경제통계 무응답 대체 (Imputation) 매뉴얼

2011. 12.

통계청
STATISTICS KOREA

경제통계 무응답 대체 (Imputation) 매뉴얼

2011. 12.

통계청
STATISTICS KOREA

목 차

Ⅰ 개요

1. 무응답이란?

○ 자료 수집 과정에서 항목 전체 또는 일부 항목이 측정되지 않은 경우 이를 무응답(nonresponse)이라 하며, 항목의 값이 관측되지 않고 빠져있다는 의미로 결측값(missing value)이라고도 부름

○ 통계조사에서 발생하는 무응답은 조사단위(사업체, 기업체 등) 자체가 응답을 하지 않는 단위 무응답(unit nonresponse)과, 조사표의 일부 항목에 대해서만 응답을 하지 않는 항목 무응답(item nonresponse)으로 나눌 수 있음

○ 전수조사에서는 단위 무응답, 항목 무응답 모두 재조사(call back)를 하거나 결측값을 대체(imputation)하나,

○ 표본조사에서는 항목 무응답은 대체(imputation)하고, 단위 무응답은 표본을 교체하거나 조사단위(사업체, 기업체)의 가중치를 조정(nonresponse weighting adjustment)하는 방식이 많이 사용됨

※ 가중치 조정은 단위무응답이 발생한 표본단위가 속한 계층 내에서 다른 표본단위의 가중치를 조정하는 방식으로 이루어짐. 예를 들어 단위무응답이 발생한 현대플라스틱(주)의 가중치가 100이고 이 표본단위가 속한 계층의 다른 표본의 가중치가 ○○(100), △△(100), ××(100)이라면, ○○(133), △△(133), ××(133)으로 조정함

2. 무응답 대체란?

○ 통계조사에서 자주 발생하는 무응답은 조사 자료의 품질을 저하시키므로 결측(missing)된 자료에 특정 값을 넣어줄 필요가 있는데, 이를 무응답 대체(imputation)라 함

○ 무응답 대체의 목적은,

① 공식통계의 기본 원칙 중 하나인 **One number principle***을 지키는 데에 일차적인 목적이 있으며,

* One number principle : 하나의 모수에는 하나의 추정치를 생산하자는 뜻. 만약 결측된 자료 자체를 사용자들이 각자 분석하도록 한다면, 이 원칙에 위배되어 하나의 모수에 여러 개의 추정치가 공존하게 되고 이는 통계 자체의 신뢰성을 훼손시키게 됨

② 무응답으로 인한 추정치의 편향을 보정하고,

③ 관측된 다른 정보들을 반영함으로써 모수추정에 대한 효율을 높이는 데에 있음

3. 무응답 대체 매뉴얼이란?

○ 무응답 대체 매뉴얼이란, 자료를 수집하는 과정에서 측정되지 않은 무응답 값을 적절한 방법을 통하여 실제값과 유사한 값으로 채우는 대체(imputation) 과정에 필요한 사항들을 규정한 업무지침서임

- 현재 대부분의 통계조사에서 무응답 대체가 이루어지고 있으나, 체계적인 방법론 대신 주관적인 판단에 의한 경우가 많음. 이러한 대체 방법은 많은 경험에 근거하여 이루어진다는 장점은 있으나

근본적인 해결방안은 아니며, 따라서 통계적 이론에 근거하여 대체의 정확도를 높일 수 있는 체계적인 가이드라인을 제시하는 것이 본 매뉴얼의 주요 목적임

- 본 매뉴얼은 무응답 대체의 개념과 방법, 대체 과정 등을 기본적인 수준에서 다루었으므로, 추후 보다 구체적이고 발전된 형태로 보완해 나아갈 필요가 있음

II 무응답 대체의 방법

1. 구분

○ 결측값 대체의 방법은 크게 결정적 대체법과 확률적 대체법으로 나뉨. 결정적 대체법에서는 대체값이 유일한 값으로 정해지는 반면, 확률적 대체법에서는 확률적인 변동을 부여하여 대체값을 선택하므로 매번 다른 대체값이 나타날 수 있음

2. 결정적 대체법

○ 의의 : 결정적 대체법은 무응답 항목에 유일하게 결정된 대체값을 대입하는 방법으로, 반복해서 대체를 수행하더라도 동일한 결과를 얻게 됨

○ 종류

① 연역적 대체 : 조사표 상의 다른 항목에 근거하여 실제값을 추리해낼 수 있는 경우 사용되는 방법

- 예를 들어 A, B, C, D 네 항목의 합에서, A=60, B=40, C=빈칸, D=빈칸이고 A+B+C+D=100이라면 C=0, D=0으로 실제값을 추리함

② 평균 대체(**mean imputation**) : 전체를 몇 개의 대체군*으로 분류한 후 무응답 항목을 해당 대체군의 (이상치를 제외한) 응답 평균값으로 대체하는 방법

* 대체군 : 사업체 조사의 경우, 동일한 유형(법인, 개인사업자), 업종(중분류 또는 소분류), 규모(매출액, 종업원수) 등을 바탕으로 대체군을 형성

- 간단하여 이용하기 쉽다는 장점이 있으나, 대체 후 **평균값의 빈도수가 많아져 분포가 왜곡된다는 단점이 있음**

⇒ 따라서 가급적이면 다른 방법을 우선 적용하고, **무응답 항목이 극소수인 경우 등에 제한적으로 사용**

③ **순차적 핫덱 대체 : 동일 조사의 다른 응답자로부터 얻은 1개 이상의 제공 레코드들이 일정한 순서로 정렬되어 있어서 한 번에 한 레코드씩 순차적으로 결측 항목에 들어가는 방법**

(참고 : 핫덱 대체 방법 정리)

▪ **핫덱 대체** : 결측된 응답자(=수용자, recipient)의 값을 **동일 자료 내의 다른 응답자(=제공자, donor)의 값으로 대체하는 것을 의미**

☞ **콜덱 대체** : 핫덱 대체는 무응답 값을 동일한 조사에서 비슷한 성향을 가진 응답자의 값으로 대체하나, 콜덱 대체는 대체할 자료를 외부 출처(기존에 실시된 다른 조사자료 등)에서 가져온다는 점에서 차이가 있음

▪ **대체군을 이용한 핫덱 대체** : 무응답 값의 대체를 위하여 자료 내의 응답값을 이용하되, 유사한 성질을 가진 대체군 내에서 제공자(donor)를 선택하는 경우를 의미

☞ 사업체 조사의 경우, 동일한 유형(법인, 개인사업자), 업종(중분류 또는 소분류), 규모(매출액, 종업원수) 등을 바탕으로 대체군을 형성함. 이 때 다양한 변수들을 고려할수록 편향(bias)은 줄어드나, 조건들을 모두 만족시키는 제공자를 찾지 못하게 될 수도 있음. 따라서 대체군 형성 변수들을 고려할 때에는 무응답이 발생한 변수와 밀접한 연관성을 가지는 변수들을 우선 포함시키면서 적정 수준을 찾는 것이 중요

▪ 순차적 핫덱 대체와 랜덤 핫덱 대체 : 동일 조사의 다른 응답자(=제공자, donor)의 값으로 대체하는 점은 같으나, 순차적 핫덱 대체는 제공된 값들이 한 번에 한 레코드씩 일정한 순서로 결측 항목에 들어가는 반면, 랜덤 핫덱 대체는 제공된 값들이 무작위로 선택되어 결측 항목에 들어가게 됨

☞ 따라서 순차적 핫덱 대체는 여러 번 반복하더라도 동일한 결과를 얻게 되는 반면, 랜덤 핫덱 대체는 매 번 다른 값으로 대체될 수 있음

▪ 순차적 핫덱 대체, 랜덤 핫덱 대체, 평균 대체 비교

사업체 A, D, E, G, H, I가 속한 대체군에서 추출한 종사자수(제공 레코드)
X_1=10 X_2=11 X_3=15

⇨

사업체	종사자수
A	(결측값)
B	15
C	15
D	(결측값)
E	(결측값)
F	10
G	(결측값)
H	(결측값)
I	(결측값)

⇨

순차적 핫덱 대체	랜덤 핫덱 대체	평균 대체
(10)	(11)	(12)
15	15	15
15	15	15
(11)	(10)	(12)
(15)	(11)	(12)
10	10	10
(10)	(15)	(12)
(11)	(15)	(12)
(15)	(11)	(12)

☞ 사업체 A, D, E, G, H, I의 종사자수 항목이 결측되었고 이들 사업체가 속한 대체군에서 일정 조건에 의해 X_1=10, X_2=11, X_3=15 라는 제공 레코드가 선택되었다고 가정할 때, 순차적 핫덱 대체는 X_1, X_2, X_3, X_1, X_2, X_3 … 값이 차례대로 결측 항목에 들어가고, 랜덤 핫덱 대체는 무작위로 들어가며, 평균 대체는 해당 대체군의 응답 평균값 $\bar{x}$= $(X_1+X_2+X_3)/3$ = 12가 들어가게 됨

④ **회귀 대체(Regression random imputation)** : 무응답이 발생한 항목(반응변수) y의 대체값을 찾아내기 위하여, 응답이 있는 항목들(설명변수) $x_1, x_2, \cdots, x_n$ 을 **회귀모형**에 적합시키는 방법

- 예를 들어 A사업체의 매출액(y)이 결측되었고 종사자수(x=1,000명)는 알고 있다면, 대체군에 해당되는 다른 사업체들의 매출액과 종사자수의 관계를 회귀모형에 적합시킴. 그 결과 **y=3x라는 회귀식**이 추정되었다면, 이 추정식으로부터 A사업체의 매출액을 y=3×1,000=3,000으로 대체함

⑤ **비 대체(ratio imputation)** : 부가적 정보로부터 둘 이상의 변수들 간에 존재하는 관계(比)를 파악하여 이를 이용하는 방법

- 예를 들어 A사업체에서 1달간 지불된 급료(a)가 결측되었고 임금을 받은 종사자수(y)는 알고 있다면, 해당 대체군의 **1달간 지불된 평균 급료와 평균 피고용인수의 비($\bar{a}/\bar{y}$)**를 계산하여 $a = (\bar{a}/\bar{y}) \times y$로 구함

- 사업체 조사에서 가장 많이 사용되는 대체법 중 하나이며, 대체군은 산업분류가 주로 사용됨

⑥ **최근방 이웃 대체(Nearest neighbor imputation)** : 전체를 몇 개의 대체군으로 분류한 뒤, 보조변수를 이용하여 각 대체군에서의 응답자료를 순서대로 정리한 후 **무응답 개체와 보조변수가 가장 유사한 응답 개체**를 찾아 대응되는 항목값으로 대체하는 방법

- 만약 재고량이 결측되었다면 대체군 내에서 **판매량(보조변수)**을 이용하여 판매량 순대로 나열한 후 **판매량이 가장 유사한 제공자**를 찾아 **이 제공자의 재고량 값**으로 대체함

- 예를 들어 지역별/산업별/사업체 크기별로 나누어 층화한 다음과 같은 대체군이 있을 경우 사업체B의 총급여는?

사업체	종사자수	총급여
A	100	60,000
B	90	(결측값)
C	120	75,000
D	110	67,000

⇒ 사업체B와 종사자수(보조변수)가 가장 비슷한 응답 사업체를 찾음 → 위 경우 사업체A가 이에 해당됨 → 사업체A의 총급여를 적용 → 따라서 사업체2의 무응답 대체값은 60,000이 됨

⑦ 과거자료 대체**(Historical unit imputation)** : 이전 조사에서 얻은 자료를 대체값으로 그대로 이용하는 방법으로, 과거의 값에 사전적으로 정한 조정원칙(예를 들어 해당 산업의 평균 매출액 상승률)을 적용하여 수정할 수도 있음

- 예를 들어 A사업체의 정기 휴무일수가 결측되었고 전년 조사에서 A사업체의 정기 휴뮤일수가 130일이었다면, 130일을 대체값으로 적용함

- 동일 항목의 조사값이 조사시점에 영향을 받지 않고 안정된 값을 보이는 경우 유용함

3. 확률적 대체법

○ 의의 : 대체값의 결정 과정에 랜덤하게 결정되는 부분이 있어서 반복적으로 임퓨테이션을 실시할 때 각각 다른 값으로 대체되는 경우를 의미함

○ 종류

① 랜덤 핫덱 대체(Random hot deck imputation) : 무응답 값의 대체를 위해 자료내의 응답값을 이용하되, 제공자가 대체군 내에서 임의로 선택되는 방법

② 랜덤 회귀 대체(Regression random imputation) : 회귀 대체와 유사하나, 마찬가지로 기존의 대체값에 확률오차를 포함시켜 대체값으로 이용함으로서 대체값의 선택에 무작위적 요소를 더함

③ 랜덤 비 대체(Ratio random imputation) : 비 대체와 유사하나 기존의 비 대체값에 확률오차를 포함시켜 대체값으로 이용하는 방법으로, 확률오차는 대체로 인한 변동의 감소를 보정해주는 효과가 있음

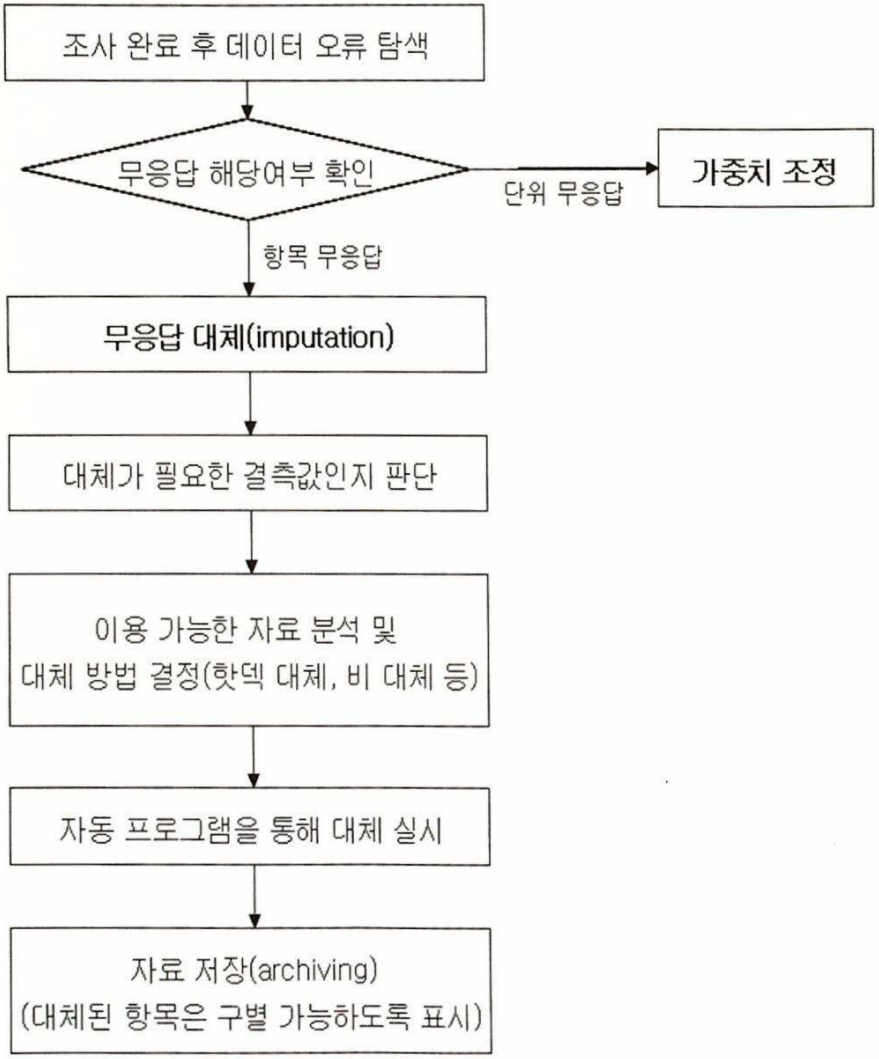
4. 기타 방법

○ 주관적 대체(Subjective imputation) : 대체를 수행하는 담당자가 자신의 경험, 지식, 다른 분석자료 등을 이용하여 각 항목에 가장 적절할 것으로 생각되는 값을 정하는 것을 의미(전문가의 판단)

- 단위 무응답보다는 항목 무응답의 경우, 적절한 대체 방법을 찾지 못한 경우에 최종적인 방법 중 하나로 검토할 수 있으며, 제한적으로 적용할 여지가 있음

III 무응답 대체 과정

1. 무응답 대체 순서



2. 대체 과정별 매뉴얼

1 데이터 오류 탐색

○ 데이터 오류는 관측 값과 실제 값이 다른 경우로 이상치, 결측값 등의 형태로 나타남

- 이상치는 대다수의 데이터에서 멀리 떨어진 관측치를 의미하며 오류를 포함할 가능성이 높음

※ 이상치는 보통 전체 데이터나 대부분의 데이터를 동시에 이용하여 오류를 탐색하는 매크로 에디팅(macro editing) 단계에서 처리함

- 결측값은 값이 있어야 할 항목에 값이 없는 경우를 의미하며, 응답자가 응답을 하지 않거나 데이터 입력 과정에서 실수로 응답 데이터가 누락되어 발생

※ 구조적 결측값(structurally missing value)은 해당되는 사항이 없는 문항에 대하여 응답하지 않은 경우로, 반드시 응답해야 하는 항목이 결측된 일반 결측값과는 구분해야 함

(예시 : 기업활동조사의 경우)

▪ 전년도 종사자수가 50명인 기업체A의 금년 종사자수가 130명으로 조사된 경우

☞ 이상치에 해당되는지 확인(재조사 등을 통해 정당한 사유가 있는 경우인지 파악)

▪ 연구개발비 항목이 빈 칸인 경우

☞ 결측값에 해당되는지 확인(연구개발 활동을 하지 않아 연구개발비가 0인 경우와 구분)

2 무응답 해당여부 확인

○ 결측값의 종류가 응답자가 전혀 응답을 하지 않은 **단위 무응답(unit nonresponse)**인지, 일부 항목에만 응답을 하지 않은 **항목 무응답(item nonresponse)**인지 구분

- 단위 무응답은 응답자의 가중치를 조정하는 **가중치 조정법**이 많이 사용되고, 항목 무응답은 주로 **대체(imputation)** 처리함

(예시 : 기업활동조사의 경우)

▪ 전년도에 조사가 되었고, 금년도에도 조사대상에 해당되는 기업이나 불응한 경우

☞ 단위무응답에 해당되므로 재조사 또는 무응답 대체 검토

▪ 다른 항목은 모두 조사되었으며 연구개발비 항목만 결측된 경우

☞ 항목무응답에 해당되므로 무응답 대체 검토

3 무응답 대체(imputation)

① 대체가 필요한 결측값인지 판단

○ 항목 무응답의 경우 대체가 필요한 결측값인지 판단

- 결측값 대체가 필요한 변수들을 식별하는 방법으로 **Fellegi**와 **Holt**방식에 따르면 ①각 레코드에 있는 자료들은 최소한의 변수값만을 바꿈으로써 모든 에디팅 기준들을 만족시킬 수 있어야 하고, ②가능한 한 자료 파일의 도수분포 상태를 유지해야 하며, ③대체의 기준은 특정한 대체법에 한정되지 않음

(예시 : 기업활동조사의 경우)

- 결측이 지나치게 많은 변수나 레코드는 수정하는 것보다 제거하는 것이 더 나을 수도 있음

☞ 기업활동조사의 경우 재조사가 불가능하고, 기업 공시자료도 없으며, 통계 전체에 미치는 영향이 적은 기업체의 경우 대체하지 않고 제외시킴

② 이용 가능한 자료 분석 및 대체 방법 결정

❍ 이용 가능한 자료로는 동일 조사의 다른 응답자의 응답값, 결측 변수의 과거 조사값, 센서스나 행정 데이터와 같은 외부 자료 등이 있음

❍ 이용 가능한 자료와 결측값의 특성 등을 고려하여 가장 적절한 대체 방법을 선택함 (여러 가지 방법을 사용해서 한 번 이상의 대체를 시행하고 결과를 비교해본 뒤 선택 가능)

- 예를 들어 총 고용규모는 응답했으나 다른 항목에 하나 이상의 무응답이 존재하며 이 사업체의 과거자료가 없는 경우, 최근방 이웃 대체와 핫덱 대체 방법을 같이 사용할 수 있음

⇒ 동일 조사의 응답 데이터들을 산업별로 층화한 후, 그 층 내에서 고용규모의 크기대로 정렬하여 총 고용규모가 가장 비슷한 응답 사업체를 찾고, 이 사업체의 해당되는 응답값을 결측값에 적용

(예시 : 기업활동조사의 경우)

▪ 종사자수

i) 전년 조사자료가 있는 경우 전년 자료값에 해당 산업의 평균 종사자수 증감률을 적용하고(과거자료 대체),

ii) 전년 조사자료가 없는 경우 해당 산업의 종사자수 평균값을 적용함(평균 대체)

▪ 유·무형자산 당기취득액

i) 우선적으로 기업 공시자료(대한상공회의소 자료)를 확인하여 있으면 그대로 적용함

ii) 기업 공시자료가 없는 경우 다음으로 전년 조사자료를 확인, 전년 조사자료가 있는 경우 전년 자료값에 해당 산업의 평균 당기취득액 증감률을 적용하고(과거자료 대체),

iii) 전년 조사자료도 없는 경우 해당 산업의 당기취득액 평균값을 적용함(평균 대체)

▪ 기업의 경영방향(외부위탁, 가맹점 유치 등)

i) 전년 조사자료가 있는 경우 전년 자료값을 그대로 적용하고 (과거자료 대체),

ii) 전년 조사자료가 없는 경우 해당 산업(대체군) 내에서 보조변수 (종사자수)를 이용하여 가장 유사한 기업을 찾고, 이 기업의 조사내용을 적용함(최근방 이웃 대체)

③ 자동 프로그램을 통해 대체 실시

❍ 각 조사별 특성(연간조사/월간조사, 전수조사/표본조사, 무응답 발생 빈도 등)에 따라 무응답 대체 프로그램 설계 및 활용

❍ 자동 프로그램으로 무응답 대체 실시 후, 대체된 값이 다른 변수와의 연관관계에 있어 논리적으로 문제가 없는지 사후 내검 실시

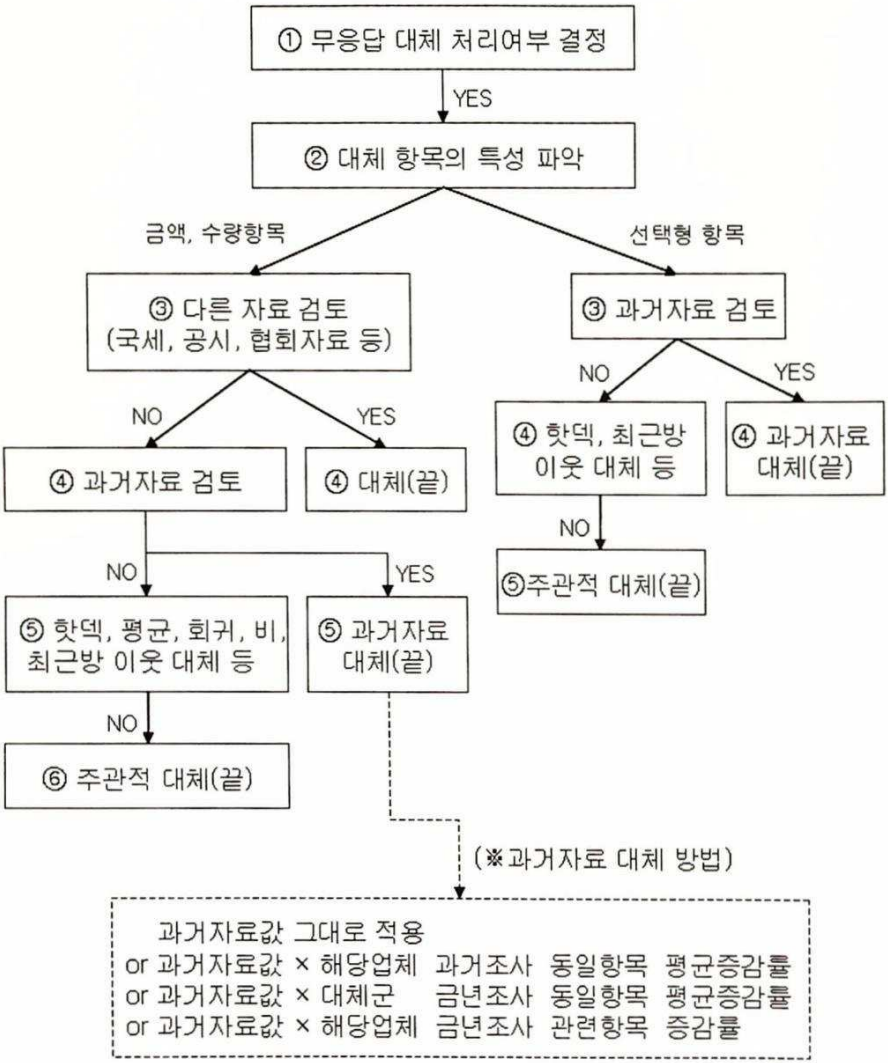
④ 자료 저장(archiving)

❍ 자료 처리 과정의 데이터와 대체 방법, 사용된 프로그램 등 중요한 정보를 기록으로 남겨 저장하는 것

❍ 특히 무응답 대체를 거친 데이터의 경우 기존의 데이터 값과 구별할 수 있도록 별도로 표시

- 자료 저장은 ①대체 관련 정보를 이용자들에게 공개함으로써 통계자료에 대한 신뢰성을 높일 수 있고, ②2차 분석이 이루어지거나 대체된 데이터가 다른 통계에 입력값으로 사용될 때 참고자료가 되며, ③대체 전·후의 통계 품질을 측정하는 데에도 유용하게 활용될 수 있음

3. 무응답 대체 템플릿(Template)



※ 템플릿을 각 조사에 맞게 활용하여 무응답 대체를 시행

1 금액 조사항목(매출액, 영업비용, 영업이익, 자산규모 등), 수량 조사항목(종사자수, 보유대수, 보유건수, 이용건수 등)

① 무응답 대체 처리여부를 결정

② 대체항목의 특성 파악 : 금액, 수량 조사항목

③ 다른 자료 검토 : 해당 업체의 해당 항목을 조사한 다른 자료가 있는가?

☞ 국세자료, 공시자료, 협회자료, 다른 통계자료 등 무응답 값을 알 수 있는 자료가 있는 경우 우선 적용

④ 과거자료 검토 : 해당 업체의 해당 항목을 조사한 과거자료가 있는가?

☞ 과거자료가 있는 경우 : i)변동이 거의 없는 안정적 항목은 과거 자료값을 그대로 적용해도 되나, 대부분은 과거자료값에 ii)해당업체 과거조사 동일항목 평균증감률을 적용하거나, iii) 대체군의 금년조사 동일항목 평균증감률을 적용하거나, iv)해당 업체 금년조사 관련항목 증감률*을 적용하는 등의 방식으로 조정을 거침

* 예시 : A사업체 금년 재료비 항목(무응답) = A업체 전년 재료비 × A업체 금년 매출원가 증감률

☞ 과거자료가 없는 경우 : i)해당 대체군에서 핫덱 대체, 평균 대체, 비 대체, 회귀 대체, 최근방 이웃 대체 등을 선택하여 적용하고, ii)최종적 방법으로 주관적 대체를 실시

2 선택형 조사항목(국외진출 여부, 가맹점 유치여부, 사무실 사용 여부 등)

① 무응답 대체 처리여부를 결정

② 대체항목의 특성 파악 : 선택형 조사항목

③ 다른 자료 검토 : 해당 업체의 해당 항목을 조사한 과거자료가 있는가?

☞ 과거자료가 있는 경우 : 과거 자료값의 적용을 검토

※ 단, 해당 업체의 동일 조사 내에서의 다른 응답 항목과 연관관계를 고려해야 함. 예를 들어 '국외진출 여부' 항목을 대체할 때, 해당 업체의 과거 조사결과가 '아니오'에 체크되어 있더라도, 금년 조사에서 '국외지사 종사자수' 항목에 종사자가 있는 경우라면, '국외진출 여부'는 '예'라고 대체해야 함

☞ 과거자료가 없는 경우 : 동일 조사 내에서 유사한 업체가 있는지 확인. 무응답 항목의 성격에 따라 i)대체군 내에서 산업분류, 조직형태, 종사자수 등이 가장 유사한 업체의 응답값을 찾아 핫덱 대체 또는 최근방 이웃 대체 등을 실시하고, ii)차선책으로 대체군 내에서 응답이 가장 많은 항목 즉 최빈값으로 대체하거나, iii)최종적 방법으로 주관적 대체를 실시

IV 기타 참고사항

○ 좋은 무응답 대체(imputation)란?

- 적절한 무응답 대체 방법을 선택하는 데에 있어서는 ①대체로 인해 실제 분포를 왜곡할 가능성에 주의해야 하고, ②모든 항목값을 관찰하지 않음으로서 발생하는 편의를 줍혀야 하며, ③대체된 데이터가 내적 일치성이 있어 자동화되고 반복될 수 있어야 함. 또한 ④사후적 평가가 가능하도록 대체 기록을 남겨야 함

○ 향후 과제

- 경제통계 조사별, 항목별로 가장 적절한 대체방법을 찾고 일관된 수정 논리에 따라 무응답 대체를 수행함으로써, 결측된 실제값과 대체값의 차이를 줄여 나가려는 노력이 필요함

- 무응답 대체는 통계조사 과정에서 데이터 오류(data error)를 탐색하고 수정하는 에디팅(editing) 과정 중 한 부분으로서, 향후 무응답 대체를 포괄하는 에디팅 매뉴얼의 개발이 요구됨

- 나아가 호주, 캐나다와 같이 무응답 대체를 비롯한 에디팅의 전(全) 과정을 자동으로 처리할 수 있는 자동 에디팅 프로그램을 도입할 필요가 있음

(참고 : Statistics Canada의 에디팅 프로그램)

- **Statistics Canada**에서는 **Banff**라는 데이터 처리 시스템을 개발하여 자동으로 데이터 에디팅을 실시하고 있음

- 특징 : **SAS**에 기초하고 있으며, 시스템을 구성하는 **9**개의 **SAS** 프로시저는 독립적 또는 복합적으로 활용 가능

- 프로시저(순서에 상관없이 사용 가능)

1. **Edit Specification** : 각 data field 간의 상관관계를 분석하여 필요한 최소한의 edits rule을 결정

2. **Edit Summary Statistics Tables** : 결정된 edits system이 실제 데이터에 부합하는지 여부를 체크

3. **Outlier Detection** : 이상치 존재 여부를 판단, 이상치 중에 대체가 필요한 데이터를 나타내어 줌

4. **Error Localization** : 어떠한 field에서 수정이 이루어져야 하는지, 대체가 필요한 field를 결정

5. **Deterministic Imputation** : 대체가 필요하다고 판정된 field에 대해 사용자가 지정한 값을 이용하여 대체 실시

6. **Donor Imputation** : 최근방 이웃 접근법을 사용하여 가장 적절한 값을 선택

7. **Estimate Imputation** : 한 번의 프로시저를 실행하면서 여러 변수들에 대해 대체를 실행. 첫 번째 시도한 대체가 성공적이지 않을 경우 다른 estimator를 사용해 다시 대체 시도

8. **Pro-Rating** : 부분의 합이 데이터의 전체 합과 일치하는지를 확인하는 과정(전체 합은 정확하다고 보고 개별 항목을 수정)

9. **Mass Imputation** : 두 단계의 조사(two-phase survey)와 같이 보조샘플에서만 자세한 정보가 조사되는 경우, 두 번째 단계에서 선택되지 않은 단위들의 무응답을 가장 가까운 제공자(donor)를 선택하거나 랜덤으로 제공자를 선택하여 대체