사업체대상 조사의 무응답 대체기법 연구
(도·소매업조사 중심으로)

결과보고서

2007. 11. 29.

연구 수행 기관 : 연 세 대 학 교

연구 책임자 : 김 재 광
공동 연구자 : 신 동 완
공동 연구자 : 최 필 근
연구 보조원 : 정 상 아
연구 보조원 : 김 소 영

차 례

표 차례

그림 차례

Ⅰ. 연구개요

1.1. 연구 목표

본 연구는 도소매업 조사의 특성에 맞는 통계적으로 유효하면서도 효율적인 무응답 처리 기법의 개발을 통하여 무응답 대체에 대한 학술적 및 실용적 가치가 높은 방법론을 연구 개발하여 통계청에서 취급하는 다른 사업체 조사에서의 무응답 대체 방법론에도 적용할 수 있도록 하는 것을 목표로 한다.

1.2. 연구 범위

본 연구의 근본 취지는 도소매업 조사의 무응답 처리 방법에 대한 방법론 개발이다. 도소매업 자료 중에 가장 중요한 무응답 형태는 본지사 자료이다. 그러나 다수의 지사가 있는 경우 회계상 또는 다른 이유로 본사단위로 사업실적이 집계되고 지사단위로는 실적이 파악되지 않고 무응답 되는 경우가 많다. 그러나 도소매업 조사의 기본 단위는 사업체이고 각 지사가 하나의 사업체로 구분되기 때문에 본사의 총합 실적으로부터 각 지사의 사업실적의 무응답 대체가 중요한 문제이다.

따라서 다음의 두 가지 문제에 초점을 맞추어 연구를 진행하였다.

(1) 해외 사례 연구: 미국과 캐나다 등의 통계 선진국의 경우 사업체 조사에서의 무응답 처리 방법론에 대한 사례를 소개하고 그에 대한 이론적 배경을 설명하였다.
(2) 본.지사 매출액 관련 무응답 처리 연구 : 사업체에서 본사의 매출액이 있고 지사의 매출액을 알지 못하여 무응답 되는 경우 종업원 수와 같은 보조 정보를 이용하여 지사의 매출액을 추정하여 배분하는 방법론에 대한 연구

따라서 본 보고서에서는 위의 두 가지 주제에 대한 연구 내용과 연구 결과들을 기술하게 될 것이다.

1.3. 연구의 기본 방향

위에서 기술한 두 가지 연구 주제에 대한 연구 방법론을 개발하고자 할 때 본 연구진은 다음과 같은 연구의 기본 방향을 설정하였다.

(1) 연구 방법론의 합리성 : 주어진 자료와 관측된 변수들을 최대한 잘 활용하는 통계적 모형을 사용하여 무응답 처리 방법론을 개발하되 그 모형 개발에 사용되는 여러 가정과 결론을 도출하는 과정이 합리적이 될 수 있도록 하였다.

(2) 연구 방법론의 적절성 : 선택된 통계적 모형이 자료를 얼마나 잘 설명해 주는가의 여부에 대한 객관적이고 과학적인 평가를 통하여 제안하고자 하는 방법론의 적절성을 보장하고자 하였다.

(3) 연구 방법론의 용이성 : 최종적으로 선택된 방법론은 실무자들이 쉽게 사용할 수 있는 방향으로 개발되어 도소매업 조사 자료의 무응답 처리에 간편하게 적용될 수 있도록 하였다.

1.4. 연구의 한계점

모든 연구가 그렇듯이 본 연구도 어려 가지 중요한 한계점과 문제점을 가지고 있다. 이 중, 어떠한 부분들은 어쩔 수 없는 부분도 있고 또 다른 부분은 반론의 여지도 있는 부분이지만 이러한 한계점과 문제점에 대한 것들을 명확히 밝힘으로써 앞으로의 유사한 연구의 계획과 진행에 반영될 수 있고자 한다.

(1) 연구 기간의 제한: 무응답 처리 연구는 통계학 응용 분야에서도 난이도가 매우 높고 까다로운 분야로 외국의 경우에서도 수년간 이상의 연구를 통해 계속 보완해 나가는 실정이나 본 연구 과제의 경우 상대적으로 단기간의 연구 기간이 주어진 상태이므로 상대적으로 충분한 연구가 이루어지지 못한 상태에서 본 결과 보고서가 작성되었다. 따라서 이 결과 보고서가 제공하는 연구 결과들은 이러한 시간적 제한 속에서 이루어졌다.

(2) 자료의 제한 : 지사의 매출액 imputation 모형 개발에 사용된 자료는 14개 기업체의 본.지사 자료 뿐이었다. 이 중 1개의 이상값(outlier)을 제외하면 불과 13개의 본사 자료만으로 분석을 해야 하였고 그 자료들마저 비밀보호 차원에서 많은 정보들이 삭제된 상태에서 제공을 받게 되었다. 따라서 이 자료들로부터 각 지사들의 매출에 대한 정확하고 신뢰성 있는 모형을 얻어내는 것은 쉽지 않은 일이고 특히 이 자료가 전체 모집단에서 얻어진 랜덤 표본이라는 보장이 없었으므로 이 자료의 모집단 대표성에 대해 정당화하기 힘든 것도 사실이다. 보

다 신뢰성 있는 연구 결과를 얻어내기 위해서는 좀 더 많은 자료들을 얻을 수 있었으면 하는 아쉬움이 있다.

(3) 부실 보고에 대한 처리의 한계: 아무리 훌륭한 통계적 모형과 기법을 사용한다고 하여도 자료 수집이 부실하여 허위 또는 부실 보고가 이루어진 자료를 이용하여 분석하는 데에는 당연히 한계가 있을 것이다. 이러한 부실 보고는 일종의 측정오차(measurement error)로 볼 수 있으며 이러한 측정 오차에 대한 연구도 향후 연구에 필요할 것으로 판단되나 본 연구에서는 여러 가지 사정으로 진행하지 않았다.

Ⅱ. 사업체 조사 무응답 대체 해외 사례 연구

2.1 서론

사업체를 표본 추출 단위 및 조사 단위로 하는 사업체 조사에서 무응답 대체 방법론을 개발하는 일은 통계학의 깊이 있는 이론뿐만 아니라 다년간의 축적된 경험과 실무적 지식을 바탕으로 이루어져야 한다. 따라서 본 연구에서는 이에 대한 연구의 시작으로써 무응답 대체에 대한 기본 이론을 소개하고 관련된 유사 조사에서의 해외 사례를 연구하여 통계 실무자들이 무응답 대체에 대한 기본 이해를 높이고 국제적인 기준을 숙지하고자 하는 것을 본 연구의 목표로 한다.

이를 위하여 먼저 무응답 대체와 관련된 여러 가지 이슈들과 기본적인 내용들을 설명하고 이와 관련된 여러 가지 실무적 통계학적 방법론들의 이론적 배경을 설명한다. 그 다음으로는 해외 사례를 소개하는데 미국과 캐나다 등지에서 현재 실시되는 몇 개의 사업체 조사에 대하여 그 현황을 비교적 상세하게 소개하였다.

2.2 무응답 대체의 기본 이론

거의 대부분의 표본조사에서는 무응답이 발생한다. 무응답은 조사 단위 자체가 응답을 하지 않는 단위무응답(unit nonresponse)과 일부 항목에 대해서만 응답을 하지 않는 항목무응답(item nonresponse)으로 나누어질 수 있다. 이러한 무응답의 처리 방법으로는 단위무응답의 경우 재조사(call-back 또는 follow-up survey)나 무응답가중치조정(nonresponse weighting adjustment)을 통하여 처리하고 항목 무응답의 경우에는 결측값대체 (imputation)를 통하여 처리한다.

무응답이 미치는 영향을 알아보기 위하여 다음과 같은 자료 구조를 생각해 보자.

층	모집단 크기	모평균	표본 크기
응답자	N_R	$\overline{Y_R}$	n_R
무응답자	N_M	$\overline{Y_M}$	n_M
모집단	N	$\overline{Y}$	n

즉, 전체 모집단은 응답자들로 이루어진 모집단들과 무응답자들로 이루어진 모집단

으로 구성되어 있을 것이다. 만약 단순임의추출로 전체 모집단에서 표본을 추출한다면 표본 내의 응답자에 대해서만 관측할 수 있을 것이다. 이 때 얻어지는 응답자 평균을 $\overline{y_R}$ 이라고 하고 이 응답자 평균을 사용하여 전체 모집단 평균을 추정한다고 할 때 다음의 결과를 얻게 된다.

$$Bias(\overline{y_R}) \doteq \frac{N_R}{N}(\overline{Y_R} - \overline{Y_M})$$

$$Var(\overline{y_R}) \doteq \frac{1}{n_R} S_R^2$$

여기서 두 가지 문제점이 발생하게 된다. 하나는 편향된 추정을 한다는 것이다. 편향이 0이 되는 경우는 응답자의 평균과 무응답자의 평균이 같게 되는 경우뿐이다.

또 다른 문제점은 표본수의 감소($n_R < n$)로 인한 추정의 효율 저하이다. 이렇게 발생하는 편향을 보정하고 감소된 효율을 다시 높이고자 하는 것이 무응답 처리의 기본 방향이다.

항목무응답에 대한 처리 방법으로 결측값대체법(imputation)이 많이 사용된다. 결측값대체는 결측(missing)된 자료에 값을 넣어줌으로서 여러 다른 사용자들이 그 자료를 사용하여도 일관된 점추정값을 구현할 수 있도록 해주는데 일차적인 목표가 있다. 만약 결측된 자료 자체를 사용자들에게 알아서 분석하도록 한다면 여러 다른 사용자들이 각기 다른 분석 결과를 얻어낼 수가 있고 이는 공식 통계의 기본 원칙 중의 하나인 One number principle(하나의 모수에 하나의 추정치를 생산하자는 뜻)에 위배되기 때문이다. 만약 One number principle이 지켜지지 않아 하나의 관심모수에 여러 개의 추정치가 공존하게 된다면 통계 자체의 신뢰성이 크게 훼손될 것이다. 그 외 결측값대체의 또다른 목적으로는 무응답 편향을 보정하고 또한 관측된 다른 정보들을 결측값대체에 반영하여 추정의 효율을 높이는 것 등이 있다.

그렇다면 어떻게 결측값대체를 하는 것이 좋은 방법일까? 이해를 돕기 위하여 다음의 예를 생각해 보자.

< Example >

초모집단 모형이 이변량 정규분포모형을 따르는 유한모집단에서 단순임의추출로 자료를 추출하였다고 하자. 이 경우 n개의 표본자료 (x_i, y_i) 는 다음의 이변량 정규분포모형을 따를 것이다.

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx}\ \sigma_{xy} \\ \sigma_{xy}\ \sigma_{yy} \end{pmatrix}\right]$$

.

또한 x_i는 모두 관측되지만 y_i는 첫 번째 $r(<n)$개의 관측치에 대해서만 관측된다고 하자. 이 경우 y_i의 응답확률이 x_i에는 의존할 수 있지만 y_i에는 의존하지 않는다고 가정하면

$$x_i \sim N(\mu_x, \sigma_{xx})$$

$$y_i | x_i \sim N(\beta_0 + \beta_1 x_i, \sigma_{ee})$$

으로 모형을 표현할 수 있고 이 경우 $\beta_0 = \mu_y - \beta_1 \mu_x$, $\beta_1 = \sigma_{xy}/\sigma_{xx}$, 그리고 $\sigma_{ee} = \sigma_{yy} - \beta_1^2 \sigma_{xx} = \sigma_{yy}(1-\rho^2)$ 으로 표현된다. 이러한 모형 하에서 결측자료에 대하여 최적예측치(best predictor)로 결측치대체를 실시하면 결측된 y_i의 대체값은

$$\hat{y_i} = \overline{y_r} + (x_i - \overline{x_r})\hat{\beta_1} \qquad (1)$$

으로 표현되고 이 경우 $(\overline{x_r}, \overline{y_r})$ 는 첫 r개의 응답 자료만을 이용하여 얻어진 (x_i, y_i)의 표본 평균이 될 것이다. 이렇게 대체(imputation)된 자료를 바탕으로 μ_y를 추정하면

$$\widehat{\mu_{yI}} = \frac{1}{n}\left\{\sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y_i}\right\}$$

으로 표현되고 이는 최대우도추정량(maximum likelihood estimator)과 동일해진다. 이 경우 추정량의 분산은

$$Var(\widehat{\mu_y}) \doteq \frac{1}{n}\sigma_{xx}\beta_1^2 + \frac{1}{r}\sigma_{ee} = \frac{\sigma_{yy}}{r}\left[1 - \left(1 - \frac{r}{n}\right)\rho^2\right]$$

이 된다. 따라서 $1-(1-r/n)\rho^2$은 무응답 대체 단계에서 전체 표본에서 관측된 x 를 반영하여 줌으로써 얻어지는 효율의 증가분이 될 것이고 이러한 결측값 대체법은 매우 효율적인 추정을 구현한다고 할 수 있을 것이다.

그러나 이 결측값대체법은 σ_{yy}의 추정에 대해 비편향추정량을 구현해 주지 않는다. 즉,

$$E(\widehat{\sigma_{yyI}}) = \sigma_{yy} - \frac{n-r}{n}\sigma_{yy}(1-\rho^2) < \sigma_{yy}.$$

따라서 식 (1)의 회귀 대체법은 μ_y에 대해서는 최적이지만 σ_{yy}에 대해서는 과소 추정을 유발한다.

대체를 하고자 할 때 크게 두 가지 시나리오를 생각해 볼 수 있다. 하나는 나중에 분석하게 될 관심모수를 미리 아는 경우이고 다른 하나는 그렇지 않게 되는 경우일 것이다. 전자의 경우에는 그 관심모수에 적합하도록 결측대체값을 적절히 결정할 수 있다. 예를 들어서 모평균의 추정만이 관심이면 (1)과 같은 회귀 대체법을 사용하면 되지만 만약 모평균 뿐만 아니라 모집단 분산 등과 같은 다른 모수의 추정에도 관심이 있다면 (1)의 예측치(predictor)에 잔차(residual)을 합쳐서 y 의 주변 분포를 보존하도록 하게 된다. 이렇게 특정 모수에 대해서 뿐만 아니라 여러 가지 모수에 대해 추정을 하게 될 것을 고려하여 범용(general purpose)무응답 대체를 하는 것은 원자료를 공표하는 경우에 사용하게 될 결측값대체법의 이상적인 목표라 할 수 있을 것이다.

결측치 대체의 방법은 크게 확률적 대체법(stochastic imputation)과 결정론적 대체법(deterministic imputation)으로 나누어 진다. 결정론적 결측치 대체는 주어진 응답자의 자료에 대해서 오직 하나의 대체값이 가능하다는 것을 의미한다. 확률적 대체는 대체값의 결정 과정에 랜덤하게 결정되는 부분이 있다. 만약 같은 자료에서 결측치 대체가 반복된다면, 결정론적 방법은 각 경우마다 같은 값으로 대체되는 반면, 확률론적 방법은 상이한 값으로 대체된다.

결측치 대체의 방법에는 다음과 같은 것들이 있다.

모 형	결정론적 대체법	확률적 대체법
	연역적 대체	
Cell mean model	평균값 대체	무작위 핫덱 대체 (Random hot deck imputation)
Ratio model	비추정 대체	Ratio random 대체
Regression model	회귀 대체	Regression random 대체
Nonparametric regression model	최근방 이웃 대체 (Nearest neighbor imputation)	

2.2.1. 연역적 대체

연역적 결측치 대체는 결측 또는 비일관된 값이 실제 값으로 추리될 수 있다고 보는 방법이다. 종종 이는 조사표 상의 다른 항목에 주어진 응답 패턴에 근거한다. 보통 연역적 결측치 대체는 가장 먼저 수행되는 방법이다. 예를 들어 네 항목의 합에서, 만약 두 항목의 값이 각각 60과 40인데 그리고 나머지 두 항목이 빈 칸으로 나와 있고 총합이 100이라면 2개의 결측값은 0이라는 것을 유추해 낼 수 있다.

2.2.2. 평균값 대체

평균값 결측치 대체에서는 결측값이나 비일관된 값이 대체군내에서 평균값으로 대체된다. 예를 들어, 주택 조사의 조사표가 한 아파트 한달 월세값을 가지고 있지 않다고 가정한다. 이 결측값은 월세를 정확하게 답변한 응답자들의 한달 평균 값에 의해 대체될 수 있다. 대체군은 대체를 필요로 하는 조사표와 같은 범주에 있는 것으로 판단되는 응답자로 구성된다. 사업체 조사의 경우에는 동일한 유형(법인, 개인사업자), 업종(중분류, 또는 소분류), 규모(종업원수 기준)을 바탕으로 대체군을 형성하는 것이 일반적이다.

평균대체는 동일한 무응답 가중치 조정을 동일한 결측치 대체군(Imputation cell)에 있는 모든 응답자에게 적용시키는 것과 같다. 무응답은 균등한 것으로 그리고 무응답자는 응답자와 유사한 특성을 가진 것으로 가정한다. 평균값 결측치 대체가 합리적인 점추정(총합의 추정, 평균의 추정, 등등)을 하는 반면, 이는 군 평균에 인위적인 못질을 가함으로써 분포 및 항목들 다변량간의 관계를 파괴한다. 이 인위적인 개입은 만약 전통적인 표본분산의 공식이 사용된다면, 최종 추정치의 추정된 표본분산을 작아지게 한다. 자료의 분포를 왜곡시키는 것을 방지하기 위해 평균값 결측치 대체는 아무런 부가적 정보가 없을 때, 또는 극소수의 레코드들만이 대체될 필요가 있을 때 종종 최종적 수단으로써 쓰인다.

2.2.3. 비추정 대체

비추정 결측치 대체(Ratio imputation)는 부가적 정보나 다른 기록들로부터 타당한 응답을 사용할 수 있을 때 둘 또는 그 이상의 변수들 간에 존재하는 관계를 사용하는 비 모델(ratio model)을 만들어 사용한다. 예를 들어 비 대체는 아래의 모델을 사용한다.

$$y_i = R \bullet x_i + \epsilon_i$$

여기서 y_i 는 변수 y에 대한 i번째 단위의 값, x_i 는 관련된 변수 x의 i번째 값, R은 직선의 기울기(즉, x_i 가 한 단위 변할 때 y_i 의 변화량), 그리고 ϵ_i 는 평균이 0이고 분산이 $x_i\sigma^2$인 오차항이다. 즉, 이 모델은 y_i 가 대략 x_i 와 선형적 관계에 있으며 y_i 의 분산은 x_i의 값에 비례하게 증가하는 것으로 가정된다.

그렇다면 y_i 값은 다음과 같이 대체될 수 있다.

$$\tilde{y_i} = \frac{\bar{y}}{\bar{x}} x_i$$

이 때, $\tilde{y_i}$는 레코드 i에 대한 변수 y의 결측치 대체값, $\bar{x}$는 결측치 대체군에 응답된 x값의 평균, $\bar{y}$는 결측치 대체군에 응답된 y값의 평균이다. 따라서 대체값은 해당 레코드의 보조 변수값에 $\bar{y}/\bar{x}$ 로 표현되는 승수를 곱하여 계산한다.

예를 들어, 취업, 지불된 급료 그리고 시간에 대한 조사표에서 2주간의 지불된 급료 y에 대한 타당하지 못한 입력 값을 가지고 있으나, 임금을 받은 피고용인의 숫자 x는 적합하게 보고되었고 그 회사의 산업은 이미 알려져 있다고 가정하자. 이 회사와 동일한 산업 분류내에서 지불된 급료와 임금을 받은 노동자의 숫자 모두 정확하게 보고된 자료를 바탕으로 지불된 급료와 피고용인의 숫자 사이의 비율을 계산할 수 있을 것이다. 이 비율(급료 대 피고용인의 숫자)은 차후 급료 값을 결정할 결측치 대체가 필요한 조사표 상의 피고용인의 숫자에도 적용될 수 있다. 이는 위에서 설명한 비추정 결측치 대체의 대표적 사례이며 이러한 비추정 결측치 대체는 사업체 조사에서 가장 많이 사용되는 대체법 중의 하나이다. 여기서 보통 보조변수 x_i 로는 종업원수가 사용되는 경우가 많고 결측치 대체군은 산업 분류가 사용된다.

2.2.4. 핫덱 결측치 대체

핫덱 결측치 대체는 결측값이나 비일관된 값을 보이는 응답자(수용레코드.수용자 :recipient record)를 모든 에디팅을 통과한 자료의 응답자(제공레코드.제공자 :donor record)들로부터 정보를 얻어서 사용하는 것이다. 수용레코드와 유사한 제공레코드를 찾기 위해, 대체가 필요한 변수와 관계있는 변수들을 찾아 대체군을 만들어야한다. 모든 에디팅과정을 통과한 대체군 내의 레코드들은 대체군 내에 있는 대체를 필요로 하는 수용레코드들에게 사용될 제공 레코드이다. 핫덱 결측치 대체는 정량적 자료와 정성적 자료 모두에 사용될 수 있으나, 일반적으로 대체군을 만들어 내기 위해 정량적

변수를 사용한다. 핫뎃 결측치 대체의 주요한 두 가지 형태는 순차적과 무작위 핫덱 결측치 대체이다.

순차적 핫덱 결측치 대체에서는 레코드들이 어떤 순서로 정렬되어있어서, 한 번에 한 레코드씩 대체군 내에서 순차적으로 처리된다. 이 대체방법은 조사표상의 결측항목을 대체군에 속한 제공 레코드의 대응값으로 대신한다. 순차적 핫덱은 만약 매 단계마다 같은 분류법이 사용된다면 결정 대체법이다. 무작위 핫덱 결측치 대체에서는, 제공레코드가 대체군 내에서 임의로 선택된다. 무작위 핫덱은 확률 결측치 대체법이다.

핫덱 결측치 대체를 설명하기 위해, 응답자의 흡연 지위를 대체하는 사례를 상정해 보자. 흡연자와 비흡연자 두 가지 응답이 가능하다고 가정하자. 제공레코드를 찾기 위해, 한 사람의 흡연 상태와 연관되어 있는 연령대와 성에 근거하여 대체군이 만들어진다. 결측치 대체를 필요로 하는 레코드가 15-24세의 여성들이라고 가정하자. 대체군인 제공레코드 집합은 그들의 흡연 상태를 말한 15-24세의 여성 응답자들이다. 제공레코드는 임의로 선택될 수도 있고(무작위 핫덱) 또는 제공자의 목록을 어떤 방식으로든 정렬하고 하나를 순서대로 선택함으로써 선택될 수도 있다(순차적 핫덱).

제공자 결측치 대체법(donor imputation methods)의 장점은 유사한 제공자(즉, 회사, 가계 등)가 유사한 속성을 가지고 있기 때문에 결측치 대체 값은 실제 값과 매우 근접한다는 점이다. 그리고 제공자 결측치 대체의 경우, 일반적으로 자료의 다변량 분포가 지켜진다.

그러나 몇몇 단점도 있다. 순차적 핫덱의 단점은 이 방법이 종종 동일한 제공레코드를 여러 번 사용할 수 있다는 점이다. 만약 한 제공자가 반복적으로 사용되면, 이는 자료의 분포를 왜곡할 수 있고, 인위적으로 추정 표본추출분산을 낮출 수도 있다. 또 다른 단점은 적절한 부가적 정보나 최소한 부분 응답(예를 들어, 가계소득, 나이, 성, 등등)이 대체군을 형성하기 위해 필요하다는 점인데, 이런 조건들이 결측치 대체가 필요한 레코드들에 늘 존재하지는 않는다는 점이다. 또한, 만약 대체군이 작거나 대체군 내에서 무응답율이 높아 아무런 제공레코드를 찾지 못하는 경우가 있으므로 주의를 기울여야 한다 (이는 대체군을 사용하는 모든 대체군에 적용되는 것이다).

제공레코드를 찾는 것이 항상 가능하다는 것을 보증하기 위해, 계층적 핫덱 결측치 대체(hierarchical hot-deck imputation)가 사용될 수 있다. 초기의 가장 상세한 대체군에서 제공레코드를 찾을 수 없을 때, 대체군은 제공레코드를 찾을 수 있는 수준에

도달할 때까지 계층적인 방법으로 쪼개진다.

이러한 핫덱 결측치 대체법은 사업체 조사보다는 가계 조사 자료의 결측치 대체법으로 많이 사용되는 방법이다. 콜드덱 결측치 대체는 핫덱 대체와 유사하나, 차이라면 핫덱 결측치 대체가 현 조사(current survey)로부터 제공레코드를 사용하는 데 비해, 콜드덱 결측치 대체는 다른 소스에서 제공레코드를 찾아 사용한다는 점이다. 종종 콜드덱 결측치 대체는 센서스 자료로부터 또는 동일한 이전조사 자료로부터 나온 과거 자료를 사용한다. 만약 제공레코드가 무작위 방법으로 선택된다면, 그 결측치 대체법은 확률적인 것이 되고, 그렇지 않다면 결정론적인 것이 된다.

광범위한 정량적 자료를 사용한 조사에서는(예를 들어 판매량과 재고량에 대한 사업체 조사), 정량적 자료 중에서 대응변수를 통해 제공레코드를 찾는 것이 바람직하거나 또는 필요할 수 있다. 최근방 이웃 결측치 대체법은 대응 변수에 근거하여 제공레코드를 선택한다. 이 결측치 대체법의 목적은 수용레코드와 정확하게 일치하는 대응변수 값을 갖는 제공레코드를 반드시 찾는 것이 아니다. 대신 대체군 내에서 대응 변수 값이 수용레코드와 가장 가까운 제공레코드, 다시 말해 최근방 이웃(the nearest neighbor)을 찾는 것이 목적이다. 이 근접성은 두레코드의 대응 변수에 해당하는 관찰값들 사이의 거리를 측정함으로써 정의된다.(예를 들어 재고 결측값을 대체하기 위해, 대체군 내에 보고된 판매량의 측면에서 최근방 이웃값을 찾는 것)

대응변수의 척도가 매우 다른 경우(예를 들어 통화변수와 땅의 면적변수)에는 최근방 이웃법을 수행할 때 주의를 기울여야 한다. 대부분의 경우 변수의 변환 형식은 척도를 표준화하기 위해 이루어져야 한다.

2.2.5. 확률적 대체법

또한 정량적 자료에 무작위 잔차(random residuals)를 더함으로써 결정론적 대체를 확률적으로 만들 수 있다. 예를 들어, 비추정 대체값에 무작위 잔차를 더하는 것 등이다.

$$\tilde{y}_i = \hat{y}_i + e_i^*$$

인 곳에서 $\tilde{y}_i$는 레코드 i에 대한 변수 y의 대체 값이고, $\hat{y}_i$는 y_i 의 결정론적 대체치 그리고 e_i^*는 응답자로부터 선택된 또는 분포에서 도출된 무작위 모델 잔차이다.

e_i^*를 선택하는 한 방법은 아래와 같다. 대체군 내의 응답자 묶음은 응답자들의 잔차로부터 계산되어질수 있다. 보통 대체군 내의 모든 $e_{i(r)}$ 값으로부터 무작위로 추출하여 e_i^*을 설정한다.

2.2.6. 기타 사항

응답자에 대한 추적 재조사를 통해서도 해결되지 않는 무응답 또는 유효하지 못한 자료가 있기 때문에, 에디팅 기준을 만족시키지 못하는 변수 값은 결측치 대체되어야 한다. 다른 모든 에디팅 오류(edit failure)의 경우, 가급적 많은 응답자의 자료를 보전해두는 것이 최선의 방법이므로, 모든 에디팅 오류를 대체하는 것이 반드시 권장되는 방법은 아니다. 대신 한 레코드에서 최소한의 변수들만을 대체하는 것이 최선이다. Fellgei와 Holt방식은 결측치 대체가 필요한 변수들을 식별하는 방법이다. 어떤 변수들이 결측치 대체되어야 하는지를 결정하기 위해서는 다음의 세 가지 기준이 쓰인다.

1. 각 레코드에 있는 자료들은 최소한의 변수 값만을 바꿈으로써 모든 에디팅 기준들을 만족시킬 수 있어야 한다.
2. 가능한 한, 자료 파일의 도수분포상태는 유지되어야 한다.
3. 대체의 기준은 특정한 방법을 내포하지 않는 에디팅 기준으로부터 나와야 한다.

Fellegi/Holt의 에디팅 접근법의 핵심적 특징은 에디팅 기준이 특정한 대체법에 한정되지 않는 다는 것이다. 오류로 에디팅 되지 못한 각 레코드는 납득할만한 대체 값의 범위와, 대체할 변수집합을 결정하는 오류처리영역(error localisation)의 단계로 나아가게 된다. 대부분의 시행에서는 하나의 제공레코드는 에디팅 된 변수들 중에서 대체가 필요하지 않은 변수를 기초로 하여 에디팅을 통과한 레코드 중에서 선택한다. 이 방법은 오직 하나의 정확한 대응을 찾기 위한 것이며, 이 방법은 에디팅에 직접적으로 포함되지 않은 여타 변수들을 고려하는 것까지 확장될 수도 있다. 간혹, 알맞은 제공자를 찾지 못하면 디폴트 대체 법이 사용되어야만 한다.

예를 들어, 어떤 조사가 결혼을 한 16세 이하의 사람들을 식별하기 위하여 나이/결혼상태 에디팅기준을 갖고 있다고 하자. 대학 학력을 지닌 18세 이하의 사람들을 식별하는 나이/교육 수준 에디팅을 갖는 조사를 생각하고, 이 조사 자료가 두 에디팅 모두를 만족시키지 못하는 레코드를 가지고 있다고 가정하자. 가령 대학 학력을 가진 10살의 결혼한 여성이 있었다고 생각하라. 이 레코드가 두 에디팅 기준을 모두 통과

하기 위해서는 개인의 결혼 지위와 교육 수준이 모두 바뀌거나, 또는 단순히 나이만 바뀌면 된다. Fellegi/Holt의 작업틀(framework)은 후자를 권한다.

2.3. 해외 사례

2.3.1 미국 BLS(Bureau of Labor Statistics) 의 사업체 조사 무응답 대체법

미국의 Bureau of Labor Statistics (BLS)는 자신들이 실시하고 있는 개별적인 조사에 따라 무응답 대체 방법과 그에 관련된 추정 방법에 대한 연구를 수행해왔다. 즉, 각 조사별로 표본 설계와 조사 주기, 각 조사에서 나타나는 무응답의 특성, 주요 보조변수들의 사용 가능 여부, 조사되는 항목의 개수 등에 따라 무응답 대체 방법을 발전시켜 왔다.

아래의 세 가지 예제를 통해 BLS의 사업체 조사에서 나타나는 무응답 조정에 대한 이해와 방법론에 대해 살펴보도록 한다.

2.3.1.1 Job Openings and Labor Turnover Survey

미국의 Bureau of Labor Statistics 에서 매달 실시하는 Job Openings and Labor Turnover Survey(JOLTS)는 고용에 관한 지표의 변화를 조사하는 사업체 조사이다. 이 조사는 콜롬비아 특별지구를 포함한 미국 전역의 비농업 부문 사업체를 대상으로 실시되며 표본으로 선정된 사업체에 대해서 총 고용규모, 총 미채용 규모(total number of job openings), 총 급여, 퇴직, 일시 해고 또는 해고, 기타 사유로 인한 이직 등에 대한 조사가 이루어진다. 데이터 수집은 Computer-Assisted Telephone Interviewing (CATI) 시스템을 통한 인터뷰와 Touchtone Data Entry (TDE)를 통한 자기기입방식으로 이루어진다. 이렇게 얻어진 데이터는 노동력 부족에 대한 수요자 측면, 즉 사업체의 입장에서 바라보는 노동 시장에 대한 경제적 지표 역할을 하는데, 특히 여기서 조사되는 미채용율(job opening rate)은 고용 시장의 탄력성을 나타내는 중요한 척도가 된다. 이 조사를 통해서 얻어진 자료들은 국가 경제 정책이나 향후 계획 수립에 쓰일 뿐만 아니라 교육과 구직자들을 위한 훈련에도 쓰인다.

▶ Sampling Design

JOLTS는 Bureau of Labor Statistics의 Quarterly Census of Employment and Wages (QCEW) program을 통해 수집한 The Longitudinal Data Base (LDB)에서 얻은 약 9백만 개의 비농업 부문 사업체 리스트를 사용하여 이 중 16,000개의 사업체를 표본으로 선정하여 조사를 실시하고 있다.

JOLTS의 표본은 3가지 기준에 따라 층화되는데, 이는 4개의 Census region에 따른 지역별 층화, 6개의 고용규모에 따른 사업체 크기별 층화, 그리고 사업체 소유 여부와 표준산업구분(Standard Industrial Classification; SIC)을 이용한 산업별 층화를 의미하며, 아래의 <표 2-1>은 이 중 산업별 층화의 예를 나타내고 있다.

<표 2-1> Industry Composition of JOLTS Strata

Division	SIC
Private	Mining
	Construction
	Durable Goods Mfg
	Nondurable Goods Mfg
	Transport & Utilities
	Wholesale Trade
	Retail Trade
	Finance, Insurance & RealEstate
	Services
Government	Federal
	State & Local

JOLTS는 표본으로 선정된 사업체의 응답에 대한 부담을 덜어주고, 조사 비용 절감을 위해 rotating panel design을 사용하고 있다. JOLTS는 18개의 uncertainty panel 들과 1개의 certainty panel을 사용하고 있다. 16,000개의 표본들 중 규모가 큰 사업체들과 같이 표본추출확률이 매우 큰 표본들은 certainty panel에 속하게 되고 나머지 표본들은 각 층 내에서 같은 크기로 나뉘어져 18개의 uncertainty panel에 각각 속하게 된다. 이렇게 한 후 certainty panel에 속한 사업체들에 대해서는 매달 조사를 실시하고 나머지 18개의 uncertainty panel에 대해서는 18개월에 걸쳐 매달 하나의 panel에 속한 사업체에 대해서만 조사를 실시한다.

▶ Imputation

JOLTS에서 발생하는 무응답은 조사 항목 중 총 고용규모에 대한 응답 여부에 의해 단위 무응답 (unit nonresponse)과 항목 무응답(item nonreponse)로 구분된다.

조사가 이루어진 달에 총 고용규모 항목에 대해 대답하지 않은 사업체는 대개의 경우 다른 항목에 대해서도 응답할 확률이 매우 적기 때문에 이런 경우는 단위 무응답으로 분류된다. 이러한 단위 무응답의 보정을 위해 JOLTS에서는 산업별, 지역별, 사업체 크기별로 나누어진 각 cell들을 부분적으로 병합하여 무응답 표본이 속한 cell의 가중치를 조정하는 가중치 조정법(weighting adjustment)을 사용하는데 이를 위해 nonresponse adjustment factors (NRAFs)가 사용된다. NRAFs는 조사가 이루어진 달의 응답자수와 전체 표본 수에 기초하여 매달 계산되는데 noncertainty panel과 certainty panel을 구분해서 계산하게 된다. 이 factor는 표본에 속한 모든 사업체들(viable sample case)의 selection weight와 응답한 사업체들(usable sample case)의 selection weight의 합의 비(ratio)로써 계산되며 이를 수식으로 나타내면 다음과 같다.

$$NRAF_{ch,p\in(1,18)} = \frac{\sum_{i\in ch,viable,p\in(1,18)} w_{i,ch}}{\sum_{i\in ch,usable,p\in(1,18)} w_{i,ch}}$$

$$NRAF_{ch,p=0} = \frac{\sum_{i\in ch,viable,p=0} w_{i,ch}}{\sum_{i\in ch,usable,p=0} w_{i,ch}}$$

여기서 w는 selection weight를 의미하며 i는 각 개별 사업체, ch는 병합(collapsed)된 층, p는 0부터 18까지의 panel을 의미하는데 0번 panel이 certainty panel이 된다.

이러한 단위 무응답과 달리 총 고용규모에 대해서는 응답을 했지만 그 외의 다른 항목에 대해서 하나 이상의 무응답이 발생한 경우에는 항목 무응답으로 분류되며 이러한 항목 무응답의 보정에는 nearest neighbour imputation method를 사용하는 hot-deck imputation procedure가 쓰인다.

이를 위해 먼저 응답 데이터들을 모든 panel에 대해서 합치고 나서 지역별/크기별/산업별로 층화를 한 후, 그 층 내에서 고용규모의 크기대로 다시 정렬을 한다. 이렇게 한 후 항목 무응답이 발생한 사업체(recipient)는 자신이 속한 cell 내에서 총 고용규모가 가장 비슷한 응답 사업체와 매칭이 된다. 즉, e_i를 항목 무응답이 발생한 사업체의 고용규모라고 한다면, 응답된 표본들의 집합 D에 속하는 모든 $j \neq j^*$에 대해서 $|e_i - e_{j^* \in D}| < |e_i - e_{j \in D}|$를 만족하는 e_{j^*}를 “donor"로 선정하게 된다. 즉, cell 내에서 고용규모를 이용한 nearest neighbor imputation 방법을 사용하는 것이다.

이렇게 매칭이 된 응답 사업체(donor)의 데이터를 사용하여 imputation이 필요한 항목의 값과 총 고용규모의 비(ratio)를 계산 한 후, 이를 무응답 사업체의 총 고용규모에 곱해줌으로써 무응답 대체가 이루어지게 된다. 이러한 방식으로 ratio를 사용해 imputation을 실시하는 것은 매칭 된 사업체에서 얻은 값을 사용해 그대로 무응답 대체를 하는 것 보다 좀 더 scale-invariant한 imputation이 이루어지게 하는 효과가 있다.

▶ Example

지역별/산업별/사업체 크기 별로 나누어진 하나의 cell 내에서 다음과 같은 자료구조가 있다고 생각해보자.

사업체	고용규모	총급여
Unit 1	100	60,000
Unit 2	90	* (missing)
Unit 3	120	75,000
Unit 4	110	67,000

모든 사업체에 대해 고용규모는 조사가 되었고 Unit 2에서 총 급여에 대해 무응답이 발생했을 때, nearest neighbor imputation 방법을 적용해서 총 급여에 대해 imputation을 하게 되면 다음과 같다.

먼저, 속한 cell 내에서 총 고용규모가 가장 비슷한 응답 사업체를 찾는다. 이 경우 Unit 2와 고용 규모가 가장 비슷한 사업체는 Unit 1이다. 이렇게 Donor가 선정이 되면 donor의 값을 이용해 (무응답 대체가 필요한 항목)/(고용규모) 의 ratio를 계산하는데 이 경우에는 (60,000)/(100)=600 이 된다. 따라서 무응답 대체 값은 90 * (60,000)/(100) = 90 * 600 = 54,000 이 된다.

2.3.1.2 Current Employment Survey

Current Employment Survey(CES)는 BLS와 State Workforce Agencies가 State-federal cooperative program에 의해 매달 함께 시행하는 대규모의 사업체 조사로써 전국적으로 선정된 표본 사업체의 급료총액기록(payroll record)을 바탕으로 하여 총 고용규모와 노동 시간, 급여 등에 대한 추정치를 국가단위, 주 단위 그리고 주요 대도시 단위 별로 제공한다. 정보 수집은 전화, fax, computer-assisted interview, mail 등의 각종 전자 미디어를 통해 이루어지는데 이는 설문 응답 시간을 단축시킴으로써 응답률을 높이는 결과를 가져 온다.

▶ Sampling Design

CES의 표본 추출은 The Longitudinal Data Base(LDB)에서 얻은 리스트를 통해 이루어지는데, 여기서 LDB는 unemployed insurance(UI)를 적용 받는 미국 내 약 9백만 개의 사업체 정보가 포함되어 있는 데이터베이스이다. 이 수치는 미국 경제의 거의 모든 부분을 포함하는 것으로써, LDB는 The Quarterly Census of Employment and Wages (QCEW) program을 통해 매 분기별로 고용과 임금에 관한 모든 정보가 수집된다. LDB의 리스트를 통해 약 160,000개의 기업과 정부기관이 표본으로 추출되는데 이는 대략 400,000개의 개별적인 worksite를 포함하며, 미국 내 비농업 부문 근로자의 약 1/3을 포함하는 수치이다.

CES의 표본은 산업별, 크기별, 주(State)별로 층화되는데, 13개의 산업구분과 8개의 사업체 크기별 구분 기준을 사용하여 각 주(State)별로 총 104개의 cell이 발생하게 된다.

▶ Benchmark adjustment

여타 다른 조사들과 마찬가지로 CES에서도 매년 sample-based estimates와 complete population counts간의 재조정을 통한 benchmark를 실시한다. 이러한 benchmark adjustment가 필요한 이유는 다음과 같다. 표본 조사는 그 규모가 모집단 조사에 비해 매우 작기 때문에 시의성 있는 추정치를 제공할 수 있다. 하지만 신생 사업체의 등장 또는 사업체의 소멸 등과 같은 현상을 적절히 반영하기 어려운 것도 사실이다. 따라서 주기적으로 모집단 자료를 사용해 표본 조사의 추정치들을 재조정하는 것이 필요한데 이를 통해 표집오차와 비표집 오차를 줄이는 효과도 함께 가질

수 있다.

CES에서는 매년 unemployment insurance (UI) tax records에서 얻은 전체 모집단 자료(UI universe count)를 이용하여 표본을 사용해 추정한 총 고용규모에 대해 benchmark adjustment를 실시한다. UI universe count는 매 분기별로 조사되는 자료로써 미국 내의 비농업 부문 근로자의 거의 97퍼센트를 포함하는 모집단 자료이다. UI 모집단 자료로 커버되지 않는 부분에 대해서는 여러 가지 다른 정보(Retirement Board and County Business Patterns 등)를 이용한 benchmark를 사용한다.

CES의 Benchmark는 매년 3월에 실시되는데, UI-based benchmark level로 sample-based employment estimate를 대체한 후, 그 차이에 대해서는 “wedge back" procedure를 통해 재조정이 시행된다. 예를 들어, 2007년 3월에 실시된 benchmark adjustment 후 752,000 만큼의 차이가 발생했다면, 이 값의 1/12만큼을 2006년 4월의 월별 고용규모 추정량에 더하고, 2/12를 2006년 5월의 월별 고용규모 추정량에 더해준다. 이렇게 하면 마지막 2007년 2월에는 발생한 차이의 11/12만큼이 더해지게 되는데, 이는 기본적으로 전체 추정 오차는 일정한 비율로 증가한다는 것을 가정하고 있다.

▶ Imputation

CES에서 발생하는 무응답은 두 가지로 구분된다. 첫 번째 경우는 일종의 단위 무응답이 발생한 경우인데, 여기에 대해 알아보기 전에 CES에서 사용하는 총 고용규모 추정량에 대해 먼저 살펴보기로 한다.

CES는 주(State)별/크기별/산업별로 나뉘어진 각 cell 내에서 총 고용규모를 추정하는데, 이 때 연속한 두 달 사이의 고용 증가율을 사용하는 “weighted link relative" estimator를 사용한다. 이것은 기본적으로 ratio estimator의 일종이고 식으로 표현하면 다음과 같다.

$$\hat{E}_c = \frac{\sum e_{i,c}}{\sum e_{i,p}} \times \hat{E}_p$$

이 때 $\hat{E}$=estimated employment, e=sample unit employment, c=current month, p=previous month 를 나타낸다. 즉, 지난 달 추정된 고용 규모에 승수를 곱하여 계산하는 비추정의 형태인데 여기서 승수는 동일 cell 내에서 응답 사업체들의 고용 증

가율 (=이 달 고용 규모 / 지난달 고용 규모)로 계산된다.

단위 무응답이 발생하게 되면 그 해당 사업체의 데이터는 총 고용규모를 추정하는 것에서 제외된다. 즉, 그 사업체가 속해있는 cell 내에서의 고용 증가율 계산에서 해당 무응답 사업체를 제외시키는 것이다. 이것은 같은 cell 내에서 무응답 사업체는 응답 사업체와 같은 특징을 갖고 있다는 것을 가정하며 이는 결국 무응답이 발생한 사업체가 속한 cell 내에서 가중치 조정을 하는 것과 같은 효과를 얻을 수 있게 된다.

무응답이 발생하는 두 번째 경우는 다음과 같다. CES는 기본적으로 국가단위와 주 단위의 추정량을 제공하는 것을 목표로 한다. 하지만 많은 주 정부에는 좀 더 작은 단위 (대도시 단위)의 통계량에도 관심이 있다. 이를 위해 CES 프로그램은 각 worksite 별로 정보 제공을 요청하게 되는데, 때로는 여러 대도시 지역을 포함한 합산된 정보만이 제공 가능한 경우가 발생하기도 한다. 이러한 경우 local area estimation은 가장 최근의 ES202 총 고용규모 데이터를 이용하여 주어진 합산된 데이터를 각 개별적인 worksite에 비례 배분하는 방법이 쓰이는데 이는 missing local information을 대체하는 가장 간단한 형태의 ratio-based imputation이라고 할 수 있다.

2.3.1.3. National Compensation Survey

National Compensation Survey(NCS)는 Bureau of Labor Statistics에 의해 실시되는 사업체 조사로써 기존의 Employment Cost Index (ECI), Occupational Compensation Survey Program (OCSP), and Employee Benefits Survey (EBS)로 나누어져 있던 3가지의 조사들이 합해진 것이다. 이 조사에서는 각 직업에 따른 임금 수준과 복리후생(benefit) 등에 관한 정보가 수집되고 개인 사업체와 주 정부, 지방 정부를 포함하며, 이 결과는 각 대도시단위별, 주 단위별, 국가 단위별로 이용 가능하다.

▶ Sampling design

NCS의 표본은 LDB를 표본 추출 틀로 사용하며 총 18,239개의 사업체를 표본으로 사용하고 있는데 이는 거의 8천 9백만 명의 근로자를 포함하고 있는 수준이다. NCS의 표본은 1명 이상의 근로자가 있는 개인 사업체와 50명 이상의 근로자가 있는 주 정부와 지방 정부를 대상으로 한다.

NCS의 표본 추출은 3단계로 이루어진다. 1단계는 지역 선택에 관한 부분인데 NCS의 표본은 총 154개의 대도시권과 비도시권으로 구성되어 있는데 이는 the Office of Management and Budget에 의해 지정된 326개의 전국 지역 구분을 대표한다. 두 번째 단계에서는 표본을 산업분류와 소유권에 기초하여 나눈 후 표본 사업체를 선정하게 된다. 표본 사업체는 각 층 내에서 총 고용규모를 기준으로 probability proportional sampling method를 사용해서 선정된다. 마지막 세 번째 단계에서는 선택된 표본 사업체 내에서 직업에 따른 표본 추출이 이루어진다. 이 때, 직업에 따른 표본 추출은 다음의 4단계를 거쳐서 이루어진다.

1. Probability-proportional-to-size selection of establishment jobs
2. Classification of jobs into occupations based on the Census of Population System
3. Characterization of jobs as full versus part time, union versus nonunion, and time versus incentive
4. Determination of the level of work of each job.

NCS에서 발생하는 무응답의 처리에 관해 살펴보기 전에 NCS가 가지고 있는 특징을 살펴보도록 한다.

첫째로 NCS는 five-year rotation sample design을 사용하고 있다, 이것은 표본의 일부에 대해서는 연간 급여에 대해 조사를 실시하고 나머지 표본에 대해서는 분기별 급여와 복리후생에 관한 조사를 실시하는 것이다. 이러한 순환적인 특성에 의해 만약 무응답이 발생했을 경우 그 무응답이 발생하는 시점이 무응답 개체에 대한 정보가 상대적으로 적은 초기 시점인지 혹은 그 이후의 시점인지 구분이 가능하다.

두 번째 특징은 NCS의 다단계 표본 설계에 의해 사업체 단위 또는 직업 단위에서 모두 무응답이 발생할 수 있다는 것이다. 따라서 사업체 수준에서의 무응답과 직업 수준에서의 무응답에 대한 무응답 처리를 위해서는 각기 다른 기준의 접근방식이 필요하다.

세 번째 특징은 조사 항목의 특성에 따른 것인데, ‘복리후생’이라는 항목은 '급여'에 비해 명확하게 정의되지 않는 특성이 있기 때문에 급여에 대한 항목들에 비해 무응답이 많이 발생하는 경향이 있다. 따라서 급여에 대해 응답하지 않은 표본은 복리후생에 관해서도 응답하지 않는 경향이 높게 나타난다. 즉, 복리후생에 대한 무응답은 급

여부분의 무응답에 nested되어 있다고 볼 수 있다.

마지막으로 복리후생에 대한 정보들 간에의 복잡한 상호관계가 존재하기 때문에 무응답 조정을 하는 과정에서 이러한 상호관계가 그대로 유지되어야 하는 것이 매우 중요하다.

▶ Imputation

NCS에서는 급여에 관한 항목과 연금에 관한 항목에 대해 각기 다른 imputation 방법을 사용하고 있다.

급여에 관한 항목 무응답 대체를 살펴보면, 무응답 처리를 위해 먼저 NCS가 사용하고 있는 five-year rotation sample design을 통해 무응답의 발생 시점을 초기 시점(initiation)과 이후 시점(update period) 구분한다. 초기 시점에 발생한 무응답에 대해서는 가중치 조정을 통한 무응답 대체가 이루어지는데, 사업체 수준과 직업 수준에 따라 unit Non-Response Adjustment(NRA)가 이루어진 뒤, 이 둘을 사용해 최종 Non-Response Adjustment Weight(NRAW)가 결정된다.

먼저 사업체 수준에서 NRA는 다음과 같은 순서로 이루어진다.

1. 주어진 표본을 사업체의 크기와 산업 구분별로 나누어서 Establishment Non-Response Cells(ENRC)를 만든다.
2. 각 ENRC 내에서 다음과 같이 ENRA Factor를 계산한다.

$$ENRAF = \frac{A+B}{A}$$

A=weighted employment of all usable establishments in the ENRC

B=weighted employment of all viable but not usable establishments in the ENRC

3. 이렇게 계산된 ENRAF가 1/4과 4의 범위 안에 있는지 확인한 후, 이 범위를 벗어난다면 크기별로 나누어진 cell를 병합(collapse)한 후 다시 ENRAF를 계산한다.
4. 계산된 모든 ENRAF가 1/4 과 4의 범위 안에 존재하게 될 때까지 이 과정

을 반복한다. 만약 크기별로 나누어진 cell들이 모두 병합된 후에도 이 조건이 만족되지 않으면 그 다음으로는 산업별로 나누어진 cell을 병합해 나가도록 한다.

이렇게 해서 얻어진 ENRC를 Final ENRC (FENRC)라고 부르기로 한다. 여기서 한 가지 주의할 점은 이러한 collapsing은 표본에 의존하는 경향이 있기 때문에 제일 처음 시작한 ENRC가 같더라도 표본이 다르다면 Final ENRC는 달라질 수 있다.

그 다음으로 직업 수준에서 NRA를 시행하게 되는데, 이 과정은 사업체 수준에서의 NRA와 같은 방식을 따른다. 이렇게 해서 얻어진 ENRAF와 Occupational Non-Response Adjustment Factor(ONRAF)를 사용해서 다음과 같이 최종 NRA weight가 결정된다.

$$NRA\,W = ENRAF \times ONRAF \times W$$

이 때, W=sample weight 를 의미한다.

이후 시점(update period)에서 발생한 무응답에 대해서는 과거의 급여 데이터와 현재 급여 데이터의 log-ratio를 사용한 regression-based imputation을 사용한다. 즉, j 번째 cell의 i 번째 표본의 t 시간에서의 급여(W_{ijt})는 다음과 같이 나타낼 수 있다.

$$\ln W_{ijt} = \beta_{jt} \ln W_{ij(t-1)} + \epsilon_{ijt}$$

$$with\ E(\epsilon_{ijt}) = 0 \ \text{and}\ E(\epsilon_{ijt}^2) = \sigma_t^2 \ln(W_{ij(t-1)})$$

이 때, β_{jt}의 weighted least squared estimator는 다음과 같다.

$$\widehat{\beta_{jt}} = \frac{\left(\frac{1}{n_j}\sum_i \ln W_{ijt}\right)}{\left(\frac{1}{n_j}\sum_i \ln W_{ij(t-1)}\right)}$$

따라서 k 번째 표본에서 발생한 급여에 대한 무응답은 다음과 같이 imputation이 될 수 있다.

$$\widehat{W_{kjt}} = \exp\left(\widehat{\beta_{jt}} \ln W_{kj(t-1)}\right)$$

연금에 대한 무응답 대체에는 주로 nearest neighbour method가 사용되는데 복리후생의 종류에 따라 조금씩 다른 방법이 쓰인다.

Social Security나 Medicare premium payments와 같은 항목은 급여 수준에 따라서 정해진 법적 기준이 있기 때문에 이런 종류의 항목 무응답은 급여 내역을 알고 있다면 법률에 의거하여 직접 계산 된 값을 사용해 imputation을 할 수 있다.

그 외의 다른 무응답 항목에 쓰이는 방법 중 하나로써 random within-cell hot-deck method가 있는데 그 절차는 다음과 같다.

먼저 모든 표본에 대해서 알려진 보조변수를 사용해 표본을 cell 단위로 나눈 후, 각 cell 내에서 무응답 항목을 대체하기 위한 donor를 응답 표본들 중에서 무작위로 선정한다. 이렇게 선정된 donor 표본의 값을 사용해 무응답 표본의 모든 무응답 항목을 대체한다. 이렇듯 하나의 donor를 사용해 한 표본 내에서 발생한 모든 무응답 항목을 일괄적으로 대체하게 되면 항목들 간에 존재하는 상호 관계를 유지한 채로 무응답 대체를 할 수 있게 된다. Cell을 나누는 기준이 되는 보조변수들은 다음과 같다.

1. ownership(private or public sector)
2. size class
3. major industry division
4. major occupational group
5. tow-digit NAICS (North America Industry Classification System) code
6. union/non-union status
7. full-time/part-time status
8. Census region

그런데 만약 특정 cell에서 적합한 donor를 찾을 수 없는 경우가 발생하게 되면 적합한 donor를 찾을 수 있을 때까지 cell을 순차적으로 병합(collapse)해 나가게 된다. 예를 들어 위에서 언급한 8가지의 변수들을 사용해서 cell을 나누었을 때 donor를 찾을 수 없었다면 Census region이라는 보조변수에 대해서 cell을 병합하도록 한다. 즉, 이제는 7가지의 보조변수만을 사용해 cell을 나눈 후, 다시 적합한 donor를 찾아

보도록 한다.

또한 같은 donor가 사용될 수 있는 횟수를 제한할 수 있는데 여기서는 3번으로 정해져 있다. 기존의 EBS 에서는 같은 donor는 한 번 이상 쓰일 수 없었는데, 연구결과에 따르면 이 횟수를 3번으로 늘림으로써 imputation 과정에서 발생하는 cell collapsing 횟수를 크게 줄일 수 있었다. 이는 결국 bias의 감소를 의미한다.

또 다른 방법으로는 nearest neighbour within-cell hot-deck method가 있는데 이는 다음과 같다. 먼저 표본들을 다음과 같은 기준에 의해 나누도록 한다.

1. ownership (private or public)
2. benefit area
3. subset of key provisions for each benefit area
4. size class
5. major industry division
6. major occupational group
7. two-digit NAICS code
8. union/non-union status
9. full-time/park-time status
10. Census region

이렇게 한 뒤 사업체의 고용규모를 기준으로 하여 고용규모가 가장 비슷한 표본을 donor로 선정하여 그 값을 사용해 imputation을 하도록 한다. 만약 주어진 cell내에서 적합한 donor를 찾을 수 없다면 순차적으로 cell을 병합해 나가게 되는데 위에서 언급된 10가지 변수들 중 1-3번 변수에 대해서는 cell을 병합하지 않는다. 만약 4-10번 변수들에 대해 모두 cell을 병합했는데도 적합한 donor가 나타나지 않으면, 이런 경우에는 같은 donor를 사용할 수 있는 횟수를 3번에서 그 이상으로 늘리는 방법을 쓰도록 한다.

2.3.2 미국 Census Bureau의 사업체 조사에서의 무응답 대체법

U.S. Census Bureau는 서비스, 도/소매업, 운송업, 제조업, 건설업, 금융업, R&D, 의료 보험비, 해외무역 등의 경제 분야에 관한 데이터를 수집하기 위한 조사를 매월, 매 분기별, 매년 또는 5년 단위로 시행하고 있다.

이러한 조사에서 발생하는 무응답 데이터의 처리에는 여러 가지 다양한 방법들이 쓰이는데, Census Bureau에서는 그러한 다양한 방법들의 공통적인 특성들을 모아 Standard Economic Processing System(StEPS) 이라는 일반화된 데이터 처리 과정을 제공하고 있다.

2.3.2.1 Standard Economic Processing System

Standard Economic Processing System(StEPS)이란 U.S. Census Bureau에서 개발한 일반화된 processing system을 지칭하는 것으로써, 조사 데이터의 수집과 에디팅, 데이터의 검토와 수정, 대체(imputation), 추정, 그리고 시스템 운영 등에 관한 기준과 표준적인 처리 방법을 제공하며, 현재 8개의 survey program이 이 시스템을 사용하고 있다.[1)]

StEPS에서 수행하는 imputation은 크게 두 가지로 구분된다. 첫째는 simple imputation으로써 imputation을 이용하여 추정된 값을 실제 조사된 값과 동일하게 취급하는 경우이다. 가장 자주 쓰이는 simple imputation method는 data filling으로써 이 방법은 응답된 데이터들의 논리적인 상호관계를 통해 무응답 데이터의 값이 추론하여 무응답 된 항목을 채워 넣는 것을 말한다. 또한 이 방법은 데이터의 변화가 전체 값에 큰 영향을 미치지 않는 경우에 사용되곤 한다.

예를 들어 전체데이터의 총계는 조사 되었고 총계를 이루는 여러 개의 변수들 중 어느 하나의 변수 값이 조사되지 않았을 때, 조사된 변수 값들의 합과 총계 값이 크게 차이가 나지 않는다면 무응답 된 변수의 값은 총계 값과 응답된 변수 값들의 합의 차이로 대체될 수 있다. 이러한 simple imputation은 일반적으로 데이터의 에디팅과 검토, 수정 작업 이전에 시행된다.

다음은 general imputation으로써 이것은 데이터의 재검토와 수정 이후에도 해결되지 않은 무응답을 처리하기 위한 과정이다. StEPS는 단위 무응답 또는 부분적인 단위 무응답을 포함하고 있는 데이터를 사용한 추정을 위해 각 개별 항목을 대체하거나 balance complex를 조정하는 방법으로 general imputation을 사용한다. 즉, general imputation은 크게 item imputation과 adjustment of balance complex의

1) Manufacturer's Shipments, Inventories, & Orders Survey (M3), Current Industrial Reports (CIR), Survey of Industrial Research and Development (R&D). Plant Capacity Utilization Survey (PCU), Manufacturing Energy Consumption Survey (MECS), Survey of {Plant Capacity Utilization (PACE), Annual Survey of Manufactures-E-commerce Business, Service Annual Surveys (SAS), Annual Retail Trade Survey (ARTS), and Annual Trade Survey (ATS).

두 가지 카테고리로 나누어 볼 수 있다. 각 카테고리 내에서 여러 가지 다양한 imputation method가 존재하는데 아래의 <표 2-2>는 다양한 개별 항목의 대체 방법을 소개하고 있다.

<표 2-2> 개별 항목의 대체 방법

Group	Name	Definition	Formula
Logical and Direct Substitution	SUM	Sum of auxiliary variables	$v' = z_1 + z_2 + \cdots + z_n$
	RESIDUA	Auxiliary variables minus the sum of other auxiliary variables	$v' = z_1 - (z_2 + \cdots + z_n)$
	PRODUCT	Product of two auxiliary variables	$v' = z_1 \times z_2$
	VALUE	Value of the auxiliary variable	$v' = z_1$
Mean	MEAN	Mean value of an auxiliary variable	$v' = \overline{z_1}$
Ratio	RATIO	Ratio prediction for imputed item	$v' = z_1(S(v)/S(z_1))_I$
	ATREND	Auxiliary variable multiplied by a trend	$v' = z_1(z_2/z_3)$
	AUXRAT	Auxiliary variable times a ratio-of-identicals	$v' = z_1(S(z_2)/S(z_3))_I$
Regression	SIMPREG	Auxiliary variable times a regression coefficient	$v' = \beta_1 z_1$
	MULTREG	Multiple regression prediction for imputed item	$v' = \beta_1 z_1 + \cdots + \beta_n z_n$

v = the item-name of the value being imputed

v' = the imputed value of v

z_j = the value of the j^{th} auxiliary variable

$S(f)$ = the sum of item f over a defined set of records

$(S(f_1)/S(f_2))_I$ = the ratio-of-identicals of items f_1 and f_2 where the numerator and denominator are both summed over the identical set of response cases.

위의 표에서 쓰이고 있는 보조변수는 다른 데이터셋에서 얻은 대체되어야 하는 항목과 같은 항목일수도 있고, 또는 같은 데이터셋에서 얻은 다른 항목일 수도 있다. mean과 ratio-of identical을 제외한 나머지 방법에서 쓰이는 보조변수는 무응답 대체가 필요한 항목을 갖고 있는 표본에서 얻은 값이며, mean과 ratio-of-identical의 계산에 쓰이는 보조변수는 관련된 imputation cell 내의 적절한 모든 표본에서 얻을

수 있는 값이다.

<표 2-3>는 StEPS를 사용하고 있는 8개의 서베이에서 위에 소개된 개별 항목의 대체 방법 중 어떤 방법을 실제로 사용하고 있는지를 보여주는 빈도수 테이블이다. 가장 많이 사용되는 방법은 ratio method로써 ATREND, RATIO, AUXRAT방법이 있고, 이 중 ATREND 방법은 imputation된 자료들만을 사용해서 만든 보조변수의 비(ratio)를 통해 무응답 대체를 하게 된다. RATIO와 AUXRAT는 적합한 모든 표본들을 사용해서 만든 비(ratio)를 이용하는데 AUXRAT 방법은 RATIO 방법을 좀 더 일반화 시킨 것이다. 그 중에서도 RATIO 방법은 보조변수의 합과 대체해야 하는 변수의 합의 비를 사용한 것으로써 간단한 상황에서 쓰인다.

<표 2-3> 개별항목 대체방법의 사용 빈도

		Item Imputation Methods								
SECTOR	Primary Survey	ATREND	AUXRAT	PRODUCT	RATIO	RESIDUA	SIMPREG	SUM	VALUE	GRAND TOTAL
Manu-facture	ASMECB	0	0	2	0	0	0	0	2	4
	CIR	0	132	0	878	0	0	0	2	2279
	M3	0	0	0	0	0	0	1	1269	13
	PACE	0	72	0	0	0	0	0	12	72
	PCU	1	2	1	0	0	0	0	0	7
	RD	123	57	0	0	2	0	2	3	292
Manufacture Total		124	263	3	878	2	0	3	1394	2667
Service	ARTS	55	54	0	0	0	1	1	40	151
	ATS	12	26	0	0	1	1	0	19	59
	SAS	155	414	0	0	0	7	0	164	740
Service Total		222	494	0	0	1	9	1	223	950
Grand Total		**346**	**757**	3	**878**	3	9	4	1617	3617

StEPS에서는 사용자에게 imputation에 사용 될 방법을 지정할 수 있도록 하고 있는데 이것은 보조변수의 종류, 선택된 대체 방법들의 시행 순서, 혹은 imputation이

사용되는 상황 등에 따라서 구체화 될 수 있다. 사용자는 하나의 항목을 대체하는 경우에 여러 가지 다양한 보조변수를 사용하거나 여러 가지 방법을 사용해서 한번 이상의 imputation을 시행 할 수도 있다. General imputation은 사용 가능한 데이터의 종류에 기초해 어떠한 방법을 사용할 것인지를 결정하도록 하며, 이때 사용 가능한 데이터는 이전 조사에서 얻어진 데이터일수도 있고 또는 센서스 자료 같은 외부 자료가 될 수도 있다.

이러한 다양한 imputation method는 하나의 서베이에서 복합적으로 쓰이기도 한다. 즉, 항목의 특성에 따라 각기 다른 방법들을 사용하게 되는 것인데 다음의 <표 2-4>는 서베이에서 여러 가지 imputation method를 사용하는 항목들의 분포를 보여주고 있다.

<표 2-4> 여러 가지 방법을 사용하는 항목들의 수

		Number of Methods									
SECTOR	Primary Survey	1	2	3	4	5	6	7	8	10	Grand Total
Manu-facture	ASMECB	4	0	0	0	0	0	0	0	0	4
	CIR	259	1010	0	0	0	0	0	0	0	1269
	M3	1	6	0	0	0	0	0	0	0	7
	PACE	0	0	0	0	0	0	0	9	0	9
	PCU	0	2	1	0	0	0	0	0	0	3
	RD	11	102	4	9	0	3	0	0	0	140
Manufacture Total		286	1120	5	9	0	3	0	9	0	1432
Service	ARTS	5	44	8	2	2	0	0	2	0	63
	ATS	2	2	9	1	0	1	0	2	0	17
	SAS	19	220	58	10	0	6	3	0	1	317
Service Total		26	266	75	13	2	7	3	4	1	397
Grand Total		312	1386	80	22	2	10	3	13	1	1829

2.3.3 Statistics Canada의 사업체 조사에서의 무응답 대체 방법

Banff와 The Generalized Edit and Imputation System (GEIS)는 Statistics Canada에서 개발한 데이터 처리 시스템으로써 캐나다의 농업, 산업, 가구 부문 조사에서 널리 쓰이고 있다.

GEIS는 1980년대 중반에 개발이 되어 1980년대 후반부터 쓰이기 시작했고, 최근들어 GEIS를 보완하여 개발한 것이 Banff이다. 따라서 Banff는 GEIS와 거의 같은 방식의 방법론을 사용하고 있고, 점차적으로 GEIS에서 Banff로 바뀌어 가고 있는 추세이다. 또한 두 시스템 간에 존재하는 몇 가지 구조적인 차이에 의해 Banff는 이전의 GEIS에 비해 좀 더 사용자 친화적이며 융통성이 있다는 평을 받고 있다.

2.3.3.1 GEIS

GEIS는 기본적으로 economic survey에서 쓰이기 위해 개발된 것으로 연속형 변수의 처리를 위해 디자인되어 있으며, 크게 editing, error localization, imputation의 세 부분으로 나누어 볼 수 있다. 이 중 imputation 부분을 살펴보면 GEIS는 세 가지 형태의 imputation을 정의하고 있는데, 첫 번째 형태는 logical imputation으로써 이는 무응답이 발생한 항목을 다른 항목을 사용하여 유추 가능할 때 사용한다.

두 번째는 prediction imputation method로써 이는 무응답 항목을 응답 된 다른 보조 변수들의 함수를 사용해 대체하는 모든 imputation 방법을 지칭하는데, 예를 들어 여기에는 mean, ratio, previous value imputation 등이 포함된다.

마지막으로 donor imputation은 무응답 된 항목을 대체하는 값으로써 다른 항목의 값을 사용하는 방법을 일컫는데 여기에는 hot-deck imputation method와 nearest neighbour imputation method가 포함된다.

2.3.3.2 Banff

Banff는 자동화된 데이터의 에디팅과 imputation을 가능하게 하며 9개의 SAS 프로시저로 구성되어 있다. 각 프로시저는 독립적으로 활용 가능하며 또한 필요에 따라 복합적으로도 적용 가능하다. Banff system은 2002년에 처음 등장하였으며 초반의 목적은 GEIS에서 쓰이는 방법론들의 재생산이었다. 따라서 Banff에서 쓰이는 방법론

들은 GEIS가 사용하던 방법론과 거의 동일하다. 하지만 두 시스템간에는 몇몇의 구조적인 차이점이 존재하고 있다.

첫 번째로 GEIS가 underlying database로써 Oracle을 사용하고 있는 것과는 달리 Banff는 SAS에 기초하고 있다. 또한 GEIS의 모듈은 상호 연관되어 있는 반면에 Banff의 SAS 프로시저는 서로 독립적이다. 마지막으로 GEIS는 UNIX 체제에서 실행되는 것과는 달리 Banff는 Windows 환경에서도 사용 가능하다. 이러한 차이점으로 인해 Banff는 GEIS에 비해 사용이 더 쉽고 유연성이 있다는 평가를 받고 있다.

Banff의 가장 중요한 장점 중의 하나가 바로 각 프로시저들 간의 독립적인 활용이 가능하다는 것이다. 즉, 사용자는 데이터 처리 과정에 9개의 프로시저 중 어느 것이라도 단독적 또는 복합적으로 적용을 할 수 있으며 순서에도 구애받지 않고 적용이 가능하다. 또한 하나의 프로시저 실행 후 만들어진 결과물은 다른 프로시저의 입력 데이터로써 사용이 가능하다. Banff의 명령어는 SAS의 명령어와 같은 형태이며 출력물은 SAS dataset의 형태를 띈다.

2.3.3.3 The Banff Procedure

Banff의 프로시저는 순서에 상관없이 사용할 수 있지만 모든 프로시저를 사용한다고 할 때 일반적으로 사용되는 순서에 따라 그 과정을 대략적으로 살펴보면 다음과 같다.

1. Proc Verifyedits - Edit Specification and Analysis

이 과정은 데이터분석을 위한 기초 단계로 쓰일 수 있으며 각 data field간의 상관관계를 알아보기 위한 과정이다. 이러한 상관관계를 edits 라고 한다면 이것은 설문지 또는 데이터의 분석을 통해서 알 수 있을 것이다. 이러한 과정을 데이터에 존재하는 제약조건들을 edits가 정확히 나타내고 있는지를 확인하는데 도움을 준다. Banff의 edits rule은 반드시 linear form으로 표현되어야 하며 데이터는 numeric, non-negative, continuous 해야 한다. 따라서 $x_1,x_2,...,x_n$ 의 변수를 이용하여 사용자가 m 개의 edits rule을 지정하였다면 이것은 다음과 같은 형태로 표현될 수 있다. 이 때, non-negative edits rule은 Banff에 의해 자동적으로 추가된다.

$$
\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \le b_1 \\
&\vdots \\
&a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \le b_m \\
&x_1 \ge 0 \\
&\vdots \\
&x_n \ge 0
\end{aligned}
$$

이렇게 edits system이 정해지고 나면, Verifyedits 프로시저는 정의된 edit system의 consistency를 체크하게 된다. 즉, 이 중에서 불필요한 edits가 존재하는지 또는 정의되지 않은 또 다른 edits rule이 숨어있는지를 체크하게 된다. 이러한 과정을 통해 필요한 최소한의 edits rule을 결정함으로써 이후에 일어나는 다른 프로시저들의 효율을 높일 수 있게 된다.

2. Proc Editstats - Edit Summary Statistics Tables

Verifyedits 프로시저는 데이터와 상관없이 linear edits 자체만을 고려하여 equation system의 consistency 여부를 확인하는 과정이라고 한다면 Editstats는 결정된 edits system이 실제 데이터에 부합하는 지를 체크하는 과정이다. 이 프로시저에서는 다음과 같은 다섯 개의 결과표가 만들어 진다.

1. the number of records which passed, missed and failed each edit
2. the distribution of records which passed, missed and failed a given number of edits
3. the number of records with pass, miss and fail overall record states
4. the number of times each filed was involved in an edit which passed, missed or failed
5. the number of times each field contributed to the overall record states

여기서 "miss" 는 무응답 데이터에 의해 edit가 실패 했을 때를 의미하고 "fail"은 하나 이상의 non-missing value에 의한 editing 실패를 나타낸다. 만약 어느 특정 edit rule에서 editing 실패율이 높게 나타났다면 사용자는 edit rule을 수정하는 등의 조치를 취해야 할 것이다. 또한 이러한 과정을 imputation이 적용된 후에 다시 한 번 실행함으로써 imputation의 효율성을 비교할 수 있다.

3. Proc Outlier - Outlier Detection

이 과정은 자료에서 이상값의 존재 여부를 판단하는 과정으로써 결과물에서 발견된 이상치들에 대해서 imputation이 필요할 것임을 나타내 준다. 또한 imputation이 필요한 정도는 아니지만 나머지 데이터들과는 확연히 다른 특징을 나타내기 때문에 imputation 과정에서 추정에 쓰이거나 donor의 자격에 적합하지 않은 값들을 나타내기도 한다.

이 프로시저에서 선택할 수 있는 방법은 Ratio, Historical Trend 그리고 Current method로써 세 가지가 있다. 만약 신뢰성이 있는 보조변수 정보가 사용 가능하다면 Ratio method를 사용하는 것이 좋고, 특히 사용 가능한 보조변수가 과거의 자료에 기초한 것이라면 Historical Tend method를 사용하는 것이 좋다. 즉, Historical Trend method는 Ratio method의 특수한 형태라고 할 수 있다. 사용 가능한 보조 변수가 없을 때에는 Current method를 사용해야 하는데 이것은 다른 데이터를 사용해서 얻은 값의 범위와 비교하여 이상치를 판별하는 방법이다. 또한 이상치를 판별하는 기준이 되는 bound는 데이터의 parameter의 함수로 나타낼 수 있는데, 세 가지 방법 모두 사용자가 이러한 모수를 사용해서 outlier intervals의 조정이 가능하다.

4. Proc Errorloc - Error Localization

앞서 Proc Verifyedit 프로시저를 통해 결정된 edit rule을 각 데이터가 만족하지 못한다면, edits를 모두 만족시키기 위해서 데이터의 어떠한 field에서 수정이 이루어져야 하는지 여부가 Error Localization 프로시저를 통해 결정된다. 하지만 실제적인 데이터의 변화는 이루어지지 않는다. 이 과정은 단지 imputation이 필요한 field를 결정해줄 뿐, 실제적인 imputation이 실행되지는 않는다.

Banff는 edit rule을 만족시키기 위해서 수정해야 하는 field를 결정할 때, 최소한의 변화량를 만드는 field의 수를 찾는 것이 아니라, edit rule을 만족시키는 최소한의 field의 수를 정한다. 이러한 방법을 Rule of Minimum Change라고 하는데 이는 원본 데이터를 최대한 보존하기 위한 방법이다. 사용자는 수정을 가할 수 있는 최대한의 field의 수를 지정할 수 있는데, 만약 Banff가 찾은 최소한의 field수가 이를 초과하게 되면 Banff는 답을 제공하지 않고 사용자가 지정한 수를 변화 시킬 것을 알린다.

5. Proc Deterministic - Deterministic Imputation

Proc Deterministic 프로시저에서는 imputation이 필요하다고 판정된 field에 대해서 사용자에 의해 지정된 값을 사용해서 imputation을 실시한다. 이 과정은 Banff 스스로 imputation에 필요한 값을 찾는 것이 아니라 이후에 실시될 다른 imputation 방법에 의해 무응답 대체가 이루어 질 때, imputation이 필요한 부분을 줄이고자 하는 것에 유용하다.

6. Proc Donorimputation - Donor Imputation

무응답 대체가 필요한 단위(recipient)에 적합한 값을 찾기 위해서 Proc Donorimputation 프로시저는 nearest neighbour approach를 사용해 그와 가장 비슷한 값을 선택해서 무응답 대체를 하게 되는데, 사용자에 의해 지정된 post-imputation edits를 건너뛸 수 있도록 하는 적합한 donor를 찾게 된다. 적합한 donor가 선택되면 무응답 대체가 필요한 모든 항목에 대해서 같은 donor에서 얻은 값을 사용하게 되는데 이를 통해 imputation이 이루어지고 난 후에도 항목들 간의 상호관계가 유지될 수 있다.

이 프로시저에서는 각각의 recipient에 대해 어떠한 field를 사용해 donor와의 거리를 계산할 것인지를 결정한다. 이러한 “matching field"는 recipient가 갖고 있는 사용 가능한 값이어야 하는데 경우에 따라서는 이러한 matching field를 갖지 않는 경우도 발생한다.

이렇게 해서 recipient와 donor간의 거리가 계산이 되면 n 개의 가장 가까운 donor가 결정이 되고, 이 중 가장 가까운 donor부터 만약 이 값이 imputation 값으로 쓰이게 되면 recipient가 post-imputation을 필요로 하지 않는지를 확인한다. 이를 만족시키게 되면 imputation이 이루어지고 다음 recipient로 넘어가게 되며, 이 조건이 만족되지 않으면 다음 donor를 사용해 다시 한 번 같은 과정을 반복하게 된다. 이 과정은 적합한 donor를 찾을 때 까지 또는 사용자가 지정한 trial의 최대 반복횟수까지 반복이 된다. 만약 사용자가 지정한 최대 반복횟수에 이를 때 까지 적합한 donor가 결정되지 않는다면 Banff는 이 프로시저를 중단하고 실패 메시지를 내보낸 후 그 다음 recipient로 넘어간다.

만약 적절한 matching field가 존재하지 않는다면 사용자는 사용 가능한 donor들 가운데서 무작위로 하나의 donor를 선택하도록 지정할 수 있다. 이 과정에서도 역시 선택된 donor의 값을 사용해 recipient의 무응답 항목을 대체했을 때 이렇게 대체된 데이터가 post-imputation edits를 필요로 하는지의 여부를 확인해야 한다.

7. Proc Estimator - Estimator Imputation

Proc Estimator 프로시저는 다양한 imputation estimator를 사용해 한 번의 프로시저를 실행하면서 여러 변수들에 대해 imputation을 실행할 수 있다. 만약 첫 번째 시도한 imputation이 성공적으로 수행되지 못하면 또 다른 estimator를 사용해 다시 imputation을 시도하게 된다.

Proc Estimator 프로시저는 estimator function과 linear regression estimator의 두 가지 추정량을 사용한다.

Estimator function에는 mean, ratio, trend 등이 속하고, linear regression estimator는 다음과 같은 선형회귀모형의 형태로 표현된다.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

여기서 y는 대체되어야 하는 값을 나타내고 x_i들은 독립변수를 나타낸다. 그리고 회귀 계수를 나타내는 β_i는 최소 자승법(method of least squares)을 사용해 추정된다.

Donor imputation 프로시저와 달리 Estimator imputation 프로시저에는 post-imputation edits가 존재하지 않는다. 따라서 결과적으로 대체된 값들은 원래의 edit rule을 만족하지 않을 수 도 있으므로 estimator imputation이 모두 끝난 뒤에는 Proc Errorloc 프로시저를 통해 그러한 일이 실제로 벌어지는 지를 확인해 봐야 한다. 또한 Proc Outlier 프로시저를 통해 이상치로 판명된 데이터들은 parameter 의 계산에서 자동적으로 제외된다.

8. Proc Prorate - Pro-Rating

Pro-Rating 프로시저는 부분의 합이 데이터의 전체 합과 일치하는지를 확인하는 과정이며 이때 사용되는 equality edits는 사용자에 의해 지정이 된다. Pro-rating edit rules은 변수들 간의 논리적 포함 관계에 제한이 없기 때문에 각각 개별적인 부분들의 합은 부분 합과 일치해야 하고 그 부분 합들의 합은 그 다음 단계의 합과 일치해야 한다. 전체 합은 정확하다고 가정하고 있기 때문에 개별 항목들만이 수정될 수 있다.

9. Proc Massimputation - Mass Imputation

Proc Massimputation 프로시저는 two-phase survey이 같이 second-phase sample (또는 subsample)에서만 자세한 정보가 조사되는 경우에 특히 유용하다. subsampling weight에 기초한 고전적인 추정량의 계산이 어렵기 때문에 이를 대신하여 second-phase에서 선택되지 않은 단위들의 무응답 정보를 imputation하기 위해서 donor imputation method를 사용한다. 이러한 방법으로 first phase에 속한 모든 표본 단위에 대한 완전한 데이터 파일이 만들어진다. 이러한 mass imputation의 경우에는 imputation이 필요한 데이터는 알려져 있으며 또한 imputation이 필요한 field 역시 알려져 있고 이는 모든 자료에 대해 동일하다. 전체 표본에서 얻은 핵심 정보(core information)들과 sub-sample에서 얻은 추가적인 정보들은 사전에 에디팅이 되고 imputation이 이루어져야 한다. 따라서 Proc Massimputation에서는 post-imputation edits가 필요하지 않다.

Donor를 선택하는 과정은 Proc Donorimputation 프로시저와 거의 비슷하지만 이 두 프로시저에는 한 가지 차이가 존재하는데 이는 post-imputation의 필요 여부 이다. 따라서 Proc Massimputation 프로시저는 가장 가까운 거리에 있는 donor를 선택해 단순히 imputation을 실시하거나, matching field가 존재하지 않을때는 무작위로 donor를 선택해서 imputation을 실시하게 된다.

2.4 결론

무응답 대체는 항목무응답을 처리하는 방법으로 흔히 사용되는 방법으로 종종 주요 보조 변수를 반영하여 그 상관관계를 유지하는 것이 중요하다. 사업체 조사에서는 무응답 대체군(imputation cell)로써 산업 분류와 종업원 규모를 사용하고 동일 산업 분류내에서는 과거자료를 이용한 ratio imputation 이 사용되고 과거 자료가 없는 경우에는 nearest neighbor imputation 과 같은 핫덱 방법이 주로 사용된다. 이러한 무응답 대체법의 결정은 무응답의 형태, 가능한 보조 변수의 종류, 그리고 다른 변수들과의 상관 관계와 관심 항목의 주변 분포 유지 등을 고려하여 결정되고 특히 여러 항목이 결측되는 경우에는 핫덱 대체가 종종 사용된다.

	survey name	imputation methods
Bureau of Labour Statistics	Job Opening and Labour Turnover Survey	nearest neighbour imputation
	Current Employment Survey	ratio imputation
	National Compensation Survey	regression imputation donor imputation
Census Bureau	Annual Retail Trade Survey	ratio imputation
	Current Industrial Reports	ratio imputation
	Service Annual Surveys	ratio imputation regression imputation
	Annual Trade Survey	ratio imputation
Statistic Canada	Survey of Supplier of Business Financing	donor imputation historical imputation
	Survey of Employment, payrolls and hours	ratio imputation mean imputation trend imputation
	Labour force survey	deterministic imputation donor imputation
	National Construction Industry Wage Rate Survey	mean imputation ratio imputation donor imputation

Ⅲ. 지사 매출액 무응답 대체 연구

3.1. 서론

3.1.1. 연구 목표

‘도.소매업 총조사’에서 상당수의 자료가 본지사로 구성되어 있어 무응답 문제의 상당 부분이 본지사 자료에서 발생하게 되고 따라서 이러한 경우에 대한 적절한 무응답 대체법 연구가 필요하다. 이 도.소매업 총조사는「사업체」단위로 조사되지만, 상대적으로 규모가 큰 본사 매출액은 파악이 쉬운 반면 규모가 작은 지사 매출액은 대부분 파악이 불가능 하다. 대개의 경우, 본사는 각 지사의 개별 매출액은 제공하지 않고 전체 지사의 매출액 합계는 파악할 수 있어 본사가 가지고 있는 지사의 인력규모를 이용한 비례배분 방법에 의존하여 지사 매출액을 배분해 왔다. 이러한 비례 배분 방법은 동일한 본사 내에서는 매출이 종업원수에 비례한다는 가정을 바탕으로 한다. 이러한 가정들이 실제로 얼마나 정확한지에 대하여 실제 조사된 자료를 바탕으로 통계학적으로 분석하여 점검하고 지사 매출액에 대해 보다 정확한 추정이 얻어지게 되는 최적의 배분 방법을 제시하고자 한다.

3.1.2. 연구 기본 방향

지사 매출액 대체법에 대한 연구는 새로운 배분법을 제안하고 이를 기존의 인력 비례 배분법과 비교하는 방향으로 이루어진다. 두 방법의 비교는 각 지사의 매출액이 파악된 자료에 근거하여 판단한다. 지사의 매출액이 파악된 자료에서 그 매출액을 모른다고 가정하고 여러 가지 가능한 대체 방법론 중에서 각 방법으로 대체값을 구한 후 실제값과 비교함으로써 더 나은 방법을 결정한다. 이를 위하여 지사의 매출액이 파악된 자료가 필요한데 이 연구에서는 2005년도 지사 매출액이 파악된 본사에 해당하는 지사 자료를 사용하였다.

본 연구를 위해서 기본적으로 사용되어지는 가정은 다음과 같다.

① 본지사 조사가 완료된 업체의 본.지사 매출액 정보 및 인력 자료는 믿을 수 있는 값이다.

② 전체 ‘도.소매업 총조사’ 의 본지사 자료에서 각 본사별 해당 지사들의 매출액

합계와 각 지사별 인력에 대한 값들을 얻을 수 있다.

③ 매출액은 기본적으로 종업원 수의 함수로 표현된다. 각 지사별 매출액에 대한 통계적 모형은 다음과 같다.

$Y_{ij} = f(X_{ij}) = \beta_i X_{ij}^{\alpha} e_{ij}$, $i = 1...14$ (본사수), $j = 1...n_i$ (본사i의 지사수)

(Y_{ij} : 본사 i내 지사 j의 매출액, X_{ij} : 본사 i내 지사 j의 종업원 수, e_{ij} : 오차)

④ 전체 도소매업에서의 매출액과 인력과의 관계는 지사매출액이 파악된 자료-대상 업체에서의 매출액과 인력간의 관계와 같은 구조를 지닌다.

첫 번째 가정은 관측된 지사매출액 자료를 바탕으로 지사 매출이 결측된 사업체의 지사 매출에 대하여 통계적 방법론을 사용함으로써 보정할 수 있는 최소한의 가정이다. 물론 조사된 자료가 모두 정확하다고는 할 수 없지만, 본사내의 지사들의 매출액이 정확히 조사되었다는 가정은 앞으로 진행 될 연구에 대한 신뢰성의 근거가 된다.

두 번째 가정은 무응답 대체법을 적용하는 경우에 사용하는 보조변수(설명변수)로써 종업원 수를 사용하고자 할 때 이 종업원 수는 모든 지사별로 관측된다는 가정이다. 이는 현 조사에서 지사별 종업원 수가 정확히 조사되기 때문에 별 무리가 없는 가정이다.

세 번째 가정은 지사 매출액을 예측하기 위한 가장 설명력 있는 변수는 지사별 종업원 수라는 가정이다. 현실적으로 본사에서 지사들의 정보를 파악할 때 다른 정보보다도 종업원 수를 파악하는 것이 가장 쉽고 정확하다. 물론 지사별 연간 급여액 등 보다 많은 정보를 가진 변수를 사용할 수 있다면 더 좋으나 대부분의 경우에는 종업원 수는 모두 기재되어 있지만 연간급여액은 모두 기재되어 있지 않다. 또한 연간급여액이라는 것은 종업원 수보다 측정오차가 더 생길 가능성이 높으므로 연간 급여액을 바탕으로 모형을 세우는 것은 위험할 수 있다. 그렇기 때문에 종업원 수가 신뢰성 있으며 현실적으로 비교적 쉽게 획득 가능한 정보이므로 이를 사용하여 지사 매출액을 예측하는 것이 타당하다.

네 번째 가정은 파악된 자료는 전체모집단을 대표한다는 것이다. 파악된 자료가 완

전하게 전체를 대표한다고 할 수는 없지만 어느 정도는 전체적인 구조를 반영하고 있다는 점에서 정확한 기준은 아니지만 근사적으로는 타당한 가정이라 볼 수 있다.

위와 같은 가정 하에서 새로운 최적의 무응답 대체 방법을 개발하고 이를 기존의 방법과 비교하고자 한다. 최적의 무응답 대체법을 개발하기 위해서는 지사 매출액과 지사 종업원수와의 관계에 대한 이해가 필요하다. 이를 위해 다음과 같은 방법을 사용하였다.

① 종업원 수를 설명변수로 하는 지사 매출액의 회귀 모형으로서 여러 가지 다양한 모형을 모색해보고 이 중에서 본지사 매출이 모두 파악된 자료에 근거하여 예측력이 가장 좋은 모형을 선택한다.

② 이 모형 하에서 지사 매출액이 파악된 자료를 바탕으로 회귀 계수를 추정하고 각 지사별 매출에 대한 최적의 무응답 대체 회귀 모형을 도출해 본다.

③ 얻어진 무응답 대체 회귀 모형을 기존의 종업원수 비례 배분법을 개선하는 새로운 비례 배분법을 제시하고 비교 평가한다.

④ 지역별 1인당 매출액의 차이를 고려하여 지역별 구분변수를 추가하여 ①~③의 분석을 시행하여 기존 모형과 ③에서 얻어진 모형을 개선하는 최적모형을 제시하고 비교 평가한다.

위의 ①은 예측력을 높일 수 있는 종업원 수와 지사 매출액의 함수 모형을 설정하는 것이고, ②는 회귀 계수 추정을 통하여 조사된 자료의 특성을 파악하여 자료의 특성에 맞는 타당한 모형을 구축하는 것이다. ③은 비례 배분법 보다 더 정확한 새로운 배분법을 제시하고 이를 기존의 비례 배분법과 비교하는 것이다. ④는 각 사업체가 위치한 지역별로 1인당 매출액을 다르게 분배하는 방법을 찾는 것으로 기존 모형과 ③ 모형을 기준으로 비교하는 것이다.

3.1.3. 한계점

① 자료에 대한 정보 미흡

지사 매출액이 파악된 14개 사업체들의 자료가 추출되어진 방법은 알 수가 없어, 이 업체들이 전체 업체를 대표한다고 보기 어렵다. 따라서 주어진 자료를 바탕으로

구축한 모형의 일반화를 정당화하기 힘들다.

② 변수 미흡

지사점별 매출액에 영향을 끼치는 많은 변수들이 존재할 것으로 예상하지만, 현재 주어진 자료에서는 종사자수 이외에 매출액 배분에 유의한 지역, 인건비, 영업비용 등의 추가적으로 사용 가능한 변수가 없다. 이에 따라 지사 매출액에 대한 모형에 사용되는 설명변수의 제약을 가져오게 되었다.

③ 이상한 데이터

지사점별 매출액은 있고 종사자수는 없는 데이터와 그 반대의 경우의 데이터가 자료에 포함되어 있었다. 이러한 자료들은 분석에서 제외하였다. 또한 종사자수에 비해 현저히 적은 매출액 자료는 기입에 오류가 있었던 것으로 판단하여 분석에서 제외하였다.

3.2. 자료 현황

3.2.1. 보유자료

현재 보유하고 있는 자료는 2005년 도소매업 조사 중 본.지사 매출액의 조사가 완료된 자료로서 본사 14개와 각 본사의 지사들로 구성되어 총 2,388개의 지사 자료가 있다. 정확한 방법론에 의한 연구를 위해서 불완전한 지사 매출액 데이터 (4,5,10,13번 회사) 제외한 모든 데이터를 사용하였다. 이 자료는 새로운 지사 무응답 대체법을 개발하고 이를 기존의 방법과 비교 평가하기 위하여 사용되었다.

3.2.2. 전체 자료 기본 분석

① <표 3-1>은 14 개 본사 세분류표로써, 업종에 따라 본사 수와 종업원의 범위를 나타내고 있다. 본사 표기가 없는 자료가 존재하고. 지사수가 19-737개로서 업종에 따라 큰 차이를 보이고 있고, 종업원 수도 1-1030 명으로 큰 차이를 나타내고 있다. 회사 일부는 세분류 표에 기재되지 않았다.

<표 3-1> 본사 세분류표

회사	본/지사수 (개)	이용가능한 변수	종업원수 범위	분류번호	세분류
1	1/119	연간급여액, 종업원수	1~6	22110	서적출판업
				52621	서적 및 잡지류 소매업
2	1/90	급여총액, 종업원수	2~98	92310	가전제품 수리업
3	1/19	급여총액, 종업원수	1~4	36910	귀금속 제조업
				52650	귀금속 소매업
6	1/14	연간급여액, 종업원수	2~9	24312	농약 제조업
				51723	비료 및 농약 도매업
7	1/737	연간급여액, 종업원수	1~35	67199	금융서비스업
8	1/567	연간급여액, 종업원수	2~36	67199	금융서비스업
9	-/234	연간급여액, 종업원수	4~13	분류번호 없음	
11	1/42	연간급여액, 종업원수	7~159	분류번호 없음	
12	-/301	연간급여액, 종업원수	8~1030	분류번호 없음	
14	1/265	종업원수	1~109	분류번호 없음	

② <표 3-2>는 전체 종업원 수의 분포이고, <표 3-3>은 회사별 종업원 수 그룹별 빈도이다. <표 3-3>의 종업원 수를 나눈 기준은 전체 종업원 수를 크기순으로 나열할 때 하위 25%, 50%, 75% 에 대응하는 값이다. 상위 25% 안의 종업원 수는 그 차이가 커서 다시 두 부분으로 나누었다.

<표 3-2> 종업원 수 분포

종업원 수	
분위수	실제값
100% (최대값)	1,030
75% (제 3사분위수)	16
50% (중위수)	11
25% (제 1사분위수)	8
0% (최소값)	1

<표 3-3>을 보면 각 사업체의 특성에 따라 인력규모의 차이가 존재하는 것을 알 수 있다. 특정 업체는 인력 규모가 소규모에 편중되어 있는 반면, 다른 업체는 인력 규모가 대규모에 편중되어 있는 것으로 보인다.

<표 3-3> 회사별 종업원 수 분포

회사	종업원 수 (명)					합계
	1~7	8~11	12~16	17~74	75~1030	
1	119	0	0	0	0	119
2	24	50	10	5	1	90
3	19	0	0	0	0	19
6	11	3	0	0	0	14
7	106	234	312	85	0	737
8	37	314	192	24	0	567
9	166	65	3	0	0	234
11	1	1	0	33	7	42
12	0	3	2	179	117	301
14	78	39	53	92	3	265
합계	561	709	572	418	128	2,388

③ 각 종업원 수 그룹에서의 1인당 매출액 분포는 <표 3-4>와 같다. 종업원수 규모에 따라 지사 수가 다른 것을 감안하면 종업원수가 적은 경우에 1인당 매출액 평균값이 크고, 종업원수가 많은 경우에 1인당 매출액 평균값이 작은 것으로 보아, 종업원 규모가 매출액에 영향을 미칠 수 있을 것이라 예상할 수 있다.

<표 3-4> 종업원 수 그룹별 1인당 매출액 분포 (단위 : 백만원)

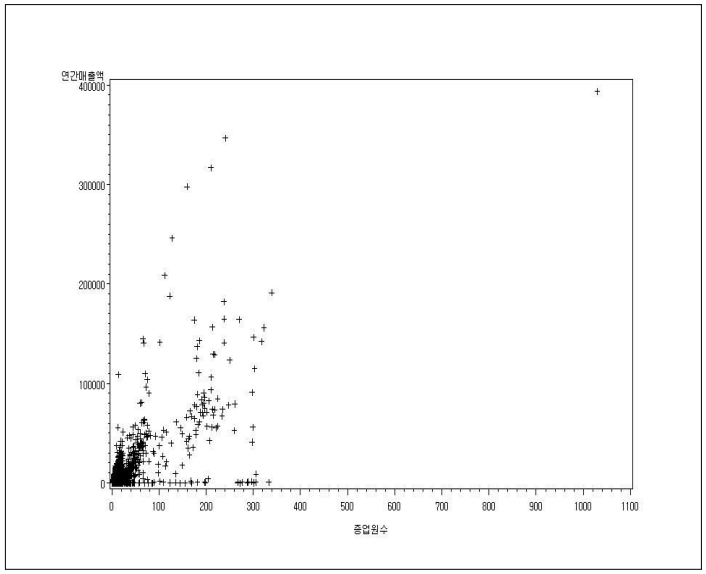
종업원수 (명)	지사 수	평균값 (백만원)	표준편차	합계	최소값	최대값
1~7	561	482	510	270,462	0.096	2,438
8~11	709	479	356	339,901	0.100	3,717
12~16	572	585	458	334,341	0.267	7,765
17~74	418	453	420	189,203	0.034	2,204
75~1030	128	388	396	49,703	0.024	1,925
합계	2,388					

또한 각 회사별 종업원 수 그룹에서의 1인당 연간매출액 평균은 <부록 표 1>과 같다. 각 그룹별로 지사 수가 다른 것을 감안하면, 대체로 종업원 규모가 클 때 1인당 매출액 평균값이 작게 나타난다.

④ 각 회사별 종업원 수와 연간매출액의 이차원적 분석

<그림 3-1>은 전체 사업체에 대한 지사 종업원 수와 지사 매출액을 나타낸다. 이를 보면 지사 종업원 수가 커짐에 따라 지사 매출액도 커지는 경향이 있음을 알 수 있다. (상관계수=0.694) <그림 3-1>에 의해 전체적으로 종업원 수와 지사 매출액의 양의 상관관계가 존재하고, 선형관계보다는 비선형에 더 적합한 것으로 보아, 종업원 수의 함수로 지사 매출액을 설명할 수 있을 것이다.

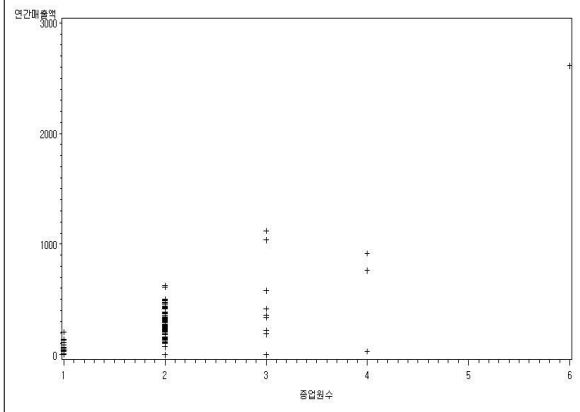
<그림 3-1> 연간매출액과 종업원수간의 산점도
(상관계수 = 0.694)



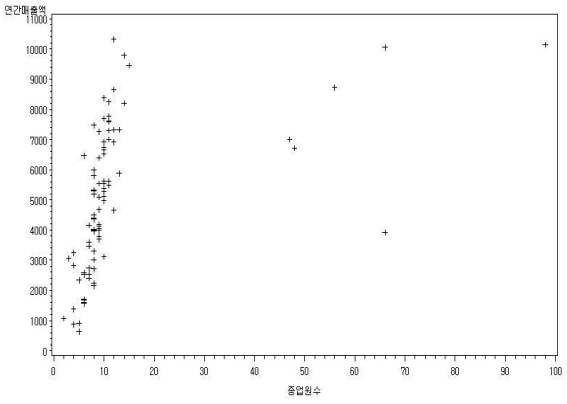
<그림 3-2>는 각 사업체 별 종업원 수와 지사 매출액을 나타낸다. <그림 3-2>에서 보듯 각 사업체에 따라 종업원 수와 지사 매출액의 관계 또한 약간의 차이는 존재하나 모두 양의 상관관계를 나타내고 있다. 이것으로 미루어, 지사 매출액을 추정하는데, 업체 간의 차이는 존재하며 동일 업체 내에서는 인력규모가 매출액에 영향력을 끼칠 수 있을 것으로 보인다.

<그림 3-2> 회사별 매출액과 종업원수간의 산점도

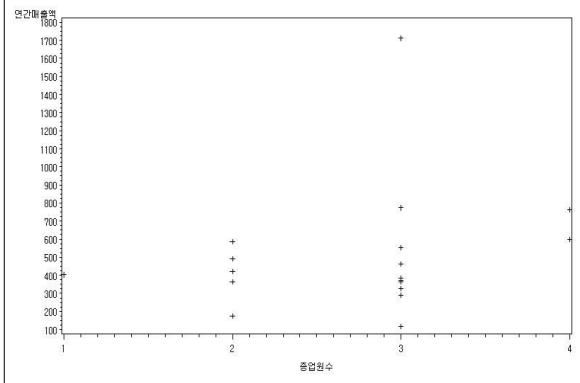
회사1(상관계수=0.709)



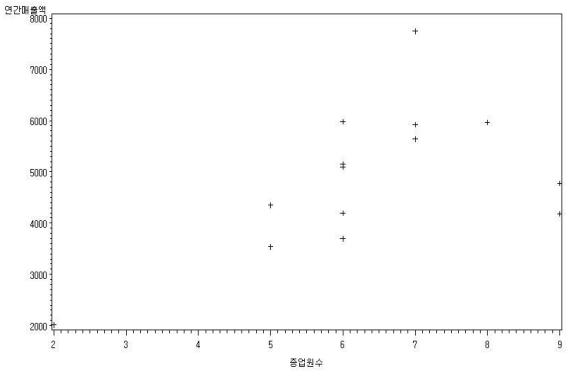
회사2(상관계수=0.447)



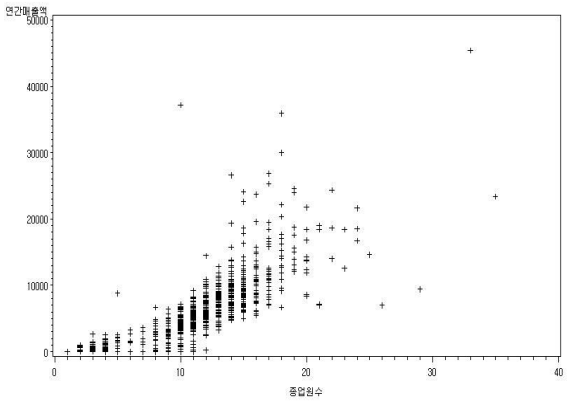
회사3(상관계수=0.229)



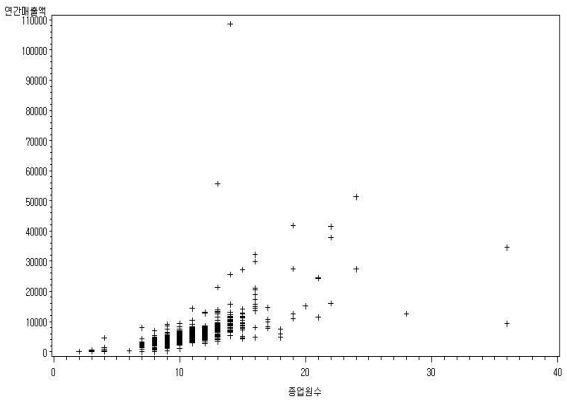
회사6(상관계수=0.554)



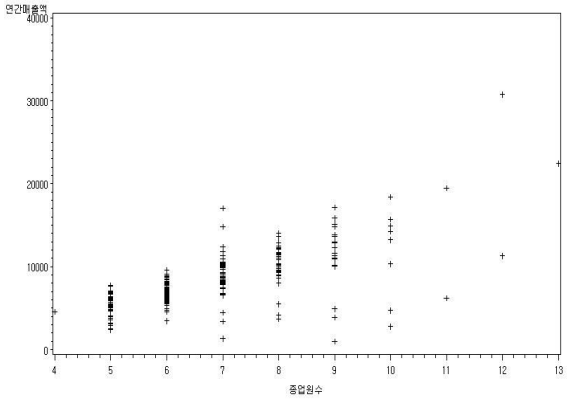
회사7(상관계수=0.767)



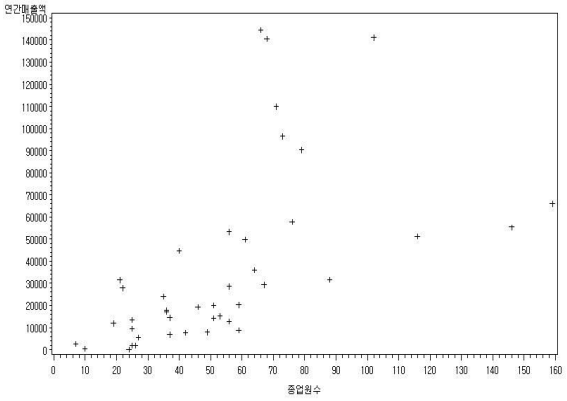
회사8(상관계수=0.567)



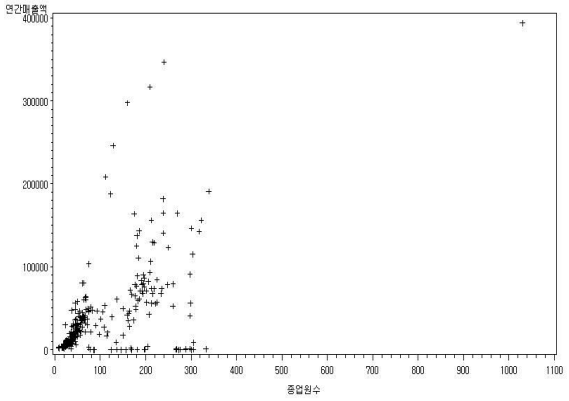
회사9(상관계수=0.676)



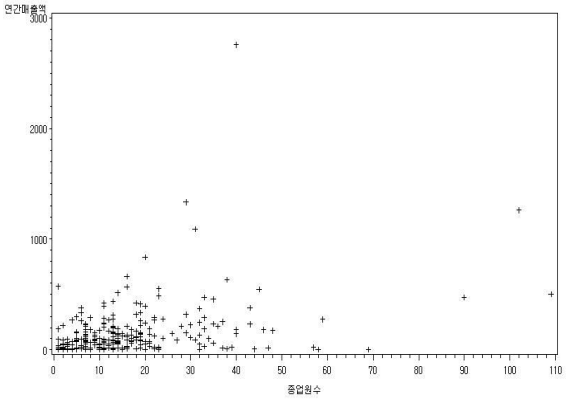
회사11(상관계수=0.559)



회사12(상관계수=0.600)



회사14(상관계수=0.357)



⑤ 각 사업체가 위치한 지역별 1인당 연간매출액의 분포는 <표 3-5>와 같다. 서울 및 기타 지역에 비해 광역시의 사업체수가 적으며, 1인당 연간매출액 평균은 서울, 광역시, 기타 순으로 줄어드는 것으로 보아, 사업체가 위치한 지역이 매출액에 영향을 미칠 수 있을 것이라 예상할 수 있다.

<표 3-5> 지역별 1인당 매출액 분포 (단위: 백만원)

지역	지사 수	평균값 (백만원)	표준편차	합계	최소값	최대값
서울	844	529	475	446,372	0.034	7,765
광역시	471	502	412	236,523	0.024	2,438
기타	1,073	467	415	500,717	0.050	2,192
합계	2,388					

3.3. 지역별 1인당 매출액이 동일하다는 전제하의 방법론

지사매출이 파악된 자료를 대상으로 매출액과 인력 사이의 관계에 대한 다양한 회귀모형을 모색해 본다. 이 중 최적의 모형을 선택한 후 이 모형에 근거하여 기존의 사용하고 있는 비보정 방법(종업원 수를 이용한 비례 배분법)을 살펴보고 이를 간단히 수정할 수 있는 승수 비보정 방법을 제시한다. 먼저 지역별 1인당 매출액이 큰 차이가 없다는 전제하에 방법을 모색해 보고 나아가 이를 매출액 차이를 허용하여 수정해 본다.

표 <3-5> 의 지역별 1인당 매출액을 보면 서울 (529백만), 광역시 (502백만), 기타 지역(467백만)에 따라 약간의 차이가 있으나 표준편차값이 400만 보다 큰 것을 감안할 때 이 차이는 통계적으로 유의한 차이가 아니라 판단된다. 따라서 지역별 1인당 매출액이 같다는 전제하에 무응답 자료의 매출액 예측 방법을 모색해본다.

3.3.1. 회귀분석 방법

지사매출액이 파악된 자료를 대상으로 매출액과 인력과의 관계에 대한 다양한 통계모형을 모색해본다. 여기서 얻어지는 통계적 방법론은 실제 상황에서는 지사매출액자료가 없기 때문에 사용 불가능하지만 이때 얻어진 통계 모형은 매출액과 인력 사이에 관련구조에 대한 우리의 이해를 증가시켜 현재 사용 중인 비보정법을 개선시킬 수 있

는 새로운 비보정법인 승수 비보정법을 제시해준다.

<그림 3-1>, <그림 3-2>에서 살펴본 바와 같이 연간매출액과 종업원 수는 상당히 높은 상관관계를 가지고 있으므로, 회귀분석 방법을 통해 지사점 매출액과 지사점 종업원수간의 관계를 살펴보자.

지사점 연간매출액을 설명할 수 있는 이용 가능한 변수로 종업원 수와 회사별 특성을 이용해 여러 모형을 세워보았다. 이들 모형을 이용하여 지사 자료를 분석하여 자료를 가장 잘 설명해주는 모형을 선택한 후 이를 집중적으로 검토한다. 설정한 모형의 설명력은 adj R2(수정된 결정계수)를 통해 비교할 수 있다.

①모형 설정

고려하는 모형은 다음의 4가지다. (Y : 연간매출액, X : 종업원 수, i : 회사, j : 지사) 이들 모형의 추정에는 Y_{ij} , 즉 지사별 매출액 구조가 필요함을 유의하자. 이는 일반적으로는 적용 불가능하지만 확보한 지사자료에는 적용가능하고 Y_{ij}와 X_{ij} 사이의 관계에 대해 우리의 이해를 높여줄 것이다.

모형 1 : $Y_{ij} = \delta_i + \alpha X_{ij} + u_{ij}$	$u_{ij} \sim N(0\,,\,\sigma^2)$
모형 2 : $Y_{ij} = \delta_i + \alpha_1 X_{ij} + \alpha_2 X_{ij}^2 + u_{ij}$	$u_{ij} \sim N(0\,,\,\sigma^2)$
모형 3 : $\log Y_{ij} = \delta_i + \alpha \log X_{ij} + u_{ij}$	$u_{ij} \sim N(0\,,\,\sigma^2)$
모형 4 : $\log Y_{ij} = \delta_i + \alpha_1 X_{ij} + \alpha_2 X_{ij}^2 + u_{ij}$	$u_{ij} \sim N(0\,,\,\sigma^2)$

모형 1은 연간매출액을 설명하는 변수로 종업원 수에 회사별 특성을 추가한 것으로, 회사별 특성은 분석 대상 자료인 10개 회사의 더미변수로 설정하였다. 이 때, 회사 수만큼 더미변수를 만들어주므로 절편항은 제외한다.

모형 2는 종업원 수와 지사점 매출액간의 선형 관계보다는 비선형 관계가 존재할 것으로 예상하여(<그림 3-1, 3-2>참조), 모형1에 종업원 수 2차항을 추가해 보았다.

모형 3은 평균에서 멀리 떨어진 극단적인 점들의 차이를 줄여주어 자료가 정규분포에 근접하도록 연간매출액과 종업원 수를 로그변환한 뒤, log(연간매출액)을 설명하는

변수로 log(종업원수)와 회사별 특성을 고려한 모형이다.

모형 4는 모형 2에서 매출액만 로그변환한 모형이다.

② 모형 결정

각 모형의 adj R2(수정된 결정계수)는 다음에 나타나 있다.

<표 3-6> 각 모형별 결정계수 비교

	adj R^2
모형 1	0.5822
모형 2	0.5825
모형 3	**0.9784**
모형 4	0.9733

<표 3-6>에서 모형의 설명력을 나타내는 'adj R2'가 모형 3에서 가장 높으므로, 모형3을 이용해 더욱 체계적인 분석을 시행한다.

이제 모형 3을 이용하여 좀 더 자세한 검토를 한다. 이 모형은 다음과 같이 표현할 수 있다. (10개 회사별 더미변수를 고려하여 순수한 절편항은 제외한 모형)

$\log Y_{ij} = \log\beta_i + \alpha \log X_{ij} + u_{ij}$, $i = 1...14$(본사 수), $j = 1...n_i$ (본사 i의 지사 수)

여기서 $\beta_i = \exp(\delta_i)$

(Y_{ij} : 본사 i내 지사 j의 매출액, X_{ij} : 본사 i내 지사 j의 종업원 수,

u_{ij} : 오차 $\sim N(0, \sigma^2)$)

이 로그모형을 다시 표현하면 다음과 같음을 유의하자.

$$Y_{ij} = f(X_{ij}) = \beta_i X_{ij}^{\alpha} e_{ij}$$

여기서 $e_{ij} = \exp(u_{ij})$이다.

이 모형의 회귀분석 결과에 따른 계수추정치는 <표 3-7>과 같다.

<표 3-7> 로그변환 회귀모형 결과

adj R^2 = 0.97837

계수 추정치			
변수	자유도	계수 추정치	p-value
logx	1	1.08283	<.0001
회사1	1	4.43475	<.0001
회사2	1	5.96171	<.0001
회사3	1	5.01725	<.0001
회사6	1	6.50066	<.0001
회사7	1	5.69978	<.0001
회사8	1	5.88149	<.0001
회사9	1	6.87738	<.0001
회사11	1	5.70575	<.0001
회사12	1	5.06278	<.0001
회사14	1	1.48377	<.0001

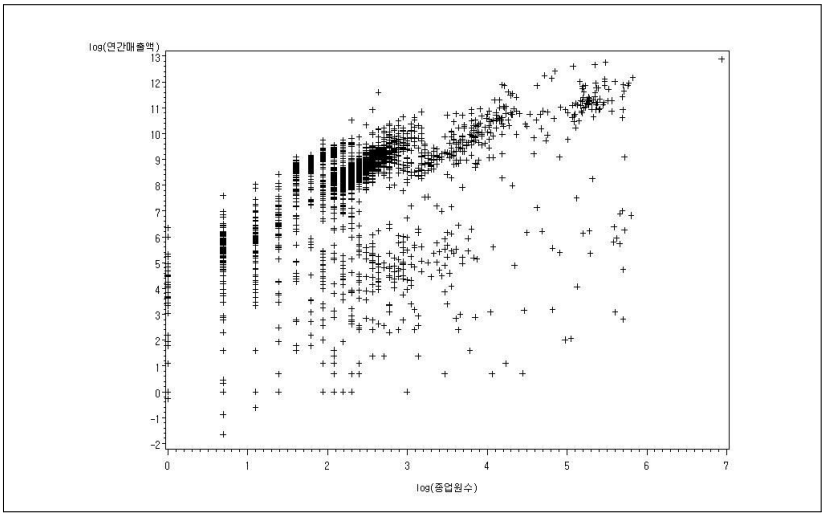
회귀식은

$$\log \widehat{Y_{ij}} = \widehat{\log \beta_i} + \hat{\alpha} \bullet \log X_{ij} = \widehat{\log \beta_i} + \underset{(0.0427)}{1.083} \log X_{ij}$$

가 되며 이들 추정계수는 각각의 p-value가 통상적인 유의수준 0.05보다 작아 의미가 있다(Significant)는 것으로 나타났다.

이 추정식이 개선의 여지가 있는지를 살펴보기 위해 $\log Y_{ij}$와 $\log X_{ij}$의 산점도를 그려보았다.

<그림 3-3> log(연간매출액)과 log(종업원수) 간의 산점도
(상관계수=0.469)



<그림 3-3>을 살펴보면 종업원수가 작은 경우 큰 기울기를 갖고 종업원수가 큰 경우는 작은 기울기를 가짐을 알 수 있다. 따라서 종업원 규모를 나누어 회귀분석을 시행하는 것이 타당하다.

즉 다음과 같은 종업원수 규모에 따라 서로 다른 기울기를 갖는 회귀모형을 검토한다.

$$\begin{aligned} \log Y_{ij} &= \log\beta_{1i} + \alpha_1 \log X_{ij} \ , & X_{ij} \le A \\ &= \log\beta_{2i} + \alpha_2 \log X_{ij} \ , & X_{ij} > A \end{aligned}$$

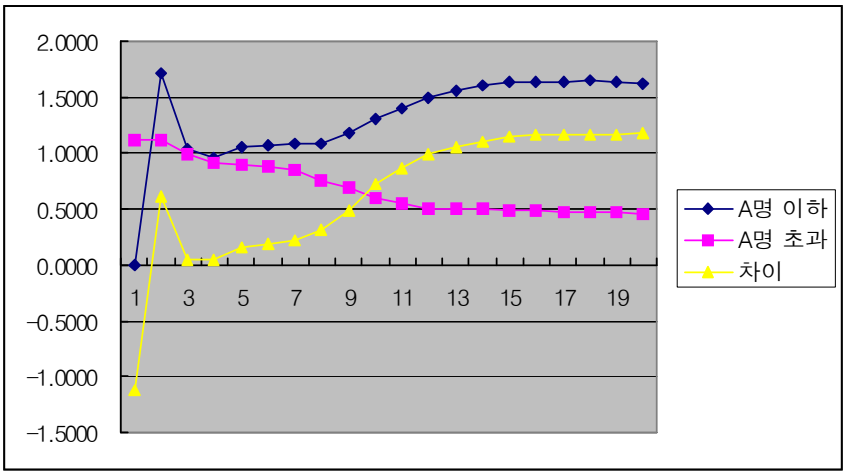
여기서 A는 종업원수 규모를 나누는 기준이다.

③ 종업원수의 소규모와 대규모의 기준 A를 알 수 없으므로, 우선 종업원수 20명까지의 여러 가지 기준으로 규모를 나누어 보았다. <표 3-8>은 규모별에 따른 회귀선 기울기의 추정치와 모형적합도이다.

<표 3-8> 규모 구분에 따른 추정결과

종업원수(A)	회귀선 기울기의 추정치		$\hat{\alpha}_1 - \hat{\alpha}_2$(차이)
	$\hat{\alpha}_1$(A명 이하)	$\hat{\alpha}_2$(A명 초과)	
1	0.0000	1.1150	-1.1150
2	1.7186	1.1095	0.6091
3	1.0347	0.9947	0.0400
4	0.9542	0.9119	0.0423
5	1.0481	0.8968	0.1513
6	1.0667	0.8819	0.1848
7	1.0799	0.8569	0.2230
8	1.0802	0.7606	0.3196
9	1.1793	0.6855	0.4938
10	**1.3090**	**0.5914**	**0.7176**
11	1.4000	0.5418	0.8582
12	1.5020	0.5101	0.9919
13	1.5556	0.5012	1.0544
14	1.6127	0.5042	1.1085
15	1.6369	0.4927	1.1443
16	1.6361	0.4792	1.1568
17	1.6410	0.4723	1.1688
18	1.6456	0.4747	1.1709
19	1.6404	0.4726	1.1678
20	1.6290	0.4543	1.1747

<그림 3-4> 규모 구분에 따른 추정치의 변화



두 회귀선의 기울기 차이가 크게 되는 A 를 선택하는 것이 타당하다. 추정결과 <그림 3-4>를 보면 A가 커짐에 따라 차이가 급격히 커지다가 10 이후로는 안정적인 값을 갖는다. 따라서 A = 10 은 적절한 선택이 될 것이다. 그러므로 추정된 모형은

$$\log Y_{ij} = \log\beta_{1i} + \underset{(0.0899)}{1.3090}\log X_{ij}, \qquad X_{ij} \le 10$$

$$= \log\beta_{2i} + \underset{(0.0627)}{0.5914}\log X_{ij}, \qquad X_{ij} > 10$$

이 된다. 이 모형에 의하면 종업원수 규모가 10이상인 지사의 경우

$$\log Y_{ij} = \log(\beta_{2i} X_{ij}^{0.5914})$$

즉

$$Y_{ij} = \beta_{2i} X_{ij}^{0.5914}$$

이 되어 종업원수 X_{ij} 에 의해 매출액 Y_{ij} 를 예측하고자 할때 종업원수 에 비례해서 배분하기 보다는 승수 0,5914 를 적용하여 승수 적용 종업원수 $X_{ij}^{0.5914}$ 에 비례해서 배분하는 것이 적당하다는 것을 시사해준다. 이는 종웝원 규모가 큰 경우 매출액을 종업원수에 비례해서 배분하는 것보다는 좀 작게 배분하는 것을 의미한다.

3.3.2. 승수 비보정 방법

- 종업원 수에 의한 지사별 매출액 무응답 대체

앞에서 검토한 회귀분석법은 지사별 매출액 자료가 있을 때에만 구현 가능한 방법이다. 그러나 실제적으로는 본사별로 지사별 총합계 매출액만 보고되기 때문에 회귀분석법은 사용 불가능하다. 여기서는 먼저 기존에 사용하는 비보정법을 살펴보고 이를 앞의 회귀분석법과 연관하여 검토한 후 새로운 비보정법을 제시한다.

① 기존에 사용하던 Ratio 방법은 본·지사점 별로 모두 조사가 완료된 종업원 수를 보조변수로 활용하는 것으로, 종업원 수가 늘어남에 따라 연간매출액이 일정하게 증가한다고 가정하여 매출액을 각 지점별 종업원수에 비례해서 배분하는 방법이다.

기존의 Ratio 방법의 정의는 다음과 같다.

업체내 1인당 매출액 = 업체내 지점 총 매출액 합계 /업체내 지점 종업원수 합계
해당 지점 매출액 = 업체내 1인당 매출액 × 해당 지점 종업원수

이를 식으로 쓰면 다음과 같다. 먼저 다음과 같은 변수를 정의한다.

X_{ij} = 본사 i 내 지사 j 의 종업원수

Y_{ij} = 본사 i 내 지사 j 의 매출액

n_{ij} = 본사 i 의 지사수

그러면 본사 i의 총인력과 총매출은 각각

$$X_{i.} = \sum_{j=1}^{n_i} X_{ij}$$

$$Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}$$

와 같이 표현된다. 본사 i 의 일인당 매출액은

$$\hat{\beta}_i = Y_{i.}/X_{i.}$$

이고 무응답된 본사 i 내 지사 j의 매출액 Y_{ij}의 추정치는

$$\widehat{Y_{ij}} = X_{ij} \cdot \hat{\beta}_i$$

가 된다. 이들을 정리하면 다음과 같다.

$$\widehat{Y_{ij}} = X_{ij} \times \frac{Y_{i.}}{X_{i.}}$$

이는 지사 종원원수와 회사전체 종업원수의 단순 비에 따라 회사 전체 매출액을 비례배분하는 방법이므로 **단순비보정법**이라 부를 수 있을 것이다.

② 3.3.1.의 회귀분석법에서 검토한 방법 중 최선의 방법은 log-log 모형이다. 이를 달리 표현하면

$$Y_{ij} = \beta_i X_{ij}^{\alpha}$$

이다. 그런데 기존의 비보정법에서 상정하는 모형은

$$Y_{ij} = \beta_i X_{ij}$$

로 위 모형에서 승수 $\alpha = 1$인 경우에 해당한다.

따라서 일반적으로 우리는 모형 $Y_{ij} = \beta_i X_{ij}^{\alpha}$ 에 적용할 수 있는 새로운 비보정법을 제시하고 이를 **승수 비보정법**이라 칭한다.

연간매출액 증가율이 종업원 수에 관계없이 일정한 것이 아니라 종업원 규모에 따라 차이가 있다는 것이 더 합리적이므로, 이 승수모형은 비보정모형보다 현실성을 좀 더 감안한 모형이라 볼 수 있다.

새로운 승수 보정 배분 모형에서 무응답 지사매출액은 다음과 같이 추정한다.

업체내 1인당 매출액: $\hat{\beta}_i = Y_{i.}/X_{i.}^{(\alpha)}$ 여기서 $X_{i.}^{(\alpha)} = \sum_{j=1}^{n_i} X_{ij}^{\alpha}$ 이다.

지점별 매출액 = 업체내 1인당 매출액 × 지사점 승수 보정 종업원수
즉, 본사 i 내 지사 j 의 예측 매출액은

$$\widehat{Y}_{ij} = X_{ij}^{\alpha} \times \frac{Y_{i.}}{X_{i.}^{(\alpha)}}$$

으로 표현된다.

③ 기존 비 보정방법과 승수 비보정 방법의 비교

기존 비보정 방법을 보완한 승수-비보정 방법을 적용하여, 실제 지사별 매출액의 배분이 최적으로 이루어지는 승수 값을 찾아보았다. 실제 지사 매출액과 추정된 지사 매출액을 각 회사별로 예측력 판단 측도인 CV [2], adj R^2 [3]을 구해서 평균값을 구해 보았다.

첫 번째 판단측도인 CV 는 각 회사별 '실제 연간매출액/추정한 연간매출액'의 CV 를 계산해 평균을 구한 것이다. 구체적으로는 다음과 같이 구한다. 먼저 각 회사별로

CV_i = i 번째 회사의 (실제 지사매출액 / 추정 지사매출액) 의 CV

즉 $Z_{ij} = Y_{ij}/\widehat{Y}_{ij}$ 라 하고 $\overline{Z_i}$, S_i 를 Z_{ij}, $j = 1,...,n_i$ 의 표본 평균, 표본 표준편차 라 할 때

$$CV_i = 100 \times S_i/\overline{Z_i}$$

이고 CV 는 CV_i 들의 단순평균이다. 따라서 CV 값이 작을수록 Z_{ij} 의 변동이 작고 결국 실제지사매출액과 추정 지사매출액이 비슷하게 되어 $\widehat{Y}_{ij}$ 을 계산한 예측방법의 예측력은 뛰어나다고 할 수 있다.

두 번째 판단측도인 adj R^2 (Adjusted R^2) 는 회사별로 Y_{ij} 를 $\widehat{Y}_{ij}$ 에 상수항 없

2) CV : 변동계수, 상대퍼짐성에 대한 측도. 작을수록 평균에 가깝게 분포.
3) adj R^2 : 수정된 결정계수, 모형 적합의 측도. 클수록 모형의 설명력이 높음.

이 단순회귀 분석을 실시할 경우의 Adjusted R^2 인 $adj\ R_i^2$ 를 구한 후 이들의 평균을 구한 것이 adj R^2 이다. 따라서 adj R^2 값이 클수록 Y_{ij} 과 $\widehat{Y}_{ij}$ 은 비슷한 값을 갖게 되어 $\widehat{Y}_{ij}$ 을 계산한 예측방법의 예측력은 뛰어나다고 할 수 있다.

<표 3-9> 승수 변화에 따른 통계량 변화

승수	CV의 평균	adj R^2 의 평균
0.1	84.52	0.6107
0.2	80.83	0.6311
0.3	77.68	0.6485
0.4	75.16	0.6623
0.5	73.42	0.6721
0.6	**72.57**	0.6778
0.7	72.72	**0.6797**
0.8	73.87	0.6782
0.9	76.00	0.6741
1	79.05	0.6682
1.1	82.93	0.6611
1.2	87.53	0.6531

<표 3-9>를 보면, 승수가 1인 경우가 기존의 비례배분에 해당 된다. 각 회사별 '실제 연간매출액/추정한 연간매출액'의 CV를 계산해 평균을 구한 'CV 의 평균'은 승수가 0.6 일 때 가장 작고, 추정한 매출액이 실제 연간매출액을 얼마나 설명하고 있는지를 나타내는 adj R^2를 회사별로 계산해 평균을 구한 'adj R^2 의 평균'은 승수가 0.7 일 때 가장 높다.

따라서 최적 승수는 0.6과 0.7 사이에 있음을 알 수 있으며, adj R^2 를 기준으로 종업원 수에 0.7승을 한 0.7-승수 비보정 방법이 기존의 비례배분 방법보다 예측력이 높다고 할 수 있다. 이 결과는 앞의 회귀 분석에서 10명보다 큰 중규모이상의 인력규모 를 갖는 지사에 대한 매출액 예측에서 추정승수가 0.59 임과 비슷한 점을 주목한다.

그러므로 승수비보정 방법에서 파악되는 모형은 회귀모형에서 중규모이상의 인력규모 를 갖는 지사에 대한 모형에 상응함을 알 수 있다.

승수 $\hat{\alpha}$ 은 다음과 같은 탄력성의 의미를 갖는다.

$$\hat{\alpha} \cong \frac{\text{매출액의 증가율}}{\text{종업원수의 증가율}}$$

즉, 종업원 수가 1%증가할 때 연간매출액이 $\hat{\alpha}$%만큼 변화한다는 뜻이다. 그래서 종업원 수가 1% 증가 할 때 0.7% 매출액이 증가한다는 것을 의미한다.

3.4. 지역별 매출액 차이를 고려한 방법론

<표 3-5> 를 살펴보면 지역별로 1인당 매출액이 차이가 있음을 알 수 있다. 즉 서울이 가장 크고 기타 지역이 가장 작다. 이러한 지역별 1인당 매출액 차이를 모형에 반영시켜 앞 3.3. 절의 승수보정 지사 매출액 예측 방법을 수정한 후 어느 정도의 개선이 있는 가 살펴보고자한다.

3.4.1. 회귀분석 방법

①모형 설정

고려하는 모형은 다음의 4가지로, 앞의 3.3.1. 절의 모형에 사업체의 지역별 효과를 보기 위해 지역변수를 추가한다. 이 때, 모형에 추가한 지역변수는 사업체의 위치에 따라 서울/광역시/기타로 구분하는 더미변수 L_1, L_2 두 개 이다.

더미변수 L_1은 서울을 나타내며 서울의 경우 1 이고 그 이외의 지역은 0 이다.

더미변수 L_2는 광역시를 나타내며 광역시의 경우 1 이고 그 이외의 지역은 0 이다.

기타지역은 L_1, L_2 모두 0 인 경우에 해당된다.

모형 1 : $Y_{ij} = \delta_i + \alpha X_{ij} + \eta_1 L_{1ij} + \eta_2 L_{2ij} + u_{ij}$	$u_{ij} \sim N(0\,,\,\sigma^2)$
모형 2 : $Y_{ij} = \delta_i + \alpha_1 X_{ij} + \alpha_2 X_{ij}^2 + \eta_1 L_{1ij} + \eta_2 L_{2ij} + u_{ij}$	$u_{ij} \sim N(0\,,\,\sigma^2)$
모형 3 : $\log Y_{ij} = \delta_i + \alpha \log X_{ij} + \eta_1 L_{1ij} + \eta_2 L_{2ij} + u_{ij}$	$u_{ij} \sim N(0\,,\,\sigma^2)$
모형 4 : $\log Y_{ij} = \delta_i + \alpha_1 X_{ij} + \alpha_2 X_{ij}^2 + \eta_1 L_{1ij} + \eta_2 L_{2ij} + u_{ij}$	$u_{ij} \sim N(0\,,\,\sigma^2)$

모형 1은 연간매출액을 설명하는 변수로 종업원 수에 회사별 특성과 지역변수를 추가한 것으로, 회사별 특성은 분석 대상 자료인 10개 회사의 더미변수로 설정하고, 지역변수는 각각 서울, 광역시를 나타내는 더미변수를 사용한다. 이 때, 회사 수만큼 더미변수를 만들어주므로 절편항은 제외한다.

모형 2는 종업원 수와 지사점 매출액간의 선형 관계보다는 비선형 관계가 존재할 것으로 예상하여(<그림 3-1, 3-2>참조), 모형1에 종업원 수 2차항을 추가해 보았다.

모형 3은 평균에서 멀리 떨어진 극단적인 점들의 차이를 줄여주어 자료가 정규분포에 근접하도록 연간매출액과 종업원 수를 로그변환한 뒤, log(연간매출액)을 설명하는 변수로 log(종업원수)와 회사별 특성, 그리고 지역변수를 고려한 모형이다.

모형 4는 모형 2에서 매출액만 로그변환한 모형이다.

② 모형 결정

각 모형의 adj R^2(수정된 결정계수)는 다음에 나타나 있다.

<표 3-10> 각 모형별 결정계수 비교

	adj R^2
모형 1	0.5844
모형 2	0.5848
모형 3	**0.9784**
모형 4	0.9734

<표 3-10>에서 모형의 설명력을 나타내는 'adj R^{2}'가 모형 3에서 가장 높으므로, 모형3을 이용해 더욱 체계적인 분석을 시행한다.

이제 모형 3을 이용하여 좀 더 자세한 검토를 한다. 이 모형은 다음과 같이 표현할 수 있다. (두 개의 지역변수와 10개 회사별 더미변수를 고려하여 순수한 절편항은 제외한 모형)

$\log Y_{ij} = \log \beta_i + \alpha \log X_{ij} + \eta_1 L_{1ij} + \eta_2 L_{2ij} + u_{ij}$,

$i = 1...14$ (본사수), $j = 1...n_i$ (본사i의 지사수)

여기서 $\beta_i = \exp(\delta_i)$

(Y_{ij} : 본사 i내 지사 j의 매출액, X_{ij} : 본사 i내 지사 j의 종업원 수,

L_1 : 사업체가 서울에 있으면 1, 아니면 0

L_2 : 사업체가 광역시에 있으면 1, 아니면 0

u_{ij} : 오차 $\sim N(0, \sigma^2)$)

이 로그모형을 다시 표현하면 다음과 같음을 유의하자.

$$Y_{ij} = f(X_{ij}) = \beta_i X_{ij}^{\alpha} \bullet \exp(\eta_1 L_{1ij} + \eta_2 L_{2ij}) e_{ij}$$

여기서 $e_{ij} = \exp(u_{ij})$이다.

이 모형의 회귀분석 결과에 따른 계수추정치는 <표 3-11>과 같다.

<표 3-11> 로그변환 회귀모형 결과
(추가로 지역별 더미변수 고려한 경우
-서울/광역시/기타)

adj R^2 = 0.97840

계수 추정치			
변수	자유도	계수 추정치	p-value
logx	1	1.08987	<.0001
회사1	1	4.47628	<.0001
회사2	1	5.99342	<.0001
회사3	1	5.06477	<.0001
회사6	1	6.52282	<.0001
회사7	1	5.73034	<.0001
회사8	1	5.91697	<.0001
회사9	1	6.91108	<.0001
회사11	1	5.69055	<.0001
회사12	1	5.07206	<.0001
회사14	1	1.49917	<.0001
서울	1	-0.03654	0.5511
광역시	1	-0.16251	0.0168

이 때, 회귀식은

$$\log \widehat{Y_{ij}} = \widehat{\log \beta_i} + \hat{\alpha} \bullet \log X_{ij} + \hat{\eta_1} L_{1ij} + \hat{\eta_2} L_{2ij}$$

$$= \widehat{\log \beta_i} + 1.083 \log X_{ij} - 0.03654 L_{1ij} - 0.16251 L_{2ij}$$
$$(0.043) \qquad (0.0613) \qquad (0.0679)$$

$$\Rightarrow \begin{cases} \text{서울}(L_1 = 1,\ L_2 = 0): & \log \widehat{Y_{ij}} = \widehat{\log \beta_i} + 1.083 \log X_{ij} - 0.03654 \\ \text{광역시}(L_1 = 0,\ L_2 = 1): & \log \widehat{Y_{ij}} = \widehat{\log \beta_i} + 1.083 \log X_{ij} - 0.16251 \\ \text{기타}(L_1 = 0,\ L_2 = 0): & \log \widehat{Y_{ij}} = \widehat{\log \beta_i} + 1.083 \log X_{ij} \end{cases}$$

가 되며 이들 추정계수는 서울의 더미변수를 제외하고는 각각의 p-value가 통상적인 유의수준 0.05보다 작아 의미가 있다(Significant)는 것으로 나타났다. 특히 광역시를 나타내는 더미변수의 계수의 유의확률은 0.0168로써 유의수준 5%에서 유의함을 알 수 있다. 따라서 지역을 나타내는 더미 변수는 지사 매출을 예측함에 있어서 종업원 수에 부가되는 추가 예측력을 지녔다고 할 수 있다.

앞에서 이미 확인한 것처럼 <그림 3-3>를 보면 종업원 규모에 따라 기울기가 다르므로, 종업원수 규모에 따라 서로 다른 기울기를 갖는 회귀모형을 검토한다.

$$\begin{aligned} \log Y_{ij} &= \log \beta_{1i} + \alpha_1 \log X_{ij} + \eta_{11} L_{1ij} + \eta_{12} L_{2ij}\ , & X_{ij} \le A \\ &= \log \beta_{2i} + \alpha_2 \log X_{ij} + \eta_{21} L_{1ij} + \eta_{22} L_{2ij}\ , & X_{ij} > A \end{aligned}$$

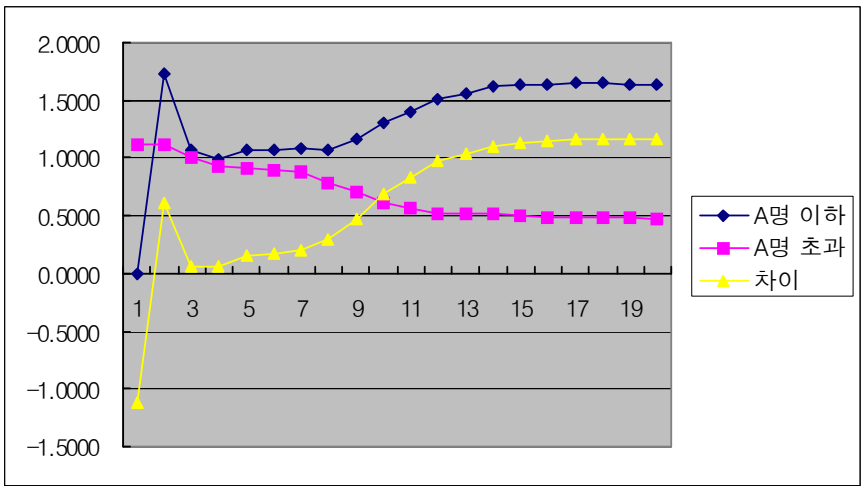
여기서 A는 종업원수 규모를 나누는 기준이다.

③ 종업원수의 소규모와 대규모의 기준 A를 알 수 없으므로, 우선 종업원수 20명까지의 여러 가지 기준으로 규모를 나누어 보았다. <표 3-12>는 규모별에 따른 회귀선 기울기의 추정치와 모형적합도이다.

<표 3-12> 규모 구분에 따른 추정결과

종업원수(A)	회귀선 기울기의 추정치		$\widehat{\alpha_1}-\widehat{\alpha_2}$(차이)
	$\widehat{\alpha_1}$(A명 이하)	$\widehat{\alpha_2}$(A명 초과)	
1	0.0000	1.1237	-1.1237
2	1.7335	1.1158	0.6177
3	1.0714	1.0032	0.0682
4	0.9898	0.9220	0.0677
5	1.0697	0.9095	0.1602
6	1.0737	0.8983	0.1755
7	1.0813	0.8752	0.2061
8	1.0746	0.7802	0.2944
9	1.1721	0.7075	0.4646
10	**1.3038**	**0.6183**	**0.6855**
11	1.4024	0.5622	0.8402
12	1.5080	0.5251	0.9829
13	1.5613	0.5155	1.0458
14	1.6187	0.5134	1.1053
15	1.6420	0.5034	1.1386
16	1.6412	0.4861	1.1551
17	1.6453	0.4802	1.1651
18	1.6491	0.4852	1.1639
19	1.6431	0.4847	1.1584
20	1.6309	0.4651	1.1658

<그림 3-5> 규모 구분에 따른 추정치의 변화



두 회귀선의 기울기 차이가 크게 되는 A 를 선택하는 것이 타당하다. 추정결과 <그림 3-5>를 보면 A가 커짐에 따라 차이가 급격히 커지다가 10 이후로는 안정적인 값

을 갖는다. 따라서 A = 10 은 적절한 선택이 될 것이다. 그러므로 추정된 모형은

$$\log Y_{ij} = \log\beta_{1i} + 1.3038\log X_{ij} + 0.03848 L_{1ij} - 0.05793 L_{2ij}\ , \qquad X_{ij} \le 10$$
$$(0.09022) \qquad (0.08876) \qquad (0.09295)$$
$$= \log\beta_{2i} + 0.6183\log X_{ij} - 0.16435 L_{1ij} - 0.21559 L_{2ij}, \qquad X_{ij} > 10$$
$$(0.06342) \qquad (0.07747) \qquad (0.08991)$$

이 된다. 추정된 모형에 의하면 $\log(X_{ij})$의 계수 즉 X_{ij} 의 승수가 지역더미 변수를 넣지 않았을 때인 3.3.1 절에서의 추정계수값이 비슷하다. 종업원수가 10인 이상의 경우 3.3.1 절에서의 계수는 0.5914 이고 위 모형에서는 0.6183이다. 3.3.1 절에서와 마찬가지로 이는 종업원 규모가 큰 경우 매출액을 종업원수에 비례해서 배분하는 것보다는 좀 작게 배분하는 것을 의미한다. 또 하나 주목할 점은 $X_{ij} \leqq 10$ 일 때보다 $X_{ij} > 10$ 인 경우 지역더미 변수의 계수값들의 절대값이 크다는 사실이다. 이는 종원원수 규모가 큰 경우가 지역효과가 더 크다는 것을 의미한다.

3.4.2. 지역감안 승수 비보정법

① 앞 절에서 검토해본 모형 중 데이터에 가장 부합하는 모형은

$$\log Y_{ij} = \log\beta_i + \alpha\log X_{ij} + \eta_1 L_{1ij} + \eta_2 L_{2ij} + u_{ij}$$

이다. 이를 달리 표현하면

서울의 경우

$$Y_{ij} = c_1\beta_i X_{ij}^{\alpha}, \quad c_1 = \exp(\eta_1)$$

광역시의 경우

$$Y_{ij} = c_2\beta_i X_{ij}^{\alpha}, \quad c_2 = \exp(\eta_2)$$

기타지역의 경우

$$Y_{ij} = c_3\beta_i X_{ij}^{\alpha}, \quad c_3 = 1$$

이 된다. 여기서 X_{ij}^{α} 의 계수는 승수보정 종업원 1인당 매출액에 해당된다. 따라서 이 모형은 승수 보정 1인당 매출액이 지역에 따라 다를 수 있음을 감안한 모형이라 할 수 있다. 이를 **지역감안 승수 비보정법**이라 부를 수 있을 것이다. 승수 비보정법은 $c_1 = c_2 = c_3 = 1$ 인 경우에 해당되고 단순비보정법은 $\alpha = 1,\ c_1 = c_2 = c_3 = 1$ 에 해당된다.

위 모형에서 무응답 지사매출액은 다음과 같이 추정한다.

업체내 1인당 승수보정 매출액: $\hat{\beta}_i = Y_{i.}/X_{i.}^{(\alpha)}$
여기서

$$X_{i.}^{(\alpha)} = c_1 \sum_{j \in \text{서울}} X_{ij}^{\alpha} + c_2 \sum_{j \in \text{광역시}} X_{ij}^{\alpha} + c_3 \sum_{j \in \text{기타}} X_{ij}^{\alpha},$$

이다.

지점별 매출액
= 지사점 승수 보정 지역감안 종업원수 × 업체내 1인당 승수보정매출액

즉, 본사 i 내 지사 j 의 예측 매출액은

$$\widehat{Y}_{ij} = c_1 X_{ij}^{\alpha} \times \frac{Y_{i.}}{X_{i.}^{(\alpha)}}, \quad \text{서울}$$

$$\widehat{Y}_{ij} = c_2 X_{ij}^{\alpha} \times \frac{Y_{i.}}{X_{i.}^{(\alpha)}}, \quad \text{광역시}$$

$$\widehat{Y}_{ij} = c_3 X_{ij}^{\alpha} \times \frac{Y_{i.}}{X_{i.}^{(\alpha)}}, \quad \text{기타지역}$$

으로 표현된다.

② 지역계수 $c_{1,}c_2$와 승수 α 의 추정

계수의 추정은 지사자료가 확보된 자료로부터 CV 을 최소화 시키거나 또는 adj R^2 을 최대화 시킴으로써 이루어질 수 있다. 다음 <표 3-13> 에 c_1, c_2, α 값을 변화시켜 감에 따라 CV 값과 adj R^2 값을 구해보았다. 좀 더 자세한 내용은 <부록 표 3>, <부록 표4>, <부록 표5>에 정리되어 있다.

<표 3-13> 지역별 계수 변화에 따른 통계량 변화

승수	c_1(서울)	c_2(광역시)	CV의 평균	adj R^2 의 평균
0.6	1.2	0.9	**70.4639**	0.68413
0.7	1.0	1.0	72.7186	0.67966
	1.1	**0.8**	71.3056	**0.68619**
	1.2	0.9	70.8761	0.68559
0.8	1.1	0.8	72.5944	0.68469
	1.2	0.9	72.2760	0.6836

<표 3-13>을 보면 'CV 의 평균'은 승수 = 0.6, c_1 = 1.2, c_2 = 0.8 일 때 가장 작고, 'adj R^2 의 평균'은 승수 = 0.7, c_1 = 1.1, c_2 = 0.8 일 때 가장 높음을 알 수있다.

따라서 최적 승수는 0.6과 0.7 사이에 있음을 알 수 있으며, adj R^2 를 기준으로 종업원 수에 0.7승을 하고 사업체가 서울에 있을 경우 1.1, 광역시에 있을 경우 0.8을 각각 곱해주는 지역별 0.7-승수 비보정 방법이 0.7-승수 비보정 방법보다 예측력이 높다고 할 수 있다. 이 결과는 앞의 회귀 분석에서 10명보다 큰 중규모 이상의 인력규모를 갖는 지사에 대한 매출액 예측에서 추정승수가 0.62 임과 비슷한 점을 주목한다. 또한 사업체의 위치에 따라 앞에 곱해주는 계수 값이 광역시일 경우 1보다 작은 것에 주목한다. 이 결과는 <표 3-5>에서 본 것과 같은 결과이다.

그러므로 지역별 승수 비보정 방법에서 파악되는 모형은 회귀분석에서 중규모 이상의 인력규모를 갖는 지사에 대한 모형에 상응한다.

3.5. 각 방법들의 예측력 비교

기존에 사용해 오던 단순비보정법과 승수 비보정법, 그리고 지역감안승수 비보정법의 예측력을 비교해 보았다. 예측성능의 비교 기준은 CV 와 adj R^2 이다.

<표 3-14>단순비보정, 승수비보정, 지역감안 승수비보정 방법의 비교

	CV			adj R^2		
	단순 비보정법	승수0.7 비보정법	지역감안 승수0.7 비보정법	단순 비보정법	승수0.7 비보정법	지역감안 승수0.7 비보정법
평균	79.05	72.72	**71.31**	0.6682	0.6797	**0.6862**

승수비보정법의 경우 승수는 <표 3-10> 에서구한 최적 승수 $\alpha = 0.7$ 을 사용하였고, 지역감안 승수0.7비보정법의 경우 <표 3-14>에서 구한 최적계수 $\alpha = 0.7$, $c_1 =$ 1.1, $c_2 =$ 0.8 을 사용하였다. 지역감안 승수0.7비보정법이 CV 값이 가장작고 adj R^2 값은 가장 큼을 알 수 있다. 따라서 이 세 가지 방법 중에서 지역감안 승수0.7비보정법의 예측력이 가장 뛰어나다고 할 수 있다.

<표 3-14>를 통해 '종업원 수' 이외의 다른 변수가 없는 경우에 단순 비례배분을 하던 기존 방법에 비해 종업원 수에 특히 승수를 0.7에 적용하였을 때가 예측력이 더 좋고, 승수 0.7만 적용할 경우에 비해 지역별 더미변수를 추가한 경우가 예측력이 더 좋음을 알 수 있다.

IV. 연구결과 및 결론

도소매업 본지사 매출액 무응답 대체법에 대한 통계적 고찰을 통해 현재 사용하고 있는 단순비보정법을 개선보완 시킬 수 있는 새로운 방법을 모색해 보았다. 이러한 연구의 결과 및 결론은 다음과 같이 정리할 수 있다.

4.1. 매출액에 대한 예측력을 지닌 변수들에 대한 분석

먼저 매출액 예측 시 사용가능한 변수들에 대한 통계적 분석을 통해 지사 종업원수와 지사 위치 지역을 고려하게 되었다.

매출액과 상관 관계가 높고 대부분의 지사 수준의 자료가 확보 가능한 변수가 종업원수이다. 종업원수 자료에 대한 통계적인 면모를 살펴보았고 이것이 매출액과 상관관계가 높음을 매출액-종업원수 사이의 상관 분석을 통해 살펴보았다. 이들 사이의 표본 상관계수는 0.694 로 파악되었다.

또한 지사 소재 지역에 따라 일인당 매출액이 다를 수도 있다는 점을 주목하여 지역을 감안한 통계적 방법을 검토하였다.

4.2. 현재 사용하고 있는 대체법인 비보정에 대한 통계적 고찰

현재 사용하고 있는 무응답 대체법인 지사 종업원 수 규모에 따른 비례배분법

(지사 매출액) = (일인당 매출액) × (지사 종업원수)

을 통계적인 면에서 검토하였다. 이것은 매출액과 종업원수에 log 를 취한 회귀모형

$$\log(\text{지사 매출액}) = \log(\text{일인당 매출액}) + \alpha \log(\text{지사 종업원수}),\ \alpha = 1$$

해당됨을 주목하였다.

4.3. 새로운 승수 비보정법 제안

지사 매출액이 파악된 자료를 이용하여 위 회귀모형을 추정해 보았다. 규모가 아주 작은 소규모 지사보다는 일정규모 이상이 되는 지사가 주된 관심사 이므로 종업원수가 10인 넘는 지사의 경우를 집중적으로 살펴보았다. 그 결과 α 값이 대략 0.6 정

도로 나타났다. 이는 새로운 비례배분법인 승수 비보정법을 제시해 주었다.

$$\log(\text{지사 매출액}) = \log(\text{일인당 매출액}) + \alpha \log(\text{지사 종업원수})$$

즉

$$(\text{지사 매출액}) = (\text{일인당 매출액}) \times (\text{지사종업원수})^{\alpha}$$

자료로부터 승수를 추정한 결과 $\alpha = 0.7$ 을 얻어 회귀분석의 결과와 부합하는 결과를 얻었다.

4.4. 지역을 감안한 승수비보정법 제안

지역별로 1인당 매출액이 차이가 날 수있음을 주목하여 이를 통계모형에 반영시켜 검토하였다. 즉 다음의 회귀 모형

$$\log(\text{지사 매출액}) = \log(\text{일인당 매출액}) + \alpha \log(\text{지사 종업원수}) + \eta \text{ 지역변수}$$

을 검토하게 되었고 이 모형에 대응하는 비보정법인 지역감안승수비보정법을 제안하게 되었다. 이는 다음의 모형

$$(\text{지사 매출액}) = (\text{지역계수}) \times (\text{일인당 매출액}) \times (\text{지사종업원수})^{\alpha}$$

에 대응한다.

4.5 기존 방법과 제안된 방법의 비교

제안된 2 방법과 기존의 단순비보정법을 비교한 결과 지역감안승수비보정법이 다른 두 방법보다 나은 예측력을 지님을 확인하였다.

참고문헌

Carl Barsky, James Buszuwski, Lawrence Ernst, Micheal Lettau, Mark Loewenstein, Brooks Pierce, Chester Ponikowski, James Smith, Sandra West (2000), "Alternative imputation models for wage related data collected from establishment survey", *Proceedings of the Second International Conference on Establishment Surveys*, 619-628.

Chester H. Pnikowski, Erin E. McNulty (2006), "Use of administrative data to explore effect of establishment nonresponse adjustment on the National Compensation Survey estimates", U.S. Bureau of Labor Statistics.

Eric Rancourt (1999) "Estimation with nearest neighbour imputation at Statistics Canada", Household Survey Methods Division, Statistic Canada.

James A. Buszuwski, Daniel J. Elmore, Lawrence R. Ernst, Michael K. Lettau, Lowell G. Mason, Steven P. Paben, Chester H. Ponikowski (2003), "Imputation of Benefit related data for the national compensation survey", Paper presented at the 2003 Joint Statistical Meetings, San Francisco, California.

John L. Eltinge, Ralph A. Kozlow, Donal M. Luery (2003), "Imputation in Three Federal Statistical Agencies", Presentation to the Federal Economic Statistics Advisory Committee (FESAC).

Kirk Muller, George Stamas, Shail Butani (1995), "Nonresponse adjustment in certainty strata for an establishment survey", *Proceedings of the Section on Survey, American Statistical Association.*

Mark Crankshaw, George Stamas (2000), "Sample design in the Job openings and Labour turnover survey", Office of Survey Methods Research, U.S. Bureau of Labor Statistics.

Michael Sverchkov, Alan H. Dorfman, Lawrence R. Ernst, Thomas G. Moerhle, Steven P. Paben, Chester H. Ponikowski (2005), "On Calibration and Non-response Adjustment for National Compensation Survey", Office of Survey Methods Research, U.S. Bureau of Labor Statistics.

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

Survey methods and practices (2003), Statistics Canada.

National Compensation Survey at U.S. Bureau of Labor Statistic
: http://www.bls.gov/ncs/

Current Employment Survey at U.S. Bureau of Labor Statistics
: http://www.bls.gov/ces/home.htm

Job Openings and Labor Turnover Survey at U.S. Bureau of Labor Statistics
: http://www.bls.gov/jlt/home.htm

Definitions, data sources and methods at Statistics Canada
: http://www.statcan.ca/english/concepts/index.htm

부록 표

<부록 표 1> 회사별 종업원 수 그룹별 1인당 매출액 평균
(단위: 백만원)

회사	종업원수 (명)	지사 수	평균값 (백만원)
1	1~7	119	133
2	1~7	24	473
	8~11	50	574
	12~16	10	609
	17~74	5	131
	75~1030	1	104
3	1~7	19	193
6	1~7	11	851
	8~11	3	580
7	1~7	106	222
	8~11	234	392
	12~16	312	611
	17~74	85	798
8	1~7	37	262
	8~11	314	431
	12~16	192	684
	17~74	24	909
9	1~7	166	1,172
	8~11	65	1,250
	12~16	3	1,744
11	1~7	1	387
	8~11	1	56
	17~74	33	633
	75~1030	7	698
12	8~11	3	247
	12~16	2	211
	17~74	179	431
	75~1030	117	382
14	1~7	78	27
	8~11	39	10
	12~16	53	10
	17~74	92	8
	75~1030	3	7
합계		2,388	

<부록 표 2> 지역별 계수 변화에 따른 통계량 변화
(승수 = 0.6)

승수	c1(서울)	c2(광역시)	CV의 평균	adj R^2 의 평균
0.6	0.8	0.7	**74.8958**	0.67175
		0.8	74.8992	**0.67252**
		0.9	75.4478	0.67021
		1.0	76.3020	0.66533
		1.1	77.3050	0.65834
		1.2	78.3606	0.64967
	0.9	0.7	73.2930	0.67745
		0.8	**73.0297**	**0.67878**
		0.9	73.3890	0.67708
		1.0	74.1125	0.67281
		1.1	75.0274	0.66642
		1.2	76.0249	0.65831
	1.0	0.7	72.2799	0.68055
		0.8	**71.7834**	**0.68248**
		0.9	71.9719	0.68143
		1.0	72.5738	0.67784
		1.1	73.4042	0.67212
		1.2	74.3442	0.66465
	1.1	0.7	71.6943	0.68153
		0.8	**70.9981**	**0.68404**
		0.9	71.0370	0.68366
		1.0	71.5295	0.68077
		1.1	72.2816	0.67576
		1.2	73.1669	0.66898
	1.2	0.7	71.4188	0.68076
		0.8	70.5538	0.68385
		0.9	**70.4639**	**0.68413**
		1.0	70.8601	0.68194
		1.1	71.5418	0.67764
		1.2	72.3765	0.67158

<부록 표 3> 지역별 계수 변화에 따른 통계량 변화
(승수 = 0.7)

승수	c1(서울)	c2(광역시)	CV의 평균	adj R^2 의 평균
0.7	0.8	0.7	74.7532	0.67555
		0.8	**74.7166**	**0.67585**
		0.9	75.2386	0.67309
		1.0	76.0759	0.66777
		1.1	77.0679	0.66038
		1.2	78.1159	0.65136
	0.9	0.7	73.3385	0.68080
		0.8	**73.0419**	**0.68172**
		0.9	73.3789	0.67961
		1.0	74.0886	0.67495
		1.1	74.9951	0.66820
		1.2	75.9875	0.65976
	1.0	0.7	72.4796	0.68346
		0.8	**71.9565**	**0.68502**
		0.9	72.1273	0.68360
		1.0	72.7186	0.67966
		1.1	73.5431	0.67361
		1.2	74.4801	0.66584
	1.1	0.7	72.0215	0.68400
		0.8	**71.3056**	**0.68619**
		0.9	71.3314	0.68548
		1.0	71.8166	0.68227
		1.1	72.5653	0.67696
		1.2	73.4495	0.66991
	1.2	0.7	71.8525	0.68282
		0.8	70.9745	**0.68562**
		0.9	**70.8761**	0.68559
		1.0	71.2683	0.68312
		1.1	71.9490	0.67855
		1.2	72.7845	0.67225

<부록 표 4> 지역별 계수 변화에 따른 통계량 변화
(승수 = 0.8)

승수	c1(서울)	c2(광역시)	CV의 평균	adj R^2 의 평균
0.8	0.8	0.7	75.5999	0.67552
		0.8	**75.5416**	**0.67553**
		0.9	76.0491	0.67250
		1.0	76.8767	0.66694
		1.1	77.8620	0.65935
		1.2	78.9054	0.65016
	0.9	0.7	74.3629	0.68032
		0.8	**74.0510**	**0.68098**
		0.9	74.3777	0.67864
		1.0	75.0808	0.67378
		1.1	75.9832	0.66685
		1.2	76.9731	0.65828
	1.0	0.7	73.6494	0.68255
		0.8	**73.1178**	**0.68388**
		0.9	73.2827	0.68227
		1.0	73.8705	0.67815
		1.1	74.6932	0.67195
		1.2	75.6298	0.66407
	1.1	0.7	73.3122	0.68270
		0.8	**72.5944**	**0.68469**
		0.9	72.6188	0.68380
		1.0	73.1036	0.68044
		1.1	73.8529	0.67500
		1.2	74.7385	0.66786
	1.2	0.7	73.2448	0.68118
		0.8	72.3713	**0.68379**
		0.9	**72.2760**	0.68360
		1.0	72.6710	0.68099
		1.1	73.3545	0.67632
		1.2	74.1931	0.66993