹 1에 속할 확률이 각각 1과 0.87로 모두 그룹 1에 속할 확률이 높았다.

<표 2-33. 조사값이 0인 집단의 공적연금 모형 모수 추정 결과>

그룹	x:연령	60대 미만	60대 이상
1	$p_1(x)$	1	0.87
2	$p_2(x)$	0	0.13
	μ_0	5.79	5.54
	σ_0^2	0.59	0.48

<표 2-34>는 조사값이 0인 집단과 0보다 큰 집단에 관한 10-fold 교차 검증법 결과를 보여준다. 조사값이 0인 경우에 모형을 사용한 대체 방법(결정적)이 조사값

을 사용한 경우와 동일한 평균 RMSE를 가진다. 이는 공적연금 자료 특성상, 조사값이 0이면 조사+행정 자료값이 0인 경우가 매칭 집단에 압도적이었으므로, 대체 모형 또한 그룹 1에 속할 확률이 모든 연령 범주에서 1에 가까웠고, 따라서 결정적 대체 방법을 이용한 대체값 또한 모두 0으로 계산되었기 때문이다.

10-fold 교차 검증법을 통해 조사값이 0보다 큰 경우 혼합 비 대체 모형 1, 2와 절편이 없는 혼합 회귀 대체 모형 1, 2를 비교한다. 각 모형 종류별로 최대 혼합 성분 개수를 5로 두고, $G=1,2,...,5$를 혼합 성분으로 가지는 다섯 개의 모형을 평균 RMSE를 기준으로 비교한다. 각 모형 종류 별로 평균 RMSE가 가장 낮도록 하는 G값이 <표 2-34>에 제시되어 있다. 조사값을 사용했을 때의 평균 RMSE는 274.3이고, 혼합 비/회귀 대체 모형 1과 2를 이용하여 대체했을 때의 평균 RMSE는 모두 약 271 ~ 272로 감소하였으나, 감소한 크기가 매우 작은 편이다.

<표 2-35>에서는 조사값과 혼합 비 대체 모형 1($G=5$) 및 조건부 혼합 회귀 대체 모형 2($G=3$)을 이용하여 계산한 대체값에 대해 MAR 검정 결과를 보여준다. 10-fold 교차 검증 결과와 마찬가지로 조사값을 사용할 경우 X^2는 60.13이고, 모형을 이용한 대체값은 모형 1에서 X^2=60.13, 모형 2에서 X^2=60.00으로 거의 차이가 없음을 알 수 있다. 이에 따라 최종 대체 방법으로 조사값을 사용하기로 한다.

<표 2-34. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

	대체 방법		평균 RMSE	G
조사값 = 0	조사값 대체		97.4	-
	모형 대체 - 결정적		97.4	
	모형 대체 - 확률적		103.6	
조사값 > 0	조사값 대체		274.3	-
	혼합 비	모형 1	271.5	5
		모형 2 - 조건부: 교육정도 (초졸/중·고졸/대졸이상)	271.7	3
	혼합 회귀	모형 1	271.3	4
		모형 2 - 조건부: 교육정도 (초졸/중·고졸/대졸이상)	271.6	1

<표 2-35. 공적연금 대체 후 MAR 검정 결과>

		X^2
조사값 대체		60.13
조사값 = 0: 결정적 모형 대체	조사값>0: 혼합 비 모형 1	60.13
	조사값>0: 혼합 비 모형 2 - 조건부: 교육정도	60.00

◦ 최종 대체 방법: 조사값 사용

<표 2-36>는 매칭 집단 별 조사 공적연금, 조사+행정 또는 대체 공적연금의 요약통계량을 정리한 표이다. 여기서, 대체값은 조사값과 동일한 값이다.

<표 2-36. 매칭 집단별 공적연금 (조사·조사+행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	76.9
	행정		0	0	0	84.9
비매칭	조사	5705	0	0	0	41.3
	대체		0	0	0	41.3
전체 (매칭,비매칭)	(조사,조사)	49518	0	0	0	72.84
	(조사+행정, 조사)		0	0	0	72.86
	(조사+행정, 대체)		0	0	0	72.86

4) 기초연금

4-1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 연령 (60대 미만, 60대 이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: 혼합 비 대체 모형1, 조건부 혼합 비 대체 모형 2, (절편이 없는) 혼합 회귀 대체 모형1, (절편이 없는) 조건부 혼합 회귀 대체 모형 2

- 조건부 혼합 비 / 회귀 대체 모형 2의 보조 변수 x : 교육정도(초졸 미만, 중·고졸, 대졸 이상)

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석

▫ 대체: 결정적 대체 방법

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b: 교육정도 (초졸 미만, 중·고졸, 대졸 이상)

◦ 범주형 변수 y_c: 조사 기초연금 및 조사+행정 기초연금의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

4-2) 분석 결과

먼저 조사값이 0인 경우의 분석 결과를 살펴보자. <표 2-37>은 조사 기초연금이 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 연령이 60대 미만과 60대 이상인 경우 그룹 1에 속할 확률이 각각 1과 0.95로 모두 그룹 1에 속할 확률이 높았다.

<표 2-37. 조사값이 0인 집단의 기초연금 모형 모수 추정 결과>

그룹	x:연령	60대 미만	60대 이상
1	$p_1(x)$	1	0.95
2	$p_2(x)$	0	0.05
	μ_0	-	4.83
	σ_0^2	-	0.73

<표 2-38>은 조사값이 0인 집단과 0보다 큰 집단에 관한 10-fold 교차 검증법 결과를 보여준다. 조사값이 0인 경우에 모형을 사용한 대체 방법(결정적)이 조사값을 사용한 경우와 동일한 평균 RMSE를 가진다. 공적연금과 마찬가지로 조사값이 0이면 조사+행정 자료값이 0인 경우가 매칭 집단에 압도적이었으므로, 대체 모형 또한 그룹 1에 속할 확률이 모든 연령 범주에서 1에 가까웠고, 따라서 결정적 대체 방법을 이용한 대체값 또한 모두 0으로 계산되었기 때문이다.

10-fold 교차 검증법을 통해 조사값이 0보다 큰 경우 혼합 비 대체 모형 1, 2와 절편이 없는 혼합 회귀 대체 모형 1, 2를 비교한다. 각 모형 종류별로 최대 혼합 성분 개수를 5로 두고, $G=1,2,...,5$를 혼합 성분으로 가지는 다섯 개의 모형을 평균 RMSE를 기준으로 비교한다. 각 모형 종류 별로 평균 RMSE가 가장 낮도록 하는 G값이 <표 2-38>에 제시되어 있다. 조사값을 사용했을 때의 평균 RMSE는 46.7이고, 혼합 비/회귀 대체 모형 1과 2를 이용하여 대체했을 때의 평균 RMSE는 모두 45.7로 1이 감소하였다.

<표 2-39>에서는 조사값과 혼합 회귀 대체 모형 1($G=1$)을 이용하여 계산한 대체값에 대해 MAR 검정 결과를 보여준다. 참고로, 혼합 회귀 대체 모형 1과 2가

모두 $G=1$이면, 두 모형은 동일한 모형이다. 10-fold 교차 검증 결과와 마찬가지로 조사값을 사용할 경우 X^2는 27.1이고, 모형을 이용한 대체값은 모형 1에서 X^2=26.7로 근소한 차이로 감소하였다. 이에 따라 최종 대체 방법으로 조사값을 사용하기로 한다.

<표 2-38. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

	대체 방법		평균 RMSE	G
조사값 = 0	조사값 대체		15.0	-
	모형 대체 - 결정적		15.0	
	모형 대체 - 확률적		17.9	
조사값 > 0	조사값 대체		46.7	-
	혼합 비	모형 1	45.7	1
		모형 2 - 조건부: 교육정도 (초졸/중·고졸/대졸이상)	45.7	1
	혼합 회귀	모형 1	45.7	1
		모형 2 - 조건부: 교육정도 (초졸/중·고졸/대졸이상)	45.7	1

<표 2-39. 기초연금 소득 대체 후 MAR 검정 결과>

		X^2
조사값 대체		27.1
조사값 = 0: 결정적 모형 대체	조사값>0: 혼합 회귀 모형 1	26.7
	조사값>0: 혼합 회귀 모형 2 - 조건부: 교육정도	-

◦ 최종 대체 방법: 조사값 사용

<표 2-40>은 매칭 집단 별 조사 기초연금, 조사+행정 또는 대체 기초연금의 요약통계량을 정리한 표이다. 여기서, 대체값은 조사값과 동일한 값이다.

<표 2-40. 매칭, 비매칭 집단별 기초연금 (조사·조사+행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	24.44
	행정		0	0	0	24.39
비매칭	조사	5705	0	0	0	19.43
	대체		0	0	0	19.43
전체 (매칭,비매칭)	(조사,조사)	49518	0	0	0	23.87
	(조사+행정, 조사)		0	0	0	23.82
	(조사+행정, 대체)		0	0	0	23.82

5) 양육수당

5-1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 연령 (30대 미만, 30대-50대, 60대 이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: 혼합 비 대체 모형1, 조건부 혼합 비 대체 모형 2, (절편이 없는) 혼합 회귀 대체 모형1, (절편이 없는) 조건부 혼합 회귀 대체 모형 2

- 조건부 혼합 비 / 회귀 대체 모형 2의 보조 변수 x : 연령(30대 미만, 40대, 50대, 50대 이상)

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석

▫ 대체: 결정적 대체 방법

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b : 연령 (30대 미만, 30대-50대, 60대 이상)

◦ 범주형 변수 y_c : 조사 양육수당 및 조사+행정 양육수당의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

5-2) 분석 결과

<표 2-41>은 조사 양육수당이 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 연령이 30대 미만, 30-50대, 60대 이상인 경우 그룹 1에 속할 확률이 각각 1, 0.97, 그리고 1로 모두 그룹 1에 속할 확률이 높았다.

<표 2-41. 조사값이 0인 집단의 양육수당 모형 모수 추정 결과>

그룹	x:연령	30대 미만	30대-50대	60대 이상
1	$p_1(x)$	1	0.97	1
2	$p_2(x)$	0	0.03	0
	μ_0	4.65	4.51	4.76
	σ_0^2	0.84	0.96	0.69

<표 2-42>는 조사값이 0인 집단과 0보다 큰 집단에 관한 10-fold 교차 검증법 결과를 보여준다. 조사값이 0인 경우에 모형을 사용한 대체 방법(결정적)이 조사값을 사용한 경우와 동일한 평균 RMSE를 가진다. 앞서 살펴본 다른 기타 소득 항목과 마찬가지로 조사값이 0이면 조사+행정 자료값이 0인 경우가 매칭 집단에 압도적이었으므로, 대체 모형 또한 그룹 1에 속할 확률이 모든 연령 범주에서 1에 가까웠고, 따라서 결정적 대체 방법을 이용한 대체값 또한 모두 0으로 계산되었기 때문이다.

10-fold 교차 검증법을 통해 조사값이 0보다 큰 경우 혼합 비 대체 모형 1, 2와 절편이 없는 혼합 회귀 대체 모형 1, 2를 비교한다. 각 모형 종류별로 최대 혼합 성분 개수를 5로 두고, $G=1,2,...,5$를 혼합 성분으로 가지는 다섯 개의 모형을 평균 RMSE를 기준으로 비교한다. 각 모형 종류 별로 평균 RMSE가 가장 낮도록 하는 G값이 <표 2-42>에 제시되어 있다. 조사값을 사용했을 때의 평균 RMSE는 155.0이고, 혼합 비/회귀 대체 모형 1과 2를 이용하여 대체했을 때의 평균 RMSE는 모두 112 ~ 115 사이의 값으로 다소 감소한 것을 확인할 수 있다.

<표 2-43>에서는 조사값과 혼합 비 대체 모형 1($G=3$) 및 조건부 혼합 회귀 대체 모형 2($G=4$)를 이용하여 계산한 대체값에 대해 MAR 검정 결과를 보여준다. 조사값을 사용했을 때, X^2=10.0인 반면, 혼합 비 대체 모형을 이용한 대체값은 모형 1에서 X^2=17.5, 모형 2에서 X^2=21.3으로 모두 조사값 보다 더 높았으므로, 최종 대체 방법으로 조사값을 사용하기로 한다.

<표 2-42. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

	대체 방법		평균 RMSE	G
조사값 = 0	조사값 대체		23.3	-
	모형 대체 - 결정적		23.3	
	모형 대체 - 확률적		25.7	
조사값 > 0	조사값 대체		155.0	-
	혼합 비	모형 1	112.5	3
		모형 2 - 조건부: 연령 (10-20/30/40/50대이상)	113.8	4
	혼합 회귀	모형 1	115.0	3
		모형 2 - 조건부: 연령 (10-20/30/40/50대이상)	114.2	2

<표 2-43. 양육수당 대체 후 MAR 검정 결과>

		X^2
조사값 대체		10.0
조사값 = 0: 결정적 모형 대체	조사값>0: 혼합 비 모형 1	17.5
	조사값>0: 혼합 비 모형 2 - 조건부: 연령	21.3

◦ 최종 대체 방법: 조사값 사용

<표 2-44>는 매칭 집단 별 조사 양육수당, 조사+행정 또는 대체 양육수당의 요약통계량을 정리한 표이다. 여기서, 대체값은 조사값과 동일한 값이다.

<표 2-44. 매칭 집단별 양육수당 (조사·조사+행정·대체) 요약통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	3.1
	행정		0	0	0	4.1
비매칭	조사	5705	0	0	0	2.9
	대체		0	0	0	2.9
전체 (매칭,비매칭)	(조사,조사)	49518	0	0	0	3.0
	(조사+행정, 조사)		0	0	0	4.0
	(조사+행정, 대체)		0	0	0	4.0

6) 장애수당

6-1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 교육정도 (초졸 미만, 중·고졸, 대졸이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: 혼합 비 대체 모형1, 조건부 혼합 비 대체 모형 2, (절편이 없는) 혼합 회귀 대체 모형1, (절편이 없는) 조건부 혼합 회귀 대체 모형 2

- 조건부 혼합 비 / 회귀 대체 모형 2의 보조 변수 x : 교육정도 (초졸 미만, 중·고졸, 대졸이상)

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석

▫ 대체: 결정적 대체 방법

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b : 교육정도 (초졸 미만, 중·고졸, 대졸이상)

◦ 범주형 변수 y_c : 조사 장애수당 및 조사+행정 장애수당의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

6-2) 분석 결과

<표 2-45>는 조사 장애수당이 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 교육정도가 초졸 미만, 중·고졸 이상의 경우 그룹 1에 속할 확률이 각각 0.99와 1로 모든 교육정도 범주에서 그룹 1에 속할 확률이 높았다. (보조 변수로 연령대를 사용하더라도 유사한 결과가 나타남)

<표 2-45. 조사값이 0인 집단의 장애수당 모형 모수 추정 결과>

그룹	x:연령	초졸 미만	중·고졸	대졸 이상
1	$p_1(x)$	0.99	1	1
2	$p_2(x)$	0.01	0	0
	μ_0	3.96	4.23	4.21
	σ_0^2	0.41	0.78	1.19

<표 2-46>은 조사값이 0인 집단과 0보다 큰 집단에 관한 10-fold 교차 검증법 결과를 보여준다. 조사값이 0인 경우에 모형을 사용한 대체 방법(결정적)이 조사값

을 사용한 경우와 동일한 평균 RMSE를 가진다. 공적연금과 마찬가지로 조사값이 0이면 조사+행정 자료값이 0인 경우가 매칭 집단에 압도적이었으므로, 대체 모형 또한 그룹 1에 속할 확률이 모든 교육정도 범주에서 1에 가까웠고, 따라서 결정적 대체 방법을 이용한 대체값 또한 모두 0으로 계산되기 때문이다.

<표 2-46. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

	대체 방법		평균 RMSE	G
조사값 = 0	조사값 대체		6.9	-
	모형 대체 - 결정적		6.9	
	모형 대체 - 확률적		7.6	
조사값 > 0	조사값 대체		56.5	-
	혼합 비	모형 1	53.3	5
		모형 2 - 조건부: 교육정도 (초졸/중·고졸/대졸이상)	55.3	5
	혼합 회귀	모형 1	54.8	3
		모형 2 - 조건부: 교육정도 (초졸/중·고졸/대졸이상)	55.8	4

조사값이 0보다 큰 경우의 10-fold 교차 검증법 결과, 조사값을 사용했을 때의 평균 RMSE는 56.5이고, 혼합 비/회귀 대체 모형 1과 2를 이용하여 대체했을 때의 평균 RMSE는 모두 53~56 사이의 값으로 작은 크기로 감소하였다. 조사값과 혼합 회귀 대체 모형 1($G=1$)을 이용하여 계산한 대체값에 대해 MAR 검정 결과 또한, 조사값을 사용할 경우는 X^2는 10.7이고 이고, 모형을 이용한 대체값은 모형 1에서 X^2=12.8, 모형 2에서 X^2=10.9로 오히려 증가하였다. 따라서 최종 대체 방법으로 조사값을 사용하기로 한다.

<표 2-47. 장애수당 대체 후 MAR 검정 결과>

		X^2
조사값 대체		10.7
조사값 = 0: 결정적 모형 대체	조사값>0: 혼합 비 모형 1	12.8
	조사값>0: 혼합 비 모형 2 - 조건부: 교육정도	10.9

◦ 최종 대체 방법: 조사값 대체 선택

<표 2-48>은 매칭 집단 별 조사 장애수당, 조사+행정 또는 대체 장애수당의 요약통계량을 정리한 표이다. 여기서, 대체값은 조사값과 동일한 값이다.

< 표 2-48. 매칭 집단별 장애수당 (조사·행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	3.1
	조사+행정		0	0	0	3.6
비매칭	조사	5705	0	0	0	1.8
	대체		0	0	0	1.8
전체 (매칭,비매칭)	(조사,조사)	49518	0	0	0	2.9
	(조사+행정, 조사)		0	0	0	3.4
	(조사+행정, 대체)		0	0	0	3.4

7) 맞춤형 기초생활 보장 지원금

7-1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 연령 (60대 미만, 60대 이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: 혼합 비 대체 모형1, 조건부 혼합 비 대체 모형 2, (절편이 없는) 혼합 회귀 대체 모형1, (절편이 없는) 조건부 혼합 회귀 대체 모형 2

- 조건부 혼합 비 / 회귀 대체 모형 2의 보조 변수 x : 연령 (60대 미만, 60대 이상)

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석

▫ 대체: 결정적 대체 방법

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b : 연령 (60대 미만, 60대 이상)

◦ 범주형 변수 y_c : 조사 및 조사+행정 맞춤형 기초생활 보장 지원금의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

6-2) 분석 결과

<표 2-49>는 조사 맞춤형 기초생활 보장 지원금이 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 연령이 60대 미만과 60대 이상인 경우 그룹 1에 속할 확률이 각각 1과 0.99로 모든 연령 범주에서 그룹 1에 속할 확률이 1에 가깝다.

<표 2-50>은 조사값이 0인 집단과 0보다 큰 집단에 관한 10-fold 교차 검증법 결과를 보여준다. 조사값이 0인 경우에 모형을 사용한 대체 방법(결정적)이 조사값을 사용한 경우와 동일한 평균 RMSE를 가진다. 다른 기타 소득항목과 마찬가지로 조사값이 0이면 조사+행정 자료값이 0인 경우가 매칭 집단에서 압도적이었으므로, 대체 모형 또한 그룹 1에 속할 확률이 모든 연령 범주에서 1에 가까웠고, 따라서 결정적 대체 방법을 이용한 대체값 또한 모두 0으로 계산되기 때문이다.

<표 2-49. 조사값이 0인 집단의 맞춤형 기초생활 보장 지원금 모형 모수 추정 결과>

그룹	x:연령	60대 미만	60대 이상
1	$p_1(x)$	1	0.99
2	$p_2(x)$	0	0.01
	μ_0	5.42	5.68
	σ_0^2	1.63	0.91

조사값이 0보다 큰 경우의 10-fold 교차 검증법 결과, 조사값을 사용했을 때의 평균 RMSE는 264.7이고, 혼합 비/회귀 대체 모형 1과 2를 이용하여 대체했을 때의 평균 RMSE는 모두 약 263~265 근처의 값으로 작은 크기로 감소 혹은 증가하였다. 조사값과 혼합 회귀 비 모형 1($G=3$)과 모형 2($G=4$)를 이용하여 계산한 대체값에 대해 MAR 검정한 결과 또한, 조사값을 사용할 경우는 X^2는 6.7이고 이고, 모형을 이용한 대체값은 모형 1에서 X^2=6.7, 모형 2에서 X^2=7.5로 같거나 증가하였다. 따라서 최종 대체 방법으로 조사값을 사용하기로 한다.

<표 2-52>는 매칭 집단 별 조사 맞춤형 기초생활 보장 지원금, 조사+행정 또는 대체 맞춤형 기초생활 보장 지원금의 요약통계량을 정리한 표이다. 여기서,
대체값은 조사값과 동일한 값이다.

<표 2-50. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

	대체 방법		평균 RMSE	G
조사값 = 0	조사값 대체		34.9	-
	모형 대체 - 결정적		34.9	
	모형 대체 - 확률적		39.9	
조사값 > 0	조사값 대체		264.7	-
	혼합 비	모형 1	262.9	3
		모형 2 - 조건부: 연령 (60대 미만/60대이상)	263.8	4
	혼합 회귀	모형 1	262.8	3
		모형 2 - 조건부: 연령 (60대 미만/60대 이상)	265.3	1

<표 2-51. 맞춤형기초생활 보장 지원금 대체 후 MAR 검정 결과>

		X^2
조사값 대체		6.7
조사값 = 0: 결정적 모형 대체	조사값>0: 혼합 비 모형 1	6.7
	조사값>0: 혼합 비 모형 2 - 조건부: 연령	7.5

◦ 최종 대체 방법: 조사값 사용

<표 2-52. 매칭 집단별 맞춤형기초생활 보장 지원금 (조사·행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	8.3
	행정		0	0	0	10.4
비매칭	조사	5705	0	0	0	6.1
	대체		0	0	0	6.2
전체 (매칭,비매칭)	(조사,조사)	49518	0	0	0	8.0
	(조사+행정, 조사)		0	0	0	9.9
	(조사+행정, 대체)		0	0	0	9.9

8) 소득세

8-1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 연령 (30대 미만, 30대-50대, 60대 이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: 혼합 비 대체 모형1, 조건부 혼합 비 대체 모형 2, (절편이 없는) 혼합 회귀 대체 모형1, (절편이 없는) 조건부 혼합 회귀 대체 모형 2

- 조건부 혼합 비 / 회귀 대체 모형 2의 보조 변수 x : 연령 (30대 미만, 30대, 40대, 50대, 50대 이상)

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석

▫ 대체: 결정적 대체 방법

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b : 연령 (30대 미만, 30대, 40대, 50대, 50대 이상)

◦ 범주형 변수 y_c : 조사 및 조사+행정 소득세의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

8-2) 분석 결과

<표 2-53>은 조사 소득세가 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 연령이 30대 미만, 30대-50대, 60대 이상인 경우 그룹 1에 속할 확률이 각각 0.86, 0.54, 0.66으로 모두 그룹 1에 속할 확률이 높았다. 이에 따라, <표 2-34>의 조사값이 0인 집단에 관한 10-fold 교차 검증법 결과, 모형을 사용한 대체 방법(결정적)이 조사값을 사용한 경우와 동일한 평균 RMSE를 가진다.

<표 2-53. 조사값이 0인 집단의 소득세 모형 모수 추정 결과>

그룹	x:연령	30대 미만	30대 - 50대	60대 이상
1	$p_1(x)$	0.86	0.54	0.66
2	$p_2(x)$	0.14	0.46	0.34
	μ_0	1.31	2.12	2.08
	σ_0^2	1.71	2.44	2.53

10-fold 교차 검증법을 통해 조사값이 0보다 큰 경우 혼합 비 대체 모형 1, 2와 절편이 없는 혼합 회귀 대체 모형 1, 2를 비교해 보면, 조사값을 사용했을 때의 평균 RMSE는 759.4이고, 모든 후보 모형의 평균 RMSE가 약 759로 조사값을 사용할

때의 평균 RMSE와 거의 비슷하다.

<표 2-54. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

	대체 방법		평균 RMSE	G
조사값 = 0	조사값 대체		162.0	-
	모형 대체 - 결정적		162.0	
	모형 대체 - 확률적		161.8	
조사값 > 0	조사값 대체		759.4	-
	혼합 비	모형 1	759.1	1
		모형 2 - 조건부: 연령 (10-20/30/40/50대이상)	759.1	1
	혼합 회귀	모형 1	758.5	1
		모형 2 - 조건부: 연령 (10-20/30/40/50대이상)	758.5	1

<표 2-55>에서는 조사값과 혼합 회귀 대체 모형 1($G=1$)을 이용하여 계산한 대체값에 대해 MAR 검정 결과를 보여준다. 조사값을 사용할 경우 X^2는 257.8이고, 모형 1을 이용한 대체값은 X^2=288.3으로 조사값을 사용한 경우보다 값이 높았다. 이에 따라 최종 대체 방법으로 조사값을 사용하기로 한다.

<표 2-55. 소득세 대체 후 MAR 검정 결과>

		X^2
조사값 대체		257.8
조사값 = 0: 결정적 모형 대체	조사값>0: 혼합 회귀 모형 1	288.3

◦ 최종 대체 방법: 조사값 대체 선택

<표 2-56>은 매칭 집단 별 조사 소득세, 조사+행정 또는 대체 소득세의 요약통계량을 정리한 표이다. 여기서, 대체값은 조사값과 동일한 값이다.

<표 2-56. 매칭 집단별 소득세 (조사·조사+행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	15.0	61.2
	조사+행정		0	0	12.0	84.7
비매칭	조사	5705	0	0	14.0	51.1
	대체		0	0	14.0	51.1
전체 (매칭, 비매칭)	(조사,조사)	49518	0	0	15.0	60.0
	(조사+행정, 조사)		0	0	12.0	80.8
	(조사+행정, 대체)		0	0	12.0	80.8

2.2.3.3 경상소득

2.2.3.1절의 근로소득과 2.2.3.2절의 기타 소득 항목의 최종 대체 결과를 이용하여 경상소득의 대체값을 구할 수 있다. 경상소득은 다음과 같이 정의되며, 각 항목별 대체값을 다음 식에 대입하여 계산한다.

▫ 경상소득 = 근로 + 사업 + 재산 + 공적이전소득 + 사적이전소득

2.2.3.1절과 2.2.3.2절의 분석 결과에 따라 근로소득(근로)과 금융소득(재산)만 대체값을 사용하고 나머지 항목은 조사값을 그대로 사용하기로 한다. 즉, 경상소득 대체 전(조사값)과 후의 값(대체값)은 근로소득과 금융소득 대체값을 사용하기 전과 후를 의미한다.

매칭 집단에서 조사 경상소득과 조사+행정 경상소득을 비교해보면, 전체적으로 값이 조사값보다 조사+행정값이 높은 것을 알 수 있다. 비매칭 집단에서 조사 경상소득에 비해 대체 경상소득이 2사분위와 평균에서 조사값보다 높은 것을 알 수 있고, 전체 집단에서도 유사한 패턴을 보인다.

<표 2-57. 매칭 집단별 경상소득 (조사·조사+행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	672	2595	1801
	조사+행정		10	867	2692	1986
비매칭	조사	5705	0	480	2400	1571
	대체		0	516	2369	1599
전체 (매칭, 비매칭)	(조사,조사)	49518	0	640	2520	1775
	(조사+행정, 조사)		3	830	2646	1939
	(조사+행정, 대체)		8	835	2646	1942

2.2.4. 분석 결과: 부채

2.2.4.1. 분석 개요

소득 항목과 달리 부채 항목 자료는 건별 자료이므로 가구원 단위 부채 총액을 먼저 대체한 후 최근접 이웃 대체 (Nearest Neighborhood Imputation; NNI) 방법을 이용하여 건별 단위에 비(ratio) 대체를 적용할 수 있다. 분석하기 전 가구원 별 부채 총액은 행정 자료에서 취급하는 항목인지의 여부에 따라 다음과 같이 두 가지 종류로 구분할 수 있다.

▫ 부채 총액 1 – 대출유형이 대환대출, 신용카드 할부 잔액, 외상, 사인 간 거래이거나 대출기관이 대부업체일 경우 해당 부채 총액
▫ 부채 총액 2 – 그 외 부채 총액

부채 총액 1은 행정 자료에서 취급하지 않는 항목에 해당하는 총액으로 조사값을 그대로 사용하고, 부채 총액 2에 해당하는 부채 총액만 가구원 단위에서 대체한다. 그 후 부채 총액 1과 더함으로써 전체 부채 총액 대체값을 계산할 수 있다.

▫ 모형 및 대체 방법

가구원 별 부채 총액 또한, 조사값이 0인 가구원이 존재하므로, 조사값을 사용한다 하더라도 혼합 비 대체 방법을 바로 적용할 수는 없다. 2.2.3.2절의 기타 소득 항목에서 소개한 방법과 같이, 조사 부채가 0인 집단과 0보다 큰 집단을 나누어 독립적으로 대체 모형 적합 후 대체한다. (2.3.3.2 기타 소득 항목 참고)

2.2.4.2. 분석 적용 및 결과

1) 분석 절차

▫ 조사값이 0인 집단: 모형 (5) 가정

◦ 보조 변수 x : 연령 (30대 미만, 30대-50대, 60대 이상)

▫ 조사값이 0보다 큰 집단:

◦ 후보 모형: (절편이 없는) 혼합 회귀 대체 모형1, (절편이 없는) 조건부 혼합 회귀 대체 모형 2

- 조건부 혼합 회귀 대체 모형 2의 보조 변수 x : 연령(30대 미만, 30대, 40대, 50대, 50대 이상)

◦ G 선택: 각 후보 모형 별 10-fold 교차 검증법을 통해 평균 RMSE 기준 가장 좋은 혼합 성분 개수 G 선택

- 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석

▫ 대체: 결정적 대체 방법

◦ 조사값이 0 인 집단 : 부분 확률적 대체 (2.4.1.2. 기타 소득 항목 참고)

◦ 조사값이 0보다 큰 집단 : 결정적 대체

▫ MAR 검정 통계량(X^2) 비교 후 최종 모형 선택

◦ 범주형 변수 x_b: 연령 (30대 미만, 30대, 40대, 50대, 50대 이상)

◦ 범주형 변수 y_c: 조사 부채 및 조사+행정 부채의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

2) 분석 결과

2017년 조사 가구원 전체를 대상으로 부채 항목에 관하여 분석한 결과는 다음과 같다. 먼저 조사값이 0인 집단의 분석 결과를 살펴보자. <표 2-58>는 조사값이 0인 집단에 가정한 모형 (5)의 모수 추정값을 나타낸다. 연령이 30대 미만, 30~50대, 그리고 60대 이상인 경우 그룹 1에 속할 확률이 각각 0.91, 0.64, 0.82로 모두 그룹 1에 속할 확률이 높았다. 이에 따라, 만약 조사값이 0인 집단에서 결정적 대체 방법을 적용할 경우, 기타 소득 항목의 분석 결과에서 살펴본 바와 같이, 조사값과 동일한 대체값, 즉 0이 계산된다. 그러나 2.3.2.2절에서 소개한 부채 기초 자료 분석 내용을 보면, 부채는 다른 기타 소득 항목과는 달리, 조사값이 0일 때 조사+행정값의 분포가 조사값(즉, 0)과 많이 다른 것을 확인할 수 있다. 만약 결정적 대체 방법을 이용하여 비매칭 집단에 대체값으로 조사값을 그대로 사용할 경우, 실제 비매칭 집단의 조사+행정 값의 분포와 많이 다를 것으로 예상된다. 이에 따라 본 최종보고서에서는 부분 확률적 대체 방법을 적용한다.

<표 2-58. 조사값이 0인 집단의 부채 모형 모수 추정 결과>

그룹	x:연령	30대 미만	30대-50대	60대 이상
1	$p_1(x)$	0.91	0.64	0.82
2	$p_2(x)$	0.09	0.36	0.18
	μ_0	6.46	7.45	7.37
	σ_0^2	1.96	2.74	2.45

<표 2-59>는 조사값이 0인 경우와 0보다 큰 경우에 관한 10-fold 교차 검증법 결과를 보여준다. 조사값이 0인 경우, 모형을 사용한 확률적 대체 방법의 평균 RMSE가 조사값 대체 및 모형을 사용한 결정적 대체의 평균 RMSE보다 더 낮은 것을 확인할 수 있다.

조사값이 0보다 큰 경우, 10-fold 교차 검증 결과를 통해 대체값으로 조사값을 그대로 사용할 경우와 절편이 없는 혼합 회귀 모형 - 모형 1 (기본 모형)과 조건부 모형 2(확장된 모형)의 결정적 대체 결과를 비교할 수 있다. 모형 1과 모형 2 모두 조사값을 대체값으로 사용했을 때 보다 평균 RMSE가 상당히 감소한 것을 확인할 수 있으며 (표 2-59), MAR 검정 결과 또한, 조사값을 사용했을 때보다 모형 1 또는 2를 사용했을 때 값이 눈에 띄게 감소한 것을 확인할 수 있다 (표 2-60). <표 2-60>에서 조사값이 0인 집단에서는 확률적 대체 방법을 적용하므로 MAR 검정 통계량 값 또한 분석을 수행할 때마다 값이 바뀔 수 있으나 동일한 결론을 확인할 수 있으며, 반복 대체 분석 결과, 모형 1과 모형 2에서 거의 유사한 MAR 검정 통계량 값 계산 되었으므로, 간결성의 원칙(principle of parsimony)에 의해 혼합 회귀 대체 모형 1을 최종 대체 모형으로 선택한다.

<표 2-59. 대체 방법별 평균 RMSE 비교 및 G 선택 결과>

	대체 방법		평균 RMSE	G
조사값 = 0	조사값 대체		5085.8	-
	모형 대체 - 결정적		5085.8	-
	모형 대체 - 부분 확률적		5045.1	-
조사값 > 0	조사값 대체		16491.1	-
	혼합 회귀	모형 1	11276.1	4
		모형 2 - 조건부: 연령 (10-20/30/40/50대이상)	11286.9	4

<표 2-60. 부채 대체 후 MAR 검정 결과>

		X^2
조사값 대체		209.5
조사값 = 0: 부분 확률적 모형 대체	조사값>0: 혼합 회귀 모형 1	76.3
	조사값>0: 혼합 회귀 모형 2 – 조건부: 연령	76.8

◦ 최종 대체 모형: 혼합 회귀 모형 1 선택

<표 2-61. 조사값 > 0 인 집단의 부채 최종 모형 모수 추정 결과>

그룹	μ_g	σ_g^2	β_g	$\sigma_{e,g}^2$	π_g
1	8.73	0.75	1.12	31626259	0.43
2	9.37	2.27	0.27	1221098198	0.09
3	5.75	2.95	4.14	4066808	0.06
4	7.56	1.65	1.02	406453	0.42

<표 2-62. 매칭 집단별 부채 (조사·조사+행정·대체) 요약 통계량>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	43813	0	0	0	1605
	행정		0	0	1000	2465
비매칭	조사	5705	0	0	0	1300
	대체		0	0	1582	1423
전체 (매칭,비매칭)	(조사,조사)	49518	0	0	0	1570
	(조사+행정, 조사)		0	0	902	2364
	(조사+행정, 대체)		0	0	1087	2379

<표 2-61>에서와 같이 최종 대체 모형의 모수를 추정한 후 결정적 대체값을 구한 다음, 매칭 집단 별 조사 부채, 조사+행정 또는 대체 부채의 요약통계량을 정리하면 <표 2-62>와 같다. 매칭 집단을 먼저 살펴보면, 조사 부채는 1,2, 그리고 3사분위수가 모두 0이고, 평균이 1605인 반면, 조사+행정 부채의 3사분위수는 1000이고 평균은 2465로 조사 부채와 큰 차이를 보이고 있다. 비매칭 집단에서 조사 부채과 대체 부채의 요약통계량을 살펴보면, 조사 부채의 1,2, 그리고 3사분위수가 모두 0이고 평균이 1300인 반면, 모형을 이용하여 대체한 부채의 3사분위수는 1582이고, 평균은 1423이다. 전체 (매칭, 비매칭) 집단에서도 유사한 패턴을 확인할 수 있으며, (조사+행정, 조사) 부채의 3사분위수와 평균보다 (조사+행정, 대체) 부채의

3사분위수와 평균이 조금 더 높지만 그 차이가 크지는 않다.

3. 금융소득 시계열 보정

조사연도 기준 2015년부터 금융소득 및 임대소득의 행정자료 활용이 가능해짐에 따라서 (조사+행정) 재산소득/소득세가 추가적으로 관측이 되고 있다. 따라서, 2012-2014년도에[1] 해당하는 가구 재산소득/소득세를 대체하여 (조사+행정) 재산소득의 시계열을 연장할 요인이 발생하였다. 따라서 본 장에서는 2012-2014년도에 해당하는 조사+행정 재산소득과 소득세를 대체하는 일련의 통계적 방법을 소개하는 것을 목표로 한다.

3.1 데이터

<표 3-1>은 연도별 평균 가구 재산소득 및 소득세 현황을 보여주고 있다. 2012-2014년까지는 조사소득만 존재하기 때문에 조사소득이 표기되었고 2015년부터는 (조사+행정) 소득이 함께 표기되었다. 재산소득 및 소득세 모두 (조사+행정) 값 대비 조사값이 과소 집계되고 있음을 확인할 수 있다. 소득의 경우에는 실제 소득에 비하여 낮게 응답하는 경향이 있는 편이고, 소득세의 경우에는 실제 과세된 세금액을 정확하게 인지하지 못하기 때문인 것으로 판단된다.

<표 3-1. 연도별 가구 평균 재산소득 / 소득세, 단위-만원>

	2012	2013	2014	2015		2016		2017	
				조사	조+행	조사	조+행	조사	조+행
재산	189	204	190	191	336	221	378	209	343
소득세	131	135	148	154	167	156	208	175	222

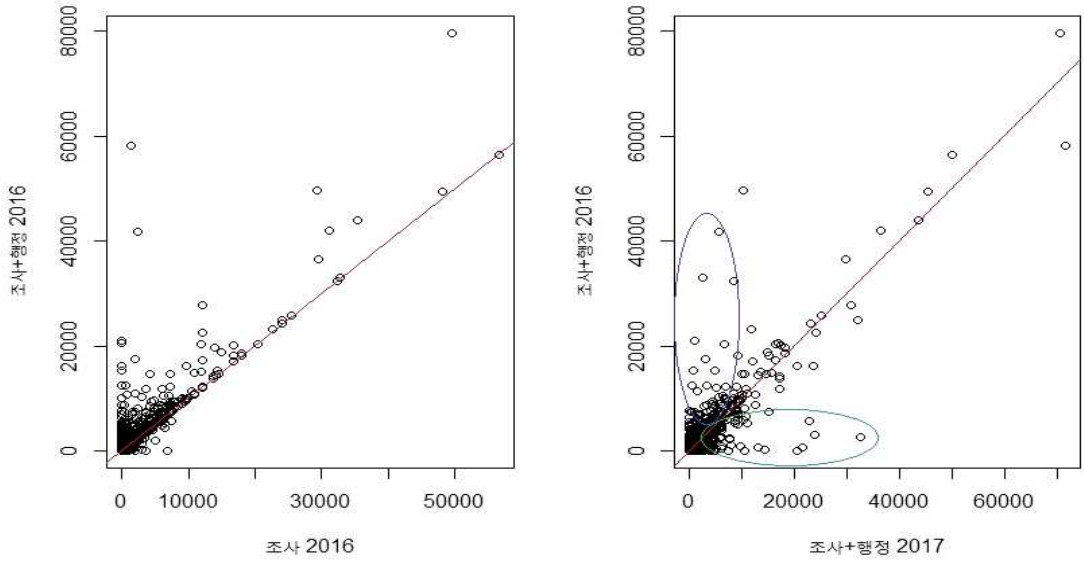
이러한 과소 추정 문제를 제외하고도, 안정적인 시계열 연장을 위해서는 조사값과 조사+행정값 관계 그리고 연도별 변화를 살펴보는 것이 필요하다. 재산소득과 소득세의 관계 및 구조가 다소 상이하기 때문에 아래에 별도로 내용을 요약하였다.

3.1.1 재산소득

1) 본 보고서는 조사연도 기준으로 작성이 되었음.

<그림 3-1>은 2016년도 기준 조사값과 (조사+행정)값의 산점도와 2016-2017년의 (조사+행정)값 변화 산점도를 나타내고 있다. <표 3-1>에서 이미 확인했듯이 저소득층 일부를 제외한 대부분의 소득층에서 조사값이 과소 추정되고 있음을 알 수 있다. (조사+행정)값의 연도별 변화는 다음의 세 가지 형태로 구분이 될 수 있음을 알 수 있다: i) 연도간 차이가 크지 않은 경우, ii) 큰 폭으로 증대된 경우, iii) 큰 폭으로 감소한 경우. ii)번과 iii)번의 변화는 하나의 통계모형으로 통합하기 쉽지 않으나 회귀선을 중심으로 양쪽에 대칭적으로 퍼져있기 때문에 예측으로 인한 오차가 상쇄될 수 있는 측면이 있다.

<그림 3-1 왼쪽 패널: 재산소득 조사 vs 조사+행정; 오른쪽 패널: 재산소득 조사+행정 2017 vs 재산소득 조사+행정 2016[2)]>



<표 3-2 조사연도별 조사/(조사+행정) 비>

	2012	2013	2014	2015	2016	2017
근로소득	0.93	0.93	0.92	0.95	0.94	0.94
재산소득				0.57	0.58	0.61

<표 3-2>는 조사연도별 조사/(조사+행정) 값 비(ratio)를 나타내고 있다. 재산소득과의 비교를 위하여 근로소득의 연도별 비 또한 같이 표기하였다. 근로소득은 조

2) 연도별 비교의 경우 두 연도에 걸쳐 모두 관측된 가구만을 활용하였다.

사연도에 상관없이 안정적인 것에 반하여 재산소득은 상대적으로 연도별로 비값에 차이가 있다. 또한 <표 3-2>는 조사에서 근로소득과는 달리 재산소득에 대한 과소 집계가 상당함을 확인할 수 있다.

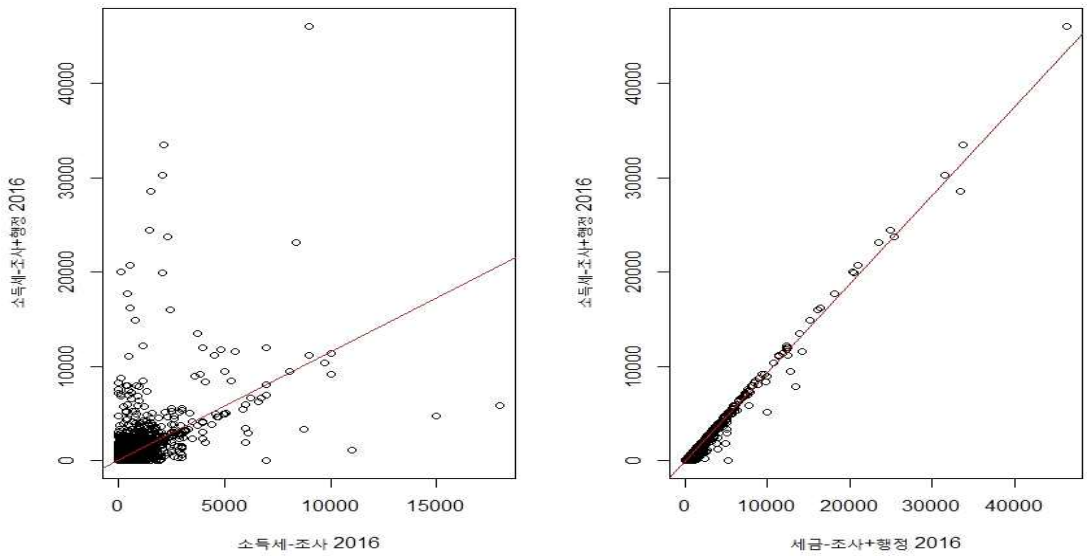
<표 3-3 조사(t)/조사(t+1), 조사+행정(t)/조사+행정(t+1) 비>

	2015/2016	2016/2017
재산소득(조사)	0.86	1.06
재산소득(조사+행정)	0.89	1.10

<표 3-3>은 조사값과 (조사+행정)값의 연도간 변화를 표기하고 있다. 연도간 변화에서는 차이가 있지만 고정된 두 개 연도 내에서는 조사값의 변화율과 조사+행정의 변화율이 유사함을 확인할 수 있으며, 이러한 유사성은 대체모형을 개발하는데 중요한 보조 정보로 사용될 수 있다. 즉, 비록 재산소득에서 조사값 자체는 과소 집계 되었지만 그 증가폭이나 감소폭은 조사+행정의 그것과 비슷한 경향을 보여주고 있으므로 이 정보가 대체 모형에 유용하게 사용될 수 있다.

3.1.2 소득세

<그림 3-2 왼쪽 패널: 소득세 조사 vs 소득세-조사+행정; 오른쪽 패널: 세금-조사+행정 vs 소득세-조사+행정>



<그림 3-2>에서 왼쪽 패널은 2016년도의 소득세의 조사값과 (조사+행정)값의

관계를 나타내고 오른쪽 패널은 세금의 (조사+행정값)과 소득세의 (조사+행정)값의 관계를 표기하고 있다. 조사와 (조사+행정)값 간에 차이가 많이 나는 가구가 많은 것을 확인할 수 있다. 소득세는 세금의 하위항목이기 때문에 세금의 (조사+행정)값과 밀접하게 연관 되어 있는 것은 당연하므로 시계열 측면에서 세금 중 소득세가 차지하는 비중이 얼마인지를 확인하고 또한 그 비중이 어떻게 변화하는지를 자세히 살펴볼 필요가 있다.

<표 3-4. 연도별 소득세/세금 비중, 단위-%>

2013		2014		2015		2016		2017	
조사	조+행	조사	조+행	조사	조+행	조사	조+행	조사	조+행
70	75	72	75	73	74	72	78	74	78

<표 3-4>는 세금 중에서 소득세가 차지하는 비중을 %로 표기하고 있다. 2015년부터 행정자료값이 취합된 재산소득과 달리 세금과 소득세는 2013년에서 2017년까지 일부 행정 자료값이 대체된 (소득+행정) 자료값이 존재한다. 하지만 2013/2014년, 2015년, 2016/2017년에 행정자료값이 대체된 규모는 상이하다. 조사값 기준으로는 약 70-74% 정도를 차지하며 조사연도에 걸쳐서 안정적인 편이다. (조사+행정) 값 기준으로는 소득세 비중이 74~78%로 약간 올라가나 연도별 변화폭은 안정적인 편이다. 이는 비추정을 사용할 수 있는 주요한 근거가 될 수 있다.

세금 정보 외에 추가적으로 소득세의 조사값이 (조사+행정)값과 추정에 도움이 될 수 있는지 확인해볼 필요가 있다. 재산소득에서 고려한 것과 마찬가지로, 소득세의 연도별 조사/(조사+행정)비와 연도별 조사값 및 (조사+행정) 값 비를 계산하였다.

<표 3-5 소득세 조사연도별 조사/(조사+행정) 비>

2013	2014	2015	2016	2017
0.78	0.87	0.92	0.75	0.79

<표 3-6 조사(t)/조사(t+1), 조사+행정(t)/조사+행정(t+1)>

	2015/2016	2016/2017
소득세(조사)	0.99	0.89
소득세(조사+행정)	0.80	0.94
세금(조사)	0.99	0.91
세금(조사+행정)	0.84	0.94

<표 3-5>는 조사연도별 소득세의 조사와 조사+행정 값의 비를 나타내고 있다.

조사연도별로 그 비가 상당히 달라짐을 확인할 수 있다. 이는, 재산소득과는 달리 조사값만을 단순하게 이용하는 대체모형은 예측오차가 커질 수 있음을 의미한다. 또한, <표 3-6>에 의하면, 소득세 조사값의 연도간 비와 (조사+행정)값의 비가 상이한 것을 확인할 수 있다. 하지만 세금의 조사값 비변화와 (조사+행정)값 비변화와 소득세의 변화비는 유사한 것을 확인할 수 있다. 즉 2015/2016년 소득세의 변화비가 0.80이고 세금의 변화비는 0.84 이고, 2016/2017의 소득세 와 세금의 변화비는 0.94로 동일하다. 이는 <표 3-4>의 결과와도 일치 되는 것으로 세금중 소득세 부분이 안정적인 비율을 유지하고 있음을 재확인시켜준다.

이러한 결과를 종합하면, 재산소득과 달리 소득세의 (조사+행정)값 대체모형은 세금 (조사+행정)값을 활용하는 방향으로 개발해야 한다는 결론을 도출할 수 있었다.

3.1.3 데이터 구조

재산소득 및 소득세의 (조사+행정) 값의 시계열 연장을 위한 대체모형을 개발하기 위해서는 데이터 구조를 파악해야 한다. 가계금융복지조사는 패널조사로 2014년까지는 고정패널로 실시되다가 2015년부터는 연동패널이 적용되어, 약 1/5해당 되는 패널이 새로운 패널로 매년 대체되고 있다.

<표 3-7 가계금융복지조사 연동패널 구조>

조사연도	2012	2013	2014	2015	2016	2017
고정패널	A1	A1	A1			
	B1	B1	B1	B1		
	C1	C1	C1	C1	C1	
	D1	D1	D1	D1	D1	D1
	E1	E1	E1	E1	E1	E1
연동패널				A2	A2	A2
					B2	B2
						C2

<표 3-7>은 가계금융복지조사의 연동패널 구조를 보여주고 있다. 패널 A1-E1은 2012-2014년에 걸쳐서 모두 관측이 되었는데, 이 중에서 패널 A1은 2015년부터는 패널 A2로 대체되었다. 따라서 패널 A1의 2015년 (조사+행정) 재산/소득세 값은 관측되지 않았다. 즉 대체모형 적용 대상 패널 가구는 2012-2015년에 걸쳐 모두 관측된 가구와 최소 한 번은 관측되지 않은 가구로 구분되며, 이는 각 응답 패턴에 맞게 횡단면 및 종단면 대체모형을 동시에 고려해야 한다는 것을 의미한다.

<표 3-8 2012-2015 가구응답 패턴 (총 24,869 가구)>

패턴	가구 수	패턴	가구 수
(1, 1, 1, 1)	13,034	(0, 1, 1, 1)	354
(1, 1, 1, 0)	3,939	(0, 1, 1, 0)	111
(1, 1, 0, 0)	1,103	(0, 1, 0, 0)	53
(1, 0, 1, 1)	101	(0, 0, 1, 1)	206
(1, 0, 1, 0)	53	(0, 0, 1, 0)	65
(1, 0, 0, 0)	1,514	(0, 0, 0, 1)	4,336

<표 3-8>은 2012-2015년에 한 번이라도 관측된 가구들의 4개 연도 응답 패턴이다. (1,1,1,1)은 네 개 조사연도에 걸쳐 모두 응답했다는 것을 의미하며 13,034가구가 이에 해당된다. (1,0,0,0)은 2012년에만 응답하고 2013년 이후부터는 서베이에 참여하지 않은 가구를 의미한다. 여기에서 2015년에만 응답한 4,336 가구는 대체가 필요하지 않은 가구로 대체대상에서 제외된다.

3.2 대체 (imputation model) 모형

2014년 (조사+행정)값을 생성하는 작업을 가정하면, 이용 가능한 정보는 2015년도의 (조사+행정)값과 2014년의 조사값이다. 익년도 정보는 종단면 대체모형(longitudinal imputation model)을 통하여 이용하고, 익년도 (조사+행정)값이 없는 가구의 경우엔 대상연도의 정보만을 이용하는 횡단면 대체모형 (cross-sectional imputation model)을 사용할 수 있다. 종단면 대체모형으로서, 본 연구용역에서 사용할 기본 방법은 `비' 대체 (ratio imputation) 방법이다. 2012-2014년 해당하는 금융소득값은 보조변수의 연도간 변화율과 익년도 금융소득의 곱으로 표현할 수 있는데, 이는 '비'추정 방법의 대표적인 형태이다. 재산소득의 경우 조사값과 (조사+행정)값의 연도별 변화가 유사하고, 소득세의 경우에는 세금의 보조변수가 존재하기 때문에, 비추정량이 좋은 성질을 가질 수 있다. 또한 비추정량은 시계열 속성을 유지하는데 있어서 장점이 있어서, 비추정량을 종단면 대체의 기본 대체 방법론으로 선택하였다.

$$Y_{it} = R_t Y_{i,t+1} \qquad (3\text{-}1)$$

여기에서 $R_t = \overline{X_t}/\overline{X}_{t+1}$로 정의하되고, Y_t는 t 조사연도의 (조사+행정) 값, X_t는 보조변수의 t 조사연도 값을 의미한다. 재산소득 대체에서는 재산소득의 조사값이 X가 되고, 소득세에서는 세금이 그 역할을 하게 된다. 식 (3-1)을 보면, 대체 대상

연도의 (조사+행정)값은 익년도의 (조사+행정)값에 대체 조사연도에 맞춰 계산된 R_t를 곱하여 얻어짐을 알 수 있다.

재산소득의 횡단면 대체모형의 경우 익년도 조사+행정값을 활용할 수 없다. 즉 해당 연도에 관측된 변수만을 이용해서 대체를 해야 되기 때문에 '비'추정 방법대신 회귀 추정량을 이용한 회귀 대체 (regression imputation) 형태로 진행할 예정이다.

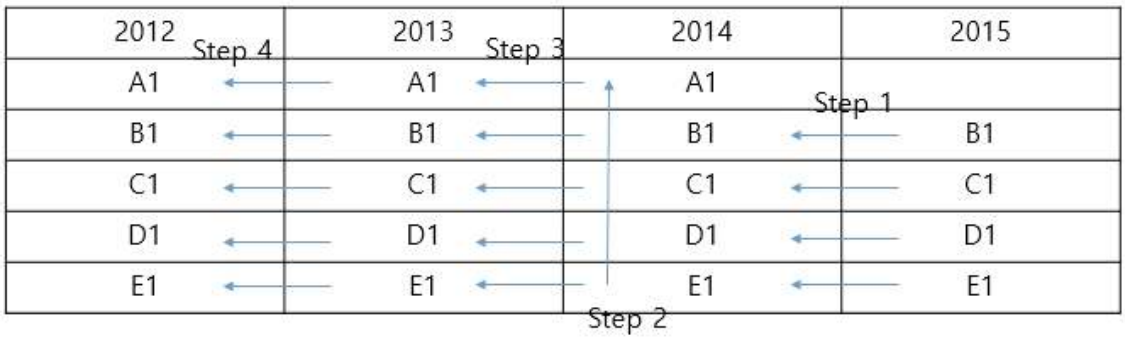
$$Y_{it} = \phi_1 X_{it} + e_{it} \qquad (3\text{-}2)$$

기본 모형은 식 (3-2)의 소개된 일반 선형 회귀모형을 사용하고, 설명변수는 X_{it}에는 해당연도의 조사값을 사용한다. 재산소득의 조사값과 (조사+행정)값이 0인 경우가 많기 때문에 절편항은 포함하지 않았고, 또한 소득에 많이 쓰이는 로그 변환을 사용하지 않았다. 이는 변환을 통하여 얻는 추정의 이점보다 역변환시 발생하는 예측 오차를 줄여주는 것이 더 중요하다고 판단했기 때문이다.

소득세의 횡단면 대체모형의 경우에는 회귀 대체모형대신 비대체모형을 사용하였다. 소득세를 실제값과 그대로 응답한 그룹과 그렇지 않은 그룹으로 나뉠 수 있는데, 이러한 그룹이 혼재되어 있는 상태에서 조사값은 중요한 설명변수의 역할을 하지 못했다. 따라서 예측 오차가 편향될 수 있는 점을 고려해서 회귀 대체모형 대신 비대체모형을 차용하였다.

3.2.1 기본 전략

<그림 3-3. 대체과정 요약도>



<그림 3-3>을 보면 (조사+행정)값 대체과정이 요약되어 있다. 2015년 (조사+행정값) 및 종단면 대체 모형을 이용하여 우선적으로 패널 B1-E1에 해당하는 (조사+행정)값을 만들어내는 것이 (Step 1)이다. 그런데 패널 A1의 경우 2015년 (조사+행정)값이 없으므로 횡단면 대체모형을 이용하여 (조사+행정)값을 생성해야 하며, 이

것이 (Step 2)이다. 이후 순차적으로 종단면 대체모형을 활용하여 2013년, 2012년 (조사+행정) 값을 생성하면 되는데, 이 과정이 (Step 3)과 (Step 4)가 된다. 일부 패널의 경우 익년도 (조사+행정)값이 존재하지 않는 경우가 있는데, 이러한 패널에 대해서는 횡단면 대체모형을 활용하여 해당 (조사+행정)값을 생성한다.

<표 3-9 응답패턴별 대체모형 적용 유형 (20,533 가구)>

유형	2012-2015 조사연도 응답패턴	가구 수
유형 1	(1,1,1,1), (0,1,1,1), (0,0,1,1)	13,594
유형 2	(1,0,0,0), (0,1,0,0), (0,0,1,0)	1,632
유형 3	(1,1,1,0), (1,1,0,0), (0,1,1,0)	5,153
유형 4	(1,0,1,1), (1,0,1,0)	154

<표 3-9>는 응답패턴별로 횡단면 및 종단면 모형이 적용되는 사례들을 세 가지 유형으로 정리한 결과이다. 첫 번째 유형은 종단면 대체모형만 적용되는 경우이다. 무응답인 조사연도가 없는 경우나, 조사값이 보고되기 시작한 조사연도부터 2015년도까지 모두 응답한 가구는 종단면 대체모형만 순차적으로 적용하면 된다. 이에 반해, (유형 2)는 횡단면 대체모형만 사용하는 경우이다. 즉 익년도의 (조사+행정)값이 없기 때문에 종단면 대체모형 대신 횡단면 대체모형만을 이용하여 (조사+행정) 값을 생성한다. (유형 3)은 첫 대체는 횡단면 모형을 이용하고 그 이후에는 종단면 대체모형만 순차적으로 이용하는 경우이다. (유형 4)의 경우에는 종단면 대체모형과 횡단면 대체모형이 반복적으로 사용되어야 하는 경우이다.

3.2.2 종단면 대체모형 (Longitudinal Imputation Model)

1) 재산소득

2014년 재산소득 예측에서는 다음의 혼합 종단면 대체모형을 사용하였다.

$$\widehat{Y_{it}} = \begin{cases} \widehat{R_t} Y_{i,t+1} & \text{if } X \le q_v \\ (X_{it} - X_{i,t+1}) + Y_{i,t} & \text{if } X > q_v \end{cases} \quad (3\text{-}3)$$

기본적으로는 비추정 모형 (3-1)을 사용하나, 조사값이 q_v가 넘어가는 경우에는 회귀 대체모형을 사용하였다. 시계열 연속성 확보를 위해서는 회귀 대체모형보다는 비대체모형이 좋지만, 비대체모형이 불편의 추정량이 아니기 때문에 $Y_{i,t+1}$값이 큰 경우에 예측 오차가 커질 위험성이 있다. 따라서, 일정한 임계치를 중심으로 회귀모

형과 비대체모형을 나눠서 적용하는 형태로 대체 모형을 개발하였다. 2014년 (조사+행정) 예측값이 생성된 후, 2012년, 2013년 대체에서는 식 (3-1)를 그대로 사용한다. 2012, 2013년 (조사+행정) 값 대체시에는 2014년에 대체된 값을 사용해기 때문에 단순 비모형 만을 사용하였다.

2) 소득세

소득세의 경우에는 <표 3-6>에서 확인한 것처럼 세금의 (조사+행정) 변화비[3]를 R_t로 사용한다. 단 소득세의 경우에는 세금의 (조사+행정)값을 활용하기 때문에, 혼합대체 모형 (3-3) 대신에 기본적인 비추정 모형 (3-1)을 그대로 사용한다.

3.2.3 횡단면 대체모형 (cross-sectional imputation model)

횡단면 예측에서는 익년도 값을 사용하지 않고 해당 연도에 관측된 값만을 이용하여 만들어야 한다. 따라서 회귀 대체모형 (Regression imputation model)을 기본 종단면 모형으로 사용하되, 적절한 설명변수를 선택하는 것이 주요 과제 내용이 된다. 재산소득과 소득세의 특성이 다르기 때문에 회귀 대체모형은 독립적으로 선택하였다.

1) 재산소득

회귀 대체모형으로는 일반적인 선형 모형을 가정한다. 기본 모형으로는 조사값만을 설명변수로 포함한 선형 모형 식 (3-2)을 가정하였다. 따라서 식 (3-2)의 X_{it}는 i번째 가구의 재산소득이 된다. (그림 3-1)에서 확인할 수 있듯이 일부 특이값을 제외하고는 조사값에 따른 (조사+행정)값의 범위가 일정함으로 오차항에는 등분산을 가정하였다. 이러한 기본 모형에 가구 변수, 소득 변수등을 추가해가며 대체 모형을 찾는 작업을 진행했다. 횡단면 대체 모형의 대상 가구가 많지 않을뿐더러, 설명변수가 많을 경우 최적합 회귀 모형은 오히려 예측 변동성이 커질 수 있기 때문이다. 모형 선택과정에서 고려된 설명변수 후보군은 <표 3-10>에 요약되어 있다.

3) 2013년까지는 세금의 (조사+행정) 변화비를 사용하고 2012년도에는 조사값의 변화비를 사용한다.

<표 3-10. 횡단면 대체모형 설명변수 후보>

설명 변수	비고
조사 근로소득	
조사 사업소득	
자산총액 5분위	
순자산액 5분위	
조사 세금	
가구주 나이	20-30대/40대-50대/60대 이상
1인 가구 여부	1인 가구 / 2인 이상
가구주 교육 정도	무학-중학교/고등학교/대학교이상
가구주 종사상 지위	상용근로자/자영업자/그 외
주택의 종류	단독주택/아파트/그 외
입주 형태	자가 / 그 외
거주지	수도권 / 광역시 / 그 외

모형의 간결함을 고려하여 최종적으로 다음의 두 모형이 최종 횡단면 회귀 대체모형으로 선택되었다.

$$Y_{it} = \phi_1 X_{it} + e_{it} \quad \text{if } (X_{it} > q_v) \tag{3-4}$$

$$Y_{it} = \phi_1 X_{it} + \phi_2 Z_{1it} + \phi_3 Z_{2it} + e_{it} \text{ if } (X_{it} \leq q_v) \tag{3-5}$$

여기에서 Z_{1it}과 Z_{2it}는 각각 i번째 가구의 경상소득과 저축액을 의미하고는 $q_v(=2000)$[4]는 종단면 대체모형에서 소개되었던 임계점으로 변동 계수 (C.V.)를 고려하여 선택되었다. 즉 2,000만원 보다 큰 구간에서는 조사값만 사용한 회귀대체모형을 사용하여, 그렇지 않은 구간에서는 경상소득과 저축액을 사용한다. 재산소득의 횡단면 대체모형을 두 가지로 나뉜 이유는 i) 조사소득이 0인 가구가 많고, ii) 모형의 간결함을 유지하고, iii) 종단면 혼합 대체모형과의 일치성을 유지하기 위해서이다.

<표 3-11 조사 재산소득이 0인 가구 수 (전체 가구 대비 비중)>

2015	2016	2017
13,729 (0.76)	13,778 (0.75)	13,865 (0.75)

<표 3-11>에서 확인할 듯이 약 70%이상의 가구의 조사소득이 0이다. 따라서 조사소득만을 변수로 사용하는 횡단면 대체모형은 만들 수가 없다. 따라서 추가 변수가 필요하기 때문에 변수 선택 과정을 통하여 경상소득과 저축액을 추가한 모형을

4) q_v는 최적 추정량은 아니며 변동계수가 안정화되는 부근에서 trial and error로 설정하였다.

선택하였다. 경상소득의 경우 재산소득의 상위 소득이지만 다른 소득이 변수로 들어가지 않기 때문에 설명 변수로 사용하는데 문제가 없다. 조사소득 값이 충분히 큰 구간에서는 대체적으로 조사값과 (조사+행정)값의 괴리가 크지 않기 때문에 변수 추가로 인한 변동성이 더 위험할 수 있다고 판단하였다. 또한, 종단면 혼합 대체 모형과의 모형 일치성도 최종 모형 선택에 고려하였다.

회귀 모형 적합에서는 소득변수에 많이 사용하는 로그변환을 사용하지 않았다. 앞서 설명하였듯이 로그 변환의 경우 전환과정에서 예측력의 편의를 가져올 위험성이 크기 때문에 본래의 스케일 값 그대로 사용하였으며, 회귀 모형의 계수값은 OLS (Ordinary Least Squares) 방법을 이용하여 추정하였다. 추정시에는 임계값에 따라서 데이터를 분리하여 추정 하지 않고 전체 데이터를 활용하여 모형 (3-4)와 모형 (3-5)의 회귀계수값을 계산하였다. 데이터를 분리하여 추정할 경우 ϕ_1 계수값의 추정치가 크게 달라 질 수 있는데, 이는 조사값이 과소추정 되는 `비'를 다르게 적용하는 형태로 모든 가구에 동일한 비율을 적용하는 종단면 비대체 모형의 사용과 배치될 수 있기 때문이다. 마지막으로 회귀 절편을 제외하고 추정하였고 예측값은 0보다 같거나 큰 값이 되도록 하였다.

2) 소득세

소득세는 재산소득과는 달리 다음의 두 가지 특징을 갖는다:

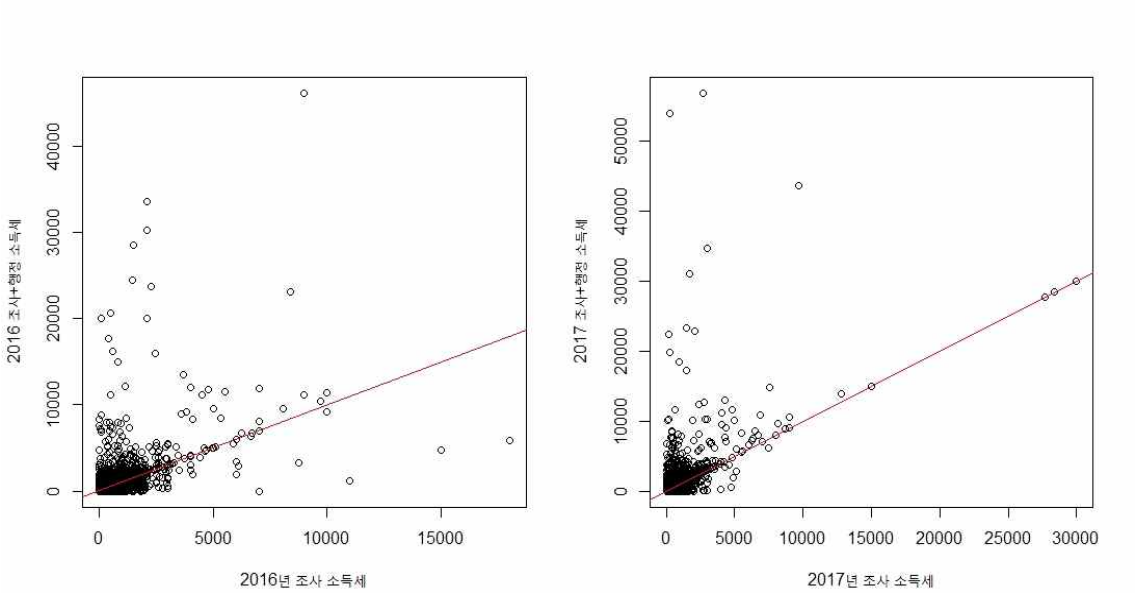
a) 2013-2014년에도 조사+행정값이 존재한다. 하지만 행정값 대체 수준이 2015-2017년에 비해서는 많이 낮은 편이다;

b) 재산소득의 경우와 달리 조사값이 조사+행정값에 대하여 가지는 설명력이 그다지 좋지 못하다. 즉, 단순 회귀대체 모형 적합이 어렵다.

따라서 이러한 특성을 고려한 종단면 대체모형을 개발하였다. <그림 3-4>에서 확인할 수 있듯이 소득세의 조사값과 조사+행정값은 대략 2개의 그룹으로 나뉘어 질 수 있는데, 조사값과 조사+행정값이 거의 일치하는 그룹과 그렇지 않은 그룹이다. 그런데 조사값과 (조사+행정)값이 유사하지 않은 가구를 예측하고 그러한 가구들의 특성을 관측된 변수로 잡아내는 것은 상당히 어려운 작업이다. 또한, 매년 그 구조와 메커니즘이 달라질 수 있음을 고려하면 종단면 대체모형이 복잡해질수록 예측 오차 또한 커질 수 있다. 따라서 본 용역에서는 종단면 대체되어 예측된 (조사+

행정)값과 조사값과의 비를 활용하는 비추정 횡단면 대체모형을 개발하였다.

<그림 3-4 2016/2017 소득세 조사 vs 조사+행정 산점도>



$$Y_{it}^{*} = R_{Y^{*},t} X_{it} \tag{3-6}$$

여기에서 X_{it}는 소득세의 조사값을 의미하고, $R_{Y^{*},t}$는 횡단면 대체된 가구들의 예측 (조사+행정)값과 조사값의 비로 정의된다,

$$R_{Y^{*},t} = \frac{\overline{\widehat{Y}_t}}{\overline{X_t}}.$$

이 때, $\overline{X_t}$는 대상연도 전체 조사값의 평균으로 계산되고, $\overline{\widehat{Y}_t}$는 종단면 모형으로 대체된 가구들의 예측 (조사+행정) 값으로 계산된다. 식 (3-6)의 특성상 종단면 대체되는 가구들의 평균 소득세는 횡단면 대체 가구들의 평균 소득세와 유사하게 된다.

3.3 모형 평가

3.2 장에서는 재산소득과 소득세 추정에 사용된 대체모형을 소개하였다. 본 장에서

는 제안된 대체모형을 평가한다. 재산소득과 소득세로 구분하여 횡단면 대체모형과 종단면 대체모형을 사용하여 예측치를 생성하여 실제값과 비교를 하였다. 따라서 실제 (조사+행정)값이 존재하는 2015-2017년 데이터만를 사용하였다.

3.3.1 재산소득

1) 종단면 대체모형

2015-2017년 조사값 및 (조사+행정)값을 이용하여 3.2장에서 제안된 재산소득의 혼합 대체모형 (3-4)을 사용하였다. 대체모형 평가에 필요한 통계적 계산은 아래에 스텝별로 정리하였다.

(Step 1) 두 조사연도에 걸쳐서 공통적으로 관측된 가구 선택. 2016-2017년도에는 14,095 가구이고, 2015-2016년도에는 13,850 가구임.

(Step 2) 비추정에 사용할 R 계산
공통적으로 관측된 가구들의 재산소득 조사값의 단순 평균 비[5]로 계산. $\hat{R}_{2015}=0.87$ 이고 $\hat{R}_{2016}=0.98$ 임.

(Step 3) 대체모형 식 (3-3)을 이용하여 예측치 생성

$$\hat{Y}_{i2015}=\begin{cases}\hat{R}_{2015}Y_{i2016} & \text{if } X_{i2015}\le 2000\\ (X_{i2015}-X_{i2016})+Y_{i2016} & Otherwise.\end{cases}$$

$$\hat{Y}_{i2016}=\begin{cases}\hat{R}_{2016}Y_{i2017} & \text{if } X_{i2016}\le 2000\\ (X_{i2016}-X_{i2017})+Y_{i2017} & Otherwise.\end{cases}$$

대체 모형을 평가하기 위해서 아래에 정의된 Root Mean Square Error (RMSE)와 Relative Bias (R.B.)를 사용하였다

5) 각 조사연도의 전체 조사값 평균 (Horvitz-Thompson 추정량)의 비로 대체 가능하다.

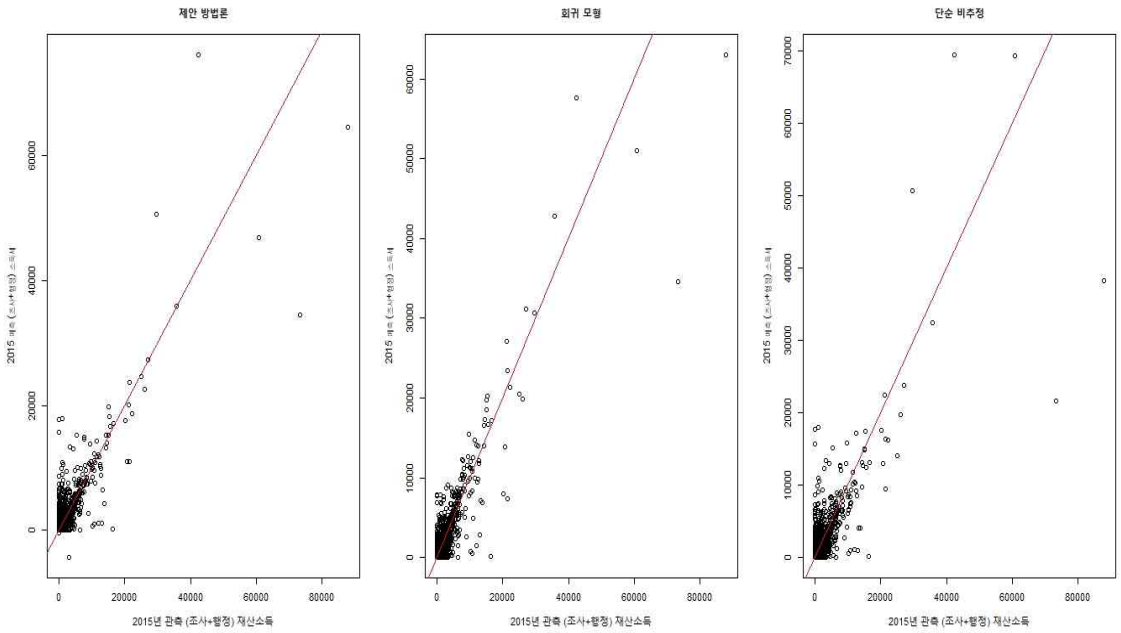
$$RMSE = \sqrt{n^{-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2},$$

$$RB = \frac{\bar{\hat{y}} - \bar{y}}{\bar{y}}.$$

<표 3-12 종단면 대체모형별 예측 결과-2015년 (조사+행정) 예측치>

모형	RMSE	RB (%)
회귀	684	5.8
단순비	969	-3.3
제안 혼합모형	841	2.2

<그림 3-5 종단면 대체 모형 관측 (조사+행정) vs 예측 (조사+행정) 재산소득>



<표 3-12>는 2015년 (조사+행정) 예측치 결과 비교이다. 회귀 대체모형은 설명변수를 재산소득 조사값 (X_{it}), 익년도 (조사+행정)값 ($Y_{i,t+1}$), 해당년도 저축액 (Z_{it})을 하는 일반 선형모형으로부터 얻어진 결과를 사용하였다. 이 때, 조사값이 0인 그룹과 그렇지 않은 그룹으로 나누어서 회귀 추정식을 사용하였다.

$$Y_{it} = \beta_{0g} + \beta_{2g} Y_{i,t+1} + \beta_{3g} Z_{it} + e_{it}, \text{ for } X_{it} = 0,$$

$$Y_{it} = \beta_{0g} + \beta_{1g} X_{it} + \beta_{2g} Y_{i,t+1} + \beta_{3g} Z_{it} + e_{it} \quad \text{for} \quad X_{it} \neq 0.$$

2016-2017년 관측 데이터를 이용하여 회귀 계수를 추정한 후, 이 값을 이용하여 2015년 (조사+행정)값을 예측하였다. 단순 비추정 모형은 식 (3-1)을 그대로 사용한 결과이다. <그림 3-5>는 2015년 재산소득의 (조사+행정) 예측치를 산점도로 표현한 결과인데, 산점도 상으로는 세 가지 대체모형간에 뚜렷한 차이가 존재하지 않는다. 평가지표 중에서 RMSE 측면에서는 회귀 대체 모형이 가장 좋으나 RB 측면에서는 제안 혼합모형이 좋다. 제안 혼합모형은 회귀 대체모형과 단순비 모형의 중간 형태이기 때문에 회귀 대체모형과 단순 비추정 모형의 장단점을 공유하고 있다. 하지만 시계열 연속성 측면과 평균 재산소득의 예측 상대오차가 작다는 측면을 고려할 때 다른 대체모형보다 선호될 수 있다.

2) 횡단면 대체모형

횡단면 대체는 식 (3-4)와 (3-5)를 이용하여 얻어질 수 있다. 2015년 (조사+행정)예측에는 2016년 데이터를 사용한 회귀 추정식을 사용하였고, 2016년 (조사+행정) 추정에는 2017년 데이터를 사용한 회귀 추정식을 사용하였다. 2016년도와 2017년도의 회귀 추정식은 <표 3-13>과 <표 3-14>에 정리되어 있다.

<표 3-13 회귀계수 추정 값 (2016년 데이터 사용)>

회귀 계수	모형 1	모형 2	모형 3
ϕ_1 (조사재산소득)	1.21 (0.0048)	1.16 (0.0052)	1.13(0.0052)
ϕ_2 (경상소득)		0.03 (0.0011)	0.01 (0.0013)
ϕ_3 (저축액)			0.01 (0.0005)
조정 결정계수	0.779	0.785	0.792

<표 3-14 회귀계수 추정 값 (2017년 데이터 사용)>

회귀 계수	모형 1	모형 2	모형 3
ϕ_1 (조사재산소득)	1.16 (0.0067)	1.11 (0.0071)	1.09 (0.0074)
ϕ_2 (경상소득)		0.02 (0.0013)	0.018 (0.0014)
ϕ_3 (저축액)			0.005 (0.0005)
조정 결정계수	0.622	0.630	0.632

3) 최종 횡단면+종단면 대체

<표 3-15 연도별 횡단면 대체모형 예측 결과>

조사연도	RMSE	RB (%)
2015	747	1.1
2016	891	-4.1

<표 3-15>는 횡단면 대체모형만을 이용하여 각 조사연도의 (조사+행정)값을 대체하였을 때의 RMSE 값과 RB 값이다. 횡단면 대체모형만으로도 종단면 대체와 유사한 결과를 얻을 수 있는 것을 확인할 수 있다. 하지만 이러한 결과는 예측 오차들이 합산되어 나타나는 결과로 횡단면 대체모형만을 이용하였을 때는 시계열 연속성은 담보되지 않는다. 따라서, 횡단면과 종단면 대체모형을 종합하여 최종적으로 (조사+예측)값을 생성하는 것이 필요한데 그 결과는 <표 3-16>에 요약되어 있다.

<표 3-16 연도별 최종 대체모형 (횡단면+종단면) 예측 결과>

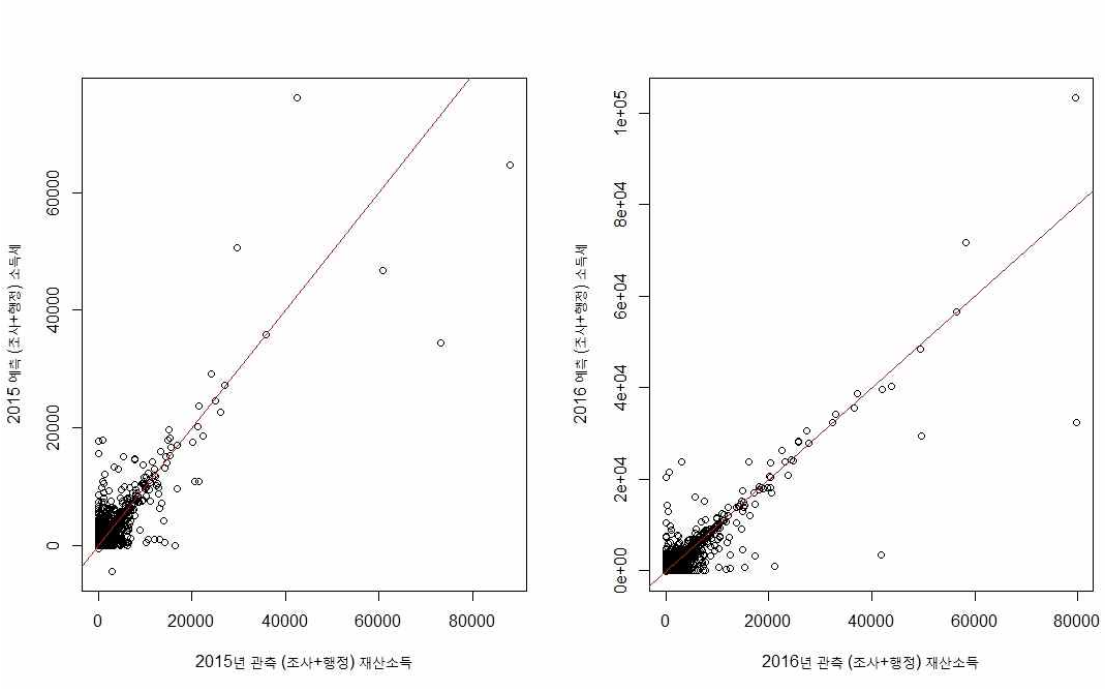
조사연도	RMSE	RB (%)
2015	793	2.7
2016	809	-3.3

<표 3-16>과 <그림 3-6>을 보면 2015년에 비하여 2016년의 결과가 상대적으로 좋지 않은데, 이는 2017년 데이터를 이용하여 얻은 회귀 추정식을 사용했기 때문이다. 2017년 데이터에서는 재산소득의 조사값과 (조사+행정)값의 선형 연관성이 약해졌기 때문에, 이 부분이 반영된 것으로 사료된다. 실제 2012-2014년 횡단면 대체 모형에서는 2015년 데이터를 활용하여 회귀계수를 추정하였다. 해당 기간의 패널과 가장 많이 겹치는 조사연도가 2015년도이기 때문에 2015년도의 관측값을 활용하였다. 2015년도를 이용하여 얻은 회귀 추정결과는 <표 3-17>에 정리되어 있다.

<표 3-17 회귀계수 추정 값 (2015년 데이터 사용)>

회귀 계수	모형 1	모형 2	모형 3
ϕ_1 (조사재산소득)	1.33 (0.0054)	1.29 (0.0057)	1.225(0.0060)
ϕ_2 (경상소득)		0.02 (0.0009)	0.06 (0.0012)
ϕ_3 (저축액)			0.01 (0.0005)
조정 결정계수	0.773	0.778	0.783

<그림 3-6 연도별 관측 (조사+행정) vs 예측 (조사+행정)>



3.3.2 소득세

재산소득과 동일한 방법으로 종단면 대체모형을 검정한다. 횡단면 대체모형의 경우 식 (3-6)으로 얻어지기 때문에 종단면 대체모형의 결과를 그대로 이어받는다. 따라서 소득세의 경우에는 횡단면 대체모형은 따로 평가하지 않고 그대로 최종 예측 결과를 실제값과 비교하였다. 종단면 대체 모형의 비교대상으로는 다중 회귀모형을 사용하였다. 설명변수로는 해당연도 조사값, 익년도 (조사+행정)값, 그리고 소득세와 밀접한 관련이 있는 경상소득을 사용하였다. 또한, 재산소득의 비교목적 회귀대체모형과 마찬가지로 조사값이 0인 경우와 그렇지 않은 경우로 나누어서 추정하였다.

<표 3-18 횡단면 및 종단면 대체모형 비 (R) 추정치>

	2015	2016
종단면 R	0.82	0.96
횡단면 R_{Y^*}	1.11	1.37

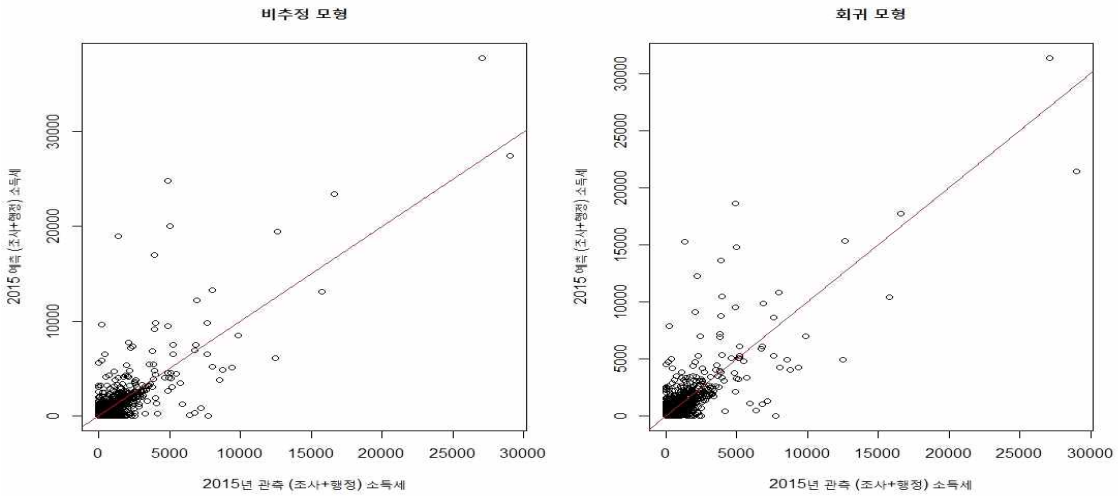
<표 3-18>은 종단면 비추정과 회단면 비추정에 사용하는 비값이다. 종단면 모형

에서의 R값은 대상연도와 익년도 세금의 관측 (조사+행정)값의 비로 구해진다. 한편 횡단면 R_{Y^*}는 대체된 가구들의 소득세 (조사+행정) 평균값과 전체 가구들의 조사값의 평균의 비로 구해진다. 2015년 추정치는 2015년과 2016년 값들로 얻어지는데 종단면 R값이 0.82인 것을 감안하면 2015년 소득세가 2016년 소득세보다 훨씬 적었다는 것을 의미한다. 횡단면 비 R_{Y^*}가 1.11이라는 것은 횡단면 대체된 가구들의 소득세 (조사+평균)이 전체 조사값의 평균보다 약 11% 정도 높았다는 것을 의미한다.

<표 3-19 대체모형별 횡단면 예측 결과>

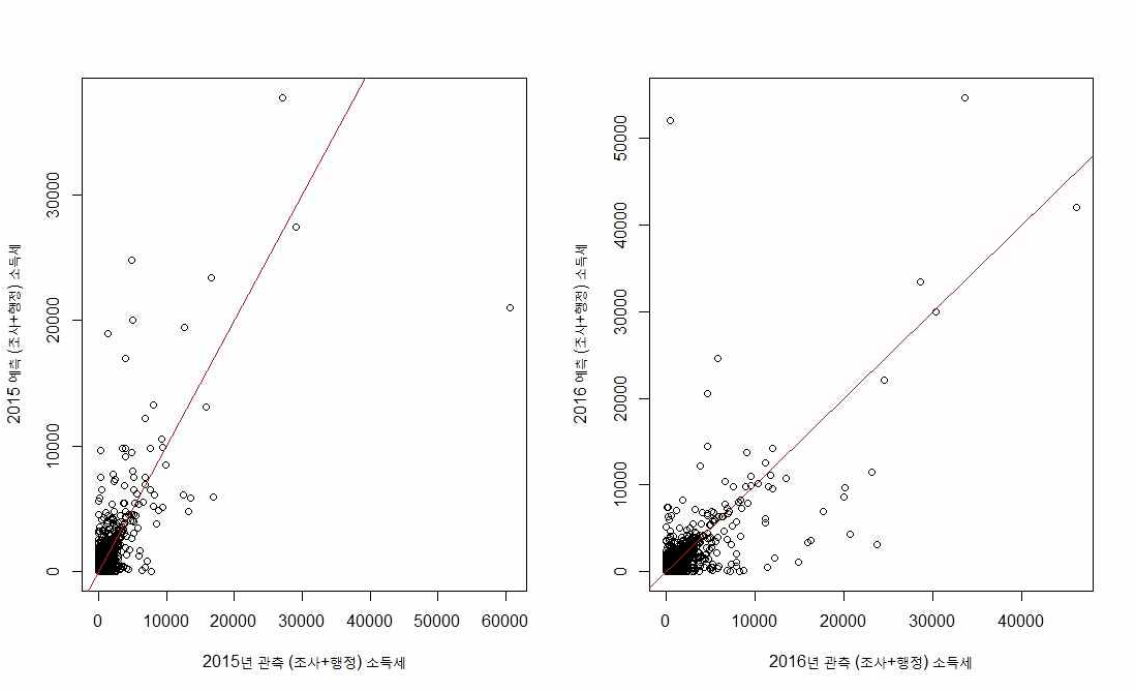
대체모형	RMSE	RB (%)
회귀모형	385	11.4
제안 비추정모형	445	5.5

<그림 3-7 종단면 대체 모형 관측 (조사+행정) vs 예측 (조사+행정) 재산소득>



<표 3-19>를 보면 회귀 대체모형과 비추정모형의 횡단면 예측 결과를 나타내주고 있다. 회귀 모형의 경우 RMSE는 비추정모형보다 작으나 RB에서 큰 값으로 과대추정하고 있다. 따라서 전체적인 평균 예측값의 정확도와 시계열 연속성 등을 고려하면, 비추정 모형이 더 합리적인 것을 확인할 수 있다. 예측 결과는 <그림 3-7>에 관측값과 함께 표시되었다.

<그림 3-7 소득세 연도별 관측 (조사+행정) vs 예측 (조사+행정)>



<그림 3-7>은 2015년과 2016년 소득세의 예측 (조사+행정)값과 관측 (조사+행정)값의 결과를 산점도로 표기한 것이다. 2016년의 관측 (조사+행정)의 우측하단부의 예측 정확도가 떨어지는 것을 확인할 수 있는데, <표 3-20>의 RMSE로 재확인할 수 있다. 그러나 평균적인 측면에서는 RB 0.9%로 편의가 적은 편이다. 2015년의 경우 종단면 대체모형을 이용한 경우보다 RB가 다소 줄어들었는데, 이는 횡단면 비추정 모형사용인한 예측값들의 평균이 관측 (조사+행정)값의 평균과 유사하기 때문이다.

<표 3-20 연도별 횡단면 대체모형 예측 결과>

조사연도	RMSE	RB (%)
2015	529	3.7
2016	667	0.9

3.4 순차적 대체

본 연구용역에서는 모형 과적합으로 인한 예측오차를 줄이기 위하여 될 수 있으면 직관적이고 단순한 형태의 대체 모형을 개발하였다. 또한 데이터 구조를 감안하

여 횡단면 대체모형과 종단면 대체모형을 분리하여 개발하였고, 재산소득과 소득세의 특성으로 인한 차이도 각 모형 개발시 반영하였다. 이렇게 개발된 모형은 3.2장에서 소개되었고 3.3장에서 관측 가능한 값들을 활용하여 검증하였다. 본 장에서는 이렇게 얻어진 대체모형을 이용하여 2012년-2014년의 재산소득과 소득세를 순차적으로 대체한다. 본 보고서에서는 대략적인 내용만 소개하고 자세한 내용은 별첨 R code를 참조하면 된다.

재산소득의 종단면 대체에서는 <표 3-17>에 표기된 모형1과 모형3의 회귀계수 추정값을 연도에 상관없이 사용한다. 조사값이 2,000만원 이상인 경우에서는 재산소득의 조사값만 사용하여 예측을 하고 그렇지 않은 경우에는 보조 변수를 사용한다. 또한, `비' 대체모형에서 사용되는 비 값은 <표 3-21>에 정리되었다. 재산소득은 종단면 대체모형에서만 비대체를 사용하고 소득세는 종단면, 횡단면 모두에서 비대체를 사용한다.

<표 3-21. 비 대체모형 ratio 추정값>

Year	재산소득 (종단면)	소득세 (종단면)	소득세 (횡단면)
2012/2013	0.89	0.95	1.24
2013/2014	1.05	1.04	1.26
2014/2015	0.98	0.97	1.13
2015/2016	0.87	0.82	1.11

1) 재산소득

- 2014년 예측: 종단면 대체는 식 (3-3)에 R_{2014} 추정값을 사용하여 예측하였고, 임계점 $q_v = 2000$을 사용하였다. 횡단면 대체에는 식 (3-4) 와 (3-5)를 임계점 q_v에 따라 사용하였다.

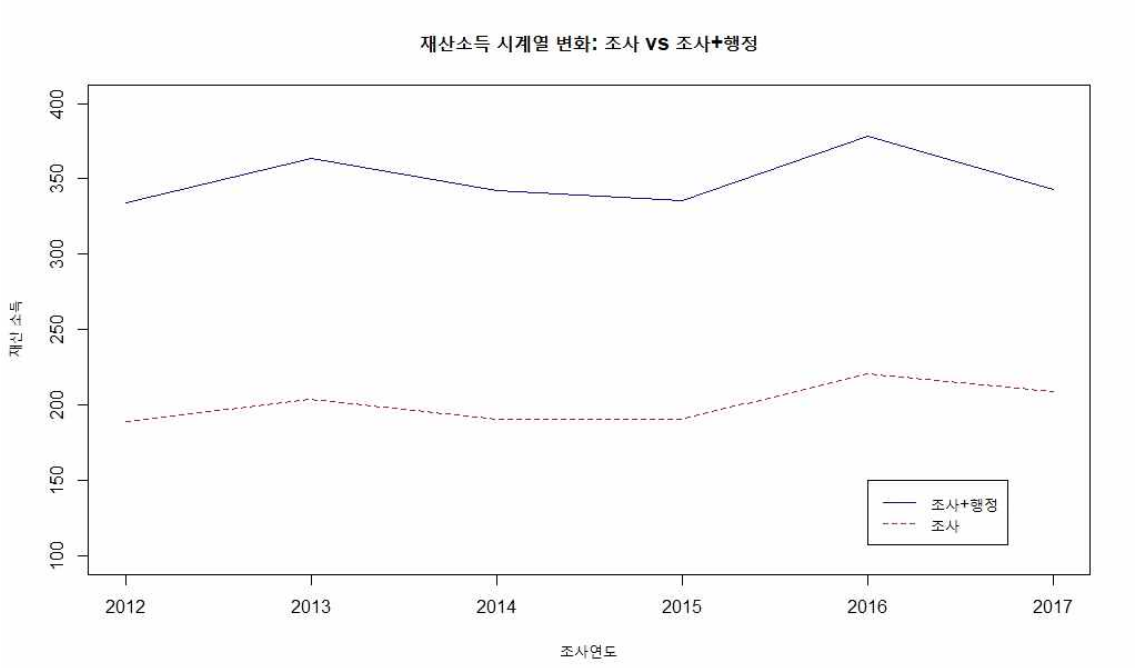
- 2012-2013년 예측: 종단면 대체에서 식 (3-1)을 사용하였다. 관측 (조사+행정) 값이 더 이상 관측되지 않기 때문에, 종단면 대체에서의 회귀 형태 모형은 적용하지 않는다. 횡단면 대체는 2014년 예측과 마찬가지 형태로 진행하였다.
- 재산소득 예측에는 회귀 대체식을 사용하기 때문에 음수값이 예측된 경우에는 0으로 대체하였다.

<표 3-22 재산소득 조사값 vs 조사+행정값>

	2012	2013	2014	2015	2016	2017
조사	189	204	190	191	221	209
조사+행정	335	364	342	336	378	343
조사/조+행	1.77	1.78	1.80	1.76	1.72	1.64

<표 3-22>를 보면 재산소득의 조사값과 (조사+행정)값의 평균값이 표시되어 있다. 평균값 계산시에는 샘플링 가중치를 사용하였다. 2012-2014년의 (조사+행정)값은 대체모형으로 생성된 예측값이고 2015년-2017년의 값은 실제로 관측된 (조사+행정)의 값이다. 예측 과정이 2015년에 관측된 데이터를 주요하게 사용하는 형태로 진행되었기 때문에, 조사/(조사+행정)의 비가 2015년 값과 유사하게 나오는 것을 확인할 수 있다.

<그림 3-8 재산소득 시계열 변화: 조사 vs 조사+행정>



<그림 3-8>은 <표 3-22>에서 얻은 결과를 시계열 그래프로 옮긴 것이다. 대체적으로 조사값의 변화를 잘 따라가고 있는 것을 확인할 수 있다.

2) 소득세

2012-2014년 모두에 걸쳐서 단순 비추정 식 (3-1)을 활용하여 (조사+행정)값을 예

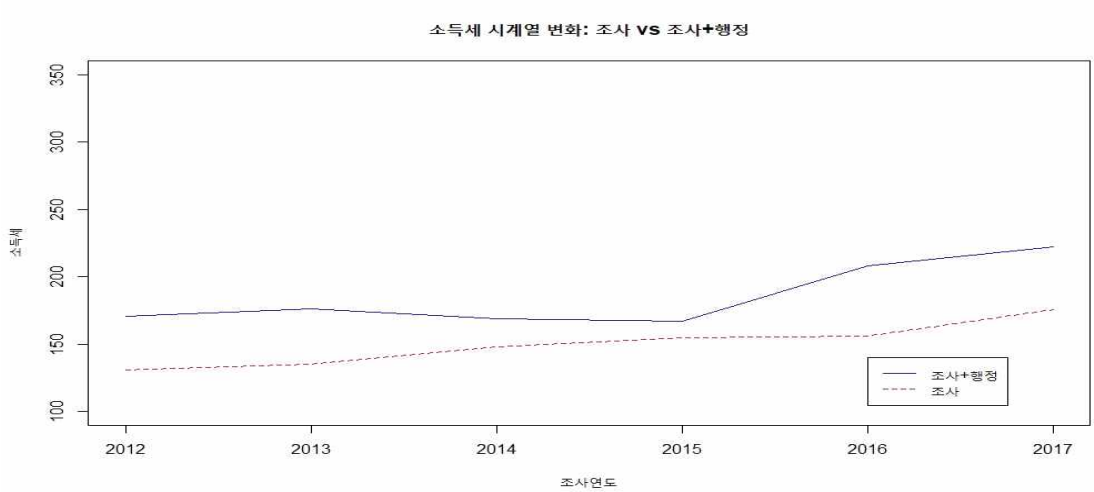
측하였으며, 횡단면 및 종단면 대체에 사용되는 비 추정치는 <표 3-21>에 정리되어 있다.

<표 3-23 소득세 조사값 vs 조사+행정값>

	2012	2013	2014	2015	2016	2017
조사	131	135	148	155	156	175
조사+행정	171	176	169	167	208	222
조사/조+행	1.30	1.30	1.14	1.08	1.33	1.27

<표 3-23>를 보면 소득세의 조사값과 (조사+행정)값의 평균값이 표시되어 있다. 평균값 계산시에는 샘플링 가중치를 사용하였다. 2012-2014년의 (조사+행정)값은 대체모형으로 생성된 예측값이고 2015년-2017년의 값은 실제로 관측된 (조사+행정)의 값이다. 소득세의 경우 세금의 (조사+행정)값의 변화를 벤치마킹하도록 설계되었다. 연도별로 행정값이 대체된 수준이 상이한데, 세금 내의 각 항목의 비중이 안정적으로 유지된다는 점을 고려하면, 세금의 변화비가 반영되는 것이 자연스럽다. 이러한 결과를 조사/(조사+행정) 변화를 통하여 확인할 수 있다.

<그림 3-9 소득세 시계열 변화: 조사 vs 조사+행정>



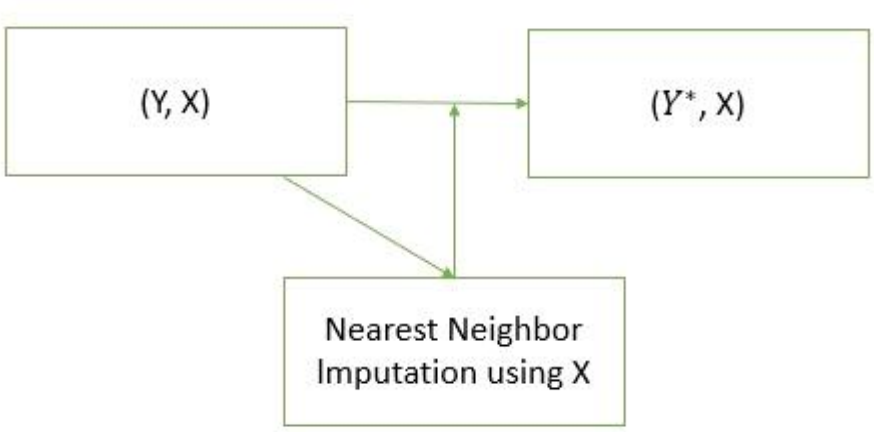
<그림 3-9>는 <표 3-23>을 시계열 그래프로 옮긴 결과이다. 세금의 (조사+행정)의 변화를 따라가는 점과 조사값와 (조사+행정)이 일치하지 않는 점을 고려할 때, 연도별로 조사값과 (조사+행정)값의 상대적 차이가 상이하다.

4. 가계금융복지조사의 시계열 연장

가계금융복지조사는 2011년[6] 소득 및 부채부터 관측하였기 때문에, 2006-2010년에 해당하는 소득은 존재하지 않는다. 그런데 가계동향조사에서는 2006-2010년에 해당하는 소득 자료가 존재한다. 따라서 가계동향조사의 2006-2016년도 원자료를 얻을 수 있는 경우에는 이를 이용하여 일종의 대체 (imputation) 방법을 이용한 자료 통합(data integration)이 가능해진다. 2006-2010년에 가계금융복지조사가 실시되었다면 수집되었을 자료값을 결측치 (missing data)로 간주하여 종단면 결측치 대체 (longitudinal mass imputation)를 실시할 수 있다. Nearest neighbor ratio imputation(NNRI)를 사용하여 가계금융복지조사의 각 표본 가구와 가장 유사한 속성을 갖는 가계동향조사 표본 가구를 찾아서 2006-2001 시계열 값을 대체하되 시계열 연속성을 위하여 `비' 보정을 실시하는 형태로 진행할 수 있다.

이를 설명하기 위해서 공통 관측 기간인 2011-2016년의 가구 소득을 X 라고 하고 2006-2010년의 가구 소득을 Y 라고 하자. 자료 X 는 두 조사 모두에서 관측이 되지만 Y 는 가계 동향조사 자료에서만 관측이 된다. 즉 가계동향조사에서는 (X, Y)가 모두 관측되고 가계금융복지조사에서는 X만 관측되는 데이터 구조를 갖는다. 이러한 구조하에서 가계금융복지조사의 X와 유사한 값을 가지는 가계동향조사 가구를 찾아서 그 가구의 시장소득과 가처분소득의 변화비 (ratio)를 빌려와서 2011-2016년도 소득에 적용하여 2006-2010년 소득값을 예측할 수 있다. 대략적인 데이터 구조와 작업 흐름이 <표 4-1>과 <그림 4-1>에 요약되어있다.

<그림 4-1 NNI 작업 흐름도>



6) 본 장은 3장과 달리 가계동향조사와의 일치성을 위하여 소득연도 기준으로 표기한다.

<표 4-1 가계금융복지조사 및 가계동향조사 자료 구조>

	가계동향	가계금융복지조사
2006-2010 (Y)	O	?
2011-2016 (X)	O	O

4.1 데이터 요약

본 장에서 주요하게 진행해야 될 과제는 가계동향조사의 시장소득과 가처분소득 값을 이용하여 2006-2010년도의 시장소득과 가처분 소득값을 비교하는 작업이다. 따라서 우선적으로 두 조사의 시장소득과 가처분소득을 비교하고 그 외 매칭에 사용할 가구 변수의 특성을 살펴보기로한다.

<표 4-2 2011-2016 시장소득 및 가처분소득, 단위-만원>

	시장소득		가처분소득	
소득연도	가계동향	가계금융복지	가계동향	가계금융복지
2011	3,593	3,902	3,479	3,657
2012	3,810	4,108	3,682	3,843
2013	3,862	4,271	3,746	4,002
2014	3,913	4,354	3,807	4.099
2015	3,902	4,432	3,837	4,198
2016	3,907	4,540	3,844	4,288

<표 4-2>는 공통관측연도의 시장소득 및 가처분 소득을 나타내고 있다. 가계금융복지조사에서 가처분 소득은 다음 연산식을 사용하였으며 모든 값은 공평한 비교를 위하여 조사값을 사용하였다:

- 시장소득=근로소득+사업소득+재산소득+사적이전소득-가구간 이전지출
  -비영리단체 이전지출,

- 가처분소득=시장소득+공적이전소득-소득세-재산세-자동차세-기타세금
  -공적연금 및 사회보험료.

가계동향조사와 가계금융복지조사는 동일한 모집단을 공유하고 있다. 하지만 실제 관측된 가구 분포는 다소 상이하다. 따라서 두 조사에서 공통적으로 관측되는 가구관련 변수- 가구주 연령, 가구주 성별, 가구주 교육정도, 가구원 수-들의 분포

를 비교할 필요성이 있다. 공통적으로 소득이 관측된 2011년을 기준 각 가구 변수들의 분포를 아래에 순차적으로 비교하였으며 각 분포는 각 조사의 가구 가중치를 사용하여 계산하였다.

<표 4-3 가구주 연령 분포 비교 (2011년 데이터 사용), 단위 %>

	20대 이하	30대	40대	50대	60대 이상
가계동향	2.6	16.7	23.7	21.3	35.7
가계금융복지	3.7	18.5	26.2	23.3	28.3

<표 4-3>은 가구주 연령의 분포를 보여준다. 60대 이상은 가계동향 조사에서 상대적으로 더 많이 표집되었으며 그 외 연령대는 가계금융복지조사에 더 많이 표집된 것을 확인할 수 있다. 이러한 연령 분포의 특성은 가구 소득과도 밀접한 연관이 있을 것으로 사료 된다.

<표 4-4 가구주 성별 분포 비교 (2011년 데이터 사용), 단위 %>

	남자	여자
가계동향	77.1	22.9
가계금융복지	77.7	22.3

<표 4-4>는 가구주 성별의 분포를 비교한다. 두 조사간 가구주 성별 분포의 연령 차이는 없는 것으로 판단된다.

<표 4-5 가구주 교육 정도 분포 비교 (2011년 데이터 사용), 단위 %>

	중등이하	고등학교	대학교이상
가계동향	35.6	34.3	30.1
가계금융복지	29.2	33.6	37.2

<표 4-5>는 가구주의 교육 정도의 분포를 나타낸다. 60대 이상의 가구주가 가계동향조사에서 더 많이 표집된 것을 감안하면, 교육 정도 분포의 차이가 자연스럽게 설명이 된다. 상대적으로 가계금융복지조사에서 관측된 가구의 가구주 교육정도 수준이 더 높은 것을 확인할 수 있다.

<표 4-6 가구원 수 분포 비교 (2011년 데이터 사용), 단위 %>

	1인	2인	3인	4인	5인 이상
가계동향	12.6	34.5	22.8	22.6	7.6
가계금융복지	18.0	25.3	20.1	27.2	9.3

<표 4-6>는 가계동향조사와 가계금융복지조사의 가구원 수 분포를 나낸다. 1인, 4인, 5인이상의 가구는 가계금융복지조사에서 더 많이 표집되었고, 2인 가구는 가계동향 조사에서 상대적으로 더 많이 표집된 것을 확인할 수 있다. 이는 가구주 연령과 밀접한 관계가 있을 것으로 사료 된다.

4.2 제안 방법론

4.2.1. NNRI

본 연구용역에서는 기본적으로 가계금융복지조사의 패널가구와 유사한 가구를 가계동향조사에서 매칭하여 소득값의 변화비를 대체하는 최근방 이웃 비대체 (NNRI, Nearest Neighbor Ratio Imputation)을 사용한다. 논의의 편의를 위하여 가계금융복지조사의 가구를 수령 가구 (recipient)로 정의하고 가계동향조사의 가구를 기부 가구 (donor)로 정의한다. 즉, 값을 빌려온다는 측면에서 수령 가구로 정의하고, 관측값을 빌려준다는 측면에서 기부 가구로 정의한다. NNRI는 다음의 일련의 과정을 통하여 진행할 수 있다:

(Step 1) 가계금융복지조사에서 수령 가구 i를 고정

(Step 2) 특정시점 t에서 수령 가구 i의 가구 변수 X_i 및 소득변수 Y_{it}와 유사한 값 (Nearest Neighbors)을 가지는 k개의 기부 가구들을 선택하여 기부 가구 집합 D_{it} 를 구성[7].

기부가구 집합 D_{it}는 구성 방법에 따라서 두 가지 경우로 나누어서 생각해볼 수 있다.

7) 만약 가계동향조사의 시장소득과 가처분소득의 변화비가 가계금융복지조사의 시장소득과 가처분소득의 변화비와 동일하다고 가정하면 모든 가계동향조사 가구를 기부 가구로 선택할 수 있다. 만약, nearest neighbor (donor)를 찾지 않고 모든 가능한 값들을 사용한다면 식 (3-1)의 일반적인 비대체모형의 형태로 표현이 되고, 이 때 구해진 비는 가계동향조사의 소득변화율을 벤치마킹하게 된다.

1) $D_i = D_{it}$ for all t

- 가계금융복지조사 각 각의 가구에 고정 기부 가구 집합을 찾는다. 이 때 시점에 걸친 관측값들을 모두 벡터화하여 비교하며 거리 함수로는 Minkowski distance를 이용한다.

$$d(Z_i, Z_j) = \left(\sum_{l=1}^{L} |Z_{il} - Z_{jl}|^p \right)^{1/p} \quad i \in \text{가계금융복지조사},\ j \in \text{가계동향조사}$$

이 때 Z_i는 가계금융복지조사의 수령 가구 i의 가구 변수와 공통 시점에 걸친 모든 소득 변수들의 벡터화된 표현이고 Z_j는 가계동향조사에 속하는 가구 j의 가구 변수와 소득변수 값 벡터이다. 일반적으로는 $p = 2$ 값을 사용하여 거리를 측정하고, 범주형 자료의 경우 더미화하여 그 거리를 계산한다.

- 그런데 가계동향가구는 패널이 아니기때문에 고정 기부 가구들을 찾는 것은 상당히 어렵다. 4.2.2장에서 고정 기부 가구들을 구성하는 하나의 방법을 소개한다.

2) $D_i \neq D_{it}$ for $t \neq S$

- 가계동향조사의 가구응답패턴 및 소득연도에 따라서 가계동향조사가구 집합을 새롭게 정의하고 그 집합으로 기부 가구를 선택함.

- (Ex) 2010년 대체 기부 가구 집합
  - 2010년 및 2011년 관측값이 있는 가계동향가구를 찾음, B_{2010}.
  - 가구집합 B_{2010}에서 기부 가구를 선택하여 $D_{i,2010}$을 각 가구별로 찾음.

- 데이터 특성을 고려하여 시점별로 구성한 기부 가구 집합을 활용할 때에는 시계열 연속성을 확보하기 어려운 단점이 있다.

(Step 3) 기부 가구집합에 속한 가구들의 소득값을 활용하여 소득 변화율을 다음과 같이 추정한다:

$$\hat{\pi}_{it} = \frac{\sum_{j \in D_{it}} X_{it}^{(j)}}{\sum_{j \in D_{it}} X_{i,t+1}^{(j)}}$$

여기에서 $X_{it}^{(j)}$는 수령 가구 i에 매칭된 가계동향조사 기부 가구 j의 t시점에서의 시장소득 혹은 가처분 소득을 의미한다.

(Step 4) 소득 변화율이 추정 되었다면, $t+1$ 시점에서의 t시점 예측치는 다음과 같이 계산된다:

$$Y_{is}^{*} = \hat{R}_{is} Y_{i,t+1}, \ s \leq t \qquad (4\text{-}1)$$

여기에서 $\hat{R}_{i,s}$는 s시점부터 t시점까지의 시계열 변화 누적곱으로 t시점 대비 얼만큼 변화했는지를 나타내며 다음과 같이 정의된다,

$$\hat{R}_{i,s} = \prod_{s \leq t}^{t} \hat{\pi}_{is}.$$

만약 기부가구가 시점에 상관없이 동일하다면 (혹은 모든 시점을 고려하여 공통의 기부 가구를 선정하였다면), $\hat{R}_{i,s}$는 다음과 같이 계산 가능하다:

$$\hat{R}_{is} = \frac{\sum_{j \in D_i} X_{is}^{(j)}}{\sum_{j \in D_i} X_{i,t+1}^{(j)}}.$$

4.2.2 기부 가구 집합

4.2.1절에서 기부가구를 구성하는 방법에 따라서 대체 방법이 달라질 수 있음을 논의하였다. 가계동향조사는 패널 조사가 아니기 때문에 각 조사연도별로 모두 관측된 가구가 존재하지 않는다. 따라서 데이터 특성을 그대로 반영하여 기부 가구를 구성하려면 연도별로 기부가구 집합 D_{it}를 생성해야 한다. 하지만 이러한 방식으로 추정을 하게 되면 매년 기부 가구 집합에 속해 있는 가구가 달라질 수 있기 때문에 시계열의 연속성을 확보하기 어려운 단점이 있다. 따라서, 본 연구용역에서는 모든 시점에 걸쳐서 기부 가구가 동일한 경우를 산정하여 기부 가구를 구성하는 방법을

소개한다.

앞서 지적했듯이 가계동향 가구는 매년 관측된 것이 아니므로 우선적으로 가계동향조사 가구들의 응답패턴을 확인해야 한다. <표 4-7>에 상위 20개 응답 패턴이 정리되어 있다. 예를 들어, 응답 패턴이 `00000111100'인 것은 2011-2014년도에는 조사에 참여하였고 그 외에는 조사에 포함되지 않은 가구를 나타낸다. 마찬가지로 `11110000000'는 2006-2009년도까지는 조사에 참여하였고 그 뒤에는 포함되지 않았음을 의미한다.

<표 4-7 가계동향조사 응답가구 패턴 (2006-2016, 62,486 가구)>

응답 패턴	가구 수	응답 패턴	가구 수
00000000001	5,129	00000111100	1,570
00000000010	1,304	00001111000	1,486
00000000011	2,739	00011110000	1,513
00000000100	1,264	00111100000	1,549
00000000110	2,409	01111000000	1,478
00000000111	2,145	10000000000	5,143
00000001000	2,894	11000000000	5,012
00000001111	1,418	11100000000	2,339
00000011110	1,531	11110000000	1,552
00000100000	3,193	11111010000	1,143

<표 4-8 가계동향조사 가구 응답 횟수 분포, 62486 가구>

1회	2회	3회	4회	5회	6회
21,819	16,062	9,872	13,204	386	1,143

<표 4-8>은 각 가구들의 응답 횟수를 분포로 나타낸다. 5회 이상에 걸쳐서 응답한 가구는 총 1,529 (=386+1143)으로 기부 가구 후보 집합으로 정하기에는 부족하다. 본 연구 용역에서는 4회이상 참여한 가구만을 기부 가구 후보군으로 설정하였다. 4회이상 응답한 가구들을 모아서 패턴 분석을 다시 해보면 <표 4-9>와 같이 정리할 수 있다. 똑같이 5회 응답했더라도 연속해서 5회를 응답했을 수도 있고 응답, 무응답을 반복해가며 5회를 응답했을 수도 있다.

<표 4-9 기부 가구 후보집단 응답패턴 예시>

패턴	2006	2007	2008	2009	2010	2011	2012
1	O	O	O	O			
2			O	O	O	O	O
3			O	O	O		O
4	O	O	O	O	O		O
5			O	O	O	O	
…	…	…	…	…	…	…	…

논의의 편의를 위하여 2006-2012년의 가계동향조사 결과만을 사용한다고 하자. 만약 <표 4-9>에서 응답이 이루어지지 않은 곳에 적정한 값을 채워 넣는다면 모든 가구들이 2006-2012년에 걸쳐서 소득값을 가지게 되고 이를 이용하여 NNRI을 진행할 수 있다. 즉 1) 공통관측년도 2011년과 2012년의 데이터를 활용하여 최근방 이웃을 찾고, 2) 최근방 이웃의 2006-2011년 소득 변화비를 계산하여, 3) 이 소득 변화비를 식 (4-1)에 적용하여 가계금융복지조사의 2006-2010년 시장소득 및 가처분 소득을 계산할 수 있다. 따라서 공통 기부가구를 구성하여 NNRI를 적용하는 통계적 방법은 다음의 일련의 과정을 통하여 실행될 수 있다:

(Step 1) 공통관측연도를 포함하여 기부 가구 구성에 사용할 대상 조사연도 기간을 설정한다. <표 4-9>의 예에서는 2006-2012년을 설정하였다.

(Step 2) 해당 조사연도 기간에 4회 이상 관측된 가계동향 조사 가구를 선정한다 (기부 가구 집단).

(Step 3) 관측값이 존재하지 않는 경우, 결측치 대체 (missing data imputation) 방법을 사용하여 적절한 값을 채워넣어 M개의 완비 데이터 집합을 생성한다. 본 연구 용역에서는 Multiple imputation by Chained Equation (MICE) 방법을 활용하였다. MICE 방법의 아이디어는 다음에 순차적으로 간략하게 요약하였다. 논의의 편의를 위하여 3개 조사연도의 소득 변수와 (Y_1, Y_2, Y_3) 가구 변수 X만을 이용하여 예시화하였다. 이 때, 소득 변수는 결측치가 발생할 수 있으며 가구 변수에는 결측치가 없다고 가정하였다.

(a) 각 가구 변수별로 대체 모형을 설정한다

$$P(Y_k \mid Y_{-k}, X, \theta_k)$$

여기에서 Y_{-k}는 변수 Y_k를 제외한 나머지 변수들만을 의미한다.

(b) $t-1$ iteration의 대체값 $Y^{*(t-1)}$이 주어졌을 때, t시점 iteration은 Gibbs sampling의 아이디어를 활용하여 순차적으로 생성한다:

$$\hat{\theta}_1^{(t)} \sim P(\theta_1 \mid Y^{*(t-1)}, X)$$

$$Y_1^{*(t)} \sim P(Y_1 \mid Y_{-1}^{*(t-1)}, X, \hat{\theta}_1^{(t)})$$

$$\hat{\theta}_2^{(t)} \sim P(\theta_2 \mid Y_1^{*(t)}, Y_3^{*(t-1)}, X)$$

$$Y_2^{*(t)} \sim P(Y_2 \mid Y_1^{*(t)}, Y_3^{*(t-1)}\ X, \hat{\theta}_2^{(t)})$$

$$\hat{\theta}_3^{(t)} \sim P(\theta_3 \mid Y_{-3}^{*(t)}, X)$$

$$Y_3^{*(t)} \sim P(Y_3 \mid Y_1^{*(t)}, Y_2^{*(t)}\ X, \hat{\theta}_3^{(t)})$$

(c) 위 Gibbs Sampling 과정을 수렴할 때까지 충분히 반복한 뒤 최종 결측치 대체값 (Y_1^*, Y_2^*, Y_3^*)을 생성한다.

(d) (b)-(c) 과정을 $M(=5)$번 반복한다.

(Step 4) 각 기부 가구 집합에서 가계금융복지조사의 각 가구에 매칭되는 최근방 이웃을 하나씩 찾는다. 따라서, 가계금융복지조사의 가구는 총 M개의 기부 가구를 갖게 되며 이 가구들이 기부가구 집합 D_i가 된다.

(Step 5) 각 가구에서 찾아진 D_i와 NNRI (식 4-1)을 이용하여, 2006-2010년에 해당하는 시장소득과 가처분소득을 만들어 낸다.

4.3 제안 방법론 평가

NNRI가 실제로 어떻게 작동될 수 있는지 알아보기 위하여 가계동향조사와 가계금융복지조사를 이용하여 임의의 데이터 집합을 만들었다. 2013-2014년을 공통관측 연도로 가정하고, 2011-2012년의 가계금융복지조사 시장소득과 가처분소득을 생성하는 형태로 NNRI를 진행하였고 실제 관측값과 비교하였다.

1) 기부 가구 구성

가계 동향조사로부터는 2011-2014년 사이에 세 번이상 응답한 가구들을 대상으로 기부 가구를 구성한다. 총 11,961가구가 선택되었으며, 결측치가 있는 가구들의 소득값은 MICE 방법을 이용하여 결측치를 대체하였다. 가구 변수의 경우 2014년을 기준으로 작성하였다. 가구주 교육정도는 중등이하/고등/대학이상으로 나누었고 가구원수는 1/2/3/4인 이상으로 구분했고 소득은 만원 단위로 전환하였다. 결측치 대체의 결과로 총 $M=5$개의 기부 가구 데이터 집합을 생성하였다.

2) 수령 가구 구성

2013년과 2014년에 걸쳐서 모두 관측된 가구들만을 대상으로 하였다. 소득이 0인 가구 일부를 제외하고 총 12,885가구가 표본 가구로 선택되었으며, 시장소득과 가처분 소득은 4.1장에서 소개된 전환식을 이용하여 계산하였다. 가구 변수는 2014년의 관측값을 이용하였고 기부 가구의 가구 변수와 동일한 형태로 변환하였다.

3) 최근방 이웃 선택

4.2장에서 소개된 거리함수 (p=2)를 이용하여 각 각의 기부 가구 집합에서 최근방 이웃을 선택하였다. 총 가구 집합이 $M=5$이므로, 가계금융복지조사의 한 가구당 5 이웃 가구들이 선택되었다.

4) NNRI

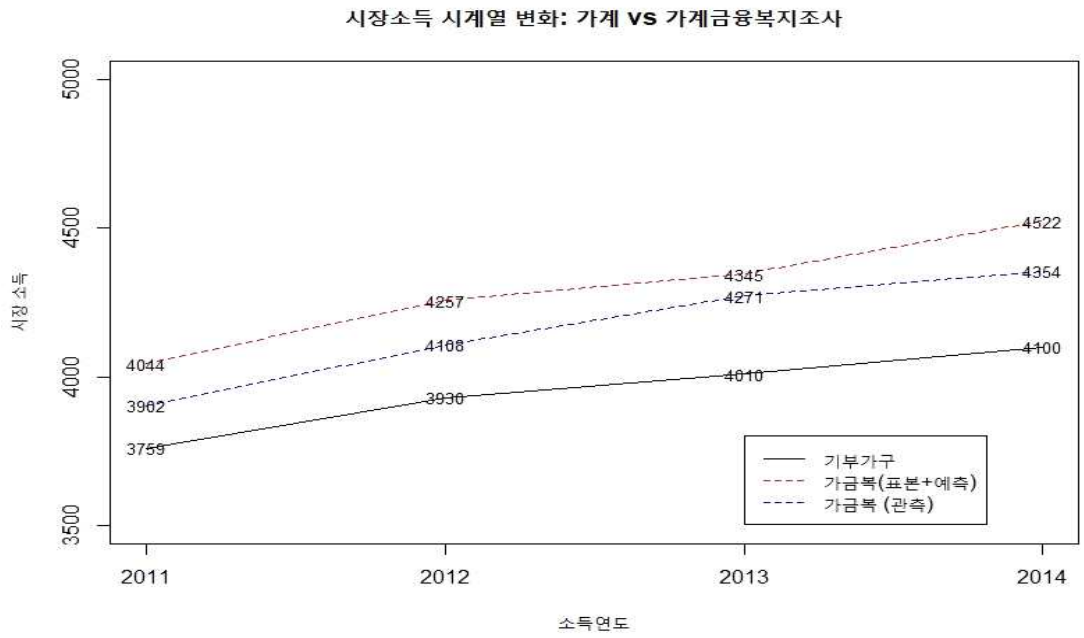
소득변화율을 2013년 기준으로 계산한 후 식 (4-1)을 이용하여 2011년과, 2012년에 해당하는 시장소득과 가처분소득을 계산하였다.

<표 4-10. 가계금융복지조사 및 가계동향 연결 결과-평균 시장소득 (만원)>

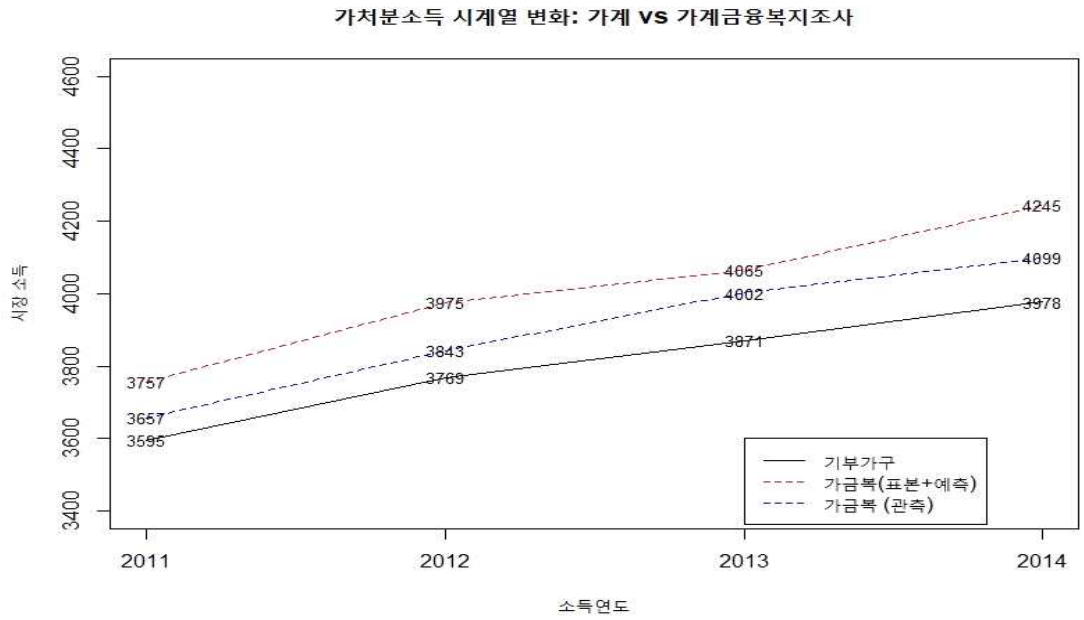
	시장소득		가처분 소득	
연도	기부 가구	가금복 (표본)	기부 가구	가금복 (표본)
2011	3759 (0.92)	4044 (0.89)	3595 (0.90)	3757 (0.89)
2012	3930 (0.96)	4257 (0.94)	3769 (0.95)	3975 (0.94)
2013	4010 (0.98)	4345 (0.96)	3871 (0.97)	4065 (0.96)
2014	4100 (1.00)	4522 (1.00)	3978 (1.00)	4245 (1.00)

<표 4-10>은 가계금융복지 표본 가구에 매칭된 가계동향조사의 기부 가구들의 시장소득과 가처분 소득 평균과 가계금융복지조사 가구의 시장 및 가처분소득 평균이 표시되어 있다. 가계금융복지조사의 2013년과 2014년의 값은 실제 관측값을 사용하여 얻은 결과이고, 2011년과 2012년은 NNRI 방법을 이용하여 경시적 대체(longitudinal imputation)를 통하여 얻어진 값들이다. 가처분소득의 경우 기부가구의 평균과 가계금융복지조사의 평균이 비슷함을 알 수 있고, 시장소득의 경우엔 가계금융복지조사의 표본 가구 평균이 2014년에서 2011년으로 갈수록 기부 가구들의 평균보다 더 빠르게 감소하고 있다. 하지만 이러한 감소속도는 가계금융복지조사의 실제 관측값의 감소속도와 유사하다. <그림 4-2>를 보면 관측 가계금융복지조사값의 소득변화비와 예측값으로 이루어진 표본 가구들의 소득변화비와 유사한 것을 재확인할 수 있다.

<그림 4-2 시장소득 연결 비교>



<그림 4-3 가처분소득 연결 비교>



<그림 4-3>은 가처분 소득의 시계열 그래프를 보여주고 있다. 시장소득과 마찬가지로 모형평가 대상으로 선택된 표본 가구들의 가처분 소득 평균이 전체 가구들을 대상으로 한 가처분 소득 평균보다는 높았다. 전반적으로 실제 관측값들의 소득 변화비와 유사한 것을 확인할 수 있으며, 표본 가구의 가처분소득에서 2011-2012년은 NNRI를 이용하여 대체된 값을 활용하여 계산되었고 2013년과 2014년은 실제 값을 사용하였다.

4.4 2006-2010년 소득 연결

4.3장에 소개된 방법을 이용하여 가계금융복지조사의 2006-2010년 시장소득 및 가처분 소득을 예측하여 연결하는 작업을 진행하였다. 공통관측연도는 2011년과 2012년으로 선택하였다. 가구 변수들의 기준값은 2012년도로 사용을 하였고, 수령가구 및 기부 가구 모두 소득값이 0이상인 가구만을 선택하였다.

1) 기부 가구 구성

가계 동향조사로부터는 2006-2012년 사이에 네 번이상 응답한 가구들을 대상으로 기부 가구를 구성하였다. 총 8,094 가구가 선택되었으며, 결측치 값은 MICE 방법을 활용하여 채워넣었다. 가구 변수는 모형 평가에서 사용한 분류 기준을 그대로 사용하였으며 총 $M=5$개의 기부 가구 데이터 집합을 생성하였다.

2) 수령 가구 구성

2011년과 2012년에 걸쳐서 모두 관측된 가구들만을 대상으로 하였다. 소득이 0인 가구 일부를 제외하고 총 17,170가구가 표본 가구로 선택되었으며, 시장소득과 가처분 소득은 4.1장에서 소개된 전환식을 이용하여 계산하였다. 가구 변수는 기부 가구와의 일치성을 고려하여 2012년의 관측값을 이용하였고 기부 가구의 가구 변수와 동일한 형태로 전환하였다.

3) 최근방 이웃 선택

모형평가에서 사용된 Mahalonobis 거리함수 (p=2)를 이용하여 각 각의 기부 가구 집합에서 최근방 이웃을 선택하였다. 총 가구 집합이 $M=5$이므로, 가계금융복지조사의 한 가구당 다섯 가구가 최근방 이웃으로 선택되었다.

4) NNRI

소득변화율을 2011년 기준으로 계산한 후 식 (4-1)을 이용하여 2006년-2010년에 해당하는 시장소득과 가처분 소득을 예측하여 대체하였다.

<표 4-11 가계금융복지조사 및 가계동향 연결 결과-평균소득 (만원)>

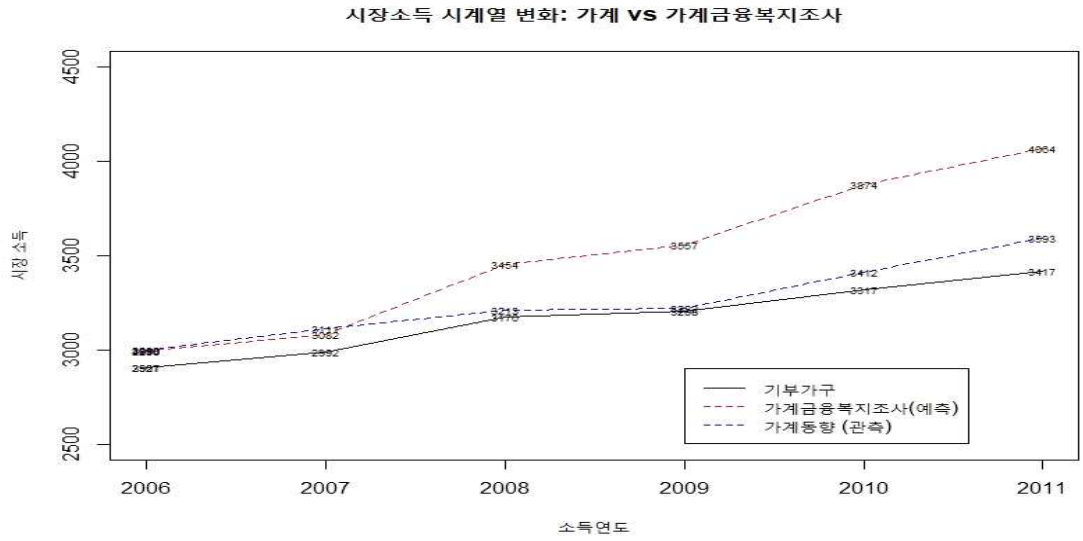
	시장소득		가처분 소득	
연도	가계동향	가계금융복지	가계동향	가계금융복지
2006	3000 (0.83)	2996 (0.74)	2889 (0.83)	2793 (0.74)
2007	3111 (0.87)	3082 (0.76)	3002 (0.86)	2920 (0.77)
2008	3213 (0.89)	3454 (0.85)	3107 (0.89)	3266 (0.86)
2009	3222 (0.90)	3557 (0.88)	3135 (0.90)	3404 (0.90)
2010	3412 (0.94)	3874 (0.95)	3309 (0.95)	3694 (0.98)
2011	3593 (1.00)	4064 (1.00)	3479 (1.00)	3788 (1.00)

( ) 값은 2011년 관측값 대비 상대적 크기임

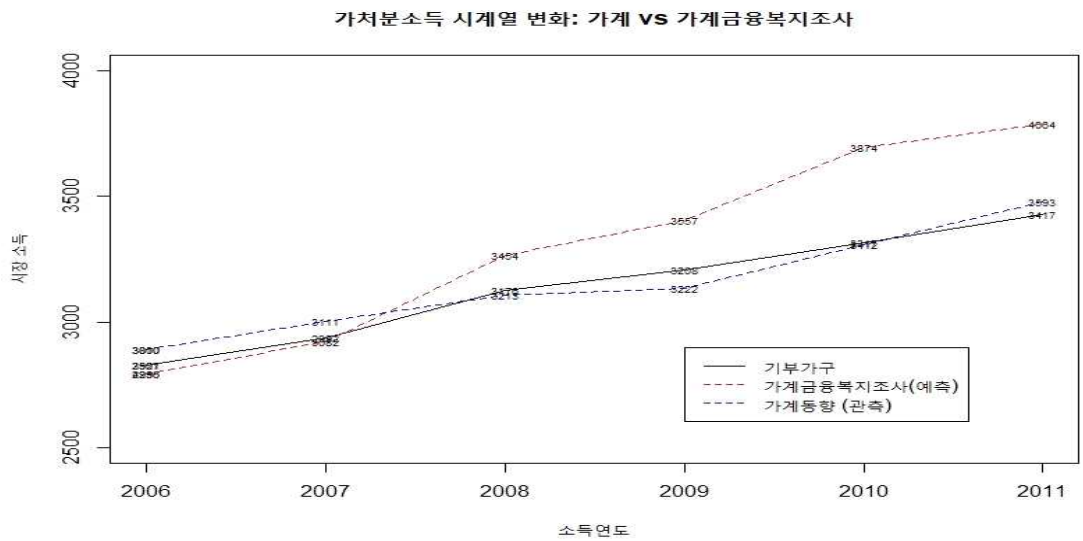
<표 4-11>은 가계동향 관측값 (기부가구가 아닌 전체 가구를 사용)과 가계금융복지조사의 예측 결과(2011년 데이터 제외)를 보여주고 있다. 2006년으로 갈수록 가계금융복지조사의 대체값 평균 감소율이 빠르게 감소하고 있다. 소득대체가 순차적으롯 진행되면서 발생된 편의가 누적되어 나타나는 것이 첫 번째 이유가 될 것이고, 또 다른 하나는 가계금융복지조사 가구들에 매칭된 가계동향조사의 가구들의 평균 소득 감소율이 전체 가계동향조사 가구들의 평균 감소율보다 큰 것이 두 번째 이유이다. 이러한 현상은 <그림 4-4>에서 확인할 수 있다. 이러한 성향은 가처분 소득에서도 그대로 나타난다. 따라서 가계동향조사를 이용하여 2006-2010년 소득값의 시계열 연장을 하기 위해서는 기부 가구 선택 및 편의 누적 문제를 필수적으

로 해결해야 할 것으로 판단되며, 이러한 내용은 장기적인 관점에서 점검해야될 것이다.

<그림 4.4 2006-2011 시장소득 비교>



<그림 4.5 2006-2011 가처분소득 비교>



5. 결론

본 연구용역 보고서는 크게 세 가지의 주요 과제내용으로 진행되었다. 1) 행정자료 매칭으로 인하여 발생할 수 있는 매칭 편향을 확인하고 보정하는 방법이 첫 번째이고, 2) 두 번째로는 금융소득의 시계열 연장방법을 고민하였고, 3) 마지막으로는 가계동향조사와의 연결방법을 연구하였다.

2장에서는 가계금융복지조사의 자료를 행정자료로 보완하는 과정에서 발생할 수 있는 매칭 편향 문제점을 다루었다. 조사 자료와 행정 자료 간 매칭 성공률이 100%가 아님으로 인해, 매칭 집단의 조사+행정 자료와 비매칭된 집단의 조사 자료 간에 발생할 수 있는 체계적인 차이를 보정하는 대체 방법론을 개발하였고, 2017년 가계금융복지조사의 조사 및 조사+행정 자료를 분석에 이용하여 소득 및 부채 항목별로 조사값과 대체값을 비교하였다.

3장에서는 재산소득과 소득세의 시계열 연장 방법을 고민하였다. 소득연도 기준 2014년부터 행정자료값의 활용이 가능해짐에 따라서 2011-2013년의 (조사+행정) 값의 시계열 연장에 대한 요인이 발생하였다. 패널가구들의 응답 패턴에 따라서 횡단면 대체모형과 종단면 대체모형으로 나누어서 대체 모형을 개발하였다. 기본적으로는 비대체모형을 사용하였으며 필요에 따라서 회귀대체모형을 보완적으로 사용하였다. 실제로 대체된 재산소득과 소득세는 조사값의 특성 및 연계변수들의 변화를 무리없이 벤치마킹하고 있는 것으로 확인되었다. 부록에 조사 소득값이 없는 경우에 대한 분석 결과도 수록하였는데 이와 관련하여 추가적인 연구가 필요할 것으로 사료된다.

4장에서는 가계동향조사결과를 이용하여 가계금융복지조사의 시장소득과 가처분소득을 연결하는 작업을 진행하였다. 공통관측연도를 이용하여 가계금융복지조사 각 가구의 최근방 이웃 (Nearest Neighbors)를 찾고 그 이웃들의 소득변화비를 활용하여 시상소득 및 가처분소득의 대체값을 만들어내는 최근방 이웃 비대체모형을 개발하였다. 대체모형의 평가처럼 연결 대상이 길지 않고 가계동향조사의 결측치 비중이 작은 경우에는 제안 방법론이 대체적으로 잘 작동하지만, 연결 대상 연도가 길어지고 가계동향조사 가구들의 소득 결측치 비중이 높은 경우에는 소득대체 연도가 누적될수록 소득 평균 감소율이 커짐을 확인하였다. 이러한 문제점은 향후 보다 심도있고 검토되어 논의해야 될 것으로 사료되며, 또한 가계동향조사를 활용하여 시계열 연장을 하는 것에는 신중해야 될 필요가 있는 것으로 판단된다.

마지막으로 본 연구보고서에 소개된 결과와 관련된 내용들은 R 코드로 작성되었으며, 해당 R 코드 내용은 <부록 1과> <부록 2>에 구별하여 정리하였다.

< 부 록 >

A. 2장 R 프로그램 패키지

본 연구 과제에서 제안하는 2.1절에서 소개한 대체 방법론을 자료 분석에 적용할 수 있는 R 프로그램 코드 패키지를 다음과 같이 제출한다.

<표 A-1. 조사값이 있는 경우 R 프로그램 파일 요약>

항목	방법론	파일번호	파일 명
근로소득	혼합 비 대체	1	MixtureRatio1.R
		2	MixtureRatio1_Funs.R, MixtureRatioFuns.cpp
		3	README_MixtureRatio1.txt
	조건부 혼합 비 대체	1	MixtureRatio1_Conditional_X.R
		2	MixtureRatio1_Conditional_X_Funs.R, MixtureRatioFuns_Conditional_X.cpp
		3	README_MixtureRatio_Conditional_X
기타소득	혼합 비 대체	1	MixtureRatio2_OtherIncome.R
		2	MixtureRatio2_OtherIncome_Funs.R, MixtureRatioFuns.cpp
		3	README_MixtureRatio2
	조건부 혼합 비 대체	1	MixtureRatio2_OtherIncome_Conditional_X.R
		2	MixtureRatio2_OtherIncome_Conditional_X_Funs.R, MixtureRatioFuns_Conditional_X.cpp
		3	README_MixtureRatio2_Conditional_X.txt
	(절편이 없는) 혼합 회귀 대체	1	MixtureGauss2_OtherIncome.R
		2	MixtureGauss2_OtherIncome_Funs, MixtureGaussFuns_noint.cpp
		3	README_MixtureGauss2.txt
	(절편이 없는)	1	MixtureGauss2_OtherIncome_Conditional_X.R

	조건부 혼합 회귀 대체		
		2	MixtureGauss2_OtherIncome_Funs_Conditional_X.R, MixtureGaussFuns_noint_Conditional_X.cpp
		3	README_MixtureGauss2_Conditional_X.txt

<표 A-2. 조사값이 없는 경우 R 프로그램 파일 요약>

항목	방법론	파일번호	파일 명
근로소득	혼합 회귀 대체	1	MixtureGauss1_Nosurvey.R
		2	MixtureGauss1_Nosurvey_Funs.R, MixtureGaussFuns.cpp
		3	README_MixtureGauss1_Nosurvey.txt
	조건부 혼합 회귀 대체	1	MixtureGauss1_Nosurvey_Conditional_X.R
		2	MixtureGauss1_Nosurvey_Conditional_X_Funs.R, MixtureGaussFuns_Conditional_X.cpp
		3	README_MixtureGauss1_Nosurvey_Conditional_X.txt

단, <표 A-1>과 <표 A-2>에서 파일 번호는 각각 다음을 나타낸다.

▫ 파일번호 1: R 프로그램 실행 파일

▫ 파일번호 2: R 프로그램 실행 파일에 필요한 소스 파일

▫ 파일번호 3: R 프로그램 실행 파일 코드 설명 (README_XXX.txt)

예를 들어, 조사값이 있는 경우 근로소득의 혼합 비 대체 모형에 관한 실행 파일의 코드 설명(README_MixtureRatio1.txt)은 아래와 같고, 함께 제출한 R 프로그램 실행 파일을 이용하여 자료 분석에 직접 적용할 수 있다.

```
==========  READ ME : MixtureRatio1.R 코드 설명
참고로 R 프로그램에서 "#"는 주석을 표시하는 명령어로, #로 시작하는 문구는 프로그램 실행에서 무시됨
(아래의 설명이 부족하다면 구글 창에 [R 함수이름]을 검색하면 다양한 예제 참고 가능)
```

```
######## 0 : 현재 R 창에 저장되어 있는 모든 객체 삭제 의미
> 두 개 이상의 프로그램 코드를 하나의 R 창에서 실행시킬 경우, 결과 저장이 오버랩되는
경우가 있으므로 이를 피하기 위해 사용
> 하나의 창에 하나의 코드를 돌릴 경우, 무시해도 되는 명령어

######## 1 : 작업 디렉토리 설정하는 명령어, "작업할 폴더 경로" 지정
> 보내드린 코드를 저장한 폴더의 경로로 지정
> 즉, 지정한 경로의 폴더 안에 "MixtureRatioFuns.cpp" 와 "MixtureRatio1_Funs.R"이 저장되어 있어야 함

######## 2-0: 코드를 실행하는데 필요한 R 패키지 설치하기
> install.packages("XXX") : XXX 패키지 설치 명령어
> 이를 실행하면 위치를 묻는 윈도우 창이 뜰텐데, 현재 머무는 국가 선택하면 됨
> 한번만 설치하면 그 다음부터 R 실행할 때에는 설치할 필요 없으므로 삭제해도 좋음

######## 2-1: R 패키지 불러오기

######## 2-2: 사용자 지정 함수 패키지 불러오기

######## 3-1: 데이터 불러오기, read.csv("데이터 경로")
> R 에서는 "pdat"라는 이름으로 데이터 저장
> head(pdat) 는 pdat의 첫 6 행을 예로 보여줌, 데이터가 어떻게 생겼는지 보고자 할 때 자주 사용
> dim(pdat)는 pdat의 dimension을 출력해주는 함수 (행 개수, 열 개수)

######## 3-2: 매칭 변수 L_YN을 이용하여 pdat 에서 Matched 변수 생성
> Matched = 1은 매칭을 0은 비매칭을 나타냄

######## 3-3: 연속형 나이를 범주형으로 변환하여 pdat에서 "gg02.1"과 "gg02.2"라는 범주형 변수 생성
> 중간 보고 발표 시 연령의 매칭, 비매칭 집단 간 효과 크기가 가장 컸음
> 따라서 연령을 이용하여 MAR 검정 통계량인 X2를 계산
> 이에 필요한 작업으로 연령을 범주형 변수로 변환

######## 3-4: 분석 대상인 2017년 자료만 선택하여 dat17 객체명으로 저장
> subset(데이터, 조건) : 데이터에서 조건을 만족하는 서브 데이터를 출력하는 함수

######## 3-5: dat17에서 행정 근로소득를 "ic.y"로 조사 근로소득을 "ic.x"로 R 변수 생성
```

######## 3-6: MAR 검정에 필요한 범주형 변수 생성
> dat17에 있는 gg02.1 변수를 dat17에 X2.var1이라는 변수로 저장
> c(x, y): x 와 y 벡터를 결합하는 함수
> X2.itv: MAR 검정 통계량 계산시 사용되는 범주형 변수 y.c 를 정의하기 위해 필요한 구간의 끝 값 저장
> quantile(x)[2:4] = x의 (0%, 25%, 50%, 75%, 100%) - 분위수에 해당하는 값 계산 후 2,3,4번째 값만 출력;
즉, 여기서는 25%, 50%, 75% 분위수만 X2.itv 라는 이름으로 저장

######## 4: 10-fold cross-validation

######## 4-1: 10-fold cross-validation에 사용할 데이터인 매칭 집단 선택하여 dat17.M으로 저장

######## 4-2: 10-fold cross-validation 실행
> MRM_CV10(데이터 명, G=최대 혼합성분 개수) : 조사값을 사용했을 때의 평균 RMSE와 G=1,2,3,4,5의 평균 RMSE 계산 및 출력
> "MRM_CV10"은 직접 코드를 작성한 사용자 정의 함수이므로, 구글 등 검색 불가능
> 함수 사용에 오류가 생긴다면 danhyang@iastate.edu로 이메일 문의

######## 5: 자료 분석 : 실제 자료에 imputation (대체) 실행

######## 5-1: Step 4의 10-fold cross-validation 의 결과로 가장 낮은 평균 RMSE를 가지는 값을 넣어준 후 자료 분석 실행

######## 5-2: 실제 자료, 즉, 비매칭 집단의 행정 자료값 imputation 실행
> output=DA.MRM(데이터 명, G): 대체 결과를 output이라는 객체명으로 R에 저장

######## 5-3: output에 어떤 결과들이 저장되어 있는지 결과 객체명 출력

######## 5-4: output에 저장된 결과 중 모수 추정값 출력

######## 5-5: output에 저장된 결과 중 MAR 검정 통계량인 X2 값 출력

######## 5-6: output에 저장된 결과 중 imp.d17 출력
> imp.d17은 비매칭 집단의 행정 자료 값을 대체한 결과까지 포함된 데이터
> write.csv(R 데이터 명, "출력 파일 저장 경로 및 파일 이름"): R 데이터를 csv 파일로 내보내기

######## 5-7: 매칭, 비매칭, 전체 집단에서 조사/행정/대체 값의 요약통계량 출력

B. 소득 항목 분석 결과 - 조사값이 없는 경우

2017년 가계금융복지조사에서 근로소득 항목을 이용하여 2.1.2절에 소개한 조사값이 없는 경우의 혼합 회귀 대체 방법을 적용해보고, 조사값이 있는 경우의 혼합비 대체 결과와 비교해보고자 한다.

1) 분석 절차

근로소득 항목은 조사값이 0보다 큰 가구원을 근로소득 대상자로 제한하여 분석한다. 혼합 회귀 대체 모형을 이용한 대체 방법론을 다음과 같이 적용할 수 있다.

- 후보 모형: 혼합 회귀 대체 모형1, 조건부 혼합 회귀 대체 모형 2
  - 예측변수: 소득세, 건강보험료
  - 조건부 혼합 회귀 대체 모형 2의 보조 변수 x: 연령(30대 미만, 30대, 40대, 50대, 50대 이상) 또는 교육정도 (초졸 미만, 중·고졸, 대졸 이상)
- G 선택: 각 후보 모형 별 10-fold 교차 검증법
  - 최대 혼합 성분 개수 $G_0 = 5$ 로 선택하여 분석
- 대체: 결정적 대체 방법
- MAR 검정 통계량(X^2) 비교 후 최종 대체 방법 선택
  - 범주형 변수 x_b: 연령 (30대 미만, 30대, 40대, 50대, 50대 이상)
  - 범주형 변수 y_c: 조사 근로소득 및 조사+행정 근로소득의 표본 사분위수를 기준으로 4개의 범주를 가지는 변수 정의

2) 분석 결과

<표 B-1>은 2017년 조사 가구원 중 근로소득이 있는 가구원을 분석 대상자로 제한하여 적용한 혼합 비 대체 모형 - 모형 1(기본 모형)과 조건부 모형 2(확장된 모형)의 결정적 대체 결과를 비교한다. 10-fold 교차 검증 결과, 혼합 회귀 모형 1의 가장 낮은 평균 RMSE는 $G = 2$일 때, 1930.6이다. 조건부 혼합 회귀 대체 모형 2(연령)에서는 가장 낮은 평균 RMSE는 $G = 2$일 때 1926.8이고, 조건부 혼합 회귀 대체 모형 3(교육정도)의 가장 낮은 평균 RMSE는 $G = 4$일 때, 1923.1이다. 혼합 회귀 모형을 이용한 대체값 모두 조사값을 사용했을 때보다 평균 RMSE가 높았다.

MAR 검정 통계량 계산 결과, 모형 1은 X^2=25.4, 모형 2(연령)은 X^2=25.3이고, 모형 3(교육정도)는 X^2=25.6으로 값이 거의 비슷했으며, 조사값을 사용했을 때 보다 검정 통계량 값이 낮음을 <표 B-1>에서 확인할 수 있다.

<표 B-1. 근로소득의 대체 방법별 평균 RMSE 및 X^2 통계량 비교>

대체 방법		G	평균 RMSE	X^2
조사값 사용		-	1652.1	41.3
혼합 회귀	모형 1	2	1930.6	25.4
	모형 2 - 조건부: 연령 (10-20/30/40/50대이상)	2	1926.8	25.3
	모형 3 - 조건부: 교육정도 (초졸 이하/중·고졸/대졸 이상)	4	1923.1	25.6

◦ 최종 대체 방법: 혼합 회귀 모형 3 - 조건부: 교육정도 ($G=4$)를 이용한 결정적 대체

<표 B-2>는 최종 대체 모형인 혼합 회귀 대체 모형 3 ($G=4$)의 혼합 비율 추정 결과를 보여준다. 4개의 혼합 성분을 가지는 모형으로, 교육정도에 따라 조사+행정 근로소득이 각 그룹에 속할 확률이 다르게 추정된 것을 알 수 있다. 교육정도가 초졸 이하이면, 그룹 4에 속할 확률이 가장 높았고, 중·고졸이상이면 그룹 1에 속할 확률이 가장 높다. <표 B-3>은 혼합 회귀 대체 모형 3의 예측 변수의 평균 및 회귀 계수의 추정 결과를 보여준다. 조사+행정 근로소득이 그룹 4에 속할 경우, 조사 건강보험료와는 관계없이 조사 소득세가 한 단위 증가할 때 12.79만큼 증가하는 평균을, 그룹 2에 속하면 조사 소득세와 관계없이 조사 건강보험료가 한 단위 증가할 때 6.58만큼 증가하는 평균을 가지는 분포를 따르는 것으로 추정되었다.

<표 B-2. 최종 대체 모형 모수 추정 결과1: 혼합 비율>

	교육정도		
그룹	초졸 이하	중·고졸	대졸 이상
1	0.21	0.47	0.46
2	0.27	0.20	0.08
3	0.02	0.08	0.32
4	0.51	0.25	0.14

<표 B-3. 최종 대체 모형 모수 추정 결과: 회귀계수 및 예측 변수의 평균>

그룹	$\beta_{g,0}$	$\beta_{g,1}$	$\beta_{g,2}$	$\mu_{g,1}$	$\mu_{g,2}$
	절편	소득세	건강보험료	로그 소득세	로그 건강보험료
1	885.78	10.00	14.92	3.85	4.40
2	1153.45	0.00	6.58	0.00	3.85
3	463.96	1.42	27.00	5.68	5.27
4	1194.66	12.79	0.00	0.51	0.00

<표 B-3>의 모수 추정값을 이용하여 결정적 대체값을 구한 후, 매칭 집단 별 조사 근로소득, 조사+행정 또는 대체 근로소득의 요약통계량을 정리하면 <표 B-4>와 같다. 여기서, 대체 근로소득으로 조사값을 사용하는 경우 (대체 1)와 조사값을 사용하지 않는 경우(대체 2)의 대체값에 대한 요약 통계량을 함께 제시하였다.

매칭 집단을 먼저 살펴보면, 조사 근로소득에 비해 조사+행정 근로소득의 1, 2사분위수는 낮은 반면, 3사분위수 및 평균은 더 높다. 비매칭 집단에서 조사 근로소득과 대체 근로소득의 요약통계량을 살펴보면, 조사 근로소득에 비해 대체 1 근로소득의 1사분위수는 높았으나 2사분위수는 낮았고, 3사분위수 및 평균은 더 높다. 반면, 대체 2 근로소득의 경우, 1,2, 그리고 3사분위에서 모두 조사 근로소득보다 낮으나, 평균은 좀 더 높다. 전체 (매칭, 비매칭) 집단에서도 유사한 패턴을 확인할 수 있다.

<표 B-4. 매칭 집단별 근로소득 (조사 · 조사+행정 · 대체) 요약 통계량 비교>

		n	1사분위	2사분위	3사분위	평균
매칭	조사	15664	1440	2400	4000	3145
	조사+행정		1200	2228	4220	3199
비매칭	조사	2047	1500	2400	3710	2929
	대체 1		1535	2356	3715	2977
	대체 2		1380	2191	3637	2968
전체 (매칭,비매칭)	(조사, 조사)	17711	1445	2400	4000	3120
	(조사+행정,조사)		1203	2280	4140	3167
	(조사+행정, 대체 1)		1240	2270	4140	3173
	(조사+행정. 대체 2)		1216	2220	4126	3172

◦ 여기서, 대체 1은 조사값이 있는 경우 대체(비 대체) 결과를, 대체 2는 조사값을 사용하지 않은 대체(회귀 대체) 결과를 나타낸다.

C. 3, 4장 R 프로그램 패키지

본 부록에서는 3장과 4장에 작성에 쓰인 R 코드 내용을 간략하게 정리하였다.

<표 C-1. 3장 내용 관련 R code 정리>

R 코드명	내용	비고
S3_C1_가계금융복지조사_데이터 요약	데이터 요약	3.1
S3_C2_가계금융복지조사_응답 패턴	응답 패턴 도출	3.1
S3_C3_가계금융복지조사_재산소득평가	모형평가-재산소득	3.3
S3_C4_가계금융복지조사_소득세평가	모형평가-소득세	3.3
S3_C5_가계금융복지조사_순차대체	순차대체 과정	3.4

<표 C-2. 4장 내용 관련 R code 정리>

R 코드명	내용	비고
S4_C1_가계동향조사_데이터 요약	데이터 요약	4.1
S4_C2_가계동향조사_응답 패턴	응답 패턴 도출	4.2
S4_C3_가계동향조사_모형_기부가구	모형-기부가구 구성	4.3
S4_C4_가계동향조사_모형_수령가구	모형-수령가구 구성	4.3
S4_C5_가계동향조사_모형_연결	모형-연결결과	4.3
S4_C6_가계동향조사_기부가구	기부가구 구성	4.4
S4_C7_가계동향조사_수령가구	수령가구 구성	4.4
S4_C8_가계동향조사_연결	연결 결과	4.4