Semi-parametric estimation of weights: Fuzzy post-stratification

Øyvind Hoveid [*]
Norwegian Agricultural Economics Research Institute
oyvind.hoveid@nilf.no

August 18, 2009

Abstract

With weights one can compute moments of a distribution of study variables over a receptor set where only auxiliary variables are known, provided study variables and auxiliaries are related exactly as in some donor set. Fuzzy post-stratification (FP) aims at consistent estimation of such weights.

A core computational trick of FP is a semi-parametric representation of the propensity to belong to the receptor set given auxiliaries. This is obtained with model-based post-stratifications and an EM-like algorithm. The fuzzyness of the resulting weights comes from the fact that the algorithm does not converge to a single vector of weights when the receptor set is finite. Rather a set of weights are plausible, and their mean is the fuzzy weights.

FP can be combined with generalized regression calibration. The technique is demonstrated for a test set from the Norwegian farm account survey.

[*] Thanks to Erwin Kalvelagen, GAMS Development Corp. for programming problem solving, to Leiv Solheim, Statistics Norway, and Seppo Laaksonen, University of Helsinki for comments, and to Knut Skarsem, NILF for preparing data.

1 Donors, receptors and weights

Several tasks within survey sampling has a common structure: Study variables and auxiliary variables are observed on a sample of units (the donor set), while auxiliary variables only are observed on an additional set of units (the receptor set).[1] Population and small-area estimation, and correction for unit and item non-response are all tasks of this structure with different receptor and donor sets. An obvious question in this context is then: *what are the moments of study variables over the receptor set given that the relationship between the two groups of variables are the same as in the donor set?* As survey statisticians usually ask only about the means of study variables over receptor units, the level of ambition is somewhat increased.

In the standard case: when the auxiliary variables represent binary indicators of memberships in a collection of strata, the answer to this general question is relatively simple and conforms with standard survey methods. More precisely, it can then be shown that three well-known methods — post-stratification, regression calibration and OLS-based generalized regression calibration — all provide identical weights on the donor units with which moments of the unknown distribution of receptor units can calculated.

With more or less continuous and dependent auxiliaries, the three methods provide different weights. How should such information then be exploited in the calculation of weights? Since the ambition is to estimate probability weights, I have chosen to transform general auxiliaries into indicators of post-strata membership using *models*.

2 An EM-like algorithm

A rule of thumb for stratification is that the variance of study variables should be relatively small within strata. With this motivation I construct *model-based nearest neighbor post-strata* around each distinct auxiliary observation in the donor set. The mentioned model is estimated with information from the donor set. A stratum consists of those points which has their predictions closest to the prediction at a certain donor point. Given this stratification, the elements in the donor and receptor sets can be counted. The model contingent weight within a stratum is simply the receptor count divided by the donor count. Receptor count divided by donor plus receptor count is the propensity of being a receptor within the model contingent stratum.

Model dependence is considered a nuisance in survey statistics. The model should satisfy some optimality criterion, and uncertainty of the model

[1] The terms ”donor” and ”receptor” are taken from Zhang and Nordbotten (2008).

need to be accounted for. These challenges will be faced here.

Clearly, the prediction model to be applied for nearest neighbor stratification should be just as reliable over the receptor set as over the donor set. Intuitively, this means that estimation should take place over the joint donor-receptor set. Only observations form the donor set is available, but the receptor set can be incorporated by means of weights on the donor observations. The weights to be applied in estimation is thus donor plus receptor count divided by donor count, ie. the inverse of the propensity to belong to the donor set. Moreover, it can be shown from insights of information theory that estimation should be conducted with maximum likelihood or equivalent methods.

The situation is now that an optimal model contingent on weights, and weights contingent on a model, both can be found. This resembles an EM-algorithm with the estimation of the model using maximum likelihood and preliminary weights is the M-step, and with the calculation of weights using the prediction model is the E-step. Mathematically, this is an issue of existence of a fixed point of weights in the mapping from preliminary weights through the weight dependent model to model dependent weights. Basically, such a fixed point is not likely to exist when the receptor set is finite. A crucial assumption of continuity of the involved mapping is not satisfied, an the EM-algorithm cannot be expected to converge. Nevertheless, it will end up in some repeating cycle, and the set of weights of the cycle are all equally plausible model-based weights. Their arithmetic mean — which depends on a set of models/stratifications — is the fuzzy weight. A convenient aspect of these weights is that the uncertainty inherited from models is considerably smaller when relying on a set of models than relying on a single one.

3 Singular value decomposition regression

When surveys has hundreds of study variables and auxiliary variables, it is convenient to treat them as two blocks of variables related in a single model so that a single vector of weights can be estimated. GLS is such a method. However, general auxiliary variables tend to be close to singularity. In small samples GLS is then known to provide high-variance parameter estimates, and some alternative method is desirable. Methods much used in the calibration literature are partial least squares (PLS) (Wold 1982) and ridge regression (Hoerl and Kennard 1970). Both reduce variance by accepting some bias, but their asymptotic properties are unclear. A promising recent method with GLS asymptotics is lasso regression (Tibshirani 1996), but no multivariate version of this seems yet developed.

To come by the problems of GLS without sacrificing the convenient asymptotic properties of that method, a simple modeling strategy *singular value decomposition regression* (SVDR) is applied. SVDR starts out (like PLS) from the singular value decomposition of the weighted covariance matrix of study variables and auxiliaries. This decomposition contains pairwise correlated components of study variables and auxiliaries, and some additional components of study or auxiliary variables. When all components are employed we are effectively doing GLS. Simpler models with less variance utilize only a few of the most correlated pairs of components. The choice of the number of components can be taken with respect to a model selection criterion which combines model fit with a penalty for the degrees of freedom of the model. Zou, Hastie, and Tibshirani (2007) recommend Schwarz' information criterion (BIC) for this purpose. Since the penalty depends inversely on the number of observations, it will vanish when the number of observations grows towards infinity, and the asymptotic properties of SVDR is that of GLS.

With SVDR as the method of estimation, it is ensured that estimated prediction models are consistent, and so are the estimated weights provided the assumption of identical stochastic relationship between study variables and auxiliaries in the donor and receptor set holds.

4 Combination with generalized regression weighting

Continuous auxiliary variables may contain more information than fuzzy post-stratification takes care of. More precisely, the donor units within a post-stratum may not be representative for the receptors. This will in particular be the case in small samples. Fuzzy post-stratified moments may then be biased, and GLS-based GREG-adjustment of the weights can be conducted. Moment constraints are then conveniently specified in terms of the components of the SVDR-regression. The selection among SVDR models will then ensure that only the most relevant calibration constraints are imposed.

The GREG-adjustment should be conducted for each model of the EM-like algorithm. The adjusted fuzzy weights will then be the arithmetic mean of adjusted weights over the same set of weights which constitute the fuzzy weights.

5 A case study

The empirical background for this paper is a business survey of Norwegian farmers ("Driftsgranskinger for jordbruket") (NILF 2007). This survey is hampered by poor design and heavy non-response. However, quite a lot of information on the non-surveyed farmers exist, and it is of importance to utilize that information in the best possible way to produce weights to make a representation of the population.

Farm accounting surveys has traditionally been stratified according to region, farm type and farm size, for design weights (FADN 2007) and post-stratified weights (Meier 2000). The results with FP are contrasted with those of such traditional post-stratification.

We will not bother the reader with a lot of study variables, and will only state results in terms of a single key variable, *net income of agriculture*, and its weighted means. All results are obtained with 50 iterations for a test subset of the survey of 137 observations. The number of receptors are 5107. Except for one model the algorithm did not converge to a repeating cycle, but since weight estimates are consistent anyway, the estimates can be applied as soon as they are relatively stable. Reported models are from the sequence which comes closest to a repeating cycle.

Variance is distinguished according to source. One part is caused by sampling. This is found by calculating weights over various resamples during the EM-algorithm. Another part is caused by modeling, and is found from the various models involved. Coefficient of bias is calculated as the bias correction relative to the un-weighted model divided by the bias corrected mean. The coefficient of ΔMSE is calculated as the square root of the difference of MSE for the corrected (assumed unbiased) and un-weighted models divided by the bias corrected mean.

Results are stated for several alternative versions of the fuzzy post-stratification. First, in table 5 there are four different model types in the EM-algorithm: "SVDR n" means SVDR-model with *n* pairs of factors. "GLS" means GLS-model of all explaining and explained factors. Moreover, there are two different frames. "Standard" refers to indicators of groups according to region, farm type and farm size. "Extended" refers to a large set of continuous variables of farm size with respect to different crops and livestock, climate variables, farmer age and household composition, and polynomials thereof, in addition to indicators above.

The coefficients of variance can be compared to that of the un-weighted mean, 0.0529. All versions of fuzzy weights have somewhat larger variance, but contrasts are not great. Contrasts are more pronounced with respect to bias-correction. Both the most simple and most complex models have smaller

Table 1: Summary statistics and coefficients of variance, bias and difference of mean squared errors of estimated population means using fuzzy weights. Extended and standard frames. Various model types

Frame	Extended				Standard
Model selection	SVDR 1	SVDR 2	SVDR 3	GLS	SVDR 3
Number of models	47	24	41	25	4
Mean R^2	0.1167	0.2210	0.2472	0.6902	0.1218
BIC	2.9655	2.6611	2.6002	24.5066	2.0980
Sampling CV	0.0782	0.0800	0.0682	0.0461	0.0681
Modeling CV	0.0096	0.0033	0.0041	0.0039	0.0135
Coefficient of bias	0.1314	0.2408	0.2351	0.1978	0.1303
Coefficient of ΔMSE	0.1251	0.2391	0.2367	0.2066	0.1311

bias correction than those of intermediate complexity. An overall impression is that fuzzy weights are not very sensitive to model selection criteria. Even GLS model which according to its BIC value is largely over-fitted, has sensible results. Among the presented extended models the (SVDR 3) has the best BIC-score.

The extra uncertainty introduced with the use of prediction models in FP seems almost negligible. One should note, however, that the uncertainty by relying on a single model is $\sqrt{\#\text{ models}}$ times the modeling uncertainty here. For the extended models this is not negligible. Hence, fuzzyness pays by reducing variance.

The contrast between the standard and the extended frame is very clear with respect to bias-correction. Stratification with respect to region, farm type and farm size seems insufficient compared to our fuzzy stratification based on three components. The single variable which seems most important in this respect is the age of the farmer.

In table 5 results are presented for some GREG-adjusted fuzzy weights.

The GREG-adjustments of fuzzy weights reported here shows some of the characteristics of this method. For the SVDR-models, the adjustment means virtually nothing as the fuzzy weights did a good job in calibrating the survey sample. Only the GLS-model is of interest as the bias correction turns out incredibly large. The over-fitted GLS model is not at all suitable as a foundation for GREG-adjustment of weights. Variances are also large. Anyway, a clear message is that GREG-adjustment is sensitive to model type.

Table 2: Coefficients of variance, bias and difference of mean squared errors of estimated population means using fuzzy weights adjusted with generalized regression calibration. Extended and standard frames

Frame	Extended		Standard
Model type	SVDR 3	GLS	SVDR 3
Sampling CV	0.0652	0.1830	0.0685
Modeling CV	0.0044	0.0232	0.0099
Coefficient of bias	0.2268	2.7539	0.1292
Coefficient of ΔMSE	0.2287	2.7579	0.1301

6 Conclusions

This article gives a statistical treatment of the problem of inference from a finite donor set of study variables and auxiliaries to a finite receptor set of auxiliaries only. An essential assumption in this respect is that the statistical relationship between study and auxiliary variables are identical in the two sets. This is basically a non-testable assumption. Only when additional variables are introduced, can one reject the assumption for the previous set of auxiliaries.

The model dependence of the fuzzy weights does not seem to be a problem, when one both accounts for the induced variance, controls for it by tacit model selection, and at last take average over models. Since, fuzzy weighting is a generalization of standard weighting over a predefined stratification, one might say that standard methods are not model free. There is always some model involved, but with standard weighting the available data — the stratum membership indicators — does not allow any questioning of it. With fuzzy post-stratification we have moved beyond the predefined stratification and let the data speak.

The test calculations suggest that fuzzy post-stratification is a productive way of dealing with general donor/receptor problems. When required, the method can be combined with GLS-based GREG-adjustment of the weights to eliminate bias. Model selection is a sensitive issue for these adjustment, and the selection procedures developed for the fuzzy weights seem even more important in this context.

References

FADN (2007): “The Farm Accountancy Data Network,” .

Hoerl, A., and R. Kennard (1970): “Ridge regression: Biased estimation for non-orthogonal problems,” *Technometrics*, pp. 55–67.

Meier, B. (2000): “A new sample, farm typology and weighting system for the Swiss Farm Accountancy Data Network (FADN),” in *Pacioli 8. Innovations in the FADN*, ed. by K. J. P. George Beers, and A. Leuftink, pp. 14–20.

NILF (2007): “Account results in agriculture and forestry,” http://www.nilf.no/Publikasjoner/Driftsgranskinger/Bm/2007/Publikasjon2007.pdf.

Tibshirani, R. (1996): “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Wold, H. (1982): “Soft modeling: the basic design and some extensions,” in *Systems Under Indirect Observation*, vol. 2, pp. 1–53. North Holland.

Zhang, L.-C., and S. Nordbotten (2008): “Prediction and imputation in ISEE: Tools for more efficient use of combined data sources,” Discussion paper, UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE. Work Session on Statistical Data Editing (Vienna, Austria, 21-23 April 2008).

Zou, H., T. Hastie, and R. Tibshirani (2007): “On the ”Degrees of Freedom” of the Lasso,” *Annals of Statistics*, 35(5), 2173–2192.