에디팅 품질관리 매뉴얼

Editing Quality Management Manual

2008. 12

한국통계학회

에디팅 품질관리 매뉴얼

Editing Quality Management Manual

2008. 12

한국통계학회

- 연구과제명 : 에디팅 품질관리 매뉴얼
- 연 구 기 간 : 2008년 7월 ~ 2008년 12월
- 연구수행자 :

연구수행기관		한국통계학회
연 구 책 임 자	김규성	서울시립대학교 통계학과 교수

▮머 리 말▮

본 매뉴얼은 통계청으로부터 연구용역을 의뢰받아 한국통계학회에서 수행한 “에디팅 품질관리 매뉴얼 개발”에 대한 최종 연구결과이다.

통계조사에서 데이터 에디팅은 조사·보고된 데이터를 처리하는 과정에서 데이터 오류를 탐색, 수정하는 활동을 말한다. 대부분의 통계조사는 데이터 품질을 향상시키기 위하여 데이터 처리 과정에 데이터 에디팅 절차를 포함하고 있다. 통계조사의 규모와 중요도에 따라 적용되는 에디팅 방법은 달라질 수 있다. 그러나 데이터 오류를 탐색하고 수정하는 에디팅의 기본 원리는 동일하기 때문에 에디팅의 기본 원리를 이해하면 통계조사에 에디팅 방법을 적용하기가 한결 수월할 것이다.

본 매뉴얼은 통계조사의 에디팅 업무에 활용하기 위하여 작성되었다. 에디팅의 개념과 조사 절차에 따른 에디팅을 소개하였고, 데이터 오류 탐색 방법과 오류 데이터 처리 방법을 설명하였다. 또한 에디팅을 문서로 기록하는 방법을 소개하였다. 본 매뉴얼은 에디팅 부분만을 집중적으로 다루고 있지만 에디팅은 통계조사의 일부분이므로 통계조사 전체 과정에서 데이터 에디팅을 이해하고 적용하여야 한다. 이에 관해서는 캐나다 통계국(2003)에서 나온 “Survey methods and practices”이나 Biemer & Lyberg (2003)가 쓴 “Introduction to survey quality”을 참고하기 바란다. 아무쪼록 본 매뉴얼이 에디팅 업무에 널리 활용되어 에디팅 업무에 도움이 되고 더 나아가 통계품질 향상에 일조할 수 있기를 기대한다.

본 매뉴얼을 개발하는데 필요한 지원을 아끼지 않은 통계청 품질관리과 관계자 여러분께 감사드린다. 그리고 본 매뉴얼 초고를 읽고 건설적인 조언을 해 주신 한국은행 박인호 박사께 깊은 사의를 표한다. 마지막으로 본 연구에 참여하여 수많은 시간을 함께 한 서울시립대학교 대학원생 이영민, 심정숙 양과 엽승현 군에게 고마움을 표한다. 이들의 열정과 정성에 힘입어 본 매뉴얼이 완성되었다.

2008년 12월

연구책임자 서울시립대학교 통계학과 교수 김규성

▮차 례▮

Ⅰ. 데이터 에디팅 개요

1. 데이터 에디팅 ·····1
2. 데이터 에디팅 방법 ·····3
3. 데이터 에디팅 제한조건 ·····6
4. 데이터 에디팅 품질관리 ·····8

Ⅱ. 데이터 에디팅 과정

1. 데이터 에디팅 설계 ·····10
2. 데이터 에디팅 절차 ·····11

Ⅲ. 데이터 오류 탐색

1. 데이터 오류 ·····17
2. 결측값 ·····18
3. 체계적 오류 ·····24
4. 영향력 있는 오류 ·····28
5. 이상치 ·····35
6. 랜덤 오류 ·····43

Ⅳ. 오류 데이터 처리

1. 개요 ·····51
2. 대체 처리 ·····52

3. 무응답 대체 방법 예제 ····· 58
4. 쌍방향 처리 ····· 64

Ⅴ. 데이터 에디팅 문서 기록

1. 개요 ····· 68
2. 데이터 에디팅 방법론 문서 ····· 69
3. 리포트 ····· 69
4. 자료 저장 ····· 73

부록A. 에디팅 가이드라인 74

부록B. 용어 해설 77

참고문헌 82

▮표 차 례▮

〈표 3-1〉 오류 처리 결과 ····33
〈표 4-1〉 데이터 기호 ····55
〈표 4-2〉 교육 기간 대체를 위한 대체 칸 구성 ····60
〈표 4-3〉 중범죄 피해 대체를 위한 대체 칸 구성 ····61
〈표 4-4〉 대체 후 완비 데이터 ····64

▮데이터 차 례▮

[데이터 3-1] 범죄 피해 관련 조사 ····20
[데이터 3-2] 면접 조사 ····22
[데이터 3-3] 가중치 데이터 ····23
[데이터 3-4] 근로자 평균 연봉 ····27
[데이터 3-5] 일변량 표본 데이터 ····37
[데이터 3-6] 학생의 팔 길이 ····40
[데이터 3-7] 사회조사 데이터 ····47
[데이터 4-1] 대체 처리 예제 ····59

▮그 림 차 례▮

[그림 3-1] 회사별 근로자 평균 연봉 막대그래프 ·····28
[그림 3-2] 근로자 수 막대그래프 ·····30
[그림 3-3] 해당 회사를 제외하고 구한 평균 연봉 ·····31
[그림 3-4] 총 연봉 예상값 ·····33
[그림 3-5] 레코드 점수 SF_i ·····33
[그림 3-6] 이상치 예제 ·····36
[그림 3-7] 마할라노비스 거리 ·····39
[그림 3-8] 표준거리 산점도 행렬 ·····41
[그림 4-1] 중범죄 피해에 대한 로지스틱 추정 ·····63

Ⅰ. 데이터 에디팅 개요

1. 데이터 에디팅

1.1 데이터 에디팅의 정의

통상적으로 통계조사는 조사/보고 계획, 현장 조사/보고, 데이터 처리 및 에디팅, 데이터 집계/추정 및 분석 그리고 통계 데이터 보급의 과정을 거쳐 이루어진다. 조사통계에서는 현장 조사 및 데이터 추정 과정을 거치고, 보고통계에서는 현장 데이터 보고 및 데이터 집계 과정을 거친다.

통계조사에서 에디팅(editing)은 데이터를 대상으로 하기 때문에 데이터 에디팅(data editing)을 가리킨다. 데이터 에디팅이란 통계조사 과정에서 데이터의 논리적으로 일치성을 결여한 오류를 찾고 수정하는 활동을 말한다.

전통적으로 데이터 에디팅은 결측값이 포함된 원 데이터(raw data)를 정비하여 결측값이 없는 완비 데이터(complete data)를 만들기 위하여 수행하였다. 여기에 더하여 최근에는 데이터를 수집, 처리 과정에서 발견되는 데이터 오류(data error)의 양을 줄이고, 데이터 품질에 대한 정보를 제공하며, 향후 조사에 대한 개선 정보를 제공하는 역할이 강조되고 있다.

데이터 에디팅은 데이터의 정확성을 향상시키는 반면 에디팅 수행을 위한 비용과 인력을 필요로 한다. 데이터 오류를 수정하면 데이터 품질이 더 높아질 것으로 생각하기 때문에 가능한 많은 데이터 오류를 찾고 수정하는 것을 당연하게 생각할 수 있다. 그러나 지나친 데이터 에디팅은 통계의 시의성을 감소시키고, 에디팅으로 인한 새로운 오류를 초래할 수 있으며, 많은 비용과 시간을 필요로 한다. 따라서 데이터 품질 향상과 데이터 에디팅에 소요되는 시간과 비용을 고려하여 데이터 에디팅을 수행하여야 한다.

1.2 데이터 오류 탐색과 처리

데이터 에디팅은 데이터 오류를 탐색하는 과정과 탐색된 오류 데이터를 처리하는 과정으로 구분된다.

데이터 오류 탐색은 데이터가 결측값(missing value)인지, 타당하지 않은 값(invalid value)인지, 불일치하는 값(inconsistent value)인지 혹은 변칙적인 값(anomalous value)인지 점검하는 것이다. 결측값은 응답자가 응답을 하지 않아서 발생하는 값이므로 조사표나 데이터 파일에서는 결측값으로 표기되어 다른 응답값과 구분되어야 한다. 타당하지 않은 값은 응답으로서 허용 가능한 범위를 벗어나는 데이터 값을 말한다. 예를 들어, 숫자를 묻는 질문에 숫자가 아닌 값이 표기되어 있으면 타당한 값이 아니다. 또한 입력된 자료가 허용된 값의 범위를 벗어나면 타당한 응답이 아니다. 불일치하는 값은 사전에 정의된 데이터 항목의 관계를 만족하지 않을 때 발생한다. 변칙적인 값은 특이한 응답이나 이상치로 나타난다.

데이터 오류 탐색에는 조사별 전문 지식과 편집규칙(edit rule)을 활용한다. 점검규칙(check rule)이라고도 하는 편집규칙은 만일 데이터가 올바르다면 반드시 만족해야 하는 데이터 항목 값 사이의 논리적 조건이나 제약조건을 말한다. 데이

터 오류 탐색 후 데이터를 최종 데이터로 채택할 수 있는 데이터와 검토가 필요한 의심스러운 데이터로 구분한 후 표기한다.

오류 데이터를 처리하는 과정에서는 결측값, 타당하지 않은 값, 혹은 불일치하는 값으로 표기된 값을 제거하고 수정한다.

1.3 데이터 에디팅의 시기

통계조사에서 면접조사의 경우 데이터 처리 과정은 조사표 응답, 데이터 입력/코딩, 데이터 에디팅, 최종 데이터 파일 준비의 절차를 거친다. 컴퓨터 보조 면접조사(CAPI)나 컴퓨터 보조 전화조사(CPTI)에서는 조사표 응답과 데이터 입력/코딩이 동시에 이루어진다. 데이터 처리 과정에서는 가공되지 않은 데이터가 입력 저장된 후, 저장된 데이터를 대상으로 에디팅이 시작된다. 데이터 오류를 탐색하고, 검토 및 수정을 하여 분석을 위한 최종 데이터 파일이 만들어지면 데이터 에디팅은 끝이 난다.

2. 데이터 에디팅 방법

데이터 에디팅 방법은 에디팅 과정에 사람이 개입하는지 여부와 에디팅 대상이 무엇인지에 따라 분류될 수 있다. 에디팅 과정에 사람이 개입하는지 여부에 따른 분류로는 수동 에디팅(manual editing), 쌍방향 에디팅(interactive editing), 자동 에디팅(automated editing)이 있다. 에디팅 적용 대상에 따른 분류로는 마이크로에디팅(microediting)과 매크로에디팅(macroediting)이 있다.

2.1 쌍방향 에디팅

수동 에디팅은 수동으로 데이터 오류를 탐색하고, 탐색된 데이터 오류를 수동으로 처리하는 것을 말한다.

쌍방향 에디팅은 데이터 획득 후에 오류 탐색 및 처리를 사람이 컴퓨터의 도움을 받아 수동으로 진행하는 에디팅을 말한다. 데이터 오류를 확인하고 수정하기 위하여 응답자를 직접 재조사 할 수도 있고, 응답자의 보조 자료로부터 새로운 정보를 얻어내어 에디팅에 이용하는 방법이다.

쌍방향 에디팅은 사용 가능한 보조 정보가 충분하고 재조사가 가능할 때 효과적인 방법이다. 이러한 조건에서 쌍방향 에디팅은 데이터 오류를 탐색하고 처리하는 정확한 방법이 된다. 그러나 쌍방향 에디팅은 많은 비용과 시간을 필요로 하고, 재조사는 추가적인 응답자 부담을 유발하며, 사람에 의한 지나친 에디팅은 사람에 의한 편향과 변동, 심지어는 에디팅 담당자에 의한 새로운 오류가 개입될 수 있는 단점이 있다.

2.2 자동 에디팅

자동 에디팅은 컴퓨터 프로그램을 이용하여 저장된 데이터의 오류를 탐색하고 처리하는 방법이다. 편집규칙을 프로그램으로 만들어 오류 탐색 및 처리에 이용한다. 보조 자료를 편집규칙에 포함시켜 오류 확인 및 수정에 사용할 수 있다.

자동 에디팅의 장점으로는 첫째, 쌍방향 에디팅에 비하여 비용이 적게 들고, 주어진 시간에 많은 양의 에디팅을 수행할 수 있다. 둘째, 시의성을 향상시킨다. 적은 자원을 가지고 빠른 결과를 만들어 낼 수 있다. 셋째, 에디팅을 일관된 방법으로 수행할 수 있다. 넷째, 프로그램에 의한 에디팅을 수행하기 때문에 에디팅 방법 재생과 에디팅 과정 기록이 쉽다. 다섯째, 에디팅 처리 과정의 품질을 관리하고 감독하는 것을 수월하게 한다.

자동 에디팅의 단점으로는 첫째, 최적의 쌍방향 에디팅과 비교하여 정확성이

떨어진다. 둘째, 단계별로 자동 에디팅 사항이 구체적으로 명시되어야 한다. 셋째, 단계별로 컴퓨터 프로그램을 개발하는 것은 시간 낭비일 수 있다.

2.3 마이크로에디팅

마이크로에디팅은 조사표나 데이터 레코드 수준에서 수행된다. 전체 데이터로 계산한 집계치나 추정치를 참고하지 않고 개별 관측치를 조사하여 오류를 탐색하거나 처리한다. 조사표나 응답자 수준에서 응답의 타당성, 일치성 등을 점검할 수 있다.

2.4 매크로에디팅

매크로에디팅에서는 전체 데이터나 혹은 대부분의 데이터를 동시에 이용하여 데이터 오류를 탐색한다. 따라서 매크로에디팅은 데이터 대부분이 준비될 때 적용 가능하다. 오류 탐색은 주로 통계적인 모형을 이용하여 그래픽 방법으로 하거나 집계치 혹은 추정치를 이용한 수치 방법으로 한다.

그래픽 에디팅(graphic editing)에서는 상자 그림, 산점도, 히스토그램 등과 같은 그래픽 표현 방법을 이용하여 데이터 오류를 탐색한다. 그래픽 에디팅을 사용하면 변수들 간의 관계를 이해하기 쉽고, 변칙적인 데이터나 이상치를 쉽게 탐색할 수 있는 장점이 있다.

2.5 편집규칙

데이터 에디팅은 주로 편집규칙에 의해서 이루어진다. 편집규칙은 필수적 편집규칙(fatal edit 혹은 hard edit)과 의문 편집규칙(query edit 혹은 soft edit, statistical edit)이 있다.

필수적 편집규칙은 확실하게 오류를 찾아내는 편집규칙을 말한다. 필수적 편집규칙을 통하여 데이터가 사용 가능하기 위해서 반드시 수정해야 하는 잘못된 데이터 항목 값을 찾아낸다. 의문 편집규칙은 의심스러운 값을 탐색하는 편집규칙이다. 의심 편집규칙에 의하여 식별된 값은 잘못된 값일 수도 있지만 반드시 그렇지만은 않다. 대부분의 경우 추가조사 없이는 오류임을 확신하기 어렵다.

[예제 1-1] 필수적 편집규칙이 적용되는 예로는 다음의 경우를 들 수 있다.

- 응답자 식별 번호 오류
- '범위 밖의 나이'와 같은 중요 변수의 타당하지 않은 값
- '자녀의 나이보다 적은 부모의 나이'와 같은 비 정상적인 값 □

[예제 1-2] 의문 편집규칙이 적용되는 예는 다음과 같다.

- 과거 데이터와 비교하여 의심스럽게 높은 값
- 어느 응답자의 응답이 과장된 것으로 의심되는 값 □

3. 데이터 에디팅 제한조건

데이터 에디팅 수행에는 가용 자원(시간, 예산, 인원), 응답자 부담, 그리고 자료의 의도적 사용 등이 제약사항이 될 수 있다.

3.1 에디팅 자원

수동 에디팅 환경에서 에디팅은 노동집약적이다. 에디팅 업무에 다음 사항을 포함한다.

- 에디팅 기준을 설정하고 에디팅 오류가 발견되었을 때 해야 할 업무를 정하고 문서화한다.
- 에디팅 할 사람을 훈련한다.
- 에디팅 작업을 감독하고 점검할 작업 메커니즘을 설정한다.

자동 에디팅 환경에서는 초기 도입 및 과정 전개를 위한 시간, 비용, 그리고 자원이 막대할 수 있다. 에디팅 업무에 다음 사항을 포함한다.

- 에디팅 기준을 개발하고 문서화한다.
- 컴퓨터 프로그램을 작성하고 가용 소프트웨어를 채택한다.
- 컴퓨터 프로그램을 테스트 한다.

3.2 응답자 부담

결측값이나 오류 데이터를 처리하기 위하여 응답자를 사후적으로 재접촉하기 위해서는 다음 사항을 고려해야 한다.

- 응답자 부담이 추가로 발생한다.
- 에디팅에 추가 비용이 발생한다.
- 사후 접촉에서 응답자가 초기 응답을 정확하게 기억하지 못할 수도 있다.

3.3 자료의 의도적인 사용

에디팅에서 수행해야할 업무의 양은 상당부분 사용하는 데이터의 특성에 좌우

된다. 그리고 데이터의 특성에 따라 에디팅의 정도를 다르게 하여야 한다.

데이터 내에서 몇몇 항목은 다른 항목에 비하여 중요하기 때문에 더 많은 시간과 자원을 들이는 것이 바람직하고 몇몇 항목은 그렇지 못하다. 몇몇 데이터는 다른 데이터에 비하여 추정치에 기여하는 정도가 크기 때문에 주의를 기울여야 한다.

4. 데이터 에디팅 품질관리

데이터 에디팅은 저장된 데이터의 오류를 탐색하고 수정하여 데이터의 품질을 향상시키는 기능을 한다. 또한 에디팅 과정을 기록하여 문서화 하면 조사과정을 투명하게 하고, 각 조사 단계에서 효과적인 품질 향상을 이루는데 큰 도움이 된다. 그러나 응답자 재조사 등 에디팅을 수행하기 위해서는 시간과 자원을 투입해야 하므로 데이터 에디팅은 효과적으로 설계되고 수행되어야 한다.

데이터 에디팅은 데이터 품질 차원 중 정확성(accuracy), 시의성(timeliness), 일관성(coherence) 등에 영향을 미친다.

통계조사에서 데이터 에디팅을 수행하면 데이터 정확성은 향상된다. 데이터 에디팅은 응답오차와 처리오차를 제거하고 무응답 편향을 줄이는데 도움이 된다. 응답자 재조사는 조사표 용어 정의를 분명하게 하고 조사표 설계, 데이터 수집 도구를 개선하는데 도움을 준다. 이는 향후 조사에서 응답자 부담과 무응답률을 낮추는 계기가 된다. 반면, 편집자가 데이터 에디팅에 지나치게 개입하면 이로 인한 새로운 오류와 불확실성이 데이터에 포함되기도 한다.

데이터 에디팅은 통계 공표를 지연시킨다. 따라서 지나친 에디팅은 시의성에 부정적인 영향을 미친다.

데이터 에디팅은 일관성에 긍정적인 영향을 준다. 데이터 정보가 에디팅에 활용될 때 데이터의 일관성은 향상된다. 특히 행정 자료에서의 활용은 데이터 일관성 면에서 긍정적인 효과를 준다.

II. 데이터 에디팅 과정

1. 데이터 에디팅 설계

데이터 에디팅을 효과적으로 수행하기 위해서는 에디팅을 설계하고, 수립된 설계에 의하여 에디팅을 수행하여야 한다. 데이터 에디팅 설계 시 고려해야 할 사항은 다음과 같다.

- 데이터 에디팅은 전체 조사과정과 연관하여 전 조사과정의 일부분이 되도록 설계한다.
- 에디팅 단계별로 요구되는 품질, 예상되는 입력물과 출력물, 그리고 시작 시점과 끝 시점 등을 정한다.
- 에디팅에 소요되는 자원은 무응답을 포함한 영향 오류 처리에 많이 배정한다.
- 편집규칙은 이전 조사 분석에 기초하여 정한다.
- 편집규칙은 주제 분야 전문가와 상의하여 정한다.
- 편집규칙의 일치성을 유지하고 중복이 되지 않도록 하며, 과도한 편집규칙은 피한다.

- 에디팅 단계별로 기록해야 할 지표를 정하고 에디팅 과정 및 결과는 문서로 남긴다.
- 에디팅 전과 후의 데이터를 메타데이터와 함께 보존한다.

2. 데이터 에디팅 절차

데이터 에디팅은 데이터 수집 과정에서 수행하는 에디팅과 데이터 입력저장 후의 수행하는 에디팅으로 구분할 수 있다.

2.1 데이터 수집 과정 중의 에디팅

데이터 수집 과정 중의 에디팅은 주로 현장에서 이루어지는 에디팅을 말하며 주로 응답의 타당성과 일치성을 점검한다. 이 단계의 에디팅의 목적은 다음과 같다.

- 데이터 수집 수단을 향상시킬 필요가 있는지 점검한다.
- 데이터 수집에 더 많은 조사원 훈련이 필요한지 점검한다.
- 명백한 오차를 발견하여 즉각적으로 응답자에게 확인한다.
- 조사표를 정비한다.

데이터 수집 과정은 조사표 응답, 조사표 수합, 데이터 입력/코딩으로 구분하여 각 과정에서 수행하는 에디팅 활동을 살펴본다.

조사표 응답 시에 면접원이 에디팅을 수행한다. 이를 위해서는 사전에 응답에 대한 편집규칙을 정하여 면접원 지침서에 수록하여야 한다.

- 항목별로 일치성을 점검한다. 일치하지 않는 응답은 곧바로 응답자에게 확인을 한다.

- 결측값을 점검한다. 항목무응답(item nonresponse)은 다시 질문하여 가능한 항목 무응답을 줄인다. 무응답과 '해당 없음'은 구분하여 표기한다. 여과질문(filter question)이 있는 경우 구조적 결측(structural missing)과 일반 결측은 구분하여 표기한다.
- 응답 범위를 점검한다. 범위를 벗어나는 응답은 다시 확인한다.
- 의심스러운 응답은 다시 확인한다. 특히 단위를 잘못 기록하여 오류가 발생하지 않도록 한다.

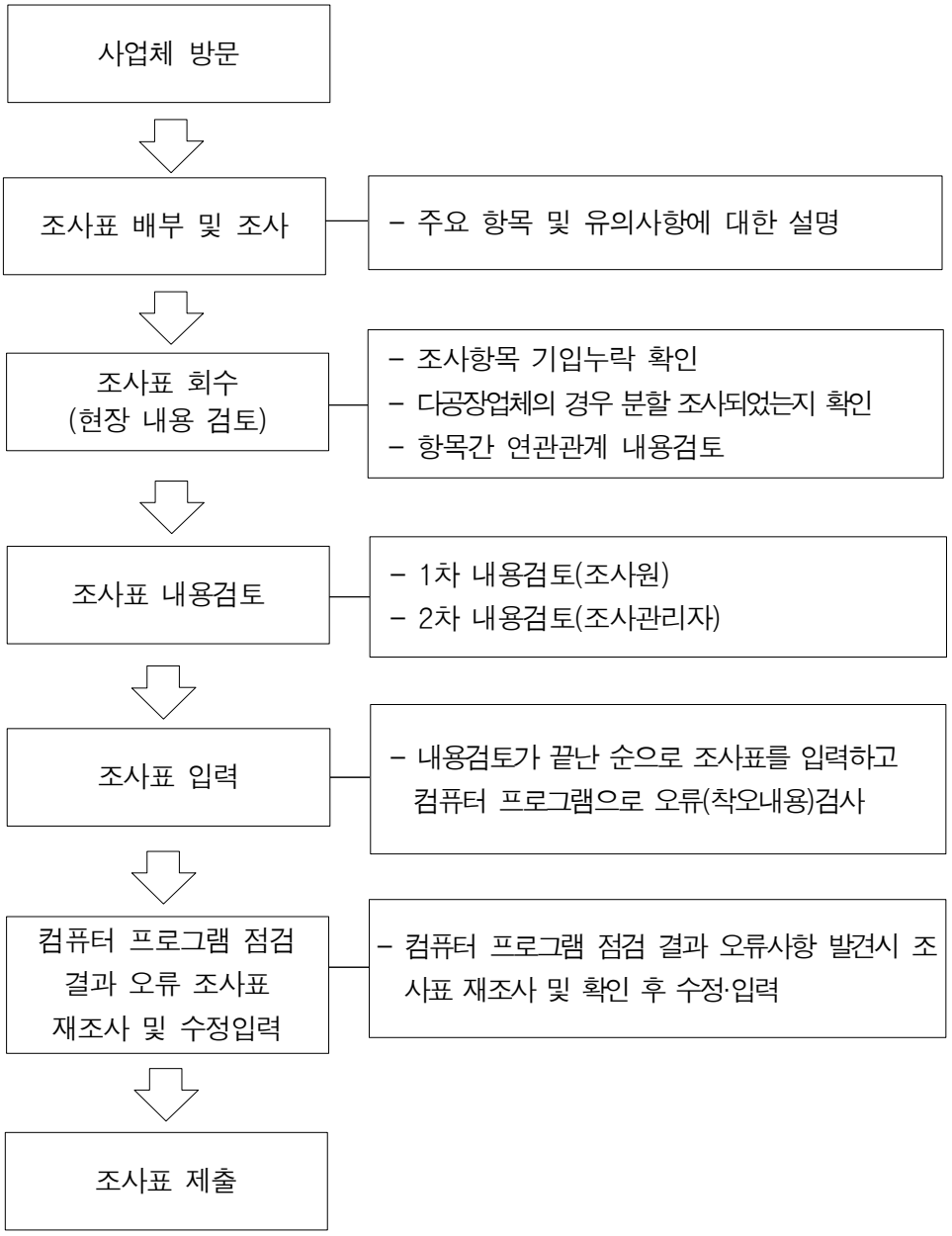
면접 중에 추가 질의는 면접원이 자발적으로 하도록 한다. 단, 응답 시간이 지나치게 길어지지 않도록 추가 질의는 중요한 항목으로 한정한다.

컴퓨터 보조 면접인 경우 편집규칙을 프로그램으로 만들어 컴퓨터에 내장한다. 일반 면접과는 달리 이전 조사 자료나 외부 자료를 내장하여 응답과 비교하면서 실시간으로 즉각적인 에디팅을 한다.

수작업으로 조사표를 작성한 경우 조사 완료 후 조사표를 수집하여 에디팅을 수행한다. 이 단계에서의 에디팅은 조사표들을 다음 단계로 계속 넘겨 처리할 것인지 여부를 결정하는 것이다. 조사표를 수용, 거부, 조치필요 등으로 분류하여, 수용할 조사표는 수용하고 응답 항목이 얼마 되지 않으면서 명백한 오류가 발견되는 조사표는 거부하고 처리를 중단하며, 추가 점검이 필요한 조사표는 분류하여 추가로 점검을 실시한다.

타자 입력이나 스캐닝 방법을 통하여 데이터를 입력하는 경우, 입력 데이터에 대한 편집규칙을 프로그램으로 만들어 레코드 단위로 실시간 에디팅을 적용한다. 레코드 항목별로 응답의 일치성, 응답 범위, 결측값 등을 점검한다. 편집규칙을 만족하지 않는 항목에 대해서는 입력 과정을 멈추고, 작업자가 특정한 조치를 취한 후에 자료 입력 저장을 계속한다. 특정한 조치는 항목 데이터를 확인하여 수정하거나, 향후 적절한 조치가 필요하다는 표기를 하는 것 등이다.

[예제 2-1] 통계청에서 수행하는 광업·제조업통계조사에서는 다음과 같은 절차를 거쳐 입력 에디팅을 수행한다.



□

2.2 데이터 입력저장 후의 에디팅

대부분의 에디팅은 이 단계에서 일어난다. 이 단계의 에디팅은 초기 에디팅, 영향 오류 식별 및 쌍방향 에디팅, 자동 에디팅, 매크로에디팅, 그리고 최종 데이터 파일 작성으로 세분된다.

초기 에디팅 단계에서는 체계적인 오류(systematic error)를 우선적으로 탐색하고 처리한다. 또한 명백한 오류를 찾아내어 제거하며 오류 데이터는 연역적 대체(deductive imputation)나 규칙기반 대체(rule based imputation)를 사용하여 처리한다.

초기 에디팅을 거친 레코드에 대하여 영향력 있는 오류(influential error)가 있는지 점검한다. 오류가 있을 것으로 의심되면서 추정치에 큰 영향을 주는 영향 데이터를 레코드 단위로 식별하며 영향 레코드로 식별되면 쌍방향 에디팅을 통하여 오류를 처리한다. 쌍방향 에디팅을 적용할 레코드 수를 줄이기 위하여 선택적 에디팅(selective editing)을 사용하고 영향 레코드(influential record)가 아닌 레코드는 자동 에디팅 처리를 한다. 자동 에디팅은 컴퓨터 프로그램을 활용하여 수행한다.

쌍방향 에디팅과 자동 에디팅을 거친 레코드 전체에 대하여 매크로에디팅을 적용한다. 매크로에디팅 단계에서는 이상치 탐색 등을 통하여 의심스러운 값들을 식별한다. 이상치가 오류일 필요는 없으나 만일 이상치가 오류라면 추정치에 큰 영향을 주기 때문에 이상치를 수정 혹은 제외한다. 매크로에디팅은 집계 수치나 추정치를 이용한 수치 방법을 사용하거나, 그림, 도표 등을 이용한 그래픽 방법을 사용한다. 의심스러운 값이 발견되면 초기 에디팅 단계로 되돌아간다. 의심스러운 집계나 추정치가 발견되지 않으면 최종 데이터 파일을 만든다.

[예제 2-2] 국민은행에서 수행하는 전국주택가격동향조사에서는 다음과 같은 에디팅 절차를 거쳐 통계를 작성한다.

• 입력에디팅 : 범위 점검 및 확인. 온라인 입력 단계에서 개별 응답 값이 입력 범위를 벗어나면 입력을 중지하고 데이터를 점검한다. 범위를 벗어나는 응답은 그 사유를 확인하고, 사유가 특이한 경우 재조사를 통하여 원인을 규명한다.

• 매크로에디팅 : 이상치 점검. 전주/ 전월 대비 지수 변동률이 이상치의 한계점을 벗어나는 경우 정당한 사유가 있는지 확인한다. 조사업소를 재조사 하거나 필요시 제3업소를 통하여 검증한다. 이상치 점검은 Hidiroglou & Berthelot 편집규칙을 이용한다.

• 출력에디팅 : 통계 결과 확인. 1차로 편집된 데이터를 이용하여 계산된 추정치를 점검한다. 개별 통계를 작성하는 과정에서 지나친 영향점을 발견하고 해당 응답에 대한 재조사를 통하여 원인을 규명한다. 원인을 규명한 후 사후층화, 기준값 수정, 가중값 조정 등의 절차를 거쳐 데이터를 검증하고 통계를 작성한다.

□

2.3 선택적 에디팅

에디팅에서 모든 레코드를 완벽하게 처리하기 위해서는 시간과 자원이 필요하다. 최종 추정에 별다른 영향을 미치지 않는 데이터 처리에 과도한 시간과 자원을 투입하는 것을 피하기 위하여 선택적 에디팅이 필요하다.

선택적 에디팅은 필수적 편집규칙만을 점검해야 한다는 생각에 바탕을 두고 있다. 따라서 선택적 에디팅은 반드시 점검해야 할 편집규칙 위주로 실시한다. 또한 조사 추정량에 미치는 잠재 영향력이 큰 레코드를 대상으로 선택적 에디팅을 실시한다.

선택적 에디팅은 비용을 절감하고, 영향력 있는 데이터의 오류를 수정함으로써 데이터 품질을 향상시킨다. 또한 에디팅 처리 기간을 줄이고, 응답자 재조사 수를 줄임으로써 응답자 부담을 줄이는 장점이 있다. 선택적 에디팅의 단점으로는 레코드 수준의 데이터 품질에 관심을 덜 기울이게 되고, 레코드 수준 데이터에 불일치하는 측면이 남게 되어 이용자로 하여금 데이터 품질이 낮다는 인상을 줄 수 있다는 점이다.

Ⅲ. 데이터 오류 탐색

1. 데이터 오류

데이터 오류는 관측 값과 실제 값이 다를 때 발생한다. 데이터 오류는 결측값, 이상치 등의 형태로 나타난다.

결측값은 응답자가 질문에 응답을 하지 않아 발생한다. 결측값은 응답자가 답을 알지 못하거나, 응답하고 싶지 않거나 또는 단순히 질문을 놓쳤을 경우와 같은 여러 가지 이유에 의해 발생한다. 이상치는 데이터 값의 통계 분포에서 꼬리 부분에 위치하는 데이터 값을 말한다. 데이터의 분포에서 이상치는 오류 데이터일 가능성이 크기 때문에 점검이 필요하다.

1.1 데이터 오류 영향

오류가 추정치에 미치는 영향에 따라 체계적인 오류(systematic error), 영향력 있는 오류(influential error), 랜덤 오류(random error)로 구분한다.

- 체계적인 오류 : 응답 레코드에서 일관되게 보고되는 오류이다. 체계적인

오류는 데이터에 체계적인 영향을 준다.

- 영향력 있는 오류 : 집계나 추정치에 영향을 미치는 오류이다. 오류의 영향은 집계하는 추정치에 따라 달라진다. 어떤 추정치에 영향이 큰 오류가 다른 추정치에는 영향이 작을 수도 있다.
- 랜덤 오류 : 체계적인 이유가 아닌 우연히 발생하는 오류이다. 데이터에 체계적인 영향을 주지는 않는다.

1.2 데이터 오류 탐색

데이터 오류는 편집규칙을 사용하여 탐색한다. 사전에 정의된 논리적, 수학적, 혹은 통계적 편집규칙에 따라 데이터를 채택 가능한 값과 채택하기 어려운 값으로 구분한다. 편집규칙을 만들 때 다음 요소들을 고려한다.

- 편집규칙은 응답자의 사회, 경제적 조건과 데이터 항목 간의 관계 이해 등 주제별 전문 지식에 기초하여 만든다. 주제 분야 전문가와 협력한다.
- 이전 조사 데이터의 활용이 가능한 경우 편집규칙은 이전 조사 데이터 분석에 기초하여 만든다.
- 편집규칙들은 서로 일치성이 있어야 하며, 중복을 피하여야 한다. 그리고 과도한 편집규칙은 피한다.

채택하기 어려운 값을 세분하여 결측값, 오류가 있는 값, 의심스러운 값으로 구분하여 표기한다.

2. 결측값

결측값은 응답자가 응답을 하지 않거나, 데이터 입력 과정에서 입력 실수로 응

답 데이터가 누락되어 파일에 저장되지 않을 때 발생한다. 즉, 값이 있어야 하는 항목에 값이 없으면 결측이 된다.

에디팅 과정에서 결측값은 편집규칙에 의한 일치성 점검에서 발견된다. 비록 결측값의 패턴은 복잡할지라도 결측값의 탐색은 보통 단순하다.

2.1 구조적 결측값

여과 질문(filter question)에 의하여 결측값의 특별한 경우가 발생한다. 여과 질문은 조사표에서 해당되는 문항을 선택하도록 하는 질문이다. 따라서 해당되지 않는 문항에는 응답을 해서는 안된다. 해당 없는 항목의 값은 반드시 응답을 해야 하는 값들과 구별하기 위하여 '구조적 결측값(structurally missing value)'이라고 한다.

2.2 결측값 표기

데이터에서 결측 항목은 공란으로 남겨 놓거나 일반 응답 값과 구분되는 값, 예를 들어 '9999' 등을 입력하고 '9999'는 결측값임을 기록한다.

일반 결측값과 구조적 결측값은 구별한다. 여과 질문에 의한 구조적 결측인지 단순한 항목무응답인지 구분하여 표기한다. 양적변수에서 '0'의 값과 무응답을 구분한다. 양적 변수의 무응답 항목에 '0'의 값을 표기하는 것은 바람직하지 않다. 무응답 '0'의 값을 응답으로 간주하여 추정치 계산을 하는 경우 추정치에 심각한 편향을 가져온다. 양적변수의 무응답은 '0'이 아닌 다른 무응답 표기를 한다.

2.3 단위무응답과 항목무응답

조사표에 응답자가 전혀 응답을 하지 않은 경우를 '단위무응답(unit non

-response)'이라고 한다. 단위무응답이 발생하는 대표적인 원인은 다음과 같다.

- 조사대상자를 접촉하지 못했을 경우
- 조사대상자가 응답을 거부 하였을 경우

응답자가 몇 개의 항목에는 응답을 하고 나머지 항목에 응답을 하지 않았을 때 발생한 결측값을 '항목무응답(item nonresponse)'이라고 한다. 항목무응답은 다음의 여러 가지 이유에 의하여 발생한다.

- 응답자가 답을 알지 못할 경우
- 응답자가 일부 항목에 응답하고 싶어하지 않을 경우
- 응답자가 응답 도중 단순히 질문을 놓쳤을 경우

[예제 3-1] 범죄 피해 관련 조사

12명의 응답자로부터 범죄피해 관련 항목을 조사하였다. 성별에서 'M'은 남성, 'F'는 여성을 나타내고, 범죄피해 항목에서 '1=범죄에 피해를 본 적이 있음', '0=없음'을 뜻한다. 무응답은 '?'로 표시하였고, 구조적 결측은 '??'로 표시하였다.

[데이터 3-1] 범죄 피해 관련 조사

응답자	나이	성별	교육 기간	범죄 피해	범죄피해장소
1	47	M	16	0	??
2	45	F	?	1	A
3	19	M	11	0	??
4	21	F	?	1	B
5	24	M	12	1	B
6	41	F	?	0	??
7	36	M	20	1	A
8	?	?	?	?	?
9	?	?	?	?	?
10	17	M	10	1	?
11	53	F	12	0	??
12	21	F	12	0	??

12명 응답자 중 8번과 9번은 전 항목에 대하여 응답을 하지 않았으므로 8번과 9번은 '단위무응답'이 된다. 반면, 2번, 4번, 6번 응답자는 교육 기간을 응답하지 않았고, 10번 응답자는 범죄 피해 장소를 응답하지 않았다. 이 경우는 '항목무응답'이 된다.

범죄에 피해를 당한 적이 있는 응답자를 대상으로 한 세부 항목에서, 범죄 피해가 없는 1번, 3번, 6번, 11번, 12번 응답자는 범죄피해 장소 항목에 응답을 하지 말아야 한다. 이러한 무응답이 '구조적 결측'이다. □

2.4 무응답 기록

무응답에 대한 지표를 기록한다. 적어도 관측값 수, 결측값 수, 그리고 구조적 결측값의 수 등을 기록한다. 무응답 지표를 이용하여 무응답 메커니즘에 대한 정보를 얻는다.

결측이 많은 변수와 레코드를 검토한다. 결측이 지나치게 많은 변수나 레코드는 수정하는 것 보다 제거하는 것이 더 나을 수도 있다.

[예제 3-2] 범죄 피해 관련 조사의 데이터에 대한 무응답 지표

[데이터 3-1]에서 나타난 무응답 지표는 다음과 같다.

- 표본수 = 12
- 단위무응답률 = (단위무응답 수)/표본수 = 2/12= 16.7%
- 항목무응답률 = (항목무응답 수)/표본수
  - 나이 : 2/12 = 16.7%
  - 성별 : 2/12 = 16.7%
  - 교육기간 : 5/12 = 41.7%

- 범죄피해 : 2/12 = 16.7%

• 범죄피해 장소는 범죄피해 응답이 1인 응답자를 대상으로 한다.

- 범죄피해 장소 : 1/5 = 20% □

2.5 무응답 영향

결측값은 잠재적으로 편향된 결과를 초래하여 데이터 품질을 떨어뜨린다.

편향은 응답자와 무응답자가 조사 항목에 대하여 각기 다른 특성을 가질 때 발생한다. 이 경우 응답자의 결과는 무응답자를 대표하기 어렵다.

[예제 3-3] 무응답 편향

어떤 면접조사에서 응답률이 80%라고 하고 응답자 평균을 $\bar{y}_R$이라고 하자.

[데이터 3-2] 면접 조사

	비율		모집단 평균	
응답자	p_R	80%	$\overline{Y}_R$	150
무응답자	p_{NR}	20%	$\overline{Y}_{NR}$	170
전체		100%	$\overline{Y}_N$	154

응답 평균이 비편향 추정량이라고 할 때, 응답 평균의 편향과 상대편향은 다음과 같다.

• 편향 : $E(\bar{y}_R)-\overline{Y}= \overline{Y}_R-[p\overline{Y}_R+(1-p)\overline{Y}_{NR}]$

$$= (1-p)[\overline{Y}_R-\overline{Y}_{NR}]$$

$$= (1-0.8)[150-170] = -4$$

• 상대 편향 : $\frac{E(\overline{y}_R)-\overline{Y}}{\overline{Y}} = (1-p)\frac{\overline{Y}_R-\overline{Y}_{NR}}{\overline{Y}} = -\frac{4}{154} = -0.026$

결과적으로 20%의 무응답으로 인하여 -2.6%의 상대편향이 발생하였다. □

2.6 무응답 처리

단위무응답과 항목무응답을 사후적으로 처리하는 주요 방법은 다음과 같다.

• 단위무응답 : 응답자 가중치를 조정하는 가중치 조정법으로 처리한다.

• 항목무응답 : 대체 방법에 의하여 처리한다.

[예제 3-4] 가중치 조정법

층1에서는 3명의 조사대상자 중 2명이 응답을 하고, 층2에서는 4명의 조사대상자 중 3명이 응답을 하였다고 하자. 응답자에 대한 가중치를 조정하여 조정된 가중치를 얻었다.

[데이터 3-3] 가중치 데이터

층 번호 (h)	표본번호 (hi)	원 가중치 (w_{hi})	응답 표기 (r_{hi})	응답값 (y_{hi})	조정된 가중치 (w_{hi}^*)
1	1	30	1	120	$30/60=50$
	2	40	0	?	0
	3	30	1	150	$30/60=50$
2	1	20	1	210	$20/70=28.6$
	2	20	1	220	$20/70=28.6$
	3	30	0	?	0
	4	30	1	320	$30/70=42.9$

모평균 추정: 수정된 가중치를 이용하여 가중평균으로 모평균을 추정한다.

$$\bar{y}_w = \frac{\sum w_{hi}^* y_{hi}}{\sum w_{hi}^*}$$

$$= \frac{50\times120+50\times150+28.6\times210+28.6\times220+42.9\times320}{50+50+28.6+28.6+42.9}$$

$$= \frac{39,526}{200.1}$$

$$= 197.5$$ □

3. 체계적 오류

3.1 체계적 오류의 정의

체계적 오류는 특정 응답항목에 대하여 모든 레코드에서 일관되게 나타나는 오류이다. 체계적 오류는 집계나 추정에서 편향을 발생시키기 때문에 통계 결과에 심각한 영향을 준다. 체계적 오류는 랜덤 오류를 탐색하기 전에 탐색되고 처리되어야 한다.

체계적 오류의 예로는 다음과 같은 경우가 있다.

- 사전에 정해진 용어 정의나 분류에 기초하여 응답을 하여야 하는데 이를 충분히 이해하지 못하고 시종일관 잘못 응답하는 경우
- 코딩 과정에서 응답을 오역하여 잘못 입력하는 경우
- 조사표에서 여과 질문과 관련한 통과 규칙을 잘못 이해하여 응답하는 경우
- 부호 오류 : 수익과 같은 음수를 가질 수 있는 변수에 대하여 음수 기호를 생략하고 응답하는 경우

- 단위측정오류 : 응답자가 잘못된 단위로 측정된 값을 보고하는 경우
- 데이터 입력 과정에서 시스템에 의한 오류

3.2 단위측정오류

총 매출액을 백만원 단위로 보고하여야 한다고 하자. 그런데 매출액을 잘못 보고하여 만원 단위로 보고하였다고 하자. 이러한 경우에 발생하는 오류를 단위측정오류라고 한다.

단위측정오류는 일치성 편집규칙에 의해서는 발견되지 않는다. 예를 들어 총 매출액에 균형편집규칙을 적용하였을 때 모든 세부 항목들이 만원 단위로 조사되었다면 균형편집규칙으로는 이러한 오류를 발견할 수 없다.

3.3 체계적 오류의 탐색

체계적 오류를 발견하기 위해서는 예상되는 체계적 오류의 종류가 무엇인지, 내재되어 있는 오류생성 메커니즘이 무엇인지를 알고 있어야 한다.

예를 들어, 단위측정오류와 같은 특별한 체계적 오류 메커니즘을 염두에 두면 범위 점검이나 비율 점검 등을 통하여 체계적 오류를 찾아낼 수 있다. 질문 항목에 대한 특별한 지식이 없는 경우, 대체로 체계적 오류를 찾아내기 어렵다. 만일 예상되는 체계적 오류에 대한 사전 이해가 없을 때에는 범위 점검(range test), 비율 점검(ratio test) 방법을 사용하여 체계적 오류를 탐색한다.

응답 값의 범위를 사전에 정한 후 응답이 범위를 벗어나면 데이터 항목을 재점검한다. 범위 점검 방법은 범주형 변수와 수치형 변수 모두에 적용할 수 있다.

[예제 3-5] 범위 점검

근로자의 성별을 묻는 질문에서 응답은 0='무응답', 1='남성', 2='여성' 중의 하나로 받는다고 하자. 따라서 응답 집합 {0, 1, 2}에 포함되지 않는 응답은 오류로 간주한다.

'월'로 응답을 해야 하는 질문에 대하여 응답 집합을 {01, 02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12}로 하였다고 하자. 이 경우 응답 집합을 벗어나는 응답은 오류로 간주하고 데이터를 재점검한다. □

비율 점검을 위한 편집규칙에 포함되는 변수는 두 개이다. 두 변수를 x, y라고 할 때 두 변수의 비율에 대한 상한(U)과 하한(L)을 사전에 지정하고, 비율 점검 결과 상한과 하한을 벗어나는 레코드는 오류가 있는 것으로 간주한다.

비율 점검 범위 : $L < \frac{y}{x} < U$

만일 두 변수에 대하여 단위측정오류 같은 응답 오류나 데이터 입력 오류가 데이터에 포함되면 y/x 비율에 즉각적인 영향을 미치게 되고, 그 값은 상한 값(U) 혹은 하한 값(L)을 벗어나기도 한다. 이러한 경우 데이터 재점검을 통하여 오류를 발견할 수 있다. 비율 점검은 두 변수의 그래픽 분석과 함께 수행하면 매우 유용하다. 비율 점검 방법은 수치형 변수에 적용 가능하다.

[예제 3-6] 비율점검

25개 회사를 대상으로 각 회사의 1년 동안 평균 근로자 수(x_i, 단위 명)와 1년 총 임금(y_i, 단위 억원)를 조사하였다고 하자([데이터 3-4]). 그러면 각 회사에서 근로자 1명의 평균 연봉은 $(y/x)\times 100$ (단위 백만원)으로 구할 수 있다.

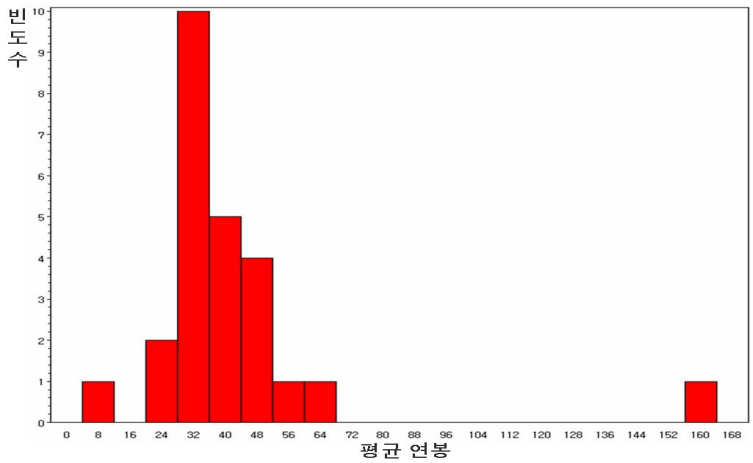
각 회사의 평균 연봉은 비율 점검을 통하여 점검할 수 있다.

$$L < \frac{y_i}{x_i} < U,\ i = 1,...,n$$

근로자의 평균 연봉의 연봉 상한(U)과 하한(L)을 고려하자. 예를 들어, 연봉 상한을 $U = 100$(10억원)으로 하고, 하한을 $L = 10$(천만원)으로 하자. 그러면 평균 연봉의 상한과 하한을 벗어나는 회사는 5번 회사(6.91)와 21번 회사(158.55)이다. 따라서 5번 사업체와 21번 사업체는 사후적으로 재점검 한다.

[데이터 3-4] 근로자 평균 연봉

회사 번호 (i)	가중치 (w_i)	근로자 수 (x_i)	총임금 (y_i)	평균연봉 $(y_i/x_i)\times 100$
1	100	49.4	15.47	31.32
2	100	41.3	14.63	35.42
3	300	31.7	12.81	40.41
4	300	32.4	12.81	39.54
5	400	78.0	5.39	6.91
6	400	72.9	22.40	30.73
7	150	38.3	17.29	45.14
8	150	75.0	24.50	32.67
9	150	45.6	20.93	45.90
10	150	90.7	25.90	28.56
11	130	41.2	16.80	40.78
12	130	48.0	13.58	28.29
13	130	32.0	8.54	26.69
14	130	46.2	16.17	35.00
15	130	22.3	10.50	47.09
16	200	48.1	18.76	39.00
17	200	51.8	16.03	30.95
18	200	28.2	14.07	49.89
19	200	41.8	13.72	32.82
20	200	75.6	20.79	27.50
21	160	8.3	13.16	158.55
22	160	46.5	19.67	42.30
23	160	47.9	15.19	31.71
24	130	134.0	83.02	61.96
25	130	149.0	84.42	56.66



[그림 3-1] 회사별 근로자 평균 연봉 막대그래프

□

3.4 체계적 오류의 탐색 및 처리

만일 체계적 오류가 범위 점검이나 비율 점검 등에 의하여 발견되면, 체계적 오류 메커니즘을 발견하기 위한 편집규칙을 추가한다. 그리고 만일 체계적 오류 메커니즘이 발견되면, 조사표, 면접관 교육, 코딩, 프로세싱 등 조사과정을 개선하여 유사한 오류를 예방한다.

체계적 오류 탐색은 지도가 필요한 조사원을 진단하는 방법이기도 하다. 즉, 체계적 오류를 많이 발생하는 조사원은 교육이 필요하다.

4. 영향력 있는 오류

4.1 영향력 있는 오류의 정의

영향력 있는 오류란 공표하는 통계에 유의한 영향을 주는 오류이다. 영향력 있는 오류는 통계 결과에 큰 영향을 미치는 영향 관측치와 연관이 있다. 영향 관측치는 올바른 값일 수도 있고, 아닐 수도 있다. 후자의 경우 영향 관측치는 영향력 있는 오류를 초래하게 된다.

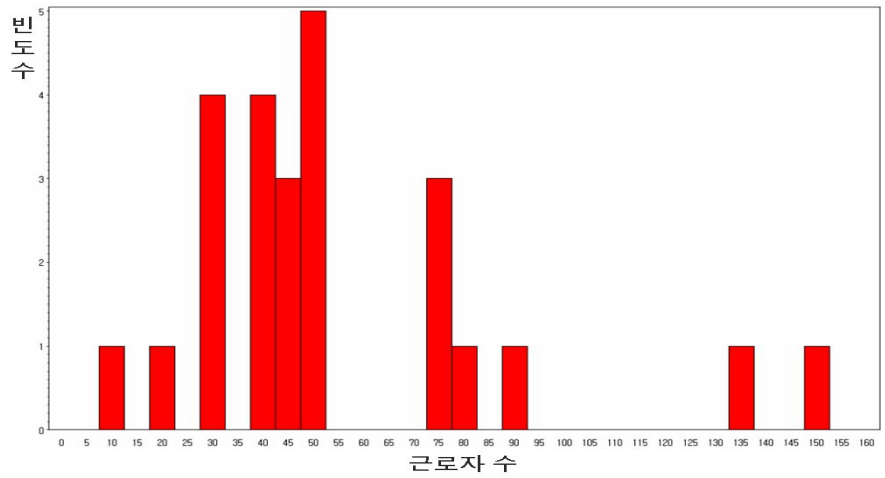
예를 들어 사업체 조사에서 소수의 대형 사업체는 영향력 있는 사업체이다. 대형 사업체는 근로자 수나 매출액 등이 다른 사업체이 비하여 월등히 크기 때문에 통계 결과에 큰 영향을 미친다. 큰 가중치가 부여된 사업체도 영향력 있는 사업체이다. 비록 이들 사업체는 크지 않다고 하더라도 큰 가중치로 인하여 추정치에 기여하는 바가 크기 때문이다.

[예제 3-7] 근로자 수 추정

[데이터 3-4]에서 평균 근로자 수를 추정한다고 하자. 평균 근로자 수는 가중치를 이용하여 가중평균으로 구한다.

- 근로자 수 : $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$
- 25개 회사 근로자 수의 가중평균 : 55.1명
- 24번, 25번 회사를 제외한 근로자 수 가중평균 : 50.0명
- 24번과 25번 회사의 근로자 수 가중평균 : 141.5명

24번과 25번 회사가 근로자 수 평균에 미치는 영향이 크므로 24번, 25번 회사가 모평균 추정에 영향을 미치는 관측치이다. 24번, 25번 회사의 근로자 수는 올바른 값일 수도 있고, 오류가 포함된 값일 수도 있다. 쌍방향 방법으로 오류 포함 여부를 점검한다.



[그림 3-2] 근로자 수 막대그래프

□

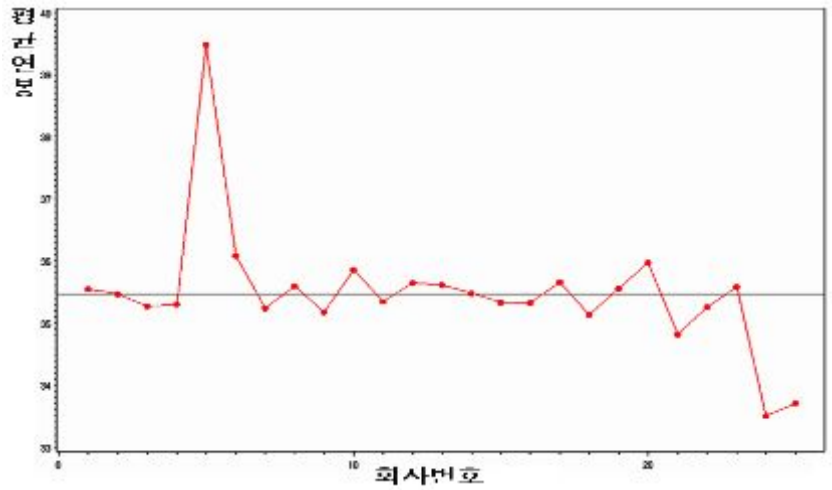
[예제 3-8] 근로자 평균 임금 추정

[데이터 3-4]에서 근로자의 평균 연봉을 추정한다고 하자. 근로자의 평균 연봉은 가중치를 이용한 가중평균으로 구한다.

- 근로자의 평균 연봉(단위 백만원) : $\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i x_i}$
- 25개 회사 근로자의 평균 연봉 : 35.4
- 5번 회사를 제외한 평균 연봉 : 39.5
- 24번 회사를 제외한 평균 연봉 : 33.5
- 25번 회사를 제외한 평균 연봉 : 33.7

평균 연봉 추정에 영향을 많이 미치는 회사는 5번, 24번, 25번 회사이다. 5번 회사는 연봉이 적어서 영향을 많이 주며, 24번, 25번 회사는 연봉이 많아서 영향을 많이 준다.

아래 그림은 해당 회사를 제외하고 구한 평균 연봉 산점도이다. 예를 들어 5번 회사를 제외하고 평균 연봉을 구하면 39.5이다. 전체 평균은 35.4(실선)이다.



[그림 3-3] 해당 회사를 제외하고 구한 평균 연봉

□

4.2 영향력 있는 오류 탐색

영향력 있는 오류 탐색은 일반적으로 선택적 에디팅을 사용한다. 선택적 에디팅은 가장 중요한 추정치에 초점을 맞추어 오류가 있을 것으로 여겨지는 레코드만 점검하는 방법으로 에디팅 자원을 효율적으로 사용하는 방법이다.

선택적 에디팅에 의하여 잠재적으로 영향력 있는 오류들을 확인하고, 쌍방향 처리 방법으로 이러한 오류들을 처리한다.

4.3 선택적 에디팅 절차

원 데이터를 잠재적으로 오류 가능성이 있는 부분과 그렇지 않은 부분으로 나눈다. 선택적 에디팅은 잠재적으로 오류 가능성이 있는 부분에 속하는 레코드를

대상으로 한다.

- M : 잠재적으로 오류가 있을 것으로 예상되는 레코드 집합
- $\widetilde{M}$: 그렇지 않은 레코드 집합

선택적 에디팅을 적용할 변수의 우선순위는 오류의 중대성에 따르거나 보고 단위 또는 변수들의 중요성에 따라 정한다. 레코드 집합 M에 속하는 레코드에 대하여 레코드 점수를 부여한다. 레코드 점수는 모형에 기초하거나 혹은 이전 시점 데이터나 외부 데이터에 기초하여 구한 참값의 예상값과 관측치의 차이로 정의한다. 예를 들어 M에 속한 레코드에 대하여 레코드 점수(SF)는 관측값 y_i과 예상값 $\tilde{y}_i$의 차이에 가중치 $w_i/\sum w_i$ 를 곱하여 정의할 수 있다.

$$SF_i = \frac{w_i|y_i - \tilde{y}_i|}{\sum_{i \in M} w_i},\ i \in M$$

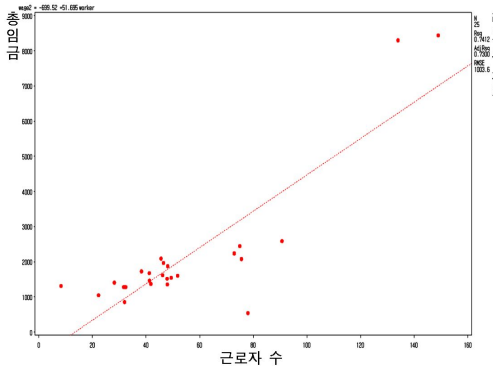
레코드 점수 SF_i가 큰 레코드 i는 영향력이 큰 레코드이다. 따라서 SF_i값이 큰 레코드를 선별하여 쌍방향 방법으로 오류를 확인하고 처리한다.

[예제 3-9] 선택적 에디팅 절차

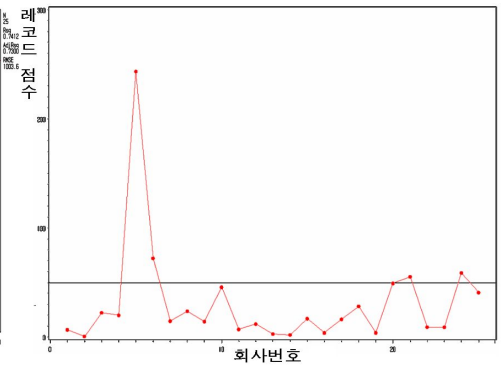
[데이터 3-4]에서 회사의 근로자 평균 연봉을 추정한다고 하자.

총 임금을 종속변수로 하고, 근로자 수를 설명변수로 하여 회귀분석을 실시한 후, 회사별로 총임금 예상값($\tilde{y}_i$)을 구한다.(<그림 3.3>)

그리고 가중치를 곱하여 레코드 점수를 구한다.(<그림 3.4>)



[그림 3-4] 총 연봉 예상값



[그림 3-5] 레코드 점수 SF_i

레코드 점수가 50이 넘는 레코드는 다음의 4개이다. 4개 레코드에 대하여 쌍방향 방법으로 오류를 확인하고 처리한다.

<표 3-1> 오류 처리 결과

사업체 번호 (i)	가중치 (w_i)	근로자 수 (x_i)	총임금 (y_i)	예측값 $(\tilde{y}_i)$	레코드 점수 SF_i
5	400	78.0	539	3,332.7	244.5
6	400	72.9	2,240	3,069.1	72.3
24	130	134.0	8,302	6,227.7	58.8
21	160	8.3	1,316	-270.5	55.3

□

4.4 선택적 에디팅의 대안

영향력 있는 오류를 탐색할 때 선택적 에디팅에 대한 대안은 편집규칙을 사용하는 것이다. 이 방법은 영향력 있는 오류가 편집규칙을 어긴다는 가정에서 출발한다.

이 방법은 3단계로 구성된다.

- 편집규칙을 통과하지 못한 레코드를 선택한다.

- 편집규칙을 만족하지 않는 변수들에 대하여 편집 제약조건을 만족하는 레코드로 만드는데 필요한 변경의 양을 추정한다.
- 필요한 변경의 양을 기초로 검토할 레코드에 우선순위를 부여한다.

4.5 선택적 에디팅 효과

선택적 에디팅의 이점은 적시성과 효율성이다. 적은 수의 레코드를 쌍방향으로 처리할 수 있기 때문이다. 그러나 중요한 변수가 많은 경우에는 마이크로에디팅 단계에서 영향력 있는 오류가 있는 모든 레코드를 탐색하기는 어렵다. 점수함수(레코드 점수)에 의해 탐색되지 않은 영향력 있는 오류들이 매크로에디팅 단계에서 탐색될 수도 있다.

4.6 매크로에디팅

매크로에디팅은 영향력 있는 오류를 식별하기 위하여 에디팅의 마지막 부분에 수행한다. 주요한 공표 총계, 공표 단위, 공표 부차 모집단을 매크로에디팅에서 고려한다.

매크로에디팅 절차는 다음과 같다.

- 추정치를 이전 시점의 추정치 또는 비교 가능한 변수를 가진 다른 자료의 결과와 비교한다. 비교할 추정치로는 평균, 총계, 비율, 상관계수, 분위수, 분산 등을 고려할 수 있다.
- 그래픽 기법을 이용한 탐색적 자료분석을 한다. 산점도, 산점도 행렬, 고차원 그래픽 기법이 이상치를 발견하는 데 주로 사용된다.
- 추정치가 사전에 정의된 채택 영역을 벗어나는 경우 의심스러운 값으로 선택한다. 의심스러운 추정치에 속하는 모든 레코드는 레코드별로 점검한다.

- 레코드 점검은 매크로 편집규칙을 만족하거나 혹은 추정치가 채택 영역에 포함된다고 판단될 때 정지한다.
- 매크로에디팅은 수용 가능한 수준에서 끝내는 것이 중요하다. 매크로에디팅은 현 데이터를 과거 데이터로 너무 접근해가는 과도 수정의 위험이 있기 때문이다.
- 매크로에디팅은 통계적 기술, 주제 분야 지식과 통계조사의 각 단계와 에디팅 과정에 대한 정보를 필요로 한다. 종종 값의 비교를 위하여 전문가의 조언도 필요하다.
- 추정치, 비교, 그래프, 예외, 이상치 등을 적절하게 기록하여 문서로 남긴다.

5. 이상치

5.1 이상치 정의

이상치는 대다수 데이터에서 멀리 떨어진 관측치를 말한다. 이상치는 영향력 있는 관측치와 직접적으로 관련되어 있다. 종종 영향력 있는 관측치는 이상치로 나타난다.

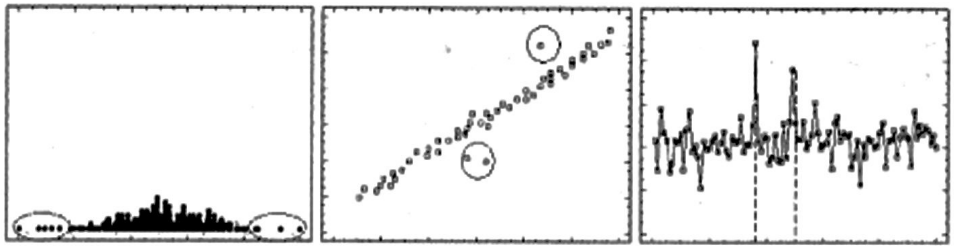
많은 통계량에 대하여 모든 영향력 있는 관측치를 개별적으로 확인하는 것은 불가능하므로 보통 매크로에디팅 단계에서 이상치를 탐색한다. 발견된 이상치는 오류를 포함할 가능성이 높으므로, 이상치가 발견되면 오류를 확인하고 처리한다.

[예제 3-10] 이상치 예제

(a) 분포 이상치 : 일변량 데이터에서 양 끝에 있는 값을 이상치로 간주한다.

(b) 패턴 이상치 : 이변량 데이터에서 직선 부근에서 벗어난 값을 이상치로 간주한다.

(c) 시계열 이상치 : 시계열 데이터에서 다른 데이터에 비하여 상대적으로 크거나 값은 값을 이상치로 간주하다.



[그림 3-6] 이상치 예제

□

5.2 일변량 이상치 탐색 방법

관측치가 평균부터의 거리가 일정 범위를 벗어나면 이상치라고 한다. 크기가 n인 표본 데이터를 고려하고 데이터의 표본 평균을 $\bar{y}$, 표준편차를 s라고 하자. 그리고 관측치로부터 평균까지의 표준거리를 d로 나타내자.

- 표본평균 : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- 표본 표준편차 : $s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}$
- 표준거리 : $d_i = \left|\frac{y_i - \bar{y}}{s}\right|,\ i = 1,...,n$

만일 데이터의 표준거리가 3보다 크면 이상치로 판정한다.

[예제 3-11] 일변량 이상치 탐색

9개 표본 데이터를 고려하자. 표본 평균은 $\bar{y}=10.78$이고, 표본 표준편차는 $s=2.223$이다. 9개 데이터에 대하여 표준거리를 구하면 아래와 같다.

[데이터 3-5] 일변량 표본 데이터

표본번호	1	2	3	4	5	6	7	8	9
관 측 값	11.74	11.46	9.72	9.98	9.60	10.02	15.89	7.90	10.71
표준거리	0.431	0.305	0.477	0.360	0.531	0.342	2.299	1.296	0.031

표준거리가 모두 3보다 작으므로 9개 데이터 중 이상치는 없는 것으로 판정한다. □

표본평균과 표본 표준편차는 이상치에 민감하게 반응한다. 따라서 이상치에 덜 민감한 로버스트 통계량을 사용하여 이상치를 탐색하기도 한다.

로버스트한 위치 통계량으로는 중위수가 종종 이용되고, 로버스트 산포 추정량으로는 절대편차의 중위수나 사분위 간 범위 등이 이용된다.

자료를 크기순으로 정렬하여 4등분한다. Q_1, Q_2, Q_3를 사분위수라고 하자.

- Q_1 : 하위 25%에 해당하는 값
- Q_2 : 상하위 50%에 해당하는 값. 중위수
- Q_3 : 상위 25%에 해당하는 값

하한 사분위 범위와 상한 사분위 범위를 구한다.

- 하한 사분위 범위 : $h_L = Q_2 - Q_1$
- 상한 사분위 범위 : $h_U = Q_3 - Q_2$

과거의 자료나 경험에서 구한 계수 c_L과 c_U를 이용하여 허용오차의 범위를 구한다.

$$(Q_2 - c_L \times h_L,\ Q_2 + c_U \times h_U)$$

만일 관측치가 이 범위를 벗어나면 이상치로 간주한다.

5.3 다변량 이상치 탐색 방법

변수가 k개인 다변량 데이터 $(y_1,...,y_k)$를 고려하자. 그리고 벡터 m을 평균벡터, 행렬 C를 공분산 행렬이라고 하자. 다변량 데이터의 이상치 탐색에서는 관측치 y로부터 중심 m까지의 거리로 Mahalanobis 거리 d^2을 사용한다.

$$d^2 = (y-m)^T C^{-1}(y-m)$$

- 변수별로 일변량 표준거리를 구하여 이상치를 탐색한다.
- 이변량 산점도를 구하여 이상치를 점검한다.
- Mahalanobis 거리를 구하여 다변량 이상치를 탐색한다. 임계값으로는 $\chi_p^2(0.005)$를 사용한다. 여기에서 p는 변수의 수이며, $\chi_p^2(0.005)$는 자유도가 p인 카이제곱 분포에서 상위 0.5%에 해당하는 값을 뜻한다. 예를 들어 자유도가 4인 경우는 $\chi_4^2(0.005) = 14.86$이다.

[예제 3-12] 다변량 이상치 탐색

20명의 학생을 대상으로 팔 길이를 4회에 걸쳐 측정하였다. 측정 시 학생의 나이는 8살(y_1), 8살 6개월(y_2), 9살(y_3), 9살 6개월(y_4)이다 ([데이터 3-6]).

변수 4개에 대하여 일변량 표준거리 d_1, d_2, d_3, d_4를 구한다.

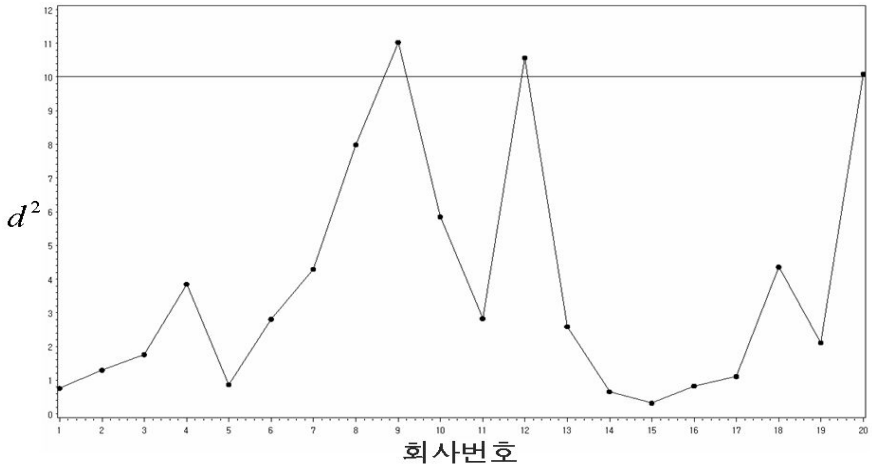
$$d_{ij} = \left| \frac{y_{ij} - \bar{y}_j}{s_j} \right|, \ i = 1,...,20, \ j = 1,...,4$$

일변량 표준거리가 4변수 모두 3보다 작다. 이상치가 발견되지 않는다. 표준거리를 이용하여 두 변수씩 쌍으로 산점도를 그린다. 아래 6개의 산점도에서 이상치는 발견되지 않는다.

Mahalanobis 거리 d^2를 구한다.

$$d_j^{\,2} = (y_i - m_i)^T C^{-1} (y_i - m_i)$$

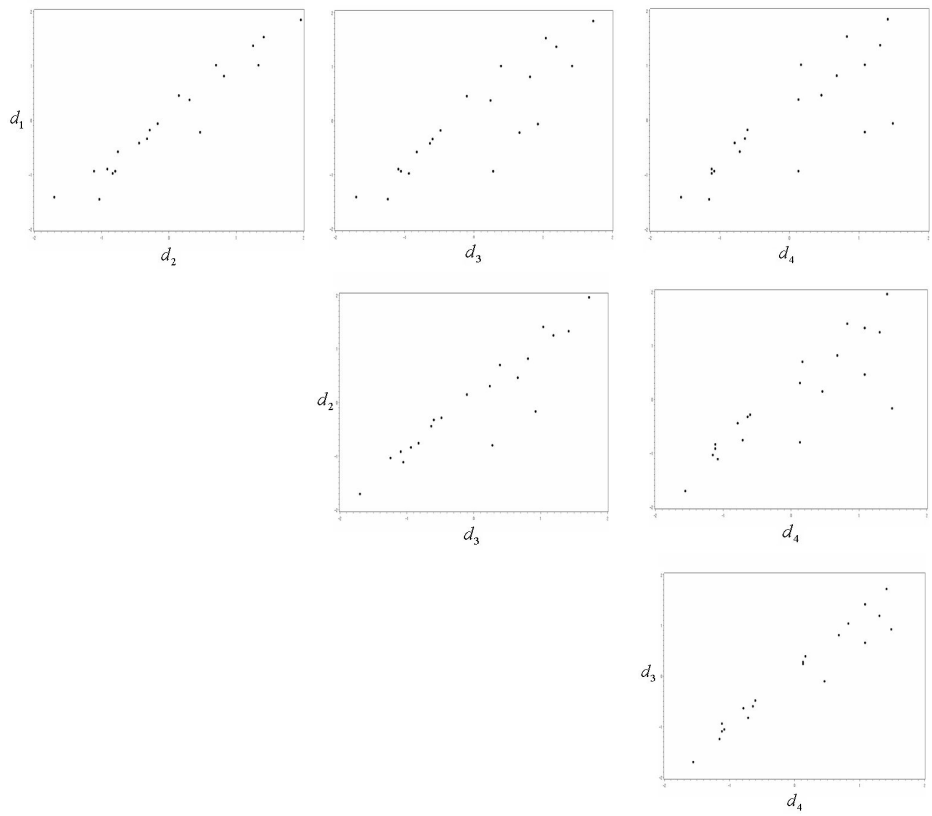
여기에서 $y_i^T = (y_{i1}, y_{i2}, y_{i3,} y_{i4})$, $m_i^T = (m_{i1}, m_{i2}, m_{i3,} m_{i4})$이다.



[그림 3-7] 마할라노비스 거리

[데이터 3-6] 학생의 팔 길이

번호	y_1	y_2	y_3	y_4	d^2
1	47.8	48.8	49.0	49.7	0.76
2	46.4	47.3	47.7	48.4	1.30
3	46.3	46.8	47.8	48.5	1.76
4	45.1	45.3	46.1	47.2	3.85
5	47.6	48.5	48.9	49.3	0.87
6	52.5	53.2	53.3	53.7	2.81
7	51.2	53.0	54.3	54.4	4.29
8	49.8	50.0	50.3	52.7	7.99
9	48.1	50.8	52.3	54.4	11.03
10	45.0	47.0	47.3	48.3	5.85
11	51.2	51.4	51.6	51.9	2.83
12	48.5	49.2	53.0	55.5	10.57
13	52.1	52.8	53.7	55.0	2.59
14	48.2	48.9	49.3	49.8	0.66
15	49.6	50.4	51.2	51.8	0.32
16	50.7	51.7	52.7	53.3	0.83
17	47.2	47.7	48.4	49.5	1.11
18	53.3	54.6	55.1	55.3	4.36
19	46.2	47.5	48.1	48.4	2.11
20	46.3	47.6	51.3	51.8	10.09



[그림 3-8] 표준거리 산점도 행렬

변수 4개를 동시에 살펴보면 9번째, 12번째 데이터와 20번째 데이터의 Mahalanobis 거리가 크게 나타난다.

$$d_9^2 = 11.03,\ d_{12}^2 = 10.57,\ d_{20}^2 = 10.09$$

20개 데이터의 마할라노비스 거리는 자유도가 4인 카이제곱 분포의 상위 0.5%에 해당하는 값, $\chi_4^2(0.005) = 14.86$보다는 작으므로 이상치라고 보기는 어렵다. 그러나 20개 데이터 중 9번째, 12번째, 20번째 데이터의 d^2값이 크게 나타나므로 관측치들은 주의를 할 필요가 있다. □

5.4 주기 데이터에서 이상치 탐색

두 시점 $(t, t+1)$에서 조사된 아래의 데이터를 고려하자.

$$(y_i(t), y_i(t+1)), \quad i = 1,...,n$$

주기 데이터에서 이상치를 탐색하는 방법은 동일한 단위에 대하여 연속적인 두 측도의 비를 이용하는 것이다. 먼저 두 관측치의 비, 관측치 비의 평균과 표본 분산을 구한다.

- $r_i = \dfrac{y_i(t+1)}{y_i(t)}$: 두 관측치의 비
- $\bar{r} = \dfrac{1}{n}\sum_{i=1}^{n} r_i$: 관측치 비의 평균
- $s_r^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2$: 관측치 비의 표본 분산

두 시점의 관측치 비의 값이 일정 범위를 벗어나면 이상치로 간주한다. 즉, 만일 아래의 식을 만족하면 이면 i번째 관측치 $y_i(t+1)$은 이상치로 간주한다.

$$r_i \notin [\bar{r} - ks_r,\ \bar{r} + ks_r]$$

여기에서 k는 주어진 상수로 보통 경험적으로 결정한다.

5.5 이상치 처리

이상치 탐색을 통하여 발견한 이상치를 처리하는 방법은 두 가지이다. 첫째, 이상치가 발견된 레코드에 대하여 재조사 등 쌍방향 처리를 하거나 대체 처리를

하여 데이터를 수정한다. 둘째, 이상치를 수정하지 않고 로버스트(robust) 추정방법을 사용하여 로버스트 추정을 한다. 로버스트 추정법에 대해서는 Maronna et. al. (2006) 등을 참고하기 바란다.

6. 랜덤 오류

랜덤 오류(random error)는 체계적인 근거에 의해 발생하는 오류가 아닌 우연적으로 발생되는 오류를 말한다. 랜덤 오류는 조사과정에서 응답자, 면접원, 그리고 다른 직원의 부주의에 의해 발생한다.

6.1 오류위치포착

편집규칙을 데이터에 적용하여 에디팅을 수행하면 편집규칙을 위반하는 불일치 레코드를 찾을 수 있다. 그 다음에는 불일치 레코드 안에서 어느 데이터가 오류 데이터인지 검토하게 된다.

불일치 레코드 안에서 어느 변수에 오류가 있는지 확인하는 활동을 '오류위치포착(error localization)'이라고 한다. 보통 오류위치포착은 쉽지 않다. 오류위치를 포착하는 방법으로는 결정론적 점검규칙(deterministic checking rule)을 적용하는 방법과 일반적인 가이드라인(general guideline)을 준수하는 방법이 있다.

결정론적 점검규칙을 적용하는 방법에서는 불일치 레코드에서 어느 변수를 오류가 있는 변수로 간주해야 하는지를 결정론적으로 정한다. 흔히 결정론적 점검규칙은 결정론적 대체방법과 함께 쓰인다.

[예제 3-13] 결정론적 점검 규칙의 예

- 신뢰성이 낮은 변수를 오류 변수로 지정한다.
- 구성 변수들의 합이 대응되는 총계변수와 일치하지 않는 경우, 총계 변수가 오류를 포함하는 것으로 간주한다. □

위의 예제에서 보듯이 결정론적 점검 규칙은 단순하고 쉽다. 또한 오류 수정과정이 분명하다. 그러나 이러한 방법은 랜덤 오류를 다루는 적절한 방법이 아일 수 있다. 왜냐하면, 모든 레코드에 대하여 동일한 변수에 오류가 있다고 간주하기는 어렵기 때문이다.

결정론적 점검 규칙의 장점은 오류 수정 과정이 단순하고 투명하다는 점이다. 반면 단점은 많은 세부적인 점검 규칙을 명시하여야 하는 점이다. 그러한 규칙 개발에는 많은 시간과 자원이 필요하다. 또한 많은 수의 상세한 점검 규칙들의 타당성을 유지하고 점검하는 것은 실제적으로도 문제가 된다. 경우에 따라서는 결정적인 점검 규칙 개발이 불가능할 수도 있다.

일반적인 가이드라인을 이용하는 방법에서는 불일치 레코드에서 잘못된 변수의 위치를 일반적인 가이드라인에 의하여 포착한다. 일반 가이드라인 중에서 가장 널리 알려진 것이 Fellegi-Holt(1976) 패러다임이다. Fellegi-Holt 패러다임은 자동 에디팅과 대체에 관한 세 가지 원칙 중 첫 번째 원칙이다.

6.2 Fellegi-Holt의 자동 에디팅과 대체에 관한 세 가지 원칙

Fellegi-Holt가 제안한 자동 에디팅과 대체에 관한 세 가지 원칙은 다음과 같다.

(ⅰ) 레코드 안에 있는 데이터는 가능한 최소 항목을 변경하여 모든 편집규칙을 만족하도록 한다.

(ii) 데이터 파일의 빈도 구조는 가능한 한 유지한다.

(iii) 대체 규칙(imputation rule)은 명시적인 설명 없이 대응되는 편집규칙에서 유도한다.

첫 번째 원칙은 Fellegi-Holt 패러다임으로 알려져 있으며, 오류위치포착에 대한 원칙을 설명하고 있다. 두 번째와 세 번째 원칙은 랜덤 오류 탐색 후 자동 대체와 관련된 것이다. 세 번째 원칙은 오류위치포착과 대체는 서로 관계되어 있고 결합하여 적용되어야 함을 의미한다.

Fellegi-Holt 방법의 장점으로는 첫째, 관측된 원 데이터를 가능한 한 많이 보존하는 것을 목표로 한다. 둘째, 모든 편집규칙을 만족하는 일치성 있는 데이터를 유도한다. 셋째, 많은 변수들에 동시에 타당하게 적용할 수 있다. 넷째, 결정적 점검 규칙보다는 덜 상세한 규칙이 요구되고, 랜덤 오류에 결정적 처리를 하여 발생하는 편향이 나타나지 않는다.

반면, Fellegi-Holt 방법의 단점으로는 첫째, Fellegi-Holt 방법은 내부적으로 일치하는 레코드를 구성하도록 할 뿐, 특정 분포의 성질을 갖는 데이터를 구축하도록 하지는 않는다. 둘째, 이 방법에 기반한 시스템은 하나의 레코드에 대한 모든 편집규칙을 필수적 편집규칙인 것으로 간주해야 한다. 셋째, Fellegi-Holt 패러다임은 충분히 강력한 편집규칙이 준비된 경우에만 성공적일 수 있고, 또한 구현하기가 쉽지 않다.

6.3 랜덤 오류 탐색

만일 적절한 소프트웨어의 사용이 가능하다면 랜덤 오류는 Fellegi-Holt 패러다임을 적용하여 탐색하고 처리한다. 만일 그러한 적절한 소프트웨어가 없다면 랜덤 오류는 결정론적 점검 규칙을 사용하여 탐색하고 처리한다.

[예제 3-14] 다음은 어느 사회조사 데이터이다. 조사 대상은 10세 이상이고, 가구주와의 관계 변수의 수준은 1=가구주, 2=배우자, 3=자녀, 4=기타이며, 학력 변수의 수준은 0=무학, 1=초등학교, 2=중고등학교, 3=대학이상이고, 혼인상태 변수의 수준은 1=미혼, 2=기혼 값을 가지는 데이터이다. 다음의 데이터에서 3번째와 15번째 데이터는 Fellegi-Holt 패러다임에 의한 편집규칙에 의해 위반되었다고 판단되는 레코드이다.

랜덤 오류 탐색은 다음 절차를 따른다.

1단계 : 주제별 전문 지식을 바탕으로 3개의 편집규칙을 만들었다고 하자.

- e_1 : [가구주와의 관계=(가구주 혹은 배우자)] ⋂ [연령<15] ⇨ 규칙위반
- e_2 : [연령<17] ⋂ [혼인상태=기혼] ⇨ 규칙위반
- e_3 : [가구주와의 관계= 배우자] ⋂[혼인상태=미혼] ⇨ 규칙위반

[데이터 3-7] 사회조사 데이터

일련번호	가구번호	가구주와의 관계	연령	학력	혼인상태	편집규칙
1	1	1	33	2	1	
2	1	4	56	2	2	
3	1	2	11	3	1	위반
4	1	4	27	3	2	
5	1	4	32	3	1	
6	2	1	24	3	1	
7	2	2	23	3	2	
8	3	3	11	3	1	
9	4	3	15	3	1	
10	5	3	11	1	1	
11	6	1	38	3	1	
12	6	2	39	3	2	
13	7	4	67	1	2	
14	7	1	34	3	1	
15	7	3	16	3	2	위반
16	8	4	14	3	1	
17	9	1	34	3	1	
18	9	2	33	3	2	
19	10	1	30	2	1	
20	10	2	30	3	2	

2단계 : 1단계의 편집규칙으로부터 완비집합을 유도한다.

편집규칙	가구주와의 관계	연령	학력	혼인상태
e_1	가구주, 배우자	10-14세	무관	무관
e_2	무관	10-14세 15-16세	무관	기혼
e_3	배우자	무관	무관	미혼

3단계 : 편집규칙 완비집합으로부터 내재되어 있는 편집규칙 유도한다. 3개의 편집규칙으로부터 4가지의 편집규칙 조합을 얻는다.

편집규칙의 조합	구성 편집규칙
2개	$\{e_1,e_2\}$ $\{e_1,e_3\}$ $\{e_2,e_3\}$
3개	$\{e_1,e_2,e_3\}$

이로부터 4가지 편집규칙 × 4개의 변수 = 16가지 편집규칙을 새로 얻는다. 먼저 $\{e_1,e_2\}$×(가구주와의 관계)의 새로운 편집규칙을 유도하는 과정을 살펴본다.

생성변수	가구주와의 관계				연령			학력				혼인정도	
	가구주	배우자	자녀	기타	10~14	15~16	17세이상	무학	초등학교	중고등	대학이상	미혼	기혼
e_1	1	1	0	0	1	0	0	1	1	1	1	1	1
e_2	1	1	1	1	1	1	0	1	1	1	1	0	1
			∪			∩			∩			∩	
가구주관계	1	1	1	1	1	0	0	1	1	1	1	0	1

주1) 1 : 해당, 0 : 비해당

주2) 생성변수는 ∪로 조합 : 1 이 하나라도 있으면 1, 아니면 0

비생성변수는 ∩로 조합 : 0 이 하나라도 있으면 0, 아니면 1

결국, $\{e_1,e_2\}$×(가구주와의 관계, 연령, 학력, 혼인정도)로 유도되는 새로운 편집규칙은 4가지이며, 새로운 편집 규칙이 내재적 편집규칙이 되기 위해서는 다음 기준을 통과하여야 한다.

a. 모두 0인 변수가 있으면 제외한다.

b. 생성변수가 모두 1값을 가지면 제외한다.

c. 이미 존재하는 편집규칙인 경우 제외한다.

	편집규칙	가구주관계				연령			학력				혼인		내재적 편집규칙 기준
		가구주	배우자	자녀	기타	10~14	15~16	17세이상	무학	초등학교	중.고등	대학이상	미혼	기혼	
생성변수	가구주관계	1	1	1	1	1	0	0	1	1	1	1	0	1	b 해당/제외
	연령	1	1	0	0	1	1	0	1	1	1	1	0	1	통과
	학력	1	1	0	0	1	0	0	1	1	1	1	0	1	b 해당/제외
	혼인정도	1	1	0	0	1	0	0	1	1	1	1	1	1	b 해당/제외

생성변수가 연령인 편집규칙만이 기준을 통과하였다. 이를 e*라고 하자.

e^*: [가구주와의관계=(가구주,배우자)]⋂[연령<17]⋂[혼인정도=미혼]⇨규칙위반

$\{e_1,e_2\}$의 조합으로 만들어지는 새로운 편집규칙은 e^*이다. 마찬가지의 방법으로 다른 $\{e_1,e_3\}$, $\{e_2,e_3\}$, $\{e_1,e_2,e_3\}$과 4가지 변수의 조합으로 새로운 편집규칙을 만들 수 있다. 예제에서 이 과정은 생략한다.

편집규칙을 가지고 편집규칙을 통과하지 못한 레코드에 대해 채택영역을 결정하는 과정을 살펴본다. 예를 들어 3번 레코드의 경우, 가구주와의 관계는 배우자이고 연령은 11세, 학력은 중고등학교이며, 혼인상태는 미혼이다.

1단계 : 채택영역의 결정 단계

편집규칙	가구주관계				연령			학력				혼인		편집 규칙
	가구주	배우자	자녀	기타	10~14	15~16	17세이상	무학	초등학교	중.고등	대학이상	미혼	기혼	
e_1	1	1	0	0	1	0	0	1	1	1	1	1	1	위반
e_2	1	1	1	1	1	1	0	1	1	1	1	0	1	통과
e_3	0	1	0	0	1	1	1	1	1	1	1	1	0	위반
e_*	1	1	0	0	1	1	0	1	1	1	1	1	0	위반

현재 레코드가 통과하지 못한 편집규칙으로는 e_1에서는 가구주관계, 연령, e_3에서는 가구주관계, 혼인상태, e_*에서는 가구주관계, 연령, 혼인상태이므로 가장

많이 위반한 가구주관계를 대체하기로 결정한다.

2단계: 대체 단계

모든 편집규칙을 통과한 레코드 중에서, 3번 레코드와 연령, 학력, 혼인정도가 동일한 기부자(8번, 16번) 중에서 가구주와의 관계를 핫덱 대체한다. 확률적으로 8번 표본이 선택되면, [가구주와의 관계=자녀]로 대체한다.

3단계 : 마찬가지로 15번 레코드는 혼인상태가 대체되었다.

4단계 : 랜덤오류의 탐색과 대체 결과

일련번호	가구번호	가구주와의 관계	연령	학력	혼인상태
3	1	3*	11	3	1
15	7	3	16	3	1*

□

6.4 자동 에디팅 소프트웨어

자동 에디팅을 위하여 에디팅 소프트웨어가 개발되어 사용되고 있다. NIM과 BANFF는 캐나다 통계청에서 개발하였고, SLICE는 네덜란드 통계청에서 개발하였다. StEPS는 미국 인구통계국에서 개발하였다. 아래에 에디팅 소프트웨어에 대한 간단한 비교가 있다.

방법	CANCEIS (NIM)	GEIS (Banff)	Cherry Pie (SLICE)	StEPS
개발연도	1999	2003	2003	1999
시스템 개발자	캐나다 통계청	캐나다 통계청	네덜란드 통계청	미국 인구통계국
에디팅 방법	있음	있음	있음	없음
대체 방법	있음	있음	없음	있음
펠레지-홀트 방법	없음	있음	있음	없음
논리적 편집규칙	있음	있음	있음	있음
논리적 대체규칙	있음	있음	없음	있음
명목형 변수	적용가능	적용불가능	적용가능	적용가능
순서형 변수	적용가능	적용불가능	적용가능	적용가능
연속형 변수	적용가능	적용가능	적용가능	적용가능

Ⅳ. 오류 데이터 처리

1. 개요

오류 데이터를 탐색하여 오류가 있는 것으로 여겨지는 레코드가 선정되면 다음 단계에서는 해당 레코드에서 오류가 있을 것으로 여겨지는 항목을 찾는다. 그리고 오류 항목 값을 그럴듯한 값으로 교체하여 오류 데이터를 처리한다.

오류 데이터 처리 방법으로는 쌍방향 처리법(interactive treatment)과 대체 처리법(imputation treatment)이 있다. 쌍방향 처리방법은 응답자를 재접촉하거나 주제별 전문지식을 이용하여 오류 데이터를 수정하는 것을 말하며, 대체 처리법은 오류 항목 값이나 결측값을 레코드나 변수 수준에서 추정치로 대체하는 것을 말한다.

오류 데이터를 처리하여 나타나는 효과는 다음과 같다.

- 오류 항목 값이나 결측값을 처리하여 완비데이터를 제공한다.
- 완비데이터에 표준적인 추정 기법 적용을 가능하게 한다.

2. 대체 처리

대체는 오류 데이터 탐색 과정에서 발견된 결측값이나, 잘못된 값, 불일치 데이터를 처리하는 한 방법이다. 오류 데이터를 제거하고 그 자리에 추정치를 대입한다. 대체 방법은 네 종류로 분류된다: 규칙 기반 대체(rule based imputation), 연역적 대체(deductive imputation), 모형 기반 대체(model based imputation), 기부자 기반 대체(donor based imputation).

무응답을 대체하여 완비데이터를 만든 후 통상적인 추정 기법을 적용하면 다음과 같은 대체 효과가 나타난다.

- 무응답 대체 처리로 인하여 추정치가 편향될 가능성이 있다.
- 무응답 대체 처리로 인하여 분산이 과소 추정된다.

이러한 현상은 무응답 대체 값을 마치 응답값인 것처럼 간주하고 처리하여 나타나는 현상이다. 따라서 무응답 대체 처리를 할 때 가능하면 추정치에 편향이 개입하지 않는 방법을 선택하도록 하는 것이 중요하다.

대체 처리에서 무응답 편향을 줄이는 좋은 방법은 대체 칸(imputation cell)을 만드는 것이다. 대체 칸은 데이터를 동질적인 여러 개의 그룹으로 분할하여 만든다. 대체 칸 형성에 사용되는 보조 정보는 모든 레코드에 대하여 알 수 있어야 한다.

- 대체 칸은 층화변수 또는 공표 영역 같은 보조 정보를 이용하여 만든다.
- 대체 칸 안에서 대체될 변수들은 가능한 동질적이 되도록 한다.

2.1 규칙기반 대체

이 방법은 주제별 전문 지식을 통하여 오류 값을 규칙에 의하여 대체한다. 규칙 기반 대체는 일반적으로 “If …, then… ”과 같은 편집규칙을 사용하여 대체한다.

[예제 4-1] 규칙기반 대체

만일 근로자 수가 0이고 근무시간이 양수이면, 근무 시간은 0이다. 즉,

‘근로자 수 = 0, 근무시간 > 0’ ⇨ ‘근무시간 = 0’ 으로 수정한다. □

규칙기반 대체의 장점으로는 첫째, 체계적인 오류가 있고 그 오류의 상황이 알려져 있으면 사용하기 적절하다. 둘째, 범주형 변수와 수치형 변수 모두에 사용할 수 있다. 셋째, 규칙기반 대체는 사용하기 쉽다.

규칙기반 대체의 단점으로는 오류 원인이 확실하지 않으면 규칙 기반 대체는 추정치에 심각한 편향을 초래할 수 있다. 위의 예에서 ‘근로자 수’ 항목은 정확한 응답이고 ‘근무시간’ 응답에 오류가 있다면 근무시간을 수정하는 것은 올바른 방법이다. 그러나 그 반대인 경우는 대체로 인하여 새로운 오류가 추가되는 결과가 된다.

2.2 연역적 대체

연역적인 대체는 논리적이고 수학적인 추론에 기초하여 오류값을 대체한다. 대체된 항목이 모든 편집규칙을 만족시키는 유일한 값일 때 적용된다.

균형 편집규칙(balanced edit rule)을 이용하면 연역적 대체가 가능하며, 패널조사에서도 연도별 상황을 파악하면 연역적 대체를 적용할 수 있다.

[예제 4-2] 균형 편집규칙

총계 항목과 합산해야 할 세부 항목이 있다고 하자. 세부 항목 중 한 항목에서 무응답이 발생하면 그 항목 값은 총계 항목을 이용하여 유일하게 결정할 수 있다.

- 균형 편집규칙 : $y_1 + y_2 + y_3 = y_{tot}$
- 응답 값 : $y_1 = 20$, $y_2 = 30$, $y_3 = ?$, $y_{tot} = 100$
- 무응답 연역적 대체 : $y_3^{*} = y_{tot} - y_1 - y_2 = 100 - 20 - 30 = 50$ □

[예제 4-3] 패널조사

패널조사에서 한 가정의 자녀의 수를 조사하였다.

- 자녀의 수 응답 : 1년차 2명, 2년차 무응답, 3년차 2명
- 무응답 연역적 대체 : 논리적으로 2년차 자녀의 수는 2명으로 대체한다. □

연역적 대체 방법의 장점으로는 첫째, 대체 값이 논리적 추론에 기초하기 때문에 신뢰성이 있다. 둘째, 연역적 대체는 가장 간단하고 저비용인 대체 방법이다. 셋째, 관측치에 균형 편집규칙과 같은 제약조건이 있을 때 유일한 값을 유도할 수 있다. 넷째, 레코드에서 오류 위치가 확실하게 포착되면 연역적 대체는 참값을 이끌어낸다. 반면, 연역적 대체 방법은 일치성 제약 조건이 변수에 부여되지 않으면 사용하기 어려운 단점이 있다.

2.3 모형기반 대체

모형기반 대체에서는 명시적인 대체 모형을 이용하여 결측값을 대체한다. 대체

모형 구축에는 다음과 같은 정보를 이용한다.

- 표집 추출틀에 있는 변수 : 그룹 크기, 경제활동 분류 등
- 이전 조사 정보 : 결측 변수의 이전 조사 값 등
- 행정 데이터 등

모형기반 대체 방법에는 회귀대체(regression imputation), 비대체(ratio imputation), 평균대체(mean imputation), 로지스틱 회귀 대체(logistic regression imputation) 등이 있다.

크기 n인 표본에서 조사변수 y에 대하여 r명이 응답을 하고 $n-r$명이 무응답을 하였다고 하자. 보조변수 x는 n개 표본 모두에서 응답이 있다.

<표 4-1> 데이터 기호

표본번호	1	2	...	i	...	$n-1$	n
보조변수	x_1	x_2	...	x_i	...	x_{n-1}	x_n
조사변수	y_1	y_2	...	?	....	y_{n-1}	y_n

주) '?' 는 무응답 표시

회귀대체에서는 회귀직선를 추정하고 추정된 회귀직선을 이용하여 무응답을 대체한다. 응답 표본수를 r개라고 하고 응답 표본을 s_r이라고 하자. 응답 표본에서 반응변수를 y, 설명변수를 x로 하여 회귀방정식을 구한다.

- $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j,\ j \in s_r,\quad \hat{\beta}_0,\ \hat{\beta}_1$: 최소제곱추정치
- $\hat{\beta}_1 = \dfrac{\sum_{j \in s_r}(x_j - \bar{x}_r)(y_j - \bar{y}_r)}{\sum_{j \in s_r}(x_j - \bar{x}_r)^2},\ \bar{x}_r = \dfrac{1}{r-1}\sum_{j \in s_r} x_j,\quad \bar{y}_r = \dfrac{1}{r-1}\sum_{j \in s_r} y_j$
- $\hat{\beta}_0 = \bar{y}_r - \hat{\beta}_1 \bar{x}_r$

표본 i와 $(i+1)$이 무응답이라고 하자. 회귀직선에 x_i와 x_{i+1}을 대입하여 회귀대체값 y_i^*, y_{i+1}^*을 얻는다. 즉,

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad y_{i+1}^* = \hat{\beta}_0 + \hat{\beta}_1 x_{i+1}$$

여기에서 $\hat{\beta_0}$, $\hat{\beta_1}$은 응답데이터에서 구한 최소제곱추정치이다. 보조변수가 2개 이상이면 중회귀모형을 적합한 후 회귀대체를 하면 된다.

비 대체에서는 응답자 평균의 비를 이용하여 무응답을 대체한다. 표본i와 표본$(i+1)$에서 발생한 무응답의 비대체값은 다음과 같다.

$$y_i^* = \left(\frac{\bar{y}_r}{\bar{x}_r}\right)x_i, \; y_{i+1}^* = \left(\frac{\bar{y}_r}{\bar{x}_r}\right)x_{i+1}$$

평균대체는 응답자 평균으로 무응답을 대체하는 것이다. 표본i와 표본$(i+1)$에서 발생한 무응답의 평균대체값은 다음과 같다.

$$y_i^* = y_{i+1}^* = \bar{y}_r$$

여기에서 $\bar{x}_r, \bar{y}_r$은 응답데이터에서 구한 표본평균이다.

로지스틱 회귀대체는 조사변수 y가 0이나 1의 값을 갖는 변수일 때 사용가능하다.

먼저 보조변수 x를 이용하여 로지스틱 회귀모형을 구한다. 조사변수 y_i의 기댓값을 p_i라고 p_i의 추정값을 $\hat{p}_i$라고 할 때 다음의 추정식을 얻는다.

$$\log\left(\frac{\hat{p_i}}{1-\hat{p_i}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

로지스틱 회귀식을 추정 확률로 표현한 후, x_i를 대입하여 추정확률 $\hat{p}_i$를 구한다.

$$\hat{p_i} = \frac{e^{\hat{\beta_0}+\hat{\beta_1}x_i}}{1+e^{\hat{\beta_0}+\hat{\beta_1}x_i}}$$

만일 $\hat{p_i} \geqq 0.5$이면 무응답은 1로 대체한다. 그렇지 않으면 0으로 대체한다. 즉,

- $\hat{p_i} \geqq 0.5 \Rightarrow y_i^* = 1$
- $\hat{p_i} < 0.5 \Rightarrow y_i^* = 0$

보조변수가 여러 개(예를 들어 k개)이면 로지스틱 중회귀모형을 적합한다.

$$\log\left(\frac{\hat{p_i}}{1-\hat{p_i}}\right) = \hat{\beta_0}+\hat{\beta_1}x_{i1}+\ldots+\hat{\beta_k}x_{ik}$$

모형기반 대체방법의 장점으로는 첫째, 모형기반 대체는 컴퓨터 프로그램을 이용하여 대용량 데이터 처리에 적용하기 쉽고, 둘째, 대체 모형의 성질이 잘 알려져 있다. 예를 들어, 만약 대체될 변수의 속성이 응답자나 무응답자에 대하여 같다면, 대체 후 평균은 모평균의 비편향 추정치가 된다. 모형기반 대체방법의 단점으로는 모형기반 대체 방법을 사용하기 위해서는 명시적인 모형이 필요하다. 만일 명시적인 모형에 대한 확신이 없으면 모형기반 대체 방법은 사용하기 어렵다.

기부자 기반 대체에서는 편집규칙을 만족하지 않은 레코드의 무응답 항목에 기부자 레코드의 응답값을 대입한다. 기부자 집합은 현행조사의 응답자이거나 다른 조사의 응답자이다.

- 핫덱 대체(hot dect imputation) : 기부자를 현행 조사 응답자 중에서 선택
- 콜드덱 대체(cold deck) : 기부자를 과거 조사 응답자나 다른 자료 응답에서 선택

기부자를 선택하는 방법은 랜덤하게 기부자를 선택하거나 최근방 기부자를 선택하는 방법 등이 있다.

- 랜덤 기부자 대체 : 기부자 집합에서 기부자를 랜덤하게 선택

- 최근방 기부자 대체 : 수령자와 거리가 가장 가까운 기부자를 선택. 만일 둘 이상의 기부자가 동일한 최소 거리를 가지면 그 중 한 기부자를 랜덤하게 선택.

최근방 기부자 대체에서 기부자는 거리 함수를 이용하여 찾는다. 거리 함수는 보조변수가 연속형일 때와 범주형일 때 다르게 표현된다. 연속형 보조변수를 사용할 때에는 보조변수 데이터 x_i와 x_j의 거리를 측정한다. 거리함수를 d라고 하면 두 데이터의 절대거리는 다음과 같다.

$$d(x_i,x_j) = |x_i - x_j|$$

범주형 보조변수를 사용할 때에는 동일 범주에 속하는 데이터를 기부자로 한다.

기부자 기반 대체 방법의 장점으로는 첫째, 기부자 기반 대체는 명시적인 대체 모형에 의한 무응답 처리가 어려운 경우에도 적용 가능하다. 둘째, 실제 응답값이 무응답 대체에 사용되기 때문에 실제성이 있다. 기부자 기반 대체 방법의 단점으로는 대체 후 데이터는 편집규칙의 일치성을 만족하지 않을 수도 있다. 그러나 계산 알고리즘이나 기부자 집합 제한 방법으로 일치성 있는 대체 후 데이터를 얻을 수 있다.

3. 무응답 대체 방법 예제

범죄피해 상황을 조사하여 아래의 데이터를 얻었다. 표본수는 20이고, 조사변수는 나이, 성별(M=남성, F=여성), 교육기간, 범죄피해 경험(1=있음, 0=없음), 중범죄피해 경험(1=있음, 0=없음)이다. 무응답은 물음표(?)로 표시되어 있다. 나이와 성별에는 무응답이 없으나 교육기간은 3명, 범죄 피해는 2명, 중범죄 피해는 4명이 응답하지 않았다.

여러 가지 대체 방법을 사용하여 무응답을 처리하기로 한다.

9번째 사람은 범죄 피해가 없다('범죄 피해=0'). 따라서 논리적으로 중범죄 피해도 없어야 한다. 따라서 '중범죄 피해'에는 연역적으로 0의 값을 대체한다.

교육기간에 응답하지 않은 3명의 무응답은 평균 대체를 한다. 평균 대체를 위하여 교육 기간과 밀접한 연관이 있는 나이 변수와 성별 변수를 이용하여 대체 칸을 만든다. 나이 35세를 전후하여 4개의 대체 칸을 만들었다.

[데이터 4-1] 대체 처리 예제

표본 번호	나이	성별	교육기간	범죄 피해	중범죄 피해
1	47	M	16	0	0
2	45	F	?	1	1
3	19	M	11	0	0
4	21	F	?	1	1
5	24	M	12	1	1
6	41	F	?	0	0
7	36	M	20	1	?
8	50	M	12	0	0
9	53	F	13	0	?
10	17	M	10	?	?
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	M	16	1	0
15	44	M	14	0	0
16	45	M	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0

주) 출처: Lohr, S. L. (1999). 273쪽.

<표 4-2> 교육 기간 대체를 위한 대체 칸 구성

		나이	
		≦ 34	≧ 35
성별	M	응답자 번호: 3, 5, 10, 14	응답자 번호: 1, 7, 8, 15, 16
	F	응답자 번호: 4^*, 12, 13, 19, 20 $(y_4^* = \bar{y}_R = 11.25)$	응답자 번호 : 2^*, 6^*, 9, 11, 17, 18 $(y_2^* = y_6^* = \bar{y}_R = 12.25)$

나이가 34세 이하인 여성으로 구성된 대체 칸에는 5명의 표본이 해당되며 이 중 1명(4번 표본)이 응답을 하지 않았다. 따라서 응답을 한 4명의 교육기간 응답의 평균으로 무응답을 대체한다.

$$y_4^* = \frac{1}{4}(12+11+12+10) = 11.25$$

나이가 35세 이상인 여성으로 구성된 대체 칸에는 6명의 표본이 있으며 이 중 2명(2번, 6번 표본)이 응답하지 않았다. 따라서 응답을 한 4명의 교육기간 응답의 평균으로 무응답을 대체한다.

$$y_2^* = y_6^* = \frac{1}{4}(13+12+14+10) = 12.25$$

평균대체에서 만든 대체 칸(<표 4-2>) 안에서 범죄 피해 변수에 대한 핫덱 대체를 실시한다. 34세 이하 여성인 대체 칸에서 19번의 대체 표본은 4, 12, 13, 20번 표본 중에 하나를 선택한다. 이때 4번, 12번, 13번, 20번 표본은 기부자가 되고, 19번은 수령자가 된다.

랜덤 핫덱 방법: 기부자 4명(4번, 12번, 13번, 20번) 중에서 한 명을 난수표를 사용하거나 컴퓨터 난수를 사용하여 랜덤하게 선정한다. 20번이 나왔다고 하자.

그러면 수령자 19번은 20번의 데이터로 대체된다.

$$y_{19}^{*} = y_{20} = 0.$$

축차 핫덱 방법: 대체칸 안의 기부자가 많고 대체 받아야 하는 수령자가 많을 때에는 축차 핫덱 방법이 편리한 방법이다. 대체 칸 안에서 바로 직전 표본이 기부자가 된다. 34세 이하 여성인 대체 칸에서 19번째 표본에 대한 기부자는 직전 표본인 13번 표본이다. 따라서 13번 표본의 값을 19번 표본에 대체한다.

$$y_{19}^{*} = y_{13} = 1.$$

중범죄 피해 변수에 대한 무응답을 대체하자. 중범죄는 범죄의 일부이므로 논리적 편집규칙이 발생한다.

- [범죄 피해=1] ⇨ [중범죄 피해=0 혹은 1]
- [범죄 피해=0] ⇨ [중범죄 피해'=0]

중범죄 피해 무응답 대체는 위의 편집 규칙을 만족시키도록 하여야 한다. 성별에 따라 중범죄 피해가 다를 것으로 보고 성별과 범죄 피해 변수를 중심으로 대체 칸을 만들었다.

<표 4-3> 중범죄 피해 대체를 위한 대체 칸 구성

		범죄피해	
		1	0
성별	M	응답자 번호: 5, 7*, 14	응답자 번호: 1, 3, 8, 15, 16
	F	응답자 번호: 2, 4, 13*	응답자 번호 : 6, 9, 11, 12, 17, 18, 20

7번 표본은 남성이면서 범죄피해가 있는 대체 칸에 속하므로, 동일한 대체 칸에 속한 5번, 14번을 기부자로 하여 핫덱 대체한다. 5번 표본은 '중범죄 피해=1'이고, 14번 표본은 '중범죄 피해=0'이므로 둘 중의 하나를 1/2의 확률로 선택하여

대체한다.

5번 표본이 선택되었다고 하면, $y_7^* = y_5 = 1$이 된다. 13번 표본은 여성이면서 범죄피해가 있는 대체 칸에 속하므로, 동일한 대체 칸에 속한 2번, 4번을 기부자로 하여 핫덱 대체한다. 두 표본 모두 '중범죄 피해=1'이므로 $y_{13}^* = y_2 = 1$로 대체한다.

10번 표본은 범죄 피해와 중범죄 피해 변수에 대하여 응답을 하지 않았다. 중범죄 피해는 나이와 성별에 따라 다를 것으로 예상되므로, 10번 표본에 대하여 나이와 성별로 최근방 대체를 한다.

10번 표본(성별=M, 나이=17)과 성별이 동일하고 나이가 가장 근접한 표본은 3번 표본(성별=M, 나이=19)이다. 따라서 3번 표본이 기부자가 된다.

- 범죄 피해 : $y_{10}^* = y_3 = 0$
- 중범죄 피해 : $y_{10}^* = y_3 = 0$

범주형 반응 변수에 로지스틱 모형을 활용하여 대체할 수 있다. 범죄 피해 변수에 대하여 10번과 19번이 응답을 하지 않았다. 나이 변수를 보조변수로 하여 로지스틱 회귀대체를 한다. 로지스틱 회귀추정 결과 나이와 중범죄 피해의 관계는 다음과 같았다. $\hat{p}$를 중범죄 피해 확률이라고 하자.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 2.5643 - 0.0896 \times \text{나이}$$

위의 로지스틱 회귀식을 추정 확률로 표현하면 다음과 같다. 10번 표본의 나이는 17세이므로 로지스틱 추정식의 나이에 17세를 대입한다.

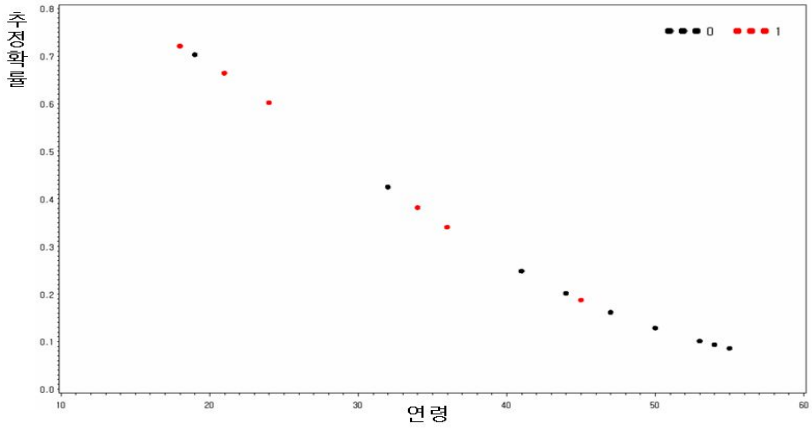
$$\hat{p} = \frac{e^{2.5643 - 0.0896 \times 17}}{1 + e^{2.5643 - 0.0896 \times 17}} = 0.73889$$

추정 확률이 0.5보다 크므로 범죄 피해 변수에는 1을 대체한다.

19번 표본의 나이는 29세이므로 로지스틱 추정식의 나이에 29세를 대입한다.

$$\hat{p} = \frac{e^{2.5643 - 0.0896 \times 29}}{1 + e^{2.5643 - 0.0896 \times 29}} = 0.49125$$

추정 확률이 0.5보다 작으므로 범죄 피해 변수에는 0을 대체한다.



[그림 4-1] 중범죄 피해에 대한 로지스틱 추정

대체 결과로 얻어진 완비 데이터는 다음과 같다.

<표 4-4> 대체 후 완비 데이터

표본 번호	나이	성별	교육기간	범죄 피해	중범죄 피해	대체방법
1	47	M	16	0	0	
2	45	F	12.25*	1	1	*평균대체
3	19	M	11	0	0	
4	21	F	11.25*	1	1	*평균대체
5	24	M	12	1	1	
6	41	F	12.25*	0	0	*평균대체
7	36	M	20	1	1*	*핫덱 대체
8	50	M	12	0	0	
9	53	F	13	0	0*	*연역적 대체
10	17	M	10	0* 1**	0	*최근방대체 **로지스틱 대체
11	53	F	12	0	0	
12	21	F	12	0	0	
13	18	F	11	1	1*	*핫덱 대체
14	34	M	16	1	0	
15	44	M	14	0	0	
16	45	M	11	0	0	
17	54	F	14	0	0	
18	55	F	10	0	0	
19	29	F	12	0* 1**	0	*랜덤핫덱/로지스틱 **축차핫덱 대체
20	32	F	10	0	0	

4. 쌍방향 처리

쌍방향 처리는 에디팅 과정에 재검토자가 개입하여 수동으로 오류 데이터를 처리하는 방법이다.

4.1 쌍방향 에디팅 방법

자동 에디팅과는 달리 쌍방향 에디팅에서는 재검토자가 오류 점검과 오류 원인

파악에 개입하여 오류를 검토하고 체계적인 오류 원인에 대한 정보를 수집한다.

- 원 데이터와 컴퓨터 파일에 저장된 데이터를 수동으로 점검하고 처리한다.
- 응답자 재조사를 통하여 표기된 오류 레코드를 수정한다.
- 쌍방향 처리에 주제별 전문 지식을 활용한다.

원 데이터 점검에서는 조사표 복사본이나 조사표 촬영사진을 통하여 원 데이터를 점검한다. 이를 통하여 재검토자는 조사표 작성 시 발생하는 다음과 같은 오류를 발견할 수 있다.

- 응답자가 몇몇 질문을 잘못 이해해서 생기는 오류
- 응답 기록시 잘못 기록하여 발생하는 오류.

예를 들어, 조사표 테이블에서 답변이 한 행씩 밀려 기록되어 결과적으로 응답한 값과 기록된 값이 아무런 연관이 없는 상황이 발생할 수 있다. 이러한 종류의 오류는 자동 에디팅으로는 탐색하기 어려운 반면, 조사표 복사본이나 컴퓨터 이미지 화면을 검사하면 쉽게 식별할 수 있다.

컴퓨터 파일로 저장된 데이터 점검에서는 컴퓨터 파일 데이터를 원 데이터와 비교하여 데이터 코딩이나 데이터 획득 과정에서 발생한 오류를 발견한다.

오차가 있는 항목의 위치가 확실하게 포착되지 않을 경우 데이터를 수정하기 위하여 응답자를 재접촉하기도 한다. 응답자를 재조사하면 비록 일치성 규칙을 만족하더라도 이상치와 같이 관측 데이터 분포의 꼬리 부분에 놓은 데이터를 확인하거나 수정하는데 유용하다. 이때 응답자 재조사 수와 길이는 응답자 부담을 줄이고, 응답률에 미치는 부정적인 효과를 피하기 위하여 최소한으로 한다. 응답자 재조사는 단순하게 의심스러운 데이터로부터 더 나은 그럴듯한 응답을 받으려고 하기 보다는, 응답자가 올바른 대답을 제공할 수 있는 문제로만 제한한다. 만

일 응답자의 이해하기 어려운 질문이라면, 재조사를 하더라도 올바른 응답을 받아낼 수 없기 때문이다. 재조사를 수행할 때, 응답자와 부딪치는 문제는 조사를 개선하기 위하여 수집한다.

주제별 지식을 쌍방향 처리에 이용하면 데이터 품질을 높일 수 있다. 쌍방향 처리를 하는 재검토자는 처리하는 레코드와 조사표 변수, 그리고 에디팅 절차에 대한 지식이 있어야 한다. 만일 그렇지 않으면 쌍방향 처리는 매우 주관적인 에디팅 결과를 초래할 수 있다.

4.2 쌍방향 처리의 범위

모든 오류 데이터에 대하여 주제별 지식을 이용한 쌍방향 처리를 하는 것은 많은 시간과 자원을 소모하기 때문에 바람직하지 않다. 쌍방향 처리는 자동 에디팅으로 해결할 수 없는 문제로 제한한다. 쌍방향 처리는 영향력 있는 오류나 이상치 같은 매우 중요한 오류 레코드로 한정한다. 지나친 쌍방향 처리는 과도한 에디팅을 초래하기 때문에 최종적으로는 도리어 데이터 품질을 떨어뜨릴 수 있다.

4.3 쌍방향 처리의 한계점

쌍방향 처리는 재검토자가 주관적인 수정을 할 위험이 있다. 데이터 수정이나 변경이 어떤 근거에 의하지 않고 단지 재검토자의 의견에 따라 이루어질 때 쌍방향 처리는 데이터에 심각한 편향을 가져온다. 쌍방향 에디팅은 재생성을 보장하지 않는다. 특히 많은 점검규칙이 있을 때 쌍방향 에디팅으로는 일치성과 완비성을 확인하기 어렵다. 쌍방향 처리는 비용과 시의성 관점에서 매우 값비싼 방법이다. 많은 비용 때문에 쌍방향 처리는 영향력 있는 레코드나 잠재적으로 영향력 있는 오류를 포함하는 소규모 표본 데이터로 대상을 한정한다. 많은 레코드가 점

검되어야 할 때 쌍방향 처리는 전문가에 의해 수행되기보다는 구체적인 데이터 성격에 숙련된 직원에 의해서 수행된다. 이러한 환경에서 수행된 쌍방향 처리 결과는 전문가에 의한 처리나 자동 에디팅 처리를 할 때보다 더 나빠질 수도 있다.

Ⅴ. 데이터 에디팅 문서 기록

1. 개요

데이터 에디팅 과정을 문서로 기록하여 조사 관리자, 에디팅 전문가 및 이용자에게 데이터 품질, 에디팅 과정, 에디팅 성과 등을 알린다. 에디팅 과정에 대한 기록 문서로는 에디팅 방법론 문서, 리포트, 그리고 에디팅 관련 중요 정보와 데이터 등이 있다.

에디팅 기록 문서에는 다음과 같은 내용을 기록한다.

- 에디팅에 사용된 자원 및 기간
- 에디팅 과정에서 나타나는 품질지표

데이터 에디팅 자원 관련 기록으로는 다음과 같은 내용을 고려한다.

- 에디팅을 수행하는데 소요된 전체 기간
- 에디팅 단계별로 소요된 기간
- 에디팅을 수행하는데 참여한 전체 인력

- 에디팅 단계별 참여 인력
- 재조사 수 및 재조사 기간 등
- 총 비용, 에디팅 단계별 비용, 장비 비용, 컴퓨터 사용 비용 등

2. 데이터 에디팅 방법론 문서

에디팅 방법론 문서의 주 이용자는 에디팅 전문가, 조사 방법론 전문가, 주제 분야 전문가들이다. 에디팅 전략에 관한 의사결정을 돕기 위하여 에디팅 관련 지표들을 방법론 문서에 기록하고 평가한다.

- 에디팅에 사용된 전략, 에디팅 과정의 흐름, 에디팅 방법 등을 기록한다.
- 에디팅 각 단계에서 에디팅 방법에 대한 설명과 함께 입력 데이터와 출력 데이터 분석을 기록한다. 이 분석은 주요 변수에 대한 기술 통계량과 에디팅으로 인한 효과 등을 포함한다.
- 에디팅 효과는 에디팅 단계별로 입력과 출력에 기초한 지표를 사용하여 측정한다.

3. 리포트

데이터 에디팅 과정에 대한 리포트를 만드는 목적은 에디팅 과정에서 중요한 점과 에디팅 전과 후의 데이터 품질에 대한 정보를 이용자에게 제공하는 것이다.

3.1 에디팅 과정에 대한 응답률 및 대체율 지표

에디팅 과정에서 단위 응답률, 항목 응답률 그리고 대체율은 매우 중요한 지

표이므로 이를 기록한다. 가중치를 이용하여 계산하면 가중 응답률, 대체율이 되며, 가중치를 이용하지 않으면 비가중 응답률, 대체률이 된다.

데이터 에디팅 관련 지표를 정의하기 위하여 다음의 기호를 사용하자.

n : 표본 레코드 수

k : 변수의 수

y_{ij} : i번째 레코드의 j번째 항목의 조사변수 값

y_{ij}^{*} : i번째 레코드의 j번째 항목 조사변수의 대체 값

x_{ij} : i번째 레코드의 j번째 항목의 보조변수 값

w_i : i번째 레코드에 부여된 가중치

r_{ij} : 응답 지시자 (응답: $r_{ij}=1$, 결측: $r_{ij}=0$)

r_{ij}^{*} : 에디팅 후의 응답 지시자 (응답: $r_{ij}^{*}=1$, 결측: $r_{ij}^{*}=0$)

[예제 5-1] 에디팅 관련 지표 및 기호

레코드 수 $n=10$, 변수의 수 $k=2$인 경우를 고려하자.

번호 (i)	가중치 (w_i)	원 관측치 (에디팅 전)		응답 지시자 (에디팅 전)		대체 후 데이터		대체 후 응답 지시자	
		y_{i1}	y_{i2}	r_{i1}	r_{i2}	y_{i1}^{*}	y_{i2}^{*}	r_{i1}^{*}	r_{i2}^{*}
1	100	M	23	1	1	M	23	1	1
2	100	M	43	1	1	M	43	1	1
3	120	F	22	1	1	F	22	1	1
4	120	F	?	1	0	F	36*	1	1*
5	150	M	37	1	1	M	37	1	1
6	150	M	?	1	0	M	38*	1	1*
7	180	M	39	1	1	M	39	1	1
8	180	?	53	0	1	F*	53	1*	1
9	130	F	51	1	1	F	51	1	1
10	130	?	?	0	0	F*	22*	1*	1*

□

(1) 단위 응답률 :

- 비가중 단위 응답률 : $\frac{1}{n}\sum_{i=1}^{n}\left(1-\prod_{j=1}^{p}(1-r_{ij})\right)$
- 가중 단위 응답률 : $\frac{\sum_{i=1}^{n}w_i\left(1-\prod_{j=1}^{p}(1-r_{ij})\right)}{\sum_{i=1}^{n}w_i}$

(2) 항목 응답률 :

- 비가중 항목 응답률 : 변수 y_j에 대하여, $\frac{1}{n}\sum_{i=1}^{n}r_{ij}$
- 가중 항목 응답률 : 변수 y_j에 대하여, $\frac{\sum_{i=1}^{n}w_i r_{ij}}{\sum_{i=1}^{n}w_i}$

(3) 비가중 대체율 : 변수 y_j에 대하여, $\frac{1}{n}\sum_{i=1}^{n}I(y_{ij}^{*}\neq y_{ij})$

(4) 가중 대체율 : 변수 y_j에 대하여, $\frac{\sum_{i=1}^{n}w_i I(y_{ij}^{*}\neq y_{ij})y_{ij}^{*}}{\sum_{i=1}^{n}w_i y_{ij}^{*}}$

[예제 5-2] 응답률 및 대체율 계산

예제 5-1에서 두 변수에 대한 응답률과 대체율을 계산하여 보자.

- 비가중 단위 응답률 $=9/10=90\%$
- 가중 단위 응답률 $=1{,}230/1{,}360=90.4\%$

- y_1에 대한 항목 응답률 :
  - 비가중 항목 응답률 $= 8/10 = 80\%$
  - 가중 항목 응답률 $= 1{,}050/1{,}360 = 77.2\%$
- y_2에 대한 항목 응답률 :
  - 비가중 항목 응답률 $= 7/10 = 70\%$
  - 가중 항목 응답률 $= 960/1{,}360 = 70.6\%$

- y_1에 대한 비가중 대체율 $= 2/10 = 20\%$
- y_2에 대한 비가중 대체율 $= 3/10 = 30\%$
- y_2에 대한 가중 대체비 $= \dfrac{120\times 36 + 150\times 38 + 130\times 22}{100\times 23 + \ldots + 130\times 22}$

$$= \frac{12{,}880}{50{,}860} = 25.3\% \qquad \square$$

3.2 에디팅 과정에 대한 메타 데이터

에디팅 과정과 관련하여 아래의 메타 데이터를 기록한다.

- 조사 단위의 범주에 대한 구체적인 정의를 기록한다.
  - 응답, 무응답, 범위 내 단위, 범위 밖 단위 등
  - 각 단위에 대한 무응답 원인 등
- 가중 방법에 대한 정밀한 설명을 한다. 이때 적절한 보조변수를 포함한다.
- 데이터 수집 방법, 수집 도구에 대한 설명한다.
- 데이터 교체 관련 정보를 기록한다.
- 대체 방법, 재가중 방법 등에 대한 정밀한 설명을 기록한다.

항목 응답과 관련하여 추가로 민감한 질문이나 조사표 길이 등 조사표 내용과 연관된 응답 부담 관련 정보를 추가한다. 항목 응답에 대한 지표는 주요 변수에 대해서만 계산한다.

4. 자료 저장

에디팅 과정에 대한 자료 저장은 에디팅 과정을 심층 고찰하기 위하여 에디팅 과정의 데이터와 중요 정보를 저장하는 것이다.

자료저장(archiving)은 새로운 에디팅 방법이나 에디팅 과정을 검사하고 새로운 품질 측정을 개발하기 위하여 필요하다. 또한 자료 저장은 새로운 이용자 요구가 발생할 때 다른 에디팅 과정을 구축할 수 있는지 확인하는데 도움을 준다.

자료 저장은 2차 분석이 이루어질 때나 혹은 에디팅된 데이터가 다른 통계에 입력 값으로 사용될 때 에디팅 과정과 관련된 질문에 답하기 위해서 필요하다. 만일 데이터 에디팅을 반복하거나 혹은 특정한 조사를 위하여 새로운 방법이나 전체 에디팅 절차를 검사하려고 한다면 데이터, 프로그램, 해당 메타데이터를 저장해야 한다. 사용된 프로그램, 쌍방향 처리를 위한 가이드라인을 포함한 관련 설명, 편집 규칙 집합, 데이터 흐름도에 대한 설명 등이 재생성을 보장하기 위하여 저장한다. 자료를 저장할 때에는 비밀보호 측면을 고려한다.

부록A. 에디팅 가이드라인

A1. 일반사항

- 데이터 에디팅은 비용과 시간이 소비된다. 주어진 자원과 시간의 제약 조건에서 데이터 정확성 향상과 자원 소비를 절충한다.
- 에디팅 자원은 중대한 데이터 오류에 초점을 맞추어 사용한다. 선택적 에디팅은 중요한 오류 처리에 적절하며, 그 처리는 쌍방향으로 응답자에 대한 재조사도 포함한다.
- 데이터 에디팅의 중요한 기능 중의 하나는 비표집오차의 특성과 그 원인을 파악하는 것이다. 이러한 정보는 현행 조사에 대한 품질 측도를 제공하는데 사용하고, 향후 조사를 개선하는데 사용하기도 한다.
- 조사의 여러 단계에서 수행되는 에디팅 활동은 서로 연결되고 조화를 이루어야 한다.
- 쌍방향 처리는 과도한 편집을 피해야 한다. 과도한 편집은 에디팅을 위한 자원과 시간이 데이터 품질 개선으로 이어지지 못할 때 발생한다.
- 오류 유형, 오류를 처리하는데 사용된 방법과, 데이터 품질에 대한 정보는 데이터의 올바른 사용을 위하여 이용자에게 제공한다.
- 에디팅을 수행하는 직원은 이론/실무적인 관점에서 적절하게 훈련되어야 한다. 에디팅 방법들에 대해 충분한 지식을 가지고 있어야 하고, 에디팅 활동과 조사 절차의 다른 부문 간의 연결에 대해 잘 알고 있어야 한다. 게다가 에디팅 과정의 문서 기록의 중요성을 알고 있어야 한다.
- 이러한 목적을 위하여 적절한 에디팅 교육이 주기적으로 조사 감독, 에디팅 전문가에게 제공되어 에디팅 방법론과 일반적인 에디팅에 관한 지식을 보급해야 한다.

A2. 데이터 오류 탐색

단계별로 편집규칙을 사용하여 데이터 오류를 탐색한다.

- 범위 오류 탐색
  - 사전에 지정한 응답범위를 벗어나는 응답을 탐색한다.
  - 응답의 범위는 이전 조사 결과를 참조하거나 전문자의 의견을 참고한다.
  - 범주형 변수는 응답범주가 아닌 응답이 있는지 검토한다.
- 일치성 및 논리성 점검
  - 응답의 일치성을 점검한다.
  - 개별 항목의 합과 총계 항목이 일치하는지 검토한다.
  - 다른 항목과 비교하여 응답이 논리적인지 검토한다.
- 허용 가능한 값이 아닌 응답이나 데이터에 부여된 코드 값이 아닌 응답이 있는지 검토한다.
- 결측값을 검토한다. 항목 무응답과 구조적 결측은 구분하여 표기한다.
- 집계치나 추정치를 외부 자료의 결과를 비교하여 비일치성이 있는지 점검한다.

A3. 데이터 오류 처리

편집규칙을 위반하는 레코드는 다음과 같이 처리한다.

- 명백한 오류는 찾아 항목값을 제거하고, 연역적 대체나 규칙 기반 대체를 하여 값을 채운다.
- 추정치에 영향을 많이 주는 오류는 쌍방향 처리한다.
  - 응답자의 응답을 재확인한다.
  - 재조사를 통하여 응답을 수정한다.

- 활용 가능한 다른 정보를 이용하여 응답을 수정한다.

• 응답자 접촉이 불가능하거나 재조사가 어려운 경우, 혹은 시간, 비용 등 조사 여건 상 쌍방향 처리가 어려운 경우 조사 관리자의 재점검을 거쳐 사전에 정한 기준에 의하여 자동 수정을 한다.

• 최종 결과물을 대상으로 집계와 추정치를 재점검한다.

A4. 대체 처리

• 연역적 대체는 하나의 참값이 예상될 때 우선적으로 고려한다.

• 규칙 기반 대체는 오류의 상황을 잘 이해할 수 있을 때 사용한다.

• 모형기반 대체에서 대체 모형에 포함되는 변수들의 타당성을 확인해야 한다. 적은 수의 대체 칸과 많은 수의 대체 모형은 피한다.

• 기부자 대체에서 거리함수는 주의 깊게 선택하고 자세하게 기록한다.

• 대체된 데이터의 일치성을 점검한다.

• 대체된 값을 표기한다.

A5. 쌍방향 처리

• 쌍방향 처리는 주제문제 전문가나, 주제 문제에 전문 지식을 가진 사람이 수행한다.

• 이전 조사 혹은 행정 데이터로부터 얻은 정보, 원 데이터를 활용한다.

• 주제 문제에 대해 지식이 없는 사람은 분명한 경우만 수정하도록 제한한다.

• 쌍방향 처리 결과로 발생한 데이터 변경은 기록하고 표기한다.

부록B. 용어 해설

- **결정론적 점검규칙 (deterministic checking rule)** : 불일치 레코드에서 어느 변수를 오류가 있는 변수로 간주해야 하는지를 결정론적으로 정하는 방법.
- **결측값 (missing value)** : 응답자가 응답을 하지 않아서 발생하는 값. 결측값은 응답자가 답을 알지 못하거나, 응답하고 싶지 않거나 또는 단순히 질문을 놓쳤을 경우와 같은 이유에 의해 발생함.
- **구조적 결측값 (structural missing value)** : 여과 질문에 의하여 응답을 하지 않아도 되는 문항에서 발생하는 값. 예를 들어 가구주에게만 묻는 질문 문항은 다른 가구원 응답은 구조적 결측이 됨.
- **규칙 기반 대체 (rule based imputation)** : 주제별 전문 지식을 통하여 오류 값을 규칙에 의하여 대체함. 규칙 기반 대체는 일반적으로 “If …, then … ”과 같은 편집규칙을 사용하여 대체함.
- **균형 편집규칙 (balanced edit rule)** : 총계는 각 부분의 합과 일치하도록 하는 편집규칙. 회계 편집규칙(accounting edit)이라고도 함.
- **그래픽 에디팅 (graphic editing)** : 상자 그림, 산점도, 히스토그램 등과 같은 그래픽 표현 방법을 이용하는 에디팅.
- **기부자 기반 대체 (donor based imputation)** : 편집규칙을 만족하지 않는 레코드의 무응답을 기부자의 레코드의 응답값으로 대체함. 기부자 집합은 현행 조사의 응답자이거나 다른 조사의 응답자임.
- **단위무응답 (unit nonresponse)** : 조사표에 응답자가 전혀 응답을 하지 않

아 발생하는 무응답. 단위무응답은 조사대상자를 접촉하지 못하거나 조사대상자가 응답을 거부하여 발생함. 단위무응답은 응답자 가중치를 조정하는 가중치 조정법으로 처리하는 것이 보통임.

- **대체 (imputation)** : 오류 데이터 탐색 과정에서 발견된 결측 값이나, 잘못된 값, 불일치 데이터를 처리하는 한 방법. 오류 데이터를 제거하고 그 자리에 추정치를 대입함. 대체 방법은 규칙 기반 대체(rule based imputation), 연역적 대체(deductive imputation), 모형 기반 대체(model based imputation), 기부자 기반 대체(donor based imputation) 등이 있음.
- **대체 칸 (imputation cell)** : 대체 처리에서 무응답 편향을 줄이는 위하여 데이터를 동질적인 여러 개의 그룹으로 분할하여 만든 칸. 대체는 대체 칸 안에서 이루어짐.
- **데이터 에디팅 (data editing)** : 통계조사 과정에서 데이터의 논리적으로 일치성을 결여한 오류를 찾고 수정하는 활동을 말함.
- **데이터 오류 (data error)** : 관측 값과 실제 값이 다를 때 발생함. 데이터 오류는 결측값, 타당하지 않은 값, 일치하지 않는 값, 이상치 등의 형태로 나타남.
- **랜덤 기부자 대체 (random donor imputation)** : 기부자 집합에서 기부자를 랜덤하게 선택하여 대체하는 방법.
- **랜덤 오류 (random error)** : 체계적인 이유가 아닌 우연히 발생하는 오류. 랜덤 오류는 조사과정에서 응답자, 면접원, 그리고 다른 직원의 부주의에 의해 발생함.
- **마이크로에디팅 (microediting)** : 조사표나 데이터 레코드 수준에서 수행하는 에디팅. 전체 데이터로 계산한 집계치나 추정치를 참고하지 않고 개별 관측치를 조사하여 오류를 탐색하거나 처리함. 조사표나 응답자 수준에서 응답의 타당성, 일치성 등을 점검할 수 있음.

- **마할라노비스 거리 (Mahalanobis distance)** : 변수가 k개인 다변량 데이터 $(y_1,...,y_k)$에서 m을 평균 벡터, C를 공분산 행렬이라고 할 때 관측치 y로부터 중심 m까지의 거리 $d^2=(y-m)^TC^{-1}(y-m)$을 말함.
- **매크로에디팅 (macroediting)** : 전체 데이터나 혹은 대부분의 데이터를 동시에 이용하여 실시하는 에디팅. 통계적인 모형을 이용하여 그래픽 방법으로 하거나 집계치 혹은 추정치를 이용한 수치 방법으로 함.
- **모형기반 대체 (model based imputation)** : 명시적인 대체 모형을 이용하여 결측값을 대체하는 방법. 회귀대체(regression imputation), 비대체(ratio imputation), 평균대체(mean imputation), 로지스틱 회귀 대체(logistic regression imputation) 등이 있음.
- **범위 점검 (range test)** : 사전에 응답 값의 범위를 정한 후 응답이 범위를 벗어나는지 점검함.
- **불일치하는 값 (inconsistent value)** : 정의된 데이터 항목의 관계를 만족하지 않는 값.
- **비율 점검 (ratio test)** : 두 변수 x, y의 비율에 대한 상한과 하한을 사전에 지정하고 비율(y/x)이 상한과 하한을 벗어나면 오류가 있는 것으로 간주함.
- **사분위수 범위 (quartile range)** : 데이터를 크기순으로 정렬하여 Q_1, Q_2, Q_3를 사분위수라고 할 때 Q_3-Q_1이 사분위수 범위임.
- **선택적 에디팅 (selective editing)** : 가장 중요한 추정치에 초점을 맞추어 오류가 있을 것으로 여겨지는 레코드만 점검하는 방법. 선택적 에디팅은 필수적 편집규칙을 위주로 실시함.
- **수동 에디팅 (manual editing)** : 수동으로 데이터 오류를 탐색하고, 탐색된 데이터 오류를 처리하는 것을 말함.

- **쌍방향 에디팅 (interactive editing)** : 데이터 획득 후에 오류 탐색 및 처리를 사람이 컴퓨터의 도움을 받아 수동으로 진행하는 에디팅을 말함. 응답자를 재접촉하거나 주제별 전문지식을 이용하여 오류 데이터를 수정함.
- **에디팅 (editing)** : 데이터 에디팅(data editing)
- **여과 질문 (filter question)** : 조사표에서 해당되는 문항을 선택하도록 하는 질문. 예를 들어 가구원 중 가구주에게만 묻는 문항은 가구주만 대답을 해야 함. 따라서 가구주임을 묻는 질문이 여과 질문임.
- **연역적 대체 (deductive imputation)** : 논리적이고 수학적인 추론에 기초하여 오류값을 대체하는 방법. 대체된 항목이 모든 편집규칙을 만족시키는 유일한 값일 때 적용됨.
- **영향력 있는 오류 (influential error)** : 집계나 추정치에 영향을 미치는 오류. 오류의 영향은 집계하는 추정치에 따라 달라짐. 어떤 추정치에 영향이 큰 오류가 다른 추정치에는 영향이 작을 수도 있음.
- **오류위치포착 (error localization)** : 불일치 레코드 안에서 어느 변수에 오류가 있는지 확인하는 활동.
- **의문 편집규칙 (query edit, 혹은 soft edit, statistical edit)** : 의심스러운 값을 탐색하는 편집규칙.
- **이상치 (outlier)** : 대다수 데이터에서 멀리 떨어진 관측치를 말함. 일변량 데이터에서는 분포의 양 끝에 있는 값을 이상치로 간주하고, 이변량 데이터에서 직선 부근에서 벗어난 값을 이상치로 간주함. 시계열 데이터에서는 다른 데이터에 비하여 상대적으로 크거나 값은 값을 이상치로 간주함.
- **자동 에디팅 (automated editing)** : 컴퓨터 프로그램을 이용하여 저장된 데이터의 오류를 탐색하고 처리하는 방법.
- **자료저장 (archiving)** : 에디팅 과정을 심층 고찰하기 위하여 에디팅 과정의

데이터와 중요 정보를 저장하는 것.

- **체계적인 오류 (systematic error)** : 특정 응답 항목에서 일관되게 보고되는 오류. 체계적인 오류는 데이터에 체계적인 영향을 줌.
- **최근방 기부자 대체 (nearest neighbor imputation)** : 수령자와 거리가 가장 가까운 기부자를 선택하여 대체하는 방법. 만일 둘 이상의 기부자가 동일한 최소 거리를 가지면 그 중 한 기부자를 랜덤하게 선택.
- **콜드덱 대체 (cold deck)** : 기부자를 과거 조사 응답자나 다른 자료 응답에서 선택하여 대체하는 방법.
- **타당하지 않은 값 (invalid value)** : 타당하지 않은 값은 응답으로서 허용 가능한 범위를 벗어나는 데이터 값. 예를 들어 입력된 자료가 허용된 값의 범위를 벗어나면 타당한 응답이 아님.
- **펠레지-홀트 패러다임 (Fellegi-Holt paradigm)** : Fellegi-Holt의 자동 에디팅과 대체에 관한 세 가지 원칙 중 첫 번째 원칙. 즉, 레코드 안에 있는 데이터는 가능한 최소 항목을 변경하여 모든 편집규칙을 만족하도록 함.
- **편집규칙 (edit rule)** : 만일 데이터가 올바르다면 반드시 만족해야 하는 데이터 항목 값 사이의 논리적 조건이나 제약조건. 점검규칙(check rule)이라고도 함.
- **필수적 편집규칙 (fatal edit, 또는 hard edit)** : 확실하게 오류를 찾아내는 편집규칙을 말함.
- **핫덱 대체 (hot dect imputation)** : 기부자를 현행 조사 응답자 중에서 선택하여 대입하는 방법.
- **항목무응답 (item nonresponse)** : 응답자가 몇 개의 항목에는 응답을 하고 나머지 항목에 응답을 하지 않았을 때 발생한 무응답. 항목무응답은 대체 방법에 의하여 처리하는 것이 보통임.

참고문헌

김규성 (2000). “무응답 대체 방법과 대체효과.” 조사연구 1권 2호: 1-14.

박진우 · 박현주 · 김진억 (2005). “주택가격동향조사를 위한 데이터 편집 사례연구.” 조사연구 6권 1호: 83-98.

변종석(2007). “Introduction to data editing.” 2007년 통계의 날 기념 워크숍 논문집, 3-73.

통계청 (2008). 『2007년 기준 광업 · 제조업통계조사 조사지침서』

Barneff, V. and Lewis, T. (1978). *Outliers in statistical data*. Wilry.

Biemer, P.P. and Lyberg, L.E. (2003). *Introduction to survey quality*. Wiley.

Fellegi, I.P. and Holt, D. (1976). “A systematic approach to automatic edit and imputation.” *Journal of the American Statistical Association* 71: 17-35.

Giles, P. and Patrick, C. (1986). “Imputation options in a generalized edit and imputation system.” *Survey Methodology* 12: 49-60.

Granquist, L. (1995). “Improving the traditional editing process.” *Business Survey Methods* Chap 21: 385-401.

Granquist, L. (1997). “The new view on editing.” *International Statistical Review* 65: 381-387.

Granquist, L. and Kover, J.G. (1997). “Editing of survey data: how much is enough?” *Survey Measurement and Process Quality*, Edited by

Lyberg, Chap 18: 415-435.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data quality and record linkage techniques*. Springer.

Hidiroglou, M.A. and Berthelot, J.M. (1986). "Statistical editing and imputation for periodic business surveys." *Survey Methodology* 12: 73-83.

ISTAT, CBS and SFSO (2007). 『Recommended practices for editing and imputation in cross-sectional business surveys』

Johnson, R.A. and Wichern, D.W. (2002). *Applied multivariate statistical analysis*. Prentice Hall.

Kovar, J.G. and Whitridge, P.J. (1995). "Imputation of business survey data." *Business Survey Methods* Chap 22: 403-423.

Lohr, S.L. (1999). *Sampling : Design and Analysis*. Duxbury.

Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust statistics: Theory and Method*. Wiley.

Rencher, A.C. (1995). *Methods of multivariate analysis*. Wiley.

Statistics Canada (2003). Survey methods and practices.

United Nations (1994). *Statistical data editing vol 1. Methods and*