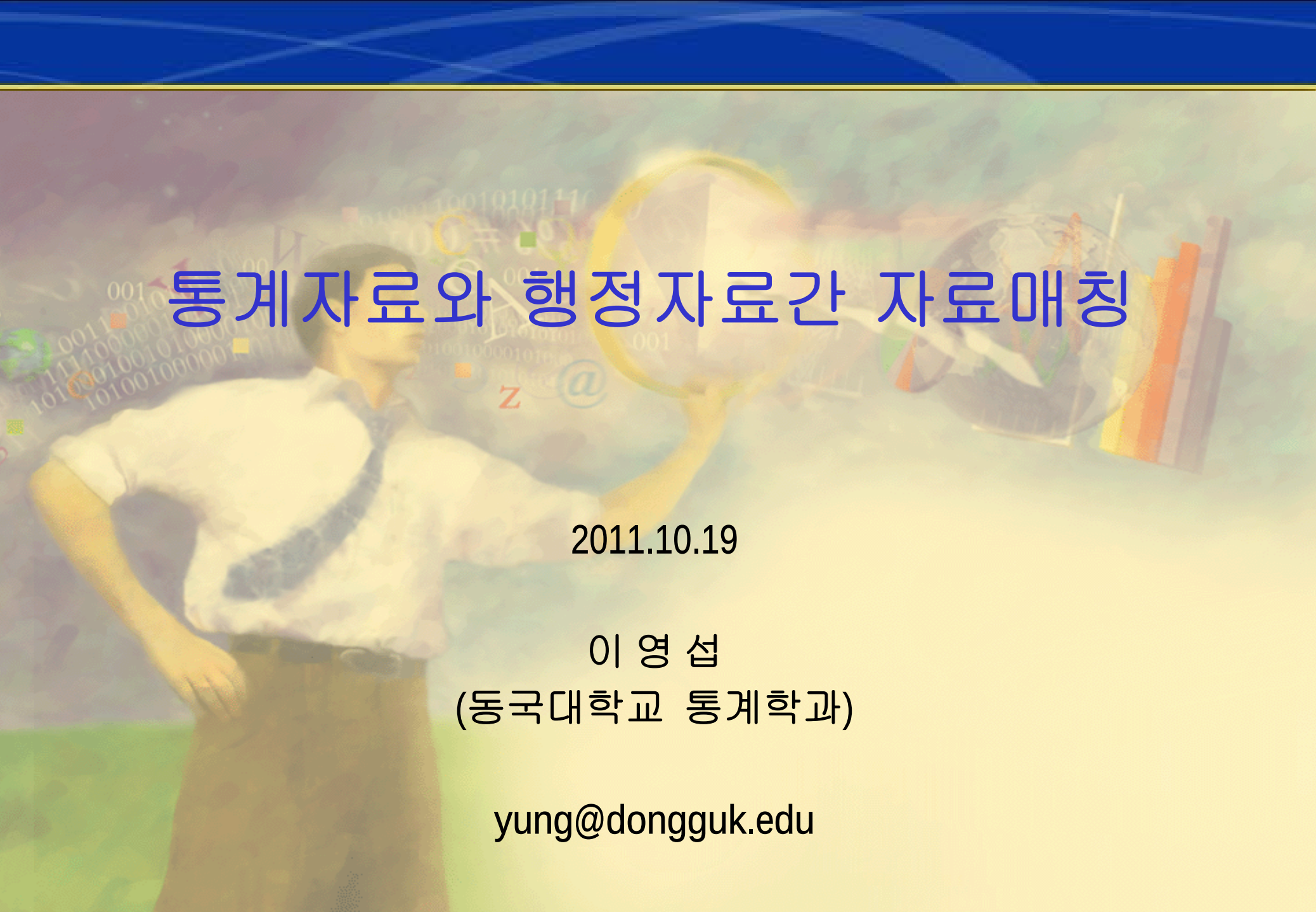

통계자료와 행정자료간 자료매칭

2011.10.19

이 영 섭
(동국대학교 통계학과)

yung@dongguk.edu

0. 목차

1. 데이터 매칭(Data Matching)

1.1 데이터 매칭의 정의

• 공공기관이나 기업이 효과적인 자료 분석을 위해서는 조사단위인 개인이나 가구들에 대한 다양한 정보(기본적인 인구통계학적인 자료, 취미와 생활 습관, 기호 등)를 얻은 후 접근해야 한다.

• 원천 데이터 소스의 다양성, 단일 자료의 불충분성, 부서간의 자료 공유의 부족으로 인하여 하나의 데이터에서 분석에 필요한 모든 정보를 얻는다는 것은 매우 어려운 일이다.

• 이러한 문제는 데이터 매칭(data matching) 또는 데이터 통합(data fusion)을 통해 많은 부분 보완할 수 있다. 일반적인 조사 데이터에 공통적으로 포함하고 있는 요소들을 기본으로 완전히 일치하지는 않지만 특성이 유사한 사람이나 집단끼리의 정보는 얻을 수 있다.

• 데이터 매칭이란 보유하고 있는 데이터 파일에 필요한 변수가 없거나, 결측값이 존재할 경우 다른 원천 데이터로부터 모아진 자료와 정보를 통합하는 것이다.

→ 데이터의 질을 상당히 높일 수 있으며 재조사로 인한 시간과 비용을 줄일 수 있다.

→ 많은 조사항목으로 인한 조사응답자의 부담을 감소시켜 조사항목에 대한 무응답률이 낮아지고 응답의 정확성이 높아져 새로운 조사를 통해 얻은 자료보다 더 좋은 자료를 얻을 수 있다.

1.1 데이터 매칭의 정의

Data Matching의 예

통계청 사회통계조사

개인ID	조사항목					공통항목					
	노동	교육	보건	정보통신	환경	성별	연령	학력	주거지	주거형태	월생활비
:	:	:	:	:	:	:	:	:	:	:	:

+

보건복지부 국민건강조사

개인ID	공통항목						조사항목						
	성별	연령	학력	주거지	주거형태	월생활비	흡연	음주	수면시간	체중	신장	혈압	혈당
:	:	:	:	:	:	:	:	:	:	:	:	:	:

=

매치된 파일

개인ID	사회통계조사항목					공통항목						국민건강조사항목						
	노동	교육	보건	정보통신	환경	성별	연령	학력	주거지	주거형태	월생활비	흡연	음주	수면시간	체중	신장	혈압	혈당
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:

1.2 데이터 매칭의 종류

• **Data Matching의 종류*** :

- **Exact Matching(정확 매칭):** 주민등록번호, 국가보험번호, 사회보장번호와 같이 ID를 나타낼 수 있는 변수가 공통으로 있을 경우, 변수 값이 완전히 일치하는 경우에 데이터를 결합하는 방법 .

장점: 같은 사람, 같은 물건을 정확하게 결합. 측정오차가 없다면 이상적인 데이터 결합.

단점: 개인의 고유한 정보이용하기 때문에 매칭 시도가 불가능하거나 사생활 침해 우려. 개인정보 보호 안됨. 정확매칭에 사용하는 개인식별 가능 변수는 사생활 침해의 여지가 있거나 직접 자료를 수집해야 하는데 이에 따른 시간과 비용의 손실이 있음.

일반적으로 자료 매칭이라 함은 정확 매칭(exact matching)을 말한다

-Statistical Matching(통계적 매칭): 공통으로 가지는 변수에 개인 식별 가능한 변수가 없을 때 수행하는 데이터 결합 방법. 정확매칭을 할 수 없는 경우나 정확매칭후 매칭이 안된 자료들에 대해서는 통계적 매칭(statistical matching)을 비롯한 다른 매칭방법 등을 적용한다.

-Judgmental Matching(판단 매칭): 공통인 변수들 사이에 정확히 일치하는 것은 없지만 자료에 대해 잘 알고 있는 경우, 또는 몇 가지 조사를 시행하고 적절하다고 판단하는 것을 결합하는 방법

-* source : "National Statistics code of Practice Protocol on Data Matching(2003)"

1.3 용어

- 서로 다른 경로로 얻어진 두 개의 파일
  - 파일 A는 (Y X)로 구성되어 있고 파일 B는 (X Z)로 구성
  - 공통변수(common variable) : 파일 A와 파일 B에 모두 관찰되는 변수
  - 유일변수(unique variable) : 파일 A에서만 관찰되는 변수 와 파일 B에서만 관찰되는 변수

- 일반적으로 데이터 매칭을 수행하면 공통변수를 이용하여 파일 B에 있는 변수 를 파일 A에 추가
  - 수용자 파일(recipient file): 파일 A
  - 제공자 파일(donor file): 파일 B
  - 결합 파일(matched file) 또는 통합 파일(fused file): 데이터 매칭을 수행한 후 생성된 파일

[그림1.1] 데이터 매칭

2. 통계적 매칭(Statistical Matching)

2.1 통계적 매칭의 구분

▶ 모수적 방법(parametric method)

- 데이터의 특징을 잘 반영하는 모형을 사용하여 접근하는 방법
- 비모수적 매칭 방법보다 일반화가 잘된다는 장점이 있지만, 자료의 크기가 아주 큰 경우에는 자료의 형태가 매우 복잡하여 모형으로 설명하는 것이 어려운 경우도 있다는 단점이 있다.
- 회귀분석 매칭 알고리즘, 단계적 매칭 알고리즘, 회귀분석과 k-최근접이웃방법의 결합 매칭 알고리즘

▶ 비모수적 방법(nonparametric method)

- 수용자 파일에서 관찰되지 않은 변수를 예측하는데, 특정모형을 가정하지 않고 전적으로 데이터에 기초하여 통계적 결합을 수행하는 접근 방법
- 사전 준비 작업이 거의 없고 수행하기 쉽다는 장점이 있는 반면 계산 시간이 오래 걸린다는 단점이 있다.
- 핫덱(hot deck) 방법, K-최근접이웃(k-nearest neighbor) 매칭 알고리즘

2.1 통계적 매칭의 구분

▶ 모수와 비모수 방법 비교

▪ 모수적 모형을 사용하면 정교한 반면 민감하다는 단점이 있고 비모수적 방법은 모수적 방법보다 좀 더 로버스트(robust)하다는 특징이 있다.

▪ Ingram et al.(2000)은 실제로 데이터 매칭 방법에 있어 회귀분석등과 같은 예측모형을 가정하는 모수적 기법이 좋은 성능을 나타낸다고 하였다.

▪ 최근에는 의사결정나무(decision trees)방법 등을 비롯한 다양한 데이터마이닝(data mining)기법들을 이용하여 예측모형을 구축하고 그 결과를 바탕으로 매칭을 시도하는 연구가 활발히 진행되고 있다.

2.1 통계적 매칭의 구분

- 유일변수 Y 와 Z를 각각 가지고 있는 파일 A와 B를 두 개의 공통변수 X_1, X_2를 이용하여 매칭

- 각 성별로 나이의 거리가 가장 짧은 자료를 결합한다. 이때 성별을 나타내는 변수 X_1을 그룹변수라 한다.

- 거리 측정 함수: 유클리디안 거리(Euclidian distance)를 이용

$$d_{ij} = \left| x_{i2}^{A} - x_{j2}^{B} \right|, \quad i = 1,2,\cdots,8, \quad j = 1,2,\cdots 6$$

,

<표2.1> 통계적 결합 예를 위해 인공적으로 만든 파일

파일 A

관측치 i	X_1^A	X_2^A	Y
A1	1	42	y_1^A
A2	1	35	y_2^A
A3	0	63	y_3^A
A4	1	55	y_4^A
A5	0	28	y_5^A
A6	0	53	y_6^A
A7	0	22	y_7^A
A8	1	25	y_8^A

파일 B

관측치 j	X_1^B	X_2^B	Z
B1	0	33	z_1^B
B2	1	52	z_2^B
B3	1	28	z_3^B
B4	0	59	z_4^B
B5	1	41	z_5^B
B6	0	45	z_6^B

2.2 통계적 매칭의 제약조건

▪ 데이터 매칭 알고리즘이 타당한 결과 도출을 위한 제약조건

첫째, 제공자 파일은 수용자 파일을 대표할 수 있어야 한다. 그러나 반드시 두 데이터가 같은 모집단에서 추출될 필요는 없다.

둘째, 공통변수 X 가 주어졌을 때, 유일변수인 Y 와 Z 사이에 조건부 독립관계가 성립

(조건부 독립성 가정-CIA ; Conditional Independent Assumption)

$$f_{YZ|X}(y,z|x) = f_{Y|Z}(y|x)\, f_Z(z|x)$$

또는,

$$f_{Z|XY}(z|x,y) = f_{Z|X}(z|x)$$

조건부 독립성이 만족되면 매칭 후의 인공 데이터의 분포가 원래의 모집단 분포와 같아지고, 매칭 후의 주변분포 역시 원래의 주변분포와 같아지게 된다.

2.3 통계적 매칭의 수행과정

1) 자료의 준비

- 제공자 파일과 수용자 파일은 서로 다른 목적과 과장을 거쳐 얻어진 자료들이므로 단위의 조화(unit harmonization)와 변수의 조화(variable harmonization) 과정을 거쳐야 한다.

Ex) 단위의 조화과정: 가구단위의 조사와 개인 단위의 조사에 대한 자료 매칭 시 조정이 필요

변수의 조화과정: 같은 값을 다르게 표현한 경우 같은 형식으로 표현(한국, KOR, 코리아)

- 불필요한 변수의 제거

2) 매칭 변수(Matching variable)의 선택

- 공통변수 중에서 실제로 매칭과정에 사용될 매칭 변수를 선택
- 사용가능한 모든 공통변수를 매칭 변수로 하면 변수의 차원이 높아져 표본이 공간상에 드물게 형성된다.
- 결과적으로 관측치 간에 결합거리가 크게 측정되어 근접거리 결합이 힘들다.
- 자료에 대한 내용을 충분히 숙지하고 일차적인 자료 분석을 한 이후에 적절한 매칭 변수를 선택하는 것이 바람직하다.

2.3 통계적 매칭의 수행과정

3) 근사성 측정

- 일반적으로 근사성 척도로서 거리를 사용
- 거리측정함수(distance function): 두 벡터 $\mathbf{a}=\left[a_1,\cdots,a_j,\cdots,a_J\right]^T$ 와 $\mathbf{b}=\left[b_1,\cdots,b_j,\cdots,b_J\right]^T$ 를 가정

  - 유클리디안 거리(Euclidian distance): $d(\mathbf{a},\mathbf{b})=\sqrt{(\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b})}=\sqrt{\sum_{j=1}^{J}(a_j-b_j)^2}$

  - 맨하탄 거리(Manhattan distance): $d(\mathbf{a},\mathbf{b})=\sqrt{\sum_{j=1}^{J}|a_j-b_j|}$

  -민코우스키 거리(Minskowki's distance): $d(\mathbf{a},\mathbf{b})=\left[\sum_{j}^{J}|a_j-b_j|^p\right]^{\frac{1}{p}}$

  -헬링거 거리(Hellinger distance): $d(\mathbf{a},\mathbf{b})=\left[\sum_{j}^{J}\left(\sqrt{a_j}-\sqrt{b_j}\right)^2\right]^{\frac{1}{2}}$

- 거리측정함수를 이용하여 수용자 파일과 제공자 파일간의 근사성을 측정하여 관측치들끼리 매칭

3.1 모수적 매칭 알고리즘

3.1.1 회귀분석 매칭 알고리즘(Regression matching algorithm)

- 매칭시키고자 하는 변수가 연속형인 경우에 회귀분석을 적용하여 매칭하는 방법

Step1

제공자 파일 B를 이용하여, 공통변수 $\mathbf{x}$ 를 설명변수로 하고, 유일변수 $\mathbf{z}$ 를 목표변수로 하는 회귀모형을 추정한다. 추정값을 $\tilde{\mathbf{z}}_b$ 라 하자.

Step2

수용자 파일 A에서 각각의 $a = 1, \cdots, n_A$ 에 대해 추정된 회귀모형을 적용하여 매개값 (intermediate value) $\tilde{\mathbf{z}}_a$ 를 계산한다.

Step3

최근접이웃방법을 적용하여 매개값 $\tilde{\mathbf{z}}_a$ 와 제공자 파일 B에서 가장 가까운 관측치의 실제값(live value) $\mathbf{z}_{b*}$ 를 수용자 파일 A의 a번째 관측치에 매칭시킨다.

- Step2와 Step3은 다양한 선택이 가능

3.1 모수적 매칭 알고리즘

3.1.2 단계적 매칭 알고리즘(한상훈 외, 2004)

수용자 파일

X	Y

여기서 $X=(X_1,\cdots,X_p)$

제공자 파일

X	Z_1 (범주형) Z_2 (연속형)

매칭↓

Case 1.

결합하려는 제공자 파일의 유일변수가 범주형인 경우

X	Y	Z_1

Case 2.

결합하려는 제공자 파일의 유일변수가 연속형인 경우

X	Y	Z_2

Case 3.

범주형과 연속형 유일변수를 동시에 결합하려는 경우

X	Y	Z_1Z_2

3.1 모수적 매칭 알고리즘

3.1.2 단계적 매칭 알고리즘

Case 1. 결합하려는 변수가 범주형인 경우

Step1

제공자 파일에서 유일변수 Z_1을 종속변수로, 공통변수들 $\mathbf{X}$를 독립변수로 하여 로지스틱 회귀모형 추정

추정된 회귀식을 각 파일에 적합시켜 얻은 값을 근사성 측정을 위한 점수로 사용:

$$D_{ab}^{F} = \left| \tilde{z}_{1a} - \tilde{z}_{1b} \right|$$

D_{ab}^{F}가 가장 작은 값을 갖는 a,b를 결합, $a = 1,\cdots,n_A$, $b = 1,\cdots,n_B$

Step1에서 측정한 근사성 정도$\left(D_{ab}^{F}\right)$가 같은 경우 → Step2

Step2

Step1에서 추정된 회귀식에 포함되지 않은 공통변수 중 범주형 변수들(K개)을 이용하여 근사성을 측정 :

$$D_{ab}^{S} = \sum_{k=1}^{K} I\left(x_{ka}, x_{kb}\right), \quad where \quad I(x,y) = \begin{cases} 0, & if \ x = y \\ 1, & if \ x \neq y \end{cases}$$

D_{ab}^{S} a,b

가 가장 작은 값을 갖는 를 결합

Step2에서 측정한 근사성이 같은 경우 → Step3

3.1 모수적 매칭 알고리즘

3.1.2 단계적 매칭 알고리즘

Step3

앞 단계에서 사용하지 않은 공통변수 중 연속형 변수들(S개)을 표준화하여 변수의 차이들로 근사성을 측정:

$$D_{ab}^{T} = \sum_{s=1}^{S} \left| x^{*}_{sa} - x^{*}_{sb} \right|$$

x^{*}는 연속형 공통변수를 표준화했다는 의미

D_{ab}^{T}가 가장 작은 값을 갖는 a,b를 결합

Case 2. 결합하려는 변수가 연속형인 경우

Case1과 동일. 단, Step1에서 선형회귀분석을 수행(Z_2를 종속변수, $\mathbf{X}$를 독립변수)

3.1 모수적 매칭 알고리즘

3.1.2 단계적 매칭 알고리즘

Case 3. 범주형 변수와 연속형 변수를 동시에 결합하는 경우

하나의 변수를 결합하는 것보다 여러 개의 변수를 한 번에 결합하는 경우가 더 일반적이다.

Step1

결합하려는 변수가 범주형인 경우와 연속형인 경우의 각 첫 번째 단계의 순위합(rank sum)으로 근사성을 측정

$$D_{ab}^{RF} = RD_{ab}^{FZ_1} + RD_{ab}^{FZ_2}, \quad a = 1, \cdots, n_A, \quad b = 1, \cdots, n_B$$

$RD_{ab}^{FZ_1}$ 은 범주형 유일변수 Z_1 을 결합할 때 step1에서의 D_{ab}^{F} 의 순위, $RD_{ab}^{FZ_2}$ 는 연속형 유일변수 Z_2 를 결합할 때 Step1에서의 D_{ab}^{F} 의 순위를 의미

순위합 D_{ab}^{RF} 값이 가장 작은 수용자 파일의 a 번째 관측치와 제공자 파일의 b 번째 관측치를 결합

Step2

결합하려는 변수가 범주형인 경우와 연속형인 경우의 각 두 번째 단계의 순위합으로 근사성을 측정

$$D_{ab}^{RS} = RD_{ab}^{SZ_1} + RD_{ab}^{SZ_2}, \quad a = 1, \cdots, n_A, \quad b = 1, \cdots, n_B$$

$RD_{ab}^{SZ_1}$ 은 범주형 유일변수 Z_1 을 결합할 때 step2에서의 D_{ab}^{S} 의 순위, $RD_{ab}^{SZ_2}$ 는 연속형 유일변수 Z_2 를 결합할 때 Step2에서의 D_{ab}^{S} 의 순위를 의미

순위합 D_{ab}^{RS} 값이 가장 작은 수용자 파일의 a 번째 관측치와 제공자 파일의 b 번째 관측치를 결합

3.1 모수적 매칭 알고리즘

3.1.2 단계적 매칭 알고리즘

Case 3. 범주형 변수와 연속형 변수를 동시에 결합하는 경우

Step3

결합하려는 변수가 범주형인 경우와 연속형인 경우의 각 세 번째 단계의 순위합으로 근사성을 측정

$$D_{ab}^{RT} = RD_{ab}^{TZ_1} + RD_{ab}^{TZ_2}, \quad a = 1, \cdots, n_A, \quad b = 1, \cdots, n_B$$

$RD_{ab}^{TZ_1}$ 은 범주형 유일변수 Z_1을 결합할 때 Step3에서의 D_{ab}^{T} 의 순위, $RD_{ab}^{TZ_2}$ 는 연속형 유일변수 Z_2를 결합할 때 Step3에서의 D_{ab}^{T} 의 순위를 의미

순위합 D_{ab}^{RT} 값이 가장 작은 수용자 파일의 a 번째 관측치와 제공자 파일의 b 번째 관측치를 결합

3.1 모수적 매칭 알고리즘

3.1.3 회귀분석과 K-최근접이웃방법의 결합 매칭 알고리즘(정성석 외 ,2004)

▪ 회귀분석 매칭 알고리즘은 추정치의 거리가 가장 가까운 하나의 관측치만을 사용함으로써 상대적으로 유사한 다른 관측치들의 정보를 무시하게 되므로 이러한 정보손실을 줄여 데이터 통합기법의 성능일 높이고자 회귀분석기법에 K-최근접이웃방법을 결합

Step1

제공자 파일의 유일변수 $\mathbf{Z}$중 임의의 s번째 변수 Z_s를 목표변수로, 제공자 파일의 공통변수 $\mathbf{X}$를 설명변수로 하여 회귀모형을 추정

Step2

추정된 회귀모형을 각 파일에 적용하여 Z_s의 예측치 $\tilde{z}_{sa}\,(a=1,\cdots,n_A)$와 $\tilde{z}_{sb}\,(b=1,\cdots,n_B)$을 계산

Step3

두 파일에서의 예측값을 이용하여 수용자 파일의 각 관측치에 대해 모든 제공자 파일 관측치와의 거리 $d(\tilde{z}_{sa},\tilde{z}_{sb})$를 계산

Step4

계산한 거리를 이용하여 수용자 파일의 각 관측치에 가장 가까운 제공자 파일의 k개의 관측치의 실제값 $z_{s1^*},\cdots,z_{sk^*}$를 선택

Step5

선택된 제공자 파일의 k개 관측치 $z_{s1^*},\cdots,z_{sk^*}$ 들의 평균(연속형인 경우)이나 최빈값(범주형인 경우)

을 구한 후 이 값을 수용자 파일의 해당 관측치에 결합

3.2 비모수적 매칭 알고리즘

3.2.1 핫덱 방법(Hot deck)

- 랜덤 핫덱(random hot deck): 수용자 파일의 각 관측치에 대해 제공자 파일의 관측치를 랜덤하게 선택하여 매칭시키는 방법

- 특히 수용자 파일과 제공자 파일의 관측치들은 대개 주어진 일반적인 특성(지형적 특성, 사회적 특성 등)에 따라 동질적인 부분집합으로 그룹화 될 수 있다.

- 각각의 수용자 관측치에 대해 주어진 지형적 특성내에서 동일지역의 관측치만이 가능한 제공자로 고려

- 자료를 동질적인 부분집합으로 그룹화 시키는 변수들을 대체군(donation class)라 한다. 일반적으로 하나 혹은 몇몇의 범주형 공통변수가 대체군이 된다.

3.2 비모수적 매칭 알고리즘

예) 파일 A에는 6개의 관측치($n_A = 6$)와 3개의 변수 '성별', '연령', '연소득' 이 존재 : 수용자

파일 B에는 10($n_A = 10$) 개의 관측치와 3개의 변수 '성별', '연령', '연지출'이 존재 : 제공자

2개의 공통변수 $X = \{X_1 = \text{'성별'} \ , \ X_2 = \text{'연령'}\}$

각각의 유일변수 $Y = \text{'연소득'}$ 과 $Z = \text{'연지출'}$ 이 존재

<표3.1> 파일 A의 관측치

a	X_1	X_2	Y
1	F	27	22
2	M	35	19
3	M	41	47
4	F	61	41
5	F	52	17
6	F	39	26

<표3.2> 파일 B의 관측치

b	X_1	X_2	Z
1	F	54	22
2	M	21	17
3	F	48	15
4	F	33	14
5	M	63	13
6	F	29	15
7	M	36	19
8	M	55	24
9	F	50	26
10	F	27	18

3.2 비모수적 매칭 알고리즘

- 파일 A의 각각의 관측치들은 파일 B의 10개의 관측치들로부터 랜덤하게 선택하여 제공자 값을 할당 받음.
- 만약 단위 b가 단위 a로 할당된다면 a에 존재하지 않는 Z 값은 b의 관측된 값으로 매칭되게 된다.
- 최종 데이터의 a번째 관측치는 (X_a, y_a, z_b) 가 된다.
- 매칭결과는 $n_B^{n_A} = 10^6$ 가지 가능한 조합

<표3.3> 랜덤핫덱방법에 의한 파일 A와 B의 매칭결과

a	b donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
1	2	F	M	54	21	22	17
2	8	M	M	21	55	19	24
3	5	F	M	48	63	47	13
4	6	F	F	33	29	41	15
5	4	M	F	63	33	17	14
6	2	F	M	29	21	26	17

3.2 비모수적 매칭 알고리즘

- 만약 공통변수 '성별'을 대체군을 정의하는데 사용한다면 파일 B에서 제공자는 수용자 파일의 각 관측치들에 대해 동일한 성별을 가진 관측치들 중에서 랜덤하게 선택
- 가능한 제공자 배열은 급격하게 줄어듬.

$$(n_M^B)^{n_M^A} + (n_F^B)^{n_F^A} = 6^4 + 4^2 = 1312$$

<표3.4> 동일 '성별' 내에서의 랜덤핫덱방법에 의한 파일 A와 B의 매칭결과

a	b donor	X_1^A	X_1^B	X_2^A	X_2^B	Y	Z
2	5	M	M	35	63	19	13
3	7	M	M	41	36	47	19
1	3	F	F	27	48	22	15
4	6	F	F	61	29	41	15
5	9	F	F	52	50	17	26
6	3	F	F	39	48	26	15

3.2 비모수적 매칭 알고리즘

3.2.2 K-최근접이웃 매칭 알고리즘(K-nearest neighbor algorithm)

▪ 가장 유사한 하나의 관측치를 매칭하는 최근접이웃방법에서 한 단계 나아가 상대적으로 유사한 k개의 관측치를 선택하여 매칭에 사용하는 방법

Step1

공통변수를 수치형으로 변환하고, 이를 이용하여 수용자 파일의 각 관측치에 대해 제공자 파일의 모든 관측치들과의 거리를 계산한다.(유클리디안 거리를 흔히 사용)

Step2

계산한 거리 중 수용자 파일의 각 관측치와 가장 가까운 제공자 파일의 관측치 k개를 선택한다.

Step3

선택된 k개 관측치에 해당하는 제공자 파일의 유일변수를 이용하여 수용자 파일의 각 관측치에 통합변수를 추가시킨다. 유일변수가 연속형이면 k개의 평균값(mean)을, 범주형이면 최빈값(mode)을 수용자 파일에 매칭시킨다.

회귀분석과 K-NN의 결합매칭 알고리즘을 이용한 사례연구(정성석 외, 2004 재계산) / 4.1 분석 설계

※ **자료** : Boston Housing 데이터 이용 (출처 : UC Irvine Data Repository)

13개의 독립변수를 이용하여 Boston지역의 집값(MEDV)을 예측하는 것이 목적

※ 데이터 파티션 :

- 각 파일에 포함될 변수 분리는 조건부독립성이 만족되도록 이루어져야 한다.
- Rässler(2002)가 제시한 회귀분석접근법으로 판단하는 방법을 이용
- 최종분석의 목표변수가 될 MEDV는 데이터 통합에 영향이 없도록 하기 위해 수용파일의 유일변수 Y 에 포함시킨다

$$Z = \beta_0 + \beta_{XZ|Y} X + \beta_{YZ|X} Y \text{ 에서 } \beta_{YZ|X} = 0 \text{ 이면 } \rho_{YZ|X} = 0$$

-**제공파일의 유일변수** Z : 위의 식으로부터 유의수준 0.05하에서 MEDV가 설명변수로서 유의하지 않은 반응변수인 INDUS, AGE, CHAS로 선택

-**공통변수** X : 제공파일의 유일변수로 선택된 INDUS, AGE, CHAS변수를 반응변수로 하여 유의한 설명변수 NOX, RM, DIS, RAD, TAX, LSTAT를 선택

-**수용파일의 유일변수** Y : 공통변수에 포함되지 않고 제공파일의 유일변수에도 포함되지 않는 변수들을 선택

4.1 분석 설계

<표4.1> 실험 데이터의 파티션 결과

변수		개체수(506)		공통변수(X)	수용파일 유일변수(Y)	제공파일 유일변수(Z)
연속	범주	수용파일	제공파일			
13	1	202	304	NOX,RM, DIS, RAD,TAX,LSTAT	PTRATIO,MEDV, ZN,B, CRIM	INDUS, AGE, CHAS(범주형)

- 제공자 파일의 유일변수 중 INDUS를 목표변수로, 공통변수를 독립변수로 하여 회귀모형적합
- INDUS는 연속형이므로 표준화→ St_INDUS
- 회귀모형 적합시 설명력있는 공통변수만 포함되도록 단계적 변수선택법을 수행

4.2 매칭 결과

<표4.2>가장 가까운 7개 예측치의 차이(통합변수가 연속형인 경우)

R	1 (INDUS=7.87)			...	100 (INDUS=5.86)			...	202 (INDUS=27.74)		
k	D	distance	INDUS(D)		D	distance	INDUS(D)		D	distance	INDUS(D)
1	131	0.003164	6.2		165	0.004296	2.25		219	18.1	0.023664
2	137	0.004478	6.2		145	0.010919	4.93		293	27.74	0.139158
3	4	0.004579	7.87		148	0.011781	5.86		228	18.1	0.185954
4	134	0.012477	6.2	...	170	0.016273	4.95	...	294	27.74	0.186472
5	111	0.015755	3.44		146	0.016933	4.93		222	18.1	0.189157
6	195	0.017690	7.38		39	0.017786	5.13		230	18.1	0.237989
7	122	0.021178	10.59		181	0.017807	2.18		85	19.58	0.259588

※ R: 수용파일의 개체, D: 제공파일의 개체, INDUS(D): 제공파일의 실제값

$$\text{distance} = |\hat{St}\_INDUS_R - \hat{St}\_INDUS_D|$$

4.2 매칭 결과

<표4.3> k에 따른 MSE의 변화(통합변수가 연속형인 경우)

반복	k=1	k=3	k=5	k=7
1	18.561	15.125	14.292	14.157
2	23.303	15.761	12.221	12.711
3	22.394	13.161	11.122	10.705
4	17.781	15.699	14.492	14.488
5	21.321	15.621	13.386	13.545
:	:	:	:	:
16	18.258	11.363	11.503	10.898
17	24.218	19.710	17.435	14.487
18	20.861	15.841	15.551	15.270
19	24.989	13.937	12.518	10.553
20	20.838	16.960	14.737	13.562
평균	21.523	14.920	13.451	12.941

- k값이 늘어남에 따라 MSE값이 감소
- MSE의 감소폭이 작아지는 시점의 k값 결정(k=3인 경우가 적합한 것으로 판단)

4.2 매칭 결과

<표4.4> 가장 가까운 7개 예측치의 차이(통합변수가 범주형인 경우)

R	1 (CHAS=0)			...	101 (CHAS=0)			...	202 (CHAS=1)		
k	D	distance	CHAS(D)		D	distance	CHAS(D)		D	distance	CHAS(D)
1	7	0.000027	0		177	0.000915	0		195	0.000675	0
2	66	0.000590	0		163	0.001186	0		213	0.000725	0
3	137	0.000728	0		228	0.002226	0		304	0.000812	1
4	3	0.001174	0	...	296	0.005875	1	...	123	0.001225	0
5	48	0.001530	0		227	0.006535	0		216	0.001637	0
6	205	0.001538	0		84	0.009045	0		291	0.001740	1
7	51	0.001826	0		102	0.010996	0		247	0.003053	0

※ R: 수용파일의 개체, D: 제공파일의 개체, CHAS(D): 제공파일의 실제값

$$\text{distance} = |\hat{P}(CHAS=0)_R - \hat{P}(CHAS=0)_D|$$

4.2 매칭 결과

<표4.5> k에 따른 error rate(%)의 변화(통합변수가 범주형인 경우)

반복	k=1	k=3	k=5	k=7
1	10.8911	3.9604	2.4752	1.4851
2	12.3762	3.9604	1.9802	1.4851
3	8.9109	3.4653	2.4752	1.4851
4	10.8911	2.9703	1.9802	1.4851
5	12.8713	3.9604	2.4752	1.9802
:	:	:	:	:
16	11.3861	3.9604	2.4752	1.9802
17	8.9109	3.4653	1.9802	1.4851
18	12.8713	3.9604	2.4752	1.4851
19	10.8911	3.4653	1.9802	1.4851
20	11.3861	3.9604	2.4752	1.9802
평균	11.5842	3.7624	2.3267	1.6089

- k값이 늘어남에 따라 error rate값이 감소
- error rate의 감소폭이 작아지는 시점의 k값 결정 (k=3 경우가 적합한 것으로 판단)

5.1 분석설계

- **본 연구에서는 조사자료와 행정자료를 효율적으로 매칭 시킬 수 있는 통계적 기법에 대한 연구를 수행**
  - ▪ 조사자료와 행정자료간 매칭 및 분석을 통하여 신규 통계의 작성 가능
  - ▪ 특정 행정자료의 활용의 범위를 확인하고 이를 이용한 새로운 자료로 재구성하여 미래 연구의 사용 가능성을 확인
  - ▪ 다양한 목적에 따라 수집되어진 자료들 간의 효율적인 통합을 위한 통계적 매칭기법을 알아보고 이들의 문제점에 대해 확인

→ 여러 행정 기관에서 제공하는 행정자료의 효율적 활용을 도모하는데 매우 중요한 역할을 할 것으로 예상.

5.1 분석설계

- 국민연금 서울지역 자료 : 223,186개 (18개 변수) – 2006년 6월 기준
- 사업체기초조사 서울지역 자료 : 741,229개 (72개 변수) – 2005년 12월 기준

<표5.1> 국민연금자료의 변수 리스트

변수	유형	길이	라벨
Addr	문자	80	소재지
Bonsa_addr	문자	80	본점소재지
Bonsa_boss_nm	문자	18	본점대표자성명
Bonsa_nm	문자	60	본점사업장명칭
Bonsa_nps_id	문자	8	본점사업장기호
Bonsa_tel	문자	14	본점전화번호
Boss_nm	문자	18	대표자성명
Bs_id	문자	10	사업장등록번호
Bs_kind	문자	60	업종

5.1 분석설계

<표5.1> 국민연금자료의 변수 리스트(계속)

변수	유형	길이	라벨
Bs_nm	문자	60	사업장명칭
Bs_type	문자	4	사업장형태(법인,개인)
Crop_id	문자	13	법인등록번호
Divid_yn	문자	1	분리적용사업장여부
Fax	문자	14	팩스번호
Mem_num	수치	8	가입자 수
Nps_id	문자	8	사업장기호
Open_date	문자	10	적용연월일
tel	문자	14	전화번호

5.1 분석설계

<표5.2> 사업체기초조사자료의 변수 리스트

변수명	유형	길이	라벨
SEQNO	문자	10	전산일련번호
ZONE_CD1	문자	2	행정구역_시도
ZONE_CD2	문자	3	행정구역_시군구
ZONE_CD3	문자	2	행정구역_읍면동
JOSA_CD	문자	3	조사구_코드
JOSA_TK_CD	문자	1	조사구_특성코드
SAUP_NU	문자	3	사업체일련번호
SAUP_NM	문자	60	사업체명
DAEP_NM	문자	20	대표자명
D_SEX_CD	문자	1	대표자_성별코드
CHANG_Y	문자	4	창업년도
CHANG_M	문자	2	창업월
SAUPRG_NU	문자	10	사업자등록번호
ADDR_G	문자	12	소재지_사업체읍면동
ADDR_L	문자	12	소재지_사업체주소리
ADDR_B	문자	20	소재지_사업체주소번지
ADDR_H	문자	4	소재지_사업체주소호
ADDR_T	문자	4	소재지_사업체주소통

5.1 분석설계

<표5.2> 사업체기초조사자료의 변수 리스트(계속)

변수명	유형	길이	라벨
ADDR_V	문자	4	소재지_사업체주소반
BUILD_N	문자	40	소재지_빌딩상가명
BUILD_D	문자	40	소재지_빌딩상가동
BUILD_L	문자	40	소재지_빌딩상가층
BUILD_H	문자	14	소재지_빌딩상가호
SAUP_R_CD	문자	1	사업장 변동
JOSIC_CD	문자	1	조직형태
BOUPIN_KIND_CD	문자	1	조직형태_회사법인
BOUPIN_NU	문자	13	법인등록번호
KUBUN_CD	문자	1	사업체구분코드
M_SAUP_NM	문자	50	사업체 구분_본사명
M_TEL_Z	문자	19	사업체 구분_본사전화지역
M_TEL_K	문자	4	사업체 구분_본사전화국
M_TEL_N	문자	4	사업체 구분_본사전화번호
M_ADDR_D	문자	40	사업체 구분_본사주소시도
M_ADDR_DC_CD	문자	40	사업체 구분_본사주소시군구
M_ADDR_G	문자	40	사업체 구분_본사주소읍면동
M_ADDR_B	문자	40	사업체 구분_본사주소번지
SAUP	문자	70	사업의 종류_주사업내용

5.1 분석설계

<표5.2> 사업체기초조사자료의 변수 리스트(계속)

변수명	유형	길이	라벨
SN1_P	문자	3	사업의 종류_주사업비중
SANGP_NM	문자	70	사업의 종류_주취급상품
SNB_CD1	문자	2	사업의 종류_주산업분류(중)
SNB_CD2	문자	1	사업의 종류_주산업분류(소)
SNB_CD3	문자	1	사업의 종류_주산업분류(세)
SNB_CD4	문자	1	사업의 종류_주산업분류(세세)
SN2_C	문자	60	사업의 종류_부사업내용
SN2_P	문자	3	사업의 종류_부사업비중
SN2_S	문자	60	사업의 종류_부취급품목
SN2_B	문자	5	사업의 종류_부산업분류
EMP_JA_M	수치	10	종사자 수_자영업주(남)
EMP_JA_F	수치	10	종사자 수_자영업주(여)
EMP_JA	수치	10	종사자 수_자영업주(계)
EMP_MU_M	수치	10	종사자 수_무급가족종사자(남)
EMP_MU_F	수치	10	종사자 수_무급가족종사자(여)
EMP_MU	수치	10	종사자 수_무급가족종사자(계)
EMP_SA_M	수치	10	종사자 수_상용종사자(남)
EMP_SA_F	수치	10	종사자 수_상용종사자(여)
EMP_SA	수치	10	종사자 수_상용종사자(계)

5.1 분석설계

<표5.2> 사업체기초조사자료의 변수 리스트(계속)

변수명	유형	길이	라벨
EMP_IM_M	수치	10	종사자 수_임시및일일종사자(남)
EMP_IM_F	수치	10	종사자 수_임시및일일종사자(여)
EMP_IM	수치	10	종사자 수_임시및일일종사자(계)
EMP_MO_M	수치	10	종사자 수_무급종사자(남)
EMP_MO_F	수치	10	종사자 수_무급종사자(여)
EMP_MO	수치	10	종사자 수_무급종사자(계)
EMP_TO_M	수치	10	종사자 수_합계(남)
EMP_TO_F	수치	10	종사자 수_합계(여)
EMP_TO	수치	10	종사자 수_합계(계)
INCOM_Y	수치	10	연간매출액_총매출액
PERIOD	수치	2	연간매출액_연간영업개월수
MONEY_M	수치	10	연간매출액_월평균매출액
GCPT_C	수치	10	자본금
SAUP_IDR	문자	10	모집단고유번호
SNB_DAEB_CD	문자	1	주산업분류(대)
ISVALID	문자	1	보고서집계포함여부

5.1 분석설계

- 수용자 파일 : 사업체기초조사자료
- 제공자 파일 : 국민연금조사자료
- 수용자 파일의 유일변수 : 총종사자수(emp_to)
- 제공자 파일의 유일변수 : 가입자수(mem_num)
- 공통변수 : 소재지, 업종, 조직형태(사업장 형태)

- **목적:** 데이터 매칭 방법 중 활용가치가 큰 통계적 매칭(랜덤핫덱방법)을 적용한다.

- **자료의 문제점:**
  - 원데이터의 관측치 개수를 살펴보면 수용자 파일로 사용될 사업체기초조사자료가 제공자 파일로 사용될 국민연금자료보다 훨씬 크다.
  - 공통변수로 사용될 수 있는 변수의 개수가 적으며 모두 범주형으로 매칭시 많은 동점이 발생한다.

- **연구방법:**
  - 정확매칭된 데이터를 가지고 다시 제공자 파일과 수용자 파일로 나누어 통계적 매칭 기법을 적용해본다. 이는 후에 매칭에 대한 평가가 용이하다는 장점이 있다.
  - 랜덤핫덱 방식을 이용하여 수용자 파일의 관측치들이 동일한 값을 갖더라도 제공자 파일에서 상이한 관측치들이 매칭되게 하여 변동이 발생 되도록 한다.

5.2 정확 매칭

※ 정확매칭

▪ 기준변수

기준변수	국민연금자료	사업체기초조사자료
사업자등록번호	bs_id(사업장등록번호)	Sauprg_nu(사업자등록번호)
대표자성명	Boss_nm(대표자성명)	Daep_nm(대표자명)

▪ 결측치 제거

관측치	국민연금자료	사업체기초조사자료
Raw data	223,186	741,229
사업자등록번호 결측 제거 후	223,186	578,548
대표자성명 결측 제거 후	223,171	578,540

➢ **사업자등록번호&대표자성명을 기준변수로 한 경우 정확매칭된 관측치 수 : 124,826개**

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (사기초기준)

case0 : 제외 없음			
그룹(사기초기준)	그룹빈도	일치개수	일치율(%)
0인	-	-	-
1인-4인	56018	24083	42.99
5인-9인	35984	7964	22.13
10인-19인	17355	2168	12.49
20인-49인	9119	543	5.95
50인-99인	3406	85	2.50
100인-299인	1733	21	1.21
300인-499인	346	2	0.58
500-999인	498	2	0.40
1000인 이상	367	0	0.00
합계	124826	34868	27.93

→ 종사자수가 증가함에 따라 일치율이 크게 감소한다.

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (사기초기준)

case1 : 무급가족 종사자 제외			
그룹(사기초기준)	그룹빈도	일치개수	일치율(%)
0인	-	-	-
1인-4인	57026	25219	44.22
5인-9인	35123	8020	22.83
10인-19인	17217	2170	12.60
20인-49인	9110	544	5.97
50인-99인	3406	85	2.50
100인-299인	1733	21	1.21
300인-499인	346	2	0.58
500-999인	498	2	0.4
1000인 이상	367	0	0
합계	124826	36063	28.89

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (사기초기준)

case2 : 임시 및 일일 종사자 제외			
그룹(사기초기준)	그룹빈도	일치개수	일치율(%)
0인	6	0	0.00
1인-4인	62660	28127	44.89
5인-9인	32967	8583	26.04
10인-19인	15573	2319	14.89
20인-49인	8018	583	7.27
50인-99인	2952	83	2.81
100인-299인	1532	17	1.11
300인-499인	328	3	0.91
500-999인	588	2	0.34
1000인 이상	202	8	0.00
합계	124826	39717	31.82

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (사기초기준)

case3 : 무급종사자 제외			
그룹(사기초기준)	그룹빈도	일치개수	일치율(%)
0인	13	1	7.69
1인-4인	57183	24603	43.03
5인-9인	35723	8011	22.43
10인-19인	17014	2177	12.80
20인-49인	8720	543	6.23
50인-99인	3298	85	2.58
100인-299인	1677	21	1.25
300인-499인	336	2	0.60
500-999인	497	2	0.40
1000인 이상	365	0	0.00
합계	124826	35445	28.40

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (사기초기준)

case4 : 무급가족 종사자와 임시 및 일일 종사자 제외			
그룹(사기초기준)	그룹빈도	일치개수	일치율(%)
0인	6	0	0.00
1인-4인	63494	29398	46.30
5인-9인	32215	8645	26.84
10인-19인	15494	2323	14.99
20인-49인	8015	584	7.29
50인-99인	2952	83	2.81
100인-299인	1532	17	1.11
300인-499인	328	3	0.91
500-999인	588	2	0.34
1000인 이상	302	0	0.00
합계	124826	41055	32.89

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (사기초기준)

case5 : 임시 및 일일 종사자와 무급 종사자 제외			
그룹(사기초기준)	그룹빈도	일치개수	일치율(%)
0인	24	2	8.33
1인-4인	63890	28725	44.96
5인-9인	32645	8641	26.47
10인-19인	15215	2327	15.29
20인-49인	7628	583	7.64
50인-99인	2839	83	2.92
100인-299인	1480	17	1.15
300인-499인	318	3	0.94
500-999인	587	2	0.34
1000인 이상	200	0	0.00
합계	124826	40383	32.35

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (사기초기준)

case6 : 무급가족 종사자와 무급종사자 제외			
그룹(사기초기준)	그룹빈도	일치개수	일치율(%)
0인	13	1	7.69
1인-4인	58186	25749	44.25
5인-9인	34860	8065	23.14
10인-19인	16883	2177	12.89
20인-49인	8711	544	6.24
50인-99인	3298	85	2.58
100인-299인	1677	21	1.25
300인-499인	336	2	0.60
500-999인	497	2	0.40
1000인 이상	365	0	0.00
합계	124826	36646	29.36

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (사기초기준)

case7 : 무급가족 종사자, 임시 및 일일 종사자, 무급종사자 제외			
그룹(사기초기준)	그룹빈도	일치개수	일치율(%)
0인	24	2	8.33
1인-4인	64716	30009	46.37
5인-9인	31898	8700	27.27
10인-19인	15139	2329	15.38
20인-49인	7625	584	7.66
50인-99인	2839	83	2.92
100인-299인	1480	17	1.15
300인-499인	318	3	0.94
500-999인	587	2	0.34
1000인 이상	200	0	0.00
합계	124826	41729	33.43

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (국민연금기준)

case0 : 제외 없음			
그룹(국민연금기준)	그룹빈도	일치개수	일치율(%)
0인	1483	0	0.00
1인-4인	69477	24083	34.66
5인-9인	27399	7964	29.07
10인-19인	13442	2168	16.13
20인-49인	7187	543	7.56
50인-99인	2224	85	3.82
100인-299인	1961	21	1.07
300인-499인	449	2	0.45
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	34868	27.93

→ 종사자수가 증가함에 따라 일치율이 크게 감소한다.

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (국민연금기준)

case1 : 무급가족 종사자 제외			
그룹(국민연금기준)	그룹빈도	일치개수	일치율(%)
0인	1483	0	0.00
1인-4인	69477	25219	36.30
5인-9인	27399	8020	29.27
10인-19인	13442	2170	16.14
20인-49인	7187	544	7.57
50인-99인	2224	85	3.82
100인-299인	1961	21	1.07
300인-499인	449	2	0.45
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	36063	28.89

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (국민연금기준)

case2 : 임시 및 일일 종사자 제외			
그룹(국민연금기준)	그룹빈도	일치개수	일치율(%)
0인	1483	0	0.00
1인-4인	69477	28127	40.48
5인-9인	27399	8583	31.33
10인-19인	13442	2319	17.25
20인-49인	7187	583	8.11
50인-99인	2224	83	3.73
100인-299인	1961	17	0.87
300인-499인	449	3	0.67
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	39717	31.82

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (국민연금기준)

case3 : 무급종사자 제외			
그룹(국민연금기준)	그룹빈도	일치개수	일치율(%)
0인	1483	1	0.07
1인-4인	69477	24603	35.41
5인-9인	27399	8011	29.24
10인-19인	13442	2177	16.20
20인-49인	7187	543	7.56
50인-99인	2224	85	3.82
100인-299인	1961	21	1.07
300인-499인	449	2	0.45
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	35445	28.40

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (국민연금기준)

case4 : 무급가족 종사자와 임시 및 일일 종사자 제외			
그룹(국민연금기준)	그룹빈도	일치개수	일치율(%)
0인	1483	0	0.00
1인-4인	69477	29398	42.31
5인-9인	27399	8645	31.55
10인-19인	13442	2323	17.28
20인-49인	7187	584	8.13
50인-99인	2224	83	3.73
100인-299인	1961	17	0.87
300인-499인	449	3	0.67
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	41055	32.89

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (국민연금기준)

case5 : 임시 및 일일 종사자와 무급 종사자 제외			
그룹(국민연금기준)	그룹빈도	일치개수	일치율(%)
0인	1483	2	0.13
1인-4인	69477	28725	41.34
5인-9인	27399	8641	31.54
10인-19인	13442	2327	17.31
20인-49인	7187	583	8.11
50인-99인	2224	83	3.73
100인-299인	1961	17	0.87
300인-499인	449	3	0.67
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	40383	32.35

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (국민연금기준)

case6 : 무급가족 종사자와 무급종사자 제외			
그룹(국민연금기준)	그룹빈도	일치개수	일치율(%)
0인	1483	1	0.07
1인-4인	69477	25749	37.06
5인-9인	27399	8065	29.44
10인-19인	13442	2177	16.20
20인-49인	7187	544	7.57
50인-99인	2224	85	3.82
100인-299인	1961	21	1.07
300인-499인	449	2	0.45
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	36646	29.36

5.2 정확 매칭- 결과 : 종사자수와 가입자수의 관계 (국민연금기준)

case7 : 무급가족 종사자, 임시 및 일일 종사자, 무급종사자 제외			
그룹(국민연금기준)	그룹빈도	일치개수	일치율(%)
0인	1483	2	0.13
1인-4인	69477	30009	43.19
5인-9인	27399	8700	31.75
10인-19인	13442	2329	17.33
20인-49인	7187	584	8.13
50인-99인	2224	83	3.73
100인-299인	1961	17	0.87
300인-499인	449	3	0.67
500-999인	440	2	0.45
1000인 이상	764	0	0.00
합계	124826	41729	33.43

5.2 정확 매칭- 결과 : 그룹별(사기초 기준) 가입자수와 종사자수의 차이 분포

■ 그룹 : 1인-4인 (도수 100이상)

DIST =종사자수-가입자수

OBS	DIST	COUNT	PERCENT	OBS	DIST	COUNT	PERCENT
1	-137	14	0.0250	24	-17	46	0.0821
2	-45	11	0.0196	25	-16	49	0.0875
3	-42	10	0.0179	26	-15	40	0.0714
4	-41	12	0.0214	27	-14	43	0.0768
5	-38	17	0.0303	28	-13	44	0.0785
6	-36	17	0.0303	29	-12	76	0.1357
7	-34	12	0.0214	30	-11	66	0.1178
8	-33	13	0.0232	31	-10	81	0.1446
9	-32	13	0.0232	32	-9	96	0.1714
10	-31	17	0.0303	33	-8	124	0.2214
11	-30	12	0.0214	34	-7	179	0.3195
12	-29	24	0.0428	35	-6	192	0.3427
13	-28	20	0.0357	36	-5	289	0.5159
14	-27	10	0.0179	37	-4	392	0.6998
15	-26	13	0.0232	38	-3	643	1.1478
16	-25	23	0.0411	39	-2	1416	2.5278
17	-24	22	0.0393	40	-1	5050	9.0150
18	-23	22	0.0393	41	0	24083	42.9915
19	-22	28	0.0500	42	1	14617	26.0934
20	-21	20	0.0357	43	2	5796	10.3467
21	-20	29	0.0518	44	3	1635	2.9187
22	-19	25	0.0446	45	4	111	0.1982
23	-18	26	0.0464				

차이가 없는 경우: 42.99% / 차이가 있는 경우: 57.01%

5.2 정확 매칭- 결과 : 그룹별(사기초 기준) 가입자수와 종사자수의 차이 분포

■ 그룹 : 5인-9인 (도수 100이상)

OBS	DIST	COUNT	PERCENT	OBS	DIST	COUNT	PERCENT
1	-53	10	0.0278	21	-10	59	0.1640
2	-42	12	0.0333	22	-9	77	0.2140
3	-31	11	0.0306	23	-8	84	0.2334
4	-28	12	0.0333	24	-7	134	0.3724
5	-27	12	0.0333	25	-6	144	0.4002
6	-26	19	0.0528	26	-5	190	0.5280
7	-24	11	0.0306	27	-4	307	0.8532
8	-23	13	0.0361	28	-3	510	1.4173
9	-22	18	0.0500	29	-2	1026	2.8513
10	-21	19	0.0528	30	-1	2726	7.5756
11	-20	13	0.0361	31	0	7964	22.1321
12	-19	27	0.0750	32	1	7009	19.4781
13	-18	29	0.0806	33	2	4625	12.8529
14	-17	21	0.0584	34	3	3860	10.7270
15	-16	34	0.0945	35	4	2836	7.8813
16	-15	33	0.0917	36	5	1713	4.7604
17	-14	41	0.1139	37	6	953	2.6484
18	-13	26	0.0723	38	7	561	1.5590
19	-12	52	0.1445	39	8	189	0.5252
20	-11	53	0.1473	40	9	27	0.0750

차이가 없는 경우: 22.13% / 차이가 있는 경우: 77.87%

5.2 정확 매칭- 결과 : 그룹별(사기초 기준) 가입자수와 종사자수의 차이 분포

■ 그룹 : 10인-19인 (도수 100이상)

OBS	DIST	COUNT	PERCENT	OBS	DIST	COUNT	PERCENT
1	-37	13	0.0749	25	-5	174	1.0026
2	-34	11	0.0634	26	-4	238	1.3714
3	-32	11	0.0634	27	-3	368	2.1204
4	-31	10	0.0576	28	-2	581	3.3477
5	-27	11	0.0634	29	-1	1186	6.8338
6	-25	12	0.0691	30	0	2168	12.4921
7	-24	14	0.0807	31	1	2052	11.8237
8	-23	14	0.0807	32	2	1559	8.9830
9	-22	16	0.0922	33	3	1140	6.5687
10	-20	25	0.1441	34	4	905	5.2146
11	-19	19	0.1095	35	5	769	4.4310
12	-18	14	0.0807	36	6	714	4.1141
13	-17	17	0.0980	37	7	680	3.9182
14	-16	24	0.1383	38	8	646	3.7223
15	-15	26	0.1498	39	9	610	3.5148
16	-14	30	0.1729	40	10	578	3.3305
17	-13	34	0.1959	41	11	443	2.5526
18	-12	37	0.2132	42	12	369	2.1262
19	-11	35	0.2017	43	13	242	1.3944
20	-10	47	0.2708	44	14	186	1.0717
21	-9	62	0.3572	45	15	172	0.9911
22	-8	70	0.4033	46	16	99	0.5704
23	-7	85	0.4898	47	17	55	0.3169
24	-6	127	0.7318	48	18	22	0.1268

차이가 없는 경우: 12.49% / 차이가 있는 경우: 87.51%

5.2 정확 매칭- 결과 : 그룹별(사기초 기준) 가입자수와 종사자수의 차이 분포

■ 그룹 : 20인-49인 (도수 30이상)

OBS	DIST	COUNT	PERCENT	OBS	DIST	COUNT	PERCENT
1	-14	33	0.36188	26	12	113	1.23917
2	-12	30	0.32898	27	13	109	1.19531
3	-11	47	0.51541	28	14	127	1.39270
4	-10	45	0.49348	29	15	126	1.38173
5	-9	67	0.73473	30	16	117	1.28304
6	-8	59	0.64700	31	17	119	1.30497
7	-7	75	0.82246	32	18	121	1.32690
8	-6	93	1.01985	33	19	140	1.53526
9	-5	128	1.40366	34	20	187	2.05066
10	-4	185	2.02873	35	21	138	1.51332
11	-3	197	2.16032	36	22	137	1.50236
12	-2	275	3.01568	37	23	124	1.35980
13	-1	404	4.43031	38	24	96	1.05275
14	0	543	5.95460	39	25	113	1.23917
15	1	615	6.74416	40	26	97	1.06371
16	2	491	5.38436	56	27	72	0.78956
17	3	417	4.57287	57	28	60	0.65797
18	4	366	4.01360	58	29	54	0.59217
19	5	319	3.49819	59	30	79	0.86632
20	6	237	2.59897	60	31	58	0.63603
21	7	223	2.44544	61	32	46	0.50444
22	8	199	2.18226	62	33	38	0.41671
23	9	177	1.94100	63	34	45	0.49348
24	10	163	1.78748	64	35	40	0.43864
25	11	147	1.61202	65	38	35	0.38381

차이가 없는 경우: 5.95% / 차이가 있는 경우: 94.05%

5.2 정확 매칭- 결과 : 그룹별(사기초 기준) 가입자수와 종사자수의 차이 분포

■ 그룹 : 50인-99인 (도수 300이상)

OBS	DIST	COUNT	PERCENT	OBS	DIST	COUNT	PERCENT
1	-5	34	0.99824	22	47	42	1.23312
2	-4	36	1.05696	23	48	47	1.37992
3	-3	55	1.61480	24	49	53	1.55608
4	-2	60	1.76160	25	50	54	1.58544
5	-1	84	2.46624	26	51	39	1.14504
6	0	85	2.49560	27	52	39	1.14504
7	1	60	1.76160	28	53	47	1.37992
8	2	76	2.23136	29	54	39	1.14504
9	3	51	1.49736	30	55	47	1.37992
10	4	60	1.76160	31	56	44	1.29184
11	5	56	1.64416	32	57	33	0.96888
12	6	48	1.40928	33	58	26	0.76336
13	7	34	0.99824	34	59	35	1.02760
14	9	31	0.91016	35	60	34	0.99824
15	10	32	0.93952	36	61	33	0.96888
16	13	30	0.88080	37	67	34	0.99824
17	41	30	0.88080	38	69	31	0.91016
18	43	41	1.20376	39	70	44	1.29184
19	44	44	1.29184	40	71	43	1.26248
20	45	35	1.02760	41	72	87	2.55432
21	46	49	1.43864				

차이가 없는 경우: 2.50% / 차이가 있는 경우: 97.50%

5.2 정확 매칭- 결과 : 그룹별(사기초 기준) 가입자수와 종사자수의 차이 분포

■ 그룹 : 100인-299인 (도수 100이상)

OBS	DIST	COUNT	PERCENT	OBS	DIST	COUNT	PERCENT
1	-18	11	0.63474	20	9	10	0.57703
2	-13	11	0.63474	21	10	21	1.21177
3	-9	10	0.57703	22	11	15	0.86555
4	-7	14	0.80785	23	12	18	1.03866
5	-6	15	0.86555	24	13	14	0.80785
6	-5	16	0.92325	25	16	14	0.80785
7	-4	18	1.03866	26	17	13	0.75014
8	-3	21	1.21177	27	23	11	0.63474
9	-2	14	0.80785	28	24	10	0.57703
10	-1	13	0.75014	29	25	13	0.75014
11	0	21	1.21177	30	26	17	0.98096
12	1	19	1.09636	31	27	12	0.69244
13	2	27	1.55799	32	97	13	0.75014
14	3	23	1.32718	33	98	12	0.69244
15	4	13	0.75014	34	99	12	0.69244
16	5	17	0.98096	35	105	11	0.63474
17	6	15	0.86555	36	106	13	0.75014
18	7	22	1.26947	37	110	10	0.57703
19	8	24	1.38488				

차이가 없는 경우: 1.21% / 차이가 있는 경우: 98.79%

5.2 정확 매칭- 결과 : 그룹별(사기초 기준) 가입자수와 종사자수의 차이 분포

■ 그룹 : 300인-499인 (도수 50이상)

OBS	DIST	COUNT	PERCENT
1	6	5	1.44509
2	309	6	1.73410
3	310	7	2.02312
4	311	10	2.89017
5	312	8	2.31214
6	319	6	1.73410

차이가 없는 경우: 0.58% / 차이가 있는 경우: 99.42%

■ 그룹 : 500인-999인 (도수 50이상)

OBS	DIST	COUNT	PERCENT
1	738	6	1.20482
2	739	15	3.01205
3	740	24	4.81928
4	744	10	2.00803
5	758	5	1.00402
6	944	6	1.20482
7	962	7	1.40562
8	980	8	1.60643
9	981	18	3.61446
10	985	5	1.00402
11	986	9	1.80723
12	987	11	2.20884
13	988	19	3.81526

차이가 없는 경우: 0.40% / 차이가 있는 경우: 99.60%

■ 그룹 : 1000인이상 (도수 50이상)

OBS	DIST	COUNT	PERCENT
1	1038	11	2.99728
2	1039	5	1.36240
3	1048	5	1.36240
4	1049	15	4.08719
5	1050	23	6.26703
6	1065	6	1.63488
7	1209	9	2.45232
8	1248	5	1.36240
9	1249	6	1.63488
10	1275	10	2.72480
11	1276	5	1.36240
12	1338	6	1.63488
13	1474	6	1.63488
14	3254	7	1.90736

차이가 없는 경우: 0% / 차이가 있는 경우: 100%

5.2 정확 매칭-다른 기준 변수 사용

▪ 기준변수

기준변수	국민연금자료	사업체기초조사자료
법인등록번호	Corp_id(법인등록번호)	Boupin_nu(법인등록번호)
대표자성명	Boss_nm(대표자성명)	Daep_nm(대표자명)

▪ 결측치 제거

구분	국민연금자료	사업체기초조사자료
Raw data	223,186	741,229
법인등록번호 결측 제거 후	115,465	95,881
대표자성명 결측 제거 후	115,462	95,880
중복 제거 후	3,736	76,575

➢ **법인등록번호가 중복인 관측치가 다수 존재**

➢ **법인등록번호&대표자성명을 기준변수로 하여 정확 매칭을 시행하면 서로 다른 업체가 매칭이 되는 경우 발생**

▪ 예시

기준변수		국민연금자료		사업체기초조사자료	
법인등록번호	대표자성명	사업장명칭	업종	사업체명	주사업내용
11011*****	이**	롯데쇼핑(주)	소매업	롯데디자인팀	인테리어디자인업
11011*****	이**	롯데쇼핑(주)롯데시네마	오락,문화,운동관련	롯데디자인팀	인테리어디자인업
11011*****	이**	롯데쇼핑(주)KKD사업본부	숙박,음식업	롯데디자인팀	인테리어디자인업

5.3 통계적 매칭 - 공통변수(common Variables)

▪ 소재지

	국민연금자료	사업체기초조사자료
변수	Addr(소재지)	Addr_b(소재지_사업체주소번지) Addr_g(소재지_사업체읍면동) Addr_h(소재지_사업체주소호) Addr_ㅣ(소재지_사업체주소리) Addr_t(소재지_사업체주소통) Addr_v(소재지_사업체주소반)
비고	텍스트로 코딩 Ex) 서울시 구로구 개봉동	-

➢ Addr_g(소재지_사업체읍면동) 변수로 통일

▪ 업종

	국민연금자료	사업체기초조사자료
변수	Bs_kind(업종)	Saup(사업의 종류_주사업내용)
비고	63가지 종류	So many

➢ Bs_kind(업종) 변수로 통일

5.3 통계적 매칭 - 공통변수(common Variables)

▪ 사업형태

	국민연금자료	사업체기초조사자료
변수	Bs_type(사업장형태(법인,개인))	Josic_cd (조직형태)
비고	법인/개인	1/2/3/4/5

▪ bs_type과 josic_cd의 교차표

Bs_type (사업장형태)	Josic_cd (조직형태)					총합
	1	2	3	4	5	
개인	58241 (96.40%)	89 (0.16%)	122 (2.79%)	34 (1.36%)	743 (39.52%)	59229
법인	2175 (3.60%)	55560 (99.84%)	4257 (97.21%)	2486 (98.64%)	1137 (60.48%)	65597
총합	60416	55649	4379	2502	1880	124826

※ ()안은 칼럼 백분율임.

➢ Bs_type(법인/개인) 변수로 통일

➢ Josic_cd가 1이면 개인/2,3,4이면 법인/5이면 삭제

5.3 통계적 매칭- 랜덤핫덱적용(예시)

수용자 파일– 사업체기초조사자료

	공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (총종사자수)
1	신정3	교육서비스업	법인	79
2	신정3	교육서비스업	법인	85
3	신정3	교육서비스업	법인	85
4	신정3	교육서비스업	법인	7
5	신정3	교육서비스업	법인	70

+

제공자 파일– 국민연금조사자료

	공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (가입자수)
1	신정3	교육서비스업	법인	18
2	신정3	교육서비스업	법인	6
3	신정3	교육서비스업	법인	5
4	신정3	교육서비스업	법인	16
5	신정3	교육서비스업	법인	16

3개의 공통변수에서 동일한 값을 갖는 관측치는 5개,

그 중에서 랜덤하게 선택

=

매칭된 파일

	공통변수1 (소재지)	공통변수2 (업종)	공통변수3 (사업형태)	유일변수 (총종사자수)	매칭된 변수 (가입자수) : A	정확매칭데이터 (가입자수) : B	차이 |A - B|
1	신정3	교육서비스업	법인	79	16	18	2
2	신정3	교육서비스업	법인	85	5	6	1
3	신정3	교육서비스업	법인	85	6	5	1
4	신정3	교육서비스업	법인	7	16	16	0
5	신정3	교육서비스업	법인	70	16	16	0

5.3 통계적 매칭- 결과

대표성 측면-가입자수

측도 (Measurements)		국민연금 데이터 (Donor File)	매칭된 파일 (Matched Data)
적률	N	120526*	120526
	평균	57.19	59.28
	표준편차	1231.21	1291.27
	평균의 표준오차	3.55	3.72
분위수	100% 최대값	82201	82201
	99%	490	482
	95%	47	46
	90%	21	21
	75% Q3	8	8
	50% 중위수	4	4
	25% Q1	2	2
	10%	1	1
	5%	1	1
	1%	0	0
	0% 최소값	0	0
MAE $\left(\frac{1}{n}\sum\lvert X_i - Median(X)\rvert\right)$		54.99	57.09

두 분포가 유사하므로 매칭결과가 원본 파일의 성질을 그대로 유지한다고 볼 수 있다.

* 사업자등록번호와 대표자성명을 기준으로 한 정확매칭 데이터 124826개에서 조직형태(josic_cd)가 명확하지 않은 관측치 1880개와 공통변수(주소, 조직형태, 사업내용)가 불일치하는 관측치 2420개를 제거

5.3 통계적 매칭- 결과

정확성 측면

차이값	빈도	백분율	누적 빈도	누적 백분율
0	38497	31.94	38497	**31.94**
1	19785	16.42	58282	48.36
2	13163	10.92	71445	59.28
3	8669	7.19	80114	66.47
4	5944	4.93	86058	71.40
5	4298	3.57	90356	**74.97**
6-10	11104	9.21	101460	84.18
11-20	7663	6.36	109123	90.54
21-30	2924	2.43	112047	92.97
31-40	1586	1.32	113633	94.28
41-50	998	0.83	114631	95.11
51-60	720	0.60	115351	95.71
61-70	500	0.41	115851	96.12
71-80	402	0.33	116253	96.45
81-90	320	0.27	116573	96.72
91-100	265	0.22	116838	96.94
101 이상	3688	3.06	120526	100.00

- **정확매칭 자료를 바탕으로 통계적 매칭 결과를 평가**

  정확매칭 자료에서의 가입자수와 통계적 매칭 자료에서의 가입자수 비교

  → 정확하게 일치하는 관측치는 31.94%

  → 5이하의 차이를 보이는 관측치는 74.97%

결론

1. 정확매칭

국민연금 (서울지역) 자료와 사업체기초조사 (서울지역) 자료에 대해 사업자등록번호와 대표자 성명을 기준변수로 사용하여 정확 매칭을 적용시켜보았다. 정확 매칭된 자료로부터 사업체기초조사자료의 종사자수와 국민연금자료의 가입자수의 일치율 및 비일치에 따른 자료의 분포를 파악해 본 결과 종사자 그룹에 따른 일치율이 크게 차이가 있으며, 종사자수가 증가함에 따라 일치율이 크게 감소하는 것으로 나타났다.

2. 통계적 매칭

사업체기초조사자료(수용자 파일, 유일변수='종사자수')와 국민연금자료(제공자 파일, 유일변수='가입자수')를 매칭시켰다(공통변수:소재지, 업종, 조직형태의 3가지 범주형 변수).

-대표성 측면: 제공자 파일인 국민연금 자료에서의 가입자수의 분포와 통계적 매칭이 된 자료에서의 가입자수의 분포를 비교해 본 결과 두 분포가 거의 유사한 것을 알 수 있었다. 따라서 통계적 매칭결과가 원본 파일의 성질을 잘 유지하고 있다고 할 수 있다.

-정확성 측면: 정확매칭된 자료를 바탕으로 통계적 매칭의 결과에 대한 정확성을 평가해 본 결과 실제값과 매칭에 의한 값이 정확하게 일치하는 경우는 5이하인 경우는 전체의 74.97%로 비교적 높은 것을 알 수 있었다.

6. 결론 및 향후과제

▪데이터매칭의 정의와 분류와 대해 살펴보았고 데이터 매칭 중에서도 특히 통계적 매칭 알고리즘에 대해 살펴보았다. 매칭의 대표성이나 정확성 측면에서의 향상을 위한 통계적 매칭 알고리즘에 대한 연구는 현재 활발히 진행 중이다.

▪ 향후 과제

▪ K-NN를 적용하여 매칭할 경우 distance의 극단값에 대한 고려:

예를 들어 k=3인 경우 distance가 각각 $D_1 \leq D_2 << D_3$인 경우 다음과 같은 방법들을 고려해 볼 수 있다.

1. 각각의 경우에 weight를 주는 방법(distance가 D_1인 경우는 weight를 크게, D_3인 경우는 weight를 작게)
2. 각각을 표준화시켜 극단값의 영향을 줄이는 방법
3. 연속형의 경우 3개의 평균을 사용하는 대신 median값으로 사용

▪ 데이터 매칭 후 평가방법에 대한 고찰

제공자 파일에서의 분포와 통합 파일에서의 분포 비교시 절대적 기준이 필요

→ t-검정이나 카이제곱-검정 등의 검정방법을 통한 결론 도출

References

1) 이영섭, 김선웅, 안홍엽, 임경은, 김희경(2009). 통계조사자료와 행정자료간의 통계적 매칭기법에 관한 연구, *통계연구*, 제 14권 제 1호, 82-98.

2) 정성석, 김순영, 김현진(2004). 데이터 보강을 위한 데이터 통합기법에 관한 연구, *응용통계연구*, 제 17권 3호, 605-617.

3) 한상훈, 안일호, 하덕주, 최종후(2004). 데이터 퓨전과 평가, *한국데이터마이닝학회 2004 추계학술대회*, 238-254.

4) D'Orazio, M., Di Zio, M. and Scanu, M.(2006), *Statistical Matching: Theory and Practice*, John Wiley & Sons.

5) Ingram, D., O'Hare, J., Scheuren, F. and Turek, J.(2000), Statistical matching: a new validation case study, Proceedings of the survey Research methods Section, American Statistical Association.

6) Kadane, J. B.(1978). Some statistical problems in merging data files, In department of Treasury, Compendium of Tax Research, 159-179.

7) Kiesl, H., Rässler, S.(2006). How valid can data fusion be?, IAB Technical Report 200615, Institute for Employment Research, Nuremberg, Germany.

References

8) Li, Qi and Racine, J. S.(2007). *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

9) National Statistics(2003). National Statistics code of Practice Protocol on Data Matching. http://www.statistics.gov.uk/about/consultations/general_consultations/downloads/Protocol_on_Data_Matching.pdf

- Rässler, S.(2002). Statistical Matching: A frequentist theory, practical applications, and alternative Bayesian approaches, Springer Verlag, New York.
- Rodgers, W. L.(1984). An Evaluation of Statistical Matching, Journal of Business and Economic Statistics 2, 91-102.
- Rubin, D. B.(1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business and Economic Statistics 4, 87-94.
- Singh, A. C., Mantel, H., Kinack, M. and Rowe, G.(1993). Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption, Survey Methodology 19, 59-79.

References

14) U.S. Department of Commerce. (1980). Report on exact and statistical matching techniques, *Statistical Policy Working Paper5. Washington, DC: Federal Committee on Statistical Methodology*.

15) Van der Putten, P,. Kok, J. N., and Gupta, A.(2002). Why the information explosion can be bad for data mining, and how data fusion provides a way out, *Second SIAM Internatoinal Conference on Data Mining*, Arlington, April, 11-13.