「무응답처리기법(Imputation)」

해외출장 결과보고서

2007. 12.

통 계 청

Ⅰ. 출장개요

□ 필요성

○ 최근 들어 사생활 보호를 중요시하며 외부에 개인정보 노출을 기피하는 경향이 점차 증대되고 있어 대규모 총조사에서 부재·불응가구 및 항목 무응답이 크게 증가

○ 2010년 인구주택총조사에서는 보다 많은 부재·불응가구와 항목 무응답이 나타날 것으로 추정되므로 통계적 무응답처리기법인 임퓨테이션(Imputation)에 대한 보다 심층적인 연구가 요구됨

□ 목적

○ 인구주택총조사 및 무응답 처리가 필요한 각종 통계조사의 정도 제고를 위해 부재·불응가구 및 항목무응답에 대한 선진통계기법 연수

- 캐나다에서 사용하고 있는 무응답처리(Imputation) 기법을 파악하여 우리나라에 맞는 무응답처리기법 연구

□ 출장지 및 출장기간

○ 출 장 지 : 캐나다 통계청

○ 출장기간 : 2007. 10.31. ~ 12.15.

□ **출장자** : 3명

○ 통계개발원 연구기획실 : 최 필근 사무관
○ 강원지방통계청 경제조사과 : 임 영일 주무관
○ 경제통계국 분석통계과 : 최 지은 주무관

□ **출장일정 및 내용**

일자	일정	내용
10.31. ~ 11. 1.	· 출발 (인천 → 오타와)	
11. 2. ~ 11. 6.	· 출장기관 방문 및 준비사항 점검	· 출입관련 행정처리 - 신원조회 및 출입증 발급 등 · 프로그램 설치 · 오리엔테이션 교육
11. 7. ~ 11.22.	· 캐나다 통계청 출·퇴근	· BANFF 프로그램 교육 · 광공업동태자료 BANFF 시스템 적용
11.23. ~ 12.13	· 캐나다 통계청 출·퇴근	· CANCEIS 프로그램 교육 · 인구주택총조사자료 CANCEIS 시스템 적용
12.14. ~ 12.15.	· 귀국 (오타와 → 인천)	

□ **소요예산**

○ 총 천원

Ⅱ. 연구결과보고

□ 연구효과

○ 2010년 인구주택총조사시 부재불응가구 및 무응답항목에 대한 고품질의 무응답처리(Imputation) 기법을 적용하여 정도 높은 인구주택총조사 자료 생산

○ 무응답처리(Imputation)가 필요한 청내 가구 및 사업체부문 통계의 선진 무응답처리기법 적용으로 자료의 정도 제고

□ 연구내용

가. 2005 인구주택총조사 자료의 CANCEIS 적용

□ CANCEIS System의 장점

○ Edit와 Imputation의 시스템화

- DA(DLT Analyzer), IE(Imputation Engine), DE(Derive Engine)의 체계적인 시스템

○ DLT(Decision Logic Tables)

- DLT를 통하여 유효하지 않은 자료의 결정, 도너의 결정
- DLT의 문법오류 체크와 DLT별 오류건수 현황

○ 가구원수(STRATA)에 의한 Imputation

- 가구원별 임퓨테이션이 아닌 가구원수가 같은 가구별로 임퓨테이션하여 가구구조의 특성을 반영하여 임퓨테이션

○ Imputation 과정의 이론적 배경

- Near minimum change imputation
- 13가지의 Distance Functions 등 다양한 이론적 배경의 뒷받침

○ Imputation 과정의 기록
- AUDIT file에 임퓨테이션되어지는 과정의 흐름이 상세하게 기록

○ Imputation된 자료의 관리
- 임퓨테이션후 나오는 결과자료에 대한 체계적인 관리

□ CANCEIS 적용시 문제점
○ 원시자료의 CANCEIS System에 맞도록 가공 수정
- 원시자료를 그대로 이용하는 것이 아니고 STRATA(가구원수), GROUP(지역)에 의해 자료를 나누어 임퓨테이션
- 원자료에서 공백에 대하여 별도의 값을 부여하여 임퓨테이션하여야 함

○ DLT 함수사용의 제약
- Rand, Max, Min, Sum, Avg, Count 등의 함수가 존재하나 Derive DLT에서만 사용가능하고 Mod 등 기타 함수사용이 제한됨

○ EDIT의 경우 불일치 자료에 대하여 곧 바로 임퓨테이션하나 이상치이며 유효한 값에 해당하는 경우는 도너에서 배제될 뿐 Edit 기능이 없음

□ 향후 계획

□ 2010 인구주택총조사의 Edit & Imputation 시스템 구축
○ 2005 인구주택총조사의 경우 Edit는 E-census내에 구축되었고 Imputation은 시스템이 아닌 PC에서 SAS 프로그램을 통하여 임퓨테이션 후 원 자료에 반영함
○ 2010 인구주택총조사는 Edit & Imputation 시스템을 구축하여 효율적인 내검 체계와 무응답에 대한 체계적인 관리와 정도 높은 자료 생산

□ Edit & Imputation 시스템의 방향
○ Editing 시스템의 이원화
- 이상치이며 유효한 값은 오류의 가능성을 가지고 있으므로 내검요원을 통한 내검후 도너에서 제외(예, 100세이상인 나이)
- 불일치 자료의 경우 Deterministic Imputation을 먼저 한 후에 Donor Imputation

○ Deterministic Imputation의 별도 관리

- 3가지 이상의 변수가 연관되는 경우 2가지 이상 변수의 조건이 같으면 다른 변수는 자동으로 수정되도록 함

○ 별도의 원자료 가공없이 시스템상의 가상공간에서의 작업 기능
- 가구원수별 또는 지역별 분리, 공백자료의 임의값 삽입 없이 시스템상의 가상공간에서 작업 후 임퓨테이션

○ 체계적인 임퓨테이션 과정 및 결과 자료 관리
- 임퓨테이션 과정의 흐름을 볼 수 있는 자료의 생성과 에디팅 및 임퓨테이션 된 자료의 연계 자료의 생성

○ Editing의 체계적인 관리
- Editing의 순서, 기능별 구성과 현황에 대한 관리
- Editing rule의 문법적 오류 체크 기능

○ 우리나라 시스템에 맞은 거리함수 및 이론적 배경
- 개발원의 연구과제로 선정하여 이론적 배경의 뒷받침
- 변수의 중요도에 따른 임퓨테이션 순서, CHAID 분석을 통한 보조변수의 선정
- Near Minimum Change Imputation의 거리함수와 도너선정

Ⅲ. 별첨

□ 연구결과 보고서

「무응답처리기법(Imputation)」

해외출장 결과보고서[Ⅰ]

2005 인구주택총조사 자료의 Canceis 적용

2008. 1.

통계개발원 연구기획실 : 최 필 근

강원지방청 경제조사과 : 임 영 일

경제통계국 분석통계과 : 최 지 은

차 례

Ⅰ. 연구개요 1

1. 연구 필요성 1
2. 연구 목표 1
3. 연구 범위 1
4. 연구 효과 2
5. 향후 방향 2

Ⅱ. CANCEIS 프로그램 사용 방법 3

1. CANCEIS의 원리와 기능 3
2. 프로그램 세부 사항 요약 8

Ⅲ. 2005인구주택총조사 자료의 CANCEIS 적용 15

1. INPUT File 15
   1.1 DATA File 15
   1.2 DATA DICTIONARY Files 16
   1.3 SYSTEM PARAMETER File 20
   1.4 HOT-DECK PARAMETER File 21
   1.5 DLT File 24
2. IMPUTATION 26
   2.1 실행파일 RUNTEST.BAT 26
   2.2 DLT Analyzer 27
3. OUTPUT File 29
   3.1 DLT Analyzer 29
   3.2 IMPUTATION Engine 36

Ⅳ. 결론 및 향후 방향 55

1. CANCEIS System 적용결과 55
2. 향후 추진 방향 56

Ⅰ. 연구개요

1. 연구 필요성

□ 2000, 2005 인구주택총조사 Imputation 적용을 통하여 나타난 여러 가지 문제점을 해결하고자 선진통계 기법 및 시스템에 대한 습득 필요

□ 이를 위해 통계 선진국인 캐나다 인구센서스의 Imputation 시스템을 파악하여 2010 인구주택총조사 Imputation 시스템 구축

2. 연구 목표

□ 본 연구는 2005 인구주택총조사 결과자료를 캐나다 통계청의 Census Imputation 시스템인 CANCEIS(CANadian Census Edit and Imputation System)에 적용하여 2010 인구주택총조사 Imputation 시스템 구축의 방향을 검토

3. 연구 범위

□ 2005 인구주택총조사의 광범위한 항목과 내검규칙을 모두 적용하기에는 한계가 있으므로 표본자료 중 서울특별시의 4인가구를 대상으로 하였으며 항목은 전수항목(남북이산가족 제외)과 표본항목중 관련되는 혼인년월과 출생자녀수를 포함하였음

4. 연구 효과

□ 2010년 인구주택총조사시 부재불응가구 및 무응답항목에 대한 고품질의 무응답처리(Imputation) 기법을 적용하여 정도 높은 인구주택총조사 자료 생산

□ 무응답처리(Imputation)가 필요한 청내 가구 및 사업체부문 통계의 선진 무응답처리기법 적용으로 자료의 정도 제고

5. 향후 방향

□ 내검규칙 적용(DLT)와 보조변수의 구성에 따른 CANCEIS 시스템을 적용한 자료와 2005 인구주택총조사 최종 결과자료를 비교분석하여 최적의 내검규칙과 보조변수 구성 검토

□ 가구별 Imputation과 가구원별 Imputation의 차이분석을 통하여 효율적인 Imputation 적용 검토

□ 이산형 또는 연속형 자료의 경우 최적의 거리함수 검토

□ 시험조사 자료의 Imputation 적용, 분석으로 2010 Edit & Imputation 시스템 구축

Ⅱ. CANCEIS 프로그램 사용 방법

1. CANCEIS의 원리와 기능

□ CANCEIS(CANadian Census Edit and Imputation System)란?

- ○ 센서스 변수들에 대해서 edit와 imputation (E&I)을 수행하는 시스템이다.
- ○ 범주형, NUMERIC, ALPHANUMERIC 변수에 대해 동시에 적용가능한 절차이다.
- ○ Nearest neighbor imputation methodology(NIM)을 이용한 minimum change donor imputation(donor를 최소한으로 변경하는 임퓨테이션 방법)
  deterministic imputation을 수행하며 derive module에서 새로운 변수를 추출하여 이용한다.

□ Canceis 핫덱 모듈에서의 내검

- ○ 좋은 응답(라벨 포함)을 정의하기 위해서 각 변수에 대해 validity set을 이용하며, validity set은 dictionary에 정의되어 있다.
- ○ 모순된 내검규칙을 구체화하기 위해서 Decision Logic Table(DLT)을 이용한다.
  시간을 절약해주는 형태로 compact DLTs가 있다.

□ DLT : Hot Deck Module

- ○ example 1:

Propositions	Edits	
AGE < 15	Y	N
MART_STAT=CLASS(ever_married)	Y	
YEARS_MARRIED > 0		Y
AGE - YEARS_MARRIED(#1) > 15		Y

- ○ example 2 :

	AGE	MART STAT	YEARS MARRIED
Failed Record :	13	MARRIED	02
Imputed Record(12) ;	13	MARRIED	12

□ DLT : Derive Module

DO DRIVE AGE
MIN AGE = 15

Common Actions

AGE (#1) < MIN AGE	Y	N
MART STAT(#1) = CLASS(ever_married)	Y	
YEARS MARRIED (#1) > 0		Y
AGE(#1) - YEARS_MARRIED(#1) > MIN_AGE		N

Propositions

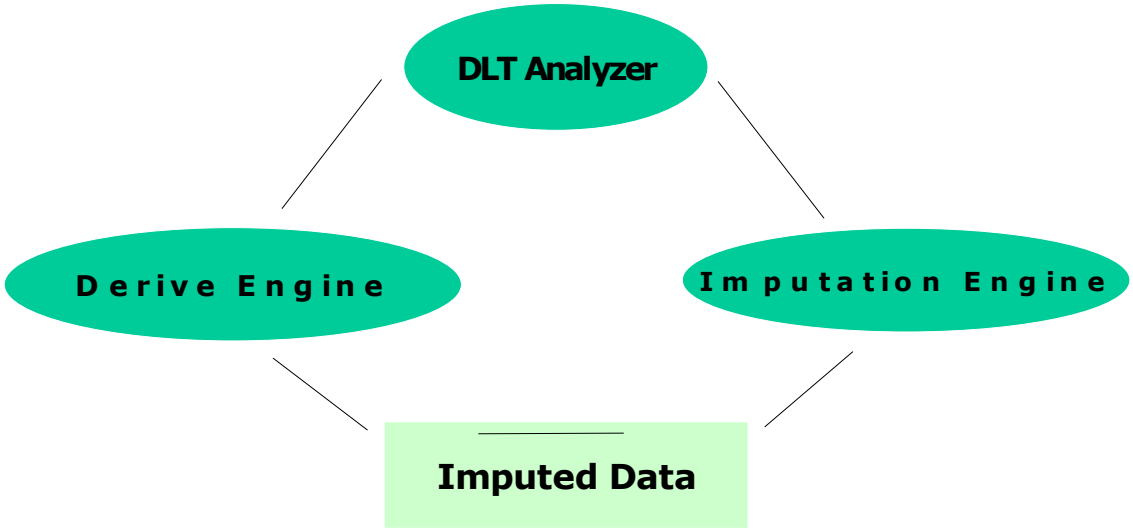
MART_STAT(#1) = SINGLE	X	
YEARS_MARRIED(#1) = AGE(#1) - 15		X

Conditional Actions

Rules

□ DLT : Hot Deck 모듈

○ System 구성요소



□ DLT Analyzer의 기능

○ DLT와 Dictionary를 비교한다.
○ 내검 명제의 논리적 구조를 체크한다.
○ 내검 명제를 표준화한다.
○ 내검 명제를 확장한다.
○ 핫덱에서는, 하나의 규칙의 명제를 가지는 통합 DLT를 만든다.

☐ Derive Engine

- ○ DLT Analyzer를 이용한다.
- ○ 각각의 record에 대해 모든 DLT를 평가한다.
- ○ 문제있는 ('problem') record에 대해 deterministic 임퓨테이션을 수행한다.
- ○ 변수들을 이끌어내고 업데이트한다.
- ○ 사용자가 구체하한 순서로 DLTs를 실행한다.

☐ Examples of Derive DLT

```
* ************************ **********
% DLT Name:                    MARSTMSTR
@ MARST(#1) = CLASS(EVER_MARRIED) ;Y;
& DO RANDOM                          ;X;

* ************************ **********
% DLT Name:                    RANDOM
$ R = RAND(0,100)
@ R < 25                        ;Y;N;N;ELSE;
@ R < 50                        ; ;Y;N; ;
@ R < 75                        ; ; ;Y; ;
& MARST(#1) = DIVORCED             ;X; ; ; ;
& MARST(#1) = NOW_MARRIED           ; ;X; ; ;
& MARST(#1) = SEPARATED             ; ; ;X; ;
& MARST(#1) = WIDOWED               ; ; ; ;X;
```

☐ Imputation Engine

- ○ DLT Analyzer에 의해 만들어진 통합(unified) DLT를 이용한다.
- ○ 모든 record의 내검을 수행한다.
- ○ Nearest neighbor donor로부터 가능한 최소한의 변수들을 임퓨트한다.
- ○ 임퓨트된 records와 reports를 생성한다.
- ○ 임퓨트된 records는 내검규칙을 패스한다.

☐ Donor Imputation Strategy

○ CANCEIS는 "Nearest-neighbor imputation methodology"- NIM에 근거하고 있다.
① donor를 찾는다
② 주어진 donor하에서, 임퓨트할 최소한의 변수를 결정한다.

○ Fellegi-Holt
① donor를 참고하지 않고 임퓨트할 최소한의 변수를 결정한다.
② donor를 찾는다.

☐ Implementation of Min Change

○ 가능한 한 많은 변수에 대해, 내검 규칙을 통과하지 못한 가구(failed edit household)와 매치되는 가구를 찾기 위해 그 가구와 지리적으로 가까운 1000개의 내검을 통과한 가구(passed edit household) 중에서 탐색한다.

○ 가장 가까운 거리에 있는 40개의 내검을 통과한 가구(nearest neighbors)가 선택되고 imputation actions를 생성하는 데에 사용된다.

☐ Nearest-Neighbour

○ Failed Household (내검 통과하지 못한 가구)

person 1	married	38
spouse	married	35
mother	**----**	**41**

○ Nearest - Neighbour (가장 가까운 이웃)

person 1	married	36
spouse	married	37
mother-in-law	**widowed**	**59**

Dfp = Distance = 3 + 0.1 + 0.1 = 3.2

* 3 : mother-in-law, widowed, 59 에 각각 1점 부여
* 0.1 : 36, 37에 각각 0.1점 부여

☐ Impute Blanks/Invalids First

○ nearest-neighbor에서 응답을 빌려옴으로써 빈칸/유효하지 않은 값(blank/invalid)을 임퓨트한다.
○ 만일 모든 내검 규칙을 패스하면 중지한다.
○ 내검규칙을 패스하지 못하면, 계속한다.

☐ Impute Essential to Impute Variables

○ 어떤 변수를 임퓨트할 필수적인 요소가 있는지 여부를 결정하기 위해서 실패한 내검(failing edits)이 평가된다.
○ 그러한 변수들은 즉각적으로 임퓨트된다.
○ 빈칸/유효하지 않은 값(blank/invalid)과 필수적 변수들이 우선적으로 임퓨트되는 것이다.

☐ Simplifying Edit Rules

○ 모든 실패한 record/donor의 짝에 대해서
- 시스템은 자동적으로 imputation action이 패스하지 못할 수 있는 내검 규칙들만 남긴다.
- 이 프로세스가 효율적으로 진행되면, 대부분의 경우 내검 규칙의 수는 획기적으로 줄어든다.

☐ Selecting One Imputation Action

○ 40개의 nearest-neighbor donors 각각에 대해서, 내검 규칙을 통과하지 못한 가구를 위한imputation action이 적용된다.
○ 가장 좋은 5개의 imputation action이 남고 그중 하나가 랜덤으로 선택된다.
○ 5개의imputation actions는 imputation process의 기대값/분산을 계산하는 데에 사용될 수 있다.

☐ Ranking Imputation Actions

○ distance measure(거리 측정)방식으로 5개의 베스트를 결정하는 것이다.

Dfpa = 0.9Dfa + 0.1Dap

○ Dfa = 내검에 실패한 가구의 imputation action과의 거리로서 최소한의 변화(minimum change)를 나타낸다.

○ Dap = imputation action과donor와의 거리로서 개연성(plausibility)을 나타낸다.

○ 모수 0.9는 min.change에 더 가중치를 둔 것을 보여준다.

☐ Discarding Imputations

○ imputation actions를 생성할 때, 어떤 것들은 부가적인 변수들을 임퓨트하지 않고 버려질 수 있다.

- Dfpa가 이미 너무 크거나
- 부가적인 변수들이 불필요하게 임퓨트되었거나
- 이 donor가 유효한 IA를 생성하지 못하는 것으로 이미 결정되었기 때문이다.

○ imputation actions가 적게 생성되거나 평가될 때 이 방법은 매우 효율적으로 간주된다.

☐ **CANCEIS의 장점**

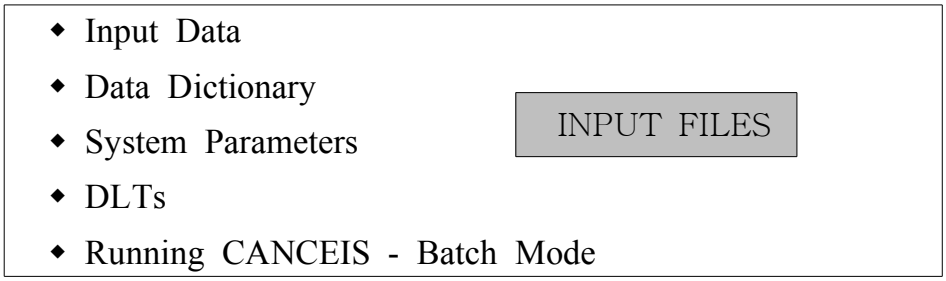
○ 메모리의 효율적 이용 - 빠른 프로세스

○ 비용의 효율성 - PC 프로세스

○ 휴대성 text input files의 사용, 특정한 database에 연계되지 않음

○ 노동력이 많이 요구되는 내검과정의 업무량을 줄이기 위해 유용한 tool을 제공

2. 프로그램 세부사항 요약



☐ Input Data

○ 두개의 모듈이 있다.

① **regular mode** (한번에 한 개의 층으로 구성)

unit	sub-unit	data			
10001	01	23	3	2	6
10001	02	52	1	1	5
10002	01	58	1	2	4
10002	02	83	4	1	4

② **job mode** (한번에 여러 개의 층으로 구성)

층	unit	sub-unit	data			
02	10001	01	23	3	2	6
02	10001	02	52	1	1	5
02	10002	01	58	1	2	4
02	10002	02	83	4	1	4
03	20001	01	79	0	3	5
03	20001	02	13	4	2	4
03	20001	03	23	4	1	4

◆ **층화**

- 층은 사용자에 의해 어떤 데이터셋에 대해서도 정의될 수 있으며 각 층은 다른 특징을 지닐 수 있다. DLT Analyzer가 층화 프로세스를 실행한다.
- impuation group은 한 층의 subset이다. 한 층의 모든 IGs는 같은 layout으로 나타나야 한다. Engines가 IGs 프로세스를 실행한다.

◆ **파일명**

- xxxfiletypeyyIGzz.txt
  - xxx = subject (3-8 characters)
  - yy = stratum number (1-5 digits)
  - zz = imputation group (1-4 digits)

○ 요구되는 포맷

- 탭은 사용불가, 한 줄에 하나의 sub-unit, numeric(또는 alphanumeric)

□ Data Dictionary

○ 변수, 가중치와 거리함수(핫덱), validity sets, value sets, classes에 대해 정의한다.

○ 수치 정보와 범주형(categorical, coded) 정보를 모두 정의한다.

○ 파일들은 필수적 또는 조건부 옵션(모듈 타입에 따라)일 수 있다.

ex.1) Coded variables(성별)

MSEX(Value Set)	VSEX(Validity Set)
1-BLANK	FEMALE
2-FEMALE	MALE
3-INVALID	
4-MALE	

❖ DLT에서는 숫자가 아니라 라벨을 써야 함

ex.2) Discrete variables(나이)

MDISC(Value Set)	VAGE(Validity Set)
(-∞, ∞)	AGEDOMAIN 0 121 1

ex.3) continuous(연속형) 변수(소득)

MCONT(Value Set)	VINCOME(Validity Set)
(-∞, ∞)	INCDOMAIN 0 1000000

ex.4) Alphanumeric 변수(postal code)

MCHAR(Value Set)	VPC(Validity Set)
모든 수, 문자, 심볼 등	PCPATTERN
	[A-Z][0-9][A-Z][0-9][A-Z][0-9]

◆ **value sets**

-input data에서 한 변수가 취할 수 있는 모든 가능한 값을 정의한다.

-코드화된 값은 사용자에 의해 정의된다.

(연관된 응답은 수치 코드와 함께 라벨을 붙인다.)

-수치 값은 시스템에 의해 모든 가능한 수를 포함하도록 정의된다.

◆ **validity sets**

-주어진 모듈하에서 유효한 값을 정의한다.

-코드화된 validity sets는 value set의 subset으로 정의된다.

□ System Parameters

○ CANCEIS를 자료에 맞추어 실행할 수 있게끔
- 내검(Editing)한다.
- 재배치(Reordering)한다.
- Donor 찾는다.
  ◆단계별로 donor의 size/number
  ◆Max DFP
  ◆ 개선 비율(improvement ratio)
  ◆ donor의 재사용
  ◆ donor로서 실패한 record의 사용
- 거리를 측정한다.
  ◆최소한의 변화(minimum change) vs. 개연성(plausibility)
  ◆ Imputation actions의 수
  ◆ NMCIA 리스트
- Outlier(아웃라이어) 진단한다.
- Output files/ reports
- Audit Trail 수행한다.

□ Input Data의 Sorting

○ nearest neighbor를 결정하는데 매우 중요하다.
○ 지리적 근접성에 의해 sort된다.
○ 임퓨트된 변수와 관계가 있는 다른 변수에 근거해sort하는 것도 가능하다.
○ 매칭 변수와 관련하여 결정될 수 있다.

□ Choosing Matching Variables

○ 좋은 donor를 찾는 데에 도움이 되는 변수
- 임퓨트되는 변수들과 관계가 있어야 함
-가중치를 적용하여 정의함
  ◆ 0 : 매치에 사용되지 않음
  ◆ positive(양의 수) : 매치에 사용됨

○ 필수 조건
- donor로서 고려되어야 하는 record의 특징 max Dfp

□ Choosing Distance Functions

○ 각각의 매칭 변수들에 대해서

- record의 전체 스코어에 어떻게 기여하는 지를 결정한다.
  - ◆ 1 : 완벽하게 매치되거나 아닌 경우
  - ◆ 3 : 이산 numeric 변수에 대해서distance function을 맞출 수 있게 함
  - ◆ 4 : 어떤 실패한/donor 값의 조합(matrix)에 대해서든 사용자로 하여금 변수에 대한 distance를 구체화할 수 있게 함. (하지만 매우 큰 matrix들만이 distance를 낮출 수 있음)

○ distance function과 변수의 가중치, Max Dfp 값의 관계를 반영한다.

□ Decision Logic Tables (DLTs)

○ DLTs의 세가지 유형

- 핫덱 (선택적)
  - ① Consistency(일관성 유지 방식)
  - ② Donor Selection(donor 선택 방식)
- Deterministic (필수적)
  - ③ Derive (논리적 관계로 유도되는 방식)

○ 구성

- DLT Header - parameters를 정의한다.
  - ▪ % DLT Name
  - ▪ % Strata
  - ▪ % Type
  - ▪ % Purpose
  - ▪ % Overlapping
  - ▪ % Symmetry
  - ▪ % Sub-unit Start/End Position
  - ▪ % Substitution Symmetry
- DLT Body - 내검 규칙을 정의한다.
  - ① Conditions
    - ▪ Conditions (@) - Hot-Deck/Derive
      내검 규칙을 만족하기 위해 반드시 충족시켜야 하는 조건을 정의
  - ② Actions
    - ▪ Conditional Actions (&) - Derive
      conditional 내검 규칙이 만족될 때 취해야 하는 action

▪ Common Action ($)

모든 record에 대해 적용될 actions를 정의

③ Rules

□ Writing Compact DLTs

○ Variable position sub-units

- #1을 사용함으로써 어떤 명제가 그것이 복수의 sub-unit로 확대될 수 있어야 함을 가리킨다.
- sub-unit numbers를 나타내는 데에만 사용되어야 한다.

  MARST(#)=SINGLE
- ex) variable position sub-unit의 compact 형태

  % symmetry : no

  % sub-unit start position : 1

  % sub-unit end position : 3

  @ sexu(#1) = sex(#2) ; Y ; N

  & Var1(#2) = A_blank ; X ; ;

  & Do table_2 ; ; X ;

○ Text Substitution(대체)

- [1]을 사용함으로써 어떤 명제가 수들의 시리즈로 확장될 수 있어야 함을 가리킨다.
- 명제의 어떤 곳에서나 사용될 수 있다.

  FLAG[1]=BLANK

ex) text substitution의 compact 형태

```
% Substitution Start Position: 1
% Substitution End   Position: 3          DLT Header

$  DECL(VAR?1TEMP,D)
$  DO TABLE1                              Common Actions

@ VAR?1TEMP = 25          ;N;             Conditions

& VAR?1TEMP = 10        ;X;
& DO TABLE2                   ;X;
& VAR?1TEMP = 25        ;X;               Conditional Actions
```

□ Running CANCEIS Batch

○ text 파일(.bat확장자)에 있는 command lines를 이용하여 실행한다.
- DLT Analyzer (statum 수준에서 실행)
- Imputation Engine(imputation 그룹 수준에서 실행)
- Derive Engine(imputation 그룹 수준에서 실행)

○ declare할 것
- Executable, Stratum, Job Number (if run in Job mode), Input files directory, Output files directory

□ Output Files

○ DLT Analyzer 로부터
- 통합 DLT의 생성
- Derive 또는 imputation engines에 필수적인 다른 파일들을 만듦

○ Derive Engine으로부터
- status files의 프로세스
- statistical report (디테일이 나와 있음)
- derivation 이후의 파일들

○ Imputation Engine으로부터
- status file의 프로세스
- statistical reports
- 내검과 임퓨테이션을 통해 프로세스된 데이타
- 임퓨테이션 이후의 파일들

Ⅲ. 2005인구주택총조사 자료의 CANCEIS 적용

2005인구주택총조사 자료를 캐나다 CANCEIS시스템에 적용하기 위하여 내검이 완벽하지 않은 2005년 12월자료를 사용하였으며, 자료의 크기상 서울의 표본자료로 조사구특성 1(일반), A(아파트)의 자료 304,170가구의 875,748명을 대상으로 하였다.

전수항목중 남북이산가족항목을 제외하였고, 표본항목중 성별 및 나이에 영향을 줄 수 있는 혼인년월과 총출생아수(남, 여)를 포함하였다.

1. INPUT File

1.1 DATA File

○ FILENAME : xxxUNITyyIGzz.txt (regular mode) 또는
xxxUNITJwwIGzz.txt (job-mode)

- job-mode를 사용하여 INGUUNITJ01IG01.txt로 설정

○ 공백을 허용하지 않아 공백자료는 특정한 값(0, -1)으로 대체하였고 유효한 공백인 자료는 별도의 값을 부여하였다

Variable name		Value	비고
STRATUM ID		04	- 4인가구
UNIT ID		110105100410004001	- 행정구역+조사구번호+조사구특성번호+거처번호+가구번호
SUB-UNIT ID		1	- 1,2,3,4
VARIABLE	R2P1	01	- 01~14, 공백은 00
	SEX	1	- 1,2, 공백은 0
	AGE	70	- 1~121, 공백은 -1
	ZODIACAL	12	- 01~12, 공백은 00
	BIRTHYEAR	1935	- 1887~2007, 공백은 -1
	EDU1	4	- 1~8, 공백은 0
	EDU2	1	- 1~5, 9, 공백은 0
	RELIGEON1	2	- 1,2, 공백은 0
	RELIGEON2	90	- 01~99, 공백은 00
	MARST	2	- 1,2, 공백은 0
	MARYEAR	1963	- 1903~2007, 공백은 -2, 유효한 공백은 -1
	CHILDB	-1	- 0~10, 공백은 -2, 유효한 공백은 -1
	CHILDG	-1	- 0~10, 공백은 -2, 유효한 공백은 -1

1.2 DATA Dictionary Files

1) VAR File

○ UNIT 파일의 순서대로 구성

Variable ID	Repeatable	Validity set ID	Status	Download ID
R2P1	R	VR2P1	1	R2P1
SEX	R	VSEX	1	SEX
AGE	R	VAGE	1	AGE
ZODIACAL	R	VZODIACAL	1	ZODIACAL
BIRTHYEAR	R	VBIRTHYEAR	1	BIRTHYEAR
EDU1	R	VEDU1	1	EDU1
EDU2	R	VEDU2	1	EDU2
RELIGEON1	R	VRELIGEON1	1	RELIGEON1
RELIGEON2	R	VRELIGEON2	1	RELIGEON2
MARST	R	VMARST	1	MARST
MARYEAR	R	VMARYEAR	1	MARYEAR
CHILDB	R	VCHILDB	1	CHILDB
CHILDG	R	VCHILDG	1	CHILDG

○ Repeatable

- R(Repeated) : sub-unit(가구원)가 있는 경우 R
- N(Non-Repeated) : sub-unit가 없는 경우 N

○ Status

- 1 : input file, output file 모두 존재
- 2 : input file은 존재, output file에는 없음
- 3 : 새로운 변수로 input file에는 없으나 output file에 존재
- 4 : 새로운 변수로 input file, output file에 모두 없음

2) SET File

○ valid set에 대한 자료형태 정보

Validity set ID	Data Type	Value Set ID
VR2P1	O	MR2P1
VSEX	O	MSEX
VAGE	D	MDISC
VZODIACAL	O	MZODIACAL
VBIRTHYEAR	D	MDISC
VEDU1	O	MEDU1
VEDU2	O	MEDU2
VRELIGEON1	O	MRELIGEON1
VRELIGEON2	O	MRELIGEON2
VMARST	O	MMARST
VMARYEAR	D	MDISC
VCHILDB	D	MDISC
VCHILDG	D	MDISC

○ Data Type : O(coded), D(Discrete), C(Continuous), H(Alphanumeric)

○ Value Set ID

- Data type이 Discrete이면 MDISC
- Data type이 Continuous이면 MCONT
- Data type이 Alphanumeric이면 MCHAR

3) CODE, VSCODE File

○ VSCODE File - 코드에 대한 정보파일로 공백자료에 대한 코드 정보도 포함

○ CODE File - 유효한 값에 대한 정보파일

CODE		VSCODE		
Validity set ID	Label	Value Set ID	Code	Label
		MR2P1	0	MISSING
VR2P1	HOUSEHOLD	MR2P1	1	HOUSEHOLD
VR2P1	SPOUSE	MR2P1	2	SPOUSE
VR2P1	SON_DAUGHTER	MR2P1	3	SON_DAUGHTER
VR2P1	SON_DAUGHTER_INLAW	MR2P1	4	SON_DAUGHTER_INLAW
VR2P1	FATHER_MOTHER	MR2P1	5	FATHER_MOTHER
VR2P1	FATH_MOTH_INLAW	MR2P1	6	FATH_MOTH_INLAW
VR2P1	GRANDCHILD	MR2P1	7	GRANDCHILD
VR2P1	GREAT_GRANDCHILD	MR2P1	8	GREAT_GRANDCHILD
VR2P1	GRANDPARENT	MR2P1	9	GRANDPARENT
VR2P1	BROTHER_SISTER	MR2P1	10	BROTHER_SISTER
VR2P1	NEPHEW_NIECE	MR2P1	11	NEPHEW_NIECE
VR2P1	AUNT_UNCLE	MR2P1	12	AUNT_UNCLE
VR2P1	OTHER_REL	MR2P1	13	OTHER_REL
VR2P1	OTHER_NON_REL	MR2P1	14	OTHER_NON_REL
		MSEX	0	MISSING
VSEX	MALE	MSEX	1	MALE
VSEX	FEMALE	MSEX	2	FEMALE
		MZODIACAL	0	MISSING
VZODIACAL	MOUSE	MZODIACAL	1	MOUSE
VZODIACAL	COW	MZODIACAL	2	COW
VZODIACAL	TIGER	MZODIACAL	3	TIGER
VZODIACAL	RABBIT	MZODIACAL	4	RABBIT
VZODIACAL	DRAGON	MZODIACAL	5	DRAGON
VZODIACAL	SNAKE	MZODIACAL	6	SNAKE
VZODIACAL	HORSE	MZODIACAL	7	HORSE
VZODIACAL	SHEEP	MZODIACAL	8	SHEEP
VZODIACAL	MONKEY	MZODIACAL	9	MONKEY
VZODIACAL	HEN	MZODIACAL	10	HEN
VZODIACAL	DOG	MZODIACAL	11	DOG
VZODIACAL	PIG	MZODIACAL	12	PIG
		MEDU1	0	MISSING
VEDU1	NO_SHOOLING	MEDU1	1	NO_SHOOLING
VEDU1	ELEMENTARY_SHOOL	MEDU1	2	ELEMENTARY_SHOOL

VEDU1	MIDDLE_SHOOL	MEDU1	3	MIDDLE_SHOOL
VEDU1	HIGH_SHOOL	MEDU1	4	HIGH_SHOOL
VEDU1	JUNIOR_COLLEGE	MEDU1	5	JUNIOR_COLLEGE
VEDU1	UNIVERSITY	MEDU1	6	UNIVERSITY
VEDU1	MASTER	MEDU1	7	MASTER
VEDU1	DOCTOR	MEDU1	8	DOCTOR
		MEDU2	0	MISSING
VEDU2	GRADUATED	MEDU2	1	GRADUATED
VEDU2	ATTENDING	MEDU2	2	ATTENDING
VEDU2	COMPLETED	MEDU2	3	COMPLETED
VEDU2	ON_LEAVE	MEDU2	4	ON_LEAVE
VEDU2	DROPPED_OUT	MEDU2	5	DROPPED_OUT
VEDU2	NO_SHOOLING	MEDU2	9	NO_SHOOLING
		MRELIGEON1	0	MISSING
VRELIGEON1	YES	MRELIGEON1	1	YES
VRELIGEON1	NO	MRELIGEON1	2	NO
		MRELIGEON2	0	MISSING
VRELIGEON2	BUDDHISM	MRELIGEON2	1	BUDDHISM
VRELIGEON2	PROTESTANTISM	MRELIGEON2	2	PROTESTANTISM
VRELIGEON2	CATHOLICISM	MRELIGEON2	3	CATHOLICISM
VRELIGEON2	CONFUCIANISM	MRELIGEON2	4	CONFUCIANISM
VRELIGEON2	WONBUDDHISM	MRELIGEON2	5	WONBUDDHISM
VRELIGEON2	CHUNGSANGYO	MRELIGEON2	6	CHUNGSANGYO
VRELIGEON2	CHONDOGYO	MRELIGEON2	7	CHONDOGYO
VRELIGEON2	TAEJONGGYO	MRELIGEON2	8	TAEJONGGYO
VRELIGEON2	TAOISM	MRELIGEON2	9	TAOISM
VRELIGEON2	DAESUN	MRELIGEON2	10	DAESUN
VRELIGEON2	ISLAM	MRELIGEON2	11	ISLAM
VRELIGEON2	WORLDJUNGGYO	MRELIGEON2	12	WORLDJUNGGYO
VRELIGEON2	MIRDAEDO	MRELIGEON2	13	MIRDAEDO
VRELIGEON2	NO_RELIGEON	MRELIGEON2	90	NO_RELIGEON
VRELIGEON2	OTHER1	MRELIGEON2	98	OTHER1
VRELIGEON2	OTHER2	MRELIGEON2	99	OTHER2
		MMARST	0	MISSING
VMARST	NEVER_MARRIED	MMARST	1	NEVER_MARRIED
VMARST	MARRIED	MMARST	2	MARRIED
VMARST	WIDOWED	MMARST	3	WIDOWED
VMARST	DIVORCED	MMARST	4	DIVORCED
VMARST	UNDER15	MMARST	9	UNDER15

4) INTERV, NUM File

○ NUM File - 유효한 구간 변수 설정

○ INTERV File - 유효한 구간의 최소값, 최대값 설정

NUM		INTERV			
Validity set ID	Interval ID	Interval ID	Minimum	Maximum	Interval Step
VAGE	CHILD	CHILD	1	15	1
VAGE	ADULT	ADULT	16	121	1
VBIRTHYEAR	BIRTHYEARDOM2	BIRTHYEARDOM2	1887	2007	1
VMARYEAR	MARYEARDOMAIN	MARYEARDOMAIN	-1	-1	1
VMARYEAR	MARYEARDOMAIN2	MARYEARDOMAIN2	1903	2007	1
VCHILDB	CHILDBDOMAIN	CHILDBDOMAIN	-1	-1	1
VCHILDB	CHILDBDOMAIN2	CHILDBDOMAIN2	0	10	1
VCHILDG	CHILDGDOMAIN	CHILDGDOMAIN	-1	-1	1
VCHILDG	CHILDGDOMAIN2	CHILDGDOMAIN2	0	10	1

○ MARYEARDOMAIN의 유효한 구간은 -1로 남성, 미혼여성의 경우 missing값으로 UNIT에서 설정한 -1값을 유효한 값으로 설정

○ MARYEARDOMAIN2의 유효한 구간은 1903에서 2007로 이외의 값은 이상치로 Imputation 함

○ Interval Step - 최소값과 최대값 사이 값들의 구간

cf) Minimum : 1, Maximum : 15, Interval Step : 1인 경우 1,2,3....15

cf) Minimum : 1, Maximum : 15, Interval Step : 2인 경우 1,3,5....15

5) CLASS, CLCODE File

○ CLASS File - 분류 ID와 연관되는 Value Det ID 정의

Class ID	Value Set ID
EVER_MARRIED	MMARST
NOT_NOW_MARRIED	MMARST
MAJOR	MDISC
MINOR	MDISC

○ CLCODE File - 각 분류 ID에 포함되는 Label 정의

Class ID	Label
EVER_MARRIED	MARRIED
EVER_MARRIED	DIVORCED
EVER_MARRIED	WIDOWED
NOT_NOW_MARRIED	NEVER_MARRIED
NOT_NOW_MARRIED	WIDOWED
NOT_NOW_MARRIED	DIVORCED

1.3 SYSTEM PARAMETER FILE

Parameter name	Value	Default value	Possible value	비고
Number of sub-units	4	0	x ≥ -1	4 : 4인가구 -1 : 여러가구를 한꺼번에 할 경우
EDITING				
Use of DLT	1	1	0,1	0(NO DLT used), 1(DLT used)
Full Edit	1	0	0,1	0(첫번째 에디팅 규칙 실패시 멈춤) 1(모든 실패한 에디팅 규칙 검사)
Secondary Edit Method	1	1	1,2	Primary edit rules - Consistency DLTs Secondary edit rules - Donor Selection DLTs 1 - both 2 - only Primary edit rules
remove redundant rules		0	0,1	0(No Check), 1(Checks and removes redundant rules)
REORDERING				
ordering method	1	1	1-3	1(Consider Original Ordering only) 2(Consider All Orderings) 3(the best sub-unit combination)
reordering weight	0.0	0	x ≥ 0	
max/min nonresponse ratio	2.0	2	x ≥ 1	
maximum nonresponse	25%	25%	0.0~100.0	
non-failing weight	1.0	1	0~9	
failing weight	1.0	1	0~9	
reorder improvement	0%	0%	0.0~100.0	
near-minimum distance	0%	0%	0.0~100.0	
STAGE CONTROL				
nb of 1st stage donors	500	500	x ≥ 0	첫 번째 단계의 도너수
nb of stage donors mult	2.0	2.0	x ≥ 1.0	두 번째 단계 이후부터의 단계별 배수
donor search id	1	1	1-5	1 - Ripple Search 2 - Bounded Ripple Search 3 - Modified Ripple Search 4 - Forward Search 5 - Backward Search
nb max of donors	20	10000	x > 0	단계별 최근접 최대 도너수
donor reuse freq	200	100	x > 0	도너 재사용 횟수
Dfp max 1st stage	-1	-1.0	-1, x > 0.0	첫 번째 단계의 Dfp 최대값
Dfpa max 1st stage	-1	-1.0	-1, x > 0.0	첫 번째 단계의 Dfpa 최대값
nb max of stages	5	10	x ≥ 1	단계 최대수
Dfpa improv ratio	0.10	0.10	0.0~1.0	
donor usage		1	1-3	1(only passed the rugular edits) 2(not fail the donor selection edits) 3(all units)
DISTANCE CALCULATIONS				
Dfpa fct id	1	1	1	
Dfpa fct param	0.9	0.75	0.5~1.0	Dfpa 계산시 α값
distance function exponent	1.0	1	x > 0	
NMCIA LIST				
nb max of nmcias	5	10	x > 0	NMCIA 상의 리스트 최대수
Dfpa max fct id	1	1	1	

Dfpa max fct param	1.1	1.1	x ≥ 1.0	Dfpa 최대값 파라메터
nmcia sel method id	1	1	1,2	1 - 임의 확률 2 - 가장 낮은 Dfpa
nmcia sel method param	1.0	1.0	x ≥ 0.0	도너 선택방법 1일때 파라메터 t값
Undecided Var Count Case Props		4	x ≥ 0	
Outliers				
Outlier detection		0	0,1,2	1(No outlier detection) 2(at the srtatum level) 3(at the GEO ID level)
minimum nb of records		20	x > 0	
exclude values		0	0,1,2	0 1(negative value) 2(negative and zero)
INPUT FILES				
geographic id		0	0,1	0(No supplied), 1(supplied)
probability of selection		0	0,1	0(No supplied), 1(supplied)
OUTPUT FILES				
Log file	1	0	0,1	0(short version), 1(long version)
Donor file	1	0	0,1	0(Not created), 1(Created)
NMCIA file	1	0	x ≥ 0	
IAEVAL file	1	0	0,1	0(Not created), 1(Created)
IASTAT file	1	0	0,1	0(Not created), 1(Created)
Adj Dfp file	1	0 0	x ≥ 0, y ≥ 0	
Unified DLT file	1	0	0,1	0(Not created), 1(Created)
failimp file		0	0,1	0(Not created), 1(Created)
Output failed imputes		0	0,1	
AUDIT TRAIL				
number of records	100%			
Audsel file	0			
stage param	1	0	0,1	0(Not printed), 1(Printed)
best Dfp's	1	0	0,1	0(Not printed), 1(Printed)
donor	1	0	0,1	0(Not printed), 1(Printed)
imputed	1			
generating list	1	0	0,1	0(Not printed), 1(Printed)
NMCIA list	1	0	0,1	0(Not printed), 1(Printed)
distance dfpi	1	0	0,1	0(Not printed), 1(Printed)
simplified DLT	1	0	0,1, y ≥ 0	
GENERAL				
Rounding Error Factor		10.0	x ≥ 0	

1.4 HOT_DECK PARAMETER FILE

1) IMPPARAM

○ 변수별 Imputation 여부와 도너와의 거리계산시 거리계산함수와 가중치 값에 대한 정보

Variable ID	Imputability	Weight	Distance funcion ID
R2P1	I	1	1
SEX	I	1	1
AGE	I	1	3
ZODIACAL	I	0.5	1
BIRTHYEAR	I	0.5	3
EDU1	I	1	1
EDU2	I	1	1
RELIGEON1	I	0.5	1
RELIGEON2	I	0.5	1
MARST	I	1	1
MARYEAR	I	0.5	3
CHILDB	I	0.5	3
CHILDG	I	0.5	3

○ Imputability - N(Non-imputable), I(Imputable)

○ Weight - 변수의 중요도에 따라 가중치 부여

○ Distance function

Distance function ID	Used for	Parameters(default)	비고
1	정수, 코드	-	일치(0), 불일치(1)
2	정수, 실수	r(1), k(121)	$1-(\frac{min(R_{fi,}R_{di})+1}{max(R_{fi,}R_{di})+1})^r$ if $\min(R_{fi,}R_{di})\geq 0$, $\max(R_{fi,}R_{di})$<k 1 otherwise
3	실수	r(0.25), k1(6), k2(2), k3(30), k4(15), k5(1), k6(1)	R_{fi}<k4 & $R_{di}\geq$k4 then D=k5 R_{di}<k4 & $R_{fi}\geq$k4 then D=k6 $\|R_{fi}-R_{di}\|\geq$x then D=1 otherwise 1-$(1-\|R_{fi}-R_{di}\|/x)^r$ x = k1 +k2(R_{fi}-k3)/10 if $r_{fi}\geq$k3
4	코드, 정수	-	Distmat file
5	코드	k1(0.008), k2(0.004), k3(0.002), k4(0)	k1(첫글자 불일치) k2(첫글자 일치,두번째 불일치) k3(첫번째, 두 번째 일치, 세 번째 불일치) k4(마지막 3글자 모두 불일치)
6	코드	k1(0.008), k2(0.008), k3(0.004), k4(0.002), k5(0)	마지막 4글자
7	정수	r(0.25),k1(2),k2(20),k3(4),k4(35), k5(1.5), k6(11), k7(15), k8(1), k9(1), k10(70), k11(0), k12(30)	
8	정수와 코드	group id, order, β(0.8), k1(1), k2(7/15),k3(3/15),k4(1/15),k5(0)	
9	두 실수	group id, order, n(40)	
10	실수	u percentile	
11	실수	u percentile	
12	문자	t(0.8), a(1), r(0.2435)	
13	문자	t(0.8), a(1)	

2) IMP

○ IMPPARAM File에서 정의한 임퓨테이션 여부를 특정 서브유니트에서 다르게 하고자 하는 경우

Variable ID	Sub-Unit Number	Imputability
R2P1	1	I
R2P1	2	I
R2P1	3	I
R2P1	4	I
SEX	1	I
SEX	2	I
SEX	3	I
SEX	4	I
AGE	1	I
AGE	2	I
AGE	3	I
AGE	4	I
ZODIACAL	1	I
ZODIACAL	2	I
ZODIACAL	3	I
ZODIACAL	4	I
BIRTHYEAR	1	I
BIRTHYEAR	2	I
BIRTHYEAR	3	I
BIRTHYEAR	4	I
EDU1	1	I
EDU1	2	I
EDU1	3	I
EDU1	4	I
EDU2	1	I
EDU2	2	I
EDU2	3	I
EDU2	4	I
RELIGEON1	1	I
…	…	…
CHILDG	4	I

3) PERMU

○ 임퓨테이션하여야 할 서브유니트에 대한 재정렬 여부

Sub-Unit Number	Pemutability
2	P
3	P
4	P

○ Pemutability : F(Fixed), P(Permutable)

1.5 DLT FILE

○ DLT - Decision Logic Table로 도너선정 또는 임퓨테이션 대상선정

○ 캐나다 센서스 DLT를 참고하여 2005 인구주택총조사 내검규칙을 적용하여 53개 DLT 작성

1) Hot-Deck DLT

□ Comments - *로 시작

○ DLT의 내용, 작성자, 작성일시 등의 정보를 기록

□ Parameter Fields - %로 시작

○ DLT Name - DLT의 이름

○ Strata - 층수(가구원수)로 해당 strata만 DLT 적용됨

○ Purpose

- Donor selection - 조건문에 해당하면 도너에서 제외
- Consistency - 조건문에 해당하면 임퓨테이션 대상에 해당

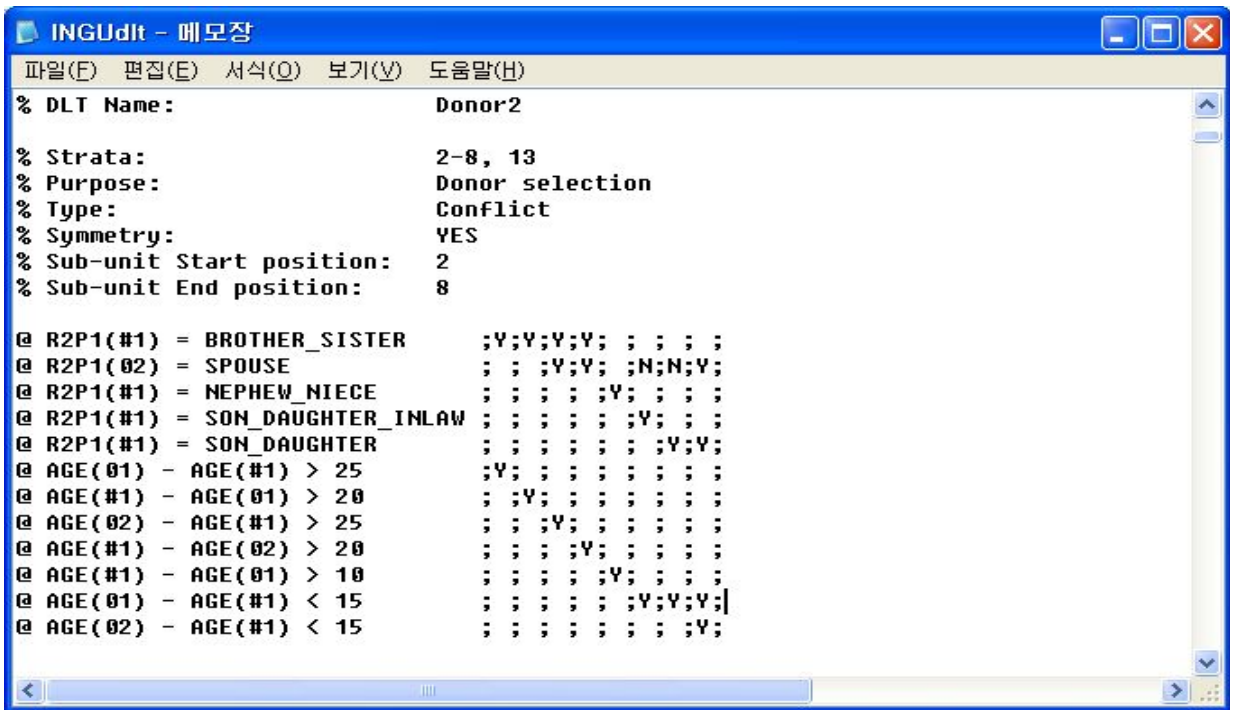
○ Symmetry - 대칭유무로 Yes이면 #1, #2이고 strata가 3인 경우 (1,2),(1,3),(2,3)이고 No이면 (1,2),(1,3),(2,3),(2,1),(3,1),(3,2)로 확장적용됨

○ Sub-unit Start position - 가구원의 시작값으로 #1의 시작값

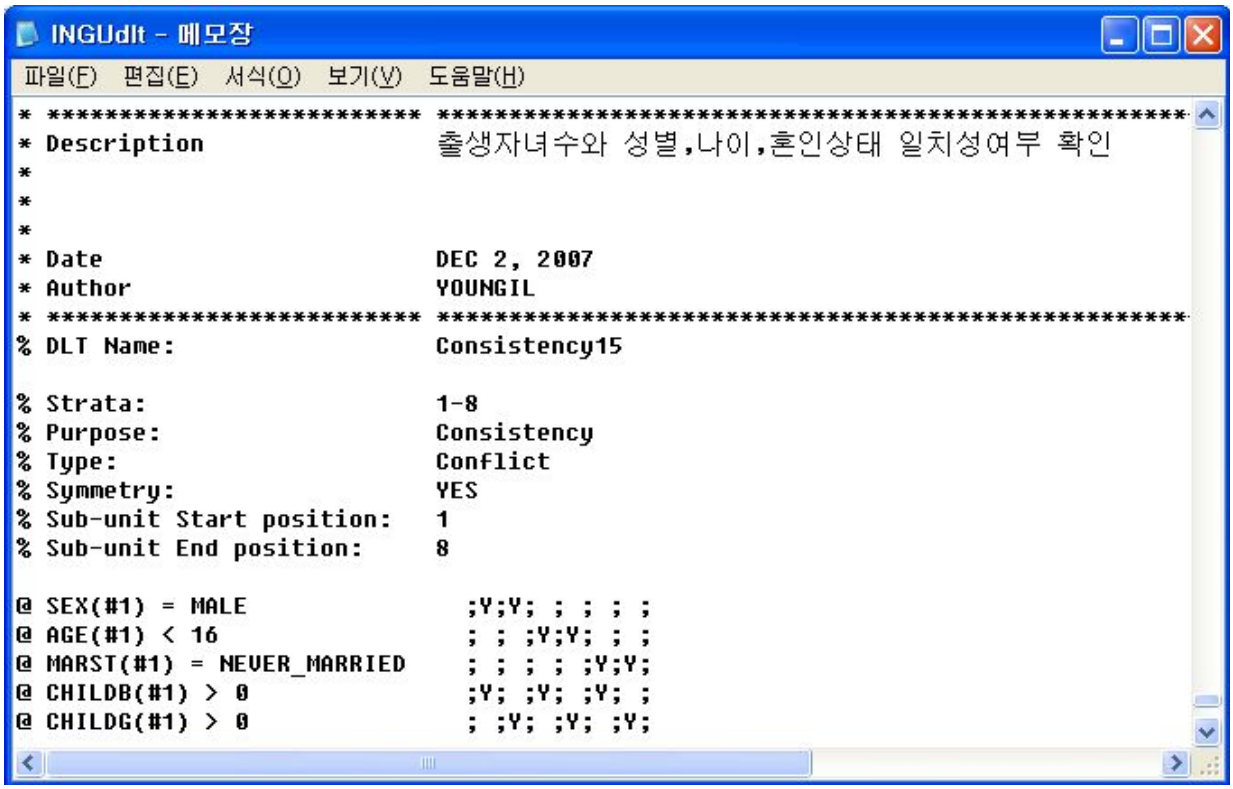
○ Sub-unit end position - 가구원의 마지막값으로 #1의 마지막값

□ Conditions - @로 시작

○ 아래 DLT의 경우 2~8번 가구원중 가구주와의 관계(R2P1)가 형제자매이고 가구주의 나이보다 25세이상 어린 경우 도너에서 제외



○ 아래 DLT의 경우 1~8번 가구원의 성별(SEX)이 남성이고 남자자녀(CHILDB)가 있는 경우 임퓨테이션 대상임



2) Derive DLT

○ Comments, Parameter Fields, Conditions는 Hot-Deck DLT와 동일

○ Purpose는 반드시 Derive로 설정하여야 함

□ Common actions - $로 시작

○ 임시변수 설정(DECL문), 변수값 설정

□ Conditional actions - &로 시작

○ 조건 실행문으로 x로 표시

3) DLT file 구성

○ Hot-Deck DLT와 Derive DLT가 모두 있는 경우 Input file과 실행파일을 따로 작성하고 Derive DLT의 output file의 결과자료를 Hot-Deck DLT의 input file로 사용하여 실행

○ Derive DLT의 Input file은 INGUD1*로, Hot-Deck DLT의 Input file은 INGUI1*로 설정

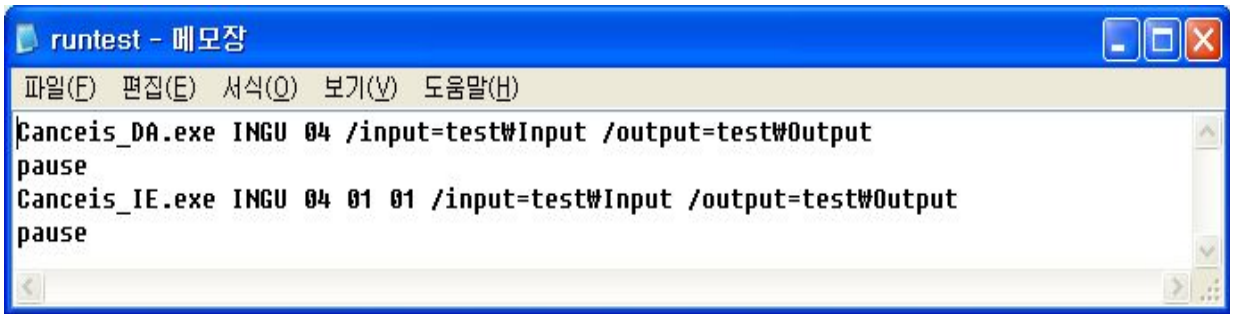
2. Imputation

2.1 실행파일 RUNTEST.BAT

○ Job Mode로 실행하였고 Hot-deck DLT만 적용하였음

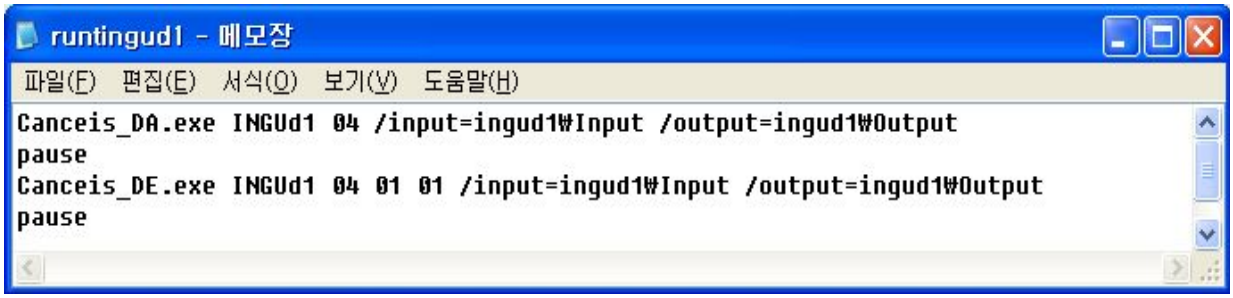
Analyser Executable Name	Data Type ID	Stratum ID	/INPUT=	/OUTPUT
Canceis_DA.exe	INGU	04		

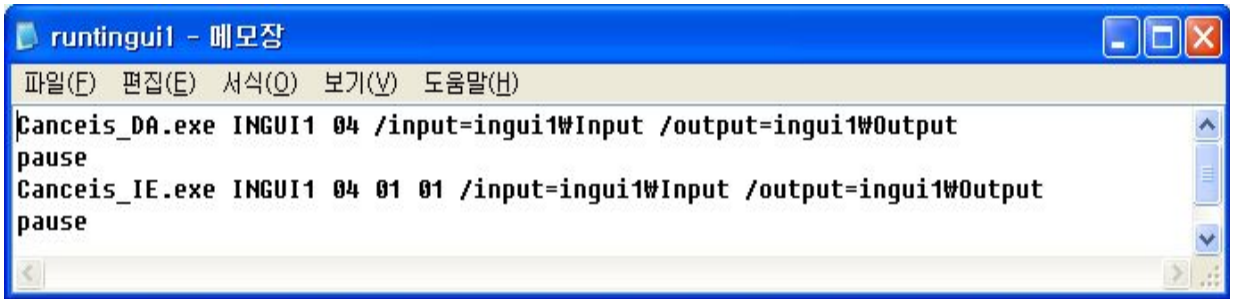
Engine Executable Name	Data Type ID	Stratum ID	Imputation Group ID	Job Mode ID	/INPUT=	/OUTPUT
Canceis_IE.exe	INGU	04	01	01		



○ Derive DLT와 Hot-deck DLT가 모두 있는 경우 실행파일을 별도로 만듬

- Derive DLT는 DA(DLT Analyzer)와 DE(Derive Engine)을 사용
- Hot-deck DLT는 DA(DLT Analyzer)와 IE(Imputation Engine)을 사용





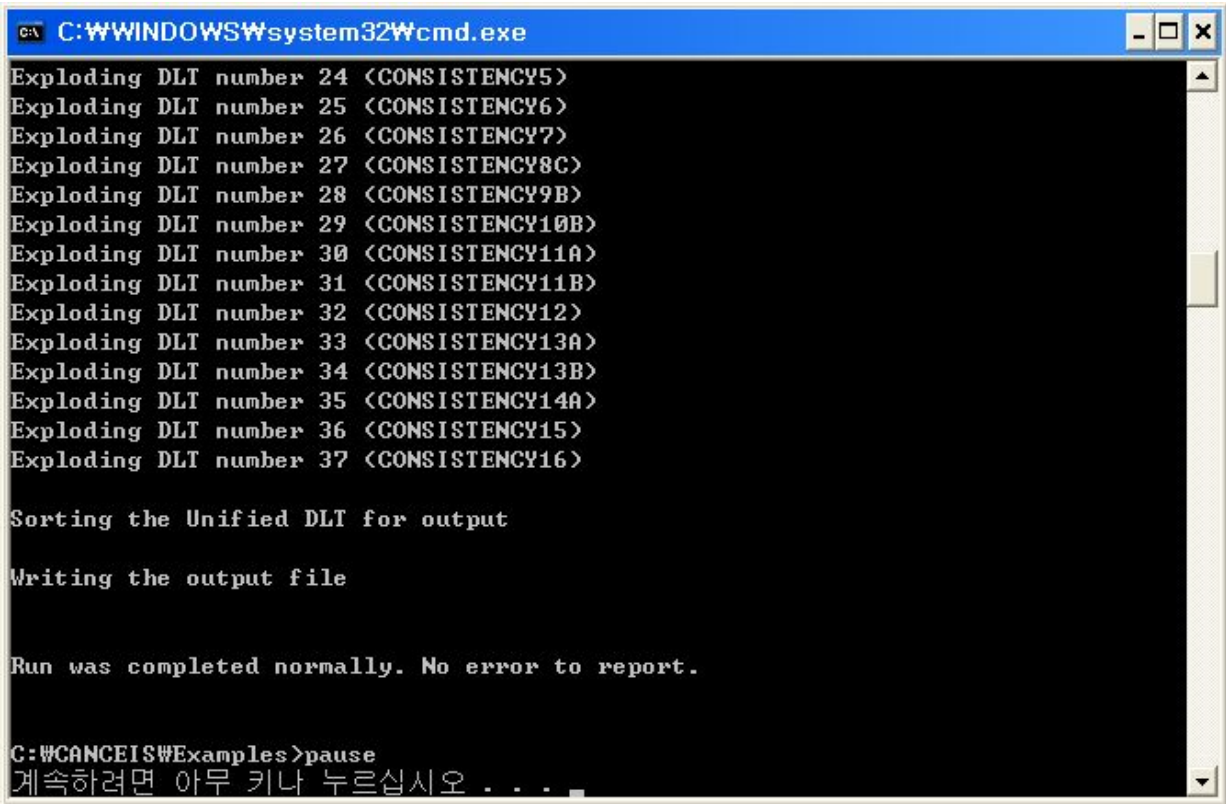
2.2 DLT Analyzer

○ Loading Parameters, Dictionaries and DLTs

○ Performing Text Substitution

○ Checking DLT Syntax

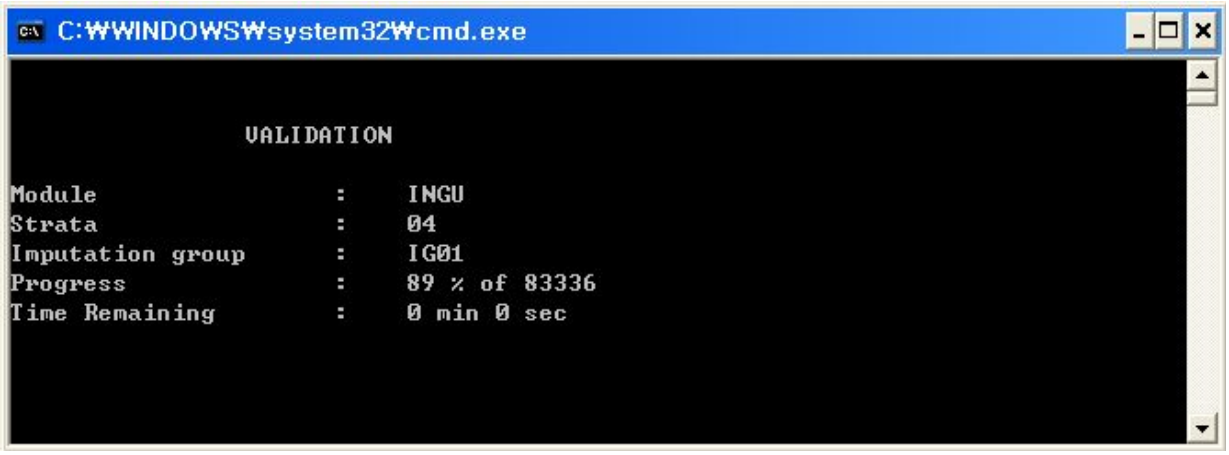
○ Explosion of generic Propositions



3. Imputation Engine

○ Valudation - 유효성 검사

- 83,336레코드에 대하여 유효하지 않은 값에 대한 검사

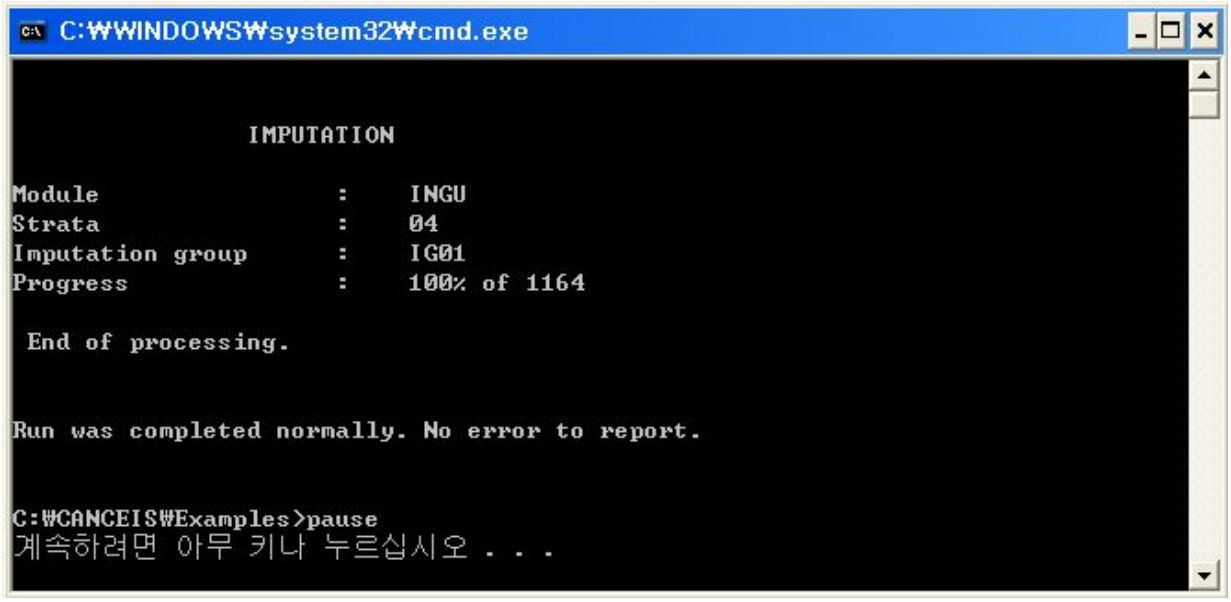


○ Editing - 에디팅

- 83,336레코드에 대한 DLT상의 도너선택과 불일치하는 임퓨테이션 대상 선정

○ Imputation

- 임퓨테이션 대상 1,164가구에 대한 임퓨테이션 현황

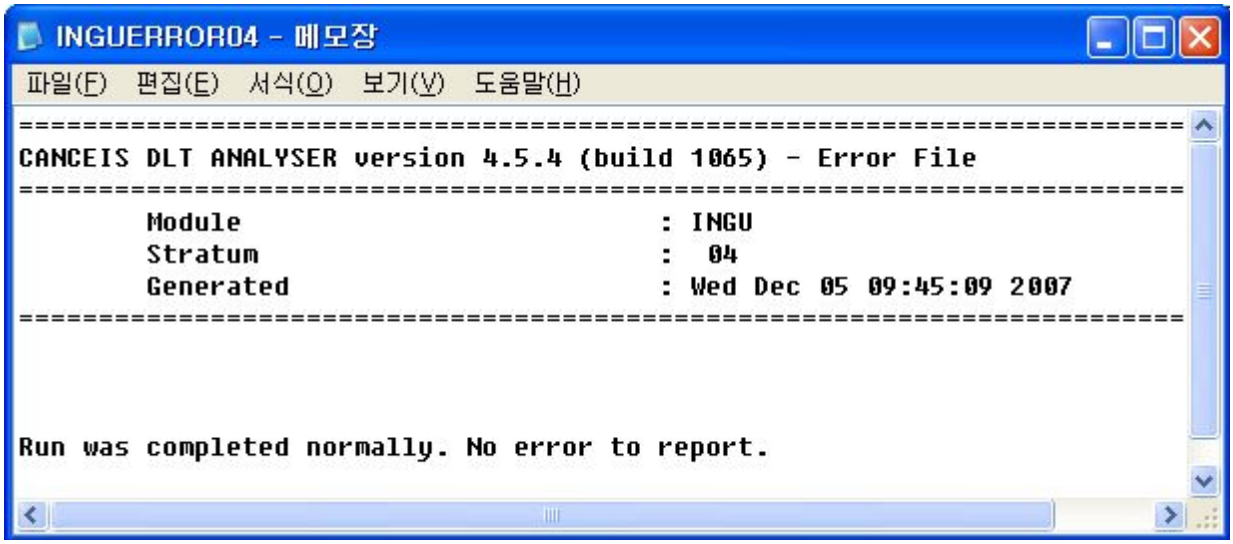


3. OUTPUT FILE

3.1 DLT ANALYZER

1) ERROR

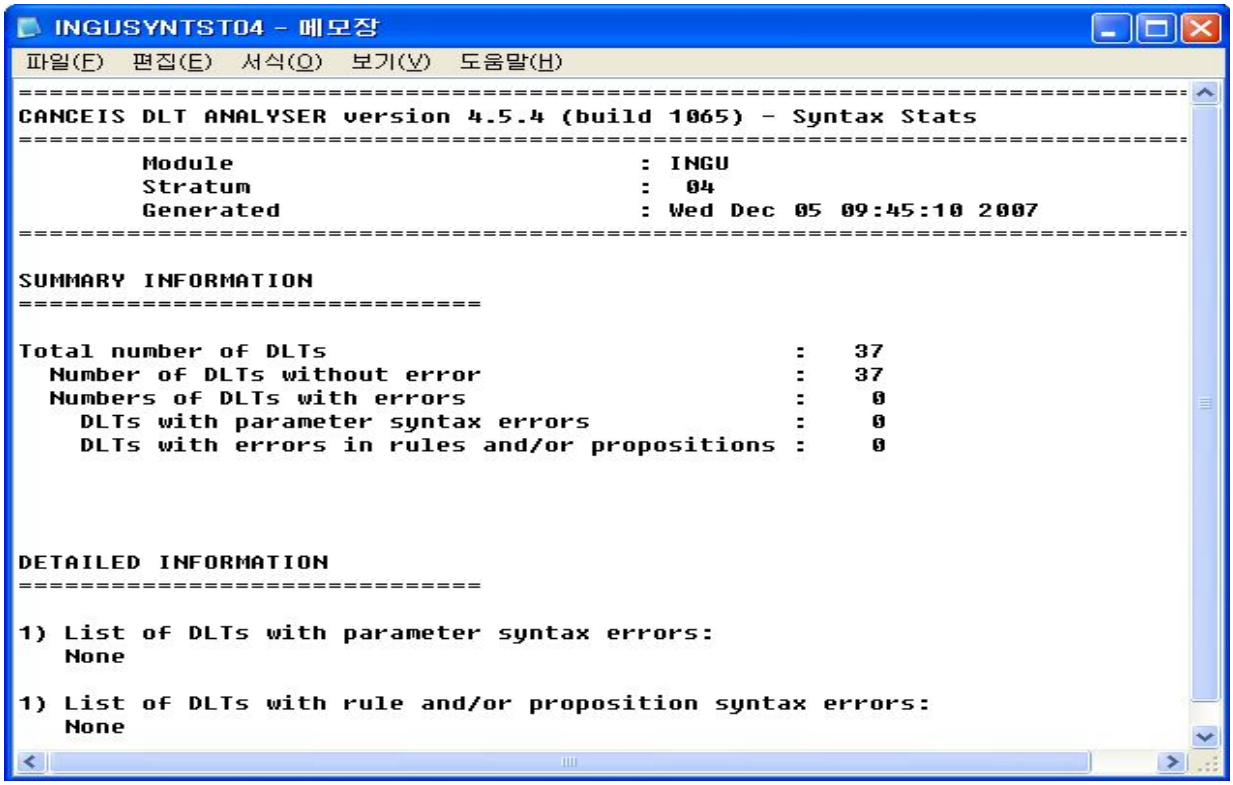
○ DLT Analyser를 실행하고 에러 및 경고를 나타냄



2) SYNTST

○ DLT 구문 통계

- 37개 DLT중 에러가 있는 DLT는 없는 것으로 나타남



3) DLTSUB

○ 53개 DLT중 TXTSUBST결과 droped DLT 16개를 제외한 37개에 대한 DLT로 Input 디렉토리에 생김

INGUDLTSUB04 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
* ************************** ********************************************************************************
* DESCRIPTION                SECONDARY EDITS. WIDOWED  AND AGE <  24+
*                            SEE COMMENT
*
*                            MATCHES CONCEPT S1
* DATE                       JUNE 16, 2005
* AUTHOR                     PATRICK MASON AND SYLVIE GUILMETTE, UPDATED FROM 1996 DLT
* ************************** ********************************************************************************
% DLT NAME:                  DONOR1

% STRATA:                    1-8, 9, 12-13, 14, 23
% PURPOSE:                   DONOR SELECTION
% TYPE:                      CONFLICT
% SYMMETRY:                  YES
% SUB-UNIT START POSITION:   1
% SUB-UNIT END POSITION:     8

@ MARST(#1) = WIDOWED ;Y;
@ AGE(#1) < 24        ;Y;
```

4) EXPLST

○ DLT중 확장모드(#1)에 대한 통계로 37개 DLT 모두 확장됨

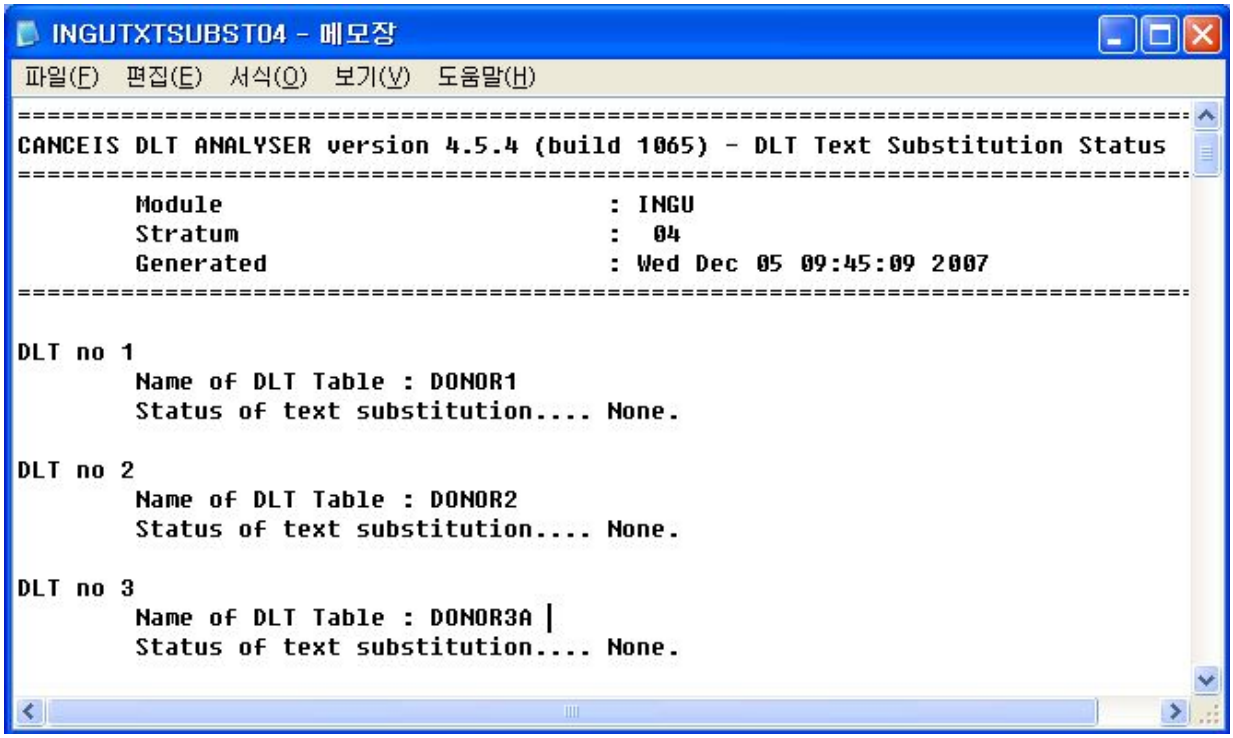
INGUEXPLST04 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

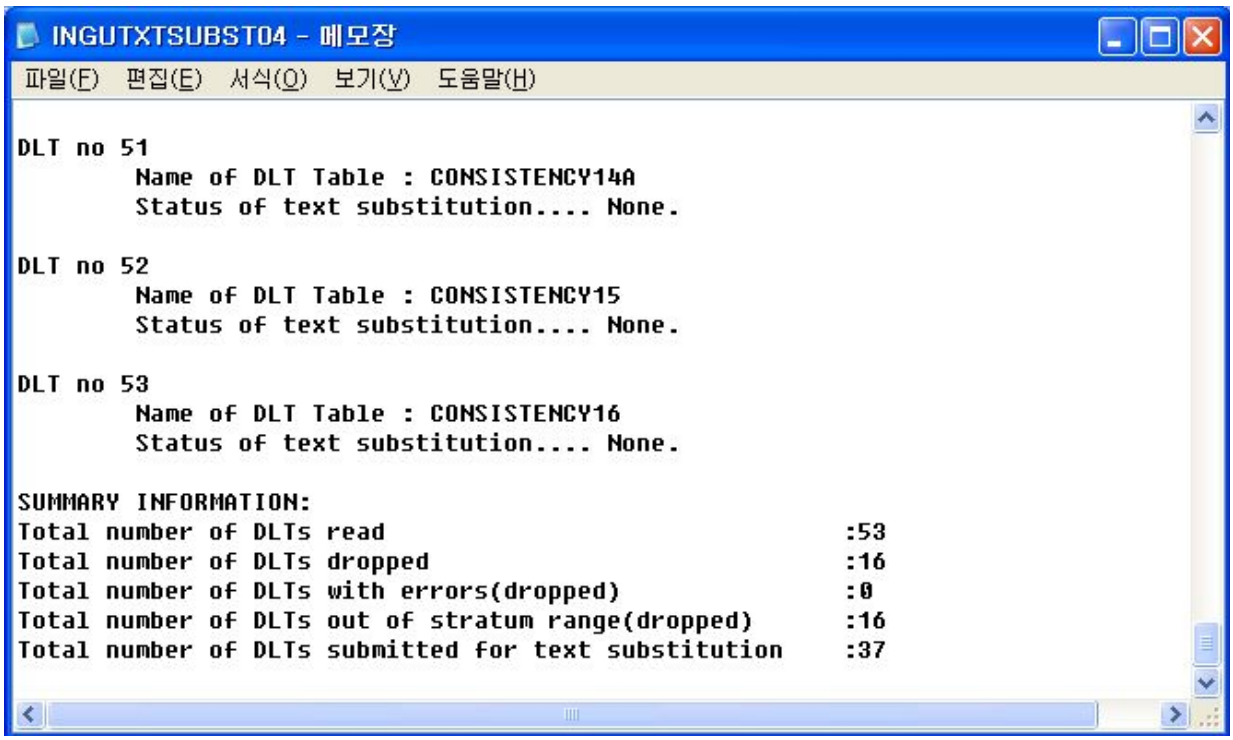
```
================================================================================
CANCEIS DLT ANALYSER version 4.5.4 (build 1065) - DLT Status
================================================================================
         Module                              : INGU
         Stratum                             :  04
         Generated                           : Wed Dec 05 09:45:10 2007
================================================================================

SUMMARY INFORMATION:
Total number of DLTs with errors remaining  :   0
Total number of DLTs dropped                :   0
Total number of DLTs submitted to explosion :  37
--------------------------------------------------------------------------------
Total number of DLTs processed              :  37


DETAILED INFORMATION:
=========================
1) List of DLTs with errors remaining:
   None

2) List of DLTs dropped:
   None

3) List of DLTs submitted to explosion:

 DLT#           DLT Name
--------------------------
    1           DONOR1
    2           DONOR2
    3           DONOR3A
    4           DONOR3B
    5           DONOR3C
    6           DONOR3D
```

5) TXTSUBST

○ DLT에 대한 Text Substitution 상태를 나타냄



○ 53개 DLT중 stratum 범위를 벗어난 16개를 제외한 37개가 Text Substitution됨

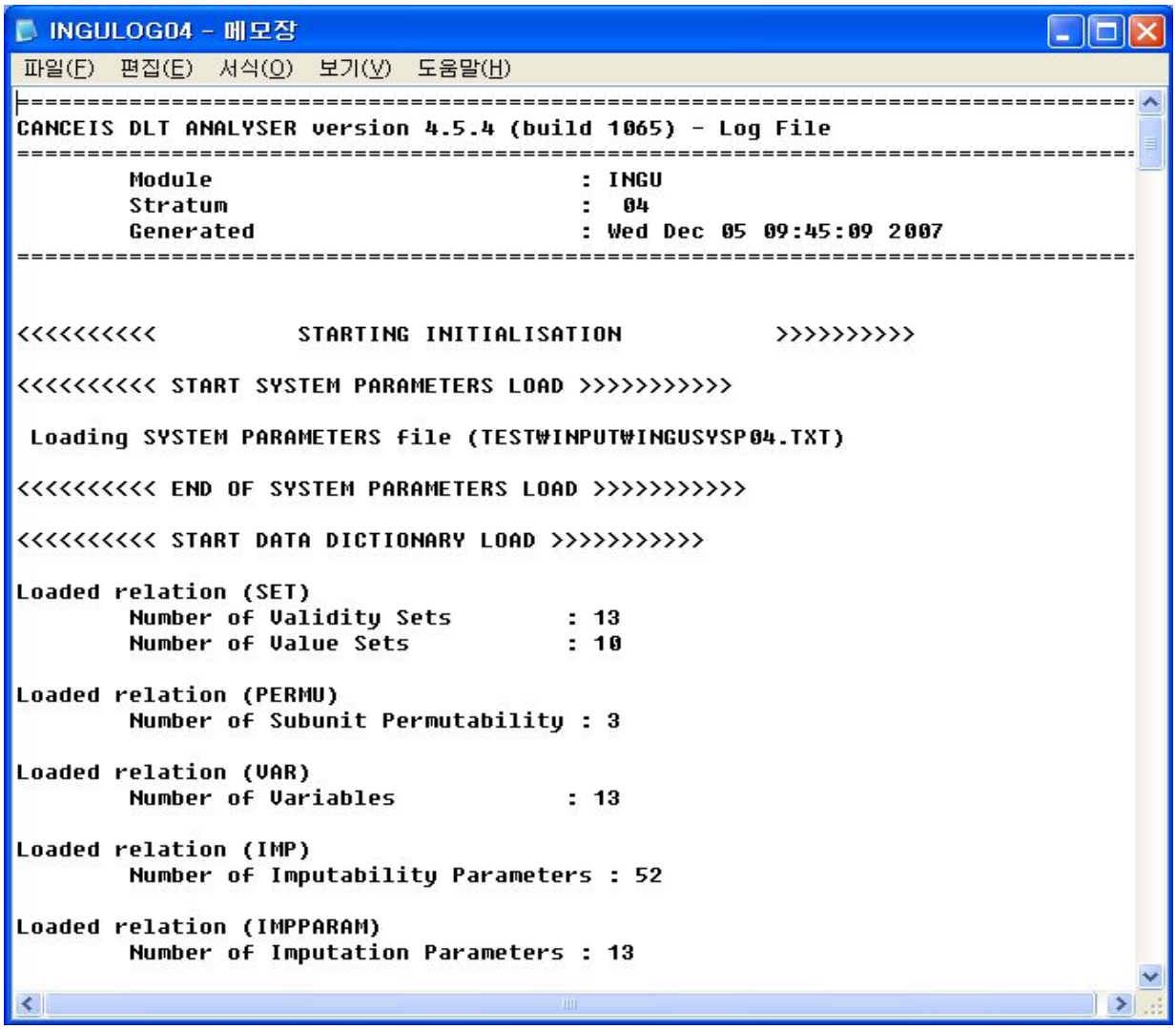


6) LOG

○ DLT Analyzer 실행에 대한 각 단계별 상태를 나타냄

- System Parameters Load
- Data Dictionary Load
- DLT Processing : Text Substitution, DLT Syntax Analyser, DLT Explosion
- 최대 메모리 사용량과 리턴 코드

Return code	내용	Return code	내용
0	에러 및 경고 없음	9	빈 파일
1	경고 있음	10	명령라인 에러
2	파일 누락	16	잘못된 수행
5	메모리 문제	32	로직 에러
6	시스템 파라메터 문제	34	데이터베이스 에러
7	레코드	40	주소창 확장 에러
8	파일 포맷 에러		



7) MASTERDLT

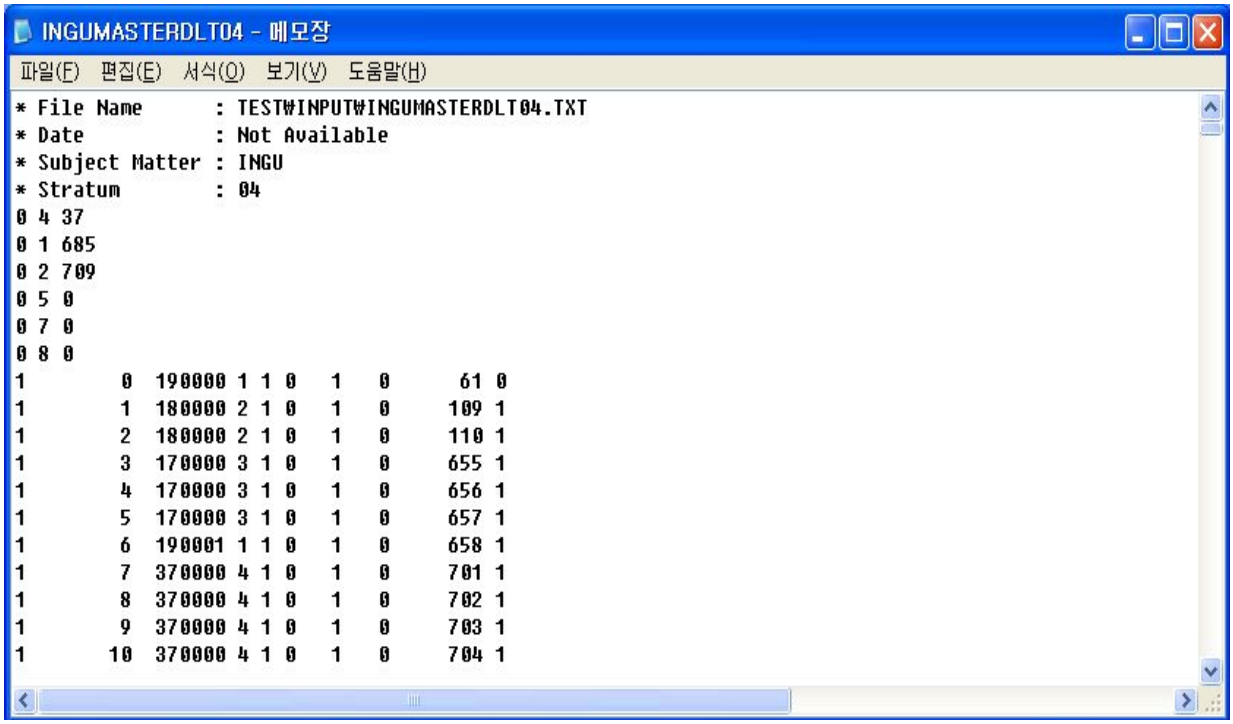
○ 통합된 DLT에 대한 정보로 6개의 Section으로 구성됨

○ Section0 - Counters

- DLT 수 : 37개
- Rule 수 : 685개
- Proposition 수 : 709개

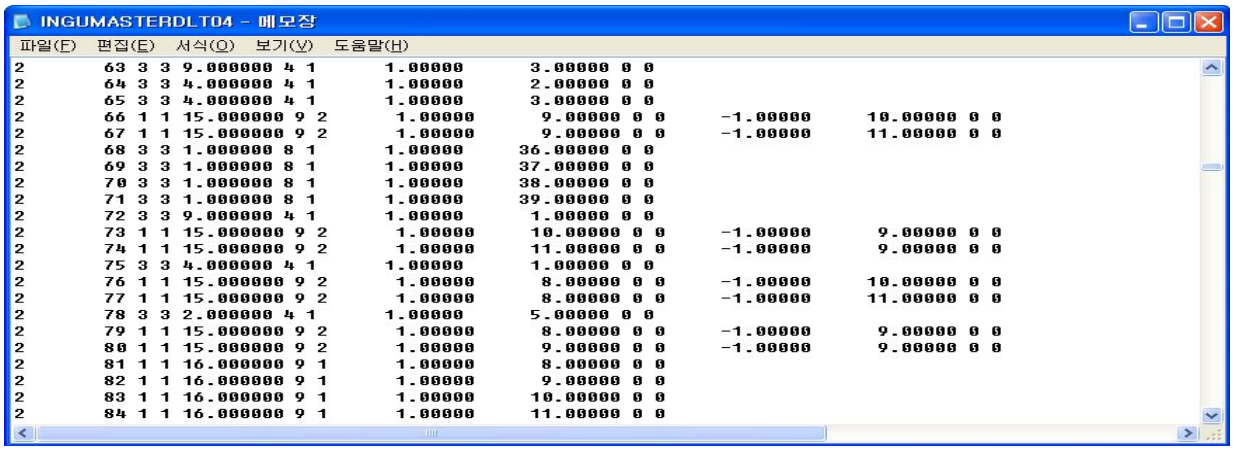
○ Section1 - Rule Descriptions

#	Field	Possible Value
0	Section ID	1
1	Rule ID	
2	Initial Rule ID	6자리중 앞 2자리는 DLT ID
3	Next Rule ID	
4	Type of DLT	1(Consistency), 2(Donor selection)
5	Number of Propositions Linked	
6	Number of Actions Linked	
7	ID of 1 Linked Proposition	
8	Yes/No flag of 1 Linked Proposition	1(Yes), 0(No)
n	ID of n linked Action	



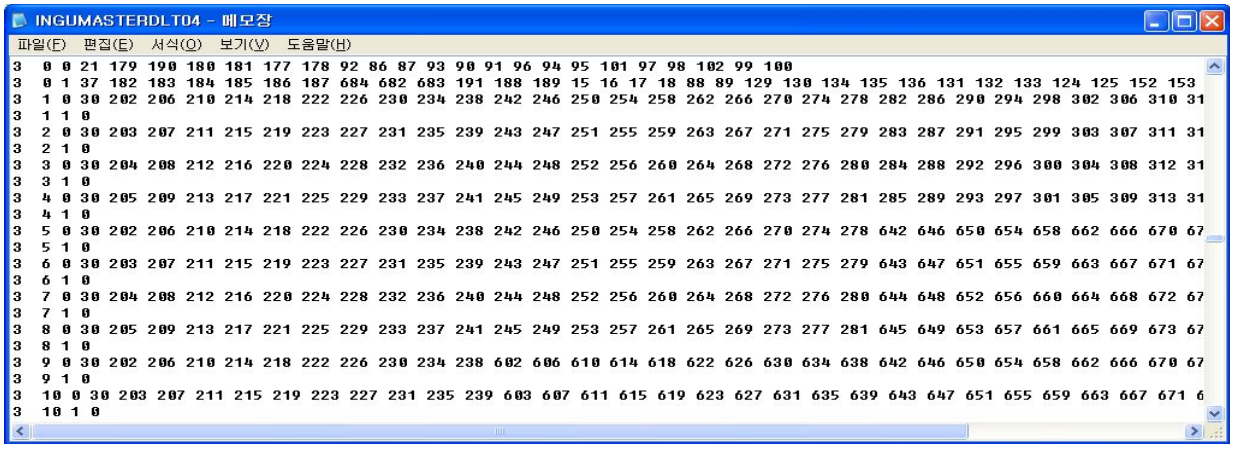
○ Section2 - Proposition Descriptions

#	Field	Possible Value
0	Section ID	2
1	Proposition Identifier	
2	Proposition Family type	1(정수), 2(실수), 3(코드), 4(문자)
3	Logical Operator	1(<), 2(=), 3(>)
4	Constant	
5	Nature of the Constant	1(실수), 2(정수), 3(문자), 4(코드), 5(실수 분류), 6(정수 분류), 7(문자 분류), 8(코드 분류), 9(숫자)
6	Number of Actions Linked	
7	Coefficient associated to the Proposition first answer box	
8	Identifier of the Proposition first answer box	



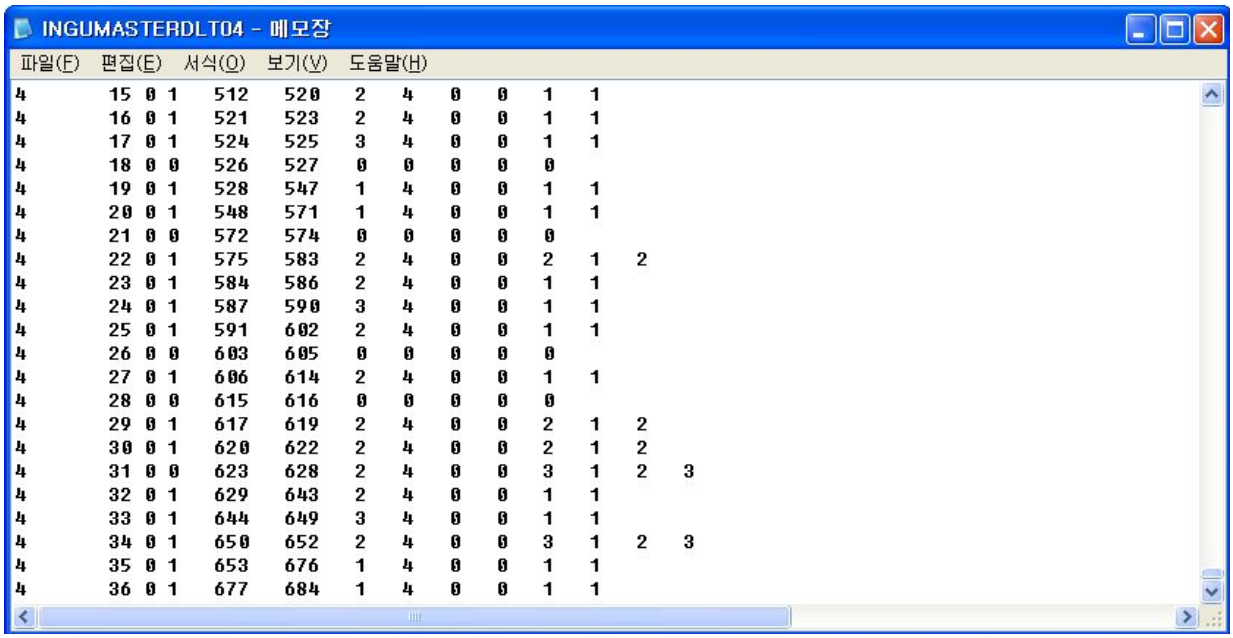
○ Section3 - Proposition-Rule linkage

#	Field	Possible Value
0	Section ID	3
1	Proposition Identifier	
2	Proposition Binary Entry	1(True), 2(False)
3	Number of Rules in which the Proposition appears in	
4	Identifier of the first rule	



○ Section4 - DLT descriptions

#	Field	Possible Value
0	Section ID	4
1	First Rule	
2	Last Rule	
3	Else flag	0(True), 1(False)
4	Overlaps	0(False), 1(True)
5	Looping	0(False), 1(True)
6	Number of variable position	
7	Order of the first Variable Position sub-unit	
m	Sub-unit start of the first Variable Position sub-unit	
m+n	Sub-unit end of the first Variable Position sub-unit	
m+n+o	Sub-unit end of the last Variable Position sub-unit	

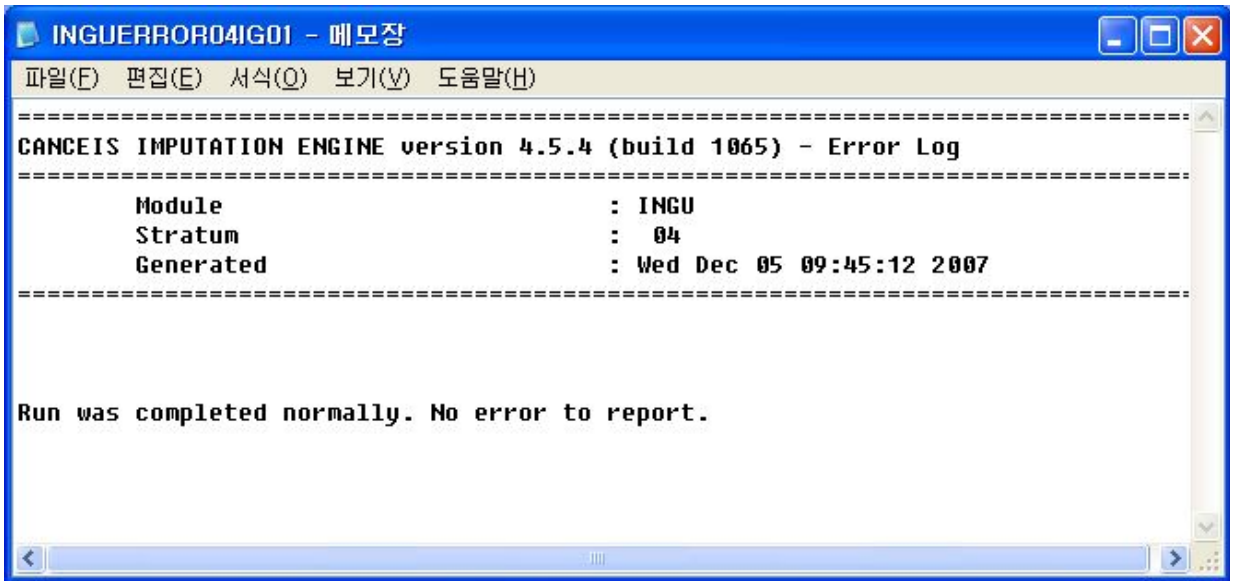


3.2 IMPUTATION ENGINE

1) Status Files

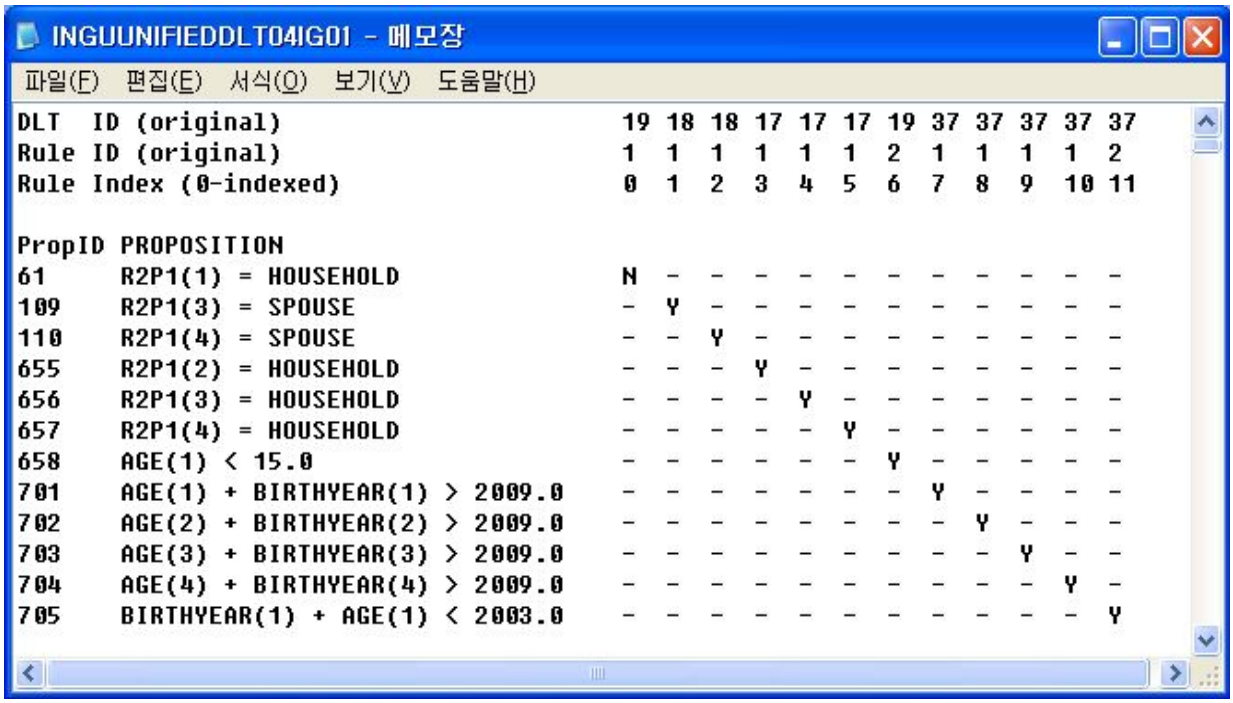
□ ERROR

○ Imputation Engine을 실행하고 에러 및 경고를 나타냄



□ UNIFIEDDLT

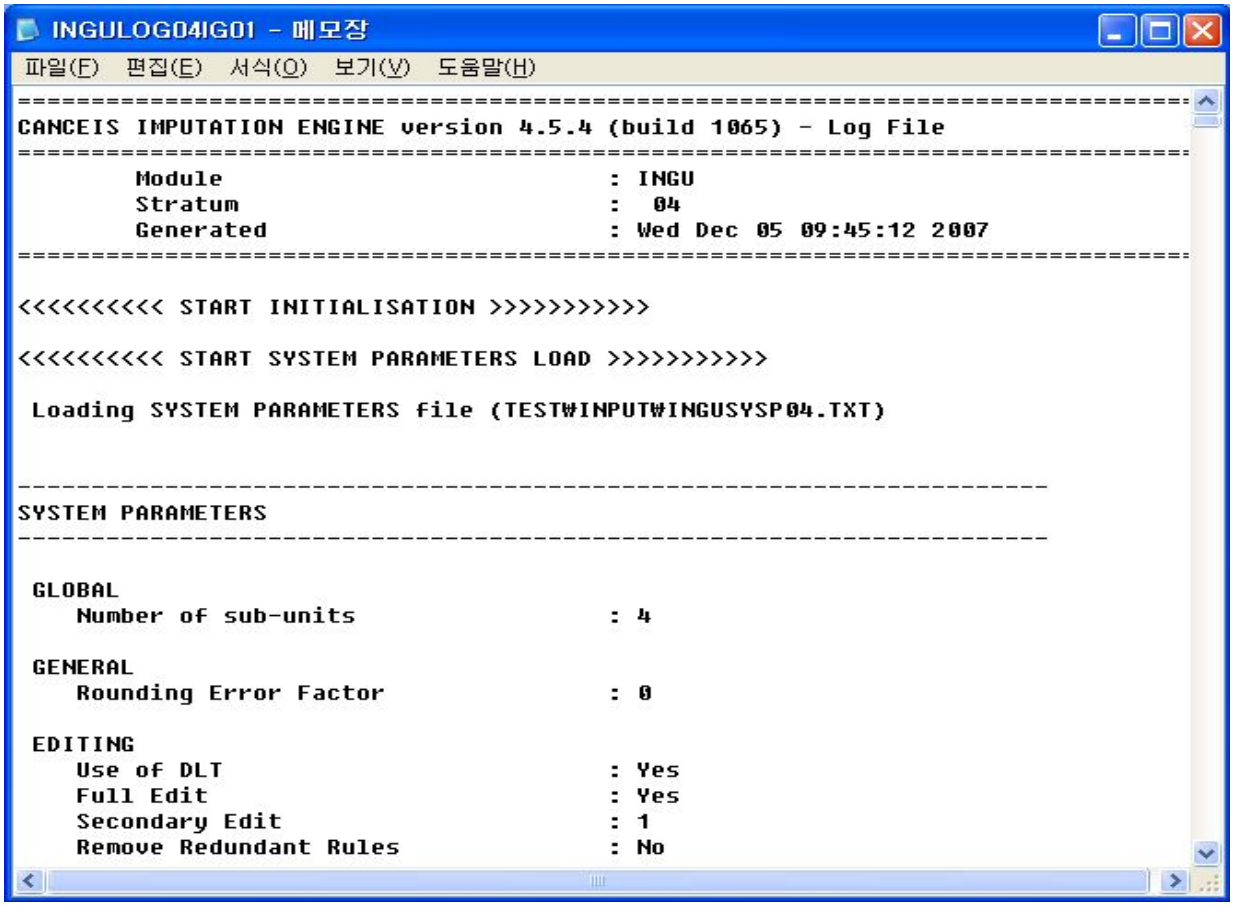
○ 에디트 규칙이 임퓨테이션과정에서 어떻게 사용되어지는지를 보여줌



□ LOG

○ Imputation Engine 실행에 대한 각 단계별 상태를 나타냄

- System Parameters Load
- Data Dictionary Load
- Load and validation of processing units
  - · Number of records found was 333344
  - · Number of punits found was 83336
- Edit of processing units
- 최대 메모리 사용량과 리턴 코드
  - · Number of fail records selected for the audit trail file : 1164
  - · Total running time of program: 00h:02m:22s
  - · Stratum: 04 Return Code: 0
  - · Used memory remaining : 72891 kb
  - · Maximum memory used : 120083 kb



2) Statistical reports

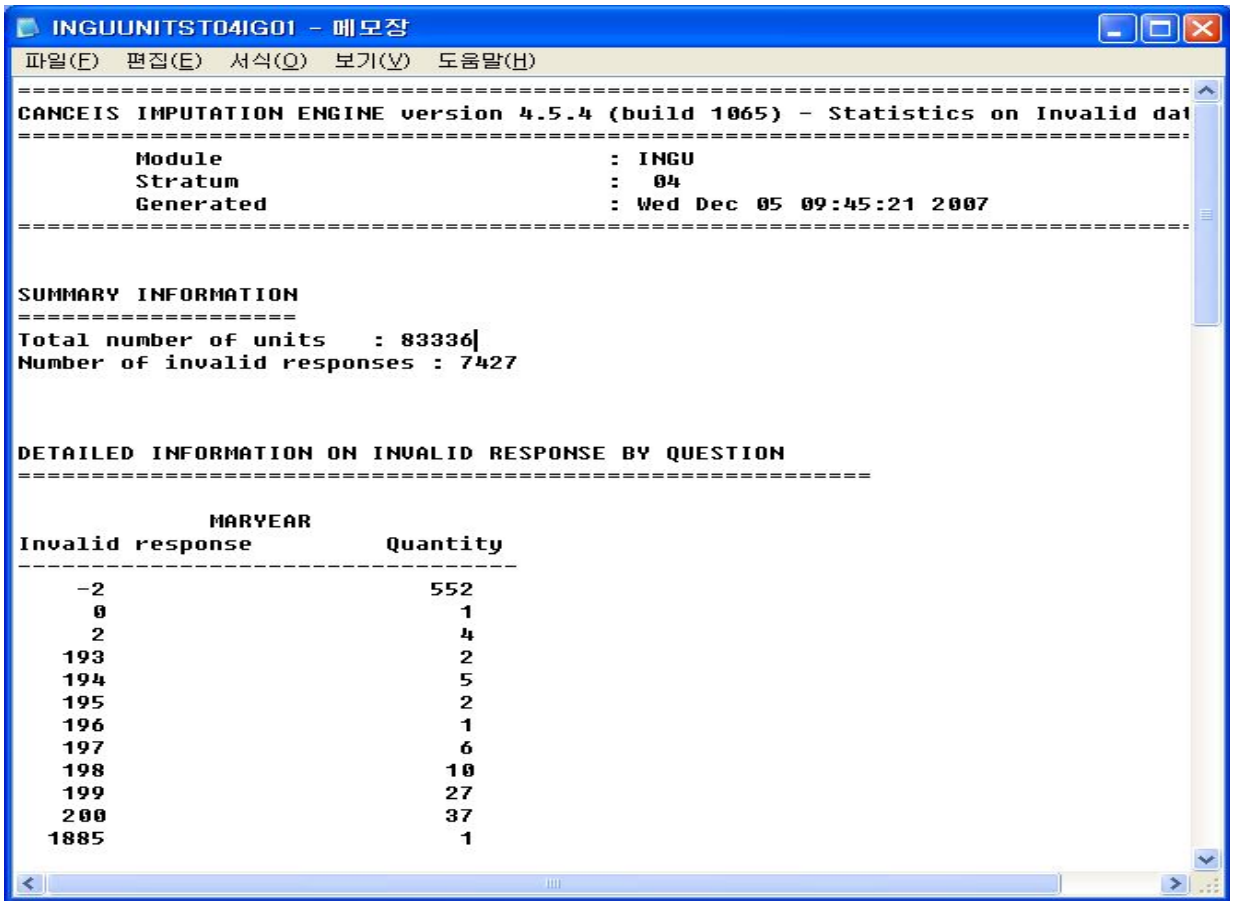
□ UNITST

○ 유효하지 않은 자료에 대한 통계

○ Interv File에서 설정한 범위를 벗어나는 자료에 대한 통계로 유효하지 않은 응답자료가 7,427건

○ 혼인년도의 경우 1903년보다 작고 2007년보다 큰 경우로

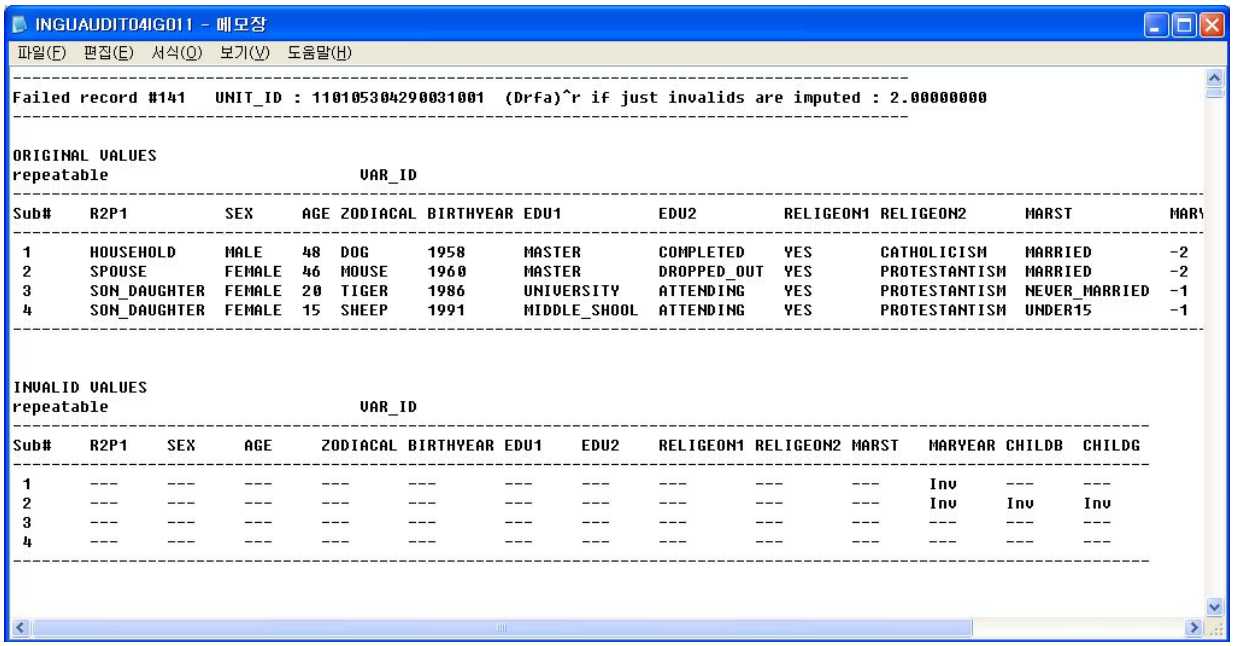
- -2가 552건, 0이 1건, 2가 4건, 193이 2건 등



□ AUDIT

○ 실패한 레코드의 가구번호와 유효하지 않은 자료에 대한 임퓨테이션시 원자료와의 거리인 Drfa는 2

- Invalid Values에서 Inv는 유효하지 않은 값을 나타냄

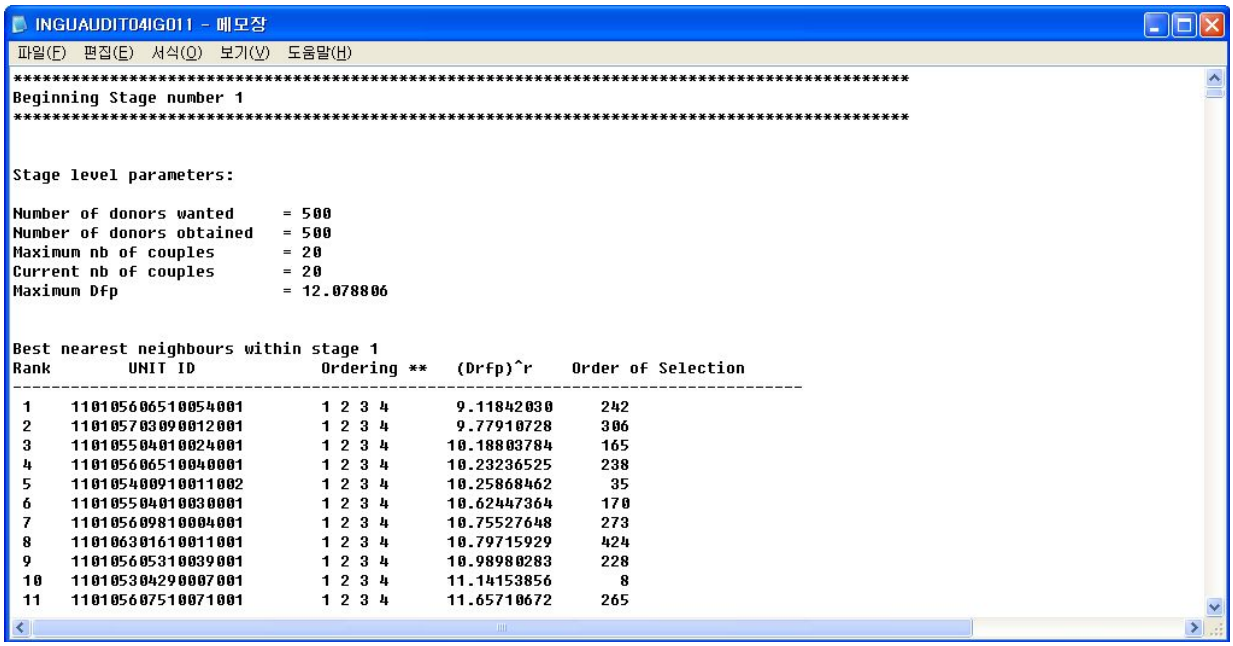


INGUAUDIT04IG011 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
-------------------------------------------------------------------------------------------------
Failed record #141   UNIT_ID : 11010530429003l001  (Drfa)^r if just invalids are imputed : 2.00000000
-------------------------------------------------------------------------------------------------

ORIGINAL VALUES
repeatable                          VAR_ID
-------------------------------------------------------------------------------------------------------------------------------------------
Sub#    R2P1          SEX     AGE ZODIACAL BIRTHYEAR EDU1          EDU2         RELIGEON1 RELIGEON2      MARST           MARY
-------------------------------------------------------------------------------------------------------------------------------------------
 1      HOUSEHOLD     MALE    48  DOG      1958      MASTER        COMPLETED    YES       CATHOLICISM    MARRIED         -2
 2      SPOUSE        FEMALE  46  MOUSE    1960      MASTER        DROPPED_OUT  YES       PROTESTANTISM  MARRIED         -2
 3      SON_DAUGHTER  FEMALE  20  TIGER    1986      UNIVERSITY    ATTENDING    YES       PROTESTANTISM  NEVER_MARRIED   -1
 4      SON_DAUGHTER  FEMALE  15  SHEEP    1991      MIDDLE_SHOOL  ATTENDING    YES       PROTESTANTISM  UNDER15         -1
-------------------------------------------------------------------------------------------------------------------------------------------


INVALID VALUES
repeatable                          VAR_ID
-----------------------------------------------------------------------------------------------------------------------
Sub#    R2P1    SEX     AGE     ZODIACAL BIRTHYEAR EDU1    EDU2    RELIGEON1 RELIGEON2 MARST   MARYEAR CHILDB  CHILDG
-----------------------------------------------------------------------------------------------------------------------
 1      ---     ---     ---     ---      ---       ---     ---     ---       ---       ---     Inv     ---     ---
 2      ---     ---     ---     ---      ---       ---     ---     ---       ---       ---     Inv     Inv     Inv
 3      ---     ---     ---     ---      ---       ---     ---     ---       ---       ---     ---     ---     ---
 4      ---     ---     ---     ---      ---       ---     ---     ---       ---       ---     ---     ---     ---
-----------------------------------------------------------------------------------------------------------------------
```

○ 첫 번째 단계에서 500개 도너중 실패한 레코드에 가장 근접한 20개의 도너에 대한 Drfp 순으로 정렬됨

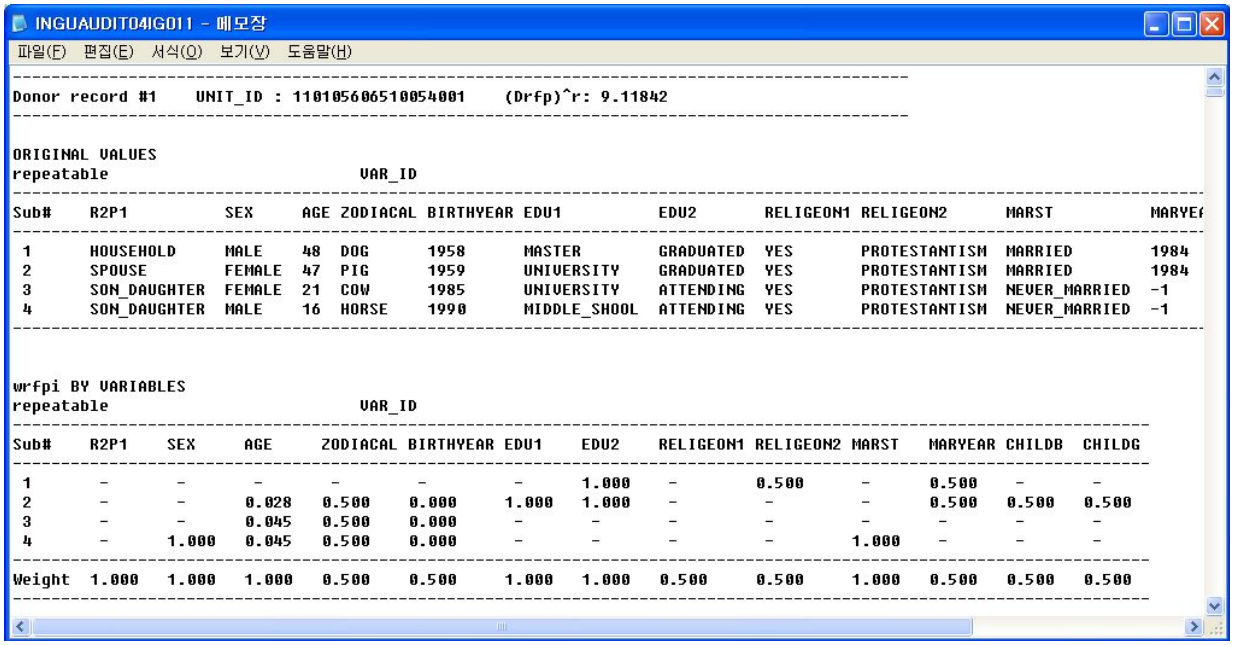


INGUAUDIT04IG011 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
*************************************************************************************************
Beginning Stage number 1
*************************************************************************************************


Stage level parameters:

Number of donors wanted     = 500
Number of donors obtained   = 500
Maximum nb of couples       = 20
Current nb of couples       = 20
Maximum Dfp                 = 12.078806


Best nearest neighbours within stage 1
Rank        UNIT ID             Ordering **   (Drfp)^r    Order of Selection
-------------------------------------------------------------------------------
 1    11010560651005400l        1 2 3 4       9.11842030     242
 2    11010570309001200l        1 2 3 4       9.77910728     306
 3    11010550401002400l        1 2 3 4      10.18803784     165
 4    11010560651004000l        1 2 3 4      10.23236525     238
 5    11010540091001100l        1 2 3 4      10.25868462      35
 6    11010550401003000l        1 2 3 4      10.62447364     170
 7    11010560981000400l        1 2 3 4      10.75527648     273
 8    11010630161001100l        1 2 3 4      10.79715929     424
 9    11010560531003900l        1 2 3 4      10.98980283     228
 10   11010530429000700l        1 2 3 4      11.14153856       8
 11   11010560751007100l        1 2 3 4      11.65710672     265
```

○ 20개 도너중 첫 번째 도너로 Drfp값은 9.11842로 아래 가중치와 거리함수에 의해 계산된 값임



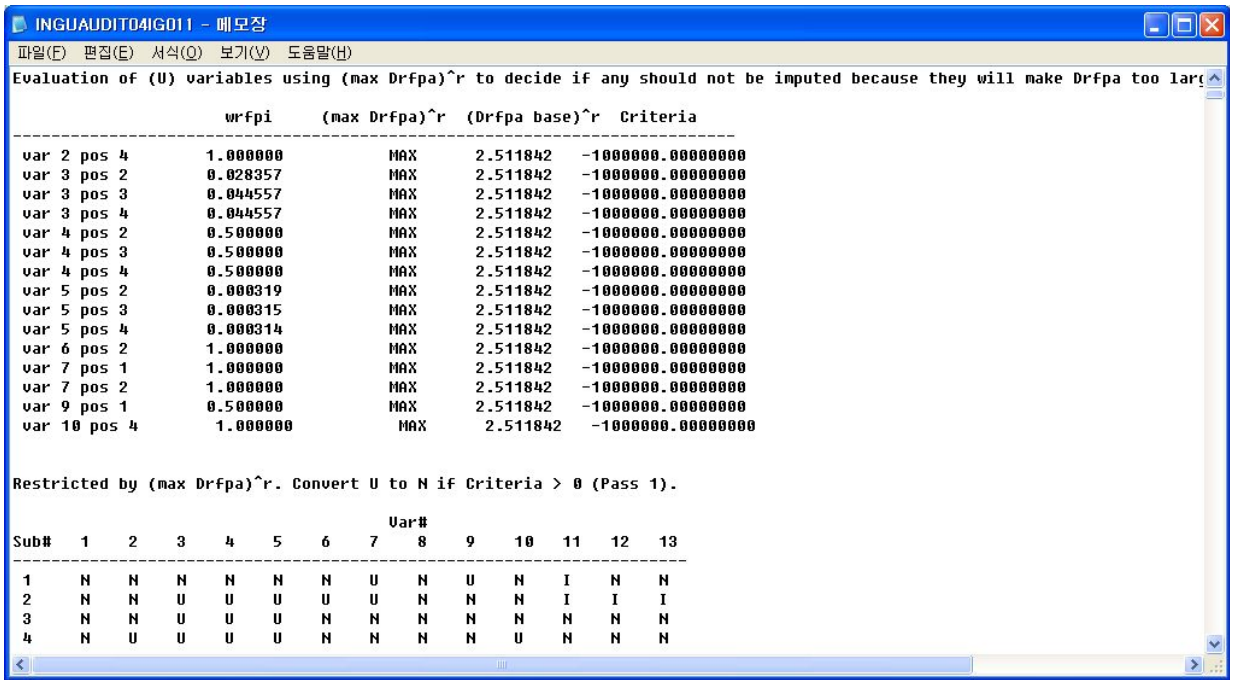
INGUAUDIT04IG011 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
-------------------------------------------------------------------------------
Donor record #1    UNIT_ID : 11010560651005400l    (Drfp)^r: 9.11842
-------------------------------------------------------------------------------

ORIGINAL VALUES
repeatable                          VAR_ID
-------------------------------------------------------------------------------------------------------------------------------
Sub#    R2P1          SEX     AGE ZODIACAL BIRTHYEAR EDU1          EDU2       RELIGEON1 RELIGEON2      MARST          MARYEA
-------------------------------------------------------------------------------------------------------------------------------
 1      HOUSEHOLD     MALE    48  DOG      1958      MASTER        GRADUATED  YES       PROTESTANTISM  MARRIED        1984
 2      SPOUSE        FEMALE  47  PIG      1959      UNIVERSITY    GRADUATED  YES       PROTESTANTISM  MARRIED        1984
 3      SON_DAUGHTER  FEMALE  21  COW      1985      UNIVERSITY    ATTENDING  YES       PROTESTANTISM  NEVER_MARRIED  -1
 4      SON_DAUGHTER  MALE    16  HORSE    1990      MIDDLE_SHOOL  ATTENDING  YES       PROTESTANTISM  NEVER_MARRIED  -1
-------------------------------------------------------------------------------------------------------------------------------


wrfpi BY VARIABLES
repeatable                          VAR_ID
-------------------------------------------------------------------------------------------------------------
Sub#    R2P1    SEX     AGE     ZODIACAL BIRTHYEAR EDU1    EDU2    RELIGEON1 RELIGEON2 MARST   MARYEAR CHILDB  CHILDG
-------------------------------------------------------------------------------------------------------------
 1       -       -       -       -        -         -     1.000     -        0.500     -      0.500     -       -
 2       -       -     0.028   0.500    0.000     1.000   1.000     -         -        -      0.500   0.500   0.500
 3       -       -     0.045   0.500    0.000       -       -       -         -        -        -       -       -
 4       -     1.000   0.045   0.500    0.000       -       -       -         -      1.000      -       -       -
-------------------------------------------------------------------------------------------------------------
Weight  1.000   1.000   1.000   0.500    0.500     1.000   1.000   0.500     0.500    1.000    0.500   0.500   0.500
-------------------------------------------------------------------------------------------------------------
```

○ Drfpa = αDfa + (1-α)Dap = 0.9*2.0 + 0.1*7.11842030 = 2.511842

- α는 SYSP File의 Dfpa fct param의 값이 0.9로 설정됨
- Dfa는 유효하지 않은 자료와의 거리는 위의 자료에서 2
- Dfpa는 도너 레코드 Drfp 9.11842030에서 Dfa 2를 뺀 값



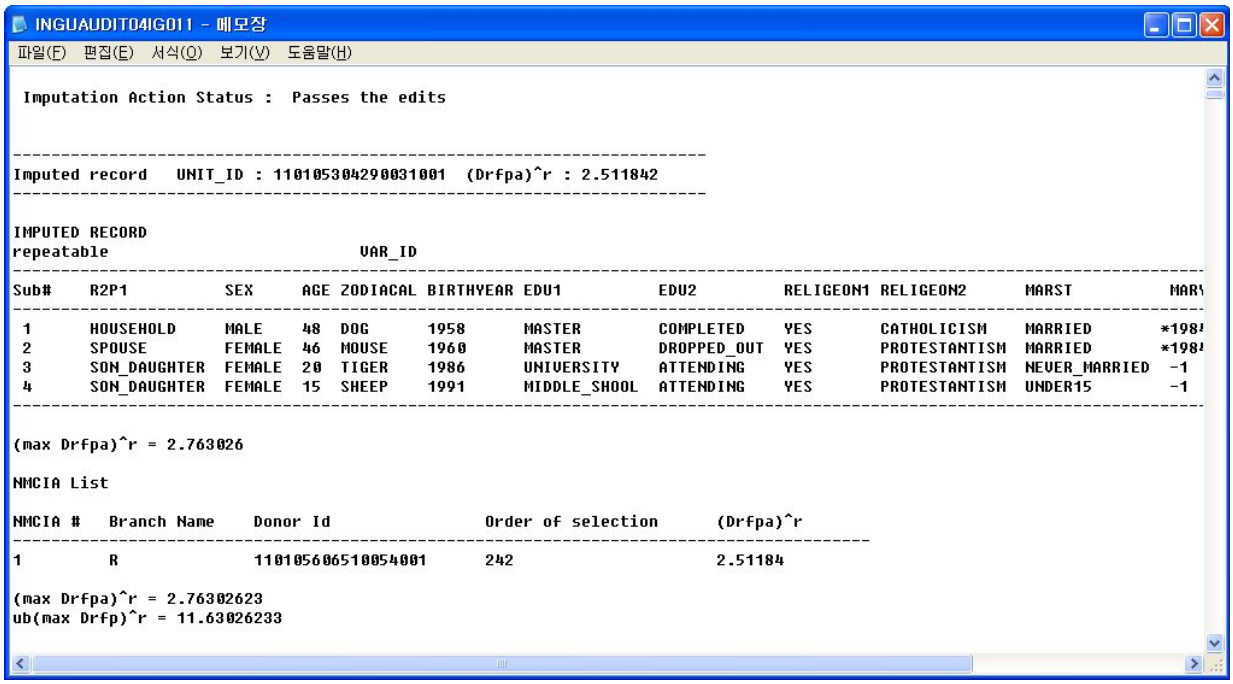
INGUAUDIT04IG011 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
Evaluation of (U) variables using (max Drfpa)^r to decide if any should not be imputed because they will make Drfpa too larg

                      wrfpi     (max Drfpa)^r  (Drfpa base)^r  Criteria
-------------------------------------------------------------------------
 var 2 pos 4        1.000000          MAX      2.511842   -1000000.00000000
 var 3 pos 2        0.028357          MAX      2.511842   -1000000.00000000
 var 3 pos 3        0.044557          MAX      2.511842   -1000000.00000000
 var 3 pos 4        0.044557          MAX      2.511842   -1000000.00000000
 var 4 pos 2        0.500000          MAX      2.511842   -1000000.00000000
 var 4 pos 3        0.500000          MAX      2.511842   -1000000.00000000
 var 4 pos 4        0.500000          MAX      2.511842   -1000000.00000000
 var 5 pos 2        0.000319          MAX      2.511842   -1000000.00000000
 var 5 pos 3        0.000315          MAX      2.511842   -1000000.00000000
 var 5 pos 4        0.000314          MAX      2.511842   -1000000.00000000
 var 6 pos 2        1.000000          MAX      2.511842   -1000000.00000000
 var 7 pos 1        1.000000          MAX      2.511842   -1000000.00000000
 var 7 pos 2        1.000000          MAX      2.511842   -1000000.00000000
 var 9 pos 1        0.500000          MAX      2.511842   -1000000.00000000
 var 10 pos 4        1.000000           MAX      2.511842   -1000000.00000000


Restricted by (max Drfpa)^r. Convert U to N if Criteria > 0 (Pass 1).

                                  Var#
Sub#   1    2    3    4    5    6    7    8    9    10   11   12   13
-----------------------------------------------------------------------
 1     N    N    N    N    N    N    U    N    U    N    I    N    N
 2     N    N    U    U    U    U    U    N    N    N    I    I    I
 3     N    N    U    U    U    N    N    N    N    N    N    N    N
 4     N    U    U    U    U    N    N    N    N    U    N    N    N
```

○ 임퓨테이션 액션에서 패스된 도너 레코드로 임퓨트 되었을 때의 값을 나타내고 NMCIA(Near Minimum Change Imputation Action) 리스트에 포함됨

- max Drfpa = Drfpa*1.1 = 2.511842*1.1 = 2.763026
- SYSP File의 Dfpa max fct param가 1.1로 설정됨

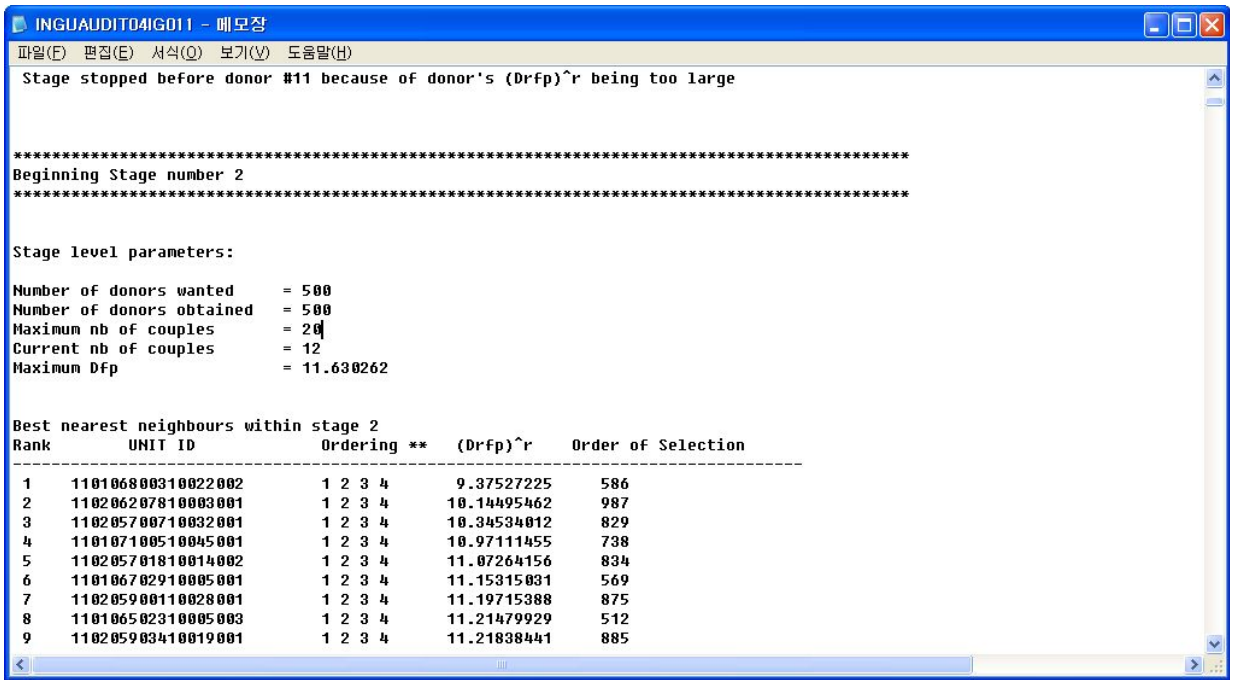


INGUAUDIT04IG011 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
 Imputation Action Status :  Passes the edits


------------------------------------------------------------------------
Imputed record   UNIT_ID : 1101053042900031001  (Drfpa)^r : 2.511842
------------------------------------------------------------------------

IMPUTED RECORD
repeatable                          VAR_ID
-------------------------------------------------------------------------------------------------------------------------------
Sub#    R2P1          SEX     AGE ZODIACAL BIRTHYEAR EDU1          EDU2         RELIGEON1 RELIGEON2      MARST           MAR
-------------------------------------------------------------------------------------------------------------------------------
 1      HOUSEHOLD     MALE    48  DOG      1958      MASTER        COMPLETED    YES       CATHOLICISM    MARRIED         *198
 2      SPOUSE        FEMALE  46  MOUSE    1960      MASTER        DROPPED_OUT  YES       PROTESTANTISM  MARRIED         *198
 3      SON_DAUGHTER  FEMALE  20  TIGER    1986      UNIVERSITY    ATTENDING    YES       PROTESTANTISM  NEVER_MARRIED   -1
 4      SON_DAUGHTER  FEMALE  15  SHEEP    1991      MIDDLE_SHOOL  ATTENDING    YES       PROTESTANTISM  UNDER15         -1
-------------------------------------------------------------------------------------------------------------------------------

(max Drfpa)^r = 2.763026

NMCIA List

NMCIA #   Branch Name    Donor Id                Order of selection      (Drfpa)^r
-------------------------------------------------------------------------------------
1         R              1101056065100540001     242                     2.51184

(max Drfpa)^r = 2.76302623
ub(max Drfp)^r = 11.63026233
```

○ 두 번째 단계에서 500개 도너중 max Dfp 11.630262보다 작은 12개가 선택됨



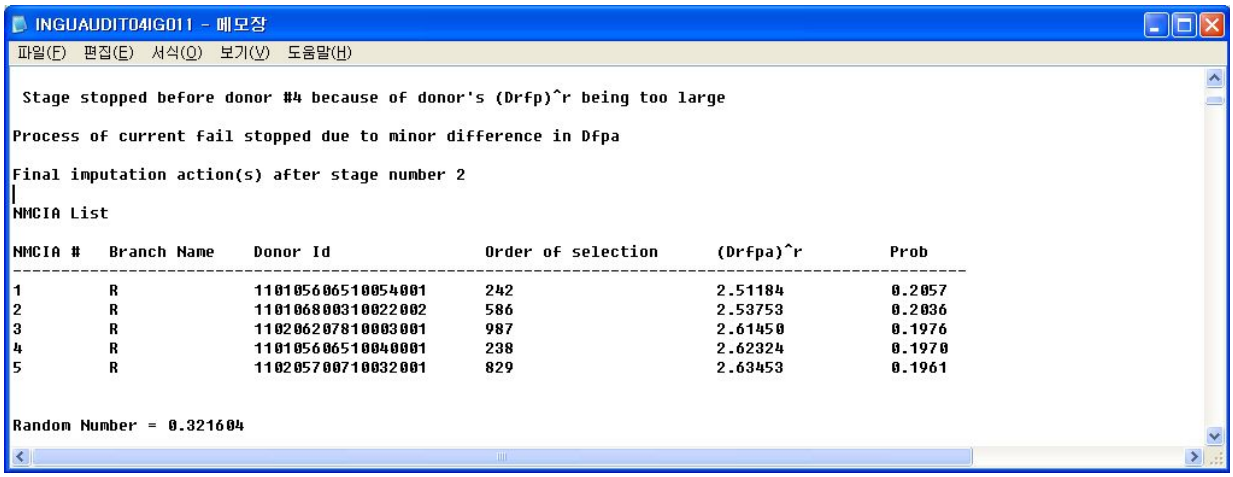
INGUAUDIT04IG011 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
 Stage stopped before donor #11 because of donor's (Drfp)^r being too large



*****************************************************************************************
Beginning Stage number 2
*****************************************************************************************


Stage level parameters:

Number of donors wanted     = 500
Number of donors obtained   = 500
Maximum nb of couples       = 20
Current nb of couples       = 12
Maximum Dfp                 = 11.630262


Best nearest neighbours within stage 2
Rank        UNIT ID             Ordering **   (Drfp)^r    Order of Selection
-------------------------------------------------------------------------------
 1    1101068003100220O2        1 2 3 4       9.37527225     586
 2    1102062078100030O1        1 2 3 4      10.14495462     987
 3    1102057007100320O1        1 2 3 4      10.34534012     829
 4    1101071005100450O1        1 2 3 4      10.97111455     738
 5    1102057018100140O2        1 2 3 4      11.07264156     834
 6    1101067029100050O1        1 2 3 4      11.15315031     569
 7    1102059001100280O1        1 2 3 4      11.19715388     875
 8    1101065023100050O3        1 2 3 4      11.21479929     512
 9    1102059034100190O1        1 2 3 4      11.21838441     885
```

○ 두 번째 단계의 4번째 도너에서 Drfp값이 너무 커서 Drfpa값이 NMCIA 리스트상의 Drfpa값보다 크므로 4번째 이후부터 멈춤

- SYSP File의 nb max of nmcias가 5로 설정됨

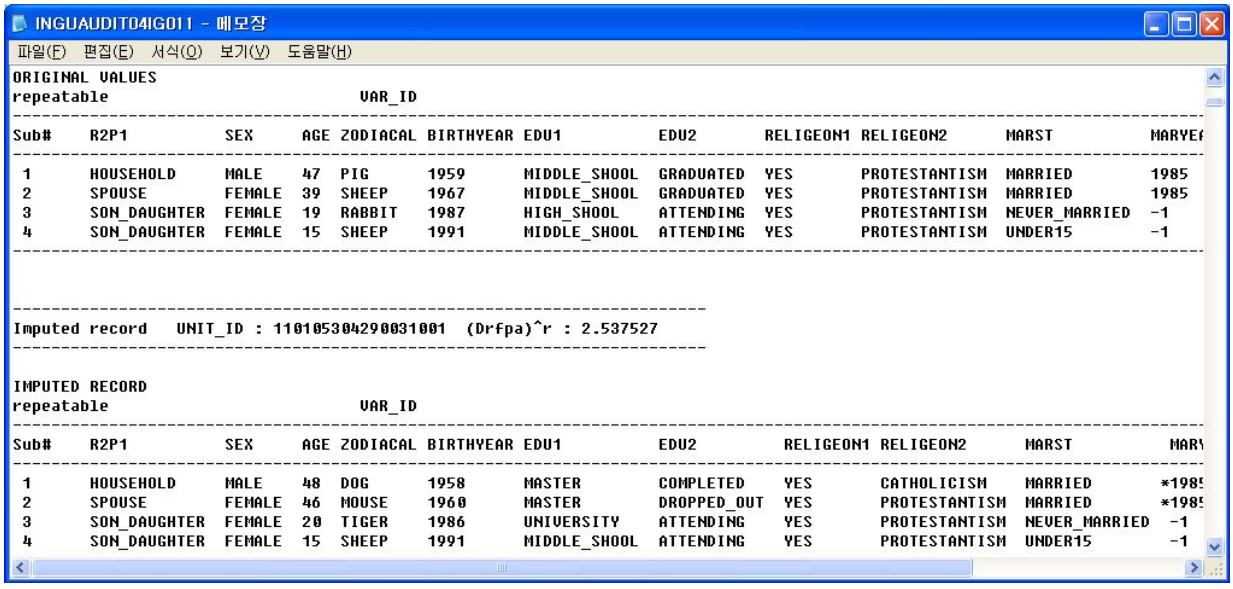


INGUAUDIT04IG011 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
 Stage stopped before donor #4 because of donor's (Drfp)^r being too large

Process of current fail stopped due to minor difference in Dfpa

Final imputation action(s) after stage number 2

NMCIA List

NMCIA #   Branch Name    Donor Id              Order of selection      (Drfpa)^r         Prob
-------------------------------------------------------------------------------------------
1         R              1101056065100540001   242                     2.51184           0.2057
2         R              1101068003100220002   586                     2.53753           0.2036
3         R              1102062078100030001   987                     2.61450           0.1976
4         R              1101056065100400001   238                     2.62324           0.1970
5         R              1102057007100320001   829                     2.63453           0.1961


Random Number = 0.321604
```

○ 랜덤넘버가 0.321604로 NMCIA상의 리스트의 5개 잠재적 도너에 대한 선택확률에 의해 2번째인 1101069003100220002가 도너로 선택됨



INGUAUDIT04IG011 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

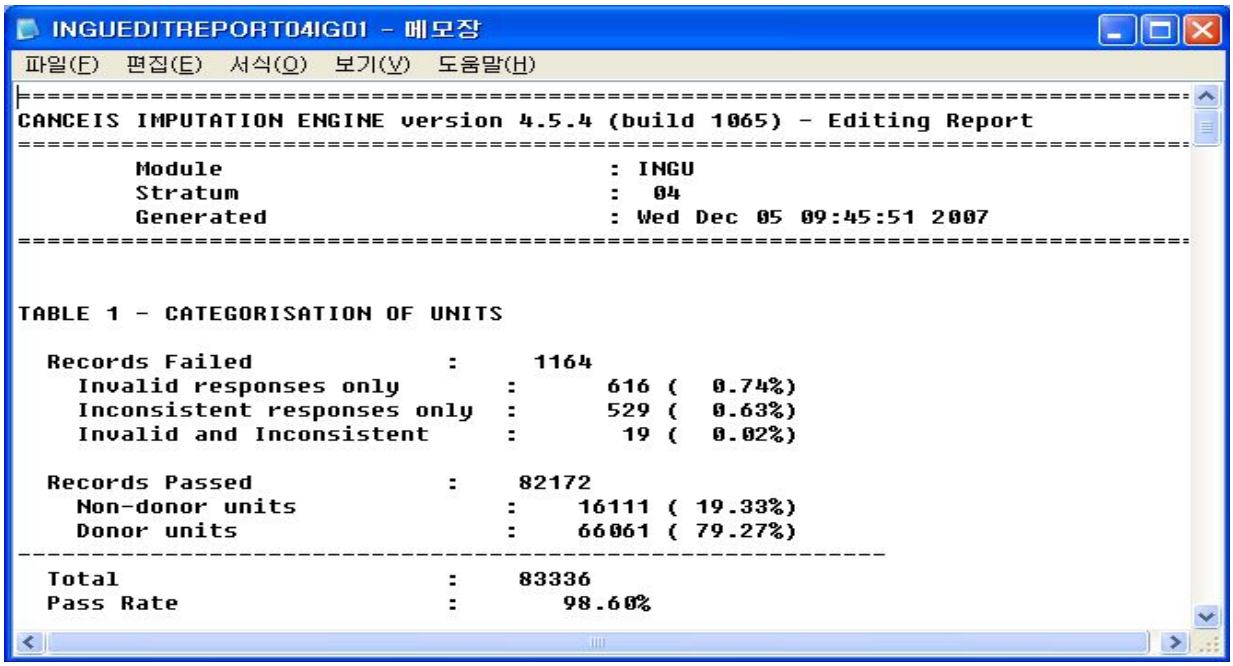
```
ORIGINAL VALUES
repeatable                          VAR_ID
-------------------------------------------------------------------------------------------------------------------------------
Sub#    R2P1          SEX     AGE ZODIACAL BIRTHYEAR EDU1          EDU2        RELIGEON1 RELIGEON2       MARST          MARYEA
-------------------------------------------------------------------------------------------------------------------------------
 1      HOUSEHOLD     MALE    47  PIG      1959      MIDDLE_SHOOL  GRADUATED   YES       PROTESTANTISM   MARRIED        1985
 2      SPOUSE        FEMALE  39  SHEEP    1967      MIDDLE_SHOOL  GRADUATED   YES       PROTESTANTISM   MARRIED        1985
 3      SON_DAUGHTER  FEMALE  19  RABBIT   1987      HIGH_SHOOL    ATTENDING   YES       PROTESTANTISM   NEVER_MARRIED  -1
 4      SON_DAUGHTER  FEMALE  15  SHEEP    1991      MIDDLE_SHOOL  ATTENDING   YES       PROTESTANTISM   UNDER15        -1
-------------------------------------------------------------------------------------------------------------------------------


-----------------------------------------------------------------
Imputed record   UNIT_ID : 1101053042900310O1  (Drfpa)^r : 2.537527
-----------------------------------------------------------------

IMPUTED RECORD
repeatable                          VAR_ID
-------------------------------------------------------------------------------------------------------------------------------
Sub#    R2P1          SEX     AGE ZODIACAL BIRTHYEAR EDU1          EDU2        RELIGEON1 RELIGEON2       MARST          MARY
-------------------------------------------------------------------------------------------------------------------------------
 1      HOUSEHOLD     MALE    48  DOG      1958      MASTER        COMPLETED   YES       CATHOLICISM     MARRIED        *1985
 2      SPOUSE        FEMALE  46  MOUSE    1960      MASTER        DROPPED_OUT YES       PROTESTANTISM   MARRIED        *1985
 3      SON_DAUGHTER  FEMALE  20  TIGER    1986      UNIVERSITY    ATTENDING   YES       PROTESTANTISM   NEVER_MARRIED  -1
 4      SON_DAUGHTER  FEMALE  15  SHEEP    1991      MIDDLE_SHOOL  ATTENDING   YES       PROTESTANTISM   UNDER15        -1
```

3) Data processed though Edit and Imputation

□ EDITREPORT

○ Table1 - Categorisation of Units

- 총 83,336 레코드 중 실패한 레코드 1,164, 성공한 레코드 82,172

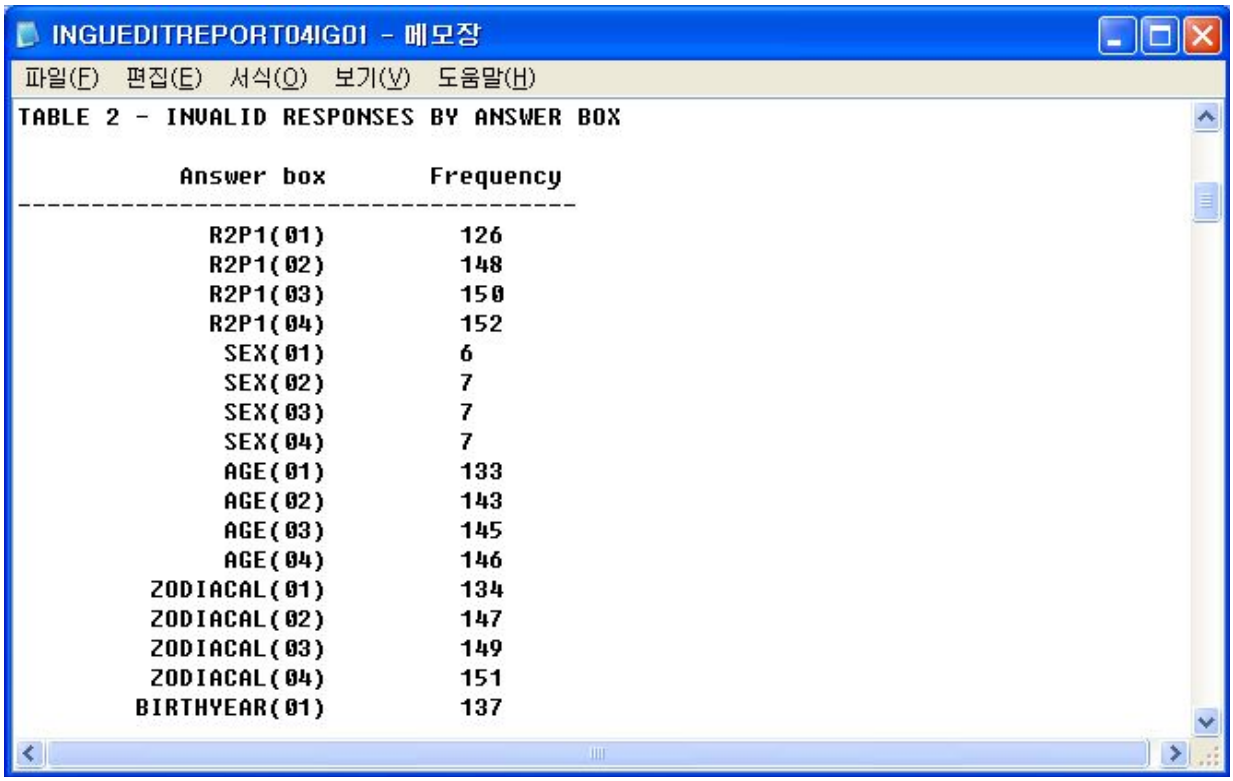


INGUEDITREPORT04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
==========================================================================
CANCEIS IMPUTATION ENGINE version 4.5.4 (build 1065) - Editing Report
==========================================================================
        Module                              : INGU
        Stratum                             :  04
        Generated                           : Wed Dec 05 09:45:51 2007
==========================================================================


TABLE 1 - CATEGORISATION OF UNITS

  Records Failed              :     1164
    Invalid responses only    :      616 (  0.74%)
    Inconsistent responses only :    529 (  0.63%)
    Invalid and Inconsistent  :       19 (  0.02%)

  Records Passed              :    82172
    Non-donor units           :    16111 ( 19.33%)
    Donor units               :    66061 ( 79.27%)
--------------------------------------------------------
  Total                       :    83336
  Pass Rate                   :       98.60%
```

○ Table2 - 변수별, 가구원별 유효하지 않은 응답 현황



INGUEDITREPORT04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
TABLE 2 - INVALID RESPONSES BY ANSWER BOX

           Answer box       Frequency
-------------------------------------
             R2P1(01)          126
             R2P1(02)          148
             R2P1(03)          150
             R2P1(04)          152
              SEX(01)          6
              SEX(02)          7
              SEX(03)          7
              SEX(04)          7
              AGE(01)          133
              AGE(02)          143
              AGE(03)          145
              AGE(04)          146
         ZODIACAL(01)          134
         ZODIACAL(02)          147
         ZODIACAL(03)          149
         ZODIACAL(04)          151
        BIRTHYEAR(01)          137
```

○ Table3 - DLT별, Rule별 실패한 건수

- 실패한 건수가 과다한 경우 DLT 확인

INGUEDITREPORT04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

TABLE 3 - RULES THAT FAIL AT LEAST ONCE

ID	DLT ID	DLT Name	Rule	Frequency
1	1	DONOR1	1	27
2	2	DONOR2	1	17
3	2	DONOR2	2	1
4	2	DONOR2	4	4
5	2	DONOR2	5	2
6	2	DONOR2	6	27
7	2	DONOR2	7	22
8	3	DONOR3A	1	17
9	3	DONOR3A	2	57
10	3	DONOR3A	3	34
11	3	DONOR3A	4	36
12	3	DONOR3A	5	53
13	3	DONOR3A	6	15
14	3	DONOR3A	7	3
15	3	DONOR3A	8	4
16	3	DONOR3A	9	1
17	4	DONOR3B	1	28

○ Table4 - 실패한 레코드가 없는 DLT Rule

INGUEDITREPORT04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

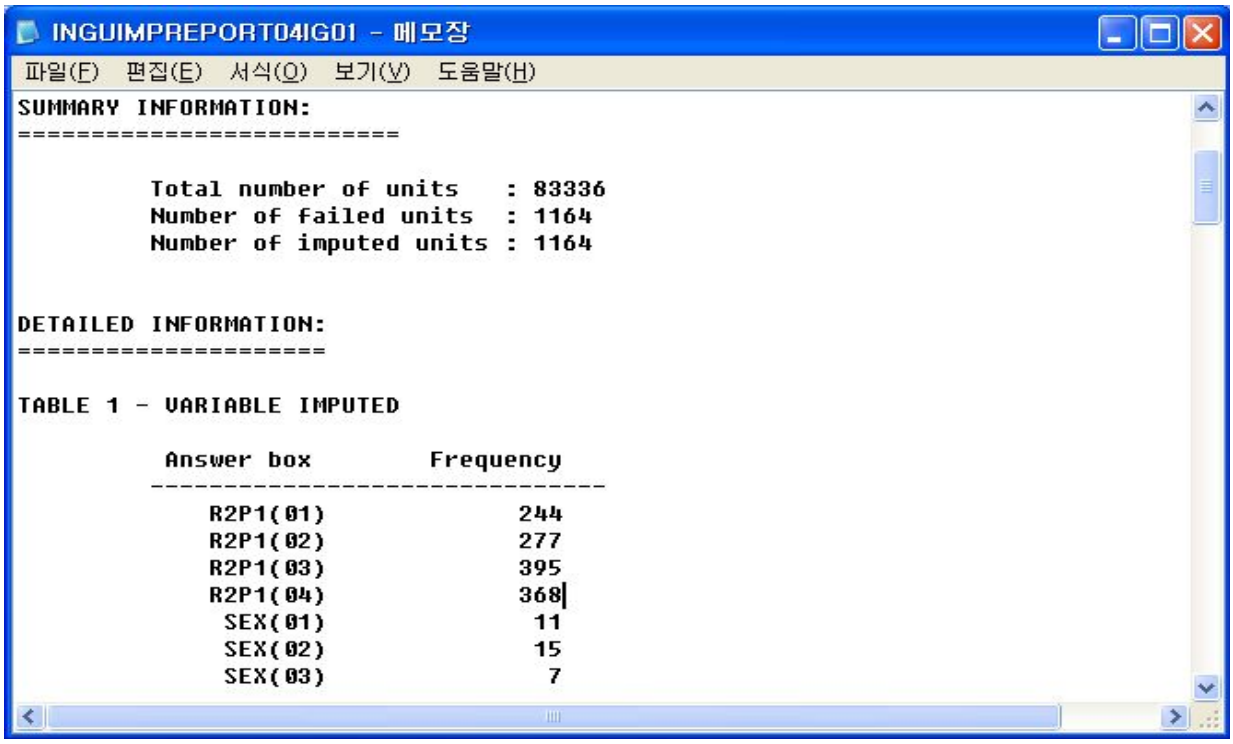
TABLE 4 - RULES THAT DO NOT FAIL ANY RECORD

ID	DLT ID	DLT Name	Rule
1	2	DONOR2	3
2	2	DONOR2	8
3	3	DONOR3A	10
4	4	DONOR3B	8
5	4	DONOR3B	9
6	4	DONOR3B	10
7	5	DONOR3C	9
8	5	DONOR3C	10
9	6	DONOR3D	9
10	6	DONOR3D	10
11	7	DONOR3E	9
12	7	DONOR3E	10
13	8	DONOR3F	9
14	8	DONOR3F	10
15	9	DONOR3G	9
16	9	DONOR3G	10
17	10	DONOR3H	9

□ IMPREPORT

○ Summary 정보 - 실패한 유니트(가구) 1,164개중 1,164개 모두 임퓨테이션 됨

○ Table1 - 변수별 가구원별 임퓨테이션 현황

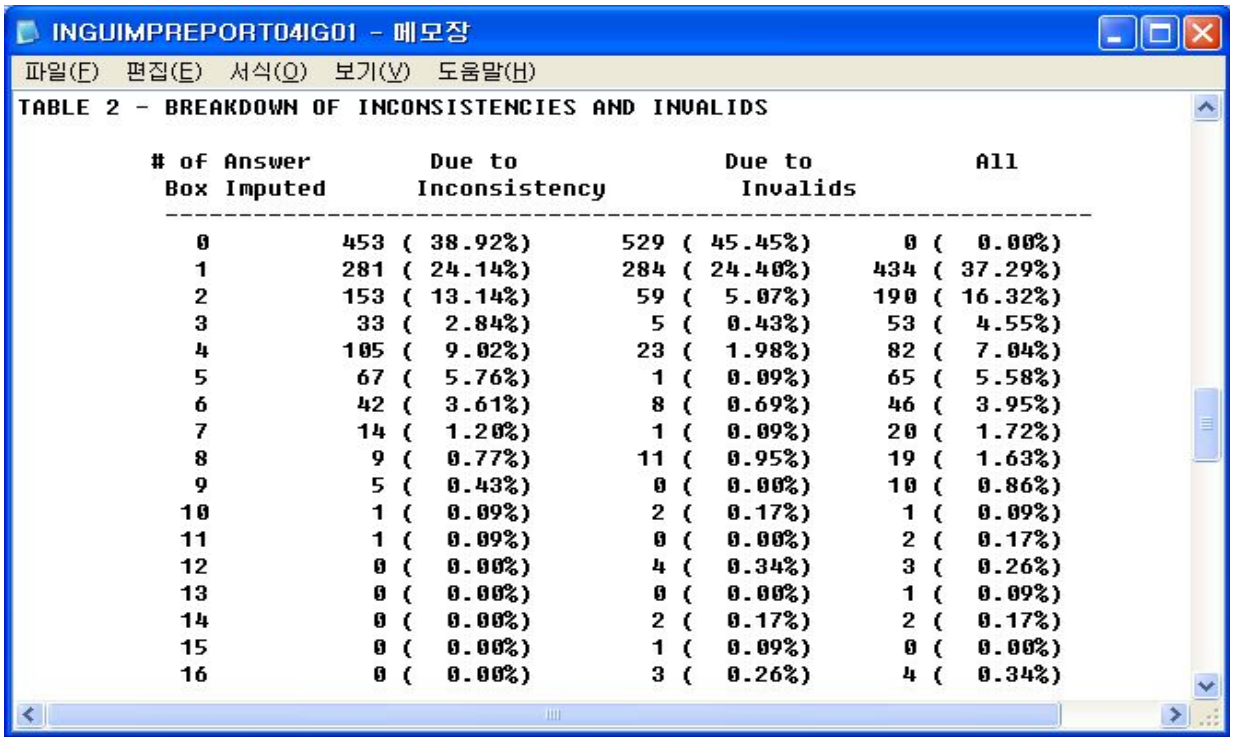


INGUIMPREPORT04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
SUMMARY INFORMATION:
============================

         Total number of units    : 83336
         Number of failed units   : 1164
         Number of imputed units  : 1164


DETAILED INFORMATION:
=====================

TABLE 1 - VARIABLE IMPUTED

          Answer box          Frequency
         --------------------------------
             R2P1(01)              244
             R2P1(02)              277
             R2P1(03)              395
             R2P1(04)              368
              SEX(01)               11
              SEX(02)               15
              SEX(03)                7
```

○ Table2 - 가구의 임퓨트된 수별 임퓨테이션 원인별 현황



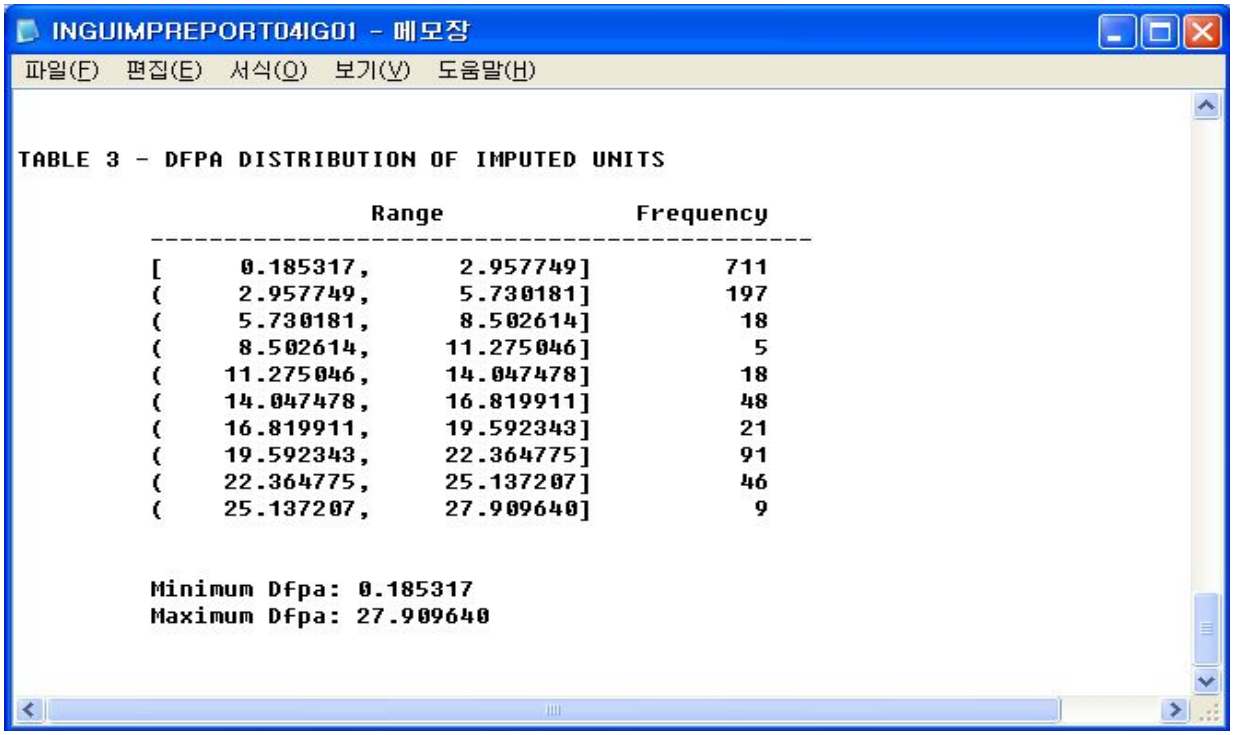
INGUIMPREPORT04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
TABLE 2 - BREAKDOWN OF INCONSISTENCIES AND INVALIDS

         # of Answer          Due to              Due to             All
          Box Imputed       Inconsistency         Invalids
          ------------------------------------------------------------------
            0          453 ( 38.92%)      529 ( 45.45%)      0 (  0.00%)
            1          281 ( 24.14%)      284 ( 24.40%)    434 ( 37.29%)
            2          153 ( 13.14%)       59 (  5.07%)    190 ( 16.32%)
            3           33 (  2.84%)        5 (  0.43%)     53 (  4.55%)
            4          105 (  9.02%)       23 (  1.98%)     82 (  7.04%)
            5           67 (  5.76%)        1 (  0.09%)     65 (  5.58%)
            6           42 (  3.61%)        8 (  0.69%)     46 (  3.95%)
            7           14 (  1.20%)        1 (  0.09%)     20 (  1.72%)
            8            9 (  0.77%)       11 (  0.95%)     19 (  1.63%)
            9            5 (  0.43%)        0 (  0.00%)     10 (  0.86%)
           10            1 (  0.09%)        2 (  0.17%)      1 (  0.09%)
           11            1 (  0.09%)        0 (  0.00%)      2 (  0.17%)
           12            0 (  0.00%)        4 (  0.34%)      3 (  0.26%)
           13            0 (  0.00%)        0 (  0.00%)      1 (  0.09%)
           14            0 (  0.00%)        2 (  0.17%)      2 (  0.17%)
           15            0 (  0.00%)        1 (  0.09%)      0 (  0.00%)
           16            0 (  0.00%)        3 (  0.26%)      4 (  0.34%)
```

○ Table3 - 임퓨테이션된 가구의 Dfpa 분포

- Dfpa의 최소값은 0.185317, 최대값은 27.909640



INGUIMPREPORT04IG01 - 메모장

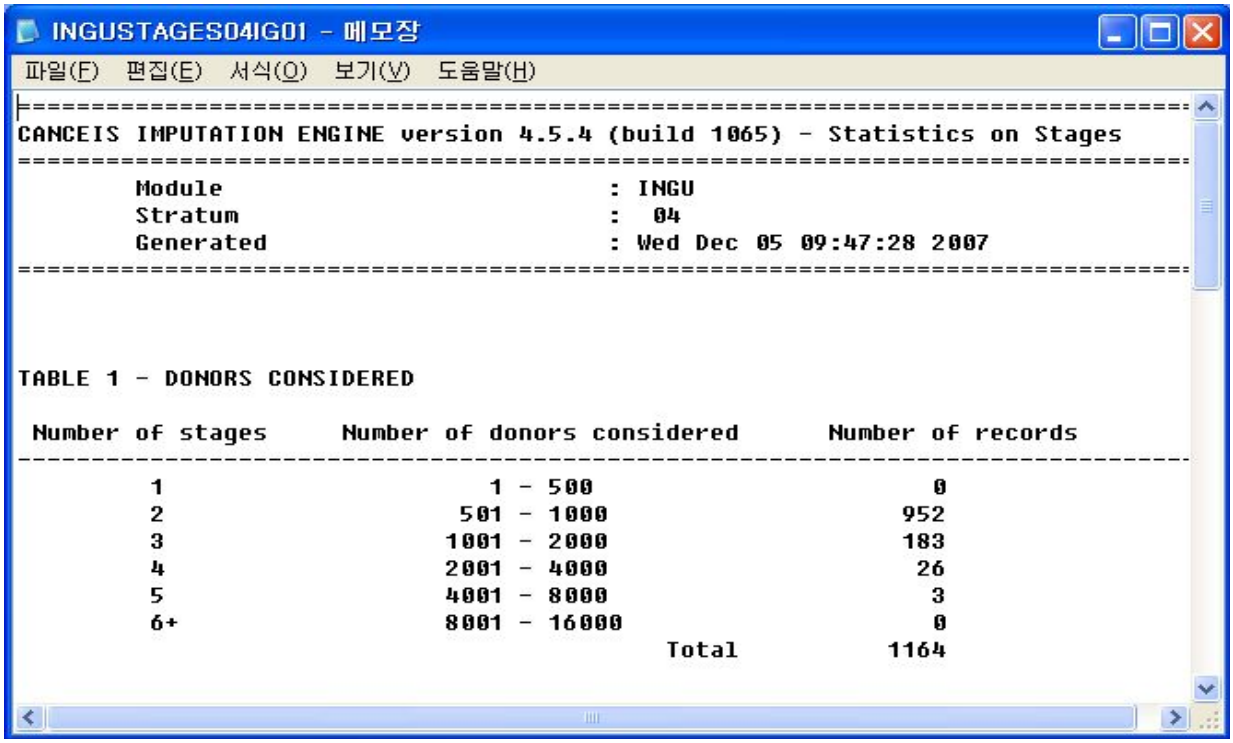
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
TABLE 3 - DFPA DISTRIBUTION OF IMPUTED UNITS

                    Range                Frequency
          ------------------------------------------
          [     0.185317,      2.957749]         711
          (     2.957749,      5.730181]         197
          (     5.730181,      8.502614]          18
          (     8.502614,     11.275046]           5
          (    11.275046,     14.047478]          18
          (    14.047478,     16.819911]          48
          (    16.819911,     19.592343]          21
          (    19.592343,     22.364775]          91
          (    22.364775,     25.137207]          46
          (    25.137207,     27.909640]           9


          Minimum Dfpa: 0.185317
          Maximum Dfpa: 27.909640
```

□ STAGES

○ Table1 - 도너를 찾는 단계별 찾은 도너 수

- 2, 3단계에서 도너가 많은 것이 좋음

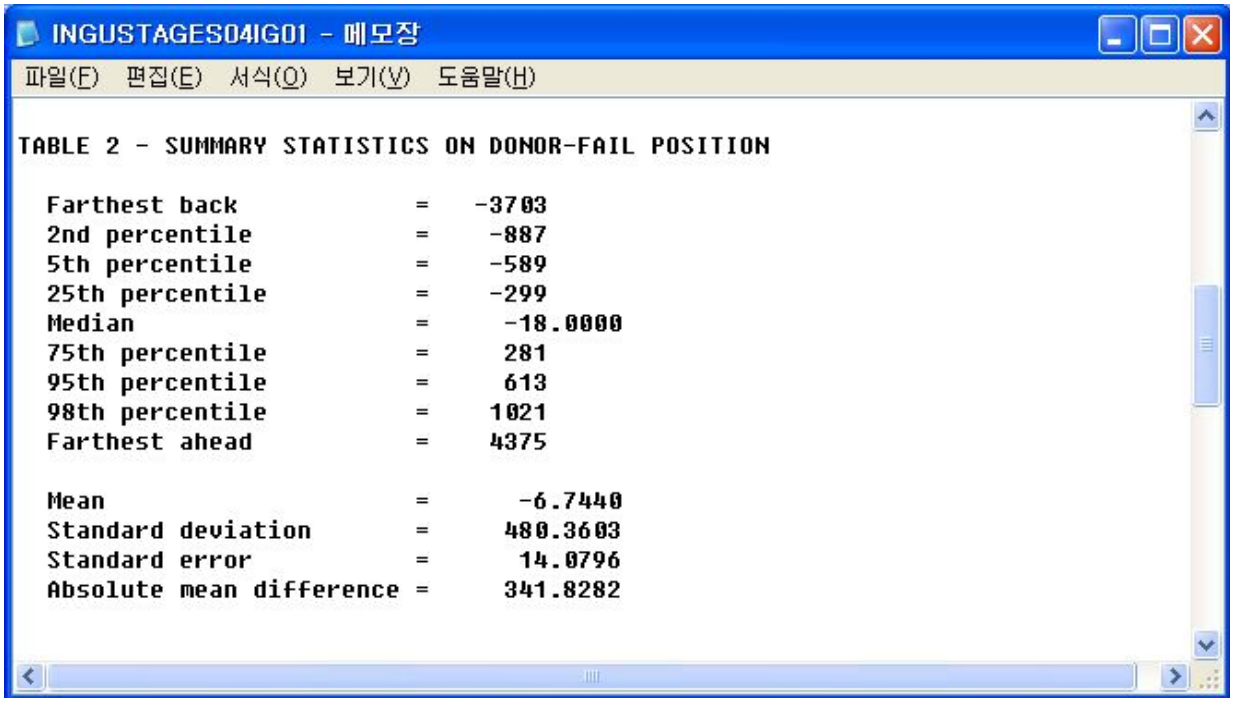


INGUSTAGES04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
================================================================================
CANCEIS IMPUTATION ENGINE version 4.5.4 (build 1065) - Statistics on Stages
================================================================================
        Module                              : INGU
        Stratum                             :  04
        Generated                           : Wed Dec 05 09:47:28 2007
================================================================================



TABLE 1 - DONORS CONSIDERED

 Number of stages     Number of donors considered      Number of records
--------------------------------------------------------------------------------
         1                       1 - 500                          0
         2                     501 - 1000                       952
         3                    1001 - 2000                       183
         4                    2001 - 4000                        26
         5                    4001 - 8000                         3
         6+                   8001 - 16000                        0
                                        Total                  1164
```

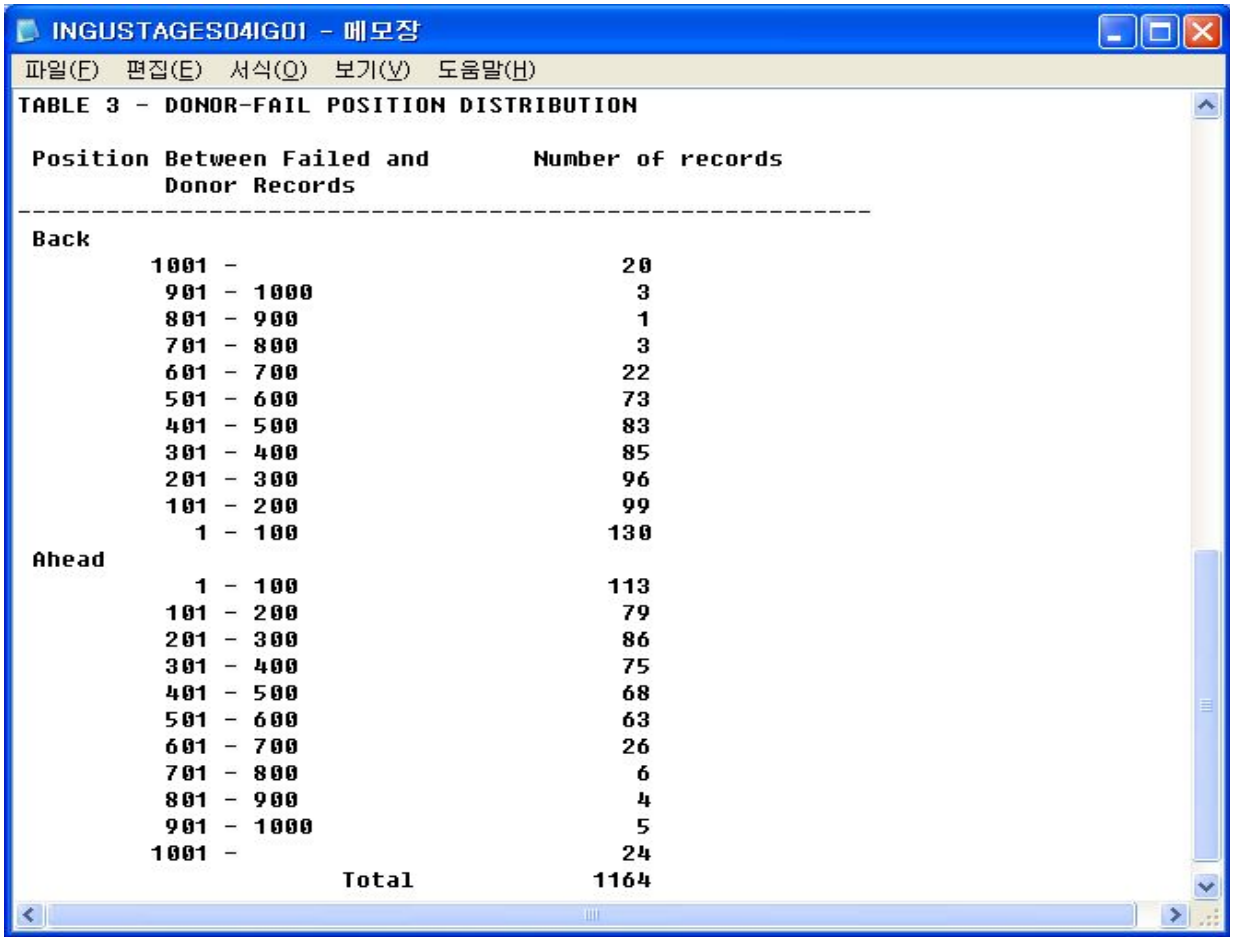
○ Table2 - 도너와 실패한 자료의 위치에 대한 통계


INGUSTAGES04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
TABLE 2 - SUMMARY STATISTICS ON DONOR-FAIL POSITION

  Farthest back             =   -3703
  2nd percentile            =    -887
  5th percentile            =    -589
  25th percentile           =    -299
  Median                    =     -18.0000
  75th percentile           =     281
  95th percentile           =     613
  98th percentile           =    1021
  Farthest ahead            =    4375

  Mean                      =      -6.7440
  Standard deviation        =     480.3603
  Standard error            =      14.0796
  Absolute mean difference  =     341.8282
```

○ Table3 - 도너와 실패한 자료의 위치 분포


INGUSTAGES04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

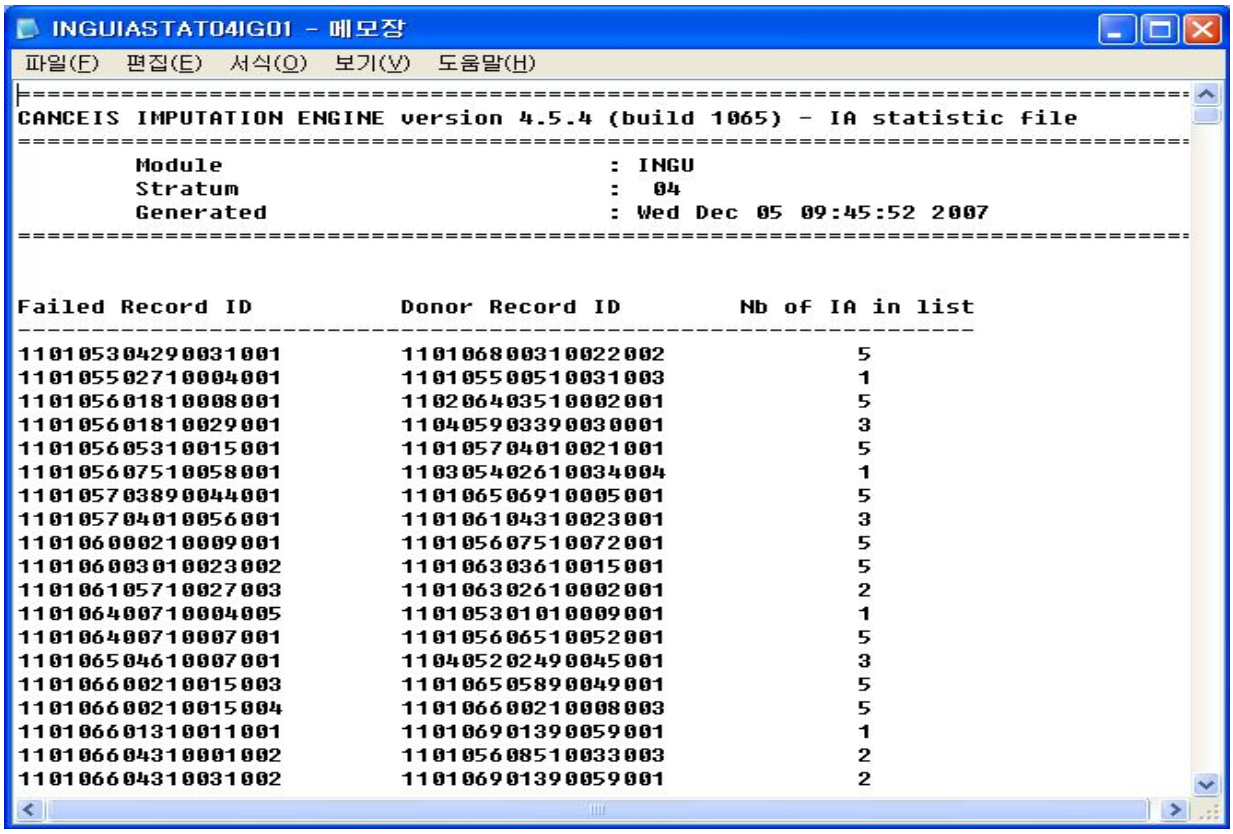
```
TABLE 3 - DONOR-FAIL POSITION DISTRIBUTION

 Position Between Failed and       Number of records
          Donor Records
-------------------------------------------------------------
 Back
         1001 -                              20
          901 - 1000                          3
          801 - 900                           1
          701 - 800                           3
          601 - 700                          22
          501 - 600                          73
          401 - 500                          83
          301 - 400                          85
          201 - 300                          96
          101 - 200                          99
            1 - 100                         130
 Ahead
            1 - 100                         113
          101 - 200                          79
          201 - 300                          86
          301 - 400                          75
          401 - 500                          68
          501 - 600                          63
          601 - 700                          26
          701 - 800                           6
          801 - 900                           4
          901 - 1000                          5
         1001 -                              24
                      Total                1164
```

□ IASTAT

○ Imputation Action에 대한 통계

- AUDIT File상의 NMCIA에 있는 Imputation Action 리스트 수

Nb of IA in list	레코드수
1	272
2	209
3	121
4	85
5	477
합계	1,164



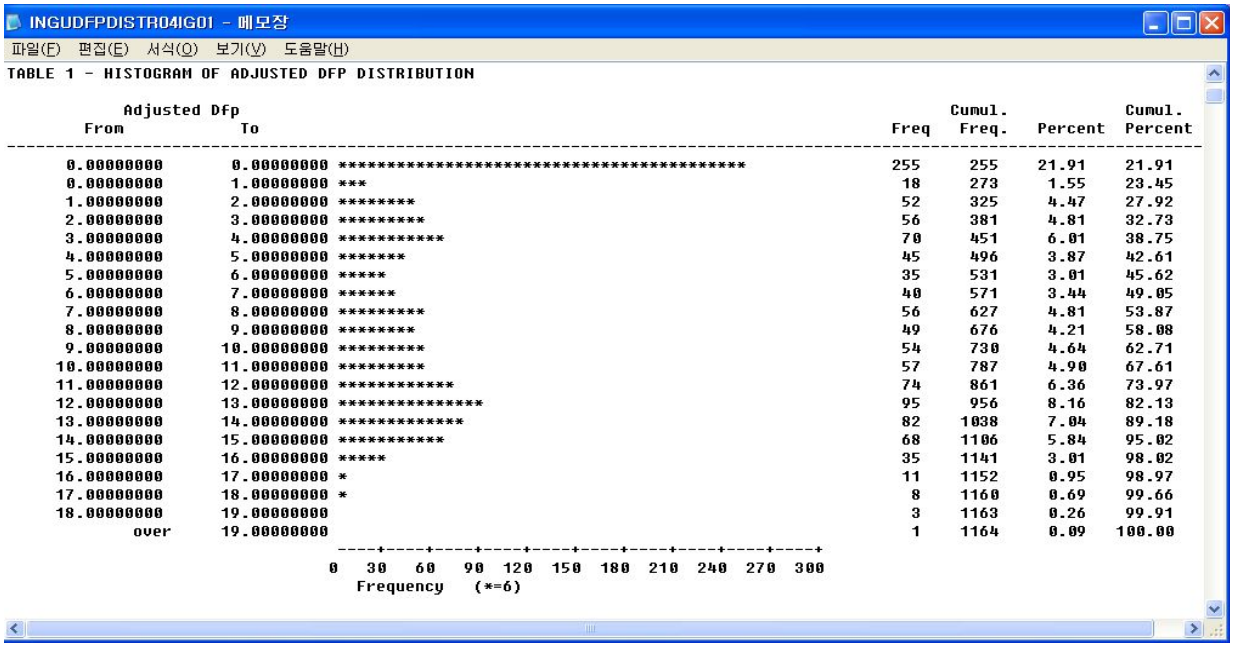
INGUIASTAT04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
==========================================================================================
CANCEIS IMPUTATION ENGINE version 4.5.4 (build 1065) - IA statistic file
==========================================================================================
        Module                                : INGU
        Stratum                               :  04
        Generated                             : Wed Dec 05 09:45:52 2007
==========================================================================================


Failed Record ID           Donor Record ID         Nb of IA in list
------------------------------------------------------------------
1101053042900310001        1101068003100220002                5
1101055027100040001        1101055005100310003                1
1101056018100080001        1102064035100020001                5
1101056018100290001        1104059033900300001                3
1101056053100150001        1101057040100210001                5
1101056075100580001        1103054026100340004                1
1101057038900440001        1101065069100050001                5
1101057040100560001        1101061043100230001                3
1101060002100090001        1101056075100720001                5
1101060030100230002        1101063036100150001                5
1101061057100270003        1101063026100020001                2
1101064007100040005        1101053010100090001                1
1101064007100070001        1101056065100520001                5
1101065046100070001        1104052024900450001                3
1101066002100150003        1101065058900490001                5
1101066002100150004        1101066002100080003                5
1101066013100110001        1101069013900590001                1
1101066043100010002        1101056085100330003                2
1101066043100310002        1101069013900590001                2
```

□ DFPDISTR

○ Table1 - 조정된 Dfp 분포의 히스토그램



INGUDFPDISTR04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

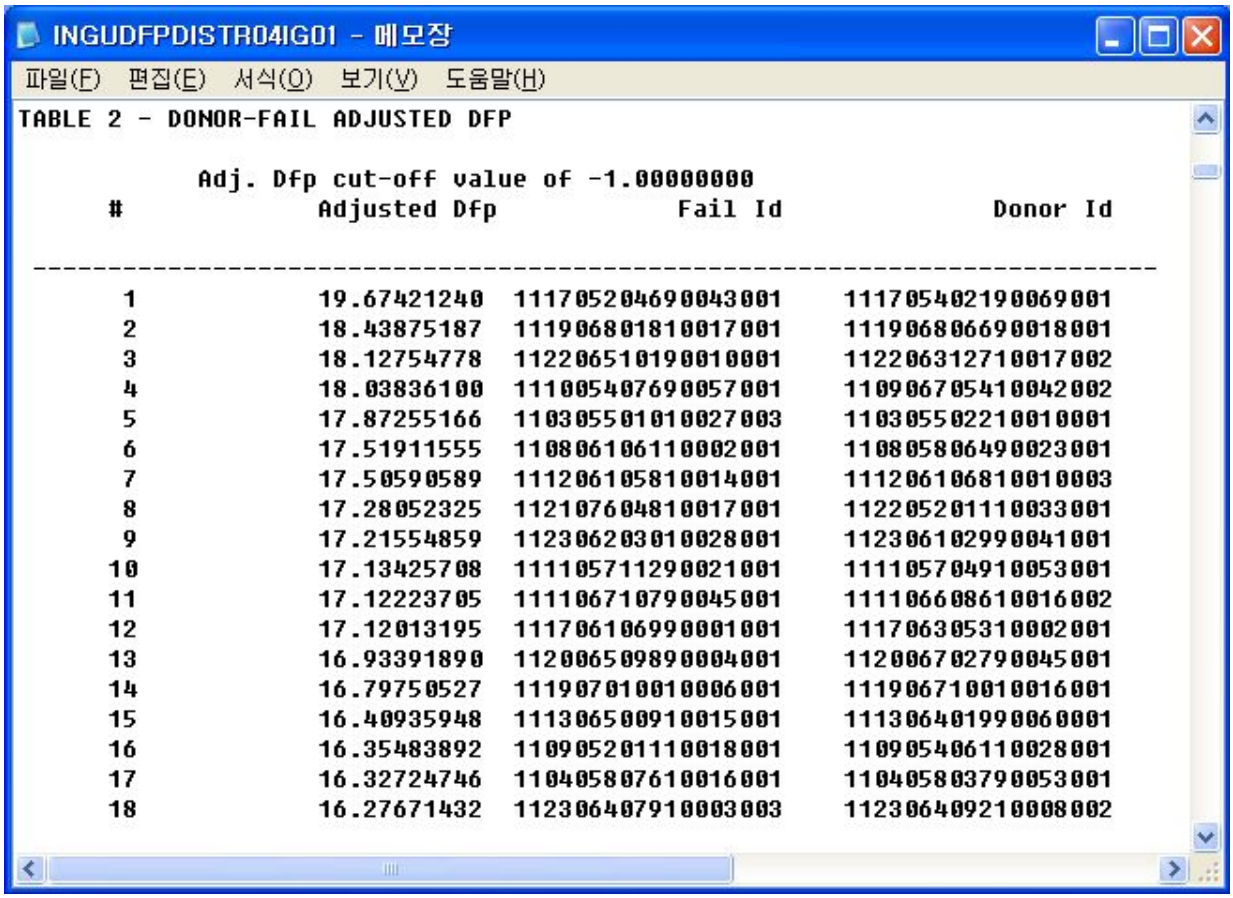
TABLE 1 - HISTOGRAM OF ADJUSTED DFP DISTRIBUTION

Adjusted Dfp From	To		Freq	Cumul. Freq.	Percent	Cumul. Percent
0.00000000	0.00000000	**********	255	255	21.91	21.91
0.00000000	1.00000000	***	18	273	1.55	23.45
1.00000000	2.00000000	********	52	325	4.47	27.92
2.00000000	3.00000000	*********	56	381	4.81	32.73
3.00000000	4.00000000	**********	70	451	6.01	38.75
4.00000000	5.00000000	*******	45	496	3.87	42.61
5.00000000	6.00000000	*****	35	531	3.01	45.62
6.00000000	7.00000000	******	40	571	3.44	49.05
7.00000000	8.00000000	*********	56	627	4.81	53.87
8.00000000	9.00000000	********	49	676	4.21	58.08
9.00000000	10.00000000	*********	54	730	4.64	62.71
10.00000000	11.00000000	*********	57	787	4.90	67.61
11.00000000	12.00000000	**********	74	861	6.36	73.97
12.00000000	13.00000000	**********	95	956	8.16	82.13
13.00000000	14.00000000	**********	82	1038	7.04	89.18
14.00000000	15.00000000	**********	68	1106	5.84	95.02
15.00000000	16.00000000	*****	35	1141	3.01	98.02
16.00000000	17.00000000	*	11	1152	0.95	98.97
17.00000000	18.00000000	*	8	1160	0.69	99.66
18.00000000	19.00000000		3	1163	0.26	99.91
over	19.00000000		1	1164	0.09	100.00

0 30 60 90 120 150 180 210 240 270 300

Frequency (*=6)

○ Table2 - 도너와 실패한 자료의 조정된 Dfp 리스트



INGUDFPDISTR04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

TABLE 2 - DONOR-FAIL ADJUSTED DFP

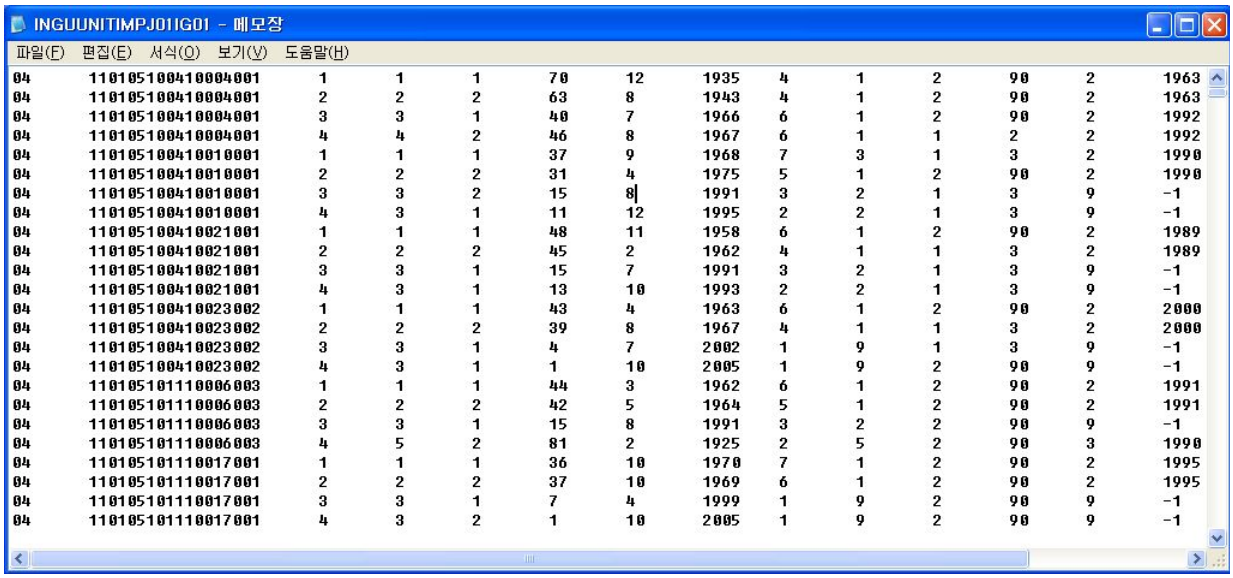
Adj. Dfp cut-off value of -1.00000000

#	Adjusted Dfp	Fail Id	Donor Id
1	19.67421240	1117052046900043001	1117054021900069001
2	18.43875187	1119068018100017001	1119068066900018001
3	18.12754778	1122065101900010001	1122063127100017002
4	18.03836100	1110054076900057001	1109067054100042002
5	17.87255166	1103055010100027003	1103055022100010001
6	17.51911555	1108061061100002001	1108058064900023001
7	17.50590589	1112061058100014001	1112061068100010003
8	17.28052325	1121076048100017001	1122052011100033001
9	17.21554859	1123062030100028001	1123061029900041001
10	17.13425708	1111057112900021001	1111057049100053001
11	17.12223705	1111067107900045001	1111066086100016002
12	17.12013195	1117061069900001001	1117063053100002001
13	16.93391890	1120065098900004001	1120067027900045001
14	16.79750527	1119070100100006001	1119067100100016001
15	16.40935948	1113065009100015001	1113064019900060001
16	16.35483892	1109052011100018001	1109054061100028001
17	16.32724746	1104058076100016001	1104058037900053001
18	16.27671432	1123064079100003003	1123064092100008002

4) Files after Imputation

□ UNITIMP

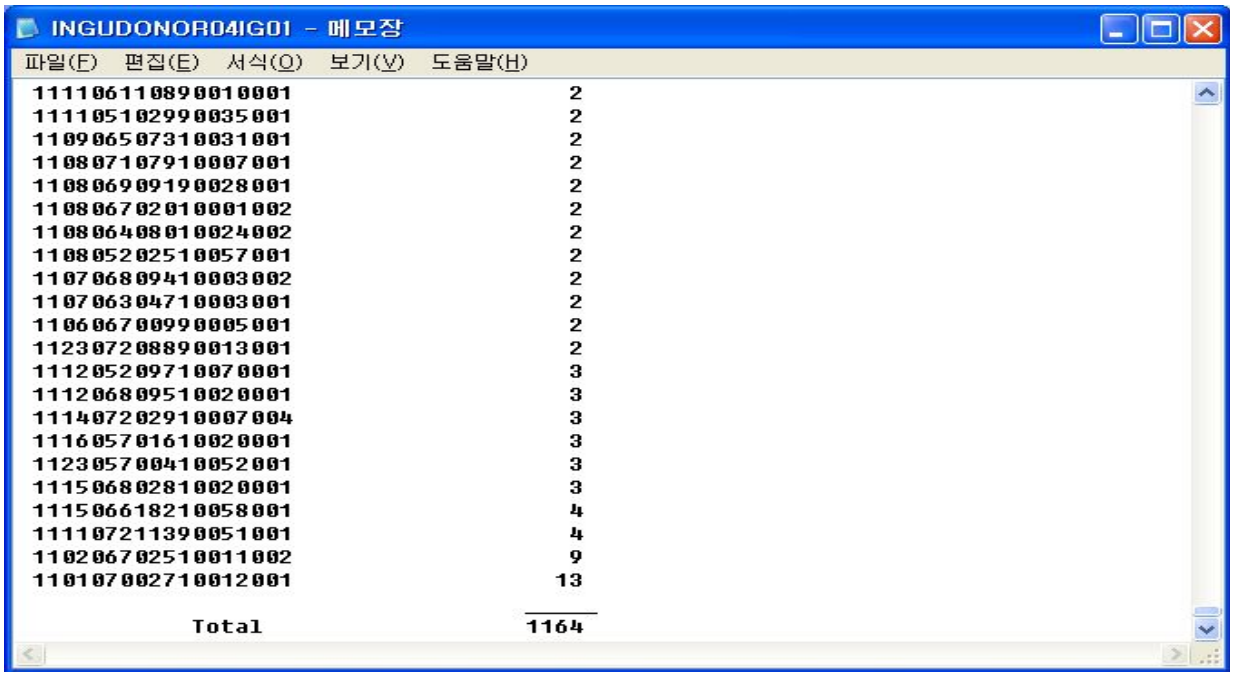
○ 임퓨테이션된 UNIT File



□ DONOR

○ 도너로서 사용된 현황으로 첫 번째는 도너 ID, 두 번째는 도너로 선택된 횟수

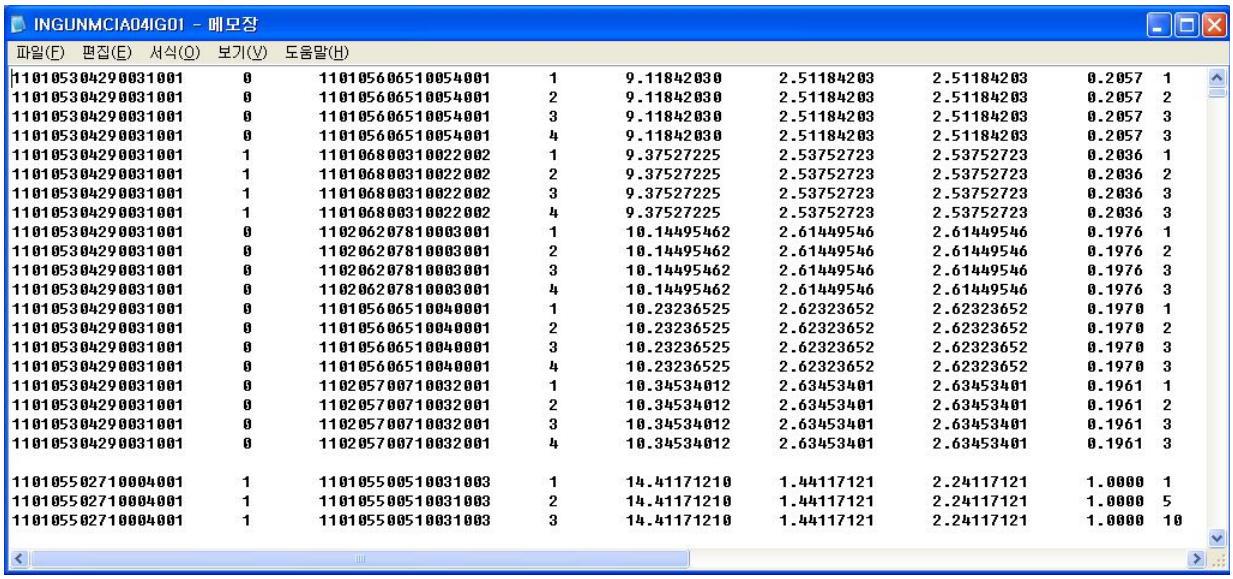
도너 선택 횟수	레코드수	도너 선택 횟수	레코드수
1	1,022	4	2
2	47	9	1
3	6	13	1



□ NMCIA

○ Near Minimum Change Imputation Action

#	Field	내용
1	UNIT ID	실패한 레코드의 고유번호
2	Used as Donor flag	도너로 사용 여부
3	Potential Donor ID	잠재적 도너 고유번호
4	Sub-unit Ordering to match Failed Unit with Donor	가구원번호
5	Dfp	실패한 레코드와 잠재도너와의 거리
6	Dfpa value after imputing invalid data only	유효하지 않은 자료의 임퓨테이션 후 거리
7	Dfpa value after imputing invalid and inconsistent data	유효하지 않은 자료와 불일치 자료의 임퓨테이션 후 거리
8	PROBSEL	Probability of selection of the unit in the sample
9~	Value of variable after imputation	임퓨테이션 후 변수 값

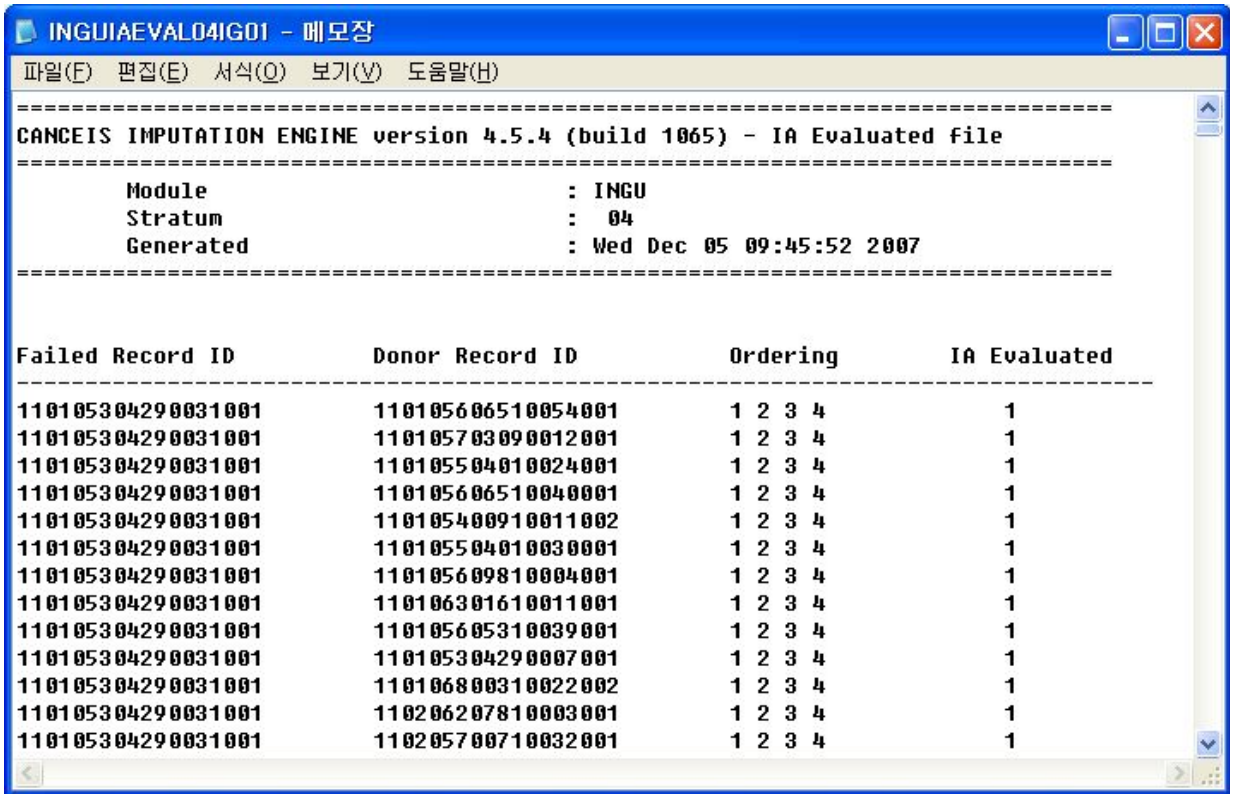


□ IAEVAL

○ Imputation Action Evaluated file

#	Field	비고
1	Failed UNIT ID	실패된 가구번호
2	Potential Donor ID	잠재 도너 가구번호
3	Sub-unit Ordering	가구원 순서
4	Nb of imputation Actions	임퓨테이션 액션수

Nb of imputation Actions	레코드수
1	30,500
3	577
5	41
7	2
9	1
합계	31,121



INGUIAEVAL04IG01 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
================================================================================
CANCEIS IMPUTATION ENGINE version 4.5.4 (build 1065) - IA Evaluated file
================================================================================
        Module                          : INGU
        Stratum                         :  04
        Generated                       : Wed Dec 05 09:45:52 2007
================================================================================


Failed Record ID          Donor Record ID           Ordering        IA Evaluated
--------------------------------------------------------------------------------
1101053042900310O1        1101056065100540O1        1 2 3 4             1
1101053042900310O1        1101057030900120O1        1 2 3 4             1
1101053042900310O1        1101055040100240O1        1 2 3 4             1
1101053042900310O1        1101056065100400O1        1 2 3 4             1
1101053042900310O1        1101054009100110O2        1 2 3 4             1
1101053042900310O1        1101055040100300O1        1 2 3 4             1
1101053042900310O1        1101056098100040O1        1 2 3 4             1
1101053042900310O1        1101063016100110O1        1 2 3 4             1
1101053042900310O1        1101056053100390O1        1 2 3 4             1
1101053042900310O1        1101053042900070O1        1 2 3 4             1
1101053042900310O1        1101068003100220O2        1 2 3 4             1
1101053042900310O1        1102062078100030O1        1 2 3 4             1
1101053042900310O1        1102057007100320O1        1 2 3 4             1
```

□ IMPSTATUS

○ 레코드별 변수별 임퓨테이션 상태 정보

UNIT ID	Classification of why unit failed	Imputation type of variable	Donor ID

○ Classification of why unit failed

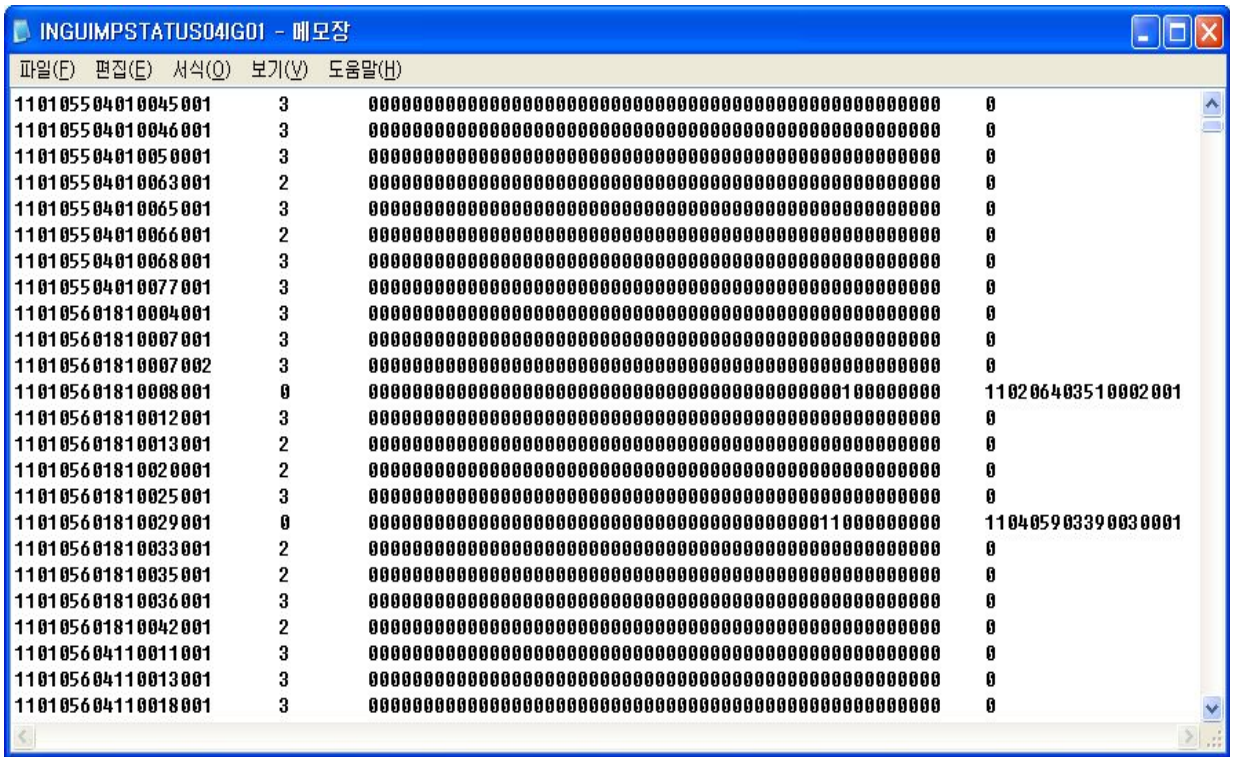
- 0 : 유효하지 않은 변수만 존재, 불일치 변수는 없는 경우
- 1 : 불일치 변수만 존재, 유효하지 않은 변수는 없는 경우
- 2 : 모든 변수가 유효하나 도너가 아닌 경우
- 3 : 모든 변수가 유효하고 도너인 경우
- 4 : 불일치 변수와 유효하지 않은 변수 모두 존재

○ Imputation type of variable

- 0 : 임퓨테이션 되지 않고 이상치도 아닌 경우
- 1 : 유효하지 않은 변수로 임퓨테이션 된 경우
- 2 : 불일치한 변수로 임퓨테이션 된 경우
- 3 : 임퓨테이션 되지 않았으나 이상치인 경우

○ Donor ID

- 0 : 임퓨테이션 변수가 없는 경우
- 4 : 임퓨테이션이 실패한 경우



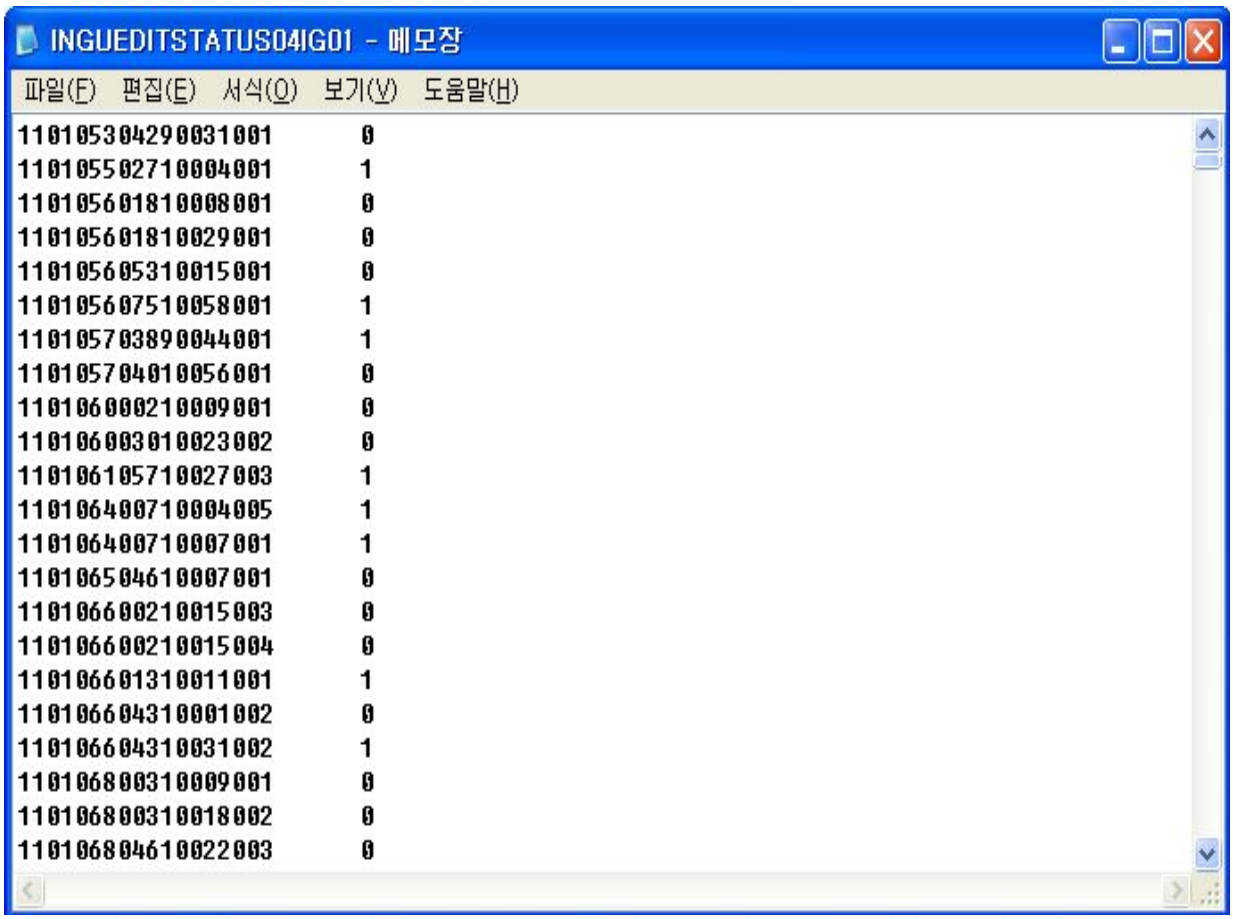
□ EDITSTATUS

○ 실패한 레코드의 실패 이유 정보

UNIT ID	Classification of why unit failed
	0 : 유효하지 않은 변수만 존재, 불일치 변수는 없는 경우 1 : 불일치 변수만 존재, 유효하지 않은 변수는 없는 경우 4 : 불일치 변수와 유효하지 않은 변수 모두 존재

○ EDITREPORT의 table1의 실패한 레코드에 대한 세부내역

- 실패한 레코드 : 1,164
  - · 유효하지 않은 응답만(0) : 616
  - · 불일치한 응답만(1) : 529
  - · 유효하지 않은 응답과 불일치 응답 모두(4) : 19



Ⅳ. 결론 및 향후 방향

1. CANCEIS System 적용 결과

□ CANCEIS System의 장점

○ Edit와 Imputation의 시스템화

- DA(DLT Analyzer), IE(Imputation Engine), DE(Derive Engine)의 체계적인 시스템

○ DLT(Decision Logic Tables)

- DLT를 통하여 유효하지 않은 자료의 결정, 도너의 결정
- DLT의 문법오류 체크와 DLT별 오류건수 현황

○ 가구원수(STRATA)에 의한 Imputation

- 가구원별 임퓨테이션이 아닌 가구원수가 같은 가구별로 임퓨테이션하여 가구구조의 특성을 반영하여 임퓨테이션

○ Imputation 과정의 이론적 배경

- Near minimum change imputation
- 13가지의 Distance Functions 등 다양한 이론적 배경의 뒷받침

○ Imputation 과정의 기록

- AUDIT file에 임퓨테이션되어지는 과정의 흐름이 상세하게 기록

○ Imputation된 자료의 관리

- 임퓨테이션후 나오는 결과자료에 대한 체계적인 관리

□ CANCEIS 적용시 문제점

○ 원시자료의 CANCEIS System에 맞도록 가공 수정

- 원시자료를 그대로 이용하는 것이 아니고 STRATA(가구원수), GROUP(지역)에 의해 자료를 나누어 임퓨테이션
- 원자료에서 공백에 대하여 별도의 값을 부여하여 임퓨테이션하여야 함

○ DLT 함수사용의 제약

- Rand, Max, Min, Sum, Avg, Count 등의 함수가 존재하나 Derive DLT에서만 사용가능하고 Mod 등 기타 함수사용이 제한됨

○ EDIT의 경우 불일치 자료에 대하여 곧 바로 임퓨테이션하나 이상치이며 유효한 값에 해당하는 경우는 도너에서 배제될 뿐 Edit 기능이 없음

2. 향후 추진 방향

□ 2010 인구주택총조사의 Edit & Imputation 시스템 구축

○ 2005 인구주택총조사의 경우 Edit는 E-census내에 구축되었고 Imputation은 시스템이 아닌 PC에서 SAS 프로그램을 통하여 임퓨테이션 후 원 자료에 반영함

○ 2010 인구주택총조사는 Edit & Imputation 시스템을 구축하여 효율적인 내검 체계와 무응답에 대한 체계적인 관리와 정도 높은 자료 생산

□ Edit & Imputation 시스템의 방향

○ Editing 시스템의 이원화

- 이상치이며 유효한 값은 오류의 가능성을 가지고 있으므로 내검요원을 통한 내검후 도너에서 제외(예, 100세이상인 나이)
- 불일치 자료의 경우 Deterministic Imputation을 먼저 한 후에 Donor Imputation

○ Deterministic Imputation의 별도 관리

- 3가지 이상의 변수가 연관되는 경우 2가지 이상 변수의 조건이 같으면 다른 변수는 자동으로 수정되도록 함

○ 별도의 원자료 가공없이 시스템상의 가상공간에서의 작업 기능

- 가구원수별 또는 지역별 분리, 공백자료의 임의값 삽입 없이 시스템상의 가상공간에서 작업 후 임퓨테이션

○ 체계적인 임퓨테이션 과정 및 결과 자료 관리

- 임퓨테이션 과정의 흐름을 볼 수 있는 자료의 생성과 에디팅 및 임퓨테이션 된 자료의 연계 자료의 생성

○ Editing의 체계적인 관리

- Editing의 순서, 기능별 구성과 현황에 대한 관리
- Editing rule의 문법적 오류 체크 기능

○ 우리나라 시스템에 맞은 거리함수 및 이론적 배경

- 개발원의 연구과제로 선정하여 이론적 배경의 뒷받침 필요
- 변수의 중요도에 따른 임퓨테이션 순서, CHAID 분석을 통한 보조변수의 선정
- Near Minimum Change Imputation의 거리함수와 도너선정

「무응답처리기법 국제공동연구」

해외출장 결과보고서〔Ⅱ〕

Banff를 이용한 광공업 동태자료 분석

2008. 1.

통계개발원 연구기획실 : 최 필근
강원지방청 경제조사과 : 임 영일
경제통계국 분석통계과 : 최 지은

차 례

Ⅰ. 연구개요 ····· 1

1.1. 연구 필요성 ····· 1

1.2. 연구 목표 ····· 1

1.3. 연구 범위 ····· 1

1.4. 연구의 한계점 ····· 2

Ⅱ. 연구에 사용된 자료 및 내검규칙(Edit Rule) ····· 2

2.1 광공업 동태조사 자료 ····· 2

2.2 내검규칙 ····· 3

Ⅲ. Banff를 이용한 광공업 동태자료 적용 ····· 4

3.1. Banff 시스템 ····· 4

3.2. 각 단계별 내용 및 적용결과 ····· 4

3.2.1. Edit Specification and Analysis (Proc Verifyedits) ····· 4

3.2.2. Edit Summary Statistics Tables (Proc Editstats) ····· 8

3.2.3. Outlier Detection (Proc Outlier) ····· 11

3.2.4. Error Localization (Proc Errorloc) ····· 15

3.2.5. Deterministic Imputation (Proc Deterministic) ····· 19

3.2.6. Donor Imputation (Proc Donorimputation) ····· 21

3.2.7. Estimator Imputation (Proc Estimator) ····· 26

3.2.8. Pro-Rating (Proc Prorate) ····· 31

3.2.9. Mass Imputation (Proc Massimputation) ····· 33

Ⅳ. 결론 및 향후 연구방향 ····· 37

4.1. 결론 ····· 37

4.2. 향후 연구방향 ····· 38

Ⅰ. 연구개요

1.1. 연구 필요성

사생활 보호 의식, 개인주의 확대, 기업비밀 보호의식의 강화로 소득, 매출액 등 통계조사 응답을 거부하는 사례가 증가하고 있으며, 특히 사업체의 경우 잦은 조사로 인하여 응답에 대한 불만 및 기피현상이 심화되고 있다. 따라서 사업체 조사에 대한 무응답처리 연구가 진행되어야 하며, 이를 위해 통계 선진국의 무응답처리 연구를 파악하고 우리나라 사업체 조사에 대한 적용가능성을 검토할 필요가 있다.

1.2. 연구 목표

본 연구는 통계 선진국 중의 하나인 캐나다 통계청에서 경제관련조사(economic survey)에서 쓰기 위해 개발된 연속형 자료의 무응답대체 시스템인 Banff 시스템을 파악하고 이 시스템을 광공업 동태조사 자료에 적용시켜봄으로써 통계청에서 취급하는 다른 사업체 조사에서의 적용가능성을 검토하는 것을 목표로 한다.

1.3. 연구 범위

본 연구의 근본 취지는 Banff 시스템 파악과 광공업 동태조사 자료에의 적용이다. 따라서 Banff 시스템의 각 절차에 대하여 알아보고, 이러한 절차에 맞춰 광공업 동태조사에 적용하여 결과를 보일 것이다. 그러나 자료의 특성에 따라서 적용되는 구체적인 방법이 다르므로 광공업 동태조사에 맞는 방법으로 분석을 할 것이다. Banff 시스템의 각 절차 중에서 자료의 특성에 따라 고려되어져야 하는 것들은 내검규칙, 이상값 체크방법, 대체군 개발, 추정식 개발 등으로써 광공업 동태조사에 적합한 것을 제시하고자 한다. 그러나 이러한 내용들은 향후 깊이 있는 연구가 되어져야 하는 것들임을 알려둔다.

1.4. 연구의 한계점

모든 연구가 그렇듯이 본 연구도 몇 가지 한계점을 가지고 있다. 이러한 한계점에 대한 것들을 명확히 밝힘으로써 앞으로의 유사한 연구의 계획과 진행에 반영될 수 있고자 한다.

(1) 연구 기간의 한계 : Banff 시스템에 대한 파악 및 연구가 3주정도 진행이 되었기 때문에 상대적으로 충분한 연구가 이루어지지 못한 상태에서 본 결과 보고서가 작성되었다. 실제 캐나다의 경우에서도 수년간 이상의 연구를 통해 계속 보완해 나가는 실정이며 본 결과 보고서를 바탕으로 사업체 조사에 대한 무응답대체 방법의 연구가 계속적으로 진행이 되어야 할 것이다.

(2) 동태자료의 한계 : 본 연구는 통계청의 광공업 동태조사 자료를 이용해 분석을 하였다. 실제 동태조사의 경우는 자료의 구조가 단순하며 특성이 명확하기 때문에 연구에 한계가 있을 수가 있다. 따라서 향후 연간조사나 경제센서스 등의 조사에 대해 Banff 시스템의 효율성을 파악할 수 있는 연구가 필요할 것이다.

(3) 무응답 자료의 한계 : 통계청의 사업체 조사들은 무응답을 허용하지 않기 때문에 실제적인 무응답이 거의 존재하지 않는다. 그러나 완전한 자료 중에는 조사담당자에 의해서 경험적으로 대체된 것들이 존재한다. 실제 무응답처리에 관한 연구를 위해서는 경험적으로 대체된 자료의 구분이 이루어져야 할 것으로 본다. 향후 이러한 자료를 가지고 세부적인 연구가 진행되어야 하겠다.

Ⅱ. 연구에 사용된 자료 및 내검규칙(Edit Rule)

2.1 광공업 동태조사 자료

본 연구에서는 2007년 8월 광공업 동태조사 자료를 가지고 분석을 하였다. 분석

을 위한 보조 자료로 동년전월자료(2007년 7월)와 전년동월자료(2006년 8월)를 이용하였다. 또한 더 정확한 대체를 하기 위하여 이미 조사되어져 있는 종업원 수와 월별 근무일수 변수를 추가로 이용을 하였다. 자료에 대한 내용은 다음과 같다.

[분석에 사용된 광공업 동태조사 자료]

SAUPID	사업체 id	JT_G	재투입(금액)
PUMID	품목 id	DO_S	국내시판(수량)
ENT_GU	기업규모	DO_G	국내시판(금액)
SIDOID	시도 id	EXP_S	수출(수량)
JSPRO_S	자체생산(수량)	EXP_G	수출(금액)
JSPRO_G	자체생산(금액)	ETC_S	기타출하(수량)
WSPRO_S	위탁생산(수량)	ETC_G	기타출하(금액)
WSPRO_G	위탁생산(금액)	GABJ_S	과부족(수량)
SSPRO_S	수탁생산(수량)	GABJ_G	과부족(금액)
SSPRO_G	수탁생산(금액)	INV_S	월말재고(수량)
GUIP_S	구입(수량)	INV_G	월말재고(금액)
GUIP_G	구입(금액)	BINV_S	전월재고(수량)
JT_S	재투입(수량)	BINV_G	전월재고(금액)

*추가자료: PRODEM(종업원수), WORKDAY(근무일수)

2.2 내검규칙

광공업 동태자료의 내검규칙은 간단하게 정의되어 진다. SAUPID, PUMID, ENT_GU, SIDOID, GABJ_S, GABJ_G을 제외한 모든 항목은 0이상이 되어야 한다. 그리고 INV_S과 INV_G의 합계를 맞추기 의한 2개의 내검규칙을 추가하면 된다. 즉,

BINV_S + JSPRO_S + WSPRO_S + SSPRO_S + GUIP_S
- JT_S - DO_S - EXP_S - ETC_S + GABJ_S = INV_S 와

BINV_G + JSPRO_G + WSPRO_G + SSPRO_G + GUIP_G
- JT_G - DO_G - EXP_G - ETC_G + GABJ_G = INV_G 이다.

따라서 총 22개의 내검규칙을 가지고 분석에 이용을 할 것이다.

Ⅲ. Banff를 이용한 광공업 동태자료 적용

3.1. Banff 시스템

Banff 시스템은 2002년에 economic survey에서 쓰이기 위해 개발된 것으로 연속형 변수의 처리를 위해 디자인되어 있으며, 자동화된 데이터의 에디팅과 imputation을 가능하게 하며 9개의 SAS 프로시저로 구성되어 있다. 각 프로시저는 독립적으로 활용 가능하며 또한 필요에 따라 복합적으로도 적용 가능하다. 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.

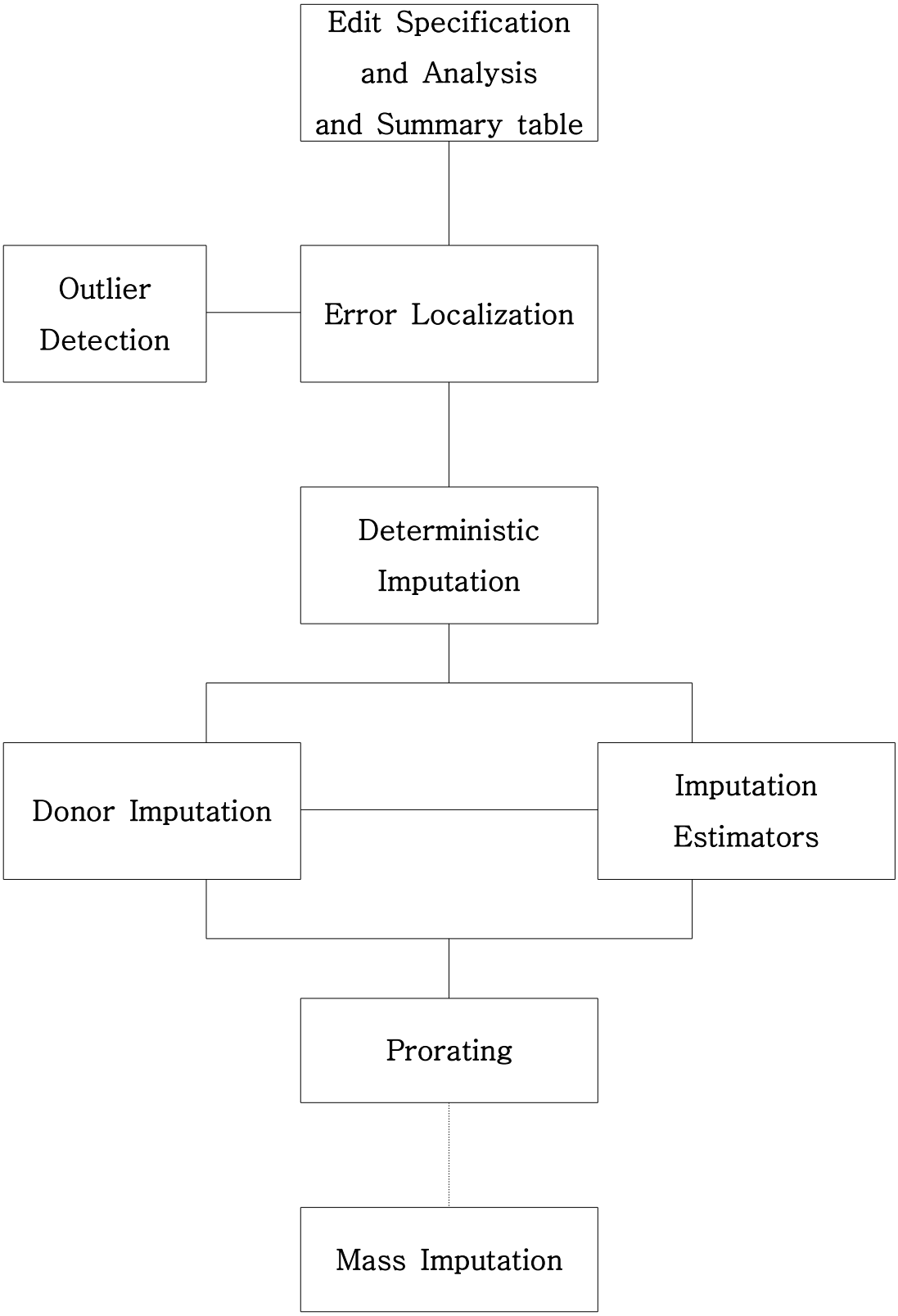
Banff는 edit, error localization, imputation system으로 구성되어 있다. Editing의 목적은 특정한 record의 수용여부, 일관성여부, outlier여부, missing여부 등을 결정하는 데에 있다. Imputation은 결측치와 오류값을 적당하고 일관성 있는 값으로 대체하는 것이다. 이 두 가지 기능은 수정의 회수를 최소화하기 위해서 대체되어야 할 record의 field를 결정하는 errror localization에 의해 연결된다. 다음의 [그림3-1]은 Banff 시스템의 모든 절차를 나타낸다.

3.2. 각 단계별 내용 및 적용결과

3.2.1. Edit Specification and Analysis (Proc Verifyedits)

1) 개요

이 과정은 데이터분석을 위한 기초 단계로 쓰일 수 있으며 각 data field간의 상관관계를 알아보기 위한 과정이다. 이러한 상관관계를 edits라고 한다면 이것은 설문지 또는 데이터의 분석을 통해서 알 수 있을 것이다. 이러한 과정을 데이터에 존재하는 제약조건들을 edits가 정확히 나타내고 있는지 확인하는데 도움을 준다.



[그림3-1] Banff 시스템의 절차

Banff의 edits rule은 반드시 linear form으로 표현되어야 하며 데이터는 numeric, non-negative, continuous 해야 한다. 따라서 $x_1,x_2,\ldots,x_n$ 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.

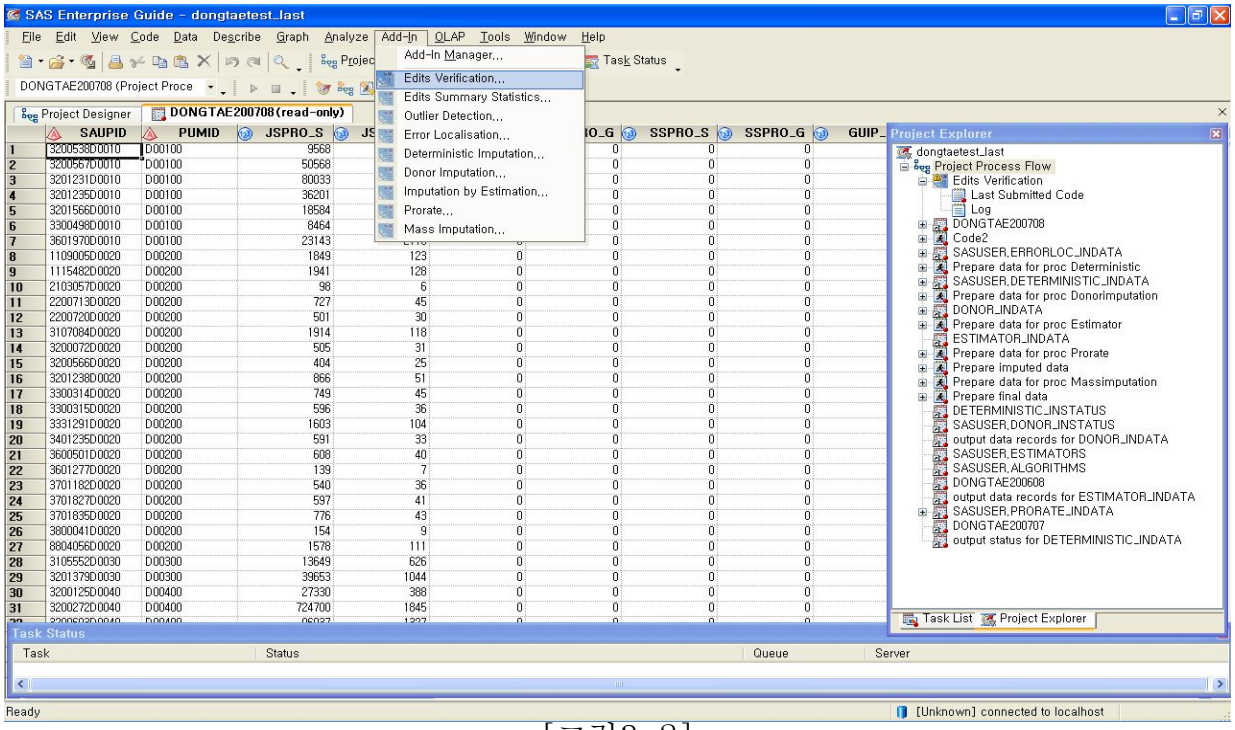
2) 절차

- 상단의 Add-In menu에서 Edit Verification task를 선택한다. Edit 창에서 Edit을 입력한다.
- Options창에서 이 절차에 의해 보고자 하는 함축된(implied) edit의 수를 설정한다. Acceptnegative 옵션을 체크하지 않으면 var>0 인 positivity edits가 적용된다. Acceptnegative 선택하면 수동적으로 positivity edit을 입력해야 한다.
- Acceptnegative 옵션이 체크되지 않은 경우 maximum cardinality (number of non-zero coordinates) 옵션을 설정할 수 있는데, 이는 edits에 의해 규정될 가능한 영역(feasible region)의 극점(extremal points)의 최대값을 의미한다. 변수가 10개일 경우 maximum도 10이 된다.
- 하단의 preview code를 클릭하여 SAS code를 볼 수 있다.

3) 적용

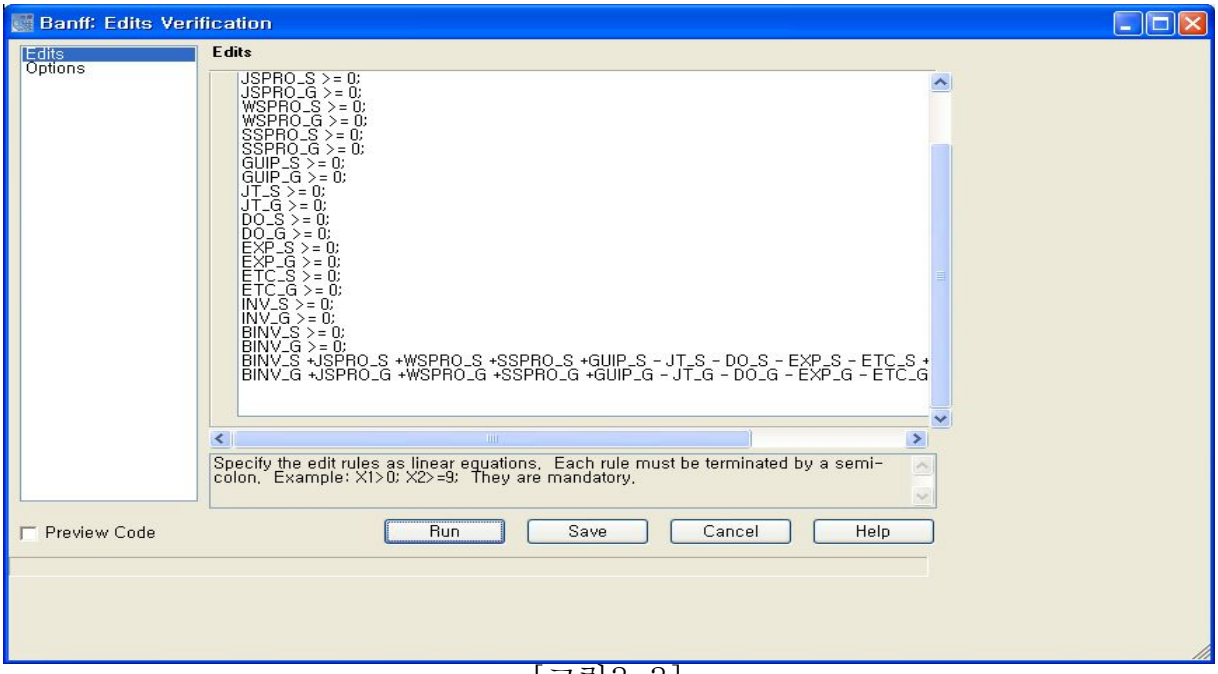
광공업 동태자료의 edit rule을 확인하기 위하여, [그림3-2]와 같이 Banff시스템과 연계된 SAS Enterprise Guide의 Add-In 메뉴로부터 Edits Verification을 선

택한다.



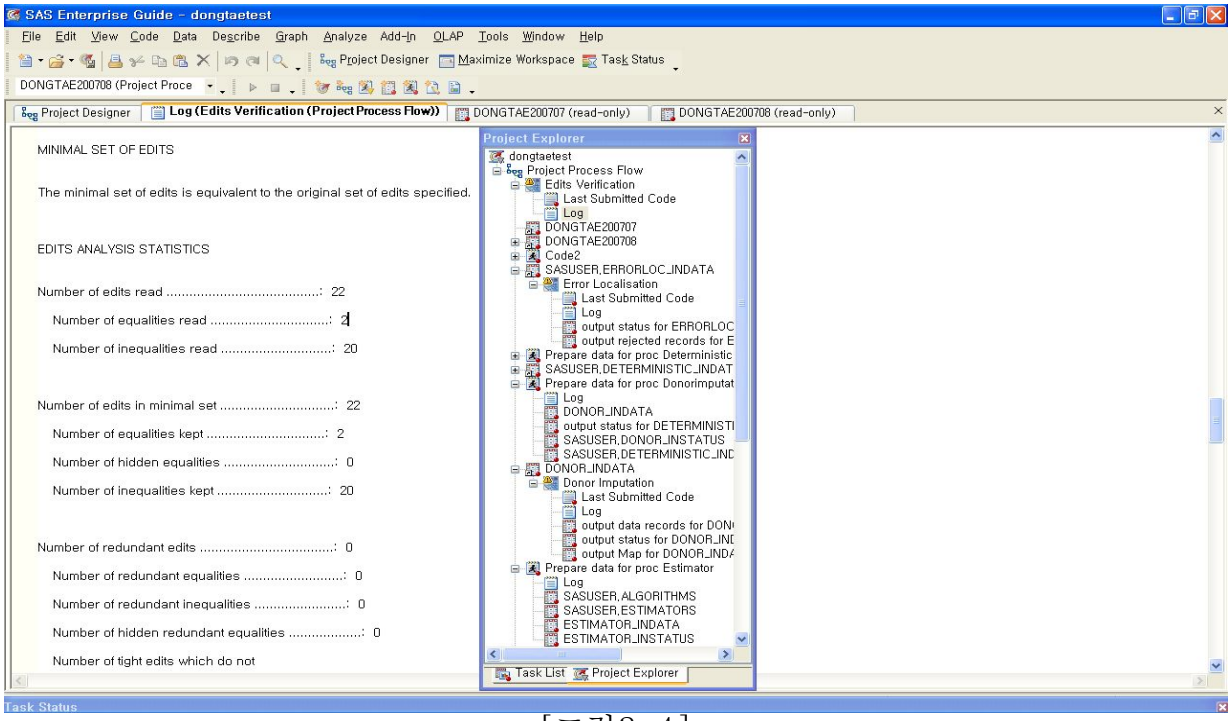
[그림3-2]

광공업 동태자료의 edit rule을 [그림3-3]과 같이 입력한 후 Edits Verification 과정을 실행하여 edit rule에 대한 점검과 향후 사용될 edits rule의 minimal set을 결정한다. 이는 중복으로 사용되는 규칙을 제거하여 효율적인 editing을 하기 위함이다.



[그림3-3]

Edits Verification과정을 실행한 결과를 보면 [그림3-4]와 같다. 총 edit rule의 수는 22개가 입력되었고 불필요한 edit rule은 없으며 향후 이 edit rule들을 사용하면 될 것이라는 것을 보여준다. 만약, 불필요한 edit rule이 있으면 이를 제거한 후 사용하면 될 것이다.



[그림3-4]

3.2.2. Edit Summary Statistics Tables (Proc Editstats)

1) 개요

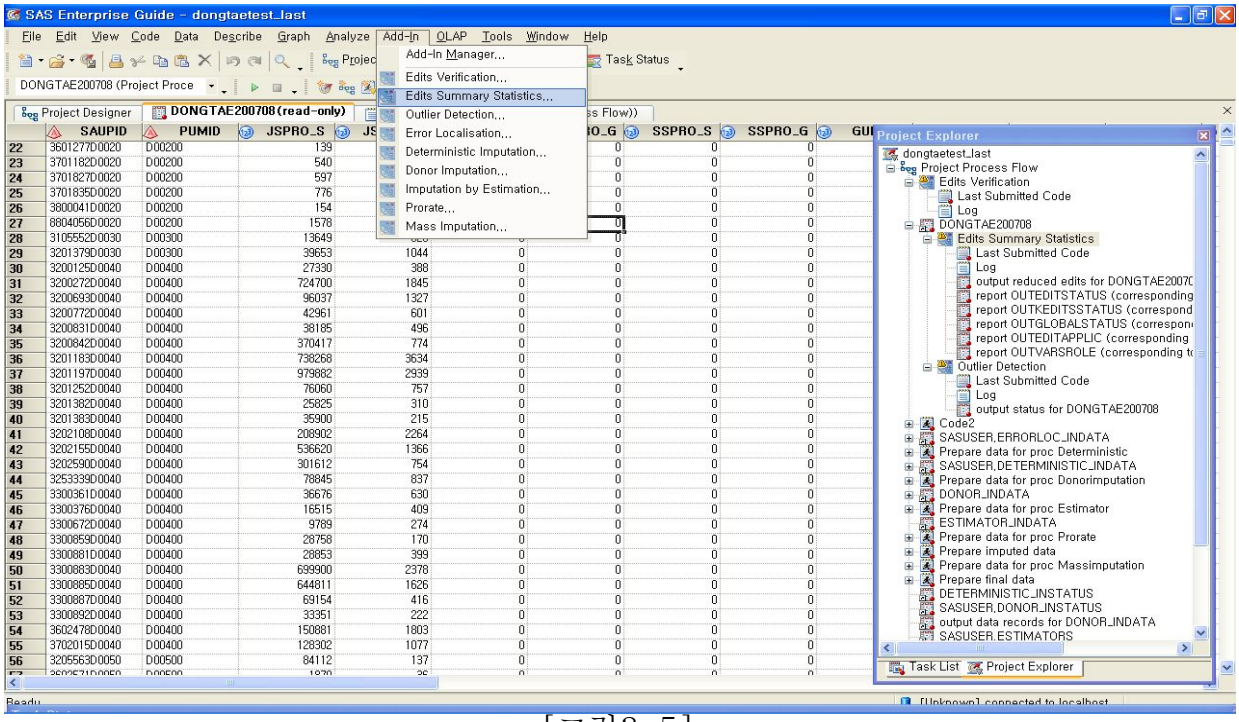
Verifyedits과정은 데이터와 상관없이 linear edits 자체만을 고려하여 equation system의 consistency 여부를 확인하는 과정이라고 한다면 Editstats는 결정된 edits system이 실제 데이터에 적용하여 각각의 record가 pass, miss, fail인지를 결정하는 절차이다. 여기서 “pass"는 모든 edit rule이 성공했을 때를 의미하고, "miss" 는 무응답 데이터에 의해 edit가 실패했을 때를 의미하고 "fail"은 하나 이상의 non-missing value에 의한 editing 실패를 나타낸다. 만약 어느 특정 edit rule에서 editing 실패율이 높게 나타났다면 사용자는 edit rule을 수정하는 등의 조치를 취해야 할 것이다.

2) 절차

- Verifyedits 절차에서 얻은 edits를 데이타에 적용하기 위해 Add-In menu에서 Edits Summary Statistics task를 연다.
- Task roles창에서 BY-group을 설정한다.
- Options 창에서 첫 번째 절차와 같이 acceptnegative 선택 여부를 결정한다.
- Edits 창에 minimal set edit rule을 입력한다.
- 하단의 preview code를 클릭하여 SAS code를 볼 수 있다.

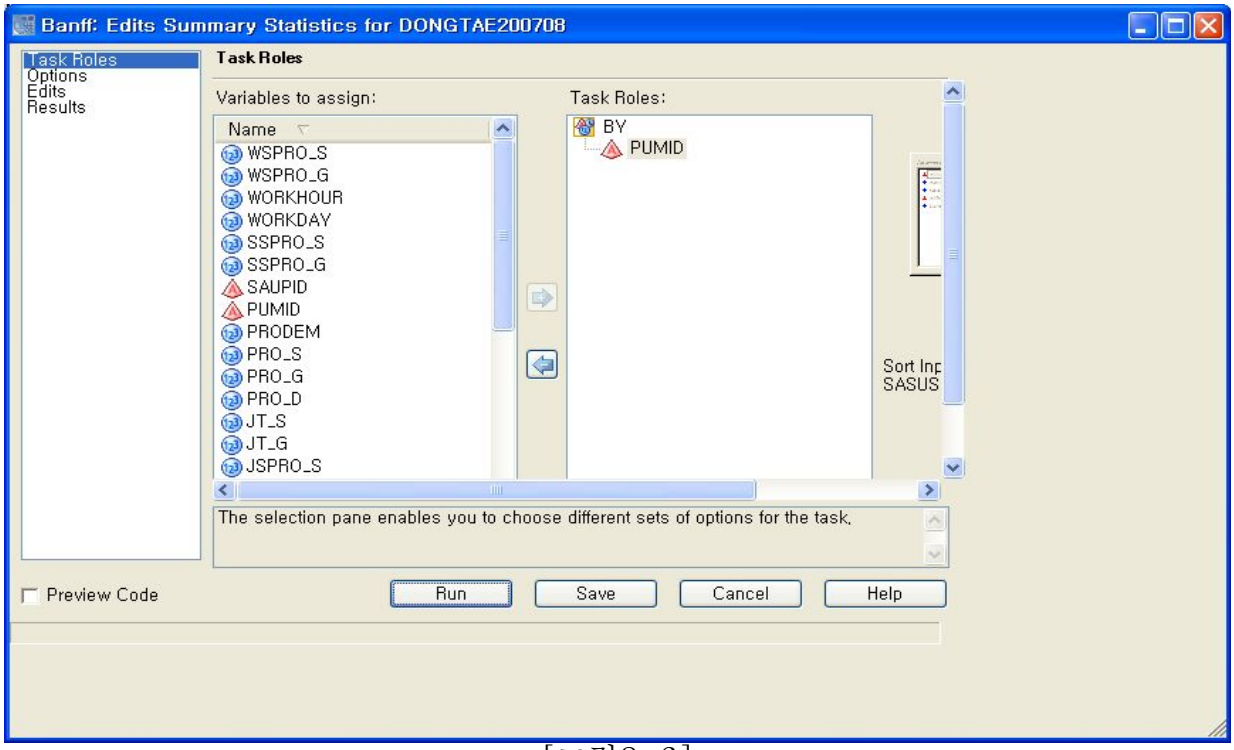
3) 적용

Edits Verification과정에서 결정된 22개의 edit rule을 사용하여 각 개체들의 edit rule의 만족여부의 결과를 볼 수 있다. [그림3-5]와 같이 SAS Enterprise Guide의 Add-In 메뉴로부터 Edits Summary Statistics를 선택한다.



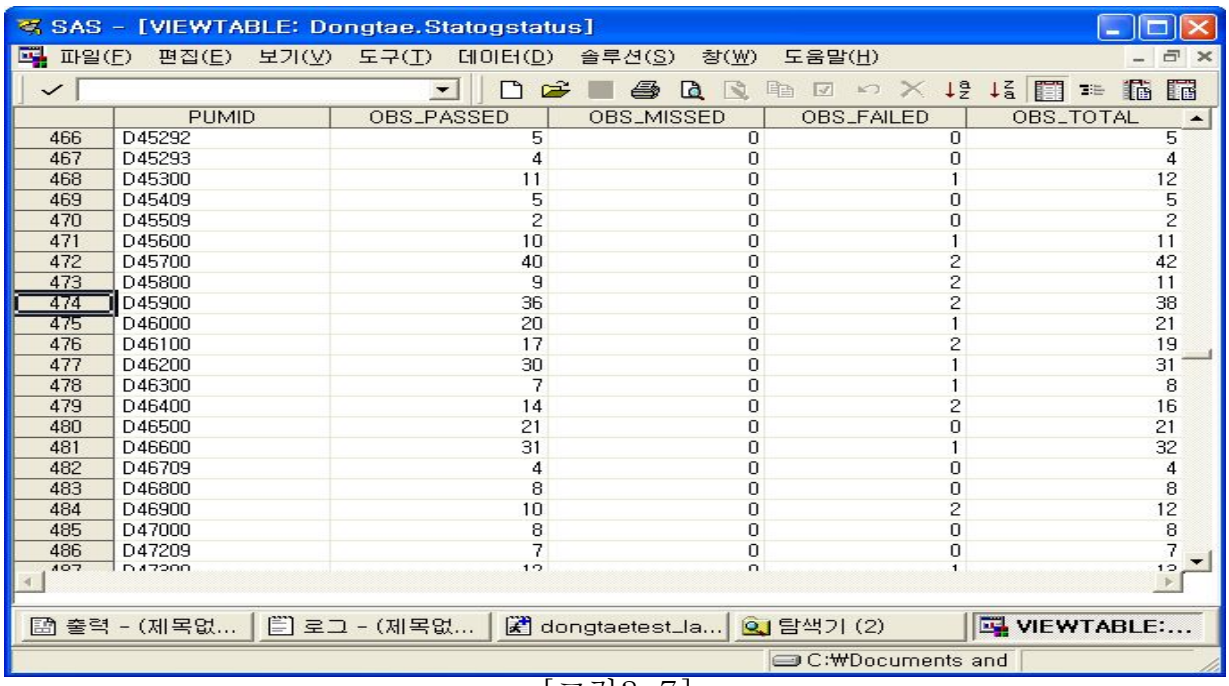
[그림3-5]

모든 과정을 수행함에 있어 비슷한 성향의 것을 묶어서 분석과 적용을 하는 것이 용이하다. 따라서 우리의 자료에서는 [그림3-6]과 같이 품목에 대하여 나누어 분석 및 imputation을 실시할 것이다. 이는 자료에 따라 지역이나 종업원수(기업규모) 등의 변수들로 나누어 분석을 할 수 있을 것이다. Edit rule은 앞에서 사용한 것을 그대로 쓰면 된다.



[그림3-6]

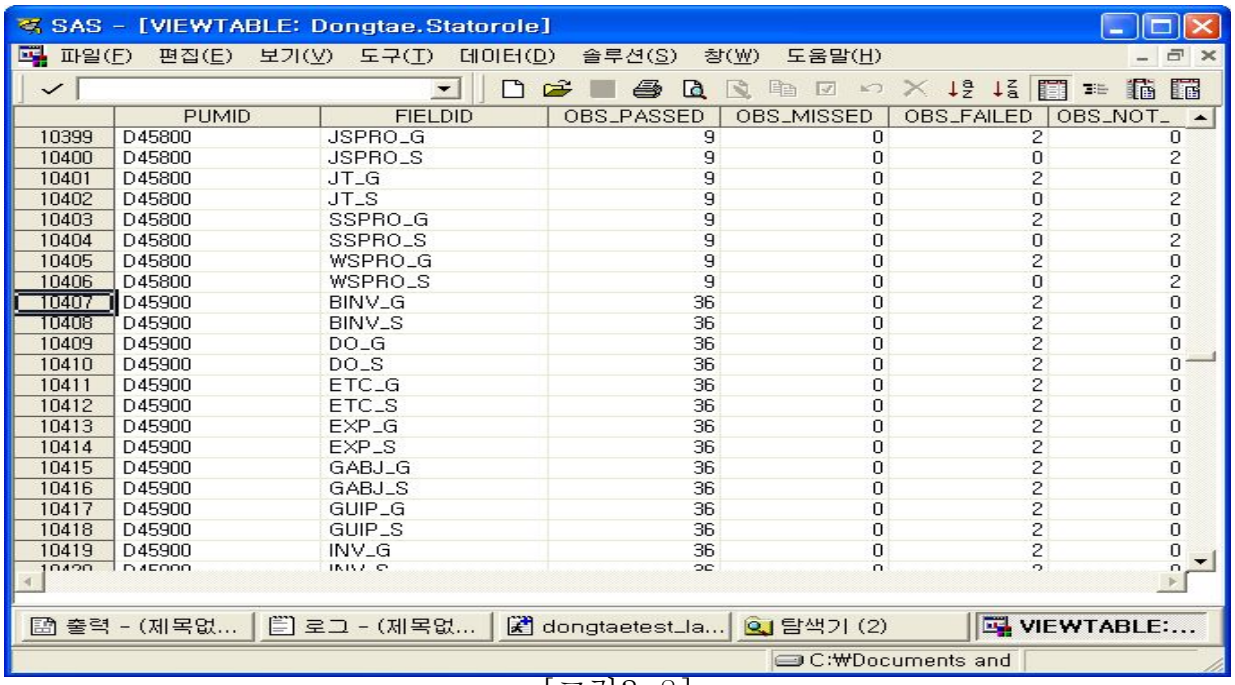
[그림3-7]은 각 품목들에 대하여 edit rule이 pass된 것과 fail된 것의 개체 수를 나타낸 결과이다. 예를 들어, 품목 D45900의 경우 총 38개의 개체중에서 36개는 모든 edit rule을 만족하지만 2개의 개체는 만족하지 않는 edit rule이 있음을 알 수 있다.



	PUMID	OBS_PASSED	OBS_MISSED	OBS_FAILED	OBS_TOTAL
466	D45292	5	0	0	5
467	D45293	4	0	0	4
468	D45300	11	0	1	12
469	D45409	5	0	0	5
470	D45509	2	0	0	2
471	D45600	10	0	1	11
472	D45700	40	0	2	42
473	D45800	9	0	2	11
474	D45900	36	0	2	38
475	D46000	20	0	1	21
476	D46100	17	0	2	19
477	D46200	30	0	1	31
478	D46300	7	0	1	8
479	D46400	14	0	2	16
480	D46500	21	0	0	21
481	D46600	31	0	1	32
482	D46709	4	0	0	4
483	D46800	8	0	0	8
484	D46900	10	0	2	12
485	D47000	8	0	0	8
486	D47209	7	0	0	7

[그림3-7]

[그림3-8]의 결과는 각 품목에 대하여 각 변수에서의 edit rule의 만족도를 보여준다. 앞의 결과와 연관을 지어서 품목 D45900의 경우, 2개체가 edit rule을 만족하지 않는데 BINV_G, BINV_S, DO_G, DO_S, ETC_G 등 대부분의 변수에서 만족하지 않고 있다는 것을 보여준다. 이는 대부분의 항목에서 무응답을 한 것으로 광공업 동태자료의 경우 무응답일 경우 -1값을 부여하기 때문에 각 항목이 0 이상이 되어야하는 edit rule을 만족하지 못하므로 이러한 결과가 나오는 것이다.


SAS - [VIEWTABLE: Dongtae.Statorole]

파일(F) 편집(E) 보기(V) 도구(T) 데이터(D) 솔루션(S) 창(W) 도움말(H)

	PUMID	FIELDID	OBS_PASSED	OBS_MISSED	OBS_FAILED	OBS_NOT_
10399	D45800	JSPRO_G	9	0	2	0
10400	D45800	JSPRO_S	9	0	0	2
10401	D45800	JT_G	9	0	2	0
10402	D45800	JT_S	9	0	0	2
10403	D45800	SSPRO_G	9	0	2	0
10404	D45800	SSPRO_S	9	0	0	2
10405	D45800	WSPRO_G	9	0	2	0
10406	D45800	WSPRO_S	9	0	0	2
10407	D45900	BINV_G	36	0	2	0
10408	D45900	BINV_S	36	0	2	0
10409	D45900	DO_G	36	0	2	0
10410	D45900	DO_S	36	0	2	0
10411	D45900	ETC_G	36	0	2	0
10412	D45900	ETC_S	36	0	2	0
10413	D45900	EXP_G	36	0	2	0
10414	D45900	EXP_S	36	0	2	0
10415	D45900	GABJ_G	36	0	2	0
10416	D45900	GABJ_S	36	0	2	0
10417	D45900	GUIP_G	36	0	2	0
10418	D45900	GUIP_S	36	0	2	0
10419	D45900	INV_G	36	0	2	0

출력 - (제목없... 로그 - (제목없... dongtaetest_la... 탐색기 (2) VIEWTABLE:...

C:₩Documents and

[그림3-8]

3.2.3. Outlier Detection (Proc Outlier)

1) 개요

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 사용가능하다면 Ratio method를 사용하는 것이 좋고, 특히 사용 가능한 보조변수가 과

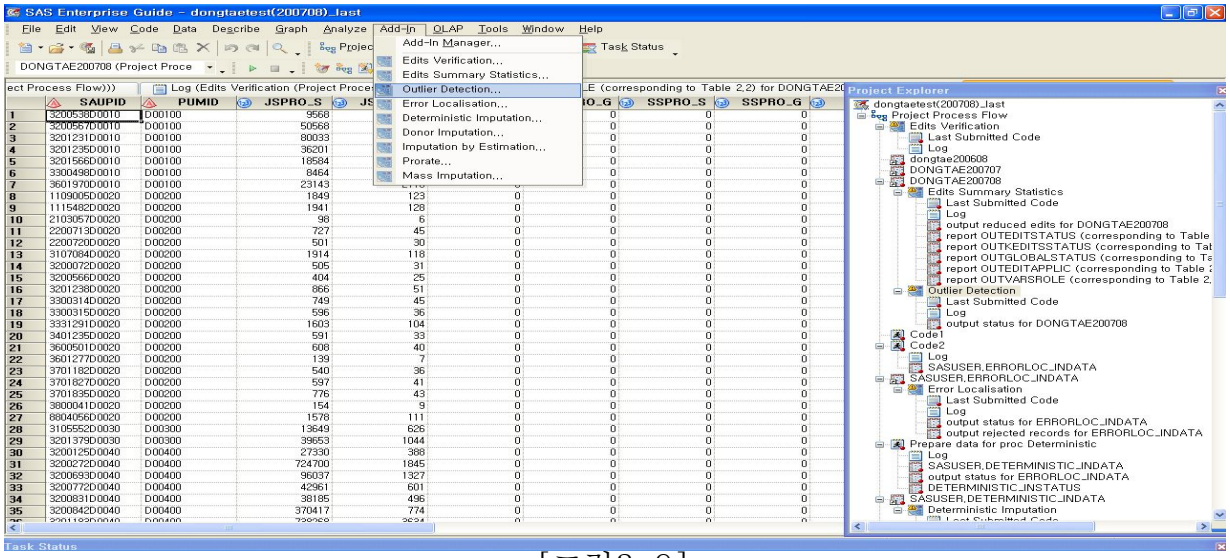
거의 자료에 기초한 것이라면 Historical Tend method를 사용하는 것이 좋다. 즉, Historical Trend method는 Ratio method의 특수한 형태라고 할 수 있다. 사용 가능한 보조 변수가 없을 때에는 Current method를 사용해야 하는데 이것은 다른 데이터를 사용해서 얻은 값의 범위와 비교하여 이상치를 판별하는 방법이다. 또한 이상치를 판별하는 기준이 되는 bound는 데이터의 parameter의 함수로 나타낼 수 있는데, 세 가지 방법 모두 사용자가 이러한 모수를 사용해서 outlier intervals의 조정이 가능하다.

2) 절차

- 상단의 Add-In menu에서 Outlier Detection task를 선택한다.
- Method창에서 data set 유형을 결정하여 outlier를 진단하기 위한 방법을 결정한다.
- Task roles창에서 id 에 key variable을, by에는 sort될 변수를, var에는 outlier를 진단하고자 하는 변수를 넣는다.
- Options창에서 MII (imputation interval multiplier), MEI(exclusion interval multiplier)를 입력한다. 음의 값이 있을 경우에 acceptnegative를 선택한다.
- Other inputs창에서 과거자료 등의 보조자료를 불러온다.

3) 적용

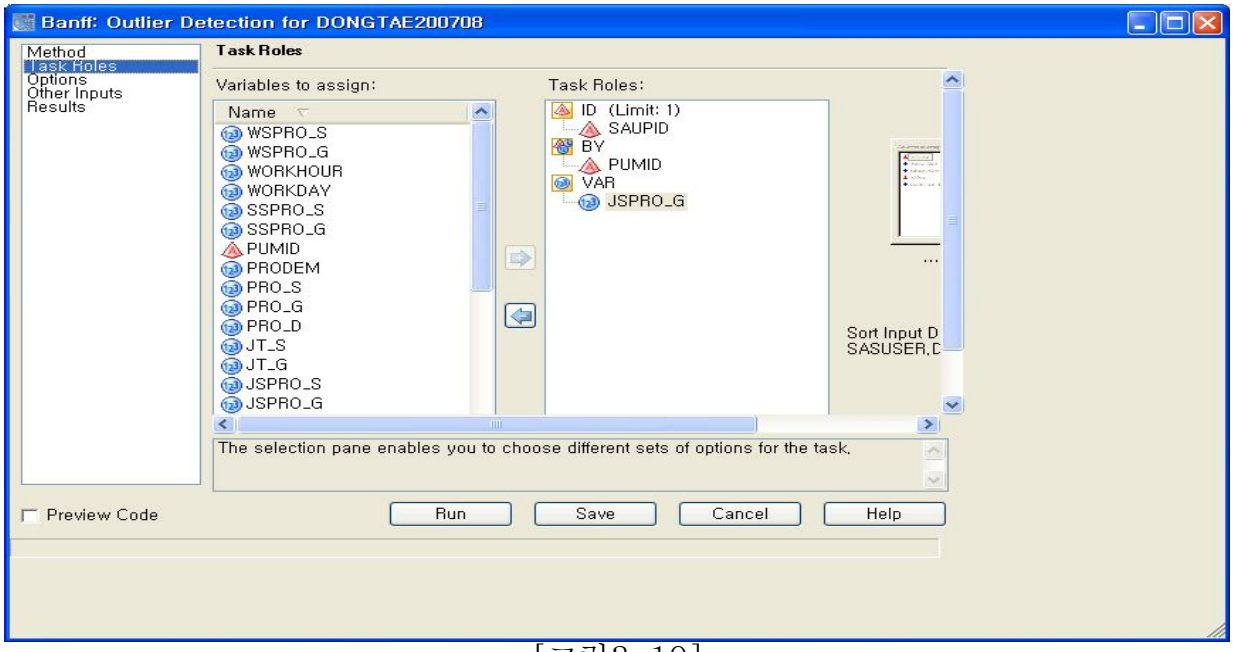
각 변수들에 대한 이상값 체크는 [그림3-9]와 같이 SAS Enterprise Guide의 Add-In 메뉴로부터 Outlier Detection을 선택하여 사용할 수 있다.



[그림3-9]

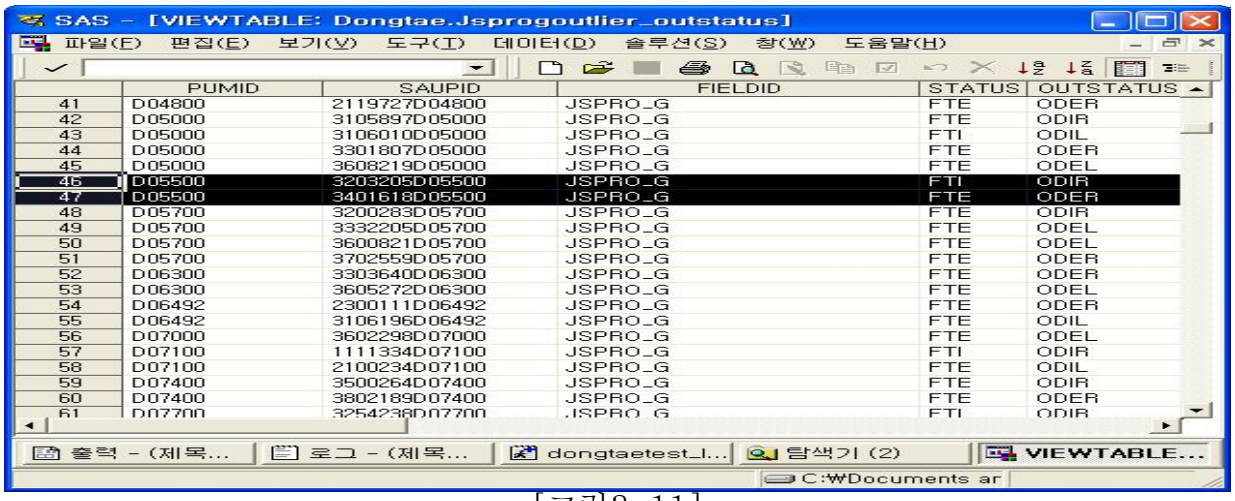
[그림3-10]과 같이 이상값 체크를 하기 원하는 변수 JSPRO_G를 선택한다. 여기서 JSPRO_G 아래에 많은 변수들을 선택하여 일괄적으로 이상값 체크를 실시할 수도 있지만 본 연구에서는 모든 변수들의 분포가 같지 않을 것을 고려해 각각의 변수를 다른 규칙에 의하여 이상값 체크를 실시하였다. 이상값의 기준은 정확하게 하나로 정해져 있지는 않지만 여러 이론에 의하여 비슷하게 정의되어진다. 실제 광공업 동태자료의 이상값 체크는 본 자료의 특성을 잘 알고 있는 실무경험자에 의하여 체크 되어질 수가 있을 것이다. 그러나 통계적인 이론과 함께 한다면 더 좋은 분석이 될 것이다. 본 연구에서는 다음의 규칙에 의하여 JSPRO_G의 이상값을 체크하였다.

(1) 현재자료의 값과 전월자료의 값의 비(ratio)를 사용한다.
(2) 이 비의 분포로부터 양쪽 끝의 총 3%를 이상값으로 1차 선택을 한다. 이상값으로 3%를 정한 것은 차후 정밀한 검토가 필요할 것으로 보인다.
(3) 또한 이상값은 아니지만 차후 대체를 위해서 이용될 자료로서는 불합리한 것으로 양쪽 끝의 총 3% 에서 5% 사이에 있는 값을 선택한다.
(4) 1차 선택이 된 이상값은 실제 계절요인에 의하여 비의 값이 차이가 날 수도 있을 것이다. 따라서 전년동월자료를 가지고 똑같은 방법으로 이상값을 2차 선택한다.
(5) 1차와 2차에서 동시에 이상값으로 선택된 것을 최종 이상값으로 선택한다. 이로서 계절성에 의한 것을 고려한 이상값이 채택되었다.



[그림3-10]

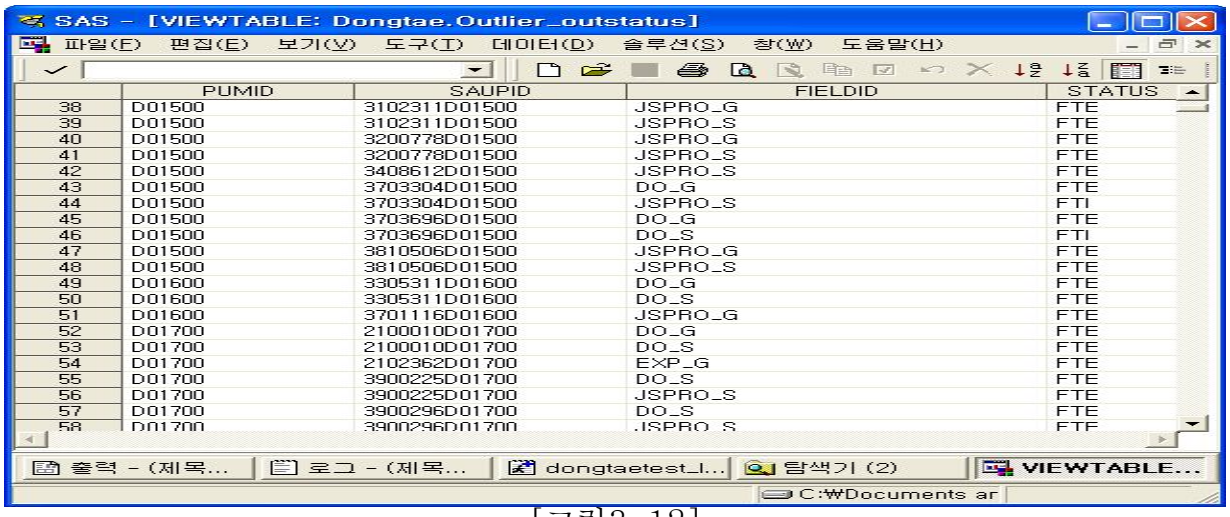
[그림3-11]은 각 사업체에 대하여 JSPRO_G의 이상값을 체크한 결과를 나타낸다. STATUS에서 FTI(Field To Impute)는 이상값으로 imputation을 필요로 한다는 의미이며, FTE(Field To Exclude)는 이상값에 가까운 값이기 때문에 대체를 위한 분석 자료로는 제외시키겠다는 것을 나타낸다. 예를 들어 품목 D05500을 생산하는 사업체 3203205D05500의 JSPRO_G 변수는 이상값으로 imputation을 실시할 것이며, 3401618D05500의 JSPRO_G 변수는 이상값에 가까운 값이기 때문에 대체를 위한 분석 자료로는 제외시킬 것이다.


SAS - [VIEWTABLE: Dongtae.Jsprogoutlier_outstatus]

	PUMID	SAUPID	FIELDID	STATUS	OUTSTATUS
41	D04800	2119727D04800	JSPRO_G	FTE	ODER
42	D05000	3105897D05000	JSPRO_G	FTE	ODIR
43	D05000	3106010D05000	JSPRO_G	FTI	ODIL
44	D05000	3301807D05000	JSPRO_G	FTE	ODER
45	D05000	3608219D05000	JSPRO_G	FTE	ODEL
46	D05500	3203205D05500	JSPRO_G	FTI	ODIR
47	D05500	3401618D05500	JSPRO_G	FTE	ODER
48	D05700	3200283D05700	JSPRO_G	FTE	ODIR
49	D05700	3332205D05700	JSPRO_G	FTE	ODEL
50	D05700	3600821D05700	JSPRO_G	FTE	ODEL
51	D05700	3702559D05700	JSPRO_G	FTE	ODER
52	D06300	3303640D06300	JSPRO_G	FTE	ODER
53	D06300	3605272D06300	JSPRO_G	FTE	ODEL
54	D06492	2300111D06492	JSPRO_G	FTE	ODER
55	D06492	3106196D06492	JSPRO_G	FTE	ODIL
56	D07000	3602298D07000	JSPRO_G	FTE	ODEL
57	D07100	1111334D07100	JSPRO_G	FTI	ODIR
58	D07100	2100234D07100	JSPRO_G	FTE	ODIL
59	D07400	3500264D07400	JSPRO_G	FTE	ODIR
60	D07400	3802189D07400	JSPRO_G	FTE	ODER
61	D07700	3254238D07700	JSPRO_G	FTI	ODIR

[그림3-11]

이상값을 체크하고자 하는 변수에 이와 같은 작업을 반복적으로 적용을 하면 된다. 본 연구에서는 JSPRO_G, JSPRO_S, DO_G, DO_S, EXP_G, EXP_S 변수에 대하여 이상값 체크를 실시하였다. 이상값을 체크했던 6개변수에서 나온 결과들을 하나로 통합한 결과가 [그림3-12]이다. [그림3-12]에 대한 설명은 앞의 것과 동일하다. 이 결과는 향후 imputation에서 이용될 것이다.


SAS - [VIEWTABLE: Dongtae.Outlier_outstatus]

	PUMID	SAUPID	FIELDID	STATUS
38	D01500	3102311D01500	JSPRO_G	FTE
39	D01500	3102311D01500	JSPRO_S	FTE
40	D01500	3200778D01500	JSPRO_G	FTE
41	D01500	3200778D01500	JSPRO_S	FTE
42	D01500	3408612D01500	JSPRO_S	FTE
43	D01500	3703304D01500	DO_G	FTE
44	D01500	3703304D01500	JSPRO_S	FTI
45	D01500	3703696D01500	DO_G	FTE
46	D01500	3703696D01500	DO_S	FTI
47	D01500	3810506D01500	JSPRO_G	FTE
48	D01500	3810506D01500	JSPRO_S	FTE
49	D01600	3305311D01600	DO_G	FTE
50	D01600	3305311D01600	DO_S	FTE
51	D01600	3701116D01600	JSPRO_G	FTE
52	D01700	2100010D01700	DO_G	FTE
53	D01700	2100010D01700	DO_S	FTE
54	D01700	2102362D01700	EXP_G	FTE
55	D01700	3900225D01700	DO_S	FTE
56	D01700	3900225D01700	JSPRO_S	FTE
57	D01700	3900296D01700	DO_S	FTE
58	D01700	3900296D01700	JSPRO_S	FTE

[그림3-12]

3.2.4. Error Localization (Proc Errorloc)

1) 개요

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 field 수가 이를 초과하게 되면 Banff는 답을 제공하지 않고 사용자가 지정한 수를 변화시킬 것을 알린다. 또한 더 믿을만한 변수에 높은 가중치를 줌으로써 이 변수가 imputation을 하기 위해 선택되지 않을 가능성을 높일 수 있다.

2) 절차

- 상단의 Add-In menu에서 Error Localization task를 선택한다.
- Outlier detection에 의한 outlying field를 missing 처리 한다.
- Edits 창에 가중치 적용을 할 경우 각 변수의 가중치를 입력한다.
- **가중치 적용사례**

	A	+	B	+	C	=	TOT
weight	2		2		2		5
values	.		290		20		300

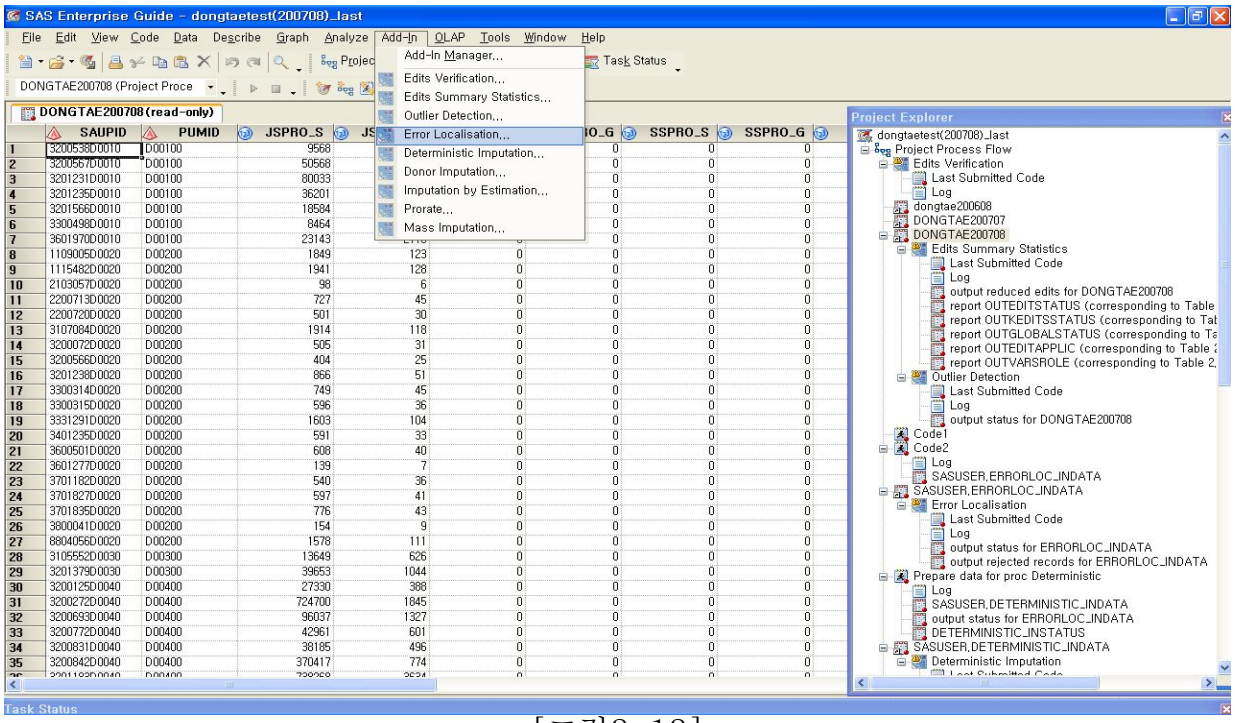
만약, TOT변수가 다른 변수보다 더 믿을 만한 변수라고 가정한다면 이 변수에 대하여 가중치를 가장 크게 적용한다. 이는 다른 어떤 두 개의 변수의 합보다 더 큰 가중치를 두었으나, 세 변수의 합보다는 작은 가중치이다. 이 의미는 두 변수 보다는 TOT가 더 믿을만 하지만, 세 변수 모두는 합계변수 TOT보다 더 믿을만

하다는 것을 의미한다. 예제의 경우, missing값을 대체해야 하고 적어도 다른 변수 하나를 대체해야 하는 경우인데 TOT보다 가중치가 작은 다른 두 개의 변수 중 하나를 랜덤으로 선택하여 대체를 하게 된다.

- Outlier로 진단된 것과 가중치 방법에 의해 error로 판정된 것은 proc errorloc 에서 FTI로 표시된다.

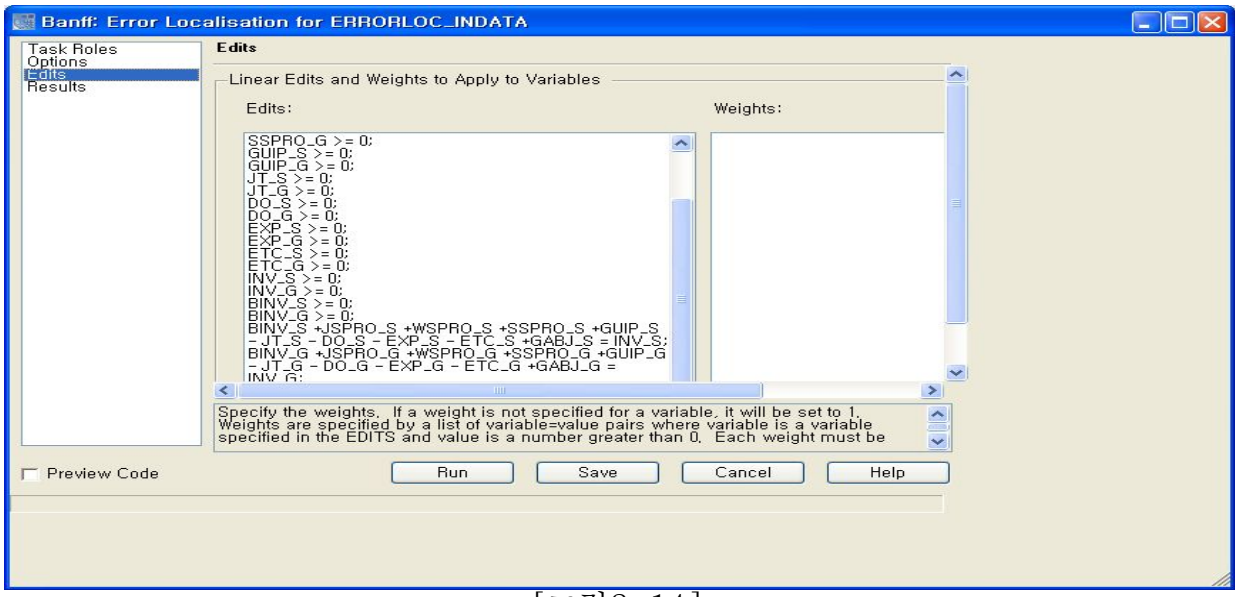
3) 적용

Edit rule을 만족시키지 못해서 수정해야하는 항목의 결정은 [그림3-13]과 같이 SAS Enterprise Guide의 Add-In 메뉴로부터 Error Localization을 선택하여 사용할 수 있다.



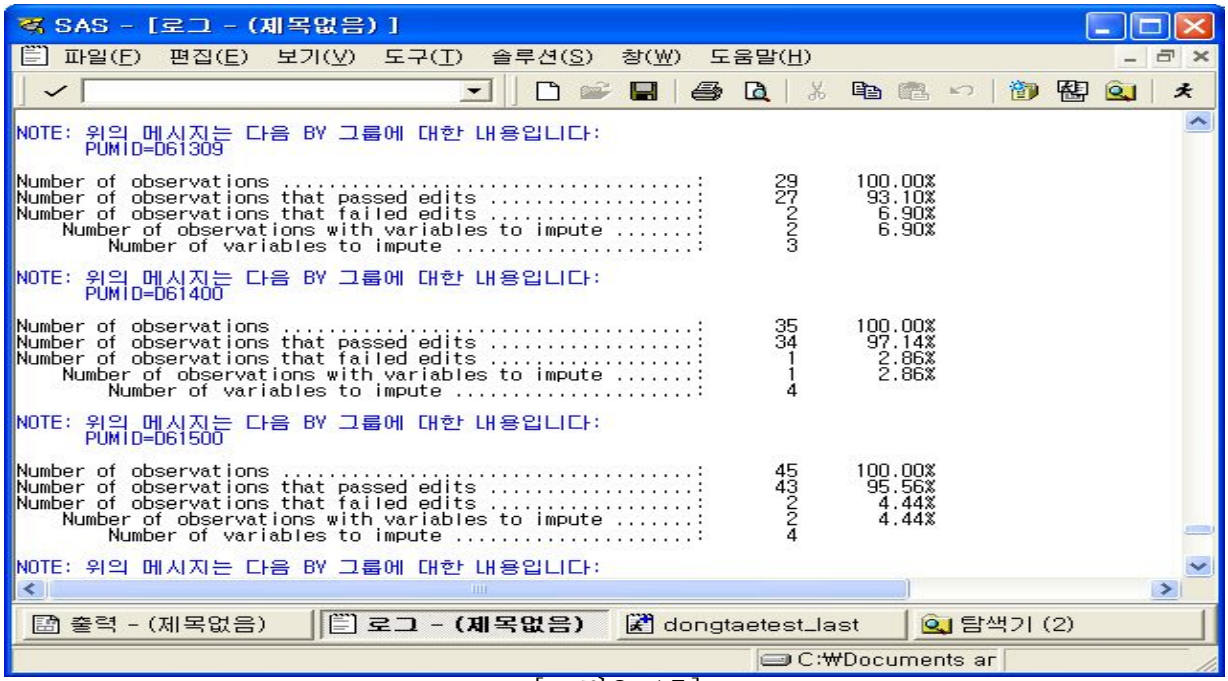
[그림3-13]

[그림3-14]처럼 앞에서 정의된 edit rule의 minimal set을 넣어 Error Localization을 수행한다. 여기서 Weights 란에서 각 변수들 중 중요한 변수를 지정할 수 있다. 가중값을 많이 줄수록 edit rule을 만족시키기 위해서 값이 변화될 가능성은 적어진다. 본 연구에서는 가중값을 주지 않아 모든 변수들이 동일한 중요성을 가진다는 것을 알 수 있다. 실행을 시켜서 나온 결과를 살펴보면 다음과 같다.



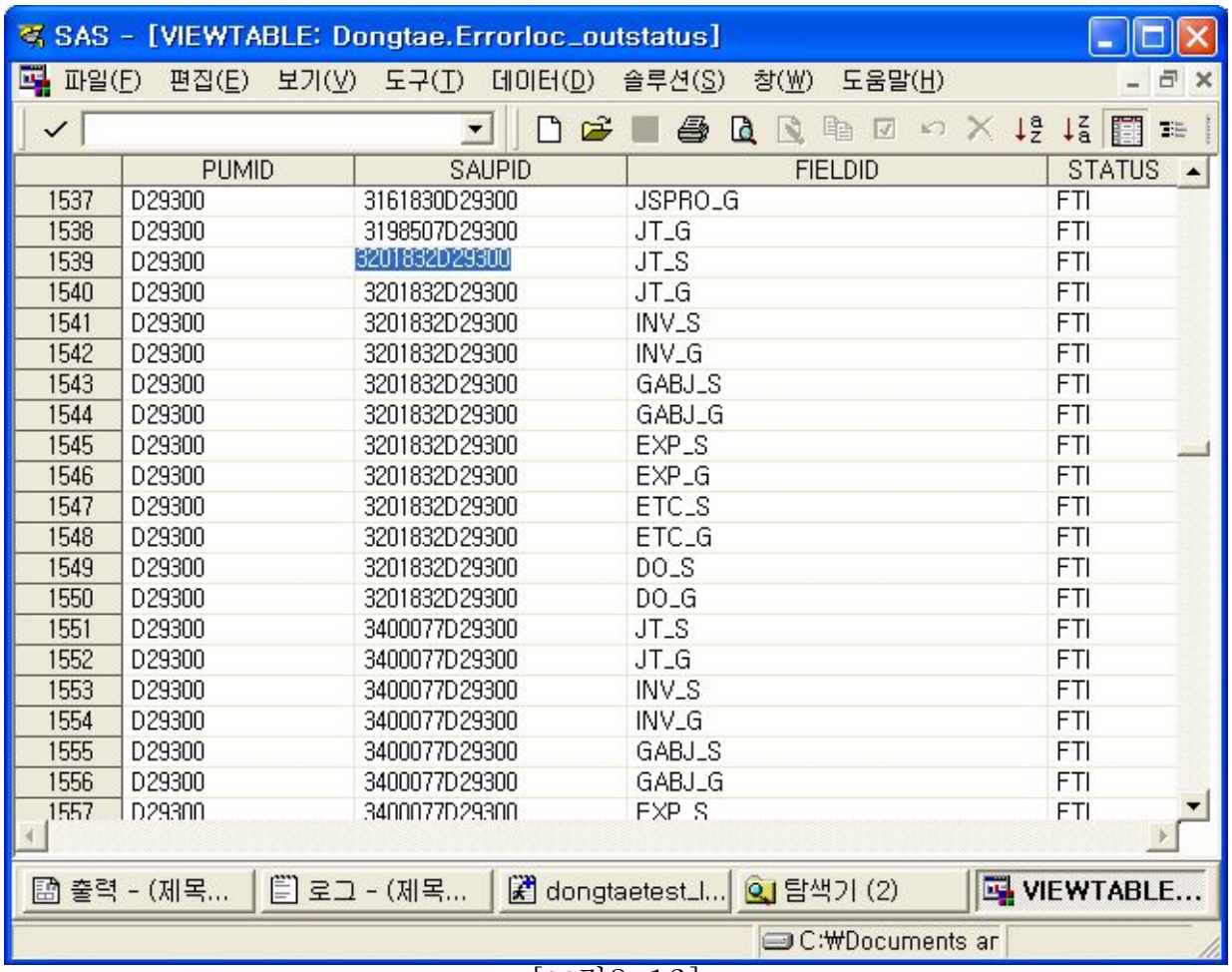
[그림3-14]

[그림3-15]는 Error Localization을 수행했을 때 모든 edit rule을 만족시킬 수 있는 최소한의 변수의 수를 나타내고 있다. 각 품목별 결과들에 대하여 살펴보면 품목 D61309의 경우 총 개체수는 29개, 모든 edit rule을 만족시키는 개체수는 27개, 만족되지 않는 개체수는 2개이다. 여기서 2개의 개체가 모든 edit rule을 만족하게 하기 위해서는 최소한 3개의 변수를 imputation을 시켜야함을 알 수 있다. 품목 D61400과 D61500의 경우는 4개의 변수를 imputation을 시켜야한다. 모든 품목별 개체에 대하여 최소한으로 바뀌어야 하는 변수의 수를 알 수 있다.



[그림3-15]

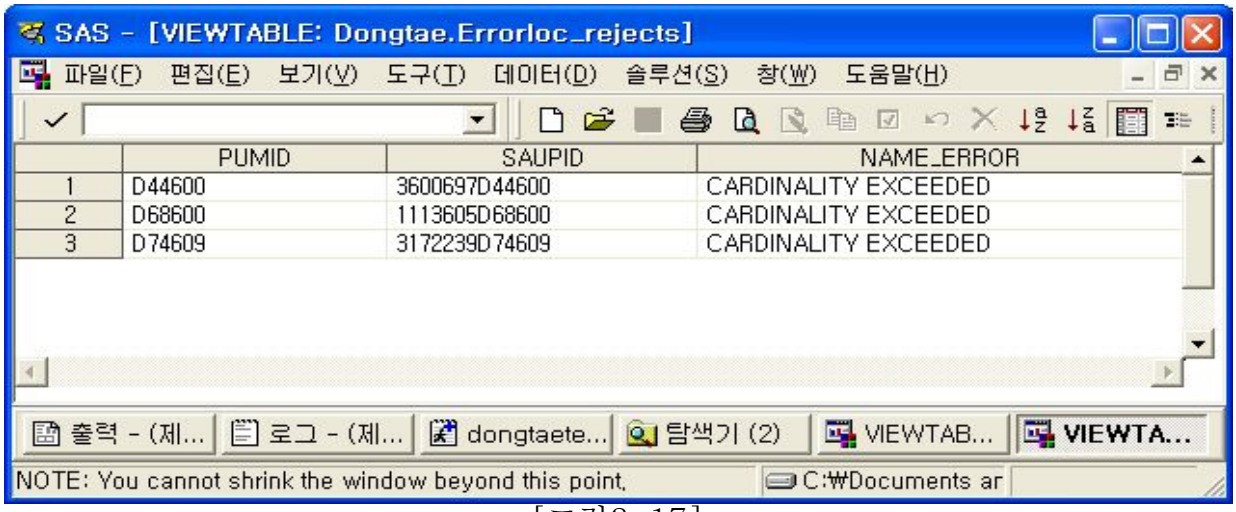
[그림3-16]은 Error Localization을 수행했을 때 imputation이 요구되어지는 변수들의 목록이 제시되어 있다. 예를 들어 사업체 3201832D29300의 경우 12개의 변수에 대하여 imputation을 실시해야 한다. 실제 이 자료값을 찾아보면 missing(-1로 입력)이 되어 있는 것을 알 수 있다. 또한 합산이 맞지 않는 경우도 포함이 되어 있다. 향후 이러한 변수 모두에 대하여 다양한 방법으로 imputation을 실시할 것이다.



	PUMID	SAUPID	FIELDID	STATUS
1537	D29300	3161830D29300	JSPRO_G	FTI
1538	D29300	3198507D29300	JT_G	FTI
1539	D29300	3201832D29300	JT_S	FTI
1540	D29300	3201832D29300	JT_G	FTI
1541	D29300	3201832D29300	INV_S	FTI
1542	D29300	3201832D29300	INV_G	FTI
1543	D29300	3201832D29300	GABJ_S	FTI
1544	D29300	3201832D29300	GABJ_G	FTI
1545	D29300	3201832D29300	EXP_S	FTI
1546	D29300	3201832D29300	EXP_G	FTI
1547	D29300	3201832D29300	ETC_S	FTI
1548	D29300	3201832D29300	ETC_G	FTI
1549	D29300	3201832D29300	DO_S	FTI
1550	D29300	3201832D29300	DO_G	FTI
1551	D29300	3400077D29300	JT_S	FTI
1552	D29300	3400077D29300	JT_G	FTI
1553	D29300	3400077D29300	INV_S	FTI
1554	D29300	3400077D29300	INV_G	FTI
1555	D29300	3400077D29300	GABJ_S	FTI
1556	D29300	3400077D29300	GABJ_G	FTI
1557	D29300	3400077D29300	EXP_S	FTI

[그림3-16]

[그림3-17]은 불응 또는 부재 사업체에 대한 정보가 나와 있다. 다시 말하면, imputation에 사용할만한 어떠한 자료도 입력되어 있지 않다. 그래서 이 사업체에 대해서는 모든 imputation이 끝난 후 지난 자료로부터 확보된 종업원수나 근무일수와 같은 자료를 이용하여 mass imputation을 실시할 것이다. Mass imputation을 실시하지 않고 지난 자료를 그대로 쓰는 방법도 있지만 실무부서와 함께 검토가 필요할 것으로 본다. 본 연구에서는 3개의 사업체가 모든 항목에 대하여 무응답을 하였다.



[그림3-17]

3.2.5. Deterministic Imputation (Proc Deterministic)

1) 개요

Edit rule을 만족시키는 오직 한 가지의 imputation값만이 존재하는 경우를 진단하는 절차이다. 이 절차는 Banff 스스로 imputation에 필요한 값을 찾는 것은 아니며, 이후에 실시될 다른 imputation 방법에 의해 무응답 대체가 이루어 질 때, imputation이 필요한 부분을 줄이고자 imputation의 첫 단계로 수행된다.

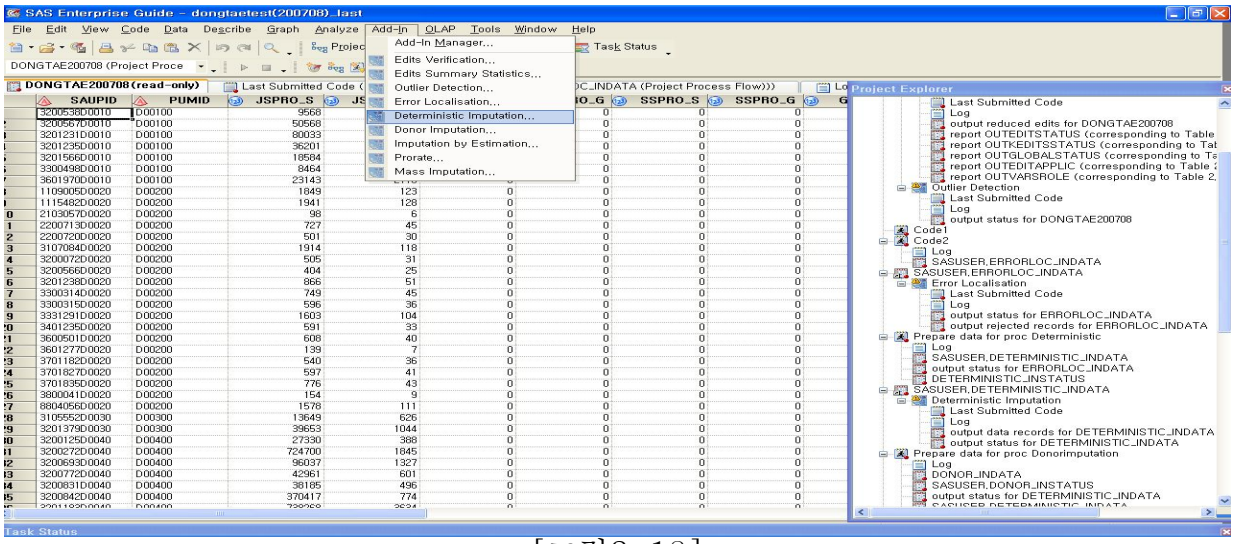
2) 절차

- 상단의 Add-In menu에서 Deterministic Imputation task를 선택한다.
- Key variable을 반드시 설정해준다.
  BY variable이 있으면 BY variable과 key variable의 순으로 자료가 정렬된다.
- Key variable 또는 fieldid에 missing value가 있는 관측치는 실행되지 않는다.
- Edit에서 이전 단계에서 사용한 edit rule을 적용한다.

3) 적용

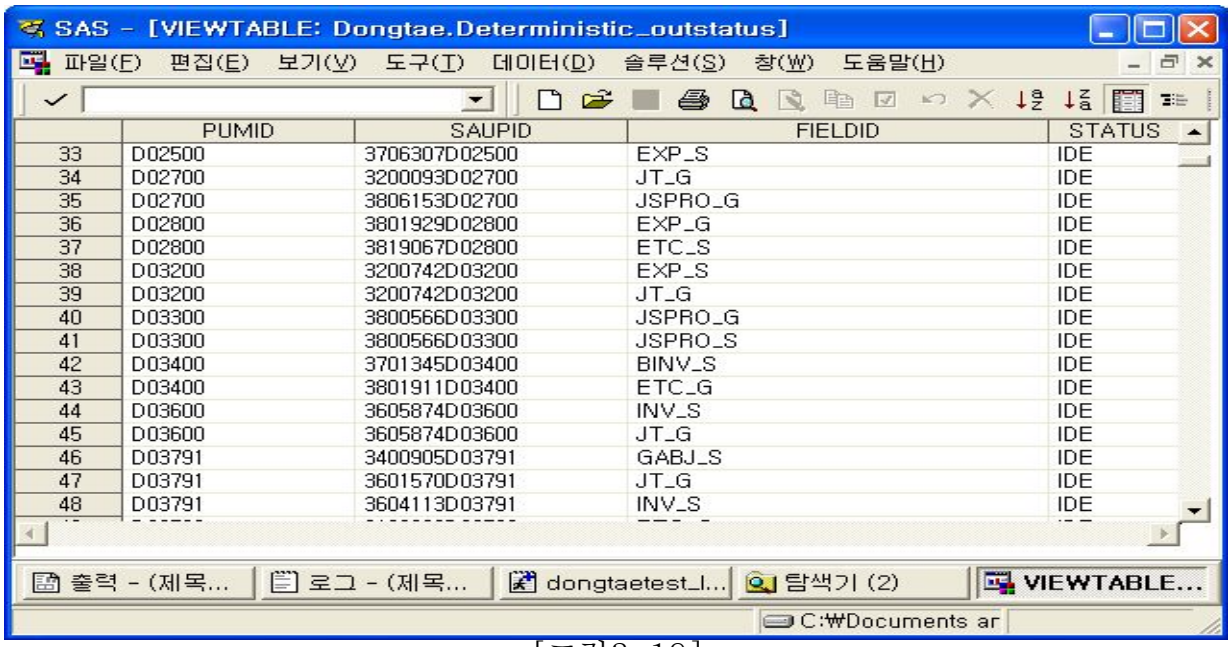
Deterministic imputation에 앞서 입력 data를 새로 만들어야 한다. 즉, 이상값 체크에서 선택된 변수와 Error Localization을 수행한 후 edit rule을 만족시키지 못하는 변수를 합친 data를 가지고 imputation 방법을 적용시켜야 한다. Data가 완성이 되면 [그림3-18]과 같이 SAS Enterprise Guide의 Add-In 메뉴로부터

Deterministic imputation을 선택하여 사용한다.



[그림3-18]

앞에서 사용한 edit rule을 적용하였을 때 이 시스템에서 스스로 imputation을 위한 값을 찾는 것은 아니며 사용자에 의해 정의된 edit rule에 의하여 직접적으로 값이 결정이 된다면 그 값을 대체시키는 것이다. [그림3-19]에는 Deterministic imputation을 실시한 후 imputation이 된 사업체와 변수에 대한 정보가 주어져 있다. 본 자료에 대한 edit rule중에서 합의 형태로 된 것이 2개가 있다. 이 식에 의하여 자동적으로 값이 결정이 되어진 변수들이라 하겠다. STATUS의 IDE는 Imputed by Deterministic imputation이라는 것을 나타낸다. 이러한 과정을 거친 후에는 실제 imputation이 필요한 변수가 상당히 줄어있음을 알 수 있다.



	PUMID	SAUPID	FIELDID	STATUS
33	D02500	3706307D02500	EXP_S	IDE
34	D02700	3200093D02700	JT_G	IDE
35	D02700	3806153D02700	JSPRO_G	IDE
36	D02800	3801929D02800	EXP_G	IDE
37	D02800	3819067D02800	ETC_S	IDE
38	D03200	3200742D03200	EXP_S	IDE
39	D03200	3200742D03200	JT_G	IDE
40	D03300	3800566D03300	JSPRO_G	IDE
41	D03300	3800566D03300	JSPRO_S	IDE
42	D03400	3701345D03400	BINV_S	IDE
43	D03400	3801911D03400	ETC_G	IDE
44	D03600	3605874D03600	INV_S	IDE
45	D03600	3605874D03600	JT_G	IDE
46	D03791	3400905D03791	GABJ_S	IDE
47	D03791	3601570D03791	JT_G	IDE
48	D03791	3604113D03791	INV_S	IDE

[그림3-19]

3.2.6. Donor Imputation (Proc Donorimputation)

1) 개요

무응답 대체가 필요한 단위(recipient)에 대하여 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 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.

2) 절차

- 상단의 Add-In menu에서 Donor Imputation task를 선택한다.
- Task Roles 창에서 ID(key변수), BY변수, DATAEXCLAVAR 변수를 지정한다.
- Options창에서 mindonors, pcentdonors, eligdon 옵션을 설정한다.

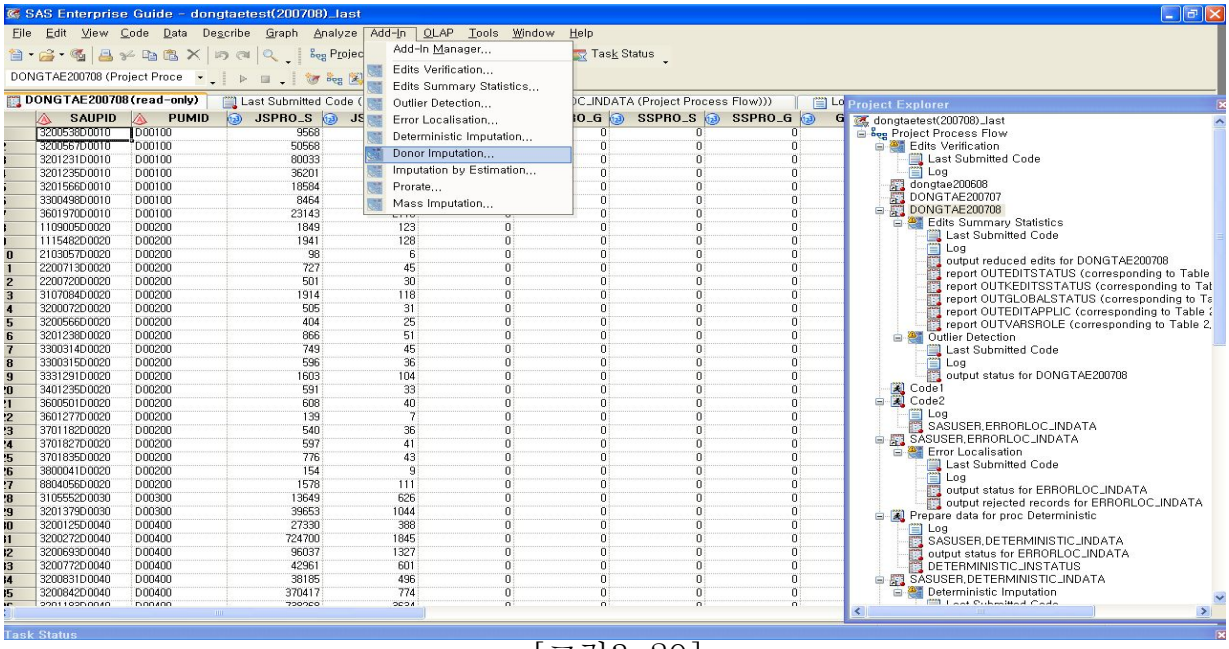
▪ minimum number of donors - default는 30
minimum percentage of donors - default는 30
maximum number of donors to try - donor를 찾기 위해 시도하는 횟수로서 필수 입력사항이다.

▪ eligdon (eligible donor : 자격이 있는 donor) 옵션을 original로 선택하면 imputation이 되어진 data는 donor로써 사용하지 않으며, any로 선택하면 donor로써 사용될 수 있다.

- Edits 창에 original edit과 post edit을 입력한다.

3) 적용

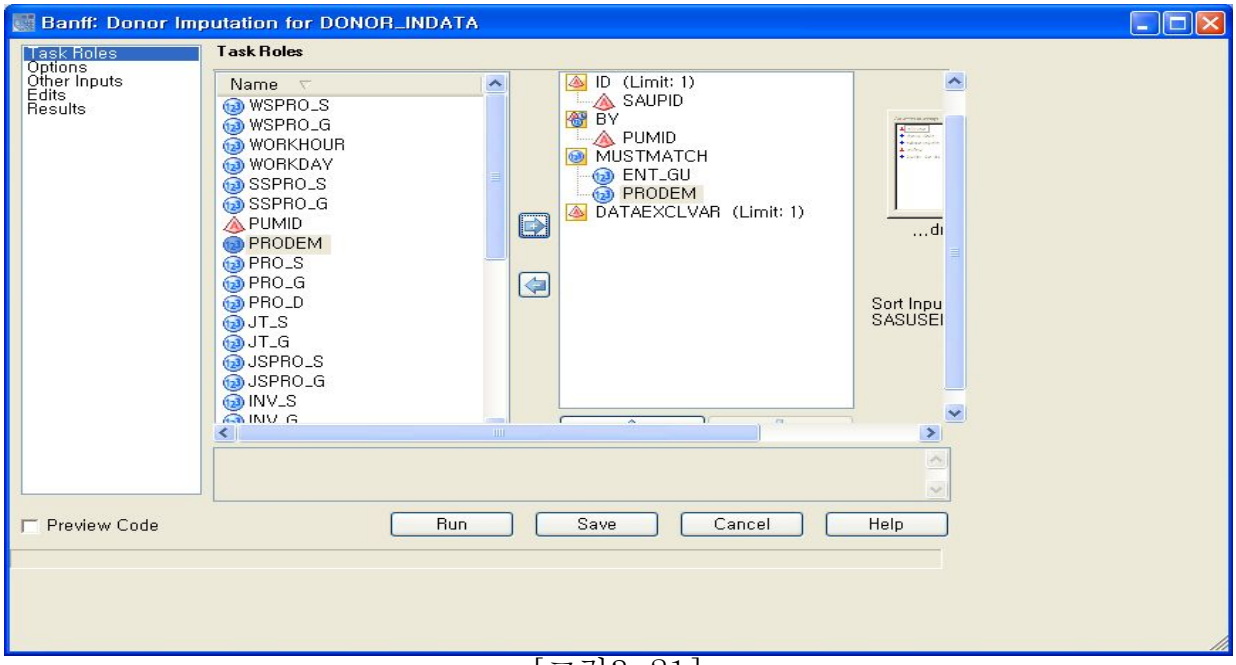
Donor imputation을 실행하기 위해서 [그림3-20]과 같이 SAS Enterprise Guide의 Add-In 메뉴로부터 Donor imputation을 선택하여 사용한다.



[그림3-20]

[그림3-21]은 Donor imputation을 하기에 앞서 matching field를 결정하는 과정을 나타낸다. Matching field란 대체하기 위한 donor를 찾을 때 가장 중요시하는 변수라고 볼 수 있다. 본 연구의 경우 ENT_GU(기업규모)와 PRODEM(종업원수)를 matching field로 선택하였다. 즉, 무응답이 있는 하나의 사업체와 가장 비슷한 사업체를 찾을 때 같은 품목을 생산하는 사업체 중에서 기업규모와 종업원수가 같은 사업체를 찾아서 대체에 적용을 하겠다는 것이다. 이러한 matching field를 선택하는 것은 매우 중요하다고 할 수 있다. 좋은 matching field를 선택할수록

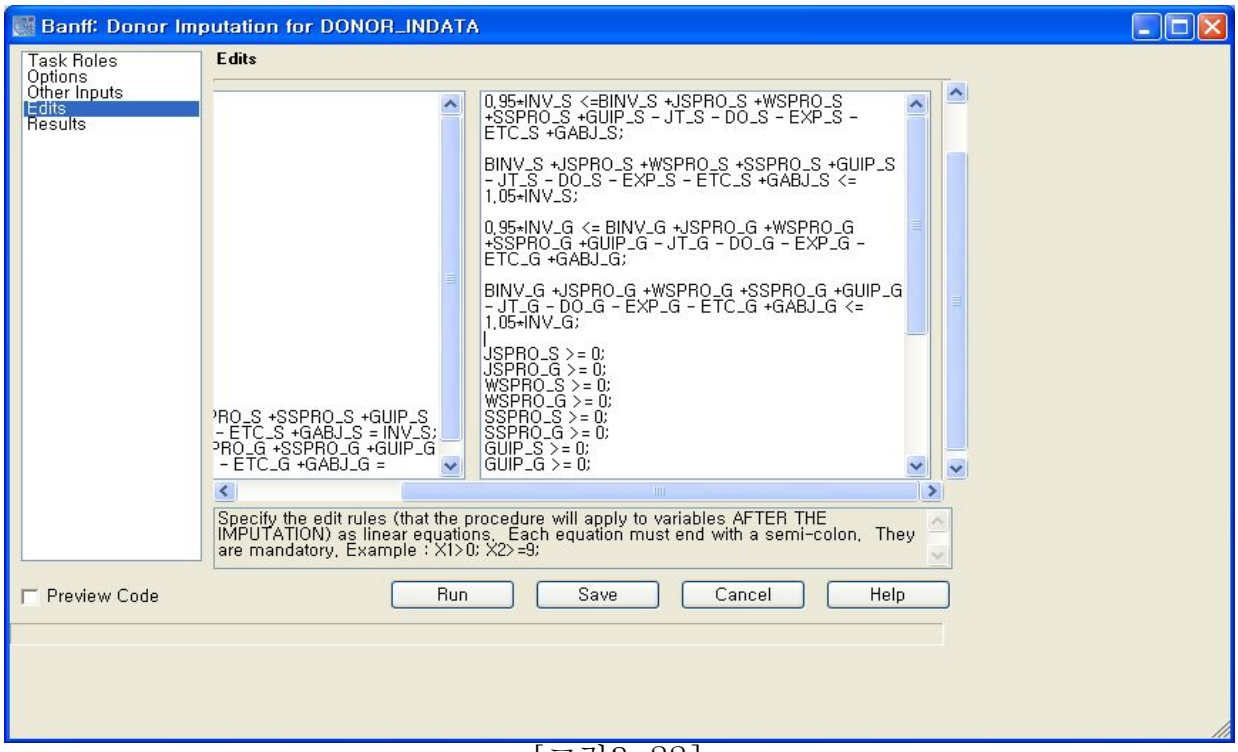
좋은 donor를 찾을 가능성이 높아진다. Matching field 선택에 관한 문제는 차후 연구 되어야 할 것으로 본다.



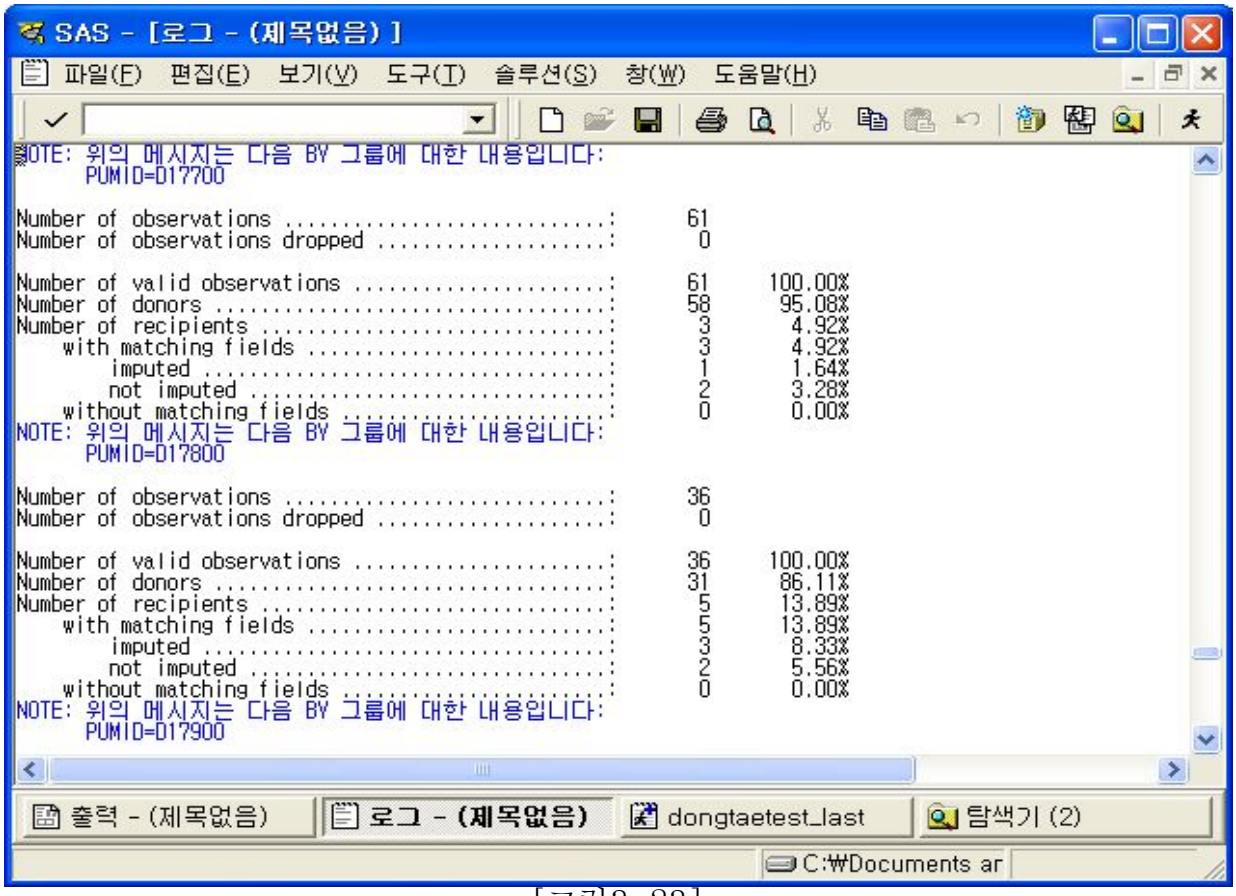
[그림3-21]

[그림3-22]는 post-imputation에 대한 조건을 완화하기 위하여 edit rule을 조금 수정을 하는 내용을 보여준다. post-imputation이란 Donor imputation을 하고 난 후 edit rule에서의 정의된 것에 위배되는 것을 의미한다. 따라서 이러한 현상이 일어나면 찾은 donor의 값을 쓰지 않고 다시 다른 donor의 값을 찾는 것을 반복한다. 그러나 대체된 후에 edit rule에서 정의된 합계는 정확히 같을 수가 없으므로 실제값의 95%에서 105%사이의 값은 같은 값으로 인정을 해주고 차후 다시 조정하는 절차를 가진다. 그러나 이 범위가 벗어나면 Donor imputation에서는 더 이상 imputation을 하지 않고 다른 imputation 방법을 사용하게 된다.

그리고 [그림3-23]은 모든 품목에 대하여 Donor imputation을 한 결과를 보여주고 있다. 두 품목의 예를 들면, 품목 D17800의 경우 총 61개의 개체가 있으며 imputation을 해야 하는 개체(recipient)는 3개이며, 이 중 1개는 imputed 되었지만 2개는 되지 못했다. 품목 D17900의 경우 총 36개의 개체가 있으며 imputation을 해야 하는 개체(recipient)는 5개이며, 이 중 3개는 imputed 되었지만 2개는 되지 못했다. 이 과정에서 imputation이 되지 못한 것은 차후 다른 방법에 의하여 imputation이 될 것이다.

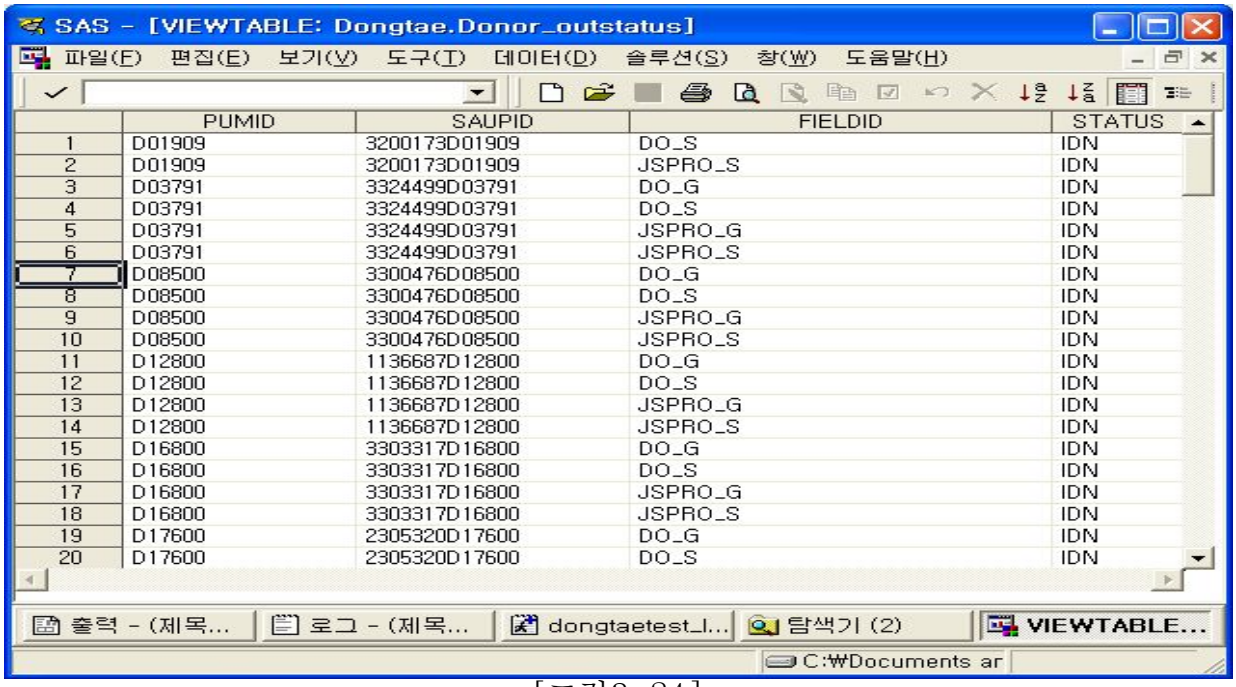

Banff: Donor Imputation for DONOR_INDATA
Task Roles
Options
Other Inputs
Edits
Results
Edits
0.95*INV_S <=BINV_S +JSPRO_S +WSPRO_S +SSPRO_S +GUIP_S - JT_S - DO_S - EXP_S - ETC_S +GABJ_S;
BINV_S +JSPRO_S +WSPRO_S +SSPRO_S +GUIP_S - JT_S - DO_S - EXP_S - ETC_S +GABJ_S <= 1.05*INV_S;
0.95*INV_G <= BINV_G +JSPRO_G +WSPRO_G +SSPRO_G +GUIP_G - JT_G - DO_G - EXP_G - ETC_G +GABJ_G;
BINV_G +JSPRO_G +WSPRO_G +SSPRO_G +GUIP_G - JT_G - DO_G - EXP_G - ETC_G +GABJ_G <= 1.05*INV_G;
JSPRO_S >= 0;
JSPRO_G >= 0;
WSPRO_S >= 0;
WSPRO_G >= 0;
SSPRO_S >= 0;
SSPRO_G >= 0;
GUIP_S >= 0;
GUIP_G >= 0;
RO_S +SSPRO_S +GUIP_S - ETC_S +GABJ_S = INV_S;
RO_G +SSPRO_G +GUIP_G - ETC_G +GABJ_G =
Specify the edit rules (that the procedure will apply to variables AFTER THE IMPUTATION) as linear equations. Each equation must end with a semi-colon. They are mandatory. Example : X1>0; X2>=9;
Preview Code
Run
Save
Cancel
Help

[그림3-22]


SAS - [로그 - (제목없음)]
파일(F) 편집(E) 보기(V) 도구(T) 솔루션(S) 창(W) 도움말(H)
NOTE: 위의 메시지는 다음 BY 그룹에 대한 내용입니다:
PUMID=D17700
Number of observations 61
Number of observations dropped 0
Number of valid observations 61 100.00%
Number of donors 58 95.08%
Number of recipients 3 4.92%
with matching fields 3 4.92%
imputed 1 1.64%
not imputed 2 3.28%
without matching fields 0 0.00%
NOTE: 위의 메시지는 다음 BY 그룹에 대한 내용입니다:
PUMID=D17800
Number of observations 36
Number of observations dropped 0
Number of valid observations 36 100.00%
Number of donors 31 86.11%
Number of recipients 5 13.89%
with matching fields 5 13.89%
imputed 3 8.33%
not imputed 2 5.56%
without matching fields 0 0.00%
NOTE: 위의 메시지는 다음 BY 그룹에 대한 내용입니다:
PUMID=D17900
출력 - (제목없음)
로그 - (제목없음)
dongtaetest_last
탐색기 (2)
C:₩Documents ar

[그림3-23]

[그림3-24]에는 Donor imputation을 실시한 후 imputation이 된 사업체와 변수에 대한 정보가 주어져 있다. STATUS의 IDE는 Imputed by Donor imputation이라는 것을 나타낸다. 사업체 3300476D08500인 경우 donor에 의하여 4개의 변수(DO_G, DO_S, JSPRO_G, JSPRO_S)가 대체 되었다는 것을 알 수 있다.


SAS - [VIEWTABLE: Dongtae.Donor_outstatus]

파일(F) 편집(E) 보기(V) 도구(T) 데이터(D) 솔루션(S) 창(W) 도움말(H)

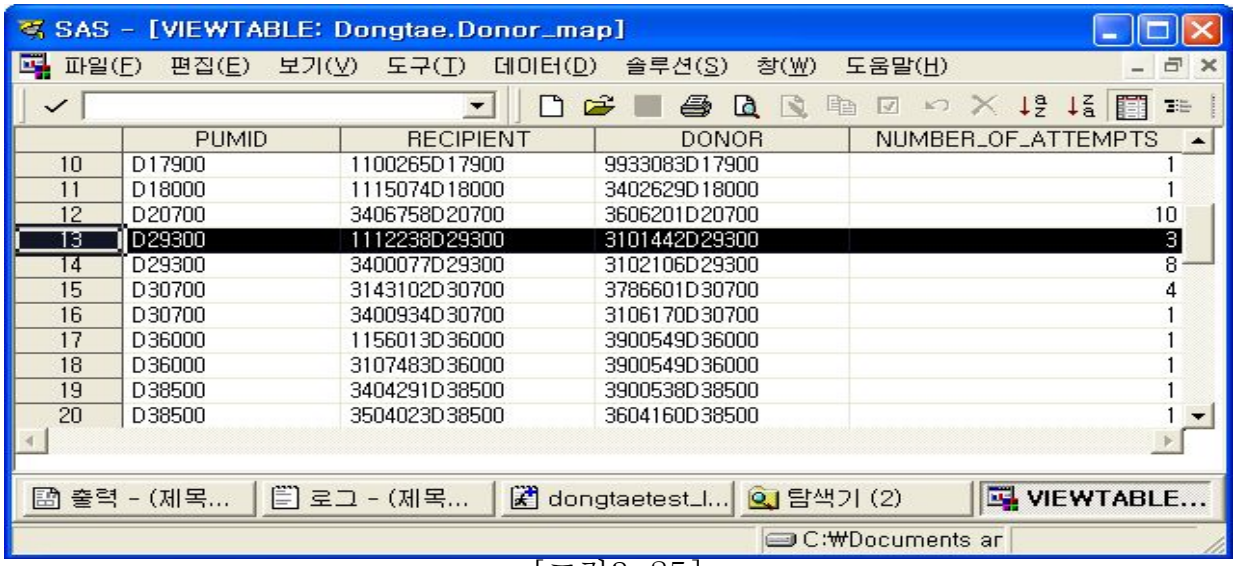
	PUMID	SAUPID	FIELDID	STATUS
1	D01909	3200173D01909	DO_S	IDN
2	D01909	3200173D01909	JSPRO_S	IDN
3	D03791	3324499D03791	DO_G	IDN
4	D03791	3324499D03791	DO_S	IDN
5	D03791	3324499D03791	JSPRO_G	IDN
6	D03791	3324499D03791	JSPRO_S	IDN
7	D08500	3300476D08500	DO_G	IDN
8	D08500	3300476D08500	DO_S	IDN
9	D08500	3300476D08500	JSPRO_G	IDN
10	D08500	3300476D08500	JSPRO_S	IDN
11	D12800	1136687D12800	DO_G	IDN
12	D12800	1136687D12800	DO_S	IDN
13	D12800	1136687D12800	JSPRO_G	IDN
14	D12800	1136687D12800	JSPRO_S	IDN
15	D16800	3303317D16800	DO_G	IDN
16	D16800	3303317D16800	DO_S	IDN
17	D16800	3303317D16800	JSPRO_G	IDN
18	D16800	3303317D16800	JSPRO_S	IDN
19	D17600	2305320D17600	DO_G	IDN
20	D17600	2305320D17600	DO_S	IDN

출력 - (제목... 로그 - (제목... dongtaetest_l... 탐색기 (2) VIEWTABLE...

C:₩Documents ar

[그림3-24]

[그림3-25]는 imputation을 해야 하는 개체(recipient)와 찾아진 각 개체들의 donor와 post-imputation을 하지 않게 되는 donor를 찾기 위하여 시도된 횟수를 보여주고 있다. 예를 들면, 사업체 1112238D29300은 사업체 3101442D29300의 값으로 imputation이 되었으며 3번째 시도에서 찾았다는 것을 나타내고 있다.


SAS - [VIEWTABLE: Dongtae.Donor_map]

파일(F) 편집(E) 보기(V) 도구(T) 데이터(D) 솔루션(S) 창(W) 도움말(H)

	PUMID	RECIPIENT	DONOR	NUMBER_OF_ATTEMPTS
10	D17900	1100265D17900	9933083D17900	1
11	D18000	1115074D18000	3402629D18000	1
12	D20700	3406758D20700	3606201D20700	10
13	D29300	1112238D29300	3101442D29300	3
14	D29300	3400077D29300	3102106D29300	8
15	D30700	3143102D30700	3786601D30700	4
16	D30700	3400934D30700	3106170D30700	1
17	D36000	1156013D36000	3900549D36000	1
18	D36000	3107483D36000	3900549D36000	1
19	D38500	3404291D38500	3900538D38500	1
20	D38500	3504023D38500	3604160D38500	1

출력 - (제목... 로그 - (제목... dongtaetest_l... 탐색기 (2) VIEWTABLE...

C:₩Documents ar

[그림3-25]

3.2.7. Estimator Imputation (Proc Estimator)

1) 개요

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.

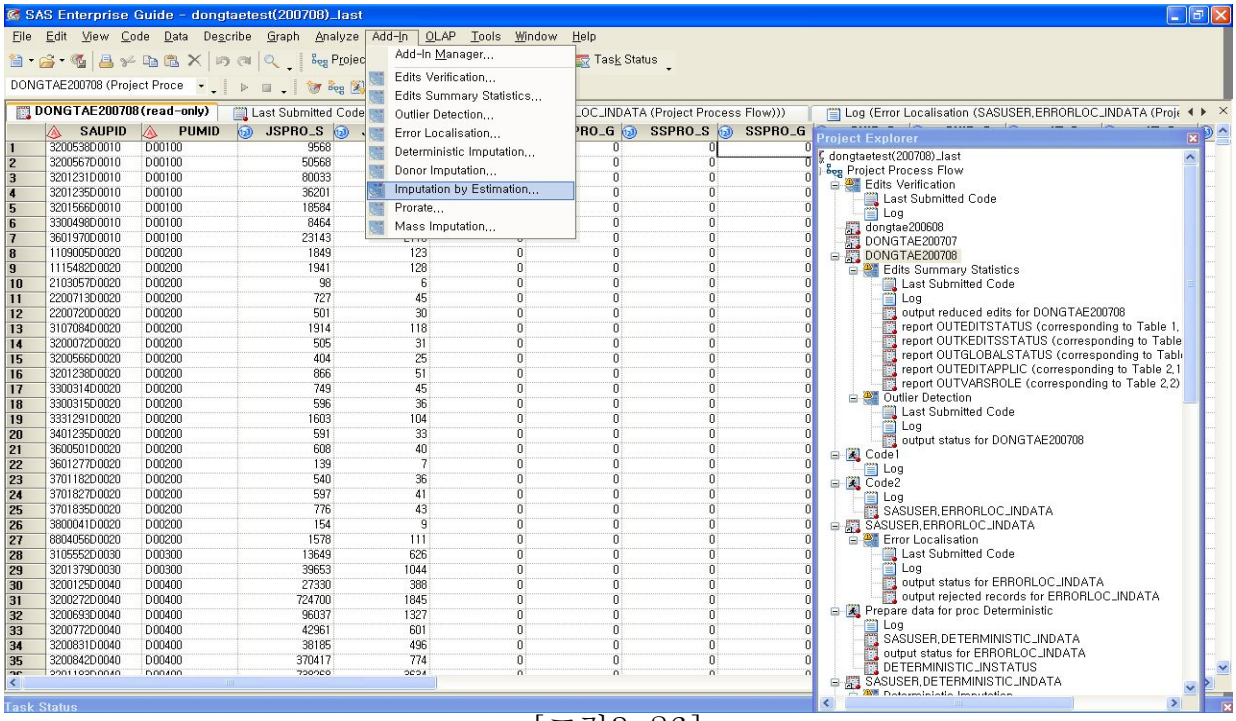
Donor imputation 프로시저와 달리 Estimator imputation 프로시저에는 post-imputation edits가 존재하지 않는다. 따라서 결과적으로 대체된 값들은 원래의 edit rule을 만족하지 않을 수도 있으므로 estimator imputation이 모두 끝난 뒤에는 Proc Errorloc 프로시저를 통해 그러한 일이 실제로 벌어지는 지를 확인해 봐야 한다.

2) 절차

- 상단의 Add-In menu에서 Imputation by Estimation을 선택하여 실행한다.
- Task Roles창에서 ID와 BY 변수를 선택한다.
- Option창에서는 대체를 위한 추정식의 선택한다. 이는 기존의 algorithm 을 사용할 수도 있고 사용자가 새로운 algorithm을 입력하여 쓸 수도 있다.
  - ▪ acceptnegative option과 rejectnegative option 중 선택해야 한다.
  : proc estimator 에서는 edit rule을 사용하지 않기 때문에 변수에 다라서 음수의 적용여부를 체크할 필요가 있다.
- Other Inputs창에서 필요한 경우 보조자료를 선택하여 사용한다.

3) 적용

Imputation estimators를 실행하기 위해서 [그림3-26]과 같이 SAS Enterprise Guide의 Add-In 메뉴로부터 Imputation by Estimation을 선택하여 사용한다. Imputation estimators 방법은 실제적으로 광공업 동태조사와 같은 양적변수가 많은 경제관련 조사에 주로 이용되는 방법이라고 할 수 있다. 그리고 imputation을 하기 위해서는 쓰고자 하는 함수 및 추정방법이 정의되어야 한다. 자료에 대한 특성에 맞는 추정방법을 사용해야하며 이것은 향후 연구가 되어야 하는 부분이라고 할 수 있다. 본 연구에서는 가장 적합할 가능성이 높은 추정방법을 제시하였으며 수정이 가능할 수도 있을 것이다.



[그림3-26]

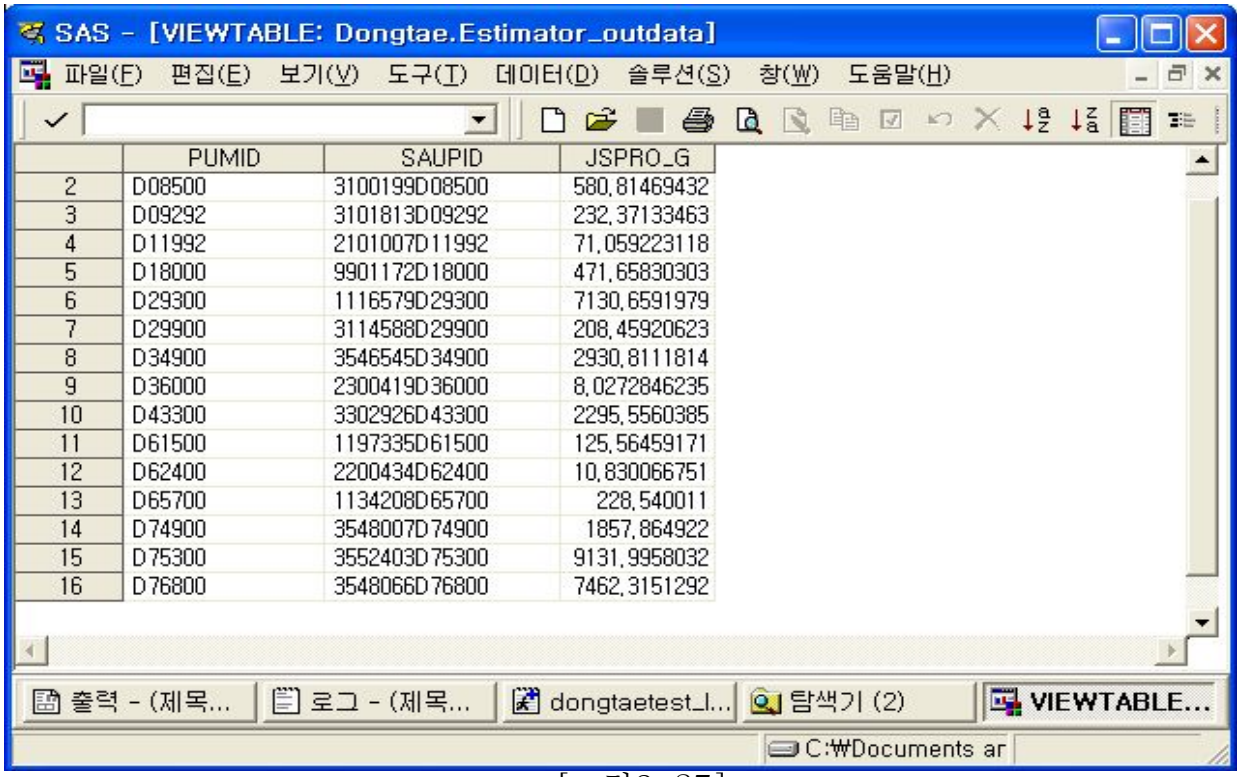
다음의 [표3-1]은 JSPRO를 imputation 시키기 위하여 본 연구에서 제시한 추정방법을 나타내고 있다. 가장 위에 있는 추정방법에 의하여 제일 먼저 추정값을 찾는다. 즉, 추정방법들 중 가장 적합한 추정방법일 것이다. 추정값을 찾은 개체는 imputation이 되고 보조변수가 존재하지 않아 추정값을 찾을 수 없는 개체는 그 다음의 추정방법에 의하여 추정된다. 이러한 과정에 의하여 모든 개체가 추정이 된다. 본 연구에서는 JSPRO(자체생산)의 경우, 전월의 JSPRO값과 PRODEM(종업원수), WORKDAY(근무일수)를 고려한 회귀모형으로 자체생산을 추정하는 것을 가장 적합한 방법으로 간주하고 추정을 하였다. 가장 적합한 추정방법을 찾는 것도

향후 연구되어야 할 내용이라고 할 수 있다.

	보조변수	추정방법
JSPRO (자체생산)	JSPRO(전월), PRODEM, WORKDAY	Regression
	PRODEM, WORKDAY, DO	Regression
	PRODEM, WORKDAY, EXP	Regression
	PRODEM, WORKDAY	Regression
	PRODEM	Regression
	PRODEM ($\widehat{JSPRO} = \frac{\overline{JSPRO}}{\overline{PRODEM}} \times PRODEM$)	Current Ratio
	JSPRO ($\widehat{JSPRO} = \overline{JSPRO}$)	Current Mean

[표3-1]

[그림3-27]에는 JSPRO_G변수에 대하여 Imputation estimators을 실시한 후 imputation이 되어진 사업체와 추정된 값에 대한 정보가 주어져 있다. 다른 변수들에 대해서도 각 변수의 특성에 맞는 추정방법에 의하여 추정을 함으로써 이와 같은 결과를 얻을 수 있다.



	PUMID	SAUPID	JSPRO_G
2	D08500	3100199D08500	580,81469432
3	D09292	3101813D09292	232,37133463
4	D11992	2101007D11992	71,059223118
5	D18000	9901172D18000	471,65830303
6	D29300	1116579D29300	7130,6591979
7	D29900	3114588D29900	208,45920623
8	D34900	3546545D34900	2930,8111814
9	D36000	2300419D36000	8,0272846235
10	D43300	3302926D43300	2295,5560385
11	D61500	1197335D61500	125,56459171
12	D62400	2200434D62400	10,830066751
13	D65700	1134208D65700	228,540011
14	D74900	3548007D74900	1857,864922
15	D75300	3552403D75300	9131,9958032
16	D76800	3548066D76800	7462,3151292

[그림3-27]

[표3-2]에서 [표3-9]은 본 연구에서 제시한 나머지 변수들에 관한 추정방법을 나타내고 있다.

	보조변수	추정방법
DO (국내시판)	DO(전월), PRODEM, WORKDAY	Regression
	PRODEM, WORKDAY, JSPRO	Regression
	PRODEM, WORKDAY, EXP	Regression
	PRODEM, WORKDAY	Regression
	PRODEM	Regression
	PRODEM ($\widehat{DO}=\frac{\overline{DO}}{\overline{PRODEM}}\times PRODEM$)	Current Ratio
	DO ($\widehat{DO}=\overline{DO}$)	Current Mean

[표3-2]

EXP(수출)의 경우는 모든 사업체에게 해당되는 것이 아니므로 0의 값이 많이 존재한다. 따라서 추정 방법도 JSPRO(자체생산)이나 DO(국내시판)과는 [표3-3]의 추정식처럼 달라져야 한다고 본다. 본 연구에서는 과거의 수출여부를 가장 중요하게 고려하였다. 그래서 과거에도 수출이 없으면 추정된 식에 0의 값이 곱해져도 0으로 imputation이 될 것이다. 이러한 특성을 가지는 JT(재투입), ETC(기타출하), WSPRO(위탁생산), SSPRO(수탁생산), GUIP(구입)은 [표3-4]에서 [표3-8]처럼 과거자료의 특성을 반영하여 추정을 하였다.

	보조변수	추정방법
EXP (수출)	JSPRO ($\widehat{EXP}=\frac{\overline{JSPRO(현재)}}{\overline{JSPRO(전월)}}\times EXP(전월)$)	Trend
	EXP(전월) ($\widehat{EXP}=EXP(전월)$)	Previous Value
	EXP ($\widehat{EXP}=\overline{EXP}$)	Current Mean

[표3-3]

	보조변수	추정방법
JT (재투입)	JSPRO ($\widehat{JT}= \frac{\overline{JSPRO(\text{현재})}}{\overline{JSPRO(\text{전월})}} \times JT(\text{전월})$)	Trend
	JT(전월) ($\widehat{JT}= JT(\text{전월})$)	Previous Value
	JT ($\widehat{JT}= \overline{JT}$)	Current Mean

[표3-4]

	보조변수	추정방법
ETC (기타출하)	JSPRO ($\widehat{ETC}= \frac{\overline{JSPRO(\text{현재})}}{\overline{JSPRO(\text{전월})}} \times ETC(\text{전월})$)	Trend
	ETC(전월) ($\widehat{ETC}= ETC(\text{전월})$)	Previous Value
	ETC ($\widehat{ETC}= \overline{ETC}$)	Current Mean

[표3-5]

	보조변수	추정방법
WSPRO (위탁생산)	DO ($\widehat{WSPRO}= \frac{\overline{DO(\text{현재})}}{\overline{DO(\text{전월})}} \times WSPRO(\text{전월})$)	Trend
	WSPRO(전월) ($\widehat{WSPRO}= WSPRO(\text{전월})$)	Previous Value
	WSPRO ($\widehat{WSPRO}= \overline{WSPRO}$)	Current Mean

[표3-6]

	보조변수	추정방법
SSPRO (수탁생산)	DO ($\widehat{SSPRO}= \frac{\overline{DO(\text{현재})}}{\overline{DO(\text{전월})}} \times SSPRO(\text{전월})$)	Trend
	SSPRO(전월) ($\widehat{SSPRO}= SSPRO(\text{전월})$)	Previous Value
	SSPRO ($\widehat{SSPRO}= \overline{SSPRO}$)	Current Mean

[표3-7]

	보조변수	추정방법
GUIP (구입)	DO ($\widehat{GUIP}=\frac{\overline{DO(\text{현재})}}{\overline{DO(\text{전월})}}\times GUIP(\text{전월})$)	Trend
	GUIP(전월) ($\widehat{GUIP}= GUIP(\text{전월})$)	Previous Value
	GUIP ($\widehat{GUIP}= \overline{GUIP}$)	Current Mean

[표3-8]

[표3-9]처럼 전월재고는 전월자료의 월말재고이므로 이 값으로 넣어주면 된다. 전월에 조사되지 않은 새로운 사업체이면 현재자료의 평균값을 넣어준다.

	보조변수	추정방법
BINV (전월재고)	INV(전월) ($\widehat{BINV}= INV(\text{전월})$)	Previous Auxiliary Variable
	BINV ($\widehat{BINV}= \overline{BINV}$)	Current Mean

[표3-9]

3.2.8. Pro-Rating (Proc Prorate)

1) 개요

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. 또한 합계를 맞추기 위해 수정된 부분은 사용자에 의해 정의된 소숫점자리(decimals)에 의해 라운드 처리된다.

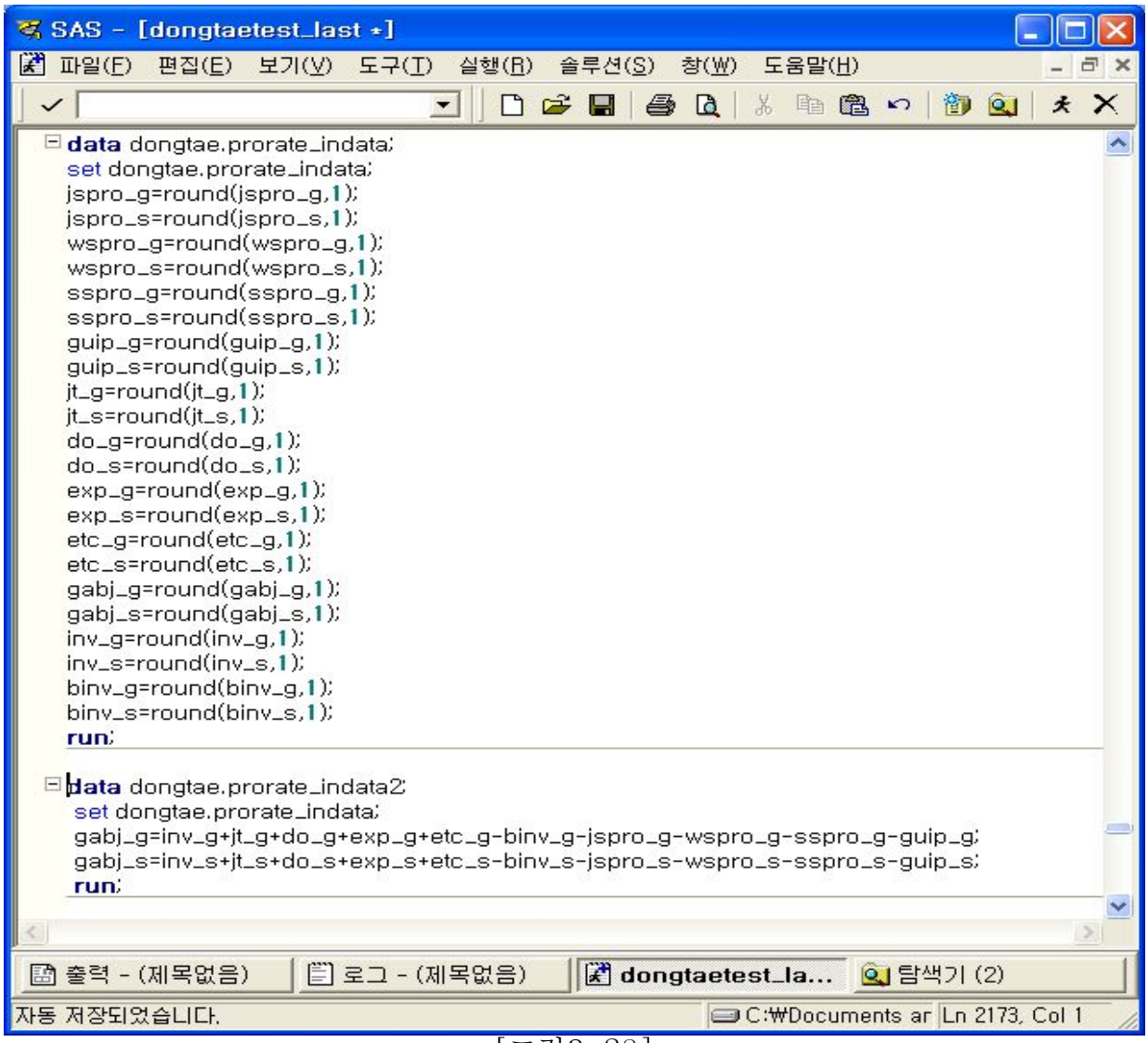
2) 절차

- 상단의 Add-In menu에서 Prorate을 선택하여 실행한다.

- Task Roles창에서 이전과 같은 ID와 BY 변수를 선택한다.
- Option창에서는 method와 modifier를 지정해 주어야한다.
- Edits창에 합의 형태로 된 edit rule을 넣어주면 된다.

3) 적용

광공업 동태조사에서는 자료의 특성상 Pro-rating절차는 필요하지 않을 것으로 본다. 본 자료에는 GABJ(과부족)이라는 항목이 있으므로 이 값을 조정함으로써 모든 합계를 일치시킬 수 있다. 따라서 모든 imputation이 끝난 후에 [그림3-28]과 같이 간단한 절차로 이 과정을 수행할 수 있을 것이다. 모든 추정된 값에 대하여 소수점이 있는 부분은 반올림하여 정수의 형태로 만들어 준 후 edit rule에 만족하게 되는 GABJ값을 결정하면 된다.



[그림3-28]

3.2.9. Mass Imputation (Proc Massimputation)

1) 개요

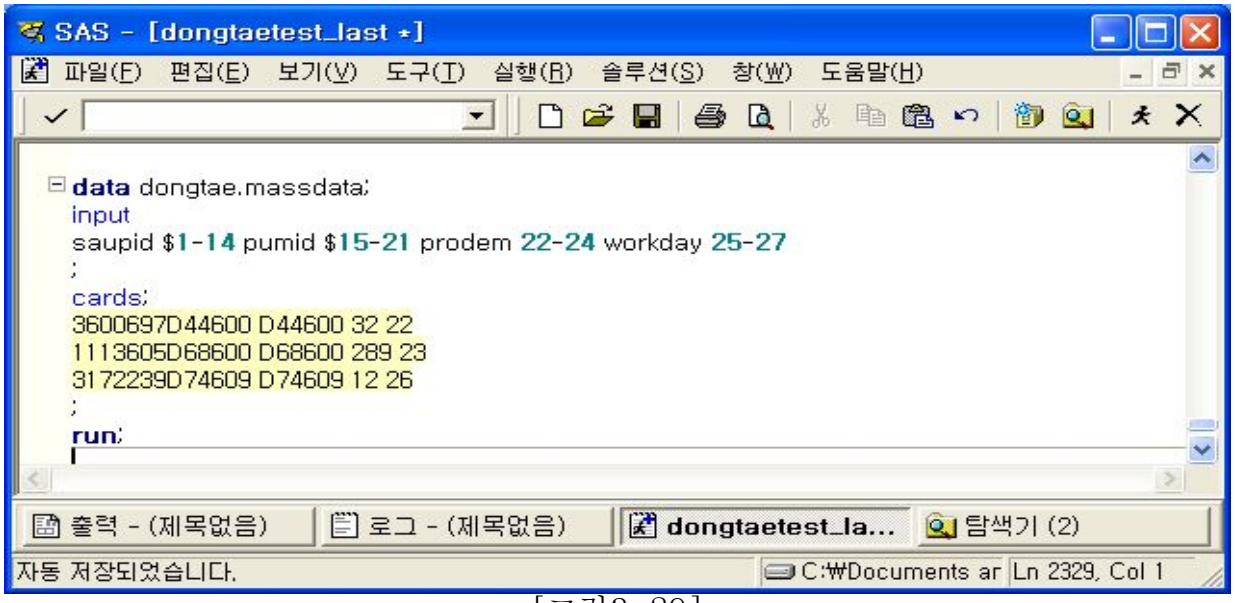
앞의 imputation 절차들을 마친 후 적용하는 마지막 단계로서, nearest neighbor approach를 이용하여 imputation이 필요한 값과 가장 유사한 관측치(valid한 값)를 찾는 절차이다. Donor를 선택하는 과정은 Proc Donorimputation 프로시저와 거의 비슷하지만 한 가지 차이가 존재하는데 이는 post-imputation을 필요로 하지 않는다는 것이다. 따라서 Proc Massimputation 프로시저는 가장 가까운 거리에 있는 donor를 선택해 단순히 imputation을 실시하거나, matching field가 존재하지 않을 때에는 무작위로 donor를 선택해서 imputation을 실시하게 된다.

2) 절차

- 상단의 Add-In menu에서 Mass Imputation을 선택하여 실행한다.
- Task Roles창에서 key변수와 BY변수, Mustmatch 변수, Mustimpute 변수를 설정한다.
  - Mustmatch변수는 matching field를 의미한다.
  - Mustimpute 변수는 imputation을 원하는 변수들의 목록을 의미한다.
- Options창에서 Mindonors, Pcentdonors를 설정한다. 이는 Donor imputation과 같은 내용이다.

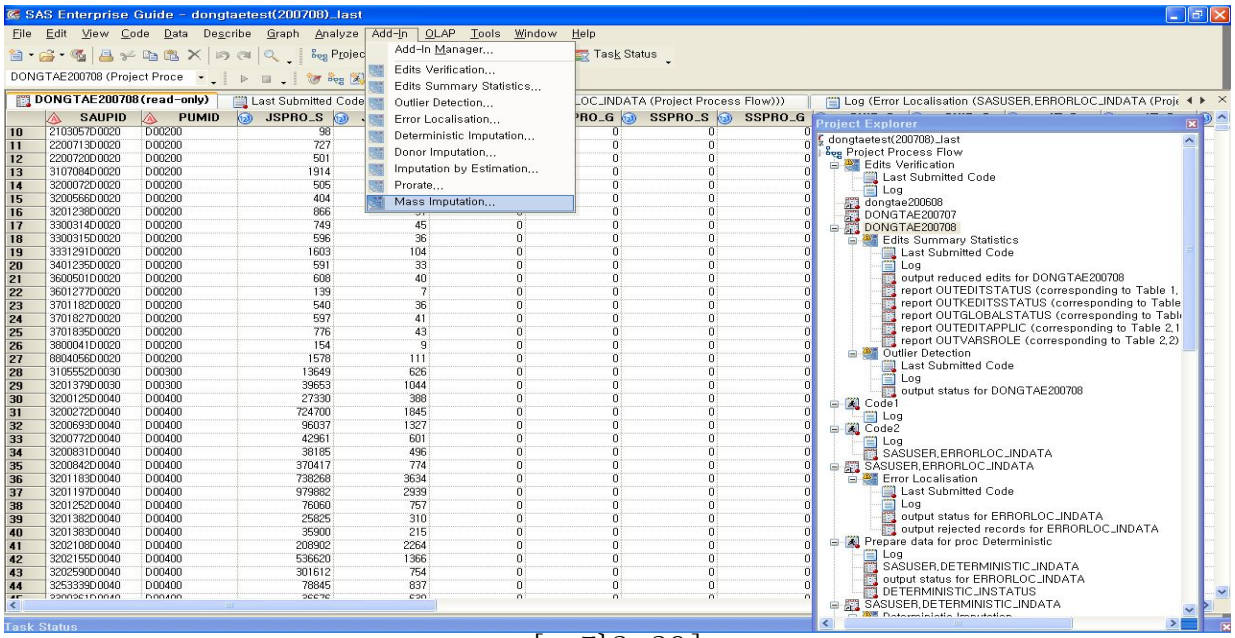
3) 적용

이전에 Error Localization을 수행한 결과 모든 항목에 대하여 무응답을 한 사업체가 3개가 있었다. 이 사업체에 대하여 Mass imputation을 실시할 것이다. 지난 과정에서 제외시켰던 3개의 업체를 [그림3-29]에서처럼 다시 입력을 해준다. 이때 이전에 조사되었던 종업원수와 근무일수의 정보만을 이용하여 imputation을 실시할 것이다. 만약 이러한 정보도 없을 때에는 무작위로 imputation을 실시하게 된다.



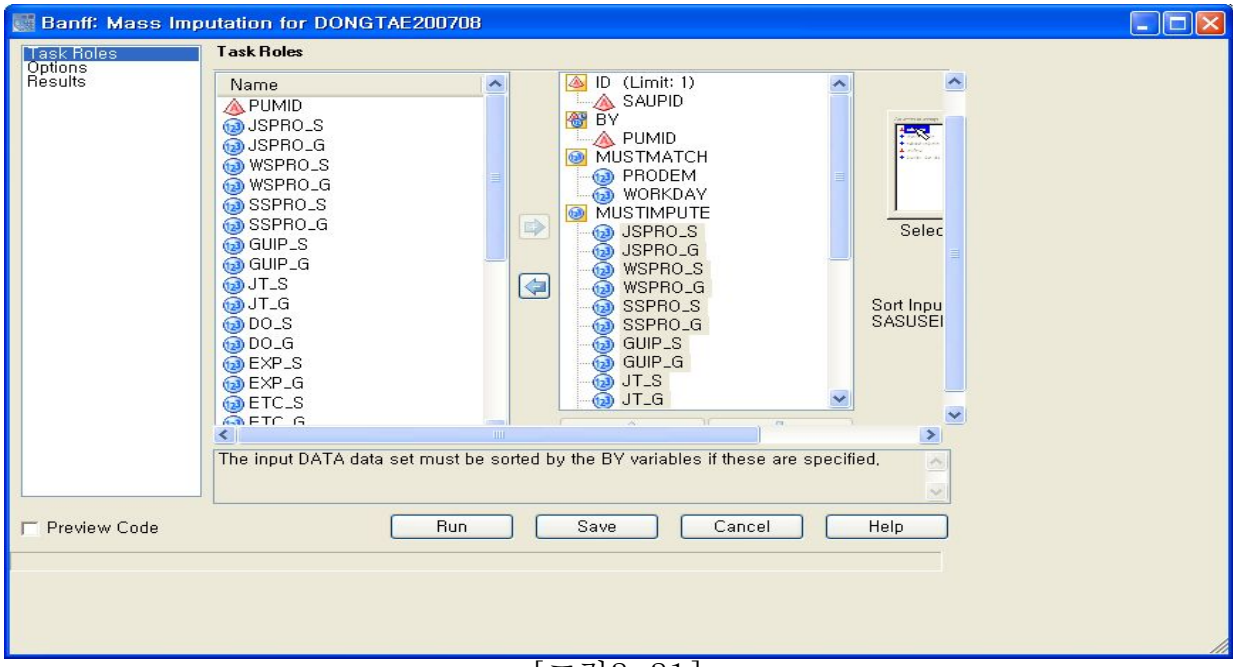
[그림3-29]

Mass imputation을 실행하기 위해서 [그림3-30]과 같이 SAS Enterprise Guide의 Add-In 메뉴로부터 Mass imputation을 선택하여 사용한다.



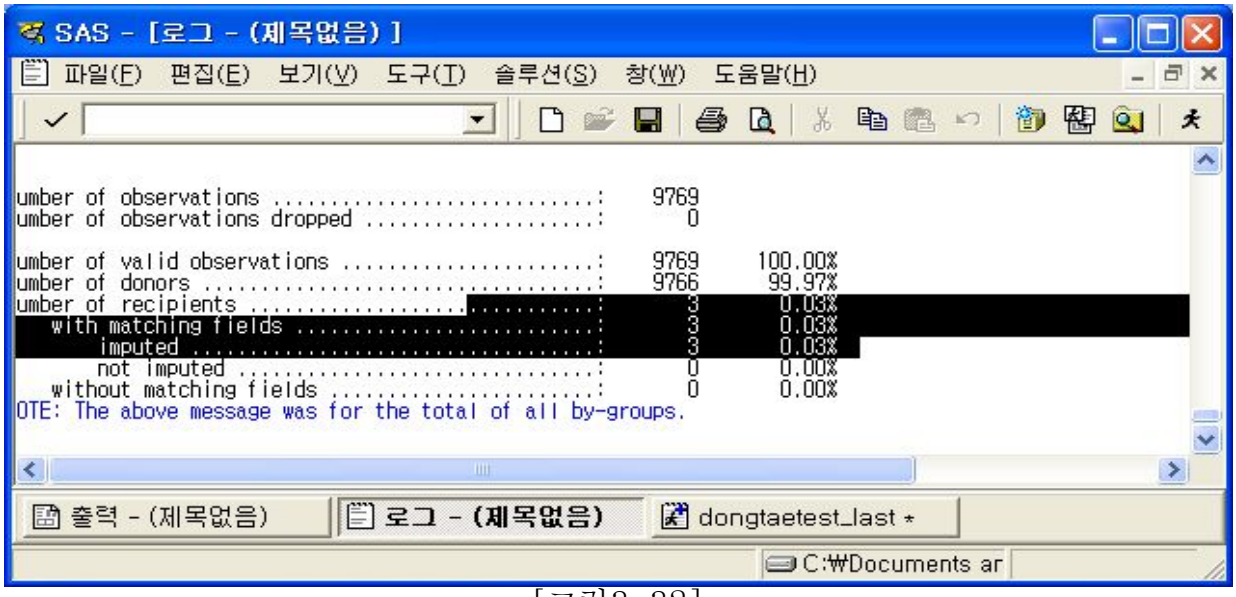
[그림3-30]

[그림3-31]은 Mass imputation을 하기에 앞서 matching field를 PRODEM(종업원수)와 WORKDAY(근무일수)로 두었으며, imputation을 원하는 변수들의 목록을 지정해주는 내용을 나타내고 있다. 여기서는 모든 항목에 대하여 무응답이므로 모든 항목을 지정해 주었다.



[그림3-31]

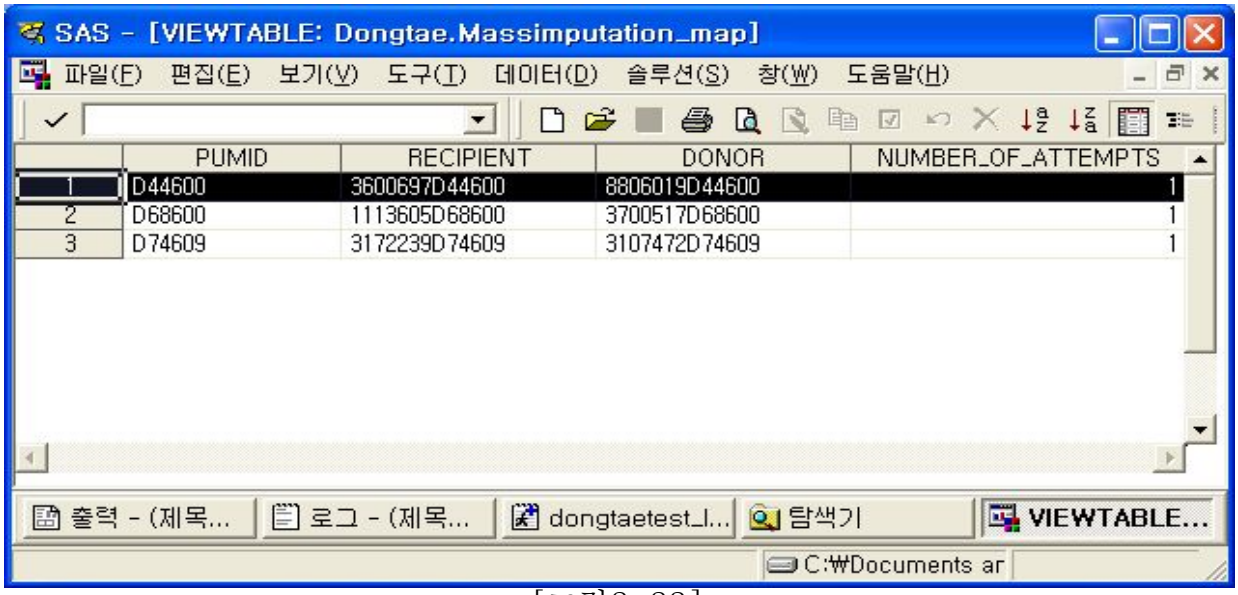
[그림3-32]에서는 Mass imputation을 한 3개의 사업체에 대하여 모두가 imputation이 되었다는 정보를 보여주고 있다.



[그림3-32]

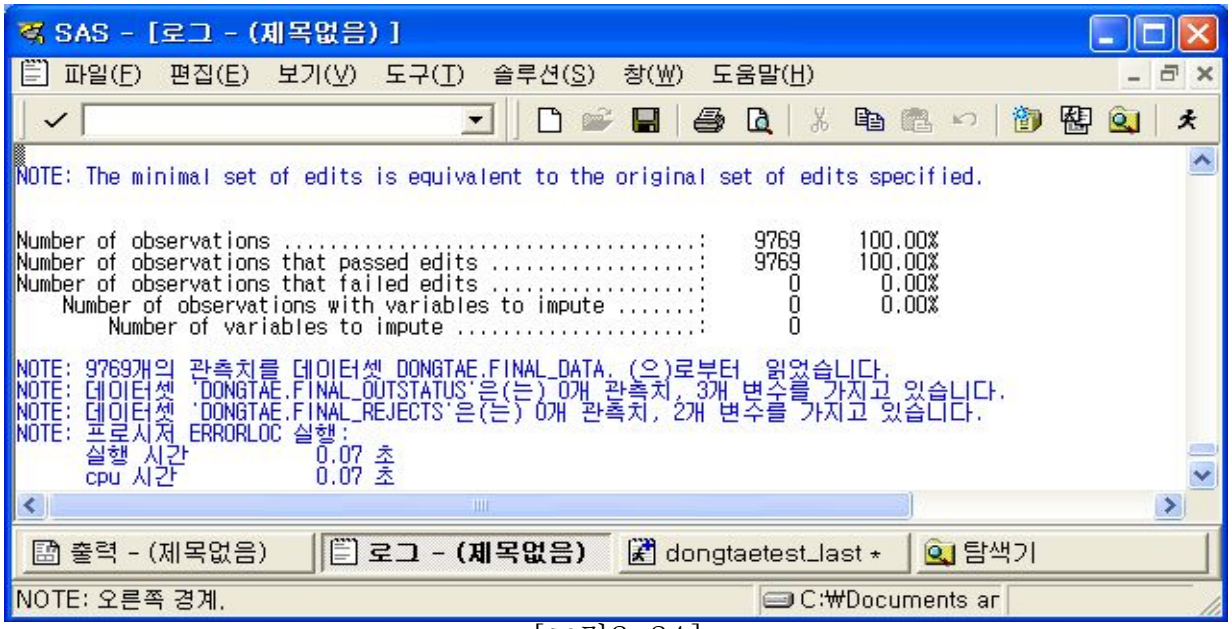
[그림3-33]에는 imputation을 해야 하는 3개의 개체(recipient)와 그것의 donor, 그리고 donor를 찾기 위하여 시도된 횟수를 보여주고 있다. Mass imputation에서는 post-imputation이 없으므로 시도횟수는 1이된다. 이는 가장 거리가 가까운 donor의 값을 그대로 쓰겠다는 것이다. 예를 들면, 사업체 3600697D44600은 사업체 8806019D44600의 값으로 imputation이 되었다는 것

을 나타낸다. 이 과정을 끝으로 광공업 동태자료에 대한 모든 imputation절차를 마치게 되며 마지막으로 edit rule을 다시 확인한다.



[그림3-33]

[그림3-34]는 모든 과정을 마친 후 다시 Error Localization을 수행한 결과이다. 총 사업체 9769개 모두가 모든 edit rule을 만족하기 때문에 광공업 동태자료에 대한 모든 imputation절차를 마무리하게 된다.



[그림3-34]

Ⅳ. 결론 및 향후 연구방향

4.1. 결론

사생활 보호 의식, 개인주의 확대, 기업비밀 보호 의식의 확대로 조사환경이 갈수록 악화되어지고 있다. 이러한 이유로 응답률이 낮아지고 조사된 자료의 품질이 저하되는 현상이 일어나고 있다. 따라서 본 연구는 자료의 품질을 향상시키기 위한 하나의 방법으로써 통계 선진국인 캐나다의 무응답처리 시스템인 Banff를 파악하고 이를 통계청의 조사에 적용을 시키고자 하였다. Banff 시스템의 가장 큰 장점은 9개의 체계화된 과정에 의하여 일괄적으로 edit와 imputation을 수행하는데 있다. 각 과정은 독립적으로 수행되며 수년에 걸친 시스템의 구축 및 보완으로 상당부분 정확한 측면을 보여주고 있다.

각 과정의 내부를 들여다보면 다양한 통계적 이론에 근거하여 구성되어 있음을 알 수 있다. 과거자료와의 관계 및 분포를 고려한 이상값 체크, Fellegi & Holt에 의해 제공된 최소변화규칙에 의한 내검과정, 적당한 donor를 찾기 위한 다양한 거리함수적용, imputation을 위한 여러 추정식의 적용 등 많은 통계적 이론을 바탕으로 시스템이 구축이 되어 있음을 알 수 있다. 또한 imputation이 되어지는 과정이나 그로 인한 결과들이 SAS의 dataset형태를 가지게 되어 결과물은 다른 프로시저의 입력 데이터로써 사용이 가능하며 체계적으로 관리도 용이한 시스템이라는 것을 알 수 있다.

그러나, Banff 시스템을 적용한 결과 보완해야 할 것들도 몇 가지 있는 것 같다. 이상값 체크시 체크 대상이 되는 변수에 대하여 모두 같은 규칙을 적용하게끔 되어 있다. 하지만 각 변수들의 분포가 다르기 때문에 다른 규칙이 적용될 수 있게 시스템이 보완되어야 할 것이다. 그리고 과거자료를 이용하여 분석할 때 여러 자료를 한꺼번에 분석할 수가 없기 때문에 연속적인 시계열은 이용하기가 힘든 점도 보완이 필요하다. 또한 부분의 합이 데이터의 전체 합과 일치하는지를 확인하는 과정인 Pro-Rating 프로시저에서는 차에 대한 것은 지원이 되지 않아 데이터 자체를 변환해야하는 번거로움이 있다. 이러한 내용에 대한 검토가 필요할 것이다.

본 연구에서는 이 시스템을 이용하여 광공업 동태자료에의 적용을 하였다. 체계적인 과정에 의한 분석은 가능하나 우리자료에 맞게 적용할 수 있는 연구가 필요

할 것으로 판단된다. 결국 우리의 자료에 맞는 내검규칙, 이상값 체크 분석, 적절한 거리함수 및 추정방법의 개발이 선행되어야 할 것이다. 이런 세부적인 연구를 시작으로 우리청도 체계적인 무응답처리 시스템을 구축하여 경제관련 조사들의 품질을 향상시킬 수 있는 방안이 마련되어야 하겠다.

4.2. 향후 연구방향

본 연구의 결과를 토대로 향후 연구방향을 제시하면 다음과 같다. Editing과 imputation을 일괄적으로 처리할 수 있는 체계적인 시스템을 장기적으로 개발을 해야 할 것이다. 센서스를 비롯해 연간조사, 분기조사 및 월간조사 모두에 적용 가능한 통합적인 시스템이 되어야 할 것이다.

이러한 시스템을 개발하기 위해서는 통계적 이론이 뒷받침되는 연구가 동시에 진행되어야 한다. 앞에서 설명한 내검규칙 및 이상값에 관한 모형과 imputation을 위한 여러 기법이 개발되어야 한다. 또한 조사특성 및 항목특성에 관한 분석을 통하여 가장 적합한 보조변수를 선정해 적용함으로써 더 효율적인 imputation이 되도록 하여야 한다.

우리 통계청은 지금까지 imputation에 관한 연구를 소홀히 한 것이 사실이다. 통계 선진국들은 오래전부터 이에 관한 연구를 계속적으로 진행하고 있으며, 연구한 것에 대하여 상당한 자부심을 가지고 있다. 이런 나라들은 우리나라와 조사환경이 다르다는 점도 고려해야겠지만 존재할 수밖에 없는 무응답을 인정하고 그것을 극복하고자하는 노력은 인정해 주어야 할 것으로 생각된다. 이번 캐나다 무응답처리 시스템 연수를 계기로 우리 통계청도 인식을 바꾸고 고품질의 통계를 생산하기 위해서라도 무응답처리에 관한 연구를 지속적으로 진행을 해야 할 것이다.