넓은 형태(wide form)의 자료를 긴 형태(long form)로 개념적 이해와 R 구현

김권현

표 1

name	gender	year2011	year2012	year2013	year2014	year2015
ChangSik Park	M	74.69	84.99	91.73	105.11	111.04
EunJung Lee	F	NA	NA	NA	NA	75.89
HaeHee Song	F	NA	NA	75.74	86.5	91.5
HoJun Park	M	NA	NA	NA	71.89	81.42
InHo Kim	M	88.24	96.91	101.85	108.13	112.45
JiSup Kim	M	70.6	83.78	94.17	100.03	106.35
Nari Yoo	F	64.78	80.76	87.3	97.13	103.8
YeoJin Lee	F	88.77	96.45	104.72	112.84	NA

어느 동네의 6세 이하의 영유아[1]의 키를 쟀다. 표 1는 2011년도부터 2015년도까지 1년마다 키를 측정한 결과이다. NA(Not Available)은 결측치를 나타낸다. NA의 존재로 아이의 나이를 짐작해 볼 수 있겠지만, 아이의 정확한 나이는 나타나 있지 않다. 보기 쉽고 깔끔하게 정돈된 듯하다. 문제는 키를 측정한 절대적 시각도 중요하지만 대부분의 경우 키를 측정할 때 아이의 나이가 더 중요하다. 키를 측정할 때 아이의 나이를 자료에 포함한다고 생각해보자. 표 2와 같이 나타낼 수 있을 것이다. 결측치의 수가 증가했다. 그런데 키를 측정할 당시의 아이의 나이를 좀 더 정확하게, 예를 들면 개월로 표시하고 싶다면 어떻게 해야 할까? 표 3과 같이 될 것이다(여기서 열의 이름의 m은 month를 줄인 것이다). 표 3에는 NA(Not Available)가 엄청나게 많다. 이렇게 측정 시기가 측정 대상에 따라 다르다면 넓은 형태(wide form)는 엄청난 공간 낭비를 초래하게 된다. 다른 방법으로 표 1에 2015년의 나이를 포함시킬 수 있다. 하지만 그런 방식이 가능한 것은 모든 아이의 측정 시기가 정확하게 일치할 때 뿐이다. 만약 아이마다 측정시기가 몇 달씩 차이가 나고, 측정 시기의 아이의 나이를 개월로 알고 싶다면 사용할 수 없다. 이때에는 다시 표 3과 같은 방식으로 나타내야 할 것이다.

표 2

name	gender	age0	age1	age2	age3	age4	age5	age6
ChangSik Park	M	NA	76.69	84.91	97.06	105.73	107.32	NA
EunJung Lee	F	NA	76.05	NA	NA	NA	NA	NA
HaeHee Song	F	72.65	82.46	91.76	NA	NA	NA	NA
HoJun Park	M	69.76	79.89	NA	NA	NA	NA	NA
InHo Kim	M	NA	NA	85.34	93.22	105.78	105.84	114.82
JiSup Kim	M	71.59	84.87	92.1	100.73	104.21	NA	NA
Nari Yoo	F	62.74	81.39	92.02	98.32	106.44	NA	NA
YeoJin Lee	F	NA	NA	88.39	100.67	107.58	111.52	NA

이런 경우에 고려할 수 있는 자료 표시 방법은 긴 형태(long form)이다. 긴 형태에서는 한 아이의 각 측정값이 한 열에 제시된다(넓은 형태에서는 한 아이의 모든 측정값이 한 열에 모두 표시가 되었다면 긴 형

1) 영아는 3세 미만의 어린이, 유아(幼兒)는 만 3세로부터 초등학교 취학시기의 어린이를 가르킨다.

태에서는 한 아이의 하나의 측정값이 한 열에 표시된다.) 가장 간단한 형태는 한 열에 아이의 이름과 측정 시기, 그리고 측정값(신장)을 적는 것이다. 따라서 긴 형태는 표 4와 같이 표현할 수 있다. 원한다면 표 5와 같이 같은 이름끼리 모아서 제시할 수도 있다. 확실히 표 3보다 공간이 적게 필요하다!

표 3

name	gen	m4	m6	m7	m10	m12	…	m55	m60	m65	m72
ChangSik Park	M	NA	NA	NA	NA	76.69	…	NA	107.32	NA	NA
EunJung Lee	F	NA	NA	NA	NA	76.05	…	NA	NA	NA	NA
HaeHee Song	F	NA	NA	NA	72.65	NA	…	NA	NA	NA	NA
HoJun Park	M	NA	69.76	NA	NA	NA	…	NA	NA	NA	NA
InHo Kim	M	NA	NA	NA	NA	NA	…	NA	105.84	NA	114.82
JiSup Kim	M	NA	NA	71.59	NA	NA	…	104.21	NA	NA	NA
Nari Yoo	F	62.74	NA	NA	NA	NA	…	NA	NA	NA	NA
YeoJin Lee	F	NA	NA	NA	NA	NA	…	NA	NA	111.52	NA

표 4

name	gender	month	height
Nari Yoo	F	4	62.74
HoJun Park	M	6	69.76
JiSup Kim	M	7	71.59
HaeHee Song	F	10	72.65
ChangSik Park	M	12	76.69
EunJung Lee	F	12	76.05
Nari Yoo	F	16	81.39
HoJun Park	M	18	79.89
JiSup Kim	M	19	84.87
HaeHee Song	F	22	82.46
ChangSik Park	M	24	84.91
InHo Kim	M	24	85.34
Nari Yoo	F	28	92.02
YeoJin Lee	F	29	88.39
JiSup Kim	M	31	92.1
HaeHee Song	F	34	91.76
ChangSik Park	M	36	97.06
InHo Kim	M	36	93.22
Nari Yoo	F	40	98.32
YeoJin Lee	F	41	100.67
JiSup Kim	M	43	100.73
ChangSik Park	M	48	105.73
InHo Kim	M	48	105.78
Nari Yoo	F	52	106.44
YeoJin Lee	F	53	107.58
JiSup Kim	M	55	104.21
ChangSik Park	M	60	107.32
InHo Kim	M	60	105.84
YeoJin Lee	F	65	111.52
InHo Kim	M	72	114.82

표 5

name	gender	month	height
ChangSik Park	M	12	76.69
ChangSik Park	M	24	84.91
ChangSik Park	M	36	97.06
ChangSik Park	M	48	105.73
ChangSik Park	M	60	107.32
EunJung Lee	F	12	76.05
HaeHee Song	F	10	72.65
HaeHee Song	F	22	82.46
HaeHee Song	F	34	91.76
HoJun Park	M	6	69.76
HoJun Park	M	18	79.89
InHo Kim	M	24	85.34
InHo Kim	M	36	93.22
InHo Kim	M	48	105.78
InHo Kim	M	60	105.84
InHo Kim	M	72	114.82
JiSup Kim	M	7	71.59
JiSup Kim	M	19	84.87
JiSup Kim	M	31	92.1
JiSup Kim	M	43	100.73
JiSup Kim	M	55	104.21
Nari Yoo	F	4	62.74
Nari Yoo	F	16	81.39
Nari Yoo	F	28	92.02
Nari Yoo	F	40	98.32
Nari Yoo	F	52	106.44
YeoJin Lee	F	29	88.39
YeoJin Lee	F	41	100.67
YeoJin Lee	F	53	107.58
YeoJin Lee	F	65	111.52

넓은 형태와 긴형태의 근본적인 차이를 주목하자. 추후에 좀 더 복잡한 상황에 대해 배우겠지만, 다음의 차이를 우선적으로 기억하자. 넓은 형태의 경우는 한 대상에서 측정한 여러 측정값을 모두 한 행에 표시한다. 그리고 열이름으로 그 측정값의 의미를 나타낸다. 다음(표 6)에서 InHo Kim의 해마다 신장은 한 행에 모두 표시 되었다.

표 6

name	gender	year2011	year2012	year2013	year2014	year2015
InHo Kim	M	88.24	96.91	101.85	108.13	112.45

반면 긴 형태에서는 InHo Kim의 측정값(신장)은 길게 늘어져 있다(표 7). 한 열에 단 하나의 측정값을 표시한다. 따라서 열이름으로 측정값의 의미을 전달할 수 없다. 굳이 한다면 표 8처럼 열이름을 지어야 할 것이다. 하지만 이 방법도 문제가 있다. InHo Kim은 2011년, 2012년, 2013년, 2014년, 2015년에 모두 측정을 했지만, EunJung Lee는 2015년에만 측정을 했기 때문에 열이름을 year2011, 2012, 2013, 2014, 2015로 하기에는 무리가 있다(표 9).

표 7

name	gender	
InHo Kim	M	85.34
InHo Kim	M	93.22
InHo Kim	M	105.78
InHo Kim	M	105.84
InHo Kim	M	114.82

표 8

name	gender	year2011, 2012, 2013, 2014, 2015
InHo Kim	M	85.34
InHo Kim	M	93.22
InHo Kim	M	105.78
InHo Kim	M	105.84
InHo Kim	M	114.82

표 9

name	gender	?
InHo Kim	M	85.34
InHo Kim	M	93.22
InHo Kim	M	105.78
InHo Kim	M	105.84
InHo Kim	M	114.82
EunJung Lee	F	76.05

넓은 형태를 긴 형태를 바꾸는 과정은 다음과 같이 단계별로 생각해 볼 수 있다. 넓은 형태에 한 행에 늘어져 있는 측정값을 한 행에 하나의 측정값으로 바꾸기 위해 우선 측정값을 한 행에 하나가 되도록 바꾼다. 예를 들어 표 10에서 한 행에 하나의 측정값으로 바꾸기 위해서는 우선 표 11로 바꿀 수 있다(공란은 NA를 생략한 것이다). 그리고 각 측정값을 한 열로 모으기 위해 전에 하나의 열을 만들어서 측정값의 의미를 적는다(표 12). 보통 이 열의 이름은 measure 또는 key로 정해준다. 이제 표 12에는 age0, age1, age2, age3, age4, age5가 열이름에도 있고, 한 열에도 있음을 확인할 수 있다. 열이름을 생략하면 표 13가 같이 변하게 될 것이다. 한 행에 측정값이 하나만 있기 때문에 나머지 NA를 모두 없애고 한 행에 모든 측정값을 표시하면 표 14와 같이 나타낼 수 있다. 이제 사라진 열이름을 새로이 만들어 주자. 보통은 value라고 많이 쓴다(표 15). 표 10을 다시 한 번 천천히 살펴보자. age가 반복되고 있음이 보이는가? 반복되는 것은 줄일 수 있다(표 15). 공통적으로 존재하는 age를 열이름으로 끌어 올렸다. 이제 표 16을 다시 보자. 측정값이라는 의미에서 value라는 열이름은 그 의미를 좀 더 구체적으로 나타내기 위해 height라고 바꿔줄 수 있을 것이다(표 17).

표 10

name	gender	age0	age1	age2	age3	age4	age5	age6
HaeHee Song	F	72.65	82.46	91.76	NA	NA	NA	NA
HoJun Park	M	69.76	79.89	NA	NA	NA	NA	NA

표 11

name	gender	age0	age1	age2	age3	age4	age5	age6
HaeHee Song	F	72.65			NA	NA	NA	NA
HaeHee Song	F		82.46					
HaeHee Song	F			91.76				
HoJun Park	M	69.76		NA	NA	NA	NA	NA
HoJun Park	M		79.89					

표 12

name	gender	measure	age0	age1	age2	age3	age4	age5	age6
HaeHee Song	F	age0	72.65			NA	NA	NA	NA
HaeHee Song	F	age1		82.46					
HaeHee Song	F	age2			91.76				
HoJun Park	M	age0	69.76		NA	NA	NA	NA	NA
HoJun Park	M	age1		79.89					

표 13

name	gender	measure							
HaeHee Song	F	age0	72.65			NA	NA	NA	NA
HaeHee Song	F	age1		82.46					
HaeHee Song	F	age2			91.76				
HoJun Park	M	age0	69.76		NA	NA	NA	NA	NA
HoJun Park	M	age1		79.89					

표 14

name	gender	measure	
HaeHee Song	F	age0	72.65
HaeHee Song	F	age1	82.46
HaeHee Song	F	age2	91.76
HoJun Park	M	age0	69.76
HoJun Park	M	age1	79.89

표 15

name	gender	measure	value
HaeHee Song	F	age0	72.65
HaeHee Song	F	age1	82.46
HaeHee Song	F	age2	91.76
HoJun Park	M	age0	69.76
HoJun Park	M	age1	79.89

표 16

name	gender	age	value
HaeHee Song	F	0	72.65
HaeHee Song	F	1	82.46
HaeHee Song	F	2	91.76
HoJun Park	M	0	69.76
HoJun Park	M	1	79.89

표 17

name	gender	age	height
HaeHee Song	F	0	72.65
HaeHee Song	F	1	82.46
HaeHee Song	F	2	91.76
HoJun Park	M	0	69.76
HoJun Park	M	1	79.89

표 15의 의미가 좀 더 명확지만 표 17도 쉽게 이해할 수 있을 것이다. 표 15의 경우와 표 17을 비교해 보자. 열이름을 생각하면, 표 15의 value는 어떤 값도 괜찮지만, 표 17의 height는 그 값에 제한이 있다. 표 15의 value에는 가능한 모든 측정값(예를 들면, 키, 몸무게, 속도, 취향 등)이 올 수 있지만, 표 17의 height는 단지 키를 잰 결과가 와야 할 것이다. 표 18은 표 17의 의미를 좀 더 명확히 나타낸 것이다.

표 18

name	gender	measure	value
HaeHee Song	F	height.age0	72.65
HaeHee Song	F	height.age1	82.46
HaeHee Song	F	height.age2	91.76
HoJun Park	M	height.age0	69.76
HoJun Park	M	height.age1	79.89

측정값의 종류가 다양해지는 경우를 한 번 생각해보자. 측정값이 키(height), 체중(weight), 비만지수(BMI)라면 표 19와 같이 될 것이다. 하지만 이 경우에 value 열의 값들이 모두 다른 단위로 측정된 값이므로 자료를 읽기 불편할 수도 있다(height는 cm, weight는 kg 단위로 측정되었고, BMI는 특별한 단위가 없다).

표 19

name	gender	measure	value
HaeHee Song	F	height	72.65
HaeHee Song	F	weight	8.22
HaeHee Song	F	BMI	22.6
HoJun Park	M	height	69.76
HoJun Park	M	weight	8.51

그래서 넓은 형태(wide form)을 긴 형태(long form)으로 바꿀 때에는 어떤 값들을 한 열로 묶을 것인지 결정을 해야 한다. 그에 따라 열의 이름을 적절하게 정할 수 있을 것이다.

여기서 중간 정리를 하자. 넓은 형태를 긴 형태로 바꿀 때에는 먼저, 어떤 열의 값들을 한 열로 나열할 것인지를 정한다. 이 때 여러 열이 합쳐서 한 열로 되었으므로 기존의 열 이름을 그대로 쓰기 어렵다. 따라서 새로운 열 이름을 만들어준다. 그리고 어떤 열의 값인지를 나타내기 위해 새로운 열이 하나 더 만들어지는데 그 열의 이름도 지정해 준다(표 20 ~ 표25).

표 20 한 열로 묶을 열을 지정한다

id1	id2	age0	age1	age2
HaeHee Song	F	NA	82.46	91.76
HoJun Park	M	69.76	79.89	NA
HaeHee Song	F	NA	82.46	91.76

표 21 한 행에 값이 하나만 있도록 행을 늘린다

id1	id2	age0	age1	age2
HaeHee Song	F	NA		
HaeHee Song	F		82.46	
HaeHee Song	F			91.76
HoJun Park	M	69.76		
HoJun Park	M		79.89	
HoJun Park	M			NA
HaeHee Song	F	NA		
HaeHee Song	F		82.46	
HaeHee Song	F			91.76

표 22 열이름을 한 열에 모아 적는다

id1	id2				
HaeHee Song	F	age0	NA		
HaeHee Song	F	age1		82.46	
HaeHee Song	F	age2			91.76
HoJun Park	M	age0	69.76		
HoJun Park	M	age1		79.89	
HoJun Park	M	age2			NA
HaeHee Song	F	age0	NA		
HaeHee Song	F	age1		82.46	
HaeHee Song	F	age2			91.76

표 23 여러 열에 흩어진 값을 한 열로 모은다

id1	id2		
HaeHee Song	F	age0	NA
HaeHee Song	F	age1	82.46
HaeHee Song	F	age2	91.76
HoJun Park	M	age0	69.76
HoJun Park	M	age1	79.89
HoJun Park	M	age2	NA
HaeHee Song	F	age0	NA
HaeHee Song	F	age1	82.46
HaeHee Song	F	age2	91.76

표 24 열 이름을 지정해 준다

id1	id2	key	value
HaeHee Song	F	age0	NA
HaeHee Song	F	age1	82.46
HaeHee Song	F	age2	91.76
HoJun Park	M	age0	69.76
HoJun Park	M	age1	79.89
HoJun Park	M	age2	NA
HaeHee Song	F	age0	NA
HaeHee Song	F	age1	82.46
HaeHee Song	F	age2	91.76

표 25 필요하다면 결측치를 생략한다

id1	id2	key	value
HaeHee Song	F	age1	82.46
HaeHee Song	F	age2	91.76
HoJun Park	M	age0	69.76
HoJun Park	M	age1	79.89
HaeHee Song	F	age1	82.46
HaeHee Song	F	age2	91.76

여기서 표 25의 id1, id2 열을 다시 한 번 보자. 사실 id2는 측정 대상을 특정한다기 보다는 측정 대상의 속성(성별)을 나타낸다. 말하자면 측정값이다. 따라서 궁극의 긴 형태라고 할 수 없다. 궁극의 긴 형태는 단 세 개의 열로 이루어진다. id, key, value.

표 25를 궁극의 긴 형태로 만들어 보면 표 26이 된다. 하지만 id로 쓰인 이름은 중복될 가능성이 있다. 동명이인의 존재는 표 27의 형태를 다소 불안하게 한다(물론 동명 이인의 경우에는 이름 뒤에 숫자를 붙여서 서로 구분하게 할 수도 있다. 하지만 이 방법은 또 다른 문제를 야기 한다). 가장 확실한 방법은 숫자를 붙이는 것이다. 그리고 이름도 측정값이다!(표 28)

표 27

id	key	value
HaeHee Song	gender	F
HaeHee Song	age1	82.46
HaeHee Song	age2	91.76
HoJun Park	gender	F
HoJun Park	age0	69.76
HoJun Park	age1	79.89
HaeHee Song	gender	F
HaeHee Song	age1	82.46
HaeHee Song	age2	91.76

표 28

id	key	value
1	name	HaeHee Song
1	gender	F
1	age1	82.46
1	age2	91.76
2	name	HoJun Park
2	gender	F
2	age0	69.76
2	age1	79.89
3	name	HaeHee Song
3	gender	F
3	age1	82.46
3	age2	91.76

궁극의 긴 형태를 소개한 이유는 이 형태가 넓은 형태로 만들기 가장 쉽기 때문이다. 앞에서 넓은 형태를 긴 형태로 변환했던 방법(표 20~ 표 25)을 반대로 하면 된다. 표 28에서 표 31은 이 과정을 도식적으로 보여준다.

이제 가장 핵심적인 내용을 모두 설명하였다. 만약 중복되는 이름이 없다면 이름이 id 역할을 하도록 할 수 있다. 그리고 id 역할을 굳이 한 열에 맡길 필요가 없다. 앞에서 봤듯이 name, gender가 id 역할을 하도록 하거나 name, gender, age0 까지 id 역할을 하게 할 수도 있다(누가 말리겠는가?) 그렇다면 넓은 형태를 긴 형태로 변환할 때 name, gender, age0의 값은 key, value 형태로 바뀌지 않는다.

표 28 먼저 한 열에 모여있는 측정값을 key 열에 따라 배열한다

id	key	name	gender	age0	age1	age2
1	name	HaeHee Song				
1	gender		F			
1	age1				82.46	
1	age2					91.76
2	name	HoJun Park				
2	gender		F			
2	age0			69.76		
2	age1				79.89	
3	name	HaeHee Song				
3	gender		F			
3	age1				82.46	
3	age2					91.76

표 29 이제 key열을 없앨 수 있다

id	name	gender	age0	age1	age2
1	HaeHee Song				
1		F			
1				82.46	
1					91.76
2	HoJun Park				
2		F			
2			69.76		
2				79.89	
3	HaeHee Song				
3		F			
3				82.46	
3					91.76

표 30 같은 id의 측정값을 한 행에 모은다

id	name	gender	age0	age1	age2
1	HaeHee Song	F		82.46	91.76
1					
1					
1					
2	HoJun Park	F	69.76	79.89	
2					
2					
2					
3	HaeHee Song	F		82.46	91.76
3					
3					
3					

표 31 필요없는 행을 지운다

id	name	gender	age0	age1	age2
1	HaeHee Song	F		82.46	91.76
2	HoJun Park	F	69.76	79.89	
3	HaeHee Song	F		82.46	91.76

넓은 형태를 긴 형태로 바꿀 때 중요한 것은 어떤 열을 한 열로 바꿀 것인가였다. 궁극의 긴 형태로 바꾼다면 value 열에는 여러 종류의 측정값이 들어갈 것이다(name, gender, height.age0, ...). 따라서 value 열에 어떤 값이 들어가길 원하는지 확실히 하면 좋다. 그리고 value도 굳이 한 열에 옮겨야만 하는 것은 아니다. 측정값의 특성에 따라 두 열에 옮길 수도 있다. 예를 들어 표 32를 긴 형태를 옮긴다고 해보자. 열 h2011, h2012, w2011, w2012은 2011, 2012년도의 키(height)와 체중(weight)을 의미한다. 키와 체중은 측정 단위가 다르다(cm, kg). 따라서 h2011~w2012를 모두 한 열로 묶을 수도 있고, 키와 체중의 두 열로 묶을 수도 있다(표 33). 표 33은 다시 한 번 보자. 표 33은 긴 형태인가? 넓은 형태인가?

표 32

name	gender	h2011	h2012	w2011	w2012
ChangSik Park	M	74.69	84.99	9.60	12.00
Nari Yoo	F	NA	80.76	7.15	10.70
YeoJin Lee	F	88.77	96.45	NA	15.00

표 33

name	gender	year	height	weight
ChangSik Park	M	2011	74.69	9.60
ChangSik Park	M	2012	84.99	12.00
Nari Yoo	F	2011	NA	80.76
Nari Yoo	F	2012	7.15	10.70
YeoJin Lee	F	2011	88.77	96.45
YeoJin Lee	F	2012	NA	15.00

이제 R에서 직접 긴 형태와 넓은 형태의 자료를 변환해 보자. R에서 쓸 수 있는 명령어는 다음과 같다.

	package	긴 형태로	넓은 형태로
stack/unstack	utils	stack	unstack
reshape	stats	reshape(direction ="long", ...)	reshape(direction ="wide", ...)
melt/dcast	reshape2	melt	dcast
gather/spread	tidyr	gather	spread

여기서는 reshape과 melt/dcast에 대해 설명하도록 하겠다. 그 이유는 stack/unstack의 경우 다소 기본적인 함수이기 때문에 stack/unstack를 한 후에 사후 처리를 해야 긴 형태/넓은 형태 변환이 완성되고, gather/spread의 경우 패키지 tidyr의 함수인데, 패키지 tidyr이 현재(2016. 3) 버전 0.4.1로 아직 개발 단계로 볼 수 있기 때문이다.

먼저 melt/dcast를 설명하고 reshape를 설명하겠다. reshape은 이용하기 어렵기로 악명이 나 있다. 사실 reshape, reshape2라는 패키지가 나온 이유도 stats::reshape이라는 함수가 쓰기 어렵기 때문이었다. 그리고 reshape은 시간에 따른 반복 측정 데이터(횡단 자료)라는 가정이 있는 듯하다.

다음의 데이터(표 35)가 데이터 프레임 dat에 저장되었을 때, 긴 형태로 바꿔보자.

표 35

name	gender	year2011	year2012	year2013
ChangSik Park	M	74.69	84.99	91.73
HaeHee Song	F	NA	NA	75.74
InHo Kim	M	88.24	NA	101.85
YeoJin Lee	F	88.77	96.45	NA

```
dat = data.frame(name = c("ChangSik Park", "HaeHee Song","InHo Kim","YeoJin Lee"),
                 gender = c("M","F","M","F"),
                 year2011 = c(74.69, NA, 88.24, 88.77),
                 year2012 = c(84.99, NA, NA, 96.45),
                 year2013 = c(91.73, 75.74, 101.85, NA))
```

긴 형태로 바꿀 때 고려할 점은 어떤 열을 하나의 열로 묶을 것이냐이다. 표 35에서는 year2011, year2012, year2013이 된다.

```
melt(dat, measure.vars = c("year2011","year2012","year2013"))
```

year2011, year2012, year2013 열은 3번째, 4번째, 5번째 열이므로 다음과 같이 써도 된다.

```
melt(dat, measure.vars = 3:5)
```

결과는 표 36과 같다. 여기서 열 이름 "variable", "value"에 주목하자. 앞에서 표 23에서 표 24로 넘어가는 과정에서 열 이름을 새롭게 만들어야 함을 보였다. melt 함수는 그 열 이름을 "variable", "value"를 기본값으로 지정해 준다. 기본값이 싫다면 variable.name, value.name으로 새롭게 지정해 줄 수도 있다.

표 36

	name	gender	variable	value
1	ChangSik Park	M	year2011	74.69
2	HaeHee Song	F	year2011	NA
3	InHo Kim	M	year2011	88.24
4	YeoJin Lee	F	year2011	88.77
5	ChangSik Park	M	year2012	84.99
6	HaeHee Song	F	year2012	NA
7	InHo Kim	M	year2012	NA
8	YeoJin Lee	F	year2012	96.45
9	ChangSik Park	M	year2013	91.73
10	HaeHee Song	F	year2013	75.74
11	InHo Kim	M	year2013	101.85
12	YeoJin Lee	F	year2013	NA

예를 들어서 melt(dat, measure.vars = 2:5, variable.name="year", value.name="height")의 결과는 표 37이 된다.

표 37

	name	gender	year	height
1	ChangSik Park	M	year2011	74.69
2	HaeHee Song	F	year2011	NA
3	InHo Kim	M	year2011	88.24
4	YeoJin Lee	F	year2011	88.77
5	ChangSik Park	M	year2012	84.99
6	HaeHee Song	F	year2012	NA
7	InHo Kim	M	year2012	NA
8	YeoJin Lee	F	year2012	96.45
9	ChangSik Park	M	year2013	91.73
10	HaeHee Song	F	year2013	75.74
11	InHo Kim	M	year2013	101.85
12	YeoJin Lee	F	year2013	NA

측정값이 저장된 열을 지정해서 긴 형태로 바꾸어 보았다. 측정값이 아닌 열은 모두 id라는 생각해서 id를 나타내는 열을 지정해 줄 수도 있다.

melt(dat, id.vars=c("name","gender"))과
melt(dat, measure.vars = c("year2011","year2012","year2013"))는 결과가 동일하다.
넓은 형태의 각 열이 id를 나타내는 열이거나 측정값을 나타내는 열임을 생각하면 된다. 만약 id.vars와 measure.vars를 모두 지정해 준다면, id.vars와 measure.vars에 포함되지 않은 열은 결과에서 제외된다.

예를 들어 melt(dat, id.vars=c("name"), measure.vars = c("year2011","year2012"))의 결과는 표 38이 된다. 그리고 na.rm=T를 통해 NA도 제외할 수 있다(표 39).

표 38

	name	variable	value
1	ChangSik Park	year2011	74.69
2	HaeHee Song	year2011	NA
3	InHo Kim	year2011	88.24
4	YeoJin Lee	year2011	88.77
5	ChangSik Park	year2012	84.99
6	HaeHee Song	year2012	NA
7	InHo Kim	year2012	NA
8	YeoJin Lee	year2012	96.45

표 39

	name	variable	value
1	ChangSik Park	year2011	74.69
3	InHo Kim	year2011	88.24
4	YeoJin Lee	year2011	88.77
5	ChangSik Park	year2012	84.99
8	YeoJin Lee	year2012	96.45

문제> reshape2::melt함수를 사용하여 표 32를 표 33으로 바꾸어 보자.

```
dat <- data.frame(name=c("ChangSik Park", "Nari Yoo", "YeoJin Lee"),
                  gender=c("M","F","F"),
                  h2011=c(74.69, NA, 88.77),
                  h2012=c(84.99, 80.76, 96.45),
                  w2011=c(9.60, 7.15, NA),
                  w2012=c(12.00, 10.70, 15.00))

datHeight <- melt(dat, id.vars=c("name","gender"),
                  measure.vars=c("h2011","h2012"),
                  variable.name="year",
                  value.name="height")

datWeight <- melt(dat, id.vars=c("name","gender"),
                  measure.vars=c("w2011","w2012"),
                  variable.name="year",
                  value.name="weight")

datHeight$year <- sub("\\D","",datHeight$year)  # 숫자가 아닌 문자를 제거한다.
datWeight$year <- sub("\\D","",datWeight$year) # 숫자가 아닌 문자를 제거한다.

merge(datHeight, datWeight)
```

2016.03.28.

김권현/서울대 협동과정 인지과학/서울대 사회과학연구원 방법론 센터 방법론 상담실