Dealing with Missing Data: Practical Use of Multiple Imputation

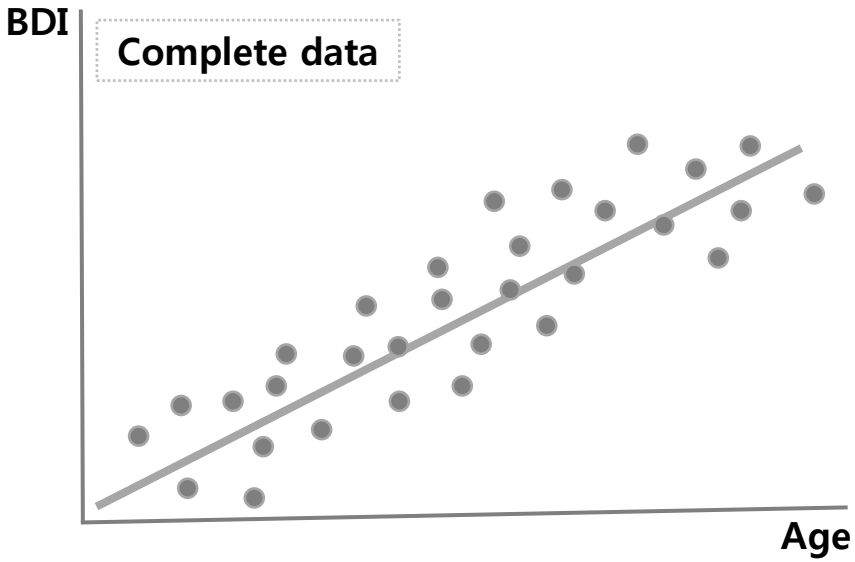
장명진
서울대병원 의학연구협력센터 의학통계실

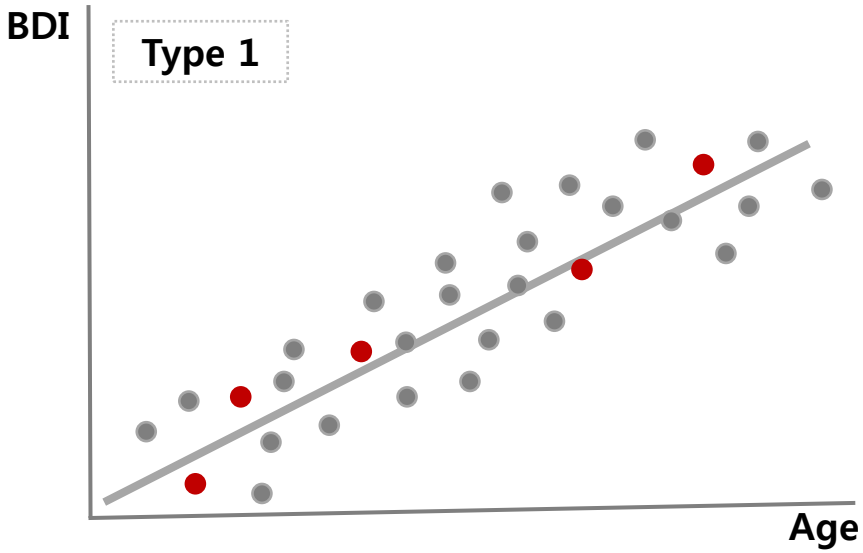
Example of missing data

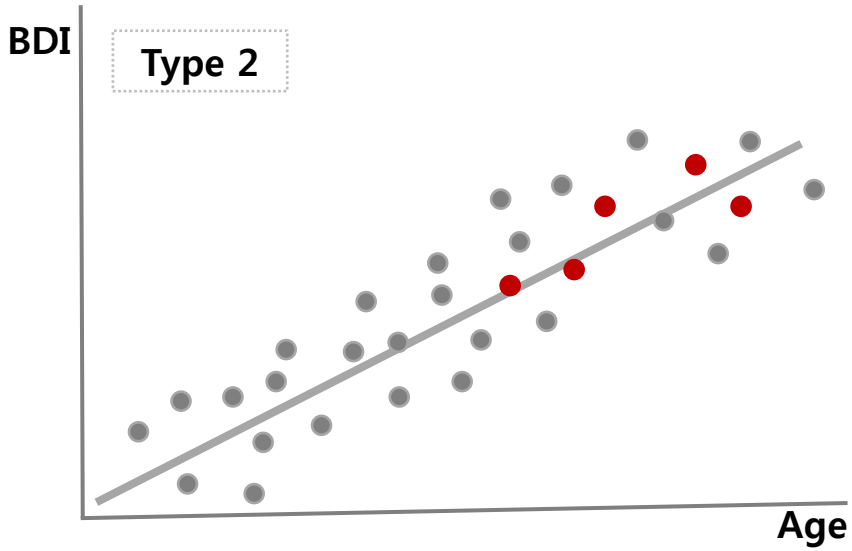
- 대상자: 40세 이상, 200명을 대상으로 우울증검사 실시
- 설문도구: 우울증 자가설문도구(Beck Depression Inventory)
- 연구목적: 우울증과 관련요인
- 결측자료 발생: 일부 대상자가 BDI 설문검사에 응답하지 않음

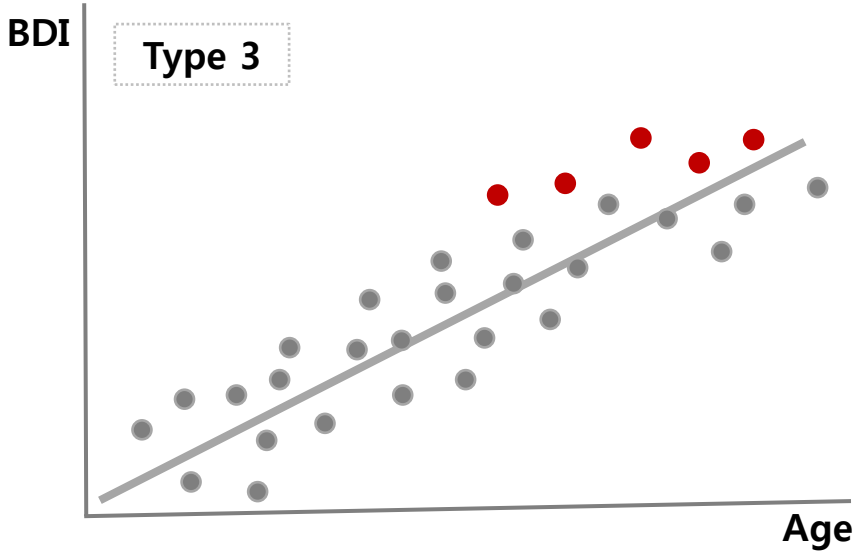
Id	Edu	…	Income	Sex	Age	Y (BDI-II, Depression score)
1	1		1	1	50	25
2	2		2	2	45	15
3	1		3	2	40	18
4	4		1	1	65	.
…	3		2	1	72	37
198	2		2	1	57	10
199	3		3	2	46	.
200	4		4	1	79	32

Types of missing





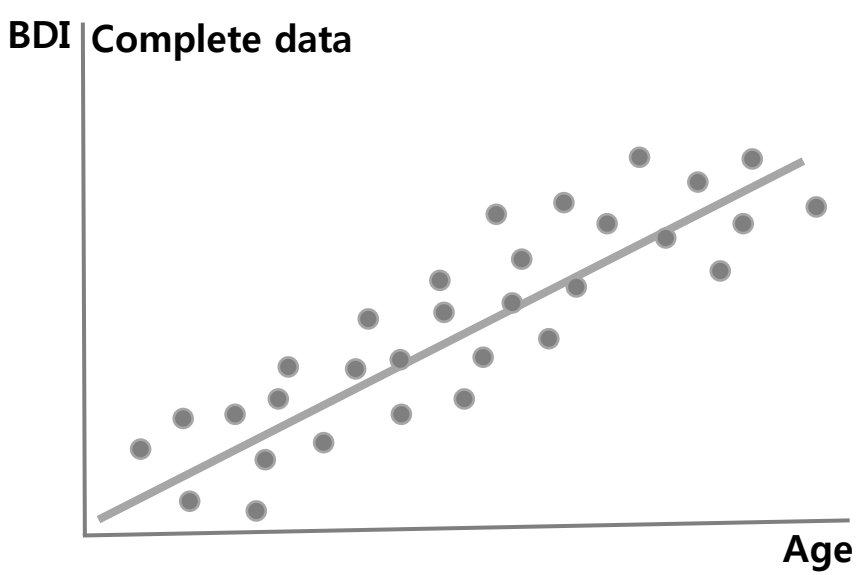


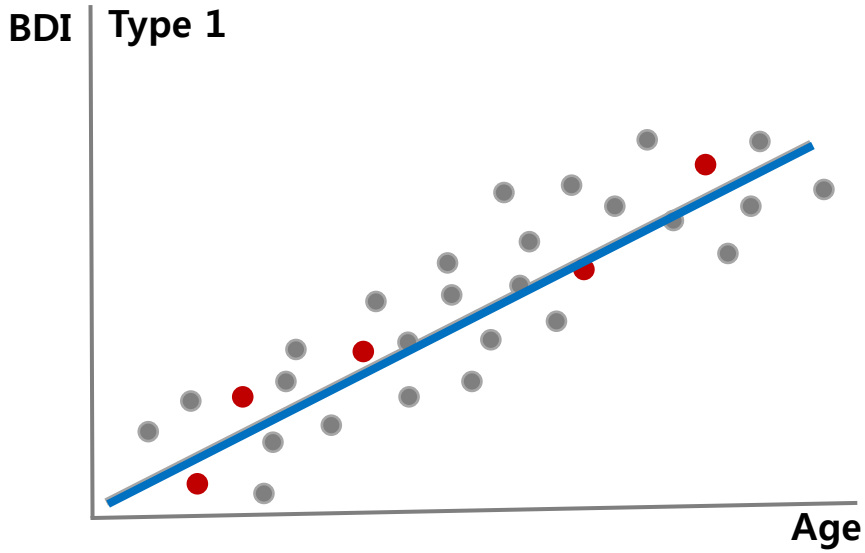


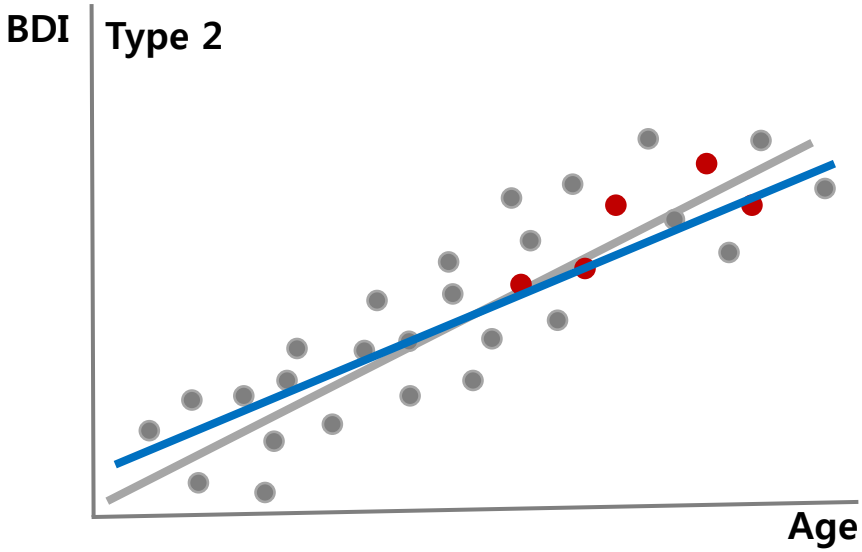
Types of missing

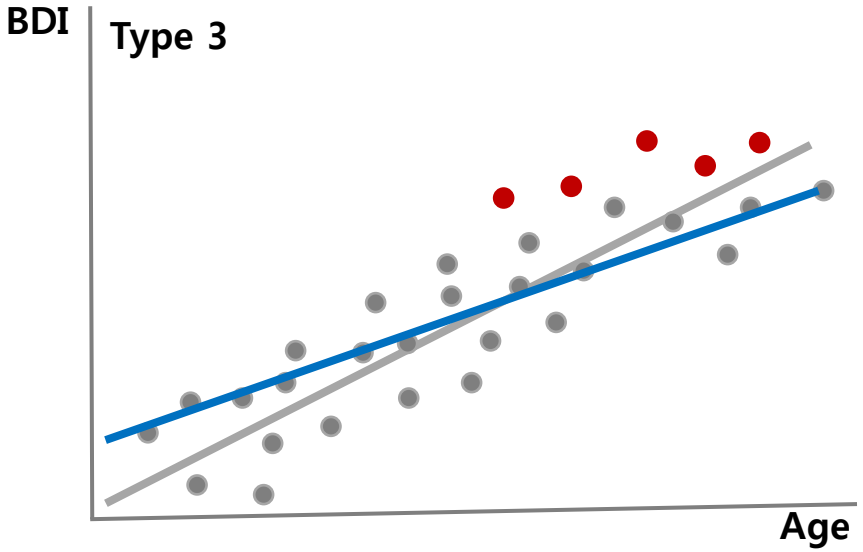
- **Type 1 : Missing Completely At Random (MCAR)**
  - Pr(Y is missing|X,Y) = Pr(Y is missing)
  - Pr(missing) is unrelated to both observed(X) and unobserved data(Y)
  - Eg: accidentally missing
- **Type 2: Missing At Random (MAR)**
  - Pr(Y is missing|X,Y) = Pr(Y is missing|X)
  - Pr(missing) depends only on observed data(X)
  - Eg: High missing in high X (high missing in older people)
- **Type 3: Missing Not At Random (MNAR)**
  - Pr(missing) depends on both observed(X) and unobserved data (Y)
  - Eg: High missing in high X and Y (high missing in older and high BDI)

Complete case analysis(관찰된 자료만 이용)









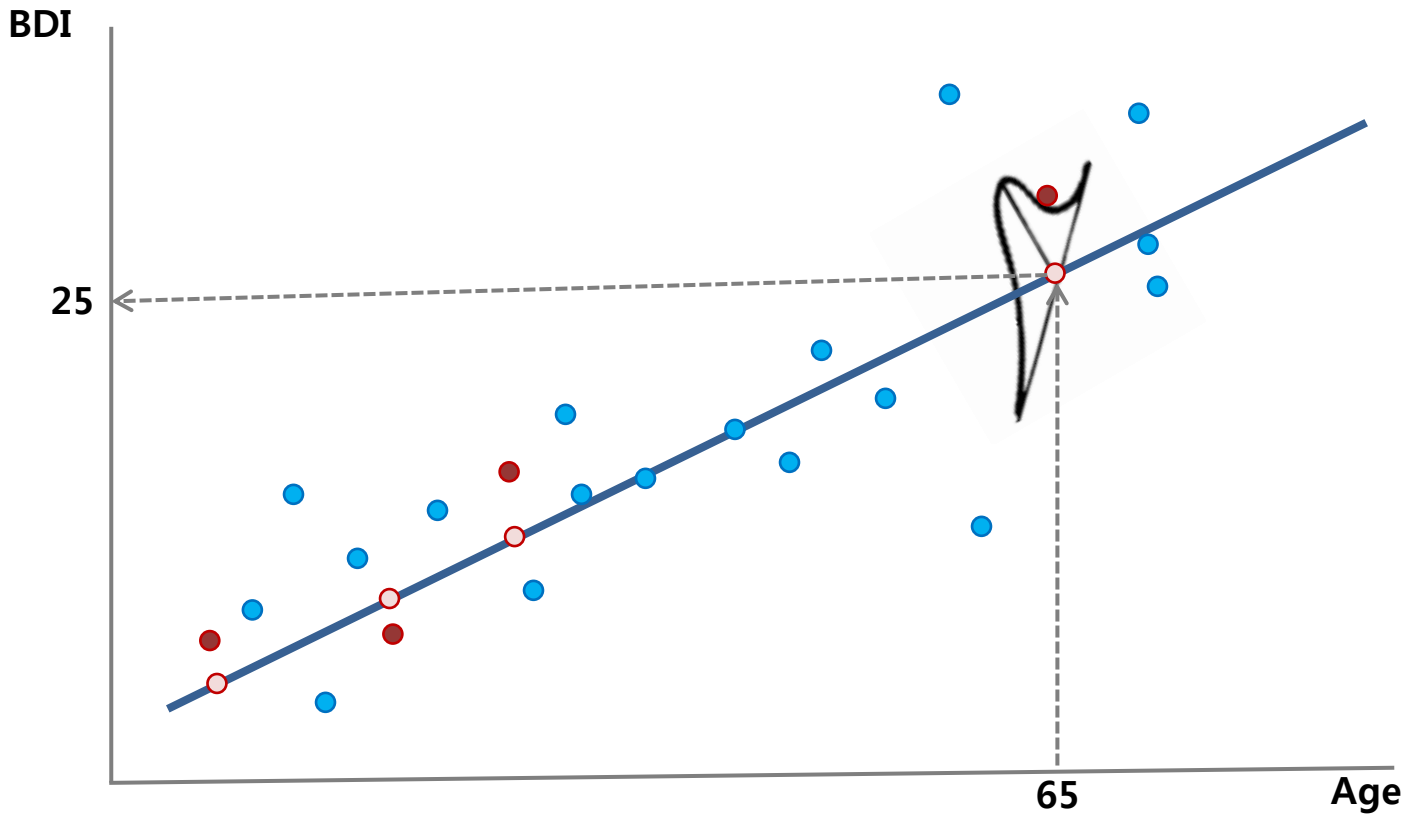
Complete case analysis: ignoring missing

- **MCAR (Type 1) : accidentally missing**
  - 응답여부가 Y, X와 무관하게 발생
  - 즉 응답자와 무응답자의 Y의 분포가 다르지 않음(not different)
  - 응답자만 가지고 분석결과는 응답자와 무응답자 모든 자료의 분석결과와 같음(unbiased estimates)

- **MAR (Type 2) : high missing in high X**
  - missing을 무시하는 경우, Y의 추정치 또는 Y와 X의 관련성 추정치가 biased

- **MNAR (Type 3) : high missing in high Y**
  - missing을 무시하는 경우, Y의 추정치 또는 Y와 X의 관련성 추정치가 biased

Missing Imputation

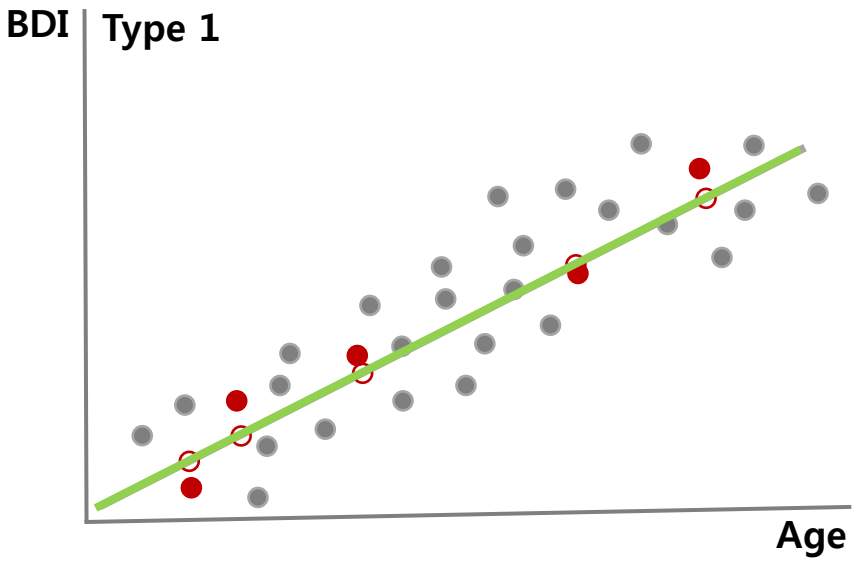


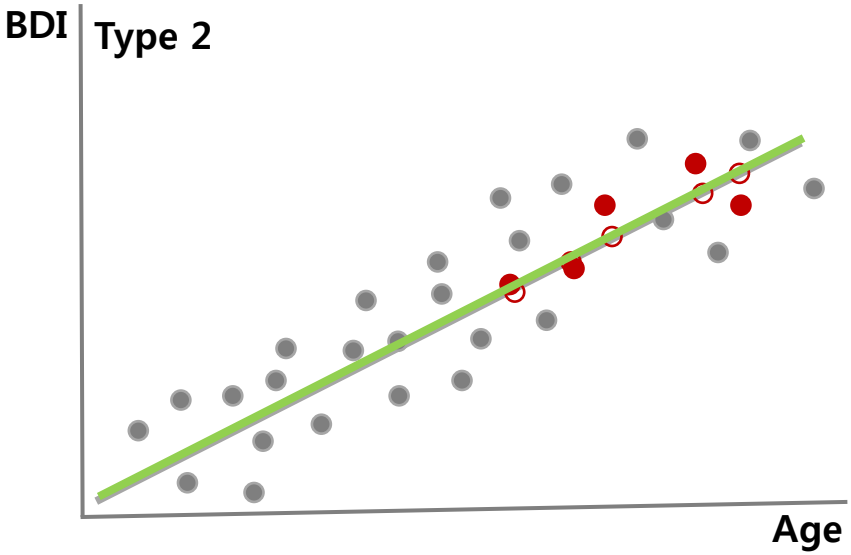
Imputation model 1: $BDI_i = \beta_0 + \beta_1 age_i$ **: 65세의 BDI missing은 모두 25로 대체**

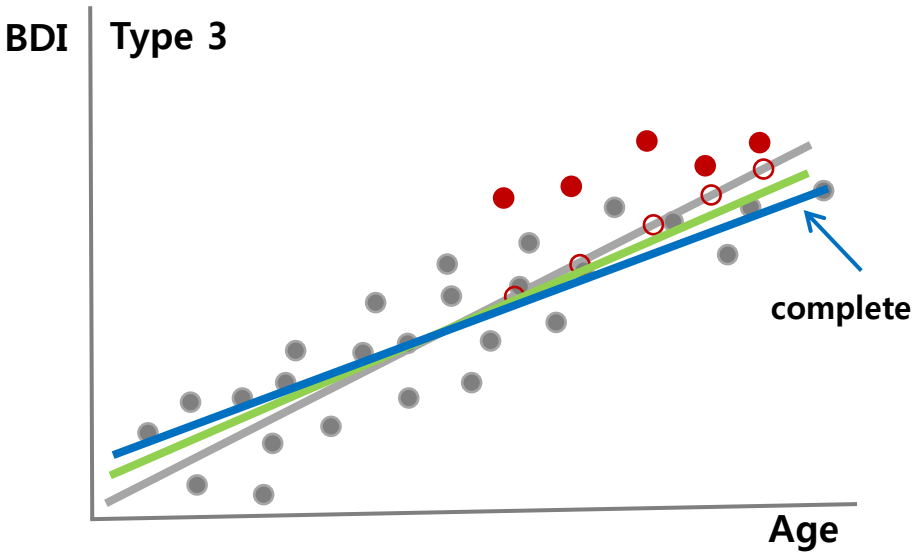
Imputation model 2: $BDI_i = \beta_0 + \beta_1 age_i + e_i, \quad e_i \sim N(0, \sigma^2)$

: 65세의 BDI의 평균(25)과 분산(σ^2)을 반영하여 imputed value 추출, 25, 23, 27, 30 등 다양

Imputed data analysis







Missing types and Imputation

- **MCAR (Type 1) : accidentally missing**
  - Both complete case analysis and imputed data analysis give unbiased estimates, but imputed data analysis gives higher statistical power than complete case analysis.

- **MAR (Type 2) : high missing in high X**
  - Proper imputation gives unbiased estimates while the estimates from complete data analysis could be biased estimates.

- **MNAR (Type 3) : high missing in high Y**
  - Imputation could not give unbiased estimates, but the bias could be less than that of complete case analysis.

Imputation: Single and Multiple imputation

- **Problem in single imputation**
  - If data are missing at random (MAR, MCAR), the estimates are unbiased.
  - However, imputed data were treated as if they were real data, so the standard error of estimates are underestimates.

- **Solution to single imputation**
  - Multiple imputation : make multiple imputed datasets and consider the uncertainty in imputation.

Example of Multiple Imputation Analysis

- **Imputed datasets**

							Imputed values		
Id	Edu	…	Income	Sex	Age	BDI	Set 1	Set 2	Set 3
1	1		1	1	50	25			
2	2		2	2	45	15			
3	1		3	2	40	18			
4	4		1	1	65	.	25	23	27
…	3		2	1	72	37			
198	2		2	1	57	10			
199	3		3	2	46	.	16	20	14
200	4		4	1	79	32			

- **Imputed data와 observed data를 이용하여 분석**

Example of Multiple Imputation Analysis

- 연구목적: Age와 BDI의 관련성 추정(BDI = β_0+ β_1 age)

- Imputed data와 observed data를 이용하여 분석
  - Set 1: BDI= 2+0.35 age : Age와 BDI의 관련성 0.35(SE=0.12)
  - Set 2: BDI= 2+0.4 age : Age와 BDI의 관련성 0.4 (SE=0.15)
  - Set 3: BDI= 2+0.25 age : Age와 BDI의 관련성 0.3 (SE=0.10)

  각 Set의 결과는 Single imputation의 결과임

- Age와 BDI의 관련성 결과(estimate and its standard error)?
  - 추정치는 각 set에서의 추정치의 평균 ???
  - 추정치의 SE는 각 set에서의 추정치의 SE의 평균 ???

Estimate and SE in multiple imputation

	Estimate($\hat{\beta}$)	SE($\hat{\beta}$)	Var($\hat{\beta}$)=SE^2
Set 1	0.35	0.12	0.0144
Set 2	0.4	0.15	0.0225
Set 3	0.3	0.10	0.01
평균	0.35		0.0156

$$\rightarrow SE(\hat{\beta}) = \sqrt{0.0156} = 0.126$$

This SE estimate does not consider the uncertainty in imputation

$$\bar{\beta} = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}^{(k)} = (0.35 + 0.4 + 0.3)/3 = 0.35$$

$$Var(\bar{\beta}) = \frac{1}{m}\sum_{k=1}^{m} Var(\hat{\beta}^{(k)}) + \left(1 + \frac{1}{m}\right)\frac{1}{m-1}\sum_{k=1}^{m}\left(\hat{\beta}^{(k)} - \bar{\beta}\right)^2$$

$$= \frac{0.12^2 + 0.15^2 + 0.10^2}{3} + \left(1 + \frac{1}{3}\right)\frac{1}{2}\left((0.35 - 0.35)^2 + (0.4 - 0.35)^2 + (0.3 - 0.35)^2\right)$$

$$= 0.0156 + 0.0033 = 0.019 \qquad \rightarrow SE(\bar{\beta}) = \sqrt{0.019} = 0.138$$

Estimate and SE in multiple imputation

- **Estimate of multiple imputation :**

$$\bar{\beta} = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}^{(k)}$$

- **Variance of multiple imputation**

$$Var(\bar{\beta}) = \underbrace{\frac{1}{m}\sum_{k=1}^{m}Var(\hat{\beta}^{(k)})}_{\text{Within-imputation variance}} + \underbrace{\left(1+\frac{1}{m}\right)\frac{1}{m-1}\sum_{k=1}^{m}\left(\hat{\beta}^{(k)}-\bar{\beta}\right)^2}_{\text{Between-imputation variance}}$$

β=0.35 SE=0.12 | β=0.4 SE=0.15 | β=0.3 SE=0.10

추정치들의 표준오차의 평균

$=(0.12^2+0.15^2+0.10^2)/3$

β=0.35 SE=0.12 | β=0.4 SE=0.15 | β=0.3 SE=0.10

Imputation에 따른 추정치들의 변동

$=\{(0.35-0.35)^2+(0.4-0.35)^2+(0.3-0.35)^2\}/2$

Imputation set이 많아지면, multiple imputation에서 추정치의 분산 감소

그럼 몇 개의 imputation set이 적절할까?

m? 일반적으로 5개 정도

Multiple imputation 예제

- 자료
  - 연구 목적: antisocial behavior와 관련요인
  - 자료원: National Longitudinal Survey of Youth
  - 581 children, surveyed in 1990
  - 10개의 변수 중 5개의 변수에 결측치 있음: SELF, POV, BLACK, HISPANIC, MOMWORK

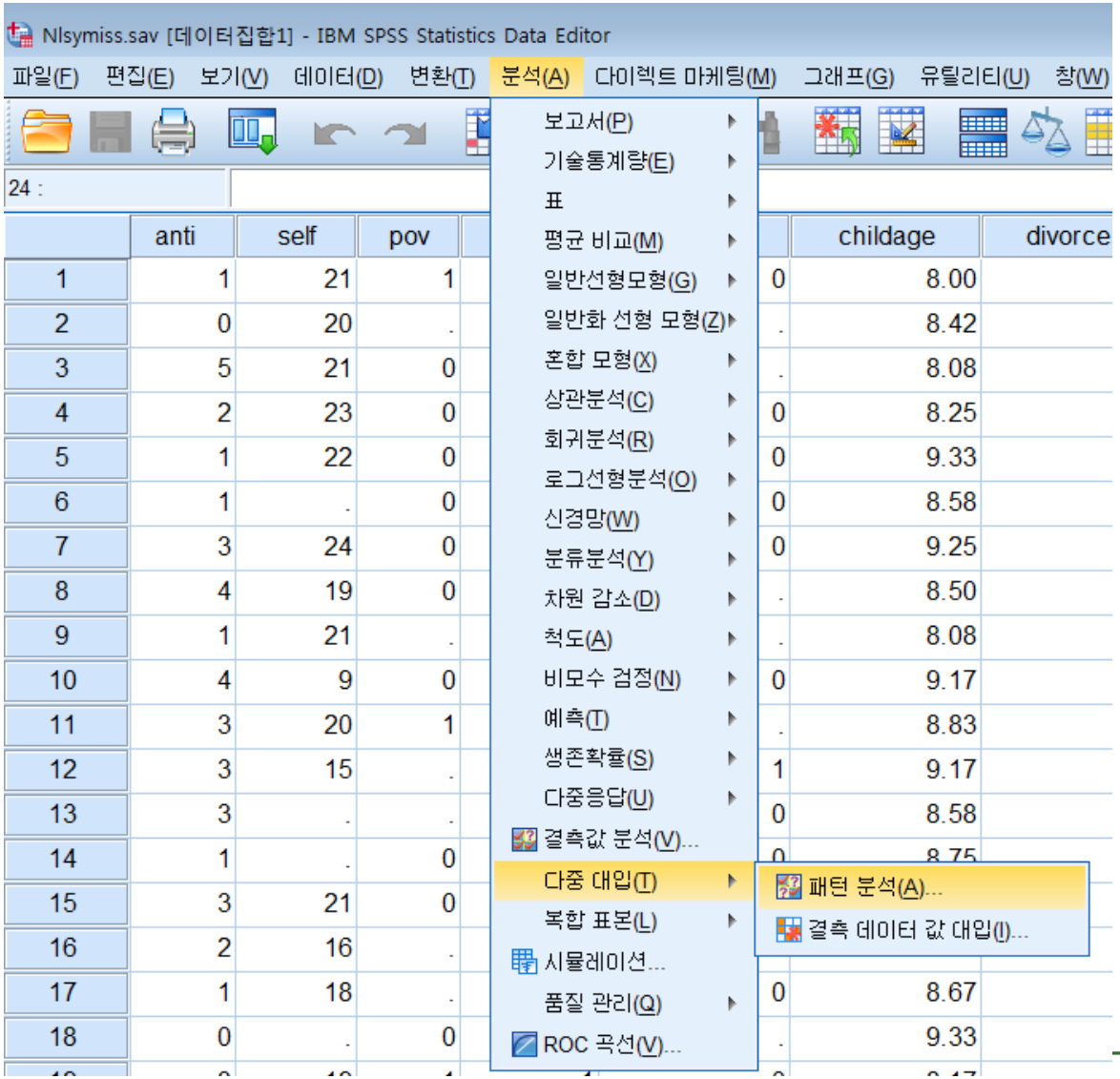
http://statisticalhorizons.com/resources/data-sets

자료 설명

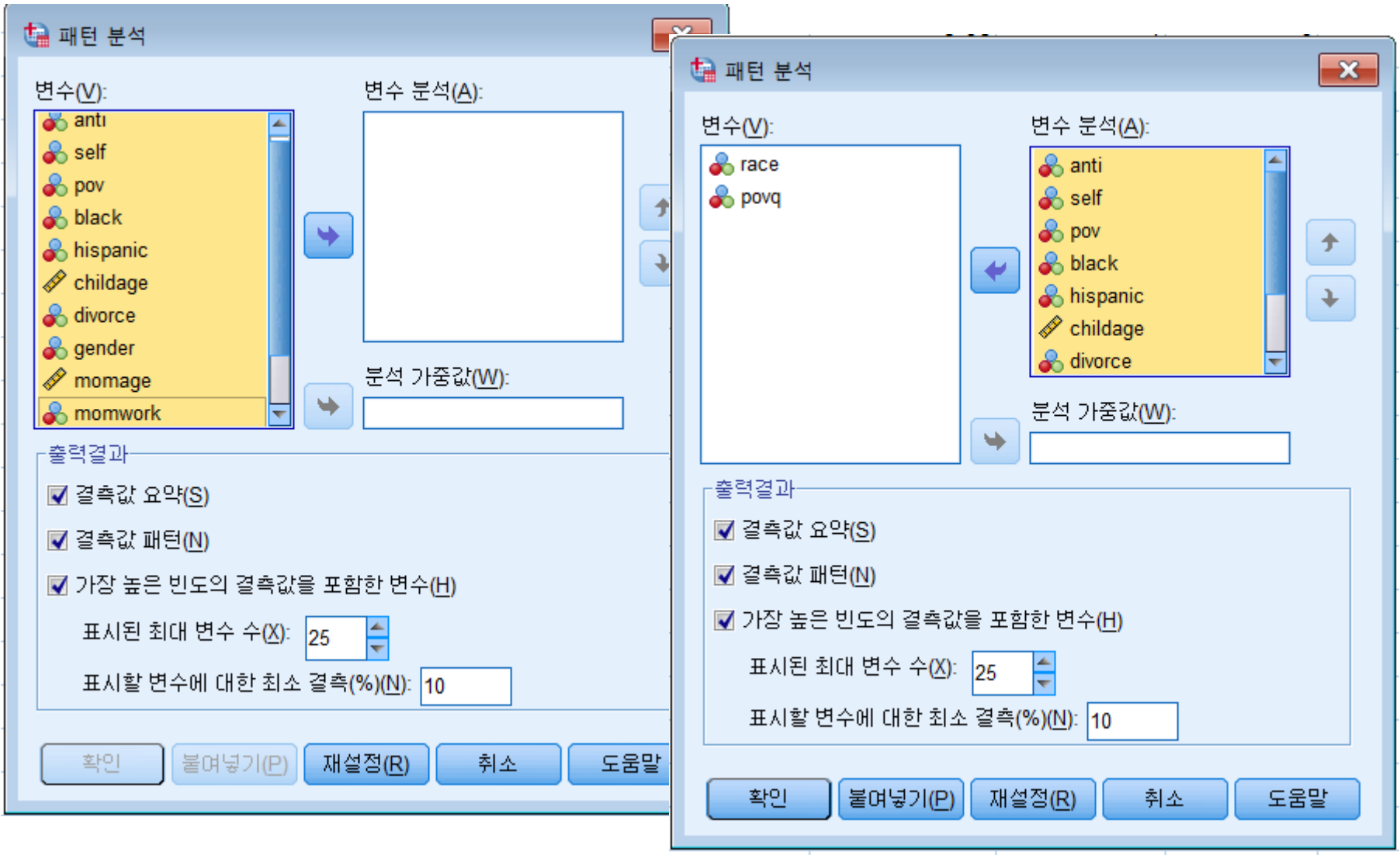
	anti	self	pov	black	hispanic	childage	divorce	gender	momage	momwork
1	1	21	1	0	0	8.00	0	1	21	0
2	0	20	.	.	.	8.42	0	1	22	1
3	5	21	0	.	.	8.08	1	0	18	0
4	2	23	0	0	0	8.25	0	0	24	0
5	1	22	0	0	0	9.33	0	1	22	0
6	1	.	0	0	0	8.58	0	0	24	0
7	3	24	0	0	0	9.25	1	1	23	.
8	4	19	0	.	.	8.50	1	0	18	0
9	1	21	.	.	.	8.08	0	0	24	0

ANTI	antisocial behavior, measured with a scale ranging from 0 to 6.
SELF	self-esteem, measured with a scale ranging from 6 to 24.
POV	poverty status of family, coded 1 for in poverty, otherwise 0.
BLACK	1 if child is black, otherwise 0
HISPANIC	1 if child is Hispanic, otherwise 0
CHILDAGE	child's age in 1990
DIVORCE	1 if mother was divorced in 1990, otherwise 0
GENDER	1 if female, 0 if male
MOMAGE	mother's age at birth of child
MOMWORK	1 if mother was employed in 1990, otherwise 0

결측 패턴 분석

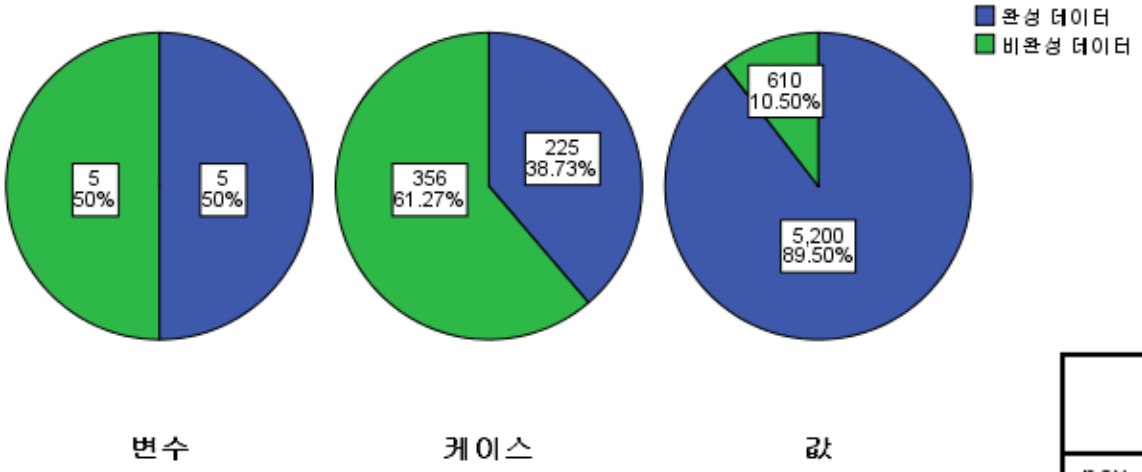

NIsymiss.sav [데이터집합1] - IBM SPSS Statistics Data Editor
파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W)
24 :
anti self pov childage divorce
1 1 21 1 0 8.00
2 0 20 . . 8.42
3 5 21 0 . 8.08
4 2 23 0 0 8.25
5 1 22 0 0 9.33
6 1 . 0 0 8.58
7 3 24 0 0 9.25
8 4 19 0 . 8.50
9 1 21 . . 8.08
10 4 9 0 0 9.17
11 3 20 1 . 8.83
12 3 15 . 1 9.17
13 3 . . 0 8.58
14 1 . 0
15 3 21 0
16 2 16 .
17 1 18 . 0 8.67
18 0 . 0 . 9.33
보고서(P)
기술통계량(E)
표
평균 비교(M)
일반선형모형(G)
일반화 선형 모형(Z)
혼합 모형(X)
상관분석(C)
회귀분석(R)
로그선형분석(O)
신경망(W)
분류분석(Y)
차원 감소(D)
척도(A)
비모수 검정(N)
예측(T)
생존확률(S)
다중응답(U)
결측값 분석(V)...
다중 대입(T)
복합 표본(L)
시뮬레이션...
품질 관리(Q)
ROC 곡선(V)...
패턴 분석(A)...
결측 데이터 값 대입(I)...

패턴분석 실행


패턴 분석
변수(V):
anti
self
pov
black
hispanic
childage
divorce
gender
momage
momwork
변수 분석(A):
분석 가중값(W):
출력결과
결측값 요약(S)
결측값 패턴(N)
가장 높은 빈도의 결측값을 포함한 변수(H)
표시된 최대 변수 수(X): 25
표시할 변수에 대한 최소 결측(%)(N): 10
확인
붙여넣기(P)
재설정(R)
취소
도움말
패턴 분석
변수(V):
race
povq
변수 분석(A):
anti
self
pov
black
hispanic
childage
divorce
분석 가중값(W):
출력결과
결측값 요약(S)
결측값 패턴(N)
가장 높은 빈도의 결측값을 포함한 변수(H)
표시된 최대 변수 수(X): 25
표시할 변수에 대한 최소 결측(%)(N): 10
확인
붙여넣기(P)
재설정(R)
취소
도움말

결측값

결측값의 전체 요약

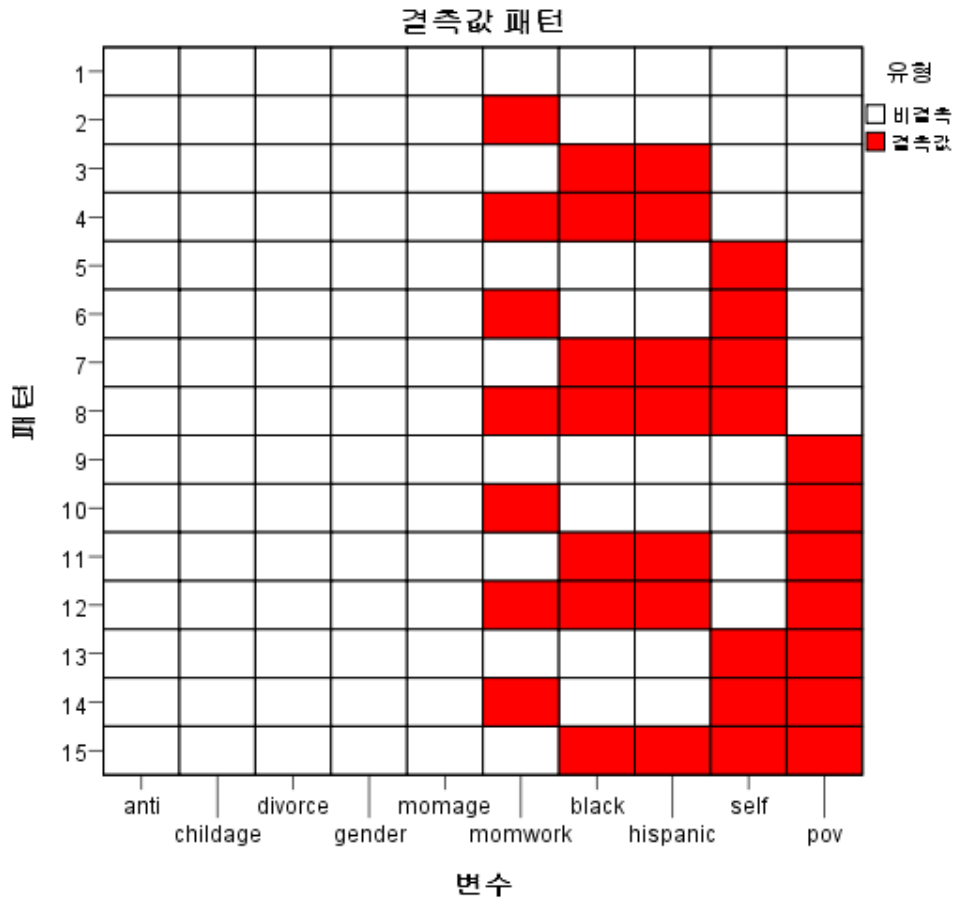


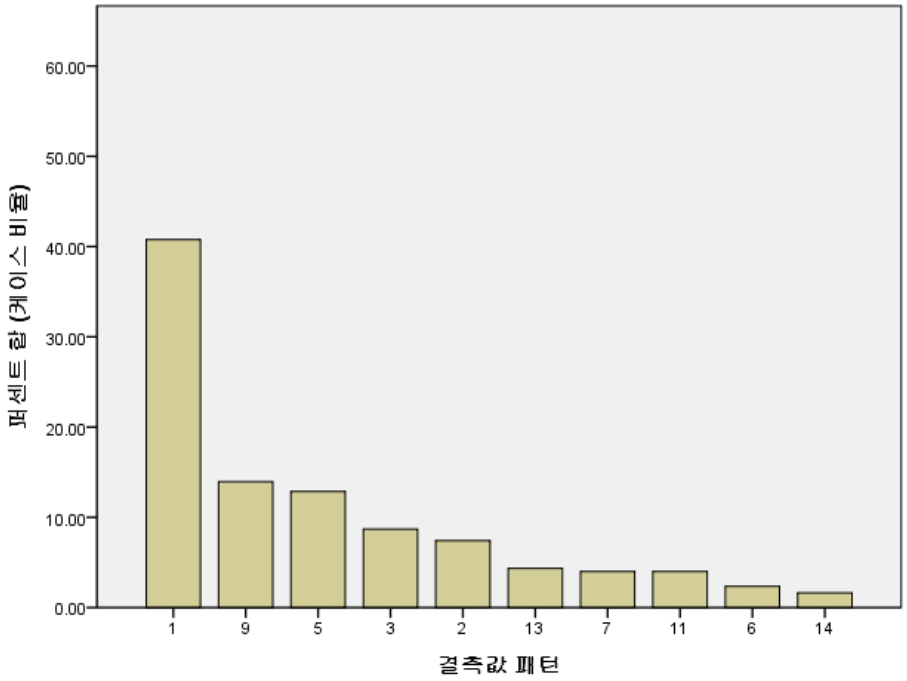
변수 요약[a,b]

	결측값		
	N	백분율	유효수
pov	150	25.8%	431
self	148	25.5%	433
hispanic	113	19.4%	468
black	113	19.4%	468
momwork	86	14.8%	495

a. 나타난 최대 변수 수: 25

b. 포함할 변수에 대한 최소 결측값(%): 10.0%

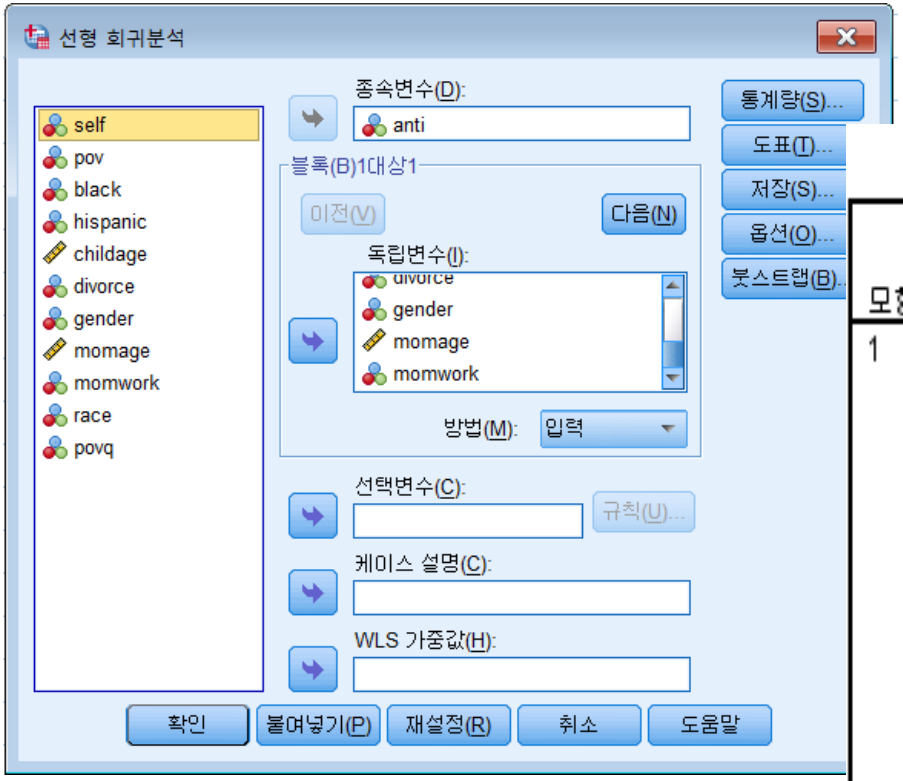




10개의 가장 빈번하게 발생하는 패턴이 차트에 표시되어 있습니다.

분석 1: complete case analysis(결측자료 무시)

- Anti-social behavior와 관련요인 분석

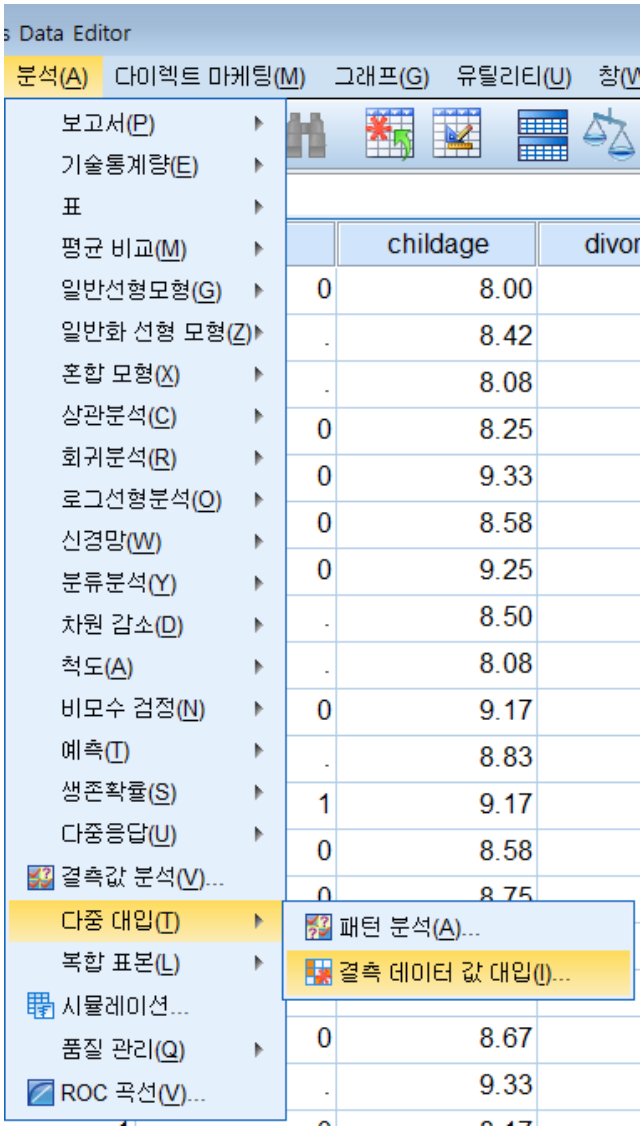


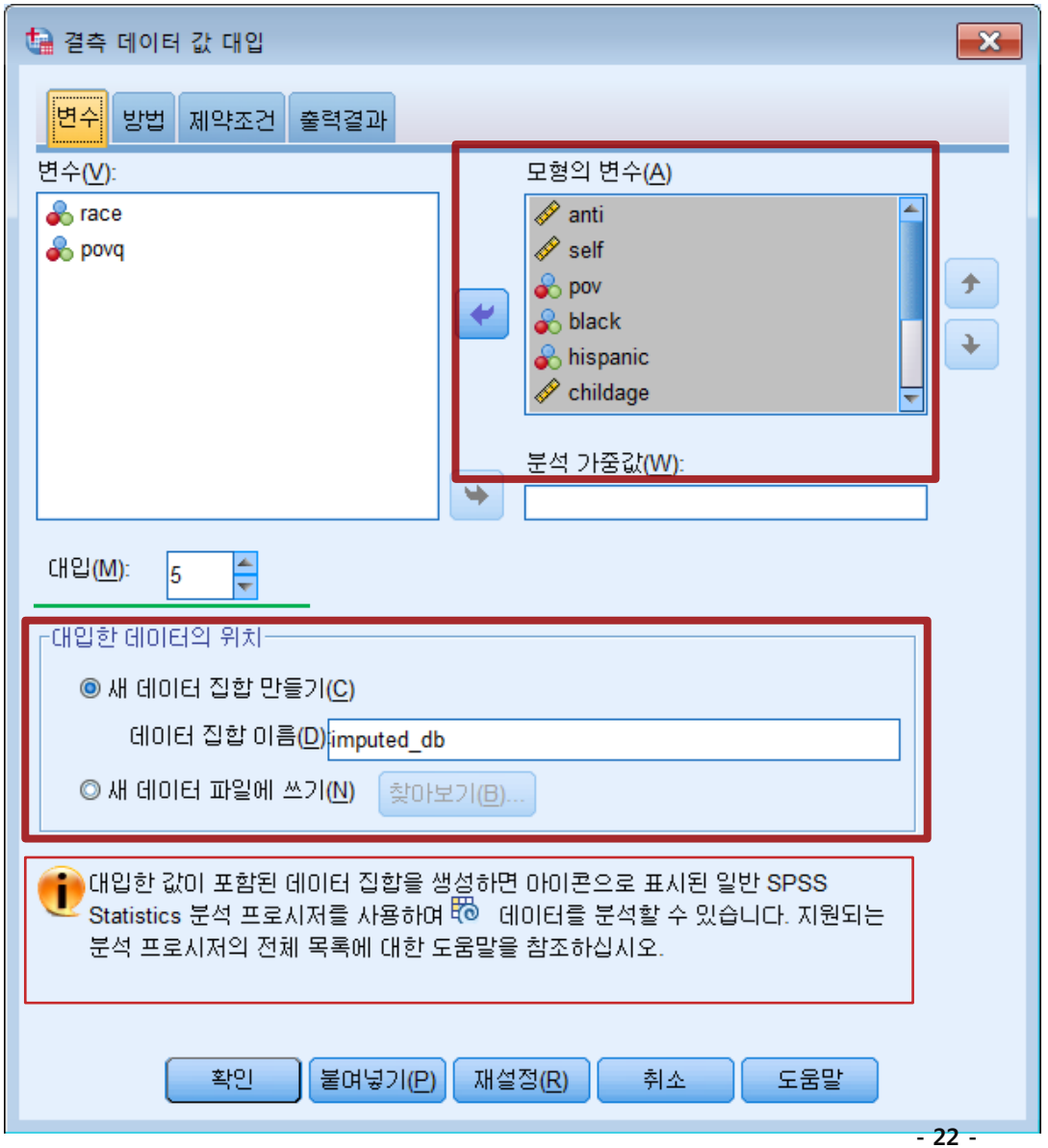
계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	2.865	1.991		1.439	.152
	self	-.045	.031	-.097	-1.445	.150
	pov	.719	.237	.230	3.031	.003
	black	.051	.249	.016	.203	.839
	hispanic	-.357	.255	-.104	-1.398	.164
	childage	.002	.171	.001	.012	.991
	divorce	.087	.245	.024	.355	.723
	gender	-.335	.198	-.110	-1.687	.093
	momage	-.012	.046	-.017	-.260	.795
	momwork	.254	.218	.081	1.170	.243

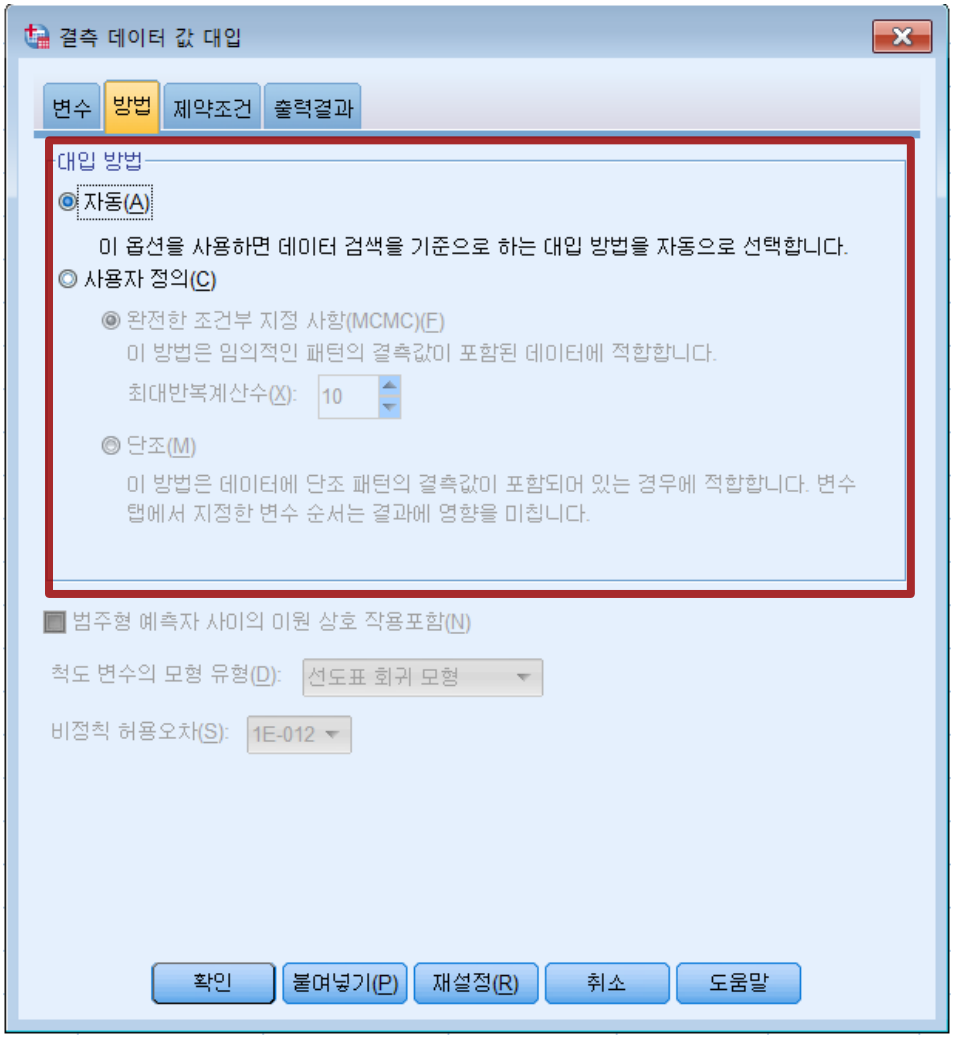
a. 종속변수: anti

분석 2: 결측값 대체_변수지정

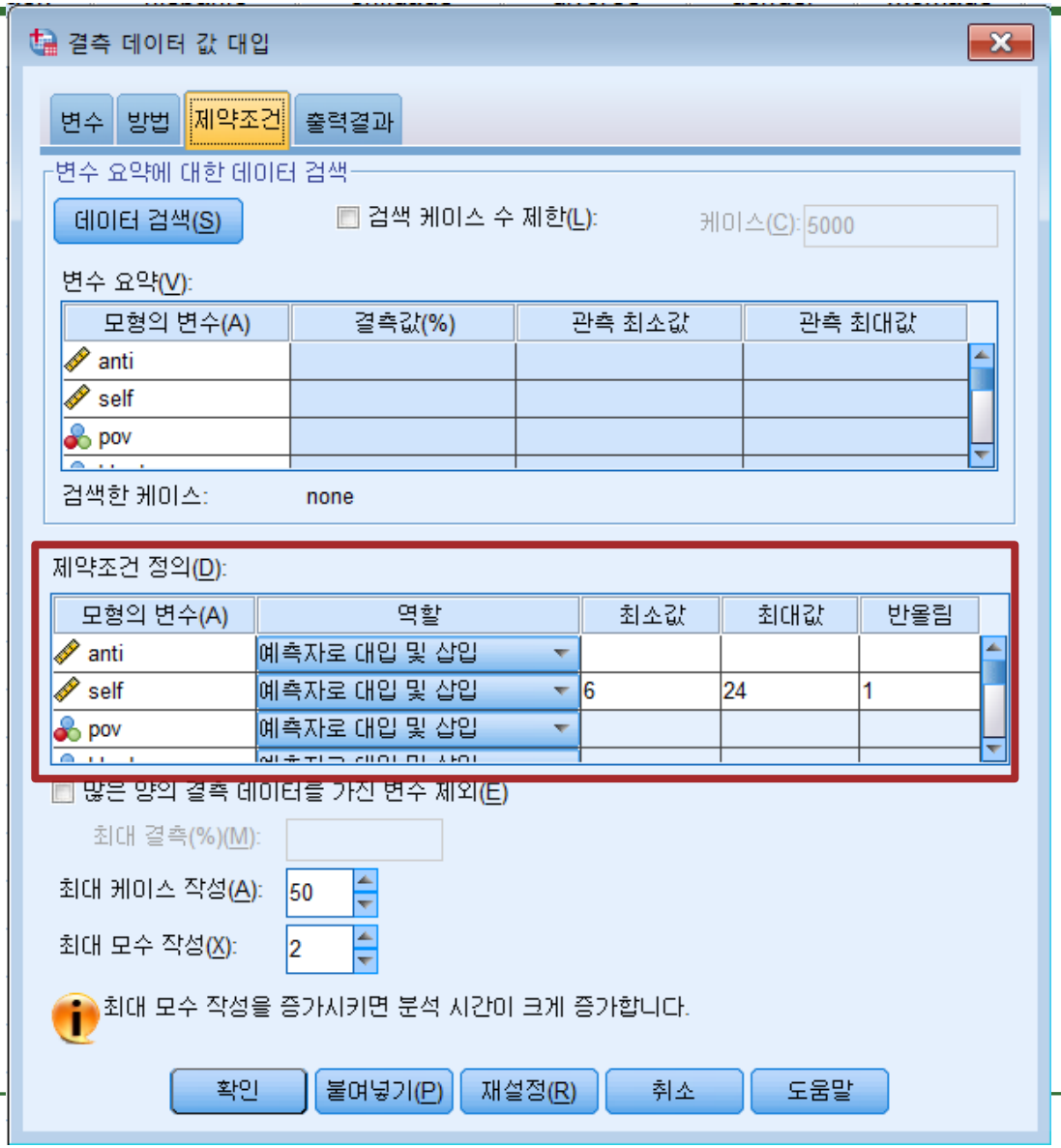




결측값 대체_방법 지정


결측 데이터 값 대입
변수
방법
제약조건
출력결과
대입 방법
자동(A)
이 옵션을 사용하면 데이터 검색을 기준으로 하는 대입 방법을 자동으로 선택합니다.
사용자 정의(C)
완전한 조건부 지정 사항(MCMC)(F)
이 방법은 임의적인 패턴의 결측값이 포함된 데이터에 적합합니다.
최대반복계산수(X):
10
단조(M)
이 방법은 데이터에 단조 패턴의 결측값이 포함되어 있는 경우에 적합합니다. 변수 탭에서 지정한 변수 순서는 결과에 영향을 미칩니다.
범주형 예측자 사이의 이원 상호 작용포함(N)
척도 변수의 모형 유형(D):
선도표 회귀 모형
비정칙 허용오차(S):
1E-012
확인
붙여넣기(P)
재설정(R)
취소
도움말

결측값 대체_제약 조건


결측 데이터 값 대입
변수
방법
제약조건
출력결과
변수 요약에 대한 데이터 검색
데이터 검색(S)
검색 케이스 수 제한(L):
케이스(C): 5000
변수 요약(V):
모형의 변수(A)
결측값(%)
관측 최소값
관측 최대값
anti
self
pov
검색한 케이스: none
제약조건 정의(D):
모형의 변수(A)
역할
최소값
최대값
반올림
anti
예측자로 대입 및 삽입
self
예측자로 대입 및 삽입
6
24
1
pov
예측자로 대입 및 삽입
많은 양의 결측 데이터를 가진 변수 제외(E)
최대 결측(%)(M):
최대 케이스 작성(A): 50
최대 모수 작성(X): 2
최대 모수 작성을 증가시키면 분석 시간이 크게 증가합니다.
확인
붙여넣기(P)
재설정(R)
취소
도움말

결측값 대체 결과

대입 제약

	대입에서의 역할		대입된 값		
	종속	예측자	최소값	최대값	반올림
anti	예	예	(지정않음)	(지정않음)	
self	예	예	6	24	정수
pov	예	예			
black	예	예			
hispanic	예	예			
childage	예	예	(지정않음)	(지정않음)	
divorce	예	예			
gender	예	예			
momage	예	예	(지정않음)	(지정않음)	
momwork	예	예			

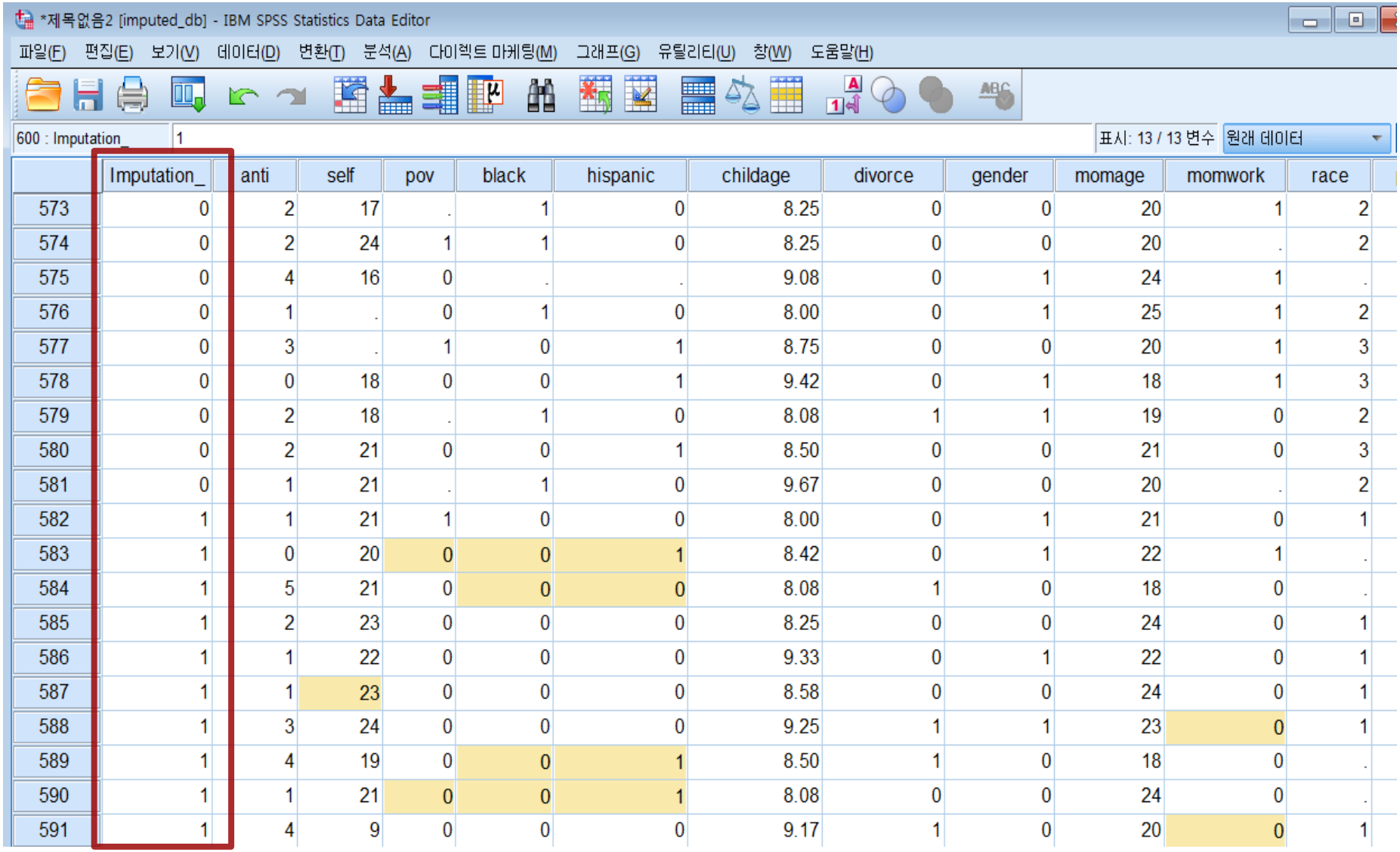
대입 결과

대입 방법		완전한 조건 지정 사항
완전한 조건 지정 방법 반복계산		10
종속변수	대입	self,pov,black,hispanic,momwork
	대입되지 않음(결측값이 너무 많음)	
	대입되지 않음(결측값 없음)	anti,childage,divorce,gender,momage
대입 시퀀스		anti,childage,divorce,gender,momage,momwork,black, hispanic,self,pov

대입 모형

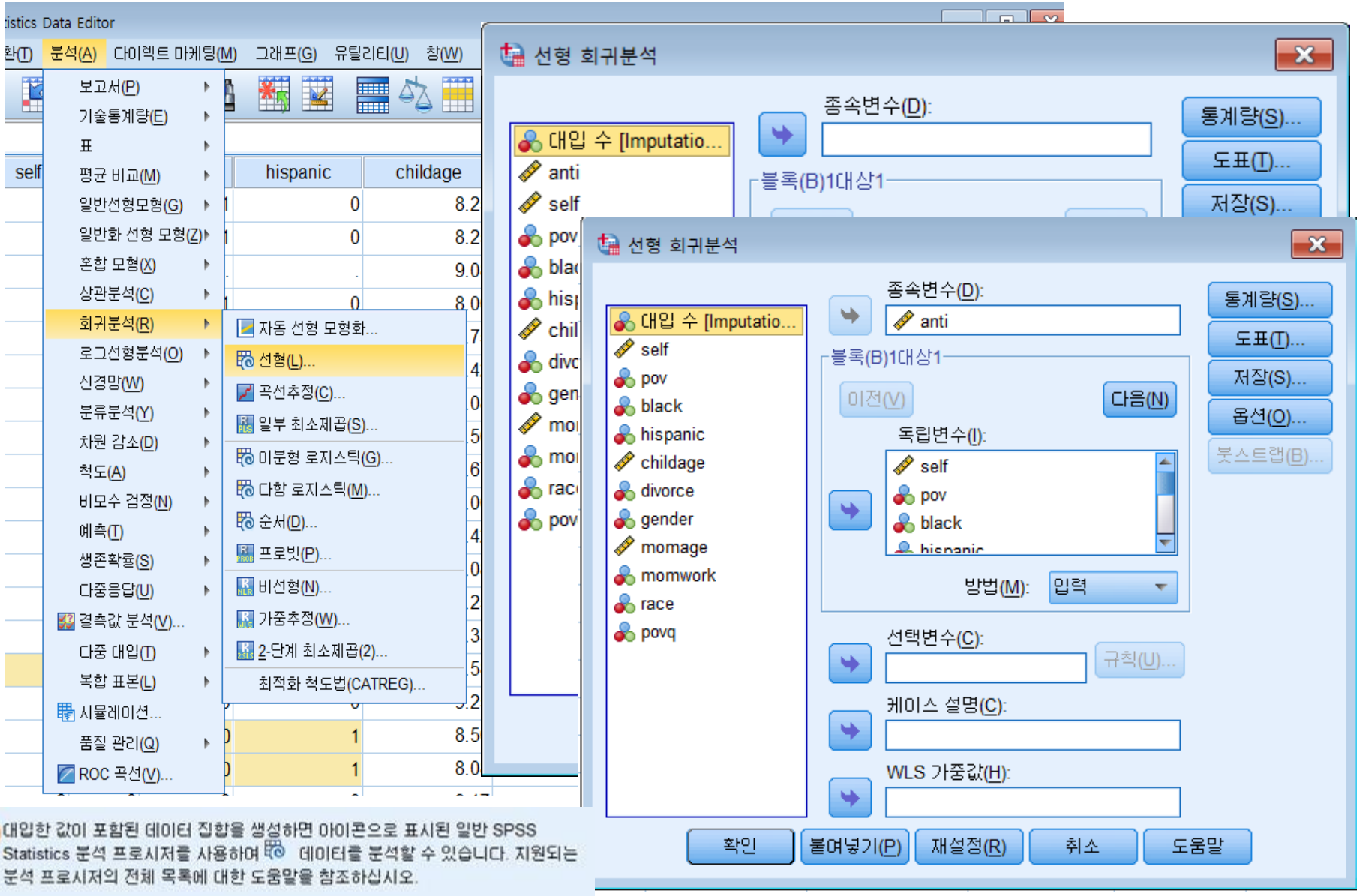
	모형		결측값	대입된 값
	유형	효과		
momwork	로지스틱	divorce, gender,black, hispanic,pov, anti,childage, momage,self	86	430
black	로지스틱	divorce, gender, momwork, hispanic,pov, anti,childage, momage,self	113	565
hispanic	로지스틱	divorce, gender, momwork, black,pov,anti, childage, momage,self	113	565
self	선형 회귀분석	divorce, gender, momwork, black, hispanic,pov, anti,childage, momage	148	740
pov	로지스틱	divorce, gender, momwork, black, hispanic,anti, childage, momage,self	150	750

MI data



	Imputation_	anti	self	pov	black	hispanic	childage	divorce	gender	momage	momwork	race
573	0	2	17	.	1	0	8.25	0	0	20	1	2
574	0	2	24	1	1	0	8.25	0	0	20	.	2
575	0	4	16	0	.	.	9.08	0	1	24	1	.
576	0	1	.	0	1	0	8.00	0	1	25	1	2
577	0	3	.	1	0	1	8.75	0	0	20	1	3
578	0	0	18	0	0	1	9.42	0	1	18	1	3
579	0	2	18	.	1	0	8.08	1	1	19	0	2
580	0	2	21	0	0	1	8.50	0	0	21	0	3
581	0	1	21	.	1	0	9.67	0	0	20	.	2
582	1	1	21	1	0	0	8.00	0	1	21	0	1
583	1	0	20	0	0	1	8.42	0	1	22	1	.
584	1	5	21	0	0	0	8.08	1	0	18	0	.
585	1	2	23	0	0	0	8.25	0	0	24	0	1
586	1	1	22	0	0	0	9.33	0	1	22	0	1
587	1	1	23	0	0	0	8.58	0	0	24	0	1
588	1	3	24	0	0	0	9.25	1	1	23	0	1
589	1	4	19	0	0	1	8.50	1	0	18	0	.
590	1	1	21	0	0	1	8.08	0	0	24	0	.
591	1	4	9	0	0	0	9.17	1	0	20	0	1

MI data를 이용한 자료분석



MI 자료분석 결과

계수[a]

대입 수	모형		비표준화 계수		표준화 계수	t	유의확률	분수 누락 정보	상대 증가 분산	상대 효율
			B	표준오차	베타					
원래 데이터	1	(상수)	2.865	1.991		1.439	.152			
		self	-.045	.031	-.097	-1.445	.150			
		pov	.719	.237	.230	3.031	.003			
		black	.051	.249	.016	.203	.839			
		hispanic	-.357	.255	-.104	-1.398	.164			
		childage	.002	.171	.001	.012	.991			
		divorce	.087	.245	.024	.355	.723			
		gender	-.335	.198	-.110	-1.687	.093			
		momage	-.012	.046	-.017	-.260	.795			
		momwork	.254	.218	.081	1.170	.243			
1	1	(상수)	2.588	1.213		2.134	.033			
		self	-.071	.020	-.141	-3.492	.001			
		pov	.581	.137	.189	4.233	.000			
		black	.116	.123	.039	.943	.346			
		hispanic	-.289	.122	-.095	-2.376	.018			
		childage	.003	.100	.001	.032	.974			
		divorce	-.071	.144	-.021	-.495	.621			
		gender	-.566	.117	-.193	-4.842	.000			
		momage	.021	.028	.031	.750	.454			
		momwork	.255	.129	.082	1.976	.049			
2	1	(상수)	2.493	1.208		2.064	.040			
		self	-.053	.020	-.105	-2.625	.009			
		pov	.870	.141	.279	6.182	.000			
		black	-.076	.138	-.025	-.548	.584			
		hispanic	-.328	.156	-.093	-2.105	.036			

계수[a]

대입 수	모형		비표준화 계수		표준화 계수	t	유의확률	분수 누락 정보	상대 증가 분산	상대 효율
			B	표준오차	베타					
4	1	(상수)	2.764	1.219		2.267	.024			
		self	-.055	.019	-.115	-2.876	.004			
		pov	.660	.140	.213	4.720	.000			
		black	.091	.144	.029	.636	.525			
		hispanic	-.348	.143	-.107	-2.432	.015			
		childage	-.036	.100	-.015	-.357	.721			
		divorce	-.103	.143	-.030	-.722	.471			
		gender	-.543	.116	-.185	-4.669	.000			
		momage	.015	.028	.022	.527	.598			
		momwork	.217	.131	.070	1.649	.100			
5	1	(상수)	2.418	1.232		1.963	.050			
		self	-.046	.020	-.095	-2.340	.020			
		pov	.624	.140	.202	4.463	.000			
		black	.136	.122	.046	1.108	.268			
		hispanic	-.283	.121	-.093	-2.337	.020			
		childage	-.031	.101	-.013	-.313	.755			
		divorce	-.109	.145	-.031	-.753	.452			
		gender	-.546	.117	-.186	-4.658	.000			
		momage	.021	.028	.031	.730	.466			
		momwork	.191	.129	.061	1.479	.140			
통합	1	(상수)	2.523	1.229		2.052	.040	.021	.021	.996
		self	-.055	.022		-2.492	.014	.223	.260	.957
		pov	.698	.188		3.712	.001	.501	.826	.909
		black	.070	.161		.435	.666	.361	.488	.933
		hispanic	-.330	.146		-2.265	.024	.143	.156	.972
		childage	-.023	.101		-.226	.821	.030	.031	.994
		divorce	-.108	.148		-.731	.465	.063	.065	.988
		gender	-.553	.117		-4.728	.000	.008	.008	.998
		momage	.021	.029		.725	.469	.039	.040	.992
		momwork	.207	.139		1.487	.138	.143	.157	.972

a. 종속변수: anti

결과 비교(complete data, imputed data)

계수[a]

대입 수	모형		비표준화 계수 B	비표준화 계수 표준오차	표준화 계수 베타	t	유의확률	분수 누락 정보	상대 증가 분산	상대 효율
원래 데이터	1	(상수)	2.865	1.991		1.439	.152			
		self	-.045	.031	-.097	-1.445	.150			
		pov	.719	.237	.230	3.031	.003			
		black	.051	.249	.016	.203	.839			
		hispanic	-.357	.255	-.104	-1.398	.164			
		childage	.002	.171	.001	.012	.991			
		divorce	.087	.245	.024	.355	.723			
		gender	-.335	.198	-.110	-1.687	.093			
		momage	-.012	.046	-.017	-.260	.795			
		momwork	.254	.218	.081	1.170	.243			
통합	1	(상수)	2.523	1.229		2.052	.040	.021	.021	.996
		self	-.055	.022		-2.492	.014	.223	.260	.957
		pov	.698	.188		3.712	.001	.501	.826	.909
		black	.070	.161		.435	.666	.361	.488	.933
		hispanic	-.330	.146		-2.265	.024	.143	.156	.972
		childage	-.023	.101		-.226	.821	.030	.031	.994
		divorce	-.108	.148		-.731	.465	.063	.065	.988
		gender	-.553	.117		-4.728	.000	.008	.008	.998
		momage	.021	.029		.725	.469	.039	.040	.992
		momwork	.207	.139		1.487	.138	.143	.157	.972

a. 종속변수: anti

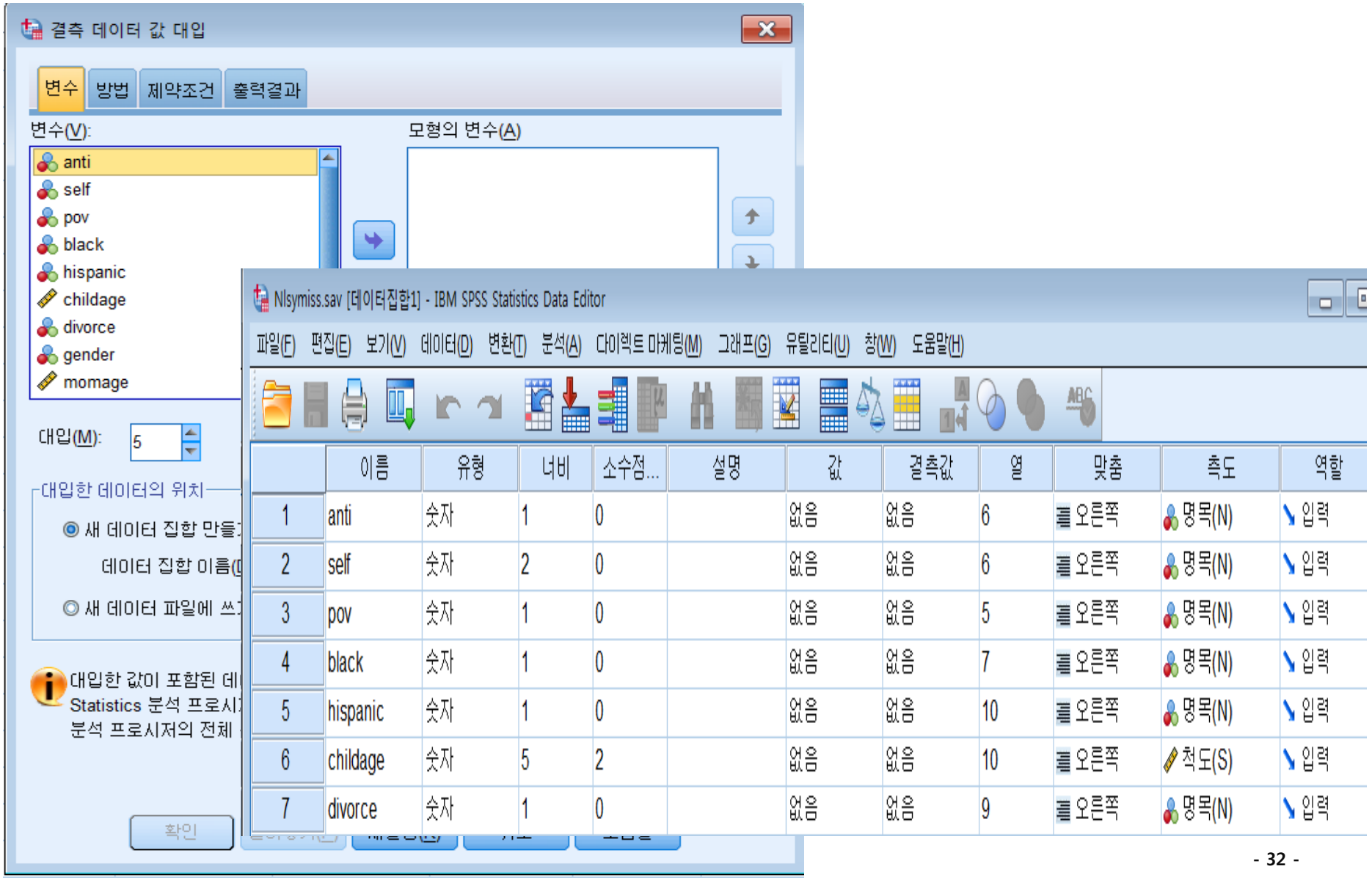
다중 대입에서 다음의 경고를 본다면

경고

self에 대한 대입 모형은 100 모수보다 많이 포함되어 있습니다. 결측값은 대입되지 않습니다. 범주형 변수의 조밀한 범주를 병합하거나 순서형 변수의 측정 레벨을 척도로 변경하거나 양방향 상호 작용을 제거하거나 입력 변수의 역할에 제한을 지정함으로써 대입 모형의 효과 수를 줄이면 문제를 해결할 수 있습니다. 또한 IMPUTE 부명령문의 MAXMODELPARAM 키워드에 허용되는 모수의 최대 수를 증가시킵니다.

이 명령 실행이 중단되었습니다

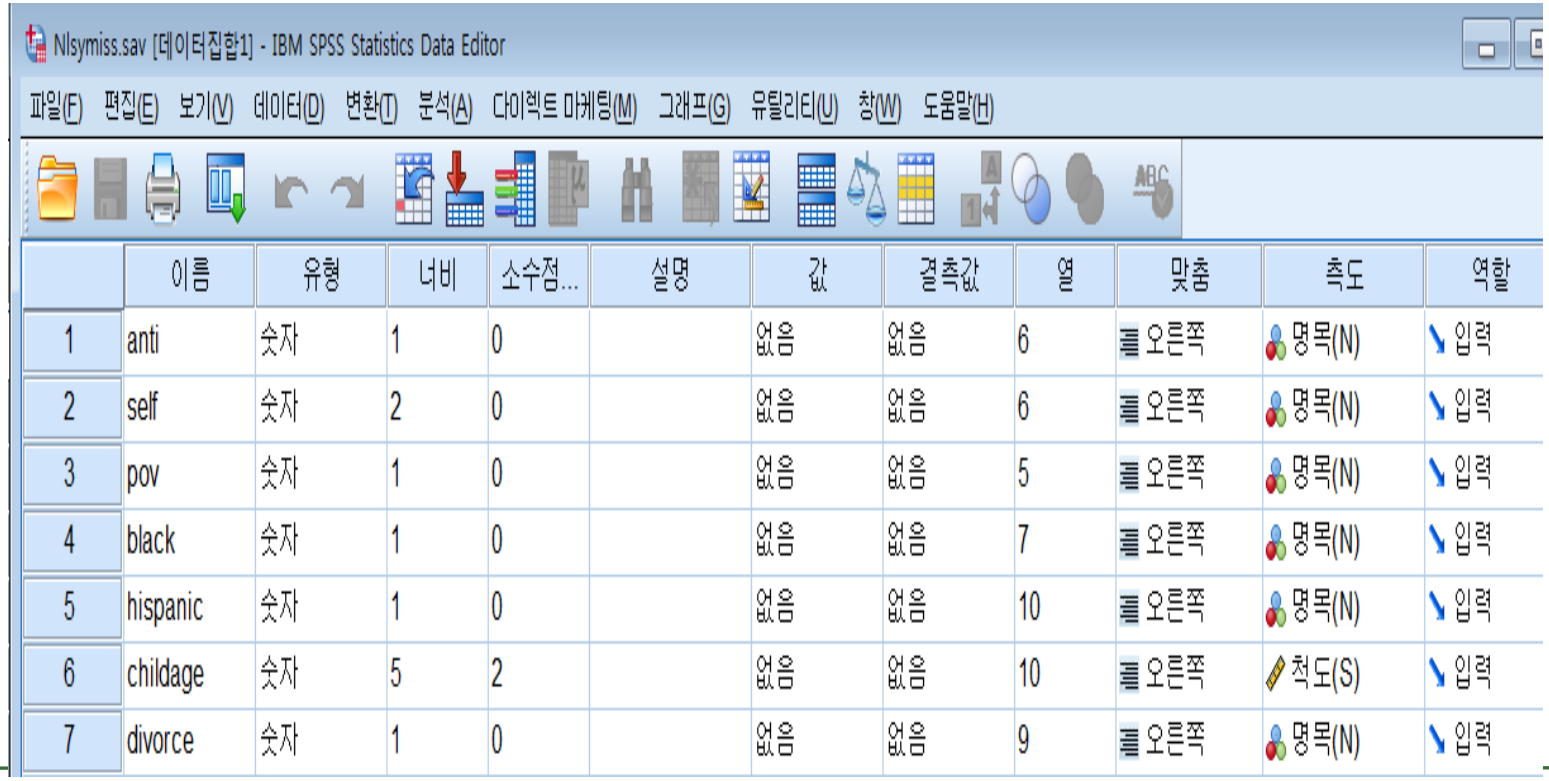
변수 유형 정리(결측값 대체 실행 전에)



	이름	유형	너비	소수점...	설명	값	결측값	열	맞춤	측도	역할
1	anti	숫자	1	0		없음	없음	6	오른쪽	명목(N)	입력
2	self	숫자	2	0		없음	없음	6	오른쪽	명목(N)	입력
3	pov	숫자	1	0		없음	없음	5	오른쪽	명목(N)	입력
4	black	숫자	1	0		없음	없음	7	오른쪽	명목(N)	입력
5	hispanic	숫자	1	0		없음	없음	10	오른쪽	명목(N)	입력
6	childage	숫자	5	2		없음	없음	10	오른쪽	척도(S)	입력
7	divorce	숫자	1	0		없음	없음	9	오른쪽	명목(N)	입력

연속형 자료: 측도 확인

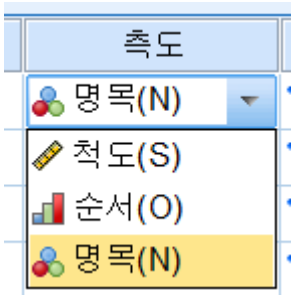
- 소수점이 없는 연속형 자료
- 예: anti-social(0,1,..., 6), self(9,10,..., 24)
- 자료의 측도가 자동적으로 명목형(nominal)로 지정될 수 있음



	이름	유형	너비	소수점...	설명	값	결측값	열	맞춤	측도	역할
1	anti	숫자	1	0		없음	없음	6	오른쪽	명목(N)	입력
2	self	숫자	2	0		없음	없음	6	오른쪽	명목(N)	입력
3	pov	숫자	1	0		없음	없음	5	오른쪽	명목(N)	입력
4	black	숫자	1	0		없음	없음	7	오른쪽	명목(N)	입력
5	hispanic	숫자	1	0		없음	없음	10	오른쪽	명목(N)	입력
6	childage	숫자	5	2		없음	없음	10	오른쪽	척도(S)	입력
7	divorce	숫자	1	0		없음	없음	9	오른쪽	명목(N)	입력

연속형 자료에 대한 측도 변경

- 측도에서 명목(N) → 척도(S)로 변경



*Nlsymiss.sav [데이터집합1] - IBM SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W) 도움말(H)

	이름	유형	너비	소수점...	설명	값	결측값	열	맞춤	측도	역할
1	anti	숫자	1	0		없음	없음	6	오른쪽	명목(N)	입력
2	self	숫자	2	0		없음	없음	6	오른쪽	명목(N)	입력
3	pov	숫자	1	0		없음	없음	5	오른쪽	명목(N)	입력
4	black	숫자	1	0		없음	없음	7	오른쪽	명목(N)	입력

*Nlsymiss.sav [데이터집합1] - IBM SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W) 도움말(H)

	이름	유형	너비	소수점...	설명	값	결측값	열	맞춤	측도	역할
1	anti	숫자	1	0		없음	없음	6	오른쪽	척도(S)	입력
2	self	숫자	2	0		없음	없음	6	오른쪽	척도(S)	입력
3	pov	숫자	1	0		없음	없음	5	오른쪽	명목(N)	입력
4	black	숫자	1	0		없음	없음	7	오른쪽	명목(N)	입력

반복측정자료의 결측

- Y: 여러 번 측정한 경우, Y의 일부자료의 missing
- X: 독립변수는 모두 관찰

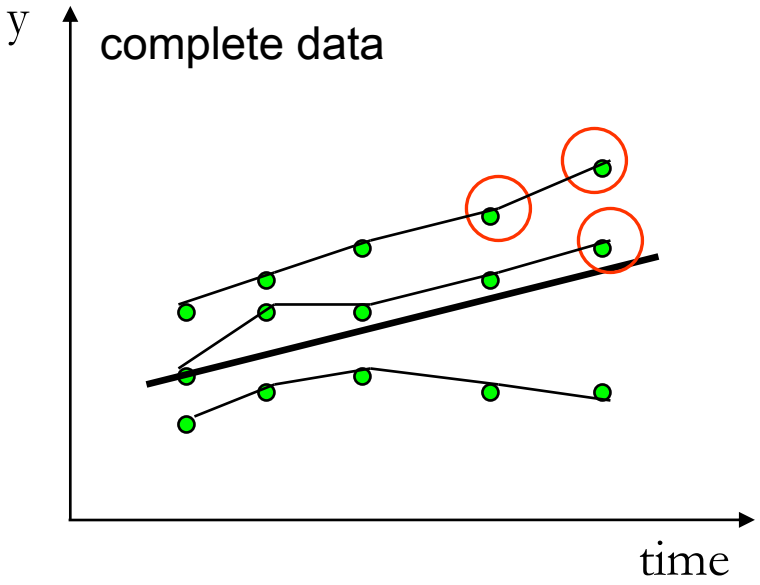
- 반복측정자료 분석방법
  - Repeated measures ANOVA (RMANOVA)
  - Mixed model
  - Generalized Estimating Equation

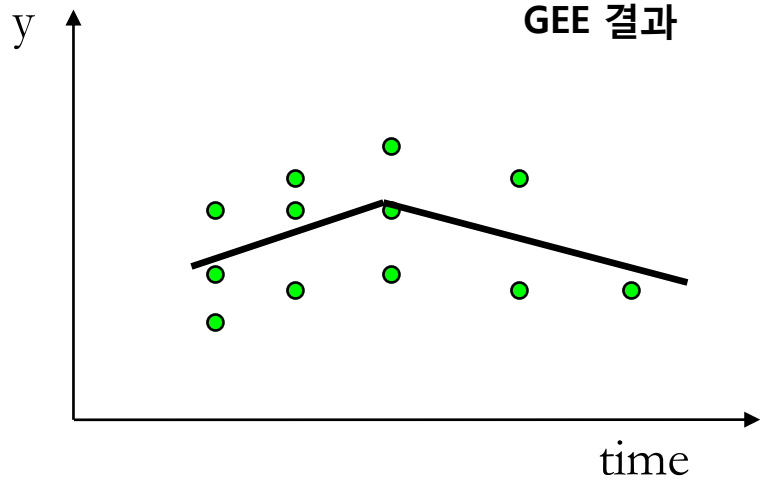
반복측정자료 분석방법 비교

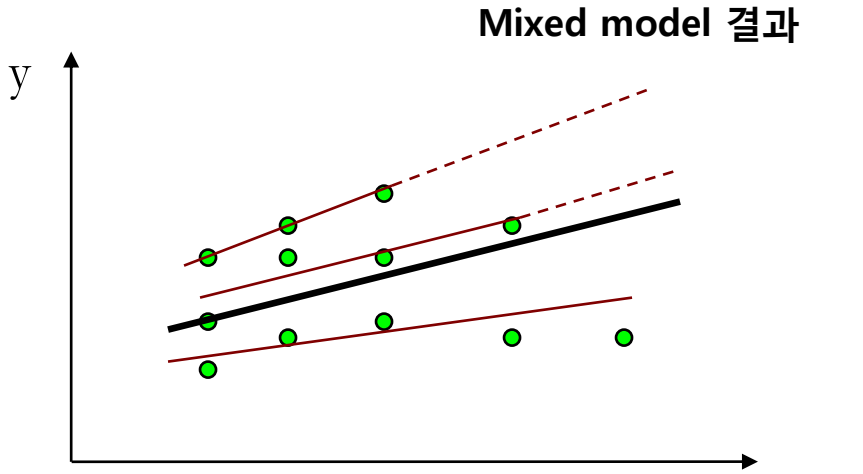
	분석자료	Missing mechanism 가정
RMANOVA	모든 시점의 Y값이 있는 자료 Id 1 : Y1, Y2=., Y3 제외	MCAR
Mixed Model	관찰된 Y 모두 포함 Id 1: Y1, Y3 자료 이용	MAR
GEE	관찰된 Y 모두 포함 Id 1: Y1, Y3 자료 이용	MCAR

- Missing이 MCAR가 아닌 경우는 RMANOVA, GEE 분석결과는 biased 일 수 있음
- Mixed model로 분석 권고

반복측정자료 분석결과 (GEE, Mixed model)







MI in SAS

1. Impute the missing data

```
proc mi data=nlsymiss nimpute=5 seed=15 out=imputed_db;
class anti pov black hispanic divorce gender momage momwork;
var anti self pov black hispanic childage divorce gender momage momwork;
fcs logistic (pov black hispanic divorce gender momwork) discrim(anti);
run;
```

2. Analyze imputed data

```
proc reg data=imputed_db outest=db_est covout ;
model anti= self pov black hispanic childage divorce gender momage momwork;
by _imputation_;
run;
```

3. Combine the parameter estimates

```
proc mianalyze data=db_est;
var self pov black hispanic childage divorce gender momage momwork;
run;
```

감사합니다!