Linear Regression for Survey Data Using Regression Weights

Pedro Luis do Nascimento Silva
IBGE, Departamento de Metodologia
Av. Chile, 500
Rio de Janeiro - RJ - Brazil - 20031-170
E-mail: pedrosilva@ibge.gov.br

Renata Pacheco Nogueira Duarte
IBGE, Departamento de Metodologia
Av. Chile, 500
Rio de Janeiro - RJ - Brazil - 20031-170
E-mail: rduarte@ibge.gov.br

1. Introduction

When the survey observations can be assumed independent and identically distributed (IID), regression models can be specified, fitted, tested and reformulated using standard methods and software. In practice, however, this assumption is seldom adequate, hence more complex model assumptions and/or special estimators are needed to accommodate features of the sample design or the population structure in the analysis. This paper considers fitting linear regression models to sample survey data incorporating auxiliary information via weights derived from regression-type estimators.

2. The Problem

Let the survey data for unit i contain the values of explanatory variables z_i used in a linear regression model to predict the value of a response variable y_i, as well as the weight w_i derived from a regression estimator considering auxiliary variables x_i. Our target superpopulation model $E(y_i|z_i) = \beta_0 + z_i'\beta$ does not involve the auxiliary variables in x_i. Yet they may still need to be considered when estimating its parameters because the regression weights w_i may be the only ones available to the analyst. An example is provided by the Brazilian Population Census, which uses a short form for data collection on key demographic variables, and a long form for more detailed socio-economic variables, which is applied only to a sample of households. Sample records carry regression-type weights, which were developed at the estimation stage to calibrate sample-weighted counts on known census counts (obtained from the short form), thus complicating model fitting by users. Regression weights are sometimes used also to compensate for differential non-response.

3. Proposed Solution

Nascimento Silva(1996, chapter 6) reviewed existing approaches to fit linear regression models in the presence of auxiliary population information. He proposed an adjustment to the Pseudo Maximum Likelihood (PML) procedure (Skinner, 1989; Binder, 1983) for the case when the sample weights are derived from regression-type estimators, as needed in the Brazilian Census case. The adjusted procedure (denoted PML-R) uses the regression-type weights for point estimation of the model parameters, and provides a simple correction to the linearization variance estimators of Binder(1983).

Asymptotic properties of the PML-R estimator were compared to those of some competitors (Ordinary Least Squares – OLS – and Pearson Adjusted – PA) in the case of Simple Random Sampling Without Replacement (SRS). The PML-R estimator for the slope β proved more efficient than the OLS estimator, in terms of a first order asymptotic approximation to the design variance. This result mimics that achieved with traditional regression estimation for means and totals. However, the gain in efficiency is expected to be minor if the target model offers a good description for the population data. PML-R estimators may also offer gains in asymptotic design efficiency when compared to the PA estimators in some cases. At the same time, they provide greater flexibility for allowing auxiliary variables of both continuous or binary type to be considered, do not require multivariate normality assumptions and require only that the population means (not variances) of the auxiliary variables are known.

4. Simulation Study

A simulation study was carried out to compare the numerical performance of the PML-R estimator with that of some competitors. Four sets of estimators were compared (OLS, PA, standard PML and PML-R), under two types of sample designs (SRS and stratified simple random sampling with unequal sampling fractions – STR – that over-samples large population units), and two modelling situations (model M1, which fits well to the population data, and model M2, for which the homoscedasticity assumption is violated). Estimators incorporating auxiliary information (PA and PML-R) were considered with four choices of a single auxiliary variable at a time.

5. Conclusions

Under the ignorable sampling scheme (SRS), all estimators performed very similarly in terms of relative root mean squared errors. For the reasonably specified regression model M1, results indicated that there is not much to be gained from incorporating auxiliary information via regression weighting and/or Pearson adjustment. Thus OLS estimators should be preferred in this case. PA estimators appeared not to be robust against misspecification of the model relating the survey and auxiliary variables, whereas PML-R estimators retained unbiasedness and precision, as well as reasonable conditional performance.

Under the STR design, the estimators which ignore the design and weights (OLS and PA) displayed substantial relative bias for the misspecified model M2, while the design-adjusted estimators (PML and PML-R) appeared to retain unbiasedness even in this case. In terms of precision, however, the estimators ignoring the sample design beat the PML and PML-R estimators, although performance was similar for the misspecified model M2.

When compared to each other, the performance of PML and PML-R estimators was very similar throughout, both in terms of bias and precision. This indicates that the design weights play the dominant part in compensating for the sample selection bias, which was substantial for parameters of the misspecified model M2. With model M1, a familiar trade-off between bias and precision appeared when choosing between the OLS or PA estimators, and the PML or PML-R estimators. When the model specifications are inadequate, one of the estimators incorporating the design weights (PML or PML-R) should be used, because while these estimators have the same level of precision as those ignoring the design weights (OLS and PA), the latter can be substantially biased.

REFERENCES

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279-292.

Nascimento Silva, P.L.d. (1996). Utilizing Auxiliary Information in Sample Survey Estimation and Analysis. University of Southampton, PhD dissertation. Southampton, UK.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In Skinner, C.J., Holt, D. and Smith, T.M.F., eds., Analysis of Complex Surveys, 59-87, Chichester, John Wiley & Sons.