요인분석을 이용한 대체방법

이재갑[1] 이우리[2] 정재구[3] 이상은[4]

요 약

표본조사에서 발생되는 무응답에 대한 대체법은 매우 다양하게 연구 되고 있다. 특히 모형을 기반으로 하는 회귀 대체법은 매우 활용도가 높다. 이 때 일반적으로 종속변수가 결측값의 변수가 되며 독립변수는 주어지게 된다. 주어지 주어진 종속변수와 독립변수의 값을 이용하여 모델을 설정하고 그에 따라 결측값을 예측하여 대체하게 된다. 이 때 예측값 즉 결측값을 구하는 과정에서 독립변수 값 자체에도 결측값이 생기게 된다는 것이다. 이때 여러 가지 방법으로 독립변수의 결측값을 대체하고 모형을 활용할 수 있다. 그러나 이 연구에서는 독립변수들을 같은 특성끼리 그룹화 시키는 요인분석(factor analysis)을 이용하여 독립변수의 결측값에따른 예측된 결측값의 변동을 최소화 하고자했다.

keywords : 요인분석, 대체법, 회귀분석, 인자점수

1. 서론

표본조사에서 응답자가 응답 자체를 안 하거나 어떤 질문 항목에 대해 응답을 안 하는 경우가 있다. 이와 같이 무응답에 대한 두 가지 형태를 다루고자 한다.

단위 무응답(unit nonresponse)은 관찰 단위가 결측된 경우이고, 항목 무응답(item nonresponse)은 적어도 한 항목이 결측된 경우이다. 사람을 대상으로 하는 설문조사에서, 단위 무응답은 응답자가 설문조사에 대해서 아무런 정보를 주지 않는 경우이고 항목 무응답은 응답자가 설문지의 특정 항목에 응답을 하지 않음을 의미한다.

이러한 무응답으로 인해 발생한 결측항목에 결측값을 할당하게 되며 이를 대체법(imputation)이라 한다. 대체법에 대한 연구는 연역적 대체(deductive imputation), 칸 평균 대체(cell mean imputation), 핫덱(hot deck), 회귀 대체(regression imputation)등 매우 다양하게 되고 있다.

특히, 모형을 기반으로 하는 회귀 대체법에서는 일반적으로 종속변수가 결측값의 변수가 되며 독립변수는 주어지게 된다. 주어진 종속변수와 독립변수의 값을 이용하여 모델을 설정하고 그에 따라 결측값을 예측하여 대체하게 된다. 그런데 이 때 예측값 즉, 결측값을 구하는 과정에서 독립변수 값 자체에도 결측값이 생기게 된다. 이 때 여러 가지 방법으로 독립변수의 결측값을 대체하고 모형을 활용할 수 있다. 그러나 본 연구에서는 독립변수들을 같은 특성끼리 그룹화

1) 경기대학교 응용정보통계학과 대학원 random@kyonggi.ac.kr
2 경기대학교 응용정보통계학과 교수 wrlee@kyonggi.ac.kr
3) 경기대학교 응용정보통계학과 교수 jkchung24@hanmail.net
4) 경기대학교 응용정보통계학과 조교수 sanglee@stat.kyonggi.ac.kr

시키는 요인분석(factor analysis)을 이용하여 독립변수의 결측값에 따른 예측된 결측값의 변동을 최소화 하고자했다.

본 연구에서는 도시가계조사에서 얻은 소득 자료를 이용하여 분석하였다.

2. 요인분석을 이용한 회귀모형

2.1 요인분석모형

X 는 $px1$ 벡터로 관측된 자료로 $E(X)=0$, $var(X)=\Sigma$라 하자.

기본 직교모형은 다음과 같다.

$$X=\Lambda F+\varepsilon \qquad (2.1)$$

여기서 Λ : 인자적재행렬
F : 공통인자
ε : 특정인자

이며 가정은 다음과 같다.

$F\sim N(0, I_m)$, $m\leq p$
$\varepsilon\sim N(0, D_\phi)$, $where\ D_\phi=(diag(\psi_1,\ldots,\psi_m)$;
ε 와 F는 서로 독립이며
$\Lambda' D_\phi^{-1}\Lambda_{j\,m\times m}=diag(J_1,\ldots j_m)\ where\ J_1\rangle J_2\rangle\cdots\rangle J_m$.

2.1.1 Λ와 D의 추정

Λ와 $\psi_1,\ldots,\psi_p$는 최대우도추정량(MLE) 방법에 의해 다음과 같이 구해진다.

$$diag(D_{\hat{\phi}}+\hat{\Lambda}\hat{\Lambda}')=diag(S)$$
$$SD_{\hat{\phi}}^{-1}\hat{\Lambda}=\hat{\Lambda}(I+\hat{\Lambda}'D_{\hat{\phi}}^{-1}\hat{D}) \qquad (2.2)$$

여기서 $S=\frac{1}{N}\Sigma_i^N(x_j-\bar{x})(x_j-\bar{x})'$ 이며

$\hat{\Lambda}$ 과 $D_{\hat{\phi}}$은 Λ와 D_ϕ 의 MLE 이다.

2.1.2 F의 추정

(2.1)로부터 $var(x_j)=\Sigma=\Lambda\Lambda'+D_\phi$

그리고 $E(f_j)$를 x_j들의 선형결합이라고 가정하자.
즉,

$$(f_j|x_j)=Ax_j=u_j\ ,\ j=1,\ldots,N \qquad (2.3)$$

여기서 $E(u_j)=0$, $var(u_j)=\sigma^2 I_m$.

$F=(f_j,\ldots,f_n)$, $U=(u_1,\ldots,u_n)$, $X=(x_1,\ldots,x_N)$ 이며, 모형을 행렬로 표현하면 다음과 같다.

$$(F|X)=AX+U$$

이때 F가 알려진 값이라면 OLS에 의해 A는 다음과 같다.

$$\hat{A}=(FA')(XX')^{-1}$$

그러면 $E(FX')=N\Lambda'$ 이 되면 대수의 법칙에 따라

$$\hat{A}\approx N\hat{\Lambda}'(XX')^{-1}=\hat{\lambda}(\frac{XX'}{N})^{-1}$$

이며 표본공분산행렬(sample coavariance matrix)를 사용해서 다시 표현하면 다음과 같다.

$$\hat{A}\approx\Lambda'(S)^{-1}$$

식 2.3 으로부터 $\hat{F}=\hat{A}X$ 혹은 $\hat{f}_j=\hat{A}x_j$ 로부터 factor score f_j는 다음과 같이 추정된다.

$$\hat{f}_j\approx\hat{\Lambda}'(S)^{-1}x_j,\ j=1,\ldots,N \qquad (2.4)$$

2.2 회귀모형

(2.4) 로 부터 다음과 같은 회귀모형을 설정한다.

$$Y=\hat{F}\cdot\beta+\varepsilon,\ \varepsilon\sim N(0,I\cdot\sigma^2)$$

여기서 $\varepsilon\sim N(0,I\cdot\sigma^2)\Rightarrow Y\sim N(F\beta,I\cdot\sigma^2)$

이 때 결측값으로 활용되는 $\hat{Y}$는 다음과 같다.

$$\hat{Y} = \hat{F}\,\hat{\beta}$$

여기서 $\hat{\beta} = (\hat{F}'\hat{F})^{-1}\hat{F}'Y$ 이다.

3. 대체법

2장에서 얻은 예측값을 결측값에 대체하여 추정량을 구한다,

$$\overline{\hat{Y}} = \frac{1}{N}\left(\sum_{i}^{N-n} Y_i + \sum_{i}^{n} \hat{Y}\right)$$

여기서 $\overline{\hat{Y}}$는 대체된 데이터의 총수입평균, $N-n$은 결측치를 제외한 관측치수, 즉 응답자수이고, Y는 실제 응답값이며, 그리고 $\hat{Y}$는 2장에서 얻은 예측값이다.

4. 모의실험

4.1 자료기술

본 연구에 사용된 data는 통계청에서 매월 조사하고 있는 '도시가계조사 - 2001년 9월'자료이며 5107개의 가구를 대상으로 한 것이다.

'도시가계조사'는 '가구의 수입과 지출에 관한 사항', '가구주 및 배우자에 관한 사항', '가구원에 관한 사항', '주거에 관한 사항'에 관하여 조사된 것이다.

본 연구에서는 그 중 '총수입'에 영향을 많이 끼칠 것으로 여겨지는 '총지출', '가구주소득', '가구주연령', '가구주교육정도', '가구원수', '직업이 있는 가구원수', '가구주직업', '소득', '연간소득', '주택가격' 의 10개 변수를 사용하였다.

연구를 위해 데이터 정제 작업을 우선적으로 하였으며 그 단계는 첫째, 총수입을 포함한 11개의 각 변수에서 결측치가 있는 관측치를 삭제하고, '가구주연령', '가구원수', '가구주직업', '가구주성별', '주택가격', '가구주소득'의 6개 변수에서는 관찰값이 0 인 경우도 삭제하였다.

이 과정을 통해 1518개 관측치를 가진 자료가 생성되었으며, 그 자료를 이 후의 연구진행에 사용하였다.

두 번째로 소득에 관련된 5개 변수들('총수입', '총지출', '가구주소득', '연간소득', '소득')은 로그변환을 하여 정규성을 따르게끔 변환과정을 하였다.

세 번째로 '가구주교육정도', '가구주직업' 의 범주형 변수는 연속형 변수로 변환하는 과정을 각각 하였다. 변환과정은 우선 각 변수의 평균과 표준편차를 가진 정규분포를 따르는 난수를 각각 생성하고, 원자료의 각 변수의 개별 범주별 해당도수만큼 난수를 나누어서 정규분포를 따

르는 연속형 변수로 만들었다.
그 중 '가구주직업'은 총소득과 분산분석결과 영향력이 낮은 것으로 판단되어 제외하였다.
마지막으로, 변수들의 단위가 서로 다르므로 '(관측값-평균)/표준편차'를 해주어 표준화변환을 하여 통일을 하였다.

4.2 요인분석을 이용한 회귀모형의 적용

기본 직교모형은 다음과 같다.

$$X_{9\times1} = \Lambda_{9\times3} F_{3\times1} + \varepsilon_{9\times1}$$

여기서 , X_1 : 총지출, X_2 : 가구주 교육정도, X_3 : 가구주소득, X_4 : 가구주연령
X_5 : 가구원수, X_6 : 직업을 가진 가구원수, X_7 : 연간소득, X_8 : 주택가격,
X_9 : 소득.

가정은 다음과 같다.

$F \sim N(0, I_3)$,
$\varepsilon \sim N(0, D_\phi)$, *where* $D_\phi = (diag(\phi_1, \phi_2, \phi_3)$;
ε 와 F는 서로 독립이며
$\Lambda' D_\phi^{-1} \Lambda_{j\,3\times3} = diag(J_1, J_2, J_3)$ *where* $J_1 > J_2 > J_3$.

f_j의 추정량은 $\hat{f}_j \approx \hat{\Lambda}'(S)^{-1} x_j$, $j = 1, \ldots, 1518$ 이며,
이에 따른 회귀모형은 $Y = \hat{F} \cdot \beta + \varepsilon$, $\varepsilon \sim N(0, I \cdot \sigma^2)$ 이다.
여기서, $\varepsilon \sim N(0, I \cdot \sigma^2) \Rightarrow Y \sim N(F\beta, I \cdot \sigma^2)$

결측값을 대체한 후의 '총수입(Y)'의 평균은 다음과 같다.

$$\overline{\hat{Y}} = \frac{1}{N}\left(\sum_{i}^{N-n} Y_i + \sum_{i}^{n} \hat{Y}\right)$$

여기서, $\hat{Y} = \hat{F}\,\hat{\beta}$
Y는 실제 응답값이며,
n은 무응답수, $N-n$은 응답자수

5. 결과

1518개 관찰치에서 임의로 5%를 결측치로 처리하여 대체한 후 실제값과 대체한 후의 값의 평균을 비교해 본다.

실제 총수입의 평균값 : $\overline{Y}=\frac{1}{1518}\sum_{i=1}^{1518} Y_i$.

결측값을 대체 후 총수입의 평균 : $\overline{\hat{Y}}=\frac{1}{1518}(\sum_{i}^{1518-76} Y_i+\sum_{i}^{76}\hat{Y})$

위의 과정을 100회 반복 시행하여 각각의 평균을 보면 다음과 같다.

log(실제 총수입의 평균)	log(대체 후 총수입의 평균)
15 .3968999	15 .3968171

참고문헌

Roderick J. A. Little, Donald B. Rubin(1987), 'Statistical analysis with missing data', John Wiley & Sons, U.S.A.

Donald B. Rubin(1987), Multiple imputation for nonresponse in surveys, John Wiley & Sons, U.S.A.

Alexander Basilevsky(1994), Statistical factor analysis and related methods, John Wiley & Sons, U.S.A

John L. Eltinge , Ibrahim S. Yansaneh(1997), Diagnostics for formation of Nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey, Survey Methodology, Vol. 23, No.1, page 33-40

Olga Troyanskaya, Michael Cantor(2001), Missing value estimation methods for DNA microarrays, Bioinformatics, Vol. 17, No. 6, page 520-525