이중 추출 자료를 이용한 측정오차분산의 추정

Measurement Error Variance Estimation Based on Subsample Re-measurements

허순영*

많은 경우, 측정오차분산은 알려지지 않은 참값 또는 참값과 연관된 공변수들의 함수로 표현될 수 있다. 이 논문은 단위 당 반복측정에 기초한 단위 내 표본분산을 이용한 선형측정오차분산의 추정에 관한 연구이다. 이 논문은 다음의 내용을 포함한다: (1) 측정오차의 크기를 나타내는 상수 δ의 추정; (2) 유한모집단으로부터의 복합표본, 작은 측정오차라는 조건하에 선형측정오차분산의 추정; (3) 부표본에 포함될 확률을 설명하기 위한 성향모형 추정. 미국의 제3차 건강영양조사자료를 사용하여 이상의 결과들을 이용한 경험적 분석을 실행하였다.

In many cases, the measurement error variances may be functions of the unknown true values or related covariates. This paper develops estimators of the parameters of a linear *measurement error variance function based on within-unit sample variances*. This paper devotes to: (1) define measurement error scale factor δ; (2) develop estimators of the parameters of the linear measurement error variance function under stratified multistage sampling design and small error conditions; (3) use propensity methods to adjust survey weights to account for possible selection effects at the replicate level. The proposed methods are applied to medical examination data from the U. S. Third National Health and Nutrition Examination Survey(NHANES III)

I. 서론

일반적으로 측정오차란 참 값과 측정값의 차로 정의된다. 일부 학자들은 측정오차를 관찰오차라 부르기도 한다(Grove, 1991). 무시할 수 없는 정도의 측정오차가 존재하는 경우, 측정오차를 고려하지 않는 기존 추정방법은 편의된 추정량을 제공한다.예를 들어, Fuller (1987, sec. 1.1.)는 단순선형회귀모형에서 설명변수들의 측정값들에 오차가 존재할 때 통상적인 최소제곱추정량들을 일반적으로 편의된다는 것을 지적했다. 또, 측정오차가 존재함으로 종속변수와 독립변수들 사이의 상관정도는 작아지게 된다.

1940년대이래, 사람들은 측정오차에관계된 여러 가지 문제에 관심을 갖게 되었다. 측정오차에 관한 논평으로 Dalenius (1981) 와 Biemer *et al.* (1991) 등을 참고할 수 있다.

Carroll 과 Stefanski (1990)은 측정오차모형의 일반화된 형태를 정의한다. 그들은 관찰

* 창원대학교 통계학과 전임강사

들이 서로 통계적으로 독립일 때, 작은 측정오차는 모형추정에서 무시할 수 있음을 보여준다. 그러나 집락추출을 적용하는 복합조사설계의 경우, 일반적으로 집락 내 관찰들은 서로 독립이 아니다.

본 논문은 유한모집단과 복합표본설계에 기초한 측정오차분산의 추정을 다룬다. 이 때, 측정오차분산은 참값의 선형함수로 가정한다. Carroll과 Stefanski (1990)의 결과를 복합표본자료에 확장하여, 추정오차가 작은 경우 선형오차분산의 모수를 추정한다.

II. 측정오차분산의 추정

1. 측정오차 모형과 측정오차분산 함수

측정값 W는 참값 X의 불편의이고 각 표본 단위에 대해 두 개의 반복측정이 이루어진다고 가정하자. 표본에 있는 t번째 단위의 참값이 x_t라 할 때, Carroll 과 Stefanski (1990)의 기호를 따라 측정오차 모형은 다음과 같이 표현할 수 있다:

$$W_{tr} = x_t + \delta U_{tr} \tag{2.1}$$

$t = 1, 2, \cdots, n;\ r = 1, 2$, 여기서 $\delta(\geq 0)$은 측정오차의 크기를 나타내는 상수이고 U_{tr}는 확률변수로

$$E(U_{tr}|x_t) = 0 \text{ 이고 } Var(U_{tr}|x_t) = \Omega(x_t, \gamma)$$

이다.

많은 경우 측정오차분산은 참값이 증가함에 따라 증가 또는 감소하는 것으로 알려져 있다. 이 경우, 측정오차분산은 다음과 같이 표현될 수 있다:

$$\Omega_t = \gamma_0 + \gamma_1 x_t \tag{2.2}$$

여기서, $\Omega_t = \Omega(x_t, \gamma)$를 나타낸다.

모형 (2.1)로부터 δ를 알 때, 표본에 있는 t번째 단위에 대해 Ω_t와 참 값 x_t의 불편추정량은 각각 $\delta^{-2}S_t^2$, $S_t^2 = (W_{t1} - W_{t2})^2/2$,과 $\overline{W}_{t\cdot} = (W_{t1} + W_{t2})/2$가 된다. 식 (2.2)로

부터 이들 추정량들을 관찰 값으로 하여 모수 (γ_0, γ_1)을 추정할 수 있다(Davidian, 1990; Davidian 과 Carroll, 1987).

2. 모수의 추정

측정오차가 무시될 수 있을 만큼 작다면, Carroll 과 Stefanski(1990: 654)에 의해서, 식 (2.2)는 다음과 같이 표현될 수 있다:

$$Y_{\delta t} = \gamma_0 + \gamma_1 X_{\delta t} + \psi_t \tag{2.3}$$

여기서 $(Y_{\delta t}, X_{\delta t}) = (\delta^{-2} S_t^2, \ \overline{W}_{t\cdot})$이고 ψ_t는 오차 항이다. 모형 (2.1)을 가정할 때, 오차항 ψ_t의 기대값 $E(\psi_t \mid x_t) = 0$이고, ψ_t의 분산 $Var(\psi_t \mid x_t)$은 Ω_t의 함수로 표현된다. 즉, 식 (2.3)은 오차 분산이 동일하지 않은 단순선형회귀모형으로 볼 수 있다.

표본이 층화다단계추출에 의해 얻어지고 w_t는 표본에 포함된 t번째 단위에 적용되는 가중치라 하자. 모형 (2.3)에서 $\gamma = (\gamma_0, \gamma_1)'$의 최소제곱추정량은

$$\hat{\gamma} = M_{XX}^{-1} M_{XY} \tag{2.4}$$

이다. 이 때,

$$M_{XX} = N^{-1} \sum_{t=1}^{n} w_t (1 X_{\delta t})'(1 X_{\delta t}), \quad M_{XY} = N^{-1} \sum_{t=1}^{n} w_t (1 X_{\delta t})' Y_{\delta t}$$

이고 N은 모집단의 크기, n은 표본의 크기이다.

3. δ의 추정

n개의 표본 단위들에 대해 참값 (Ω_t, x_t)를 알고 있다면, 식 (2.2)의 γ의 추정량은

$$\hat{\gamma}_T = M_{xx}^{-1} M_{xy} \tag{2.5}$$

이고, 여기서

$$M_{xx} = N^{-1}\sum_{i=1}^{n} w_i(1x_i)'(1x_i), \quad M_{xy} = N^{-1}\sum_{i=1}^{n} w_i(1x_i)'\Omega_i$$

이다. 표본의 크기가 충분히 클 때, M_{xx}에 대한 M_{XX}의 상대적 크기는 참값 x_i들의 분산에 대한 표본 단위 내 분산 S_i들의 상대적 크기에 의존한다(Heo, 1999). 이러한 사실에 의해서 δ는 다음과 같이 다음과 같이 추정될 수 있다:

$$\delta = \{q^{-1}\mathrm{tr}(\ M_{xx}^{-1/2} S_{uu} M_{xx}^{-1/2})\}^{1/2} \tag{2.6}$$

여기서 $S_{uu} = diag(0, N^{-1}\sum_{i=1}^{n} w_i(S_i^2/2))$ 이고 $M_{xx} = M_{xx} - S_{uu}$이다.

III. 미국식 NHANES III 자료에 응용

1. 미국 NHANES III 자료

미국 제3차 국민건강영양조사(U. S. Third National Health and Nutrition Examination Survey: U. S. NHANES III)는 시설에서 생활하지 않는 일반시민의 건강과 영양 상태를 평가하기 위해서 1996년 미국 건강통계센터(U. S. National Center for Health Statistics: NCHS)에서 실시하였다. 이 조사는 미국 전역을 49개 층으로 층화한 층화다단계설계에 의해 표본을 추출하였다. 표본에 추출된 사람들은 건강상태에 대한 설문지에 응답하고 상세한 의료 검사를 받았다.

많은 설문지 응답과 의료적 측정치에 측정오차에 관한 의문이 제기되었고, NCHS에서는 초기표본 중 소수 부표본을 취해 재검사를 하였다. 이 과정에서 여러 가지 기술적, 상황적 제약에 따라 통상적인 이중추출을 적용할 수 없었다.

본 논문에서는, 경험적 분석을 위해 NHANES III의 의료 측정치 중의 하나인 총골밀도 측정치(TOBMD)를 선택하였다. 20세 이상 성인 16,573명의 일차표본 조사자 중 1,108명만이 두 번째 측정값을 제공했다.

2. 가중치의 조성

NHANES III의 초기 표본의 가중치를 w_{1t}라 하자. 재조사에 응한 사람들에 대해 이 가중치는 조정될 필요가 있다. 초기표본에 포함된 사람이 재조사에 응하여 측정치를 얻을 확률을 알고 그 확률을 p_t라 한다면, 가중치는 다음과 같이 조정될 수 있다:

$$w_{2t} = w_{1t} / p_t \tag{3.1}$$

따라서, s_r을 두개의 측정자료들을 모두 제공하는 사람들의 집합이라고 할 때, $N = \sum_{t \in s_r} w_{2t}$ 가 된다.

식 (3.1)에서 p_t는 보조변수들의 함수로 로지스틱회귀식에 의해 추정될 수 있다. 즉, x를 이 확률을 설명하는 보조변수들이라 할 때, $p_t = p(x_t)$는 다음과 같이 추정될 수 있다:

$$\log[p(x_t)/\{1-p(x_t)\}] = x_t\beta.$$

이 성향모형(propensity model)에 대한 자세한 내용은 'Eltinge 외(1997)'을 참고할 수 있다.

NAHNES III 의 총골밀도 측정치의 경우, 응답자의 인종, 성별, 나이, 거주지가 이 확률에 영향을 줄 것으로 여겨졌다. <표 1>은 이러한 보조변수들을 기초로 여러 로지스틱회귀식을 적용하여 최종적으로 얻은 회귀모형의 설명변수들로써 지시변수들을 나타낸다. <표 2>는 최종모형의 회귀계수 추정치를 나타낸다. <표 3>은 <표 2>로부터 추정된 확률을 가지고 조정된 가중치를 보여준다.

<표 3.1> 로지스틱회귀모형의 설명지시변수들

변수명	의미
(지역의 기본그룹)	(북동부)
Other Region	중서부, 남부, 서부
(인종의 기본그룹)	(히스패닉 계열이 아닌 백인)
Black	히스패닉 계열이 아닌 흑인
MAmer	멕시칸 계의 미국인
Other	기타
ORegion*Other	중서부, 남부, 서부×기타인종
(성의 기본그룹)	(남성)
Female	여성
(연령의 기본그룹)	(70세 이상)
Age20	20-29
Age30	30-39
Age40	40-49
Age50	50-59
Age60	60-69
Age iF	Age i × Female, i = 20,30,40,50,60

<표 2> 총골밀도의 두 번째 측정자료를 제공할 확률 추정을 위한 로지스틱회귀 계수의 점추정, 표준오차, 95% 근사 신뢰구간

Predictor	β_i	$se(\beta_i)$	(β_{iL}, β_{iU})
Intercept	-2.433	0.145	(-2.724, -2.141)
Other Region	-0.110	0.121	(-0.354, 0.133)
Black	-0.183	0.089	(-0.362, -0.003)
MAmer	-0.291	0.103	(-0.499, -0.083)
Other	0.269	0.267	(-0.267, 0.805)
ORegion*Other	-1.251	0.451	(-2.158, -0.344)
Female	-0.412	0.168	(-0.749, -0.075)
Age20	-0.019	0.208	(-0.437, 0.398)
Age30	-0.258	0.210	(-0.680, 0.164)
Age40	-0.015	0.240	(-0.498, 0.468)
Age50	0.279	0.246	(-0.215, 0.773)
Age60	0.611	0.144	(0.322, 0.900)
Age20F	0.183	0.249	(-0.318, 0.684)
Age30F	0.612	0.219	(0.172, 1.052)
Age40F	0.669	0.223	(0.221, 1.118)
Age50F	0.447	0.268	(-0.092, 0.986)
Age60F	0.105	0.259	(-0.415, 0.626)

<표 3> NHANES III의 초기표본 가중치와 조정된 가중치의 요약

가중치		합 계
W_{1hij}	전체초기 표본 가중치	1.772×10^8
	반복을 제공한 초기 표본 가중치	12,976,478
W_{2hij}	반복을 제공한 조정된 표본 가중치	1.771×10^8

<표 4> 복합표본에 근거한 측정오차분산회귀모형의 회귀계수 추정치, 표준오차, 95%의 근사 신뢰구간

설명변수	$\beta_i\times10^4$	$se(\beta_i)\times10^4$	$(\beta_{iL}, \beta_{iU})\times10^4$
절편	-0.995	4.017	(-9.069, 7.078)
$\overline{W}_{t\cdot}$	13.644	4.717	(4.165, 23.123)
Age20	-7.558	2.390	(-12.362, -2.754)
Age30	-4.151	1.716	(-7.598, -0.703)
Age40	-4.082	2.237	(-8.578, 0.413)
Age50	-4.807	2.077	(-8.981, -0.634)

3. 측정오차분산의 추정

식 (2.6)으로부터 총골밀도 측정치에 대해, $\delta = 0.0653$ 이다. 경험적으로 이 값은 측정오차를 고려하지 않아도 될 정도로 작은 것으로 볼 수 있다. 따라서, 측정오차분산함수의 모수 γ는 식 (2.3)의 선형 회귀식에 의하여 추정될 수 있다.

측정오차분산을 잘 설명하는 회귀식을 찾기 위해, 앞의 로지스틱 회귀모형에서 사용한 것과 동일한 인구 통계적 변수들과 BMI(body mass index)를 설명변수로 보았다. <표 3.4>는 이러한 설명변수들을 기초로 적합한 회귀식들로부터 최종적으로 선택된 회귀식의 계수들을 보여준다. <표 3.4>로부터 측정오차분산은 다음과 같이 표현될 수 있다:

$$\Omega_t = \beta_0 + \beta_1 x_t + \beta_{21}\text{Age20} + \beta_{22}\text{Age30} + \beta_{23}\text{Age40} + \beta_{24}\text{Age50}$$

이 모형에서 60세 이상의 인구는 기본연령집단이 된다.

III. 결론

본 연구에서는 관찰 값들이 참값에 불편의이고, 선형측정오차분산을 가정하였다. 이러한 가정 하에 먼저, 측정오차의 크기를 나타내는 상수 δ를 추정하였다. 이어서, Carroll과 Stefanski (1990)의 결과를 복합표본에 적용하여 δ가 작은 경우, 선형측정오차분산의 모수를 추정하였다.

이어서, 미국 제3차 국민건강영양조사의 의료자료 중 총골밀도 자료에 위의 결과를 적용하였다. 먼저, 로지스틱 회귀식을 이용하여 재조사에 응하여 자료를 제공할 확률을 추정하고, 그에 따라 가중치를 조정하였다. 조정된 가중치를 이용하여 δ를 추정하고, 측정오차분산 모형을 추정하였다.

본 연구에서는 선형측정오차분산을 가정하였다. 그러나, 이러한 가정은 비선형 모형으로 확대될 수 있을 것이다.

참고문헌

Biemer P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (eds.). 1991. *Measurement Errors in Surveys*. New York: John Wiley & Sons.

Carroll, R. J. and Stefanski, L. A. 1990. "Approximate quasi-likelihood estimation in models with surrogate predictors." *Journal of the American Statistical Association* 85: 652-663.

Dalenius, T. E. 1981. "The survey statistician's responsibility for both sampling and measurement error." In D. Krewski, R. Platex, and J. N. K. Rao (eds.), *Current Topics in Survey Sampling*, 17-29. New York: Academic Press.

Davidian, M. 1990. "Estimation of variance functions in assays with possible unequal replication and nonnormal data." *Biometrika* 77: 43-54.

Davidian, M. and Carroll, R. J. 1987. "Variance function estimation." *Journal of the American Statistical Association* 82: 1079-1091.

Eltinge, J. L., Heo, S., and Lee, S. R. 1997. "Use of propensity methods in the analysis of subsample re-measurements for NHANES III." In Proceedings of the Survey Methods Section, 27-36. Statistical Society of Canada.

Fuller, W. A. 1987. *Measurement Error Models*. New York: John Wiley.

Grove, R. M. 1991. "Measurement error across the disciplines." In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys* 1-25. New York: John Wiley & Sons.

Heo, S. 1999. "Diagnostics for survey inference accounting for incomplete data and measurement error." Unpublished Ph.D. dissertation, Department of Statistics, Texas A&M University, College Station, TX.