범주형 자료의 진단방법에 관한 연구
A Study on Diagnostics Method for Categorical Data

이 선 규*
Lee, Sun-Kyu
조 범 석**
Cho, Bum-Suk

Abstract

In this study we are concerned with the diagnostics method of cross-classified categorical data using logistic regression model of binary response models for cell proportions. under this model, we could examine the goodness-of-fit of the models using Pearson's χ^2 test statistic and likelihood ratio statistic. Under this model, these statistics are assumed that sample survey schemes are with replacement sampling model. But, these statistics are often inappropriate for analysing contingency tables consists of complex sampling schemes obtained sample survey data.

In this study we are examined diagnostics procedures detecting any outlying cell proportions and influential observations on design space in logistic regression modeltake account of the survey design effects.

1.서 론

분할표를 구성하는 범주형 자료에 대한 분석방법은 다항표본추출 또는 적다항표본추출의 가정 하에서 집약적으로 발전되어왔다. 다차원 분할표에서 대수선형모형(loglinear model)과 로짓모형(logit model)을 이용한 분석방법은 여러가지 가설의 검정을 체계적으로 제시하는 연속형 자료에 대한 분산분석과 유사하다. 이 방법을 실행하기 위해 이용되는 *SAS*, *SPSS*와 같은 통계 패키지는 각 칸의 빈도수가 다항표본추출모형에 의해 추출되었거나, 각 칸의 관찰값이 서로 독립이고, 동일한 분포를 따른다는 가정 또는 단순임의표본추출(simple random sampling)에 의해 추출되었다는 가정 하에서 실행되고 있다.

표본조사를 실시하기 위해 이용되는 표본추출방법은 여러가지가 있지만, 대부분 층화나 집락추출을 결합한 방법이다.표본추출계획이 층화표본추출이나 집락표본추출과 같은 복합표본추출(complex sampling)인 경우, 관심있는 모집단이 몇개의 층이나 집락으로 층화되거나, 집락화(clustering)된다. 본 연구에서는 로지스틱 회귀모형하에서 범주형 자료에 대해 이용되어온 기존의 진단방법을 조사설계를 고려하여 수정하며, 층화표본추출에 의해 추출된 실증적 자료를 이용하여 수정된 통계량 및 진단방법을 적용시켜 본다.

2.모형

크기 N인 관심있는 유한모집단이 표본조사에서 측정된 하나 또는 그 이상의 범주형 설명변수의 범주에 따라 T개의 상호배반적인 칸으로 분할되었다고하자. 그리고, t번째 칸의 크기를 N_t, t번째 칸에서 성공의 값을 갖는 이항확률변수의 총합을 N_{t1}이라 하고, 성공율을 $\pi_t = N_{t1}/N_t$라고 하자.

t번째 칸의 성공율 π_t에 대한 일반적인 이항응답모형은 다음과 같다.

---

* 금오공과대학교 산업공학과 부교수
** 성균관대학교 통계학과

$$\pi_t = f_t(\underline{\theta}) \quad , \qquad t = 1, \cdots, T \tag{1}$$

여기서 $\underline{\theta}$는 $r\times 1$모수벡터 ($r \le T$)이고, $f_t(\underline{\theta})$는 Birch의 정규성조건을 만족하는 함수이다. $f_t(\underline{\theta})$에 대한 예로서 logit model, probit model, complementary log-log model 등이 있다.

만약, 유한모집단에서 n개의 최종단위로 구성된 표본 s가 특정한 표본조사설계에 의해 추출되었다고 하고, $\hat{N}$과 $\hat{N}_t$, $\hat{N}_{tl}$을 모집단크기 N의 조사추정값, t번째 칸크기 N_t의 조사추정값, t번째 칸에서 성공의 횟수에 대한 조사추정값이라고 하자.

t번째 칸에서 성공율 π_t의 조사추정값은 다음과 같다.

$$\hat{p}_t = \hat{N}_{tl} / \hat{N}_t \quad , \qquad t = 1, \cdots, T$$

독립변수의 벡터 $\underline{x}$ 가 주어진 경우에 이항응답변수 y의 조건부 분포

$$\Pr(y_t | x_t) = \pi_t^{y_t}(1-\pi_t)^{1-y_t} \quad , \qquad y_t \in (0, 1)$$

에 대한 로지스틱 회귀모형을 생각하자.

식(1)의 이항응답모형에서 로지스틱 회귀모형은 다음과 같다.

$$v_t = v(\pi_t) = \ln\left[\frac{f_t(\underline{\theta})}{1-f_t(\underline{\theta})} \right] = \underline{x}_t'\underline{\theta} \quad , \quad t = 1, \cdots, T \tag{2}$$

여기서 $\pi_t = f_t(\underline{\theta}) = [\,1 + \exp(-\underline{x}_t'\underline{\theta})]^{-1}$이고, $\underline{x}_t$는 설명변수의 범주로부터 구한 $r\times 1$벡터이며, $\underline{\theta}$는 $r\times 1$모수벡터이다.

2.1 모수추정

각 칸에서 독립적인 이항표본추출이 이용될 경우에 우도방정식은 다음과 같다.

$$X' D_n \underline{\hat{f}} = X' D_n \underline{q} \tag{3}$$

여기서 $X = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_T)'$은 계수 r을 갖는 $T\times r$행렬이고, $\underline{q}$는 표본비율 $q_t = n_{tl}/n_t$ 의 벡터이며, $D_n = diag(n_1, n_2, \cdots, n_T)$, n_t는 t번째 칸의 표본크기 ($\sum n_t = n$) 이고, n_{tl}은 t번째 칸의 표본총합이다.

실제로 표본추출에 이용되는 표본조사계획이 적이항표본추출이 아닌 경우에 식(2)에 대한 의사우도방정식은 다음과 같다.

$$\left(\frac{\partial \underline{f}}{\partial \underline{\theta}}\right)'_{\underline{\theta}=\hat{\underline{\theta}}} D^{-1}_{f(\hat{\theta})} D^{-1}_{[1-f(\hat{\theta})]} D_w \underline{\hat{p}} = \left(\frac{\partial \underline{f}}{\partial \underline{\theta}}\right)'_{\underline{\theta}=\hat{\underline{\theta}}} D^{-1}_{[1-f(\hat{\theta})]} \underline{w}$$

여기서 $\underline{\hat{p}} = \hat{N}_{tl}/\hat{N}_t$ 은 $\underline{\pi}$ 의 조사추정값이고, $\underline{\hat{\theta}}$ 은 $\underline{\theta}$ 에 대한 의사우도추정량이며, $\underline{\hat{f}} = \underline{f}(\underline{\hat{\theta}})$ 은 $\underline{f}(\underline{\theta})$에 대한 의사우도추정량이고, $\underline{w} = (w_1, \cdots, w_T)'$, $w_t = \hat{N}_t/\hat{N}$은 t번째 칸에서 모비율의 조사추정값이다.

모형(2)에 대한 의사우도방정식에서 $\frac{\partial \underline{f}}{\partial \underline{\theta}} = D^0 X$ 이고, $D^0 = D_{f(\hat{\theta})} D_{[1-f(\hat{\theta})]}$ 는 t번째 대각원소가 $d_{tt} = \left(\frac{\partial v_t}{\partial f_t}\right)^{-1}$인 대각행렬이다.

식(2)에 대한 의사우도방정식에 $\frac{\partial \underline{f}}{\partial \underline{\theta}} = D^0 X$ 를 대입하면

$$X'\,\underline{\hat{v}} = \underline{0} \tag{4}$$

이 되고, 여기서 $\underline{w} = D_{\hat{f}}^{-1} D_w \underline{\hat{f}}$ 이고, $\underline{\hat{v}} = \hat{D}^0 D_{\hat{f}}^{-1} D^{-1}_{(1-\hat{f})} D_w (\underline{\hat{p}} - \underline{\hat{f}})$이며, $\hat{D}^0$는 f대신에 $\hat{f}$을 대입하여 구한 D^0값이다.

식(4)에서 k번째 방정식은

$$\sum^{T} x_{tk}\,\hat{y}_t = \sum^{T} \frac{x_{tk}\,\hat{d}_{tt}\,w_t(\hat{p}_t - \hat{f}_t)}{[\,\hat{f}_t(1-\hat{f}_t)\,]} = 0 \qquad (5)$$

이고, 식(5)의 k번째 편미분 방정식은 다음과 같다.

$$\frac{\partial(\sum^{T} x_{tk}\,\hat{y}_t)}{\partial\theta_j} = -\sum^{T} x_{tk}\,\hat{b}_t\,x_{tj}$$

이를 행렬로 나타내면

$$\left(\frac{\partial(X'\hat{y})}{\partial\underline{\theta}}\right) = -(X'\hat{B}X)$$

이 된다. 여기서 $\hat{B}$은 대각행렬로 t번째 대각원소 $\hat{b}_t$은

$$\hat{b}_t = \hat{d}_{tt}^{\,2}\,w_t\,\hat{f}_t^{-1}(1-\hat{f}_t)^{-1}\cdot$$

$$\left\{1 + (\hat{p}_t - \hat{f}_t)\left[\,\hat{d}_{tt}\frac{\partial(\hat{d}_{tt}^{-1})}{\partial\hat{f}_t} + \frac{(1-2\hat{f}_t)}{\hat{f}_t(1-\hat{f}_t)}\right]\right\} \qquad (6)$$

이다.

로지스틱 회귀모형인 경우

$$\hat{d}_{tt} = \hat{f}_t(1-\hat{f}_t)$$

이므로

$$\hat{b}_t = w_t\hat{f}_t(1-\hat{f}_t) \qquad (7)$$

이 된다.

식(4)가 $\underline{\hat{\theta}}$에 대해 비선형이기 때문에 해를 구하기 위해서 반복적인 계산법이 필요하므로 *Newton−Raphson* 방법을 이용하면 $\underline{\theta}$의 추정값은 다음과 같다.

$$\underline{\theta}^{(i+1)} = \underline{\theta}^{(i)} - \left[\frac{\partial(X'\underline{y}^{(i)})}{\partial\underline{\theta}^{(i)}}\right] X'\underline{y}^{(i)}$$

$$= \underline{\theta}^{(i)} + (X'B^{(i)}X)^{-1}X'\underline{y}^{(i)} \qquad (8)$$

여기서 $\underline{\theta}^{(i)}$는 i번째 반복후에 구한 $\underline{\theta}$값이고, $B^{(i)}$는 i번째 반복후에 구한 $\hat{B}$값이며, $\underline{y}^{(i)}$는 i번째 반복후에 구한 $\underline{y}$값이다.

2.2 추정량의 점근적 특성

추정량 $\underline{\hat{\theta}}$과 $\underline{f}(\underline{\hat{\theta}})$의 점근적 특성에 대해 알아보자.

<정리 1> [16] Birch의 정규성조건하에서

$$(\underline{\hat{\theta}} - \underline{\theta}) = (X'\Delta X)^{-1}X'D_{\underline{W}}[\,\underline{\pi} - \underline{f}(\underline{\theta})]$$

이고

$$\underline{f}(\underline{\hat{\theta}}) - \underline{f}(\underline{\theta}) \approx D_{\underline{W}}^{-1}\Delta X(\underline{\hat{\theta}} - \underline{\theta})$$

이다. 여기서 $\Delta = diag\{W_1f_1(1-f_1), \cdots, W_Tf_T(1-f_T)\}$ 이고,

$D_{\underline{W}} = diag(W_1, \cdots, W_T)$ 이며, $W_t = N_{t1} / N_t$이다

추정량 $\underline{\hat{\theta}}$ 과 $\underline{f}(\underline{\hat{\theta}})$ 의 점근적 공분산행렬에 대해서 알아보자.

$\underline{\hat{p}}$의 공분산행렬이 $n^{-1}V$ 이므로 $\underline{\hat{\theta}}$ 의 점근적 공분산행렬은

$$Cov(\underline{\hat{\theta}}) = n^{-1}(X'\Delta X)^{-1}(X'D_{\underline{W}}VD_{\underline{W}}X)(X'\Delta X)^{-1} \equiv T \qquad (9)$$

이 되고, $f(\hat{\theta}) = \hat{I}$ 의 점근적 공분산행렬은

$$Cov(\hat{I}) = X' D_{\pi} D_{(1-\pi)} T D_{\pi} D_{(1-\pi)} X$$
$$= D_W^{-1} \Delta X T X' \Delta D_W^{-1} \quad (10)$$

이다.

2.3 적합도 검정통계량

로지스틱 회귀모형의 적합도를 검정하기 위한 통계량을 정의하자.

<정의 1> 로지스틱 회귀모형의 적합도를 검정하기 위한 *Pearson* 검정통계량은

$$X_B^2 = n \sum_{t=1}^{T} (\hat{p}_t - \hat{f}_t)^2 w_t / [\ \hat{f}_t(1 - \hat{f}_t)] \quad (11)$$

이고, 우도비 검정통계량은

$$G_B^2 = 2n \sum_{t=1}^{T} w_t \{\ \hat{p}_t \ln(\hat{p}_t / \hat{f}_t) + (1 - \hat{p}_t) \ln[\ (1 - \hat{p}_t)/(1 - \hat{f}_t)]\ \} \quad (12)$$

이다.

독립적인 이항표본추출하에서 X_B^2과 G_B^2은 자유도 $T-r$ 를 갖는 χ^2분포를 따른다는 사실은 잘 알려져 있다. [9]

일반적인 복합표본조사설계에 대해서 X_B^2과 G_B^2의 점근분포는 다음과 같다.

$$X_B^2 \sim \sum \delta_t W_t \quad , \quad t = 1, \cdots, T-r$$

여기서 $W_t \sim \chi_1^2$이고, δ_t는 행렬 $(C' \Delta^{-1} C)^{-1} (C' \Delta^{-1} D_W V D_W \Delta^{-1} C)$ 의 고유값이며, C는 계수 $T-r$ 를 갖는 $T \times (T-r)$ 행렬, $C'X = O$ 이다.

로지스틱 회귀모형의 적합도를 검정하기 위한 검정통계량의 점근분포에 대해서 알아보자. 모형(2)하에서 이들 통계량의 점근분포는 축소모형을 이용하여 구할 수 있다.

행렬 X가 (X_1, X_2) 로 분할되었다고 하자. 그러면 X_1은 $T \times s$행렬이고, X_2는 $T \times u$행렬이다.

모형(2)에 대한 축소모형은 다음과 같다.

$$\underline{v} = X\underline{\theta} = X_1\underline{\theta}_1 + X_2\underline{\theta}_2 \quad (13)$$

여기서 $\underline{v} = (v_1, \cdots, v_T)'$ 이고, $X = (X_1, X_2)$ 이며, $\underline{\theta} = (\underline{\theta}_1', \underline{\theta}_2')'$, $\underline{\theta}_1$은 $s \times 1$벡터, $\underline{\theta}_2$는 $u \times 1$벡터 $(s + u = r)$ 이다.

모형(13)에서 귀무가설 H_0^*: $\underline{\theta}_2 = \underline{0}$ 를 검정할 경우에 의사우도방정식은 다음과 같다.

$$X_1' D_w \hat{\underline{I}} = X_1' D_w \hat{\underline{p}} \quad (14)$$

여기서 $\tilde{\underline{\theta}}_1$는 $\underline{\theta}_1$에 대한 의사최우추정량이고, $\tilde{\underline{I}} = f(\tilde{\underline{\theta}}_1)$ 은 $f(\underline{\theta}_1)$에 대한 의사최우추정량이다.

식(14)에서 반복적인 계산에 의해 $\underline{\theta}_1$과 $f(\underline{\theta}_1)$의 추정값을 구할 수 있다. 모형(2)에서 정의한 F는 모형(13)에 대해서 다음과 같다.

$$F = \left(\frac{\partial f}{\partial \underline{\theta}}\right) = \left(\frac{\partial f}{\partial \underline{\theta}_1} \ , \ \frac{\partial f}{\partial \underline{\theta}_2}\right) = D^0 X$$

여기서

$$F_1 = \left(\frac{\partial f}{\partial \underline{\theta}_1}\right) = D^0 X_1 \text{이고,} \qquad F_2 = \left(\frac{\partial f}{\partial \underline{\theta}_2}\right) = D^0 X_2$$

이다.

모형(13)에서 귀무가설 H_0^*: $\underline{\theta}_2 = \underline{0}$ 를 검정하기 위한 검정통계량을 정의하자.

<정의 2> 귀무가설 H_0^*: $\underline{\theta}_2 = \underline{0}$ 를 검정하기 위한 *Pearson* 검정통계량은

$$X_B^2 = n\sum^{T}(\hat{f}_t - \hat{f}_t)^2 w_t / [\ \hat{f}_t(1-\hat{f}_t)] \qquad (15)$$

이고, 우도비 검정통계량은

$$G_B^2 = 2n\sum^{T} w_t\{(\hat{f}_t \ln(\hat{f}_t / \hat{f}_t) + (1-\hat{f}_t)\ln[\ (1-\hat{f}_t)/(1-\hat{f}_t)]\ \} \qquad (16)$$

이다.

귀무가설 H_0^*: $\underline{\theta}_2 = \underline{0}$하에서 X_B^2과 G_B^2은 점근적으로 동일하고, 이항표본추출에 대해서 자유도 u를 갖는 χ^2분포에 근사한다는 사실은 잘 알려져 있다. [9]

특정한 조사설계 $p(s)$에 대해서 귀무가설 H_0^*: $\underline{\theta}_2 = \underline{0}$ 가 사실인 경우 X_B^2과 G_B^2의 점근분포에 대해서 알아보자.

<정리 2> [16] 귀무가설 H_0^*: $\underline{\theta}_2 = \underline{0}$ 하에서

$$X_B^2 = n\hat{\underline{\theta}}_2'(\tilde{X}_2'\Delta\tilde{X}_2)\hat{\underline{\theta}}_2$$

이고

$$X_B^2 \approx \sum^{u}\delta_t W_t$$

이다. 여기서 $W_t \sim \chi_1^2$이고, δ_t는 일반화설계효과행렬 $(\tilde{X}_2'\Delta\tilde{X}_2)^{-1}(\tilde{X}_2' D_{\underline{W}} V D_{\underline{W}}\tilde{X}_2)$의 고유값이며, $\tilde{X}_2 = [\ I - X_1(X_1'\Delta X_1)^{-1}X_1'\Delta]\ X_2$ 이다.

모형(2)는 모형(13)의 완전모형이므로 (X, X_2) 가 $T\times T$행렬이 되는 $X_1 = X$ 이고, X_2가 $T\times(T-r)$ 행렬이다.

$$C = \Delta\tilde{X}_2$$

이면

$$C'X = \tilde{X}_2'\Delta X = X_2'[\ I - \Delta X(X'\Delta X)^{-1}X'\ ]\ \Delta X = O$$

이고

$$(\tilde{X}_2'\Delta\tilde{X}_2)^{-1}[\ \tilde{X}_2' D_{\underline{W}} V D_{\underline{W}}\tilde{X}_2] = (C'\Delta^{-1}C)^{-1}[\ C\Delta^{-1}D_{\underline{W}} V D_{\underline{W}}\Delta^{-1}C]$$

가 되므로 X_B^2과 G_B^2의 점근분포는 다음과 같다.

$$X_B^2 \approx \sum^{u}\delta_t W_t$$

여기서 δ_t는 $(C'\Delta^{-1}C)^{-1}[\ C\Delta^{-1}D_{\underline{W}} V D_{\underline{W}}\Delta^{-1}C]$ 의 고유값이다.

이항표본추출의 경우에 모든 칸에 대해서 고유값이 $\delta_t = 1$이므로 귀무가설 H_0^*에 대해서 X_B^2과 G_B^2의 점근분포는 χ_u^2이 된다.

X_B^2이나 G_B^2의 점근분포가 복잡하기 때문에 표본조사자료를 이용하여 분석을 실행할 경우 이들 통계량을 수정할 필요가 있다.

<정의 3> X_B^2과 G_B^2의 수정된 검정통계량은 다음과 같다.

$$X_M^2 = X_B^2 / \hat{\delta}_. \qquad (17)$$

$$G_M^2 = G_B^2 / \hat{\delta}_. \qquad (18)$$

여기서 $(T-r)\hat{\delta}_. = \sum\hat{\delta}_t = n\sum^{T}\hat{v}_{tt}w_t / [\ (\hat{f}_t(1-\hat{f}_t)]$ 이고, $\hat{\delta}_.$은 $\delta_.$의 조사추정값이며,

$\hat{v}_{tt}$ 은 $\widehat{Cov}(\hat{\underline{p}}-\hat{\underline{f}}) = n^{-1}S\hat{V}S'$ 의 t번째 대각원소이고,

$S = I - D_w^{-1}\hat{\Delta}X(X'\hat{\Delta}X)^{-1}X'D_w$, $\quad \hat{V} = \widehat{Cov}(\sqrt{n}\,\hat{\underline{p}})$ 이며,

$\hat{\Delta} = diag[\ w_1\hat{f}_1(1-\hat{f}_1), \cdots, w_T\hat{f}_T(1-\hat{f}_T)]$

이다.

축소모형에서 X_B^2과 G_B^2의 수정된 검정통계량은 다음과 같다.

$$X_M^2 = X_B^2 / \hat{\delta}_.$$

$$G_M^2 = G_B^2 / \hat{\delta}_.$$

여기서 $u\hat{\delta}_. = \sum \hat{\delta}_t = n\sum^T \tilde{v}_{tt} w_t / [\ (\hat{f}_t(1-\hat{f}_t)]$ 이고, $\hat{\delta}_.$은 $\delta_.$의 조사추정값이며, $\tilde{v}_{tt}$는 잔차의 공분산행렬의 추정값 $\widehat{Cov}(\hat{\underline{f}}-\hat{\underline{f}}) = n^{-1}D_w^{-1}\hat{\Delta}\hat{X}_2\hat{A}\hat{X}_2'\hat{\Delta}\times D_w^{-1}$의 t번째 대각원소이고,

$\hat{A} = (\hat{X}_2'\hat{\Delta}\hat{X}_2)^{-1}\ [\ \hat{X}_2'D_w\hat{V}D_w\hat{X}_2]\ (\hat{X}_2'\hat{\Delta}\times\hat{X}_2)^{-1}$ 이다.

3.범주형 자료의 진단

3.1 조사설계를 고려한 진단방법

표본추출이 적이항표본추출인 경우에 로지스틱 회귀모형에서 $\underline{r} = \hat{\underline{p}} - \hat{\underline{f}}$의 t번째 원소 r_t에 대한 표준화 잔차는 다음과 같다.

$$e_t = \frac{(\hat{p}_t - \hat{f}_t)}{[\ w_t\hat{f}_t(1-\hat{f}_t)\ ]^{1/2}} \tag{19}$$

e_t는 진단에서 중요한 역할을 하며, 이 값은 임의의 칸의 비율이 특이값인지를 식별해 준다.

조사설계를 고려한 경우에 칸 비율이 특이한 경향을 갖는 지를 탐지하기 위해서 로지스틱 회귀모형에 대한 잔차벡터를 정의하자.

잔차벡터를 $\underline{r} = \hat{\underline{p}} - \hat{\underline{f}}$ 이라고 하면 이항응답모형의 경우에

$$\underline{r} = (\hat{\underline{p}} - \hat{\underline{f}}) = [\ \hat{\underline{p}} - \underline{f}(\underline{\theta})\ ] - [\ \underline{f}(\hat{\underline{\theta}}) - \underline{f}(\underline{\theta})]$$

$$= [\ I - F(F'A^*F)^{-1}F'A^*]\ [\ \hat{\underline{p}} - \underline{f}(\underline{\theta})\ ] \tag{20}$$

이므로 로지스틱 회귀모형인 경우는 다음과 같다.

$$\underline{r} = [\ I - D^0X(X'D_{[\ W\hat{f}(1-\hat{f})\ ]}X)^{-1}X'D_W]\ [\ \hat{\underline{p}} - \underline{f}(\underline{\theta})]$$

조사설계를 고려한 표준화 잔차는

$$e_t^* = \frac{(\hat{p}_t - \hat{f}_t)}{[\ \hat{v}_{tt}(\hat{\underline{p}} - \hat{\underline{f}})\ ]^{1/2}} \tag{21}$$

이다. 여기서 $\hat{v}_{tt}(\hat{\underline{p}} - \hat{\underline{f}})$은 $\widehat{Cov}(\hat{\underline{p}} - \hat{\underline{f}}) = n^{-1}S\hat{V}S'$ 의 t번째 대각원소이고,

$S = I - F(F'A^*F)^{-1}F'A^*$이며, $\hat{V} = \widehat{Cov}(\sqrt{n}\,\hat{\underline{p}})$ 이다.

로지스틱 회귀모형인 경우 행렬 S는 다음과 같다.

$$S = [\ I - \hat{D}^0X(X'D_{[\ w\hat{f}(1-\hat{f})\ ]}X)^{-1}X'D_w\ ]$$

e_t^*는 모형(2)하에서 $e_t^* \approx N(0,\ 1)$이므로 e_t^*의 절대값을 이용해서 임의의 칸에 대한 비율이 특이한 경향을 나타내고 있는 지를 탐지할 수 있다.

이제, 특이한 칸을 식별하기 위해서 표준화 잔차를 이용한 다중비교의 Bonferroni방법을 이용하자.

$$|\ e_t^*\ | \ > \ z(1 - \frac{\alpha}{2t}) \tag{22}$$

여기서 $z(\eta)$ 는 표준정규분포의 η번째 백분율이고, α는 검정력의 크기이며, t는 분할표의 칸의 수이다. 식(22)에서 e_t^*의 절대값이 $z(\eta)$보다 클 경우에 해당칸을 특이값으로 간주할 수 있다. 탐구적 자료분석(exploratory data analysis)을 이용하면 표준화 잔차를 표본으로 간주하여 사분위수(quartiles)와 사분위 간범위(inter quartile range : IQR)를 계산하여 특이값을 식별할 수 있다.

e_t^*의 가장 인접한 사분위수에서

$$e_t^* > (1.5)\,IQR \tag{23}$$

이면 해당칸을 특이값으로 간주할 수 있다.

이항표본추출의 경우에 로지스틱 회귀모형에 대해서 Pregibon[15]은 *Pearson* 검정통계량 X^2과 우도비 검정통계량 G^2의 성분을 이용한 잔차를 제시하였다. 조사설계를 고려하여 수정된 검정통계량의 성분을 이용한 잔차는 다음과 같다.

$$\tilde{X}_{B(t)} = \frac{X_{B(t)}}{(\hat{\delta}_.)^{1/2}} = (nw_t)^{1/2}(\hat{p}_t - \hat{f}_t)/\left[\ \hat{\delta}_.\,\hat{f}_t(1-\hat{f}_t)\ \right]^{1/2} \tag{24}$$

$$\tilde{G}_{B(t)} = \frac{G_{B(t)}}{(\hat{\delta}_.)^{1/2}} = \left\{2nw_t\left[\ \hat{p}_t \ln(\hat{p}_t/\hat{f}_t) + (1-\hat{p}_t)\ln\left((1-\hat{p}_t)/(1-\hat{f}_t)\right)\right]\ \right\}^{1/2} \tag{25}$$

여기서 $u\hat{\delta}_. = \sum \hat{\delta}_t = tr[\ (E_2'A^*E_2)^{-1}(E_2'A^*\hat{V}A^*E_2)]$ 이고, $X_{B(t)}$는 $X_B^2 = \sum^T X_{B(t)}^2$ 의 원소이며, $G_{B(t)}$는 $G_B^2 = \sum^T G_{B(t)}^2$ 의 원소이고, $\hat{\delta}_.$은 $\delta_.$의 조사추정값이다.

이들 성분의 값이 클 경우 모형에 의해 설명되지 않는 칸이 존재하고 있다는 것을 탐지할 수 있다.

또한, Pregibon[15]은 영향을 크게 주는 자료값을 탐지하기 위해 투사행렬 M의 대각원소 m_{tt}를 이용하였다. 이 값이 작을 경우에 설명변수에 극단점(extreme points)이 존재하고 있다는 것을 알려준다. 조사설계를 고려한 투사행렬은 다음과 같다.

$$M^* = I - \hat{H}^* = I - \hat{B}^{1/2}X(X'\hat{B}X)^{-1}X'\hat{B}^{1/2}$$

이 값을 이용하여 극단점의 존재여부를 탐지할 수 있다.

다음으로 추정된 모형의 성분에 대한 각 관찰값의 영향을 탐지하기 위해 자료값의 집합에서 t번째 자료를 제거시키는 제거방법을 이용하여 추정된 모형에 나타나는 변화를 알아 보고자 한다. 계수의 민감성을 알아보기 위해 이용되는 통계량은 다음과 같다.

$$\frac{\left[\ \hat{\theta}_j - \hat{\theta}_j(-t)\right]}{s.e.(\hat{\theta}_j)} \tag{26}$$

여기서 $\hat{\theta}_j(-t)$는 t번째 자료를 제거시킨 후 계산된 θ_j에 대한 의사최우추정값이고, $s.e.(\hat{\theta}_j)$ 은 $\hat{\theta}_j$의 표준오차이다.

식(26)은 t번째 자료에 대한 j번째 계수의 민감성을 측정하는 측도로 이용된다. 추정값 $\hat{f}_t$에 대한 t번째 자료값의 민감성을 알아보기 위한 통계량은 다음과 같다.

$$\frac{\left[\ X_B^2 - \tilde{X}_B^{\,2}(-t)\right]}{\hat{\delta}_.}$$
$$\frac{\left[\ G_B^2 - \tilde{G}_B^{\,2}(-t)\right]}{\hat{\delta}_.} \tag{27}$$

여기서 $\tilde{X}_B^{\,2}(-t)$ 는 $\hat{f} = f(\hat{\theta})$ 대신에 t번째 자료값을 제거시킨 경우 $f(\theta)$의 추정값 $\hat{f}(-t)$ 를 대입시켜 구한 X_B^2이고, $\tilde{G}_B^{\,2}(-t)$는 $\hat{f} = f(\hat{\theta})$ 대신에 t번째 자료값을 제거시킨 경우 $f(\theta)$ 의 추정값 $\hat{f}(-t)$ 를 대입시켜 구한 G_B^2이다.

적합도의 민감성을 알아보기 위한 통계량은 다음과 같다.

$$\frac{[\ X_B^2 - X_B^2(-t)]}{\hat{\delta}_\cdot} \qquad (28)$$

$$\frac{[\ G_B^2 - G_B^2(-t)]}{\hat{\delta}_\cdot}$$

여기서 $X_B^2(-t)$와 $G_B^2(-t)$는 $T-1$개의 자료에서 계산된 X_B^2과 G_B^2 이고,

$$X_B^2(-t) = n \sum_{t \neq i} \frac{[\ \hat{p}_t - \hat{f}_t(-t)\ ]^2 w_t}{\{\hat{f}_t(-t)[\ 1 - \hat{f}_t(-t)]\ \}} \quad \text{이며,}$$

$$G_{B(-t)}^2 = 2n\sum w_t \left\{ \hat{p}_t \ln\left[\frac{\hat{p}_t}{\hat{f}_t(-t)} \right] + (1-\hat{p}_t)\ln\left[\frac{(1-\hat{p}_t)}{(1-\hat{f}_t(-t)} \right] \right\} \quad \text{이다.}$$

그리고, 조사설계를 고려한 경우에 대체방법을 이용한 통계량은 다음과 같다.

$$R_t^* = \frac{(e_t^*\ \tilde{h}_{tt}^*)}{r} \qquad (29)$$

여기서 r은 $\hat{H}^*$의 계수이다.

3.2 실증자료분석

본 연구에서 전개된 이론적 결과를 실증적 자료에 적용시켜 보기로 하자. 이용된 자료는 노동부가 1990년 7월 1일에서 1990년 7월 31까지 1개월간 한국표준산업분류에 의한 농업, 수렵업, 임업 및 어업 부문을 제외한 전 산업의 상용근로자 10인이상 사업체중 층화계통 추출방법에 의해 추출된 4100개의 표본사업체를 대상으로 성별, 나이, 학력, 근속년수 등을 조사한 직종별 임금실태 조사 보고서에 수록된 전국표본조사의 자료이다. 그러나, 본 연구에서는 편의상 이 자료가 전국을 행정구역상으로 서울특별시와 5개의 광역시 그리고 7개의 도로 13개의 층으로 나누어, 표본이 층화표본추출에의해 추출되었다고 가정하였다. 표본자료는 15세에서 64세까지의 노동력이 있는 사람에 한하여 구성되었고, 로지스틱 회귀모형을 이용하기 위해서 나이와 학력수준의 2개의 설명변수를 채택하였다.

나이수준 A_k는 구간 $[\ 10+5k, 14+5k]$, $k=1,2,\cdots,10$ 으로 10개의 그룹으로 나누었고, 각 구간의 중간점 $A_k = 12+5k$ 가 해당되는 나이그룹에 있는 모든 사람에 대한 나이값이고, 학력수준 E_i는 중졸, 고졸, 전문대졸, 대졸이상으로 $E_i = 9, 12, 14, 16$ 의 값을 변수의 값으로 갖는다.

칸 비율 π_{ki}에서 변동을 설명하기 위한 로지스틱 회귀모형은 다음과 같다.

$$v_{ki} = \ln\{\pi_{ki}/(1-\pi_{ki})\} = \theta_0 + \theta_1 A_k + \theta_2 E_i \qquad (30)$$

$$k = 1, 2, \cdots, 10 \qquad i = 1, 2, 3, 4$$

모형(30)에 노동자료를 적용시킬 경우 *Newton-Raphson* 방법에 의해 모수 $\underline{\theta}$의 추정값은

$$\underline{\hat{\theta}}' = (\ -5.8343,\ 0.0883,\ 0.3281\ )$$

이다. 모형(30)의 적합도를 검정하기 위한 검정통계량은

$$X^2 = 454.2 \qquad G^2 = 433.8$$

이고, $\hat{\delta}_\cdot = 9.712$ 이다.

노동자료에 기존의 적합도 검정통계량 X^2과 G^2의 값이 자유도 $T-r=37$ 를 갖는 χ^2의 상한 5% 점 $\chi^2_{0.05}(37) = 52.2$ 보다 크기 때문에 표본조사설계가 무시된다면 모형(30)는 기각된다.

표본조사설계효과를 고려 할 경우에 $X^2/\hat{\delta}_\cdot$과 $G^2/\hat{\delta}_\cdot$의 값을 살펴보면

$$X^2/\hat{\delta}_\cdot = 46.77 \qquad G^2/\hat{\delta}_\cdot = 44.67$$

이므로 모형은 타당하다고 생각된다.

조사설계를 고려한 표준화 잔차 e_t^*의 점근적 분포가 $N(0,1)$이므로 1.96을 초과하는 $|e_t^*|$의 갯수는 17로서 대략적으로 $0.5T=20$ 과 같다. 그러나, $\hat{p}_t$의 값이 분할표의 칸에서 1이나 1에 가까운 값을 가지게 되면 표준화 잔차를 이용하는 것은 바람직하지 못하다. 10번째와 21번째, 31번째, 32번째 칸에 대한 e_t^*의 절대값이 노동자료의 다른 모든 칸에 비해 과다하게 크므로 특이한 경향이 있는 칸으로 간주할 수 있다.

이제, 다중비교의 *Bonferroni* 방법을 이용하여 특이한 경향이 있는 칸을 식별해보자. $z\left(1-\frac{\alpha}{2t}\right)$가 $z(0.99)=0.8389$ 이므로 표준화 잔차의 절대값이 이 값들을 초과하게 되면 해당 칸을 특이한 칸으로 간주할 수 있다.

또한, 표준화 잔차를 표본으로 간주하여 사분위수와 사분위간범위를 이용하여 특이한 경향이 있는 칸을 식별해 보자. $(1.5)IQR=-6.894$가 되므로 이 값보다 큰 e_t^*를 갖는 칸은 특이한 칸으로 간주할 수 있다.

추정된 모형의 성분에 대한 각 칸의 빈도수의 영향을 탐지하기 위해 t번째 자료를 제거하여 나타난 변화가 주는 계수의 민감성을 구해 보면 다음과 같다. 이 경우에 추정량 $\hat{\underline{\theta}}$의 표준오차는

$$s.e.(\hat{\underline{\theta}})=(0.1316,\ 0.0017,\ 0.0092)$$

이다. $s.e.(\hat{\theta}_0)$ 에 영향을 미치고 있는 칸은 1번째 칸과 3번째 4번째, 12번째 칸의 빈도수이며, $s.e.(\hat{\theta}_1)$ 에 영향을 미치는 칸은 3번째, 7, 8, 9번째, 그리고 12번째 칸의 빈도수이고, 3, 4번째 칸의 빈도수가 $s.e.(\hat{\theta}_2)$ 에 영향을 미치는 칸임을 알 수 있다.

그리고, 추정값에 대한 t번째 자료값의 민감성을 살펴보면 12번째 칸의 빈도수가 X_B^2과 G_B^2에 절대적인 영향을 미치고 있다는 것을 알 수 있다.

조사설계를 고려한 경우에 대체방법을 이용하면 13개의 칸에 해당되는 R_t^*의 값이 나머지 칸에 비해 크게 나타났고, 그 중에서 20번째와 21번째, 31번째 칸에 해당되는 R_t^*의 값은 10보다 크게 나타났다.

4.결 론

범주형 자료의 분석에 가장 많이 이용되고 있는 통계량은 *Pearson* X^2 검정통계량과 우도비 검정통계량 G^2이 있으며, 범주형 자료로 구성된 분할표에서 응답변수가 단 두개의값(실패=0,성공=1)을 갖는 경우, 설명변수들 사이의 연관성을 알아보기 위해 이항응답모형이 이용되고 있다.

이항응답모형하에서 자료분석에 이용되는 *Pearson* X^2 검정통계량과 우도비 검정통계량 G^2들은 표본조사설계가 이항표본추출모형이나 적이항표본추출모형을 가정하고 있다. 그러나, 실제 이용되는 대부분의 표본조사설계는 집락과 층화를 포함한 복합표본추출방법이 이용되고 있다. 따라서, 이 가정들은 복합표본추출의 경우 적절하지 않게 된다.

본 연구에서는 이항응답모형의 대표적인 모형의 하나인 로지스틱 회귀모형에서 조사설계효과를 고려한 수정된 통계량을 이용하여 칸 비율의 특이성과 제시된 모형에 영향을 미치는 자료값을 조사하기 위한 기존의 진단측도를 수정하였으며, 층화표본추출에 의해 추출된 실증적 자료를 이용하여 본 연구에서 제시한 수정된 진단측도를 적용시켜 보았다.

본 연구와 관련하여 추가적인 연구방향을 제시하면, 동적인 그래픽스를 이용하여 자료와 모형에 대한 진단방법의 연구가 필요하다.

참 고 문 헌

[1] Anderson, E. B.(1992). Diagnostics in categorical data analysis, J. Roy. Statist. Soc.(B), Vol. 54, pp.781-791.

[2] Barnett, V., and Lewis, T.(1984). *Outliers in Statistical Data,* John Wiley & Sons.

[3] Bedrick, E. J.(1983). Adjusted chi-squared tests for cross-classified tables of survey data, Biometrika, Vol. 70, pp. 591-596.

[4] Birch, M. W.(1964). A new proof of the Pearson-Fisher theorem, Ann. Math. Statist., Vol. 35, pp. 817-824.

[5] Cochran, W. G.(1977). *Sampling Techniques,* John Wiley & Sons.

[6] Cohen, J. E.(1976). The distribution of the chi-squared statistic under clustered sampling from contingency tables, J. Amer. Statist. Assoc., Vol. 71, pp. 665-670.

[7] Collet, D.(1991). *Modelling Binary Data,* Chapman and Hall, London.

[8] Cook, R. D. and Weisberg, S.(1982). *Residuals and Influence in Regression,* Chapman and Hall, London.

[9] Cox, D. R., and Snell, E. J.(1989). *Analysis of Binary Data,* Chapman and Hall, London.

[10] Gross, W. F.(1984). A Note on chi-squared tests with survey data, J.Roy. Statist. Soc.(B), Vol. 46, pp. 270-272.

[11] Kay, R., and Little, S.(1987). Transformations of the explanatory variables in the logistic regression model for binary data, Biometrika, Vol. 74, pp. 495-501.

[12] Lesaffre, E., and Albert, A.(1989). Multiple-group logistic regression diagnostics, Appl. Statist., Vol. 38, pp. 425-440.

[13] McCullagh, P., and Nelder, J. A.(1989). *Generalized Linear Models,* Chapman and Hall, London.

[14] Morel, J. G.(1989). Logistic Regression Under Complex Survey Designs, Survey Methodology, Vol. 15, pp. 203-223.

[15] Pregibon, D.(1981). Logistic regression disgnostics, Ann. Statist., Vol. 9, pp. 705-724.

[16] Rao,J.N.K.(1984). On Chi-Squared Tests for Multi-Way Tables with Cell Proportions Estimated from Survey Data, Ann. Statist., Vol. 12, pp. 46-60.

[17] ———— (1987). On Simple Adjustments to Chi-Square Tests with Sample Survey Data, Ann.Statist., Vol. 15, pp. 385-397.

[18] Williams, D. A.(1987). Generalized linear model diagnostics using the deviance and single case deletions, Appl. Statist., Vol.2, pp.181-191.