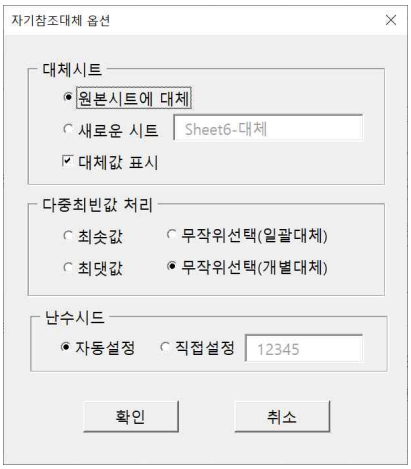
◉ 결측값 대체

○ 결측값 대체 > 자기참조대체(수치)

◦ 수량적 결측값을 평균, 중앙값, 최빈값, 최솟값, 최댓값으로 대체하거나 해당변수의 관측값을 무작위로 선택하여 대체함

◦ 분석폼
- '변수목록'에서 대체할 변수를 선택해 '대체변수' 목록으로 전달.
- '대체방법'에서 대체값으로 사용할 통계값 선택. 'Hot Deck'은 대체값을 무작위로 선택함
- '대체옵션'을 누르면 [그림 2]와 같은 '자기참조대체 옵션'이 나오고 대체시트명, 대체값표시(노란색), 최빈값이 여러 개인 경우 어떤 값으로 대체할지와 Hot Deck이나 다중최빈값 처리에서 난수시드를 지정할 수 있음

[그림 1] 자기참조대체(수치) 분석폼



[그림 2] 자기참조대체 옵션폼

【자기참조대체(수치) 예제】

	A	B	C	D
1	나이	몸무게	클레스테롤	수축기혈압
2	38	51	160	110
3	28	60	188	
4	35	80	260	165
5	30	68.55556	210.8889	135
6	45	75	270	156
7	40	62	181	115
8	33	66	174	118
9	54	78	215	143
10	47	82	240	165
11	36	63	210	122

○ 결측값 대체 > 반복측정대체(수치)

◦ 반복측정(패널, 경시) 자료에서 같은 관측자료의 정보를 이용하여 대체함

◦ 분석폼

- 자료 구성은 한 관측값에 대해 반복측정값을 각 변수열에 저장한 형태를 가지면 측정시간순으로 '순차적 대체변수'에 선정함
- '대체방법'으로는 다음과 같은 방법을 적용할 있음
  · LOCF(Last Observation Carried Forward): 앞에서 가장 최근에 관측된 값으로 대체
  · BOCF(Baseline Observation Carried Forward): 기저값으로 대체
  · NOCB(Next Observation Carried Backward): 뒤에서 가장 최근에 관측된 값으로 대체
  · MCF(Mean Carried Forward): 앞의 평균으로 대체
  · WOCF(Worst Observation Carried Forward) 앞에서 가장 좋지 않은 관측값으로 대체(작을수록 안 좋으면 Min, 클수록 안 좋으면 Max 선택)

[그림 3] 반복측정대체 분석폼

◦ '대체옵션'을 누르면 '자기참조대체 옵션'이 나오고 대체시트명, 대체값표시(노란색)를 지정할 수 있음

【LOCF 적용예제】

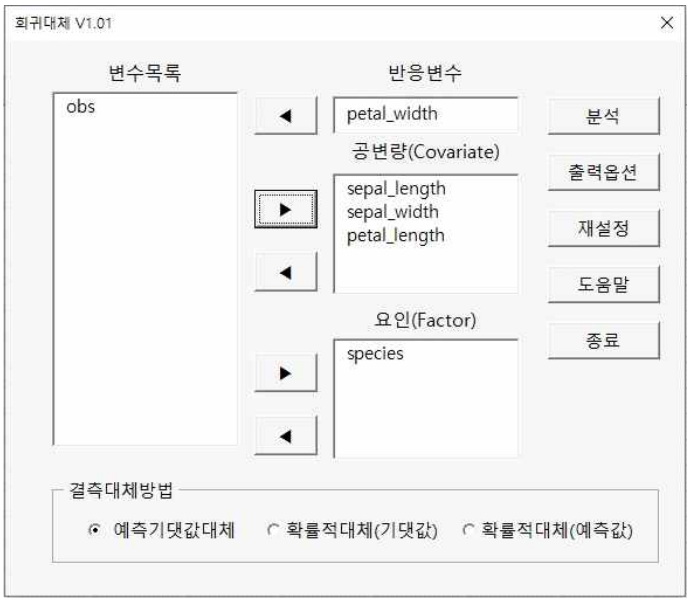
	A	B	C	D	E	F
1	PIN	Visit1	Visit2	Visit3	Visit4	Visit5
2	1	22.1	27.1	22.3	19.8	20
3	2	13.5	16.9	15.7	15.1	21.8
4	3	19	20	16.4	16.4	11.8
5	4	24.1	15.1	25.8	28.3	17
6	5	14.5	17.4	10.2	6.5	22.8
7	6	22	22	14.4	16	14.3
8	7	19.1	20.6	22.8	26.8	24
9	8	11.5	10.4	10.4	17.1	18.8
10	9	20	25.5	21.4	11	21.3
11	10	22.1	28.6	28.3	27.3	22.5
12	11	6	19.4	19.7	7.6	21.3
13	12	14.5	16.5	19.9	7.5	6.3
14	13	25.1	25.1	21.3	26.8	28
15	14	27	11.9	18.2	13.6	16.3
16	15	19	18	10.4	14.5	7.8

○ 결측값 대체 > 회귀대체(수치)

◦ 관측값들을 이용하여 회귀식을 유도하고 유도된 회귀식의 예측값 또는 예측변동성을 포함한 예측값으로 결측값을 대체함

◦ 분석폼

- '변수목록'에서 대체할 변수를 '반응변수', 설명변수 중 수치변수를 '공변량(Covariate)', 범주형변수를 '요인(Factor)'으로 전달. 요인의 경우 수준수 보다 하나 적은 가변수를 자동 생성하여 최소제곱법을 적용하여 모수를 추정함
- 결측대체방법에서의 세 방법은 다음과 같음
  · 예측기댓값 대체: 일반적으로 $\hat{y}$로 표시되는 값으로 대체
  · 확률적 대체(기댓값): $\hat{y}+t\times\mathrm{se}(\hat{y})$로 대체함. 여기서 t는 t-분포를 따르는 난수
  · 확률적 대체(예측값): $\hat{y}+t\times\mathrm{se}(y-\hat{y})$로 대체함.
- 출력옵션은 '자기참조대체 옵션'에서 '다중최빈값 처리' 부분이 빠진 것을 제외하고 동일함



[그림 4] 회귀대체 분석폼

◦ 설명변수가 없는 자료의 경우 예측값을 계산할 수 없기 때문에 대체되지 않으며 요인의 경우 모형 적합 시에 존재하지 않는 수준의 값을 가지는 자료 또한 대체값을 제공하지 않음

- 아래 그림에서 C2 자료가 결측이기 때문에 분석에서 제외되고 F9 자료가 결측이기 때문에 대체에서 제외
- F8의 '테스트'라는 수준이 회귀분석(공분산분석)에 사용된 자료에 없는 수준이기 때문에 E8의 결측값은 대체되지 않음

【회귀대체 예제】

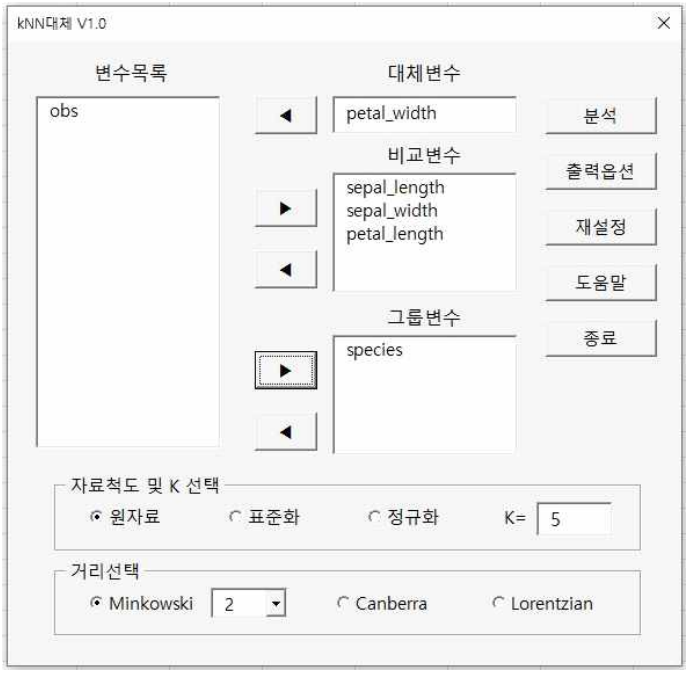
	A	B	C	D	E	F
1	obs	sepal_length	sepal_width	petal_length	petal_width	species
2	1	50		14	2	setosa
3	2	64	28	56	20.0529303	virginica
4	3	65	28	46	15	
5	4	67	31			virginica
6	5	63	28	51	18.8924378	virginica
7	6	46	34	14	3	setosa
8	7	69	31	51		테스트
9	8	62	22	45	15	versicolor
10	9	59	32	48	18	versicolor

○ 결측값 대체 > kNN대체(수치)

◦ 비교변수들 간의 거리를 계산하고 가장 근접한 k개의 대체변수 관측값의 평균으로 결측값을 대체함

◦ 분석폼

- '변수목록'에서 대체할 변수를 '대체변수', 거리를 비교할 변수를 '비교변수', 그룹별로 비교를 원하는 경우 '그룹변수'로 전달. 그룹변수의 경우 선택사항이며 그룹변수가 지정되지 않으면 전체 자료로 분석하고 그룹변수를 지정하면 그룹변수의 값으로 분류한 후 각 범주에 속한 자료들만 비교하여 대체값을 도출함
- '자료척도 및 K선택'에서 비교 전 자료의 척도를 정할 수 있으며 표준화는 각 변수별로 평균을 뺀 값을 표준편차로 나눈 값을, 정규화는 최솟값과 최댓값을 이용하여 0과 1사이의 값으로 만듬
- '거리선택'에서 여러 종류의 거리를 선택할 수 있는 Minkowski는 바로 다음의 콤보박스에 있는 숫자 p를 선택하여 거리를 계산하는데 p=1,2이면 $\left(\sum_i |x_i - y_i|^p\right)^{1/p}$, ∞이면 $\max(|x_i - y_i|)$를 구함. Canberra는 $\sum_i |x_i - y_i| / (|x_i| + |y_i|)$, Lorentzian은 $\sum \log(1 + |x_i - y_i|)$를 계산함



[그림 5] kNN대체 분석폼

- K는 값을 직접입력하여 지정할 수 있으며 비교대상 자료크기가 K보다 적은 경우에는 자료크기 만큼만 사용하고 거리가 같은 자료가 있어 근접한 자료의 수가 K보다 크면 다음과 같이 가중치를 이용하여 대체값을 구함.

$$\frac{1}{K}\sum_{i=1}^{n} d_{(i)} + \frac{K-n}{K(m-n)}\sum_{i=n+1}^{m} d_{(i)}$$

여기서 $d_{(i)}$는 비교거리의 i-번째 순서통계량이고 $d_{(n+1)} = \cdots = d_{(m)}$이고 $n+1$와 m 사이에 K가 존재함

【kNN대체 예제】

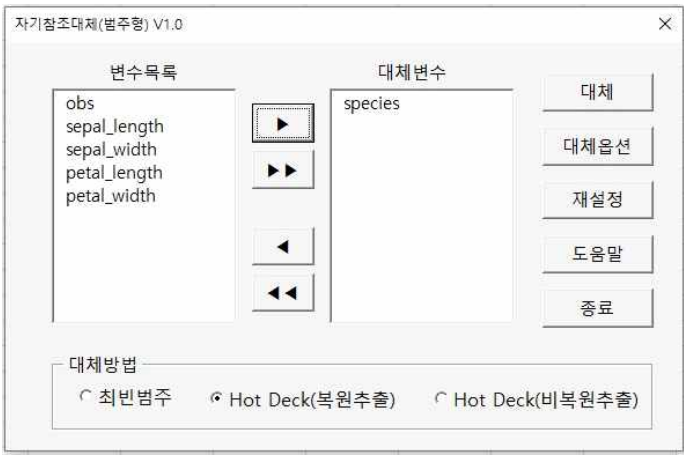
	A	B	C	D	E	F
1	obs	sepal_leng	sepal_width	petal_length	petal_widt	species
2	1	50	33	14	2.2	setosa
3	2	64	28	56	22	virginica
4	3	65	28	46	11.6	versicolor
5	4	67	31	56	21.4	virginica
6	5	63	28	51	15	virginica
7	6	46	34		3	setosa
8	7	69	31	51	23	virginica
9	8	62	22	45		
10	9	59	32	48	18	versicolor
11	10	46	36	10	2	setosa

○ 결측값 대체 > 자기참조대체(범주형)

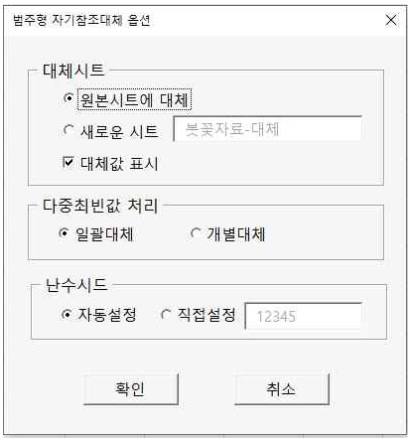
◦ 수량적 결측값을 평균, 중앙값, 최빈값, 최솟값, 최댓값으로 대체하거나 해당변수의 관측값을 무작위로 선택하여 대체함

◦ 분석폼

- '변수목록'에서 대체할 변수를 선택해 '대체변수' 목록으로 전달.
- '대체방법'에서 대체값으로 선택하는 방법을 지정함. 최빈범주는 빈도수가 가장 많은 범주의 값으로 대체하며 최빈범주가 여러개 인 경우 대체옵션을 통해 선택된 하나로 동일하게 대체할 것인지 최빈범주에서 무작위로 선택하여 대체할 것인지를 결정할 수 있음. 'Hot Deck'은 관측범주에서 무작위로 선택하며 복원추출과 비복원 추출을 선택할 수 있음
- '대체옵션'을 누르면 아래와 같은 '범주형 자기참조대체 옵션'이 나오고 대체시트명, 대체값표시(노란색), 최빈범주가 여러 개인 경우 어떤 값으로 대체할지와 Hot Deck 처리에서 난수시드를 지정할 수 있음



[그림 6] 자기참조대체(범주형) 분석폼



【자기참조대체(범주형) 예제】

	A	B	C	D	E	F
1	obs	sepal_length	sepal_width	petal_length	petal_width	species
2	1	50	33	14	2	setosa
3	2	64	28	56	22	virginica
4	3	65	28	46	15	setosa
5	4	67	31	56	24	virginica
6	5	63	28	51	15	virginica
7	6	46	34	14	3	setosa
8	7	69	31	51	23	virginica
9	8	62	22	45	15	versicolor
10	9	59	32	48	18	versicolor
11	10	46	36	10	2	virginica
12	11	61	30	46	14	versicolor