예방의학회지 제35권 제2호(2002년 5월)
Korean J Prev Med 2002;35(2):136-140

무응답이 있는 설문조사연구의 접근법 : 한국노인약물역학코호트 자료의 평가

백지은, 강위창[1], 이영조, 박병주[2]

서울대학교 자연과학대학 통계학과, 대전대학교 정보통계학과[1], 서울대학교 의과대학 예방의학교실[2]

An Approach to Survey Data with Nonresponse: Evaluation of KEPEC Data with BMI

Jieun Baek, Weechang Kang[1], Youngjo Lee, Byung-Joo Park[2]

Department of Statistics, Seoul National University College of Natural Science; Department of Informetrics and Statistics, Daejun University College of Natural Science[1]; Department of Preventive Medicine, Seoul National University College of Medicine[2]

Objectives : A common problem with analyzing survey data involves incomplete data with either a nonresponse or missing data. The mail questionnaire survey conducted for collecting lifestyle variables on the members of the Korean Elderly Phamacoepidemiologic Cohort(KEPEC) in 1996 contains some nonresponse or missing data. The proper statistical method was applied to evaluate the missing pattern of a specific KEPEC data, which had no missing data in the independent variable and missing data in the response variable, BMI.

Methods : The number of study subjects was 8,689 elderly people. Initially, the BMI and significant variables that influenced the BMI were categorized. After fitting the log-linear model, the probabilities of the people on each category were estimated. The EM algorithm was implemented using a log-linear model to determine the missing mechanism causing the nonresponse.

Results : Age, smoking status, and a preference of spicy hot food were chosen as variables that influenced the BMI. As a result of fitting the nonignorable and ignorable nonresponse log-linear model considering these variables, the difference in the deviance in these two models was 0.0034(df=1).

Conclusion : There is a lot of risk if an inference regarding the variables and large samples is made without considering the pattern of missing data. On the basis of these results, the missing data occurring in the BMI is the ignorable nonresponse. Therefore, when analyzing the BMI in KEPEC data, the inference can be made about the data without considering the missing data.

Korean J Prev Med 2002;35(2):136-140

Key Words: Algorithm, Questionnaire, Log-linear Models, Body Mass Index, Korean Elderly Pharmacoepidemiologic Cohort

서 론

연구대상이 대규모인 경우 교란변수에 대한 정보를 얻을 수 있는 효과적이고 경제적인 방법이 바로 설문조사이다. 이처럼 설문조사는 실제조사를 통해서 얻어진 자료이기 때문에, 무응답(nonresponse data)이나 결측치(missing data)가 포함될 수 있다는 것이다. 1996년 한국노인 약물역학 코호트(Korean Elderly Phamacoepidemiologic Cohort ; KEPEC) 구성원들을 대상으로 실시했던 설문조사 연구에서도 이러한 문제점은 발생하였다.

KEPEC은 부산지역의 공무원 및 사립학교 교직원 의료보험관리공단의 피보험자 및 피부양자 중 65세 이상인 노인들로 1993년 구성된 Dynamic Cohort로서, 1996년에 실시한 설문조사는 이들 대상자 중 9,366명으로부터 얻어진 자료이다. 설문의 내용은 주관적 건강상태, 활동상태, 흡연 및 음주습관, 식생활습관, 수면습관, 집밖 및 집안에서의 육체적 활동, 체격, 산과력으로 이루어져 있는데, 본 연구에서 사용한 문항은 흡연 및 음주습관, 식생활습관, 수면습관, 체격 등에 관한 내용이다 [1]. 설문대상이 노인이기 때문에 대리응답자를 사용할 수밖에 없는 경우가 많았고, 이러한 대리응답에 의해 결측치나 무응답이 포함될 수밖에 없었던 것이다.

이러한 결측치의 형태는 Little과 Rubin이 정의한대로 요약할 수 있다 [2]. 우선 결측치는 무시할 수 있는 무응답과 무시할 수 없는 무응답으로 나눌 수 있으며, 무시할 수 있는 무응답에는 MCAR (Missing complete at random)과 MAR(Missing at random)이 있다. MCAR은 반응값을 관찰할 확률이 관찰된 결과값과 관찰되지 않은 결과값 모두와 독립인 경우를 말하고, MAR은 반응값을 관찰할 확률이 관찰된 값에는 영향을 받지만 관찰되지 않은 결과값과는 독립인 경우를 말한다. 무시할 수 없는 무응답이란 반응값을 관찰할 확률이 관찰되지 않은 결과값의 영향을 받는 경우를 말한다.

무응답을 포함하고 있는 자료를 적절하게 처리하여 분석할 수 있는 통계적 방법들이 많이 소개되어 왔다. 범주형 자료에서 무응답이 발생한 경우를 다루기 위한 분석방법들은 무응답이 무시할 수 있다라는 가정과 무시할 수 없다라는 가정에 근거한 것으로 나눌 수 있다. '무시할

접수 : 2002년 1월 14일, 채택 : 2002년 3월 4일
책임저자 : 박병주 (서울시 종로구 연건동 28, 전화: 02-740-8325, 팩스: 02-747-4830, e-mail: bjpark@snu.ac.kr)

수 있는 무응답' 이라는 가정에 근거한 방법들은 모두 대수선형모형을 정의한 후 최대우도 추정법을 사용하여 모수를 추정한다. 그리고 무응답의 처리와 분석이 까다로운 '무시할 수 없는 무응답' 이라는 가정에 근거한 방법들에는 Pregibon [3], Little [4], Albert, Clogg et. al의 방법과 Park & Brown 방법 [5], EM 알고리즘을 이용하여 최대우도추정량을 구한 Fey [6], Baker & Laird [7], Chambers & Welsh의 방법 [8]이 있다.

본 연구에서는 무응답 또는 결측치가 발생한 범주형 자료를 적절하게 처리하여 분석할 수 있는 통계모형의 추정법으로 ML추정법을 적용하여, 독립변수에는 결측치가 발생하지 않고 반응변수에만 결측치가 발생한 자료의 missing mechanism을 파악하였다. 즉, KEPEC의 구성원들을 대상으로 1996년에 실시했던 설문조사 자료를 이용하여 체질량지수(Body Mass Index)를 반응변수로 했을 때의 missing pattern에 대해 평가하였다. 이는 독립변수들은 결측치를 포함하지 않고 반응값에만 결측치가 발생한 자료를 다룰 때 분석 이전에 반드시 거쳐야 할 과정으로써, 그 자료에서 발생한 결측치가 어떤 missing mechanism인지 알아보기 위한 통계모형을 적용하는 예로 KEPEC의 자료를 이용한 것이다.

연구대상 및 방법

1. 연구대상

KEPEC의 구성원 중 설문조사에 응했던 9,366명중에서 독립변수의 무응답으로 인한 677명을 제외시킨 8,689명을 대상으로 몸무게, 키, 몸매형, 음주력, 흡연력, 식이습관, 생활습관 등을 독립변수로 하고 체질량지수를 반응변수로 하였다. Table 1은 독립변수로 고려했던 변수들에 대해서 정리한 것이다. 전체 8,689명 중 반응변수인 BMI가 관측되지 않은 경우는 430명이었다. 설문대상자 8,689명 중, 남성은 2,848명, 여자는 5,841명으로 약 1대 2의 비율을 이뤘고, BMI에 영향을 주는 변수로는 나이, 흡연여부, 매운 음식의 선호도에 관한 변수를 선택하였다. 그리하여 체질량지수에 영향을 주는 유의한 변수와 체질량지수를 다음과 같이 범주화하였다.

Table 1. General characteristics of the Korean Elderly Pharmacoepidemiologic Cohort at Busan, Korea, 1993

Variables		Female		Male	
		No.	%	No.	%
Age	65 - 69	955	16.3	564	19.8
	70 - 74	2120	36.3	1120	39.3
	75 - 79	1430	24.5	671	23.6
	80 - 84	828	14.2	364	12.8
	85 +	508	8.7	129	4.5
Alcohol drinking	none	4345	74.7	833	29.4
	ex-drinker	402	6.9	613	21.6
	current drinker	1068	18.4	1390	49.0
Smoking status	none	4050	69.3	535	18.8
	ex-smoker	666	11.4	1056	37.1
	current smoker	1125	19.3	1257	44.1
Spicy-hot food	rare	1563	26.7	574	20.2
	sometimes	3649	62.5	1775	62.3
	often	629	10.8	499	17.5
Meat food	rare	2884	49.4	787	27.6
	sometimes	2812	48.1	1818	63.9
	often	145	2.5	243	8.5
Salty food	rare	733	12.6	399	14.2
	sometimes	3707	64.0	1671	59.4
	often	1354	23.4	744	26.4
Sleeping time (mean±SD)		7.8±3.5		7.7±3.3	
Insomnia	rare	1401	24.0	877	30.8
	sometimes	3023	51.7	1347	47.3
	often	1417	24.3	624	21.9

Age : 나이 (1=65세-74세, 2=75세 이상)

Smoking Status : 흡연여부 (1=전혀 안 피움, 2=끊었음, 3=피움)

Spicy Food Preference : 매운 음식 선호 (1=전혀 안 먹음, 2=조금 먹음, 3=자주 먹음)

Y : BMI (1=-19, 2=20-24, 3=25+)

Table 2는 8,689명의 자료를 범주화된 각 변수별로 BMI가 관측된 경우와 관측되지 않은 경우로 나누어 정리한 것이다.

2. 연구방법

1) 이론적 배경

두 변수 X, Y에 대해서, X는 독립변수, Y는 종속변수라 가정하자. 그리고 변수 R을 종속변수 Y의 관측여부를 나타내는 지시변수로 정의하자. 즉,

R=1 : X와 Y가 모두 관측

R=2 : X만 관측

여기서 세 변수 X, Y, R에 대하여 정의되는 모형이 대수선형모형인데, 무응답을 유발시키는 메커니즘을 설명하기 위한 모형으로써 사용된다. 예를 들어 반응변수 Y가 관측될 확률이 Y의 값에 영향을 받는다면, 다시 말해 Nonignorable Missing이라면 대수선형 모형에는 Y항과 R항의 교호작용인 YR항을 포함시켜야 한다. 그러므로 YR항을 포함하는 모형은 '무시할 수 없는 무응답을 갖는 모형(nonignorable nonresponse model)' 이고, YR항을 포함하지 않는 모형은 '무시할 수 있는 무응답을 갖는 모형(ignorable nonresponse model)' 이 된다.

2) 통계적 모형

세 변수 X, Y, R에 대해 X=x, Y=y, R=k 가 관측될 확률을 p_{xyk} 라 하고, z_{xyl} 은 관측된 반응변수에서 얻어진 관측도수, z_{x+2} 는 관측되지 않은 반응변수에서 얻어진 주변합으로 나타낸다면 우도함수

Table 2. Status data for body mass index in Korean Elderly Phamacoepidemiologic Cohort, 1993

Smoking Status	Spicy Food Preference	Age	Known BMI(Response)			Unknown BMI
			- 19	20 - 24	25 +	
1	1	1	164	302	192	38
1	1	2	223	240	111	27
1	2	1	287	676	534	80
1	2	2	365	533	277	87
1	3	1	43	105	109	8
1	3	2	44	82	48	10
2	1	1	55	92	48	5
2	1	2	77	81	33	18
2	2	1	122	289	145	16
2	2	2	142	201	112	25
2	3	1	28	79	29	7
2	3	2	29	60	24	5
3	1	1	76	102	32	8
3	1	2	91	89	22	11
3	2	1	245	393	160	33
3	2	2	259	317	88	38
3	3	1	611	120	68	8
3	3	2	60	66	29	6

는 다음과 같이 나타낼 수 있다.

$$L = \left[\prod_x \prod_y (P_{xy1})^{Z_{xy1}}\right]\left[\prod (P_{x+2})^{Z_{x+2}}\right]$$

바로 이 우도함수를 최대로 만드는 것이 최대우도 추정법이고, 무응답모형의 모수에 대한 최대우도 추정법에서 L을 최대로 만드는 값은 앞서 설명한 EM 알고리즘을 이용해서 추정할 수 있다 [7, 9].

불완전한 자료를 모델을 적합시키는데 이용하는 EM 알고리즘은 E-step과 M-step으로 이루어져 있는데, E-step은 관측이 안 된 경우의 칸 도수를 추정하는 단계이고 M-step은 관측된 칸 도수와 E-step에서 얻어진 칸 도수를 완전한 자료라고 가정한 후에, 대수선형모형을 추정하는 단계이다.

간단한 예로, 2×2 분할표를 생각해보자. R이 1인 경우는 X와 Y가 모두 관측된 경우이고, R이 2인 경우는 반응변수 Y가 관측되지 않은 경우이다. 즉, R=2인 경우는 반응변수 Y가 관측되지 않았으므로 주변합만을 알고 있는 경우이다. 그리고 z_{xyk}는 R=k(k=1, 2)인 경우의 각 칸의 관측된 반응도수이다.

즉, EM 알고리즘은 주변합만을 가지고 관측이 안 된 경우의 칸의 도수를 추정한 후, 관측된 칸 도수와 함께 완전한 자료라는 가정으로 대수선형모형을 반복하여 추정하는 과정을 말하는 것이다.

그리고 E-step은 다음과 같은 성질을 만족시킨다.

$$E(z_{xy1}) = z_{xy1}$$

$$E(z_{xy2}) = z_{x+2}\Pr(Y=y \mid R=2, x) = z_{x+2}\frac{m_{xy}\pi_{xy}}{\sum_y m_{xy}\pi_{xy}}$$

여기서, $\pi_{xy} = \Pr(R=2 \mid x, y) = M_{xy2}/M_{xy+}$이고, m_{xy}는 관측된 자료를 가지고 추정한 기대도수이고, M_{xyr}은 무응답 모형의 기대도수이다. 만약에 주변 모델이 포화모형이라면, $m_{xy}=M_{xy+}$가 되어 $E(z_{xy2})$는 $z_{x+2}M_{xy2}/M_{x+2}$ 가 된다. M-step에서는 IPF(iterative proportional fitting)이 사용되기 때문에, EM 알고리즘은 언제나 수렴값을 갖게 되는 것이다 [10].

포화모형 즉, 완전모형에서는 반응변수 Y가 모두 관측되었기에 모든 모수들에 대한 추정을 할 수 있다. 그러나 위의 2×2 분할표에서 R=2인 Y가 관찰되지 않은 경우를 포함하는 불완전한 모형에서는 현재의 분할표에서는 추정할 수 없는 모수들이 확대된 분할표에는 존재하게 된다. 완전모형과 불완전모형의 자유도에 대한 계산은 다음과 같다.

(1)완전한 모형일 때 : df(model) + df(lack of fit) = df(data)

(2)불완전한 모형일 때 : df(model) - df(inestimable parameters)+ df(lack of fit) = df(data)

현재의 분할표

R=1		R=2
z_{111}	z_{121}	z_{1+2}
z_{211}	z_{221}	z_{2+2}

확대된 분할표

R=1		R=2		(주변합)
z_{111}	z_{121}	?	?	z_{1+2}
z_{211}	z_{221}	?	?	z_{2+2}

3) 분석방법

BMI에 영향을 주는 변수인 나이, 흡연여부, 매운 음식에 대한 선호도를 범주화하였다.

X_a는 나이를 나타내는 변수로 1, 2 두 개의 범주를 갖고, X_s는 흡연여부를 나타내는 변수로 1, 2, 3의 세 개의 범주를 갖고, X_f 역시 매운 음식에 대한 선호도를 나타내는 변수로 1, 2, 3 세 개의 범주값을 갖으며 Y는 BMI값으로 반응변수는 1, 2, 3의 범주값을 갖는다. BMI에 대한 실제값이 없는 경우를 무응답으로 분류하여 지시변수 R을 사용하여 무응답 여부에 대한 분류 변수로 사용하였다. 위의 네 변수인 나이, 흡연여부, 매운 음식에 대한 선호도, BMI범주에 대한 상호작용항은 모든 모델에 포함되도록 하였으며, BMI 범주와 무응답 여부에 대한 상호작용항인 YR항을 포함하는 모든 가능한 조합을 만들어 무시할 수 없는 무응답 모형 8개와 YR항을 포함하지 않는 무시할 수 있는 무응답 모형 7개를 적합시켰다. 무시할 수 없는 무응답 모형을 적합시키기 위한 방법으로는 SAS의 MACRO를 이용하여 프로그램을 작성하여 EM알고리즘을 통한 모수 추정을 하였고, 각 범주

Tables 3. Estimated percentage for body mass index in Korean Elderly Phamacoepidemiologic Cohort

	Nonresponse model	BMI		Goodness of fit	
		BMI<20	25≤BMI	Deviance	df
Ignorable	($X_a X_s X_f Y$, $X_a R$, $X_s R$, X_f)	28	25	5.3428	48
Model	($X_a X_s X_f Y$, $X_a R$, $X_s R$)	28	25	236.8968	50
	($X_a X_s X_f Y$, $X_a R$, $X_f R$)	28	25	669.2653	50
	($X_a X_s X_f Y$, $X_s R$, $X_f R$)	28	25	9.6394	49
	($X_a X_s X_f Y$, $X_a R$)	28	25	944.5430	52
	($X_a X_s X_f Y$, $X_s R$)	28	25	241.4507	51
	($X_a X_s X_f Y$, $X_f R$)	28	25	677.0698	51
Nonignorable	($X_a X_s X_f Y$, $X_a R$, $X_s R$, $X_f R$, YR)	30	26	5.3394	47
Model	($X_a X_s X_f Y$, X_aYR)	30	26	942.8313	47
	($X_a X_s X_f Y$, $X_a R$, YR)	30	26	943.6109	51
	($X_a X_s X_f Y$,X_sYR)	30	26	241.0858	47
	($X_a X_s X_f Y$, $X_s R$, YR)	30	26	241.2687	50
	($X_a X_s X_f Y$, X_fYR)	30	26	674.6772	47
	($X_a X_s X_f Y$, $X_f R$,YR)	30	26	676.2836	50
	($X_a X_s X_f Y$, YR)	30	26	950.7371	53

Xa : Age (1=65-74, 2=75+)
Xs : Smoking Status(1=none, 2=ex-smoker, 3=current smoker)
Xf : Spicy Food Preference(1=rare, 2=sometimes, 3=frequently)
Y : BMI (1=-19, 2=20 - 24, 3=25+)
R : Dichotomous Response (1=response, 2=no response)

에 해당하는 사람들의 분포를 추정하였다.

연구 결과

Table 3은 최대우도 추정법을 사용하여 추정한 대수선형모형이다. Y항과 R항의 교호작용인 YR항의 포함여부에 따라 무시할 수 있는 무응답모형과 무시할 수 없는 무응답모형을 적합시켰으며, 각각의 모형을 적합시켰을 때 그에 따른 BMI 분포에 대한 추정값도 제시하였다. 모든 모형에 대한 적합도 검정에 대한 기준은 deviance와 자유도를 고려하여 검정할 수 있다.

무시할 수 있는 무응답모형과 무시할 수 없는 무응답모형의 포화모형을 비교해보자. 무응답을 고려하지 않고 적합시킨 대수선형모형의 deviance는 5.3428이고, 무응답을 고려하여 적합시킨 모형의 deviance는 5.3394이다. 여기서 자유도는 앞서 언급한 공식에 의해 계산된 것이다. 두 모형의 deviance의 차이는 0.0034(df=1)로서 YR항을 제거하여도 모형에는 전혀 영향을 미치지 못 하고 있음을 알 수 있다.

또한 무시할 수 있는 무응답과 무시할 수 없는 무응답을 적합시킨 후, BMI 분포에 대한 확률을 계산해 보았다. 그 결과, BMI가 20미만인 사람들의 확률은 무응답을 고려하지 않았을 때는 28%이었고 무응답을 고려했을 때는 30%로, 2%정도 차이가 났다. 그리고 BMI 25이상인 사람들의 확률은 무응답을 고려하지 않았을 때는 25%이었고 무응답을 고려했을 때는 26%로, 1%정도밖에 차이나지 않았다.

고 찰

대부분 연구대상 집단에 대한 교란변수 자료 수집 방법으로 전화나 우편설문을 통한 표본조사를 통해서 많은 형태의 범주형 자료가 수집·분석되고 있다. 이러한 가운데 완전한 자료를 수집한다는 것은 더욱 힘들기 마련이므로, 수집된 자료들이 결측치를 포함할 수밖에 없는 것은 당연한 일이다. 그러나 무응답을 포함한 자료들의 missing pattern을 무시한 채, 단순히 모델을 적합하여 변수에 대한 영향력이나 전체 표본에 대한 추론을 하는 것은 매우 위험한 일이 아닐 수 없다. 그러므로 결측치를 포함한 자료를 분석할 시에는 꼭 연구자료에 있는 결측치를 무시하고 분석해도 좋은지에 대한 검정을 거쳐야한다는 것이었다. 연구자가 결측치를 포함한 자료를 분석할 시에는 먼저 다음과 같은 과정을 고려한 후, 자료를 분석해야 한다. 첫째, 무응답에 영향을 미칠 수 있는 여러 변수들을 고려하여 자료의 집단 특성을 먼저 밝혀야한다. 둘째, 무응답이 무시할 수 있는 무응답인지 무시할 수 없는 무응답인지에 대한 논의가 필요하다. 마지막으로 무응답의 특성을 고려한 모델을 적합한 후, 무응답이 어느 집단에 속할 것인가를 예측해야 한다.

무응답을 무시한 채 그릇된 결과를 예측했던 표본조사를 통한 의학적 사례는 없지만, 여론조사에서는 찾아볼 수 있다. 한 예로 1948년 미국 여론 조사 기관에서 실시한 미국 대통령 후보의 지지도에 관한 예비조사를 들 수 있다 [7, 11]. 그 당시 후보자로는 Truman, Deway, 그 외 후보자들이 있었고, 범주값으로 7, 8, 9, 10월의 조사기간, 참여자들의 경제적인 수준을 나타내는 변수가 있었다. 후보자에 대한 선택을 못 한 경우를 무응답으로 분류하여 응답자와 무응답자를 나누었다. 당시 이 자료를 분석한 여론조사 기관에서는 Deway 후보가 승리할 것으로 예상을 했었지만 실제 선거에서는 Truman후보가 승리하는 결과를 낳았다. 또한 우리나라의 경우, 15대 국회의원 선거당시 3개 방송사가 공동으로 전화투표조사를 실시하여 당선자 예측을 한 결과 253개 선거구 중에서 39곳이 빗나가기도 하였다 [12]. 이 모두가 무응답율이 높은 자료를 무시한 채 얻어진 자료만 가지고 분석을 하여 그릇된 결과를 발표하게 되는 실수를 범하게 된 것이었다.

이처럼 결측치를 포함하고 있는 자료의 경우에는 이를 잘 처리하여야 한다는 예로 KEPEC의 BMI자료에서 무응답이 미치는 결과에 대해서 규명하고자 하였다. KEPEC의 BMI자료에서의 무응답율은 약 5%정도로 그리 높지 않았다. 무응답의 고려여부에 따른 두 완전모형 검정

결과는 두 모형의 deviance의 차이가 0.0034(df=1)로서 무응답이 미치는 영향이 전혀 없음을 알 수 있었다. 그러므로 KEPEC의 BMI에서 발생한 결측치는 무시할 수 있는 무응답임을 알 수 있고, 앞으로 KEPEC의 BMI를 이용한 모든 분석은 결측치를 무시하고 관측된 자료만 이용하여도 타당하다는 결론을 내릴 수 있었다. 이는 BMI자료 자체에 대한 무응답율이 그리 높지 않았고 의도적으로 응답을 회피할 특별한 이유가 없는 문항이므로, 전체 표본에 대한 예측을 할 때에도 미치는 영향이 높지 않았던 것이라 볼 수 있다. 만약 무응답율이 높은 자료에서 위와 같은 검정을 할 경우에는 앞의 예처럼 결과가 뒤바뀔 수도 있는 경우가 발생할 수도 있을 것이고, 대부분의 사람들이 노출하기를 꺼리는 생활수준이나 소득수준과 같은 사회경제적 변수 혹은 외모에 예민한 여학생들을 대상으로 조사한 체중에 대해서는 좀 더 다른 결론이 나올 수도 있을 것으로 예상된다.

본 연구에서는 관심있는 변수가 반응변수였기에, 설명변수에 발생한 결측치는 제거한 후 반응변수에만 발생한 결측치를 중점으로 다루기 위해 KEPEC의 BMI자료를 사용하였다. 그러나 실제로 표본조사를 통해 얻어진 자료에서 발생한 설명변수도 문제가 될 수도 있다. 예를 들어, 본 연구에서 논의된 BMI에 영향을 주는 변수선택과정에서도 결측치가 제거됨으로 인해서 BMI에 유의한 영향을 줄지도 모르는 변수가 선택되지 않게 되는 문제가 발생할 수도 있기 때문이다. 이미 설명변수를 다룰 수 있는 모형에 대한 논의가 있기는 하지만, 그리 활발하진 못하다. 따라서 앞으로 설명변수에 발생한 결측치를 좀 더 고려한 후 반응변수에 대한 결측치 문제를 해결한다면 더 나은 결과를 얻을 수 도 있고, 더 신뢰할만한 결과를 제시할 수도 있을 것으로 생각된다.

결 론

본 연구에서는 결측치를 포함하고 있는 자료의 경우에 이를 잘 처리하여야 한다는 예로 KEPEC의 BMI자료에서 무응답이 미치는 결과에 대해서 규명하고자 하였다. KEPEC의 BMI자료에서의 무응답율은 약 5%정도로 그리 높지 않았고, Y항과 R항의 교호작용인 YR항의 포함여부에 따른 무시할 수 있는 무응답 모형과 무시할 수 없는 무응답 모형을 적합시킨 결과, 무응답을 고려하지 않고 적합시킨 대수선형모형의 deviance는 5.3428이고, 무응답을 고려하여 적합시킨 모형의 deviance는 5.3394이다. 두 모형의 deviance의 차이가 0.0034(df=1)로서 무응답이 미치는 영향이 전혀 없다고 볼 수 있다. 그러므로 KEPEC의 BMI에서 발생한 결측치는 무시할 수 있는 무응답임을 알 수 있다. 각각의 모형을 적합시켰을 때 그에 따른 BMI 분포에 대한 확률은 BMI가 20미만인 사람들의 확률은 무응답을 고려하지 않았을 때는 28%이었고 무응답을 고려했을 때는 30%로, 2%정도 차이가 났다. 그리고 BMI가 25 이상인 사람들의 확률은 무응답을 고려하지 않았을 때는 25%이었고 무응답을 고려했을 때는 26%로, 1%정도밖에 차이 나지 않았다.

본 연구를 통해서 KEPEC의 BMI를 이용한 모든 분석은 결측치를 무시하고 관측된 자료만 이용하여도 타당하다는 결론을 내릴 수 있었다. 이는 BMI자료 자체에 대한 무응답율이 그리 높지 않았고 의도적으로 응답을 회피할 특별한 이유가 없는 문항이기에, 전체 표본에 대한 예측을 할 때에도 미치는 영향이 높지 않았던 것이라 볼 수 있다. 만약 무응답율이 높은 자료에서 위와 같은 검정을 할 경우에는 1948년 미국 여론 조사 기관에서 실시한 미국 대통령 후보의 지지도에 관한 예비조사의 경우처럼 결과가 뒤바뀔 수도 있을 것이다. 그리고 BMI에 영향을 주는 변수를 선택하는 과정에서 설명변수의 결측치가 제거됨으로 인해서 BMI에 유의한 영향을 줄지도 모르는 변수가 선택되지 않게 되는 문제가 발생할 수도 있을 것이므로, 설명변수에 발생한 결측치를 좀 더 고려한 후 반응변수에 대한 결측치 문제를 해결한다면 더 신뢰할만한 또 다른 결과를 얻을 수도 있을 것으로 생각된다.

참고문헌

1. Park BJ, Kim DS, Koo HW, Bae JM. Reliability and Validity of a Life Style Questionnaire for Elderly People. *Korean J Prev Med* 1998; 31(1): 49-58 (Korean)
2. Little RJA, Rubin DB. Statistical Analysis with Missing Data, New York: John Wiley & Son; 1987
3. Pregibon D. Typical Survey Data : Estimation and Imputation. *Survey Methodology* 1977; 2: 70-102
4. Little RJA. Models for Nonresponse in Sample Survey. *JASA* 1982; 77: 237-250
5. Park TS, Brown MB. Models for Categorical Data with Nonignorable Nonresponse. *JASA* 1994; 89: 44-52
6. Fey RE. Causal Models for Patterns of Nonresponse. *JASA* 1986; 81:354-365
7. Baker SG, Laird NM. Regression Analysis for Categorical Variables with Outcome Subject to Nonignorable Nonresponse. *JASA* 1988; 83: 62-69
8. Chambers RL, Welsh AH. Log-linear Models for Survey Data with Non-ignorable Non-response. *JRSS B* 1993; 55(1): 157-170
9. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *JRSS B* 1977; 39: 1-38
10. Bishop YM, Feinberg SE, Holland PW. Discrete Multivariate Analysis, Cambridge, MA : MIT Press; 1975.
11. Park TS. An Approach to Categorical Data with Nonignorable Nonresponse. *Biometrics* 1998; 54: 1579-1590
12. Park TS, Lee SY. Analysis of Categorical Data with Nonresponses. *Korean J Appl Statistics* 1998; 11: 83-95 (Korean)