Osval A. Montesinos López and Kent M. Eskridge

A regression group testing model for a two-stage survey under informative sampling for detecting and estimating the presence of transgenic corn

Abstract

Group testing (GT) regression methods are effective for estimating and classifying binary responses and reduce the required number of diagnostic tests. For this reason, these methods have been used for the detection of transgenic corn in Mexico. However, there is no appropriate methodology when the sampling process is complex and informative. We developed group testing regression models for the analysis of surveys conducted in two stages with unequal selection probabilities and informative sampling. A simulation study demonstrates that the proposed model considerably reduces the bias in estimation compared to other methods that ignore the weights.

Introduction

GT is a method for screening samples for an attribute when samples are grouped into pools (or batches) and each pool is tested for presence of the attribute where all samples in the pool are cleared of having the attribute if a pool tests negative.

A sampling process is informative when the sampling probabilities are related to the values of the outcome variable after conditioning on the model covariates and using standard methods gives biased estimates (Pfeffermann, 2006). One approach for dealing with informative sampling is to include design (sampling) weights to account for unequal selection probabilities. When the weights are incorporated in the likelihood function, pseudo-maximum likelihood (PML) is required.

Objective

To generalize the group testing methodology to surveys conducted in two stages with stratification and different cluster sizes when the sampling is informative.

Model and Simulation

The finite population values with dichotomous responses were generated from the two-level superpopulation model with linear predictor: $\eta_{ij} = \beta_0 + b_i$, with *i=1,2,…,M;* $b_i \sim N(0, \sigma_b^2)$, response variable $Y_{ij}|b_i \sim binary(\pi_i)$, *j=1,2,…,*$N_i$*;* and logit link: $log\left(\frac{\pi_i}{1-\pi_i}\right)$, with $\beta_0 = -4.4631$, $\sigma_b^2 = 0.9888$ as our true model parameter values. We simulated the individual responses, Y_{ij} using a Bernoulli distribution with mean $\pi_{ij} = 1/(1+\exp(-\beta_0 - b_i))$. The finite population consisted of $M = 300$ with 83 clusters belonging to stratum 1 and 217 to stratum 2.

Sampling process

	Population	Sample	Sampling process
First stage	$M = (M_1 + M_2)$ fields	$m = (m_1 + m_2)$ fields	PPS
Second stage	$N_{ih} = N_{ih1} + N_{ih2}$ plants	$n_{ih} = n_{ih1} + n_{ih2}$ plants	SRS

PPS=Probability proportional to size and SRS=simple random sampling.

Incorporating the weights in the PML

The weighted pseudo-likelihood was equal to

$$L = \prod_{h=1}^{2}\prod_{i=1}^{m_h}\left[\int_{-\infty}^{\infty}\prod_{h_*=1}^{2}\prod_{k=1}^{g_{ih}}\{L_{ij}(\boldsymbol{\theta}|b_i)\}^{w^*_{j|ihh*}}\varphi(b_i)db_i\right]^{w^*_{ih}} \quad (1)$$

where $w^*_{j|ihh*}$ is the weight at pools level, and w^*_{ih} is the weight at field level. Six methods of incorporating the weights were studied. The NLMIXED procedure of SAS (SAS Institute, 2011) was used for maximizing the expression (1).

Results

Table 1. Simulation means of the intercept ($\beta_0 = -4.4631$ true value) and the second level standard deviation ($\sigma_b = 0.9944$ true value). Cluster sample $m_i = 24$ (8 from stratum 1 and 16 from stratum 2) under PPS. Elementary units size $n_j = 100$ (50 from stratum 1 and 50 from stratum 2) under SRS. Pool size (*s*). Six hundred simulations were performed.

			Weighting method					
s	Parameter	Estimate	M1	M2	M3	M4	M5	M6
1	β_0	Mean	-3.3314	-4.3702	-4.9287	-4.4653	-4.4532	-4.3647
	σ_b	Mean	1.0261	1.0179	1.5583	0.9972	0.9852	1.0101
5	β_0	Mean	-3.3646	-4.3682	-4.9311	-4.4645	-4.4519	-4.3626
	σ_b	Mean	0.9456	0.9621	1.5347	0.9367	0.9225	0.9541
10	β_0	Mean	-3.41	-4.367	-4.936	-4.4665	-4.4533	-4.3605
	σ_b	Mean	0.8658	0.8809	1.5137	0.8561	0.838	0.8633

M1 unweighted maximum likelihood. M2 PML using raw weights at the cluster level. M3 PML using raw weights at both levels. M4 PML using raw weights at the cluster level and scaling method A at the individual level. M5 PML using raw weights at the cluster level and scaling method B. M6 PML using method D with weights at the cluster level.

Conclusions

When the sampling process is informative, weights at both levels should be included. However, we need to use scaled weights because using the raw weights produces more bias than ignoring the weights altogether.

We generalized the mixed regression GT methodology for a complex informative sampling process and we give NLMIXED or GLIMMIX code to run the analysis. This methodology can save considerable resources when estimating any binary response and can produce almost the same results than as individual testing.

References

Pfeffermann, D., Da Silva Moura, F.A., and Do Nascimento Silva, P.L. (2006). Multi-level modelling under informative sampling. *Biometrika,* 93, (4), 943-959.

SAS Institute. (2011). *SAS 9. 3 Output Delivery System: User's Guide*. SAS Institute.