응용통계연구 (2009)
22(4), 697–707

Fellegi-Holt 기법을 이용한 에디팅의 시도 및 분석

이의규[1] · 심규호[2]

[1]통계청 통계개발원, [2]통계청 통계개발원

(2009년 4월 접수, 2009년 5월 채택)

요 약

실제 통계조사에서는 응답 자료라 할지라도 부정확한 응답 등으로 항목 간 연관성 오류가 나타나곤 한다. 이러한 경우 사용자는 상당한 혼란에 빠질 수 있으며 이는 통계자료의 신뢰에 대한 문제이기도 하다. 따라서 특별한 사유가 없이 납득하기 어려운 레코드는 탐색되고 수정되어야 할 필요성이 있다. 이때 어떤 변수를 수정해야 할지를 레코드마다 일일이 결정하는 것은 그리 간단하지 않다. 본 연구에서는 Fellegi-Holt 방법을 이용하여 사업체 조사 자료의 에디팅을 시도하고 그 결과와 문제점을 분석한다.

주요용어: 오류위치포착, Fellegi-Holt 방법, 자동 에디팅.

1. 서론

통계자료 에디팅(statistical data editing)은 자료 수집 및 처리 단계에서 오류를 찾아내고 이를 수정하는 과정을 말한다. 과거 이 에디팅과 관련된 국내 논문은 주택가격동향조사를 위한 데이터 편집 사례연구 (박진우 등, 2005)외 극소수였으나 최근에 들어서는 그 관심이 고조되고 있다. 특히 2007년 통계의 날 기념 워크숍에서는 에디팅 관련 사례들을 소개한 바 있으며 (변종석 등, 2007), 에디팅 품질관리 매뉴얼 (김규성, 2008)이 작성된 바 있다.

에디팅은 작업 방법에 따라 수작업 에디팅과 자동 에디팅(automatic editing)으로 구분하곤 하는데, 컴퓨터의 발전과 더불어 점차적으로 사람의 힘에 덜 의존하는 쪽으로 발전하고 있다 (Granquist, 1997). 한편, 에디팅(editing)은 흔히 임퓨테이션(imputation)과 혼돈하기 쉽다. 주로 응답된 자료의 오류를 찾아 수정하는 것을 에디팅이라 여겨지는 반면에 응답하지 않은 항목 값을 통계적인 방법을 통해 대체하는 것을 임퓨테이션이라 부른다. 그러나 응답한 값이 무시할 정도로 의미가 없어 무응답으로 여겨지는 경우, 에디팅은 임퓨테이션 과정을 수반하게 되어 이 둘을 정확히 구분하기 어렵다. 또한 임퓨테이션을 수행하기 위해서는 이상치에 대한 검토가 선행되어야 하며 임퓨테이션 후에는 다시 에디팅을 해야 하는 등 에디팅과 임퓨테이션은 서로 밀접하게 연관되어 있다. 일반적으로 통계자료 에디팅은 임퓨테이션을 포함하는 자료처리 과정의 포괄적 의미로 해석할 수 있다.

통계청에서는 통계작성기관으로서 이러한 에디팅 과정을 내용검토(또는 줄여서 내검)라 부르며 이를 수행해 왔다(본 논문에서는 에디팅과 내검을 혼용하여 사용하기로 한다). 내용검토의 여러 절차 중 하나의 단계로서 각 조사마다 입력·내검프로그램을 통해 조사표를 입력하게 되는 데, 담당자가 내검규칙을 설정하고 그 규칙에 어긋나는 경우에 해당 에러코드를 자동으로 나타나게 한다. 이후, 자료 내용을 에러코드에 따라 조사표 확인이나 재접촉을 통해 수정 또는 오류의 사유를 기재하여 입력하게 된다.

[1]교신저자: (302-701) 대전광역시 서구 둔산동, 통계청 통계개발원, 통계사무관. E-mail: ekyoolee@nso.go.kr

한편 사업체대상 조사는 금액에 관한 민감한 정보를 담고 있어 정확한 답변을 얻어내기 어렵다. 반면 통계자료 사용자는 통계자료가 완전하고 논리적으로 어떤 문제가 없기를 기대한다. 만약 하나의 레코드(사업체)에서 항목 간 수량적 관계로 볼 때 응답한 값이 무시할 정도로 의미가 없다면 사용자는 상당한 혼란에 빠질 수 있으며 이는 또한 통계자료의 신뢰에 대한 문제이기도 하다. 따라서 응답한 자료라 하더라도 어떤 특정 사유가 없이 비합리적인 값이 나타난다면 적절한 전략에 의거한 수정이 불가피할 경우가 있게 된다. 즉 오류 또는 무응답이 발생한 응답자에 대해 재접촉/재조사가 힘들거나 이를 통해서도 해결되지 않을 때에는 최종적으로 내검규칙을 만족시키지 못한 항목 값은 수정될 필요가 있다.

국외에서는 이미 사업체 대상 조사의 자동 에디팅 시스템이 개발되어 사용되고 있다. 1984년에 개발된 미국의 SPEER (Winkler와 Draper, 1997), 1985년부터 착수하여 개발된 캐나다의 GEIS (Whitridge와 Kovar, 1990), 네덜란드의 CherryPi (Nordholt와 De Waal, 1999)가 대표적이다. 특히 캐나다의 GEIS는 Banff (Kozak, 2005)로 최근 발전되어 사용되고 있다. 이와 같은 배경에서 이의규와 심규호 (2007)는 자료의 자동 오류위치포착 및 수정의 근거가 되는 Fellegi와 Holt (1976) 방법의 원리와 절차를 소개한 바 있다. 본 논문에서는 종사자 4인 이하의 광업·제조업 조사 자료에 Fellegi-Holt(F-H) 기법을 이용한 에디팅을 시도하고 문제점을 분석한다.

2. Fellegi-Holt 기법의 리뷰

De Waal과 Coutinho (2005)는 자동오류포착방법으로 이상치 검색 기법(outlier detection techniques), 신경망(neural networks)에 의한 방법 그리고 수학적 최적화 문제 해결에 근거하는 방법을 제시하였다. F-H 기법은 수학적 최적화에 기초한 대표적인 방법이다. F-H는 대용량 자료의 처리가 가능한 컴퓨터의 발전과 더불어 자동에디팅 문제를 이론적으로 체계화하였으며 이후 각국에서 이를 이용한 자동에디팅 시스템 개발 및 연구 발표가 활발히 진행되어 왔다.

F-H 방법은 조사 자료에 오류가 있는지를 판단하기 위해 조사 담당자가 미리 설정한 내검규칙(edits)을 필요로 한다. 자료의 형태가 범주형 자료(categorical data)인 경우에는 논리적 내검규칙(logical edits)을 부여하고 연속형 자료(continuous data)인 경우에는 산술적인 내검규칙(arithmetic edits)을 선정하여 오류 여부를 판단한다. 만약 레코드가 모든 내검규칙을 만족한다면 오류위치포착은 그 레코드에 필요하지 않다. 그러나 적어도 하나의 내검규칙이 만족되지 않는 경우에는 수정이 요구되는 값을 식별하기 위한 오류위치포착(error localization)의 단계가 필요하다. 그런데 어떤 값이 부정확하고 대체되어야 하는 지의 결정은 그리 단순하지 않다.

이때 어떤 변수를 대체해야 할지를 결정하는 자동화 전략이 필요한데 그것이 Fellegi와 Holt에 의한 전략이다. 이는 주어진 정보를 최대한 보존하면서 모든 내검규칙을 만족하게 하는 최소개의 수정할 변수를 찾아내자는 것이다. 다시 말해 자료의 정보를 수정하는 것은 매우 치명적일 수 있으므로 가능한 정보를 보존해야 한다는 원칙에 따른 것이다.

오류위치포착에 대한 이해를 돕기 위해 하나의 간단한 예를 들어 본다. 다음과 같은 2개의 계량 산술적 내검규칙(quantitative arithmetic edits)이 주어졌다고 하자(각 변수는 음이 아닌 수).

$$E_1 : X_1 - X_2 \geq 0,$$
$$E_2 : X_2 - 3X_3 \geq 0.$$

이제 하나의 레코드가 (6, 4, 8)로 코딩되었다고 하자. 따라서 이 레코드는 두 번째 규칙을 위반한 레코드이다. 문제는 이때 어떤 필드(항목, 변수)를 수정하여야 최대한 정보를 유지하면서 모든 규칙을 만족하게 할 수 있는가이다. 이 경우 X_1이나 X_2만을 바꾸어서는 모든 내검규칙을 만족할 수 없다. 그러나

표 **2.1.** 위배된 내검규칙 행렬

	X_1	X_2	X_3
E_2		1	1
E_3	1		1

X_3를 1로 바꾼다면 모두 만족한다. 물론 두 개 이상의 변수를 모두 바꾸어서 성립이 가능할 수 있으나 최대한 자료를 보존한다는 원칙에서 하나만을 바꾸는 것이 합리적이라는 것이다.

이와 같은 결론은 주어진 내검규칙 E_1과 E_2로부터 변수 X_2의 소거를 통해 다음과 같은 식을 구함으로써 도출될 수 있다.

$$E_3 : X_1 - 3X_3 \geq 0.$$

위의 E_3를 내재적 내검규칙(implicit edits)이라 한다. 위 식에 다시 레코드의 값을 각 규칙에 대입하면 주어진 레코드는 E_2와 E_3의 내검규칙을 만족하지 못하고 있음을 알 수 있다. 따라서 전체 위배된 내검규칙은 E_2, E_3가 된다. 이들 각각에 포함된 변수를 행렬로 표현하면 표 2.1을 얻을 수 있다.

그런데 이 표에서 X_3는 위배된 모든 내검규칙에 포함되어 있음을 볼 수 있다. 즉 명시된 내검규칙으로부터는 어떤 변수를 바꾸어 주어야 할지가 명확하지 않으나 이처럼 추가된 내검규칙을 이용하면 자료의 오류위치를 효율적으로 판단할 수 있다. 더 나아가 X_3값을 미지수로 놓고 나머지 주어진 값을 조건식에 대입하여 풀면 $0 \leq X_3 \leq 4/3$일 때 모든 규칙을 만족하게 된다. 즉, $X_3 = 1$이 가능한 대체값이 될 수 있다.

이 F-H 방법의 가장 큰 특징은 오류자료의 수정할 값을 결정할 때 모든 변수가 동시에 고려된다는 것이다 (Greenberg, 1986). 특히 주어진 편집규칙으로부터 유도된 내재적 편집규칙(implied edits, implicit edits)이 오류자료의 변경할 변수들을 결정할 때 주요한 역할을 한다. 이러한 알고리즘은 If-Then-Else의 구조보다 효율적이고 편집규칙의 수정 또는 변경 시 그 관리가 용이하다 (Chen 등, 2002). 또한 변수의 신뢰성 가중치를 부여할 수도 있어 여전히 유효한 방법으로 보고되고 있다.

그러나 설정된 모든 내검규칙을 필수 규칙(hard edits)으로 간주한다는 것과 오류를 우연적 오류로 국한한다는 것이 단점으로 지적된다. 특히 요구되는 내재적 내검규칙 수가 매우 많을 수 있으며 이때 모든 내재적 규칙의 생성에 있어서 많은 시간이 소요될 수 있다는 것이다. 이러한 단점을 극복하기위해 변수의 비(ratio) 규칙을 이용하는 방법이 제안된 바 있다. 이 방법은 내재적 내검규칙의 수가 선형규칙에 비해 줄어들고 속도는 빨라지나 변수가 모두 비음(non-negative)인 경우에 국한된다 (De Waal과 Coutinho, 2005).

한편 연속형 자료에 대한 F-H 기법 기반의 오류위치포착을 위한 알고리즘은 크게 3가지 방법으로 구분할 수 있다. 첫째는 integer programming 방법으로 자료처리 속도가 다소 느린 한계가 있다. 둘째는 캐나다 통계청의 GEIS에서 채택하고 있는 변형 Chernikova의 알고리즘, 마지막으로 네덜란드 통계청의 변형 Fourier-Motzkin 소거법을 들 수 있다. 사업체조사의 자동에디팅을 위해 필요한 알고리즘에 대한 자세한 정보는 De Waal (2003)의 논문을 참조하기 바란다.

3. 예제 분석

3.1. 예제 자료의 개요

2003년 기준 산업총조사 중 종사자 4인 이하의 광업·제조업 통계조사 자료에 F-H 기법을 적용해 보았다. 종사자 4인 이하의 사업체 자료를 위한 조사표는 업체명, 종사자수 및 급여액, 건물 연면적 등 8개

표 **3.1.** 에러코드와 내검사항(일부)

에러코드	전산내검사항
CD	(2항, 4항) 2항 연간급여액 합계는 4항 ④급여총액과 일치되어야 함
EA	(4항, 5항, 6항) 조사표상의 수입부문의 합계는 비용부문의 합계보다 커야함 5항 제품출하액 + 6항 임가공수입액 < 1.2 × 4항 비용소계 (① + ⋯ + ③)
EB	(4항, 5항, 6항) 5항 제품출하액 > 10 × 4항 비용소계 (① + ⋯ + ③)
RD	(5항, 6항, 7항) 5항 제품출하액 + 6항 임가공수입액 / 조업월수 × 12개월 < 0.1 × 유형자산 연말잔액
QD	(5항, 6항) 소규모 사업체에서 생산하기 어려운 품목이 조사된 경우 (자동차, 선박, 컴퓨터 등)(34, 35, 30)

항목으로 이루어져 있다. 조사표 자료는 입력 자료의 각 항목에 대해 입력·내검 프로그램을 운용하여 검사하는데, 표 3.1과 같은 점검규칙에 따라 내용검토를 하고 있다 (통계청, 2004). 여기서 선택에러(밑줄이 없는 코드)는 에러사유를 기재하면 내검사항에서 해제되는 에러이며, 필수에러(밑줄이 있는 코드)는 나타나면 안 되는 에러이므로 필히 수정을 하는 에러이다. 특히 조사항목 중 4항(연간생산비 소계), 5항(연간 제품 출하액 합계), 6항(연간 임가공수입액 합계), 7항(유형자산 연말잔액)은 금액을 기입하는 항목이며 서로 연관성을 갖는다. 이들 항목과 관련된 선택에러 EA, EB, RD의 점검사항을 명시된 내검규칙으로 설정하고 F-H 기법의 기본원리를 이들 항목에 적용하였다.

적용 자료는 자료 이용의 제약으로 이미 내검과 임퓨테이션이 완료된 자료를 사용하였다. 사업체수는 약 19만개로 집계되었으며 규칙에 어긋난 자료는 거의 없었다. 그러나 선택 내검규칙에서 벗어난 이들 일부분의 자료에 대하여 F-H 기법을 이용한 에디팅을 시도하고 그 유용성과 문제점을 검토하고자 한다.

3.2. 내재적 내검규칙의 생성

명시적 내검규칙으로부터 내재적 내검규칙을 생성하기 위해 종사자 4인 이하 사업체조사의 해당 항목간 수량적 연관규칙을 아래와 같이 기호화 한다. 수입부문의 5항 합계(연간 제품출하액)를 X_1, 6항 합계(연간 임가공 수입액)를 X_2, 비용부문의 4항 소계(연간 원재료비 등)를 X_3, 유형자산의 7항 합계(공장의 자산)를 X_4라 하자. 이제 EA, EB, RD의 각 내검사항을 수식화하면 다음과 같다.

1) 수입부문의 합계는 비용부문의 합계보다 커야 함.

$$X_1 + X_2 \geq 1.2X_3$$

2) 출하액은 원재료비의 10배보다는 작아야 함.

$$X_1 \leq 10X_3$$

3) 수입부문의 합계가 유형자산 연말잔액의 10%보다는 커야 함.

$$X_1 + X_2 \geq 0.1X_4$$

표 3.2. 내검규칙 E_1을 위배한 사업체 일부 (단위: 백만 원)

사업체고유번호	조업월수	주요품목	출하액	임가공수입액	주요생산비	유형자산
165***	12	방충망 샷시	12	0	31 (2, 10)	10
145***	12	산업기계제작	12	0	80 (2, 10)	20
346***	12	의료기기제조	10	0	70 (1, 8)	20

즉, 주어진 명시된 내검규칙(explicit edits)은 다음과 같이 다시 정리할 수 있다.

$$E_1 : X_1 + X_2 - 1.2X_3 \geq 0$$
$$E_2 : -X_1 + 10X_3 \geq 0$$
$$E_3 : X_1 + X_2 - 0.1X_4 \geq 0$$

그리고 이로부터 유도되는 내재적 내검규칙(implicit edits)을 아래와 같이 구할 수 있다.

$$E_4 : X_2 + 10X_3 - 0.1X_4 \geq 0$$

유도되는 내재적 내검규칙은 이외에도 존재하나 실질적으로 필요치 않아 생략하였다. 한편, 내검규칙은 조업월수가 12개월을 가정한 것이다. 만약 조업월수를 고려하면 이 규칙은 다음과 같이 표현할 수 있다.

$$E_3' : \frac{12}{a}(X_1 + X_2) - 0.1X_4 \geq 0,$$

여기서 a는 해당 레코드의 조업월수이다. 즉 조업월수가 6개월이면 해당 레코드의 수입부문에 2배를 해준 값이 유형자산액의 10%보다 커야 함을 의미한다. 따라서 이 식에 따른 내재적 규칙 E_4는 다음과 같이 다시 표현된다.

$$E_4' : \frac{12}{a}(X_2 + 10X_3) - 0.1X_4 \geq 0.$$

3.3. 오류위치포착 및 수정범위

이제 각 레코드가 주어진 내검규칙에 대해 검토된다. 각 실패한 레코드는 하나 이상의 내검규칙을 만족하지 않는다. 어떤 레코드는 두 개 이상의 내검규칙에 어긋날 수 있으며 이때 실패한 내검규칙의 변수를 모두 커버(cover)하는 변수군을 가질 수 있다. 앞에서 언급하였듯이 이는 오류위치 포착단계에서 매우 중요한 정보가 된다. 각 레코드의 위배된 경우에 따라서 오류위치포착 문제를 구체적으로 살펴본다.

3.3.1. E_1을 위배한 경우 수입부문의 합계는 주요생산비의 1.2배보다 커야한다는 내검규칙 E_1에 위배되는 건수가 594건으로 나타났다 (표 3.2 참조). 다시 말해 이들 레코드는 출하액과 임가공수입액을 합한 금액이 주요생산비 수준보다도 적은 레코드를 말한다. 규칙에 어긋나는 특별한 사유가 기재되면 내검에 통과되나 만약 재접촉이 불가능한 자료에 대해서는 자동내검 및 수정을 고려할 필요가 있다.

당연하게 수입부문이나 비용부문 둘 중 하나를 선택하여 수정한다. 그러나 수입부문이 유형자산의 10%보다 커야한다는 조건(E_3)은 성립하므로 수입부문과 유형자산은 일관성을 유지한다. 따라서 주요생산비를 검토하는 전략이 적절하다. 표 3.2에서 생산비의 괄호 속에 내검규칙을 만족하게 하는 가능한 범위를 나타내었다. 이는 생산비를 미지수로 놓고 나머지 항목 값을 각 내검규칙에 대입하여 얻은 결과이다.

표 3.3. 내검규칙 E_2를 위배한 사업체 일부 (단위: 백만 원)

사업체고유번호	조업월수	주요품목	출하액	임가공수입액	주요생산비	유형자산
217***	12	산업기계	100 (1, 10)	0	1	10
233***	12	철판절단	150 (10, 10)	0	1	100
165***	12	노트북가방	389 (1, 20)	0	2	6

표 3.4. 내검규칙 E_3를 위배한 사업체 일부 (단위: 백만 원)

사업체고유번호	조업월수	주요품목	출하액	임가공수입액	주요생산비	유형자산
167***	12	여성정장	150	0	47	4,011 (1,500)
201***	12	콘크리트 블록	130	0	72	2,701 (1,300)
386***	12	콘테이너 제조	60	0	47	1,870 (600)
357***	9	타월 제조	50	0	24	1,058 (667)

3.3.2. E_2를 위배한 경우 출하액은 원재료비의 10배보다 작아야한다는 내검규칙 E_2에 위배되는 경우가 824건으로 나타났다. 이 조건을 만족하지 못한 레코드는 출하액이 주요생산비의 10배보다도 많은 레코드로서 주요생산비에 비해 과다한 출하액을 보이는 사업체이다. 이들 사업체들 중에서 소금(천일염) 사업체가 627개로 대부분을 차지한다. 따라서 이들 품목을 생산하는 사업체의 내검기준이 완화될 필요가 있다. 또한 주요품목이 도장과 같이 주요생산비가 0(백만 원)으로 기입된 경우는 출하액이 1(백만 원)이라도 규칙을 위반하게 되므로 주요생산비가 0이면서 출하액이 작은 영세사업체 또한 내검에서 제외되어야 할 것이다.

표 3.3에서 보면 주요생산비와 유형자산은 일관성을 띠므로 특정품목과 영세사업체를 제외하고는 출하액을 우선 검토하는 것이 바람직하다(또는 임가공업체로서 임가공수입액에 기재되어야 할 가능성도 있음). 출하액의 가능한 범위는 괄호안에 제시하였다.

3.3.3. E_3를 위배한 경우 수입부문의 합계가 유형자산 연말잔액의 10%보다는 커야한다는 내검규칙 E_3에 위배되는 건수가 498건으로 나타났다. 표 3.4는 내검규칙 E_3를 위배한 사업체의 일부이다. 조건을 만족하지 못한 레코드는 유형자산에 비해 수입부문의 금액이 작은 경우에 해당한다. 이 규칙에 위배된 레코드의 유형을 살펴보면, 많은 유형자산을 요구하는 쌀(벼)(도정), 떡, 고추(고춧가루), 참·들기름, 건강보조용 즙(액) 등 곡물 등을 빻고, 찧고, 짜내는 업종이 대부분을 차지한다. 이들 품목과 관련된 레코드에 대해서는 사전에 내검에 걸리지 않도록 조치할 필요가 있다.

여기서는 수입부문이나 유형자산 중 하나를 수정한다. 그러나 주요생산비가 함께 낮게 조사되어 적은 주요생산비에 수입부문의 금액이 작은 것은 일관되어 보인다(E_2 성립). 따라서 이 규칙을 위배하는 품목유형과 관련된 사업체 이외에는 유형자산에 대한 검토가 우선적으로 고려되어야 할 것이다. 내검기준을 만족하기 위한 유형자산의 허용 최대값을 괄호 속에 나타냈다.

3.3.4. E_1과 E_3를 위배한 경우 수입부문의 합계가 주요생산비의 1.2배보다 커야한다는 내검규칙 E_1과 수입부문의 합계가 유형자산 연말잔액의 10%보다는 커야한다는 내검규칙 E_3를 동시에 위배한 건수는 66건으로 나타났다 (표 3.5 참조). 이 경우에는 주요생산비 하나를 수정할 때에는 다른 내검규칙을 만족시킬 수 없으며 유형자산을 수정하더라도 다른 한 쪽 내검에 걸릴 수 있다. 물론 두 개의 항목을 수정하면 모든 내검규칙을 만족할 수 있으나 공통으로 들어간 수입부문을 수정하는 것이 가능한 정보를 보존한다는 입장에서 합리적이다 (표 3.6 참조).

예를 들면, 컴퓨터 부품잉크를 생산하는 사업체는 (1, 0, 22, 267)로 코딩되었다. 즉 2억6천7백만 원의

표 **3.5.** 내검규칙 E_1과 E_3를 위배한 사업체 일부 (단위: 백만 원)

사업체고유번호	조업월수	주요품목	출하액	임가공수입액	주요생산비	유형자산
360***	12	컴퓨터 부품잉크	1 (27, 220)	0	22	267
515***	12	멸치액젓	20 (108, 900)	0	90	1,050
506***	12	유니폼	1 (24, 200)	0	20	130

표 **3.6.** E_1과 E_3에 위배된 내검규칙 행렬

	X_1	X_2	X_3	X_4
E_1	1	1	1	
E_3	1	1		1

표 **3.7.** 내검규칙 E_2와 E_4를 위배한 사업체 일부 (단위: 백만 원)

사업체고유번호	조업월수	주요품목	출하액	임가공수입액	주요생산비	유형자산
513***	12	원격조명릴	189	0	0 (19, 157)	350
198***	12	재단판	86	0	4 (9, 72)	605
373***	12	치아제조	27	0	1 (3, 22)	110

표 **3.8.** E_2과 E_4에 위배된 내검규칙 행렬

	X_1	X_2	X_3	X_4
E_2	1		1	
E_4		1	1	1

유형자산을 갖는 사업체가 주요생산비 2천2백만 원을 들여 출하액이 1백만 원인 경우이다. 이때 출하액이 주요생산비와 유형자산에 비해 매우 작은 자료로 어떤 특정사유가 없다면 잘못 보고되거나 기재될 가능성이 크므로 수정되는 것이 바람직할 것이다. 각 내검규칙에 출하액을 미지수로 놓고 레코드의 나머지 값을 대입하면 출하액은 2천7백만 원 이상 2억2천만 원 이하의 출하액 범위를 갖게 되고 이 범위 값 안에서 모든 내검규칙을 만족한다.

3.3.5. E_2와 E_4를 위배한 경우 출하액은 원재료비의 10배보다 작아야한다는 E_2 규칙과 생성된 내검규칙 E_4를 동시에 위배한 경우는 109건이다 (표 3.7 참조). 이 규칙들에 위배된 대부분의 자료는 주요생산비가 출하액이나 유형자산에 비해 매우 작은 경우(없거나 4백만 원 이하)로 나타났다. 따라서 유형자산이나 출하액이 작으면서 주요생산비가 0이면 허용되도록 한다. 특히 소금과 잘신은 원재료비가 거의 없고 인력에 의해 제조되므로 주요생산비가 적어도 허용되도록 한다. 특정사유가 없는 경우에는 수정이 필요한데 표 3.8에서와 같이 주요생산비가 공통으로 두 규칙을 커버하므로 주요생산비를 수정하는 쪽이 바람직하다.

3.3.6. E_3와 E_4를 위배한 경우 수입부문의 합계가 유형자산 연말잔액의 10%보다 커야한다는 내검규칙 E_3와 생성된 내검규칙 E_4에 동시 위배된 건수는 78건으로 나타났다 (표 3.9 참조). 즉 수입부문과 주요생산비가 유형자산에 비해 특이하게 작은 경우이다. 그러나 수입부문의 출하액과 주요생산비가 모두 0인 경우가 많은데 주요생산비가 0이어도 되는 임가공 업체나 유형자산 역시 작은 경우에는 예외로 한다.

표 3.10에서 보듯이 임가공 수입액 또는 유형자산 연말잔액이 두 규칙을 커버함으로써 임가공수입액이나 유형자산 연말잔액을 수정하는 것이 합리적이다. 그런데 출하액과 주요생산비가 작은 경우에는 유형

표 **3.9.** 내검규칙 E_3와 E_4를 위배한 사업체 일부 (단위: 백만 원)

사업체고유번호	조업월수	주요품목	출하액	임가공수입액	주요생산비	유형자산
50***0	12	알루미늄 절단	0	9	2	1,040 (290)
50***7	12	절삭가공	0	45	1	2,721 (550)
164***	12	호박, 매실즙	0	15	2	1,022 (350)
513***	12	사출제품	470	0	219	38,651 (4,700)

표 **3.10.** E_3와 E_4에 위배된 내검규칙 행렬

	X_1	X_2	X_3	X_4
E_3	1	1		1
E_4		1	1	1

표 **3.11.** E_1, E_3, E_4에 위배된 내검규칙 행렬

	X_1	X_2	X_3	X_4
E_1	1	1	1	
E_3	1	1		1
E_4		1	1	1

표 **3.12.** E_2, E_3, E_4에 위배된 내검규칙 행렬

	X_1	X_2	X_3	X_4
E_2	1		1	
E_3	1	1		1
E_4		1	1	1

자산이 하향 조정되는 것이 설득력이 있다. 물론 임가공수입이 없는 사업체인 경우에는 유형자산 연말 잔액 항목을 수정함이 바람직하다.

3.3.7. E_1, E_3, E_4**를 위배한 경우** E_1, E_3, E_4에 위배된 레코드는 모두 2건이다. 이때는 표 3.11에서와 같이 임가공수입액이 존재하면 임가공 수입액을 수정하는 것이 정보의 손실을 최소화한다. 그러나 출하액만 존재한다면 이때는 하나의 변수가 모든 내검규칙을 커버하지 못하기 때문에 최소한 두 개를 바꾸어야 한다.

예를 들면, 이 규칙들을 위배한 남성복 맞춤 사업체 (403***)는 (1, 0, 1, 234)으로 유형자산 234백만 원, 출하액이 1백만 원, 주요생산비가 1백만 원이다. 이때 출하액이 유형자산의 10%보다 크도록 출하액을 2와 10사이의 값 그리고 유형자산을 100보다 작은 값으로 수정하면 모든 내검규칙을 만족하게 된다.

3.3.8. E_2, E_3, E_4**를 위배한 경우** E_2, E_3, E_4에 위배된 레코드는 모두 4건이다. 표 3.12를 보면 최소 두 개의 항목 값을 바꾸어야 한다. 그러나 X_2와 X_4 두 개를 바꾸어도 모든 내검규칙을 커버하지 못하므로 이 쌍은 제외된다. 이들 규칙을 위배한 레코드 중 하나인 자동차 소음방지 부품업체(366***)는 유형자산 13억2천5백만 원, 매출액 2천만 원, 주요생산비가 1백만 원이다. 이때는 유형자산을 수정해도 주요생산비와 매출액 간 내검규칙을 만족할 수 없고 매출액을 수정해도 주요생산비와 유형자산 간 내검규칙을 만족할 수 없으며 주요생산비를 수정해도 수입부문과 유형자산 간 내검규칙을 만족할 수 없다. 따라서 주요생산비와 매출액, 주요생산비와 유형자산, 매출액과 유형자산 중 하나의 변수 쌍을 수정하여

야 모든 내검규칙을 만족하게 된다. 예를 들어 생산비와 유형자산을 택하였을 경우, 생산비는 2와 16의 범위에 있으면서 유형자산은 200보다 작을 때 모든 내검규칙을 만족한다.

3.4. 분석결과

결과로부터 하나의 레코드가 명시적 내검규칙이나 생성된 내검규칙을 두 개 이상 위배한 경우, 위배된 내검규칙에서 공통변수를 찾음으로써 최소의 정보손실로 모든 내검규칙을 만족하는 항목을 선택할 수 있었다. 그러나 앞의 결과에서처럼 수정해야할 변수가 유일하지 않을 경우가 있다. 이 경우 각 항목의 응답 신뢰도를 고려하여 수정해야 할 변수를 선택해야 한다. 하나의 내검규칙을 위반한 경우에는 당연하게 수정해야 할 변수가 유일하게 결정되지 않는다. 이 경우는 주로 위배하는 품목의 정보와 만족하는 다른 내검규칙을 고려하는 전략을 취한다.

본 논문에서 적용된 자료는 이미 내검이 완료된 자료이다. 특히 여기서의 에러는 선택에러로 내검에 걸린 모든 자료가 오류를 의미하는 것은 아니다. 그러나 오류 응답일 가능성이 높은 자료에 오류위치 포착의 기본원리를 적용하여 그 유용성을 검토하였다. 만약 에디팅 전 자료에 이 오류위치 포착기법을 적용한다면 수작업 에디팅 시 어떤 항목을 우선 검토해야 하는 지의 정보를 줄 수 있을 것이다.

4. 결언

조사 자료는 철저한 내검에도 불구하고 응답자의 부정확한 응답, 단위착오 등 여러 가지 원인으로 잘못된 정보가 포함될 가능성은 여전히 존재할 수 있다. 특히 재접촉/재방문의 어려움 발생 등 조사환경이 악화될 경우, 주어진 자료에서만 오류를 찾아내고 주어진 정보를 통해서만 자료를 대체하여야 하므로 자동 오류위치포착과 수정은 더욱 필요하다. 이와 같은 이유로 재접촉 및 재조사를 가급적 금하는 국가에서는 자동내검이 발달되어 있다. 한편 우리나라의 조사환경도 점차 변화하고 있어 이에 대응할 수 있는 자동내검에 대한 연구가 필요하다고 판단된다.

현재로서는 오류위치를 포착해서 수정하는 것에 대한 우려가 적지 않다. 즉 담당자에 의해 주어진 내검규칙을 벗어나는 자료가 실제로 맞는 응답일 수 있기 때문이다. 그러나 이러한 경우는 매우 적으며 이미 여러 논문에서 자료에 대한 과도한 에디팅으로 인한 문제점이 지적된 바 있다. 즉 조사의 시의성 측면에서나 비용 측면에 있어서 효율성을 고려할 필요가 있다. 물론 종사자 5인 이상의 사업체조사의 경우, 일정 규모 이상은 세밀한 수작업 에디팅을 실시하여(선택적 에디팅) 자동에디팅의 단점을 극복할 수 있다.

또 하나의 문제점으로는 현재의 내검규칙이 조사표 내검을 위해 설정된 선택적인 내검규칙으로서 이를 자동수정에 적용하는 것은 다소 무리인 부분이 있다. 이는 필수적인 내검규칙을 개발하거나 선택적인 내검규칙의 조건들을 더욱 관대하게 재설정함으로써 개선될 수 있을 것이다. 즉 좀 더 강한 내검규칙을 부여하고 이를 위배한 경우에만 자동 수정을 고려한다면 위험을 최소화할 수 있으며 일관되고 합리적인 전략으로 대처할 수 있을 것이다. 한 가지 간과해서 안 될 것은 이와 같은 내검규칙의 설정을 위해서는 담당자의 경험과 이상치분석 등 선행 연구가 필요하다는 것이다.

자동에디팅은 자료수정을 위한 작업이라기보다는 자료의 품질을 효율적으로 관리하는 도구로 인식하여야 할 것이다. 본 연구에서는 명시된 내검규칙으로부터 내재적 내검규칙을 생성하여 위배된 내검규칙의 변수를 공통적으로 커버하는 변수를 찾는 단순한 방법을 이용하였으나 추후 선형계획법을 이용한 오류위치포착기법을 적용하고자 한다. 끝으로 본 보고서를 통해 향후 데이터 에디팅 분야의 연구가 더욱 더 활성화되기를 기대한다.

참고문헌

김규성 (2008). 에디팅 품질관리 매뉴얼, <한국통계학회>.

박진우, 박현주, 김진억 (2005). 주택가격동향조사를 위한 데이터편집 사례연구, <조사연구>, **6**, 83–98.

변종석 (2007). Introduction to Data Editing, Data Editing in Survey, 2007 통계의 날 기념 워크숍, <한국조사연구학회>.

이의규, 심규호 (2007). 사업체 대상 조사의 자동내검기법, <국가통계발전을 위한 통계기법의 개선>, 통계개발원, 81–100.

통계청 (2004). <2003년 기준 산업총조사 입력·내검 프로그램 운영요령서>, 통계청 내부자료.

Chen, B., Thibaudeau, Y. and Winkler, W. E. (2002). A comparison study of ACS if-then-else, NIM, and DISCRETE edit and imputation systems using ACS data, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

De Wall, T. (2003). *Processing of Erroneous and Unsafe Data*, Ph. D. Thesis, Erasmus University Rotterdam.

De Waal, T. and Coutinho, W. (2005). Automatic editing for business surveys: An assessment of selected algorithms, *International Statistical Review*, **73**, 73–102.

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

Granquist, L. (1997). The new view on editing, *International Statistical Review*, **65**, 381–387.

Greenberg, B. (1986). The Use of Implied Edits and Set Covering in Automated Data Editing, Bureau of the Census, Statistical Research Division Report Series SRD Research Report Number: Census/ SRD/ RR-86/02.

Kozak, R. (2005). The Banff system for automated editing and imputation, *Proceedings of the Survey Methods Section*, SSC Annual Meeting.

Nordholt, E. S. and De Waal, T. (1999). Automatic Editing in the Dutch Labour Cost Survey Using CherryPi, UN Statistical Commission and Economic Commission for Europe, Working Paper 7.

Whitridge, P. and Kovar, J. (1990). Applications of the Generalized Edit and Imputation System at Statistics Canada, Statistics Canada.

Winkler, W. E. and Draper, L. R. (1997). The SPEER edit system, statistical data editing, *UN Economic Commission for Europe*, **II**, 51–55.

A Trial of Data Editing Using Fellegi-Holt Techniques and Its Analysis

Euikyoo Lee[1] · Kyuho Shim[2]

[1]Statistics Research Institute, KNSO; [2]Statistics Research Institute, KNSO

(Received April 2009; accepted May 2009)

Abstract

In actual statistical surveys, the inconsistencies within the record are often occurred due to incorrect response. The users may be confused and statistical agencies may have a problem of reliability on statistical data in this case. It is needed to detect and correct the unconvinced record without any special reasons. However, it is not simple to determine which item should be corrected in every failed record. In this paper we briefly introduce Fellegi-Holt method, apply to a business survey, and then discuss the problems for this trial editing.

Keywords: Error localization, Fellegi-Holt method, automatic editing.

[1]Corresponding author: Statistician, Statistics Research Institute, KNSO, Dunsan-dong, Seo-gu, Daejeon 302-701, Korea. E-mail: ekyoolee@nso.go.kr