Identifying Interviewer Effect with Logistic-regression Model with Extra-binomial Variation Assumption

Allan Lau
Senior Statistician, Vocational Training Council, HKSAR China

Summary

This paper elaborates the use of logistic-regression model with extra-binomial variation assumption to identify the existence of and to assess the interviewer effect in the course of data collection.

Introduction

In Beta-binomial distribution model we postulate that P_i, the mean of y_i (the number of successes)/n_i, varies about P with variance $P(1-P)\phi$. In sample survey, it is not uncommon that interviewers are arranged to work in only one area or that interviewers of different years of experience are employed. The area served or the experience of interviewers may become the explanatory variables for the mean of P_i. To identify the significancy of these explanatory variables, we can postulate that P_i varies about some mean, say μ_i, which in turn depends on the factor or the covariate mentioned. Since μ_i falls in the range between 0 and 1, we can postulate that the logit of μ_i is related linearly with the factors or covariates of the interviewers. Symbolically,

$$\mu_i = \left(1 + \exp\left[-X_i \beta \right] \right)^{-1}$$

where X_i is the i^{th} row vector in the design matrix X for the explanatory variables,
β is the corresponding coefficient matrix.

Logistic-regression Model with Extra-binomial Variation

With this postulation, we are in effect assuming that y_i varies about mean which depends on some interviewer covariates, but with extra-binomial variation and common intra-class correlation (ϕ) at different level of the factors or at different value of the covariates. Symbolically,

$$E(y_i) = n_i \mu_i, \quad \text{and} \quad V(y_i) = n_i \mu_i (1-\mu_i)(1+(n_i-1)\phi)$$

If P_i follows a beta distribution, y_i follows a beta-binomial distribution. Define $l_i = \log[f(y_i)]$ to equal the log-likelihood function for μ_i and ϕ, the joint log-likelihood function for all μ_i and ϕ is represented by $l = \sum_i l_i$.

It can be proved that the score for βs : $S = \left(\text{K} \quad \frac{dl}{d\boldsymbol{b}_j} \quad \text{K} \right)^T$

$$\text{where } \frac{dl}{d\beta_j} = \sum_i \frac{dl_i}{d\mu_i} \frac{d\mu_i}{dx_i\beta} \frac{dx_i\beta}{d\beta_j}$$

$$= \sum_i \frac{(y_i - n_i\mu_i)}{V(y_i)} x_{ij} \left(\frac{d\mu_i}{dx_i\beta} \right)$$

$$= \sum_i \frac{(y_i - n_i\mu_i)}{1 + (n_i-1)\phi} x_{ij}$$

This is in fact true for all random variable y_i which follow a probability distribution function (p.d.f.) which is in the exponential family. In matrix form, define

$$M^* = W^{-1}M,$$

where W = diagonal matrix with element $1+(n_i-1)\phi$
M = column matrix with element $y_i - n_i\mu_i$

Then, the score : $S = X^T M^* = X^T W^{-1} M$

the information : $I = E(SS^T) = X^T W^{-1} VX$

where V = diagonal matrix with element $n_i\mu_i(1-\mu_i)$.

The maximum likelihood estimate for β_s, given ϕ, can be calculated using Fisher' s scoring method.

i.e. $\beta_m = \beta_{m-1} + I_{m-1}^{-1} S_{m-1}$ where m is the iteration number.

Model Identification Using Fisher' s Scoring Method

Fisher' s scoring method can be viewed as the weighted least square method. For example

$$\begin{aligned}\beta_m &= \beta_{m-1} + I_{m-1}^{-1} S_{m-1} \\ &= \beta_{m-1} + \left(X^T W^{-1} VX\right)_{m-1}^{-1} \left(X^T W^{-1} M\right)_{m-1} \\ &= \left(X^T W^{-1} VX\right)_{m-1}^{-1} \left(X^T W^{-1} VZ\right)_{m-1}\end{aligned}$$

where $Z = X\beta + V^{-1}M$. That is Z regresses on X and is with covariance matrix $\left(W^{-1}V\right)^{-1}$.

The goodness of fit statistics for this model is

$X^2 = \sum_i w_i (y_i - n_i\mu_i)^2 / n_i\mu_i(1-\mu_i)$ where w_i is the i^{th} element of W.

X^2 is approximately the weighted sum of squares of residuals $\left(Z - X\hat{\beta}\right)^T \left(W^{-1}V\right)\left(Z - X\hat{\beta}\right)$ and asymptotically follows a chi-square distribution with degree of freedom d.f. = k – p

where k = number of interviewers
p = number of explanatory variables

With $\phi = 0$, W is an identity matrix and the model reduces to the ordinary logistic-regression model.

$$E\left(X^2\right) = k - p + \phi \sum_i (n_i - 1)\left(1 - v_i\, q_i\right) \quad \text{---------- (a)}$$

where v_i = element of V

q_i = element of $X\left(X^T W^{-1}\, VX\right)^{-1} X^T$,

With ϕ not equal zero

$$E\left(X^2\right) = \sum_i w_i \left(1 - w_i\, v_i\, q_i\right)\left(1 + (n_i - 1)\phi\right) \quad \text{---------- (b)}$$

Using these relationship between $E(X^2)$ and ϕ, the following iterative procedure can be used to estimate ϕ and β_s:

Steps

(1) Estimate β_s assuming $\phi = 0$ and calculate X^2.
(2) Test goodness of fit by comparing X^2 with chi-square (k – p).
(3) If the model is rejected, calculate $\hat{\phi}$ using equation (a).
(4) Estimate β_s with f estimated by $\hat{\phi}$.
(5) Calculate X^2. If X^2 is close to $E(X^2)$=k–p, stop. If not, re-estimate ϕ using formula(b) and re-start step(4).

Following this iteration procedure, a computer programme written with SAS PROC IML has been developed to estimate β_s and ϕ.

References

1. S. Lynne Stokes and Joe R. Hill (1985) ' Modeling interviewer variability for dichotomous variables' . Proceeding of the Survey Research Method Section, American Statistical Association pp 344-348.
2. D. M. Smith (1983) ' Maximum likelihood estimation of the parameters of the beta-binomial distribution' Journal of the Royal Statistical Society.
3. D. A. Williams (1982) ' Extra-binomial variation in logistic linear models' Applied Statistics vol. 31, no. 2, pp 144-148.