Testing for Informativeness in Analytic Inference from Complex Surveys

F. Jay Breidt[1,3], Jean D. Opsomer[1], Wade Herndon[1], Ricard Cao[2],
Mario Francisco-Fernandez[2]
[1] Colorado State University, Fort Collins, CO, USA
[2]Universidade da Coruña, A Coruña, Spain
[3] Corresponding author: F. Jay Breidt, email: jbreidt@stat.colostate.edu

Abstract

We discuss tests for informativeness of the design in analytic inference using data from a complex survey. Design informativeness occurs if a model correctly specified for the population does not hold in the sample. We generalize existing methods through a likelihood ratio test that compares the design-based fit of the expanded model to the model-based fit. We derive the asymptotic distribution of the test statistic, which is a linear combination of independent chi-square random variables. The coefficients in the linear combination are eigenvalues of a matrix that can be consistently estimated from the data. We also consider a bootstrap version, and evaluate the tests via simulation and application to real data. Empirical results show that the new test complements existing methodology, providing good power against interesting alternatives.
Key words: bootstrap, likelihood ratio test, superpopulation, weighted estimation.

1 Introduction

Surveys are an important source of data in a wide range of disciplines. A survey is typically designed to estimate characteristics of the particular finite population from which the sample is drawn. This context is referred to as *descriptive inference* for surveys. For large-scale surveys, a combination of statistical efficiency and cost considerations often results in a complex sampling design that includes unequal inclusion probabilities, stratification and clustering. An extensive literature exists on how to incorporate these design complexities into appropriate descriptive inference methods. So-called *design-based* methods are the standard approach to construct estimates and do inference in this context It is also common for analysts to use survey data to answer scientific questions that are applicable more widely than for one particular finite population. In such situations, the questions concern characteristics of a statistical model describing relationships among variables, and the finite population is viewed as representing a realization from that model. This is referred to as *analytic inference* for surveys. Statisticians have long been aware of the fact that it is not appropriate to ignore survey considerations when doing analytic inference for survey data. Both design-based and model-based methods can be applied in this context, and there is currently still some disagreement as to which of these approaches is most appropriate. See Little (2004) and Pfeffermann (2011) for recent discussions of this topic.

We propose a new likelihood-based method for testing the hypothesis of design informativeness. Testing for design informativeness is a crucial component in choosing a suitable approach for performing analytic inference. In the analytic inference context, design informativeness can be described

succinctly as the fact that due to the design complexities, the postulated model is not correct for the sample data. If the design can be determined to be non-informative with respect to a particular postulated model, then it is reasonable to ignore the design in subsequent model fitting and analysis. On the other hand, if informativeness cannot be rejected, the analysis will need to explicitly account for the design complexities, which can be done either by staying within a design-based framework or by adjusting the model to incorporate design effects.

A number of authors have previously considered testing for informativeness, but overall, the literature on this existing topic is quite sparse. An important class of tests is based on assessing the significance of the difference between weighted and unweighted estimates of model parameters. This idea forms the basis of the procedures proposed by DuMouchel and Duncan (1983) and Fuller (1984) for the coefficients in linear regression. Pfeffermann (1993) extended this to general likelihood-based problems with explicit estimators, and Pfeffermann and Sverchkov (2003) to estimators that are defined as the solutions to estimating equations. The procedure we will propose is most closely related to this type of tests but is connected more directly with the model likelihood.

When the postulated model is a linear regression model, the test based on the difference between weighted and unweighted estimated coefficients is equivalent to an F-test for the significance of the parameters of an *extended* linear model, with the extension composed of the interactions between the covariates of the original model and the weights. See Fuller (2009, Section 6.3.1) for a derivation of this equivalence. Testing based on comparing the postulated model with an extended version of the model was also used by Nordberg (1989) for logistic regression, even though it is not equivalent to testing whether the difference between the weighted and unweighted estimates is significant in this case.

Another class of tests targets the moments of the postulated model rather than the model parameters, and tries to evaluate whether they are equal to the moments of the model that holds for the sample data. This is generally done in the regression context, so that the relevant moments are conditional on model covariates. Pfeffermann and Sverchkov (1999) show that the hypothesis of equal conditional moments for both models is equivalent to lack of correlation between the model errors and the sampling weights, and use classical correlation test statistics to test this hypothesis. This testing procedure is easy to apply but is not exact, in the sense that it is generally not clear how many moments should be compared. Pfeffermann and Sverchkov (1999) noted that "in practice, it would normally suffice to test the first 2-3 correlations." A more serious problem is the difficulty of having to interpret multiple tests simultaneously, so that the overall confidence level of the procedure is typically unknown.

A final class of tests are based on an identity in Pfeffermann and Sverchkov (1999), which shows that the difference between postulated model and the sample model can be assessed through a regression of the survey weights on the model variables. Eideh and Nathan (2006) use a Kullback-Leibler information statistics to perform this test, but the approach is more generally applicable, as noted by Pfeffermann and Sverchkov (2007). This class of tests targets the informativeness directly, but requires that a model relating the weights and the model variables be defined, so that it is subject to its own possible model specification bias.

2 Proposed Testing Procedure

We consider here the regression context. A finite population U of size N is sampled and the observed data are $\{\boldsymbol{x}_k^T, y_k\}$ for $k \in s$. The target of inference is the conditional distribution y_k given $\boldsymbol{x}_k$ in the model that generated the population values, $y_k \sim f(\cdot \mid \boldsymbol{x}_k; \boldsymbol{\theta})$. The model is specified up to a finite set of parameters denoted by $\boldsymbol{\theta}$, and we write $\boldsymbol{\theta}_0$ for the target values. The sampling design used

to select s is not known to the analyst, but sampling weights w_k are assumed available. We let I_k denote the sample membership indicator, i.e. $I_k = 1$ if $k \in s$ and 0 otherwise. In what follows, we will be considering both weighted and unweighted estimates of the model parameters, and a is used as generic notation for either the unweighted case, with $a = 1$ denoting $\{a_k\} \equiv 1$, or the weighted case, with $a = w$ denoting $\{a_k\} = \{w_k\}$. The (weighted or unweighted) estimators $\hat{\boldsymbol{\theta}}_a$ maximize the log-likelihood

$$l_a(\boldsymbol{\theta}) = \sum_{k \in U} a_k I_k \ln f\left(y_k \mid \boldsymbol{x}_k; \boldsymbol{\theta}\right).$$

The following theorem gives the asymptotic distributions of the weighted and unweighted likelihood ratio statistics:

Theorem 1 *Under suitable conditions, the likelihood ratio statistics satisfy*

$$T_1 = 2\left\{l_1\left(\widehat{\boldsymbol{\theta}}_1\right) - l_1\left(\widehat{\boldsymbol{\theta}}_w\right)\right\} = N\left(\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_w\right)^T \boldsymbol{J}_1 \left(\widehat{\boldsymbol{\theta}}_1 - \widehat{\boldsymbol{\theta}}_w\right) + o_p(1)$$

and

$$T_w = 2\left\{l_w\left(\widehat{\boldsymbol{\theta}}_w\right) - l_w\left(\widehat{\boldsymbol{\theta}}_1\right)\right\} = N\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right)^T \boldsymbol{J}_w \left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right) + o_p(1),$$

where

$$\boldsymbol{J}_a = \lim_{N \to \infty} \frac{1}{N} \sum_{k \in U} a_k \, I_k \, \boldsymbol{\mathcal{I}}(\boldsymbol{x}_k; \boldsymbol{\theta}_0)$$

and $\boldsymbol{\mathcal{I}}(\boldsymbol{x}_k; \boldsymbol{\theta}_0)$ is the Fisher Information for the kth observation. In addition,

$$N^{1/2}\left(\widehat{\boldsymbol{\theta}}_w - \widehat{\boldsymbol{\theta}}_1\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, -\boldsymbol{J}_1^{-1} + \boldsymbol{J}_w^{-1} \boldsymbol{K}_w \boldsymbol{J}_w^{-1}\right)$$

where

$$\boldsymbol{K}_a = \lim_{N \to \infty} \frac{1}{N} \sum_{k \in U} a_k^2 \, I_k \, \boldsymbol{\mathcal{I}}(\boldsymbol{x}_k; \boldsymbol{\theta}_0).$$

Hence,

$$T_a \xrightarrow{\mathcal{L}} \sum_{j=1}^{p} \lambda_{aj} Z_j^2$$

where $\boldsymbol{\lambda}_a$ is the vector of eigenvalues of $(-\boldsymbol{J}_1^{-1} + \boldsymbol{J}_w^{-1} \boldsymbol{K}_w \boldsymbol{J}_w^{-1})^{T/2} \, \boldsymbol{J}_a \, (-\boldsymbol{J}_1^{-1} + \boldsymbol{J}_w^{-1} \boldsymbol{K}_w \boldsymbol{J}_w^{-1})^{1/2}$ and $\{Z_j\}_{j=1}^p$ are independent and identically distributed $\mathcal{N}(0,1)$.

The asymptotic distribution of T_a contains unknown quantities $\boldsymbol{J}_a$ and $\boldsymbol{K}_a$, but these can be consistently estimated using plug-in methods. Alternatively, the distribution of T_a may be approximated via parametric bootstrap, which does not require estimation of these matrices. Our parametric bootstrap consists of sampling $\{y_k^*\}_{k \in U}$ as independent random variables with $y_k^* \sim f(\cdot \mid \boldsymbol{x}_k; \widehat{\boldsymbol{\theta}}_w)$ or $y_k^* \sim f(\cdot \mid \boldsymbol{x}_k; \widehat{\boldsymbol{\theta}}_1)$; both are possible because the bootstrap distribution of interest is computed under the null hypothesis.

This proposed procedure is very general and is applicable whenever a likelihood can be written down. We study several special cases, including linear regression, in which the expression simplify considerably. We also compare the procedure with other approaches available in the literature, including those mentioned in the previous section, theoretically and in simulations.

3 Application

In this section the likelihood ratio test for design informativeness will be applied to fitting a Gamma mixture model for biomass of hauls of American Plaice fish in the southern Gulf of St. Lawrence. This application makes it possible to demonstrate the flexibility of the approach in a non-standard setting. The dataset consists of 189 hauls collected in 2008, and in addition to biomass, it records date/time information, haul location, and stratum identifiers. The total number of trawlable units is known within each stratum and so sample weights are taken to be the total number of units in a stratum divided by the number of units observed from that stratum. This yields highly variable sampling rates between strata, and potential design informativeness.

For $y = Biomass$, consider the mixture model

$$y_k = \{z_k \times 0\}\{(1 - z_k) \times x_k\},$$

where $z_k \sim \text{Bernoulli}(\delta)$, and $x_k \sim \text{Gamma}(\alpha, \tau)$. Biomass is represented by a random variable that takes a value of 0 with probability δ and is positive following a Gamma distribution with probability $(1 - \delta)$. The probability density function for y_k is

$$f(y_k; \delta, \alpha, \tau) = \delta^{z_k} \left\{ (1-\delta) \frac{y_k^{\alpha-1} e^{-y_k/\tau}}{\tau^\alpha \Gamma(\alpha)} \right\}^{1-z_k}.$$

To obtain maximum likelihood estimates, $\hat{\boldsymbol{\theta}}_a = (\hat{\delta}_a, \hat{\alpha}_a, \hat{\tau}_a)$, for $a = 1$ or $a = w$, maximize the sample-level log-likelihood function

$$\begin{aligned} l_a(\delta, \alpha, \tau) &= \ln \delta \sum_{k \in s} a_k z_k + \{\ln(1-\delta) - \alpha \ln \tau - \ln \Gamma(\alpha)\} \left\{ \sum_{k \in s} a_k (1 - z_k) \right\} \\ &\quad + (\alpha - 1) \sum_{k \in s} a_k (1 - z_k) \ln y_k - \frac{1}{\tau} \sum_{k \in s} a_k y_k (1 - z_k) \end{aligned}$$

with respect to δ, α, and τ.

Figure 1 shows that the weighted data do in fact differ slightly from the unweighted data. The weighted and unweighted parameter estimates are $(\hat{\delta}_w, \hat{\alpha}_w, \hat{\tau}_w) = (0.070, 0.514, 34.342)$ and $(\hat{\delta}_1, \hat{\alpha}_1, \hat{\tau}_1) = (0.0794, 0.497, 32.465)$, respectively. It is not immediately clear whether this difference is statistically significant, so we apply the likelihood ratio test.

Following the procedure described in the previous section, we compute $T_1 = 2(l_1(\hat{\delta}_1, \hat{\alpha}_1, \hat{\tau}_1) - l_1(\hat{\delta}_w, \hat{\alpha}_w, \hat{\tau}_w)) = 1.11$. In order to obtain the approximate distribution of T_1 under the null hypothesis of no design informativeness, we derive expressions for $\boldsymbol{J}_a$ and $\boldsymbol{K}_a$ and estimate those from the data. Finally, we obtain the weights $\boldsymbol{\lambda}_a = (0.0554, 0.0554, 0.0554)$, and the 0.05 critical value for the mixture of χ^2_1 with these weights is 0.433. Recalling that the value of the test statistic was $T_1 = 1.11$, we strongly reject the null hypothesis of non-informative selection.

References

DuMouchel, W. H. and G. J. Duncan (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association 78*, 535–543.

Eideh, A. H. and G. Nathan (2006). Fitting time series models for longitudinal survey data under informative sampling. *Journal of Statistical Planning and Inference 136*, 3052–3069.

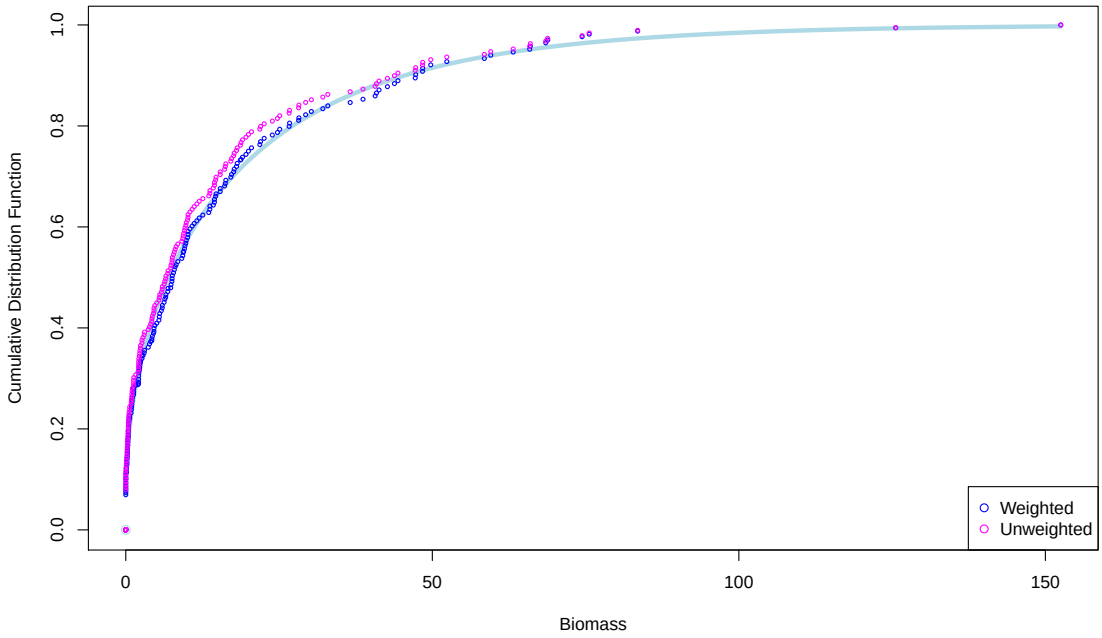

Figure 1: Weighted (lower dots) vs. unweighted (higher dots) empirical distribution functions for biomass of American Plaice, with weighted mixture model fit (solid curve).

Fuller, W. A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology 10*, 97–118.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: Wiley & Sons.

Little, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association 99*(466), 546–556.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics 5*, 223–239.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review 61*, 317–337.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? why is it a problem? how can we approach it? *Survey Methodology 37*, 115–136.

Pfeffermann, D. and M. Sverchkov (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā, Series B 61*, 166–186.

Pfeffermann, D. and M. Sverchkov (2003). Fitting generalized linear models under informative sampling. In R. L. Chambers and C. J. Skinner (Eds.), *Analysis of Survey Data*, Chapter 12, pp. 175–195. New York: Wiley.

Pfeffermann, D. and M. Sverchkov (2007). Small area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association 102*, 1427–1439.