A practical bootstrap method for testing hypotheses from survey data

by Jean-François Beaumont and Cynthia Bocci

June 2009
A practical bootstrap method for testing hypotheses from survey data

Jean-François Beaumont and Cynthia Bocci

Abstract

The bootstrap technique is becoming more and more popular in sample surveys conducted by national statistical agencies. In most of its implementations, several sets of bootstrap weights accompany the survey microdata file given to analysts. So far, the use of the technique in practice seems to have been mostly limited to variance estimation problems. In this paper, we propose a bootstrap methodology for testing hypotheses about a vector of unknown model parameters when the sample has been drawn from a finite population. The probability sampling design used to select the sample may be informative or not. Our method uses model-based test statistics that incorporate the survey weights. Such statistics are usually easily obtained using classical software packages. We approximate the distribution under the null hypothesis of these weighted model-based statistics by using bootstrap weights. An advantage of our bootstrap method over existing methods of hypothesis testing with survey data is that, once sets of bootstrap weights are provided to analysts, it is very easy to apply even when no specialized software dealing with complex surveys is available. Also, our simulation results suggest that, overall, it performs similarly to the Rao-Scott procedure and better than the Wald and Bonferroni procedures when testing hypotheses about a vector of linear regression model parameters.

Key Words: Bootstrap weights; Analysis of survey data; Hypothesis testing; Informative sampling; Linear regression; Model parameters.

1. Introduction

The bootstrap technique is becoming more and more popular in sample surveys conducted by national statistical agencies. The main reasons seem to be that it can easily deal with several situations that would be difficult to handle otherwise (e.g., nonresponse weight adjustment, calibration, non-smooth statistics, etc.) and that it is convenient for analysts. In most of its implementations, several sets of bootstrap weights accompany the survey microdata file given to analysts; no other design information is provided. These weights are usually obtained by assuming that the first-stage sampling fractions are small enough that a without-replacement sampling design can be accurately approximated by a with-replacement sampling design. The reader is referred to Rao, Wu and Yue (1992) for a succinct but clear description of a method to construct bootstrap weights under this assumption when a stratified multistage sampling design has been used.

So far, the use of the technique in practice seems to have been mostly limited to variance estimation problems (e.g., Langlet, Faucher and Lesage 2003; Yoo, Mantel, and Liu 1999; and Hughes and Brodsky 1994). On the research side, efforts have been mainly oriented towards finding an appropriate bootstrap methodology for variance estimation when the sample is drawn without replacement from a finite population (see Sitter 1992; or Shao and Tu 1995, Chapter 6, for a review of methods). Some authors have also studied the problem of determining bootstrap confidence intervals for a finite population parameter (e.g., Rao and Wu 1988; Korn, Rao and Wu 1988; Sitter 1992; and Rao et al. 1992). To our knowledge, there does not seem to be any literature on hypothesis testing using the bootstrap technique in survey sampling although this problem has been studied in the context of classical statistics. The reader is referred to Hall and Wilson (1991) for a discussion on bootstrap tests of hypotheses and to Efron and Tibshirani (1993) for an excellent account of the bootstrap technique in classical statistics. It is worth noting the work of Graubard, Korn and Midhune (1997) who applied the classical parametric bootstrap method to survey data in order to test the fit of a logistic regression model. Their procedure is valid when sampling is not informative.

The problem of hypothesis testing from complex survey data has been well studied in the last 30 years (e.g., Rao and Scott 1981; Fay 1985; Thomas and Rao 1987; Korn and Graubard 1990; Korn and Graubard 1991; Graubard and Korn 1993; Thomas, Singh and Roberts 1996; and Rao and Thomas 2003). However, except perhaps for estimating unknown variances/covariances involved in these methods, the bootstrap technique has apparently not yet been considered for testing hypotheses. The goal of this paper is thus to propose a bootstrap methodology for testing hypotheses about a vector of unknown model parameters when the sample has been drawn from a finite population. The probability sampling design used to select the sample may be informative or not. Informally speaking, sampling is informative when the model that holds for the selected...
sample is different from the model that holds for the whole population; otherwise sampling is not informative.

Our method uses model-based test statistics that incorporate the survey weights. Such statistics are usually easily obtained using classical software packages. We approximate the distribution under the null hypothesis of these weighted model-based statistics by using bootstrap weights. An advantage of our bootstrap method over existing methods of hypothesis testing with survey data is that, once sets of bootstrap weights are provided to analysts, it is very easy to apply even when no specialized software dealing with complex surveys is available.

We introduce notation and the problem in section 2. In section 3, we describe and justify our proposed bootstrap methodology for testing hypotheses with survey data. A linear regression example is given in section 4 to illustrate the theory. We briefly describe the alternative Rao-Scott (Rao and Scott 1981), Wald and Bonferroni procedures in section 5 when testing hypotheses about a vector of linear regression model parameters. They are evaluated in section 6 and compared to our proposed bootstrap procedure through a simulation study. Finally, we conclude in the last section with a short summary and discussion.

2. Preliminaries

We assume that a finite population $U$ of size $N$ has been generated according to a model, specified by the analyst, that describes the conditional distribution $F(y_U | X_U; \beta, \theta)$. The $N$-vector $y_U$ contains the population values of a dependent variable $y$, $X_U$ is an $N$-row matrix that contains the population values of a vector of independent variables $x$, $\beta$ is an $r$-vector of unknown model parameters and $\theta$ is a potential vector of additional unknown model parameters. We are interested in testing hypotheses about $\beta$ but not $\theta$. We also assume that, if the entire population $U$ could be observed, a test statistic $t(U; e)$ would be used to test the multiple linear hypothesis $H_0: H\beta = e$ against the alternative hypothesis $H_1: H\beta \neq e$. The $Q \times r$ matrix $H$ is used to define the hypothesis to be tested and $e$ is a $Q$-vector of constants specified by the analyst. Ideally, $t(U; e)$ is asymptotically pivotal; i.e., it has an asymptotic distribution that does not depend on any unknown parameter. We consider statistics that have the following quadratic form:

$$t(U; e) = (H\hat{\beta}_U - e)' (A(U))^{-1} (H\hat{\beta}_U - e),$$

where $\hat{\beta}_U$ is a consistent estimator of $\beta$ under the model and $A(U)$ is some scaling matrix. Typically, $A(U)$ is symmetric and positive definite.

As an illustrative example, let us assume that $y_k$, for all population units $k \in U$, are independently and identically distributed random variables with mean $\beta$ and variance $\theta$ and that we are interested in testing the null hypothesis $H_0: \beta = c$. In this example, $Q = 1$, $r = 1$, $H = 1$ and $X_U = 1_U$, where $1_U$ is a population vector of one’s. A common test statistic for this problem is

$$t(U; e) = \frac{(\hat{\beta}_U - c)^2}{\hat{\theta}_U / N},$$

where $\hat{\beta}_U = \sum_{k \in U} y_k / N$ and $\hat{\theta}_U = \sum_{k \in U} (y_k - \hat{\beta}_U)^2 / (N - 1)$. The statistic (2.2) has the same form as (2.1) if we let $A(U) = \hat{\theta}_U / N$. This statistic is usually assumed to follow the distribution $\chi^2_1$ or $F_{1, N-1}$ under the null hypothesis.

As is typically the case, a random sample $s$ of size $n$ is selected from the finite population $U$ according to a given probability sampling design $p(s)$. Since the dependent variable $y$ and, possibly, the independent variables $x$ are not observed for nonsample units, we may want to use the statistic $t(s; e)$ instead of $t(U; e)$. In the above example, this would lead to $t(s; e) = n(\hat{\beta}_s - c)^2 / \hat{\theta}_s$, where $\hat{\beta}_s = \sum_{k \in s} y_k / n$ and $\hat{\theta}_s = \sum_{k \in s} (y_k - \hat{\beta}_s)^2 / (n - 1)$. However, if sampling is informative with respect to the model, it may be more appropriate and is undoubtedly more common to use a weighted test statistic of the form

$$\hat{t}(s, w_s; e) = (H\hat{\beta}_{ws} - e)' (\hat{\Delta}(s, w_s))^{-1} (H\hat{\beta}_{ws} - e).$$

The $n$-vector $w_s$ contains the survey weight of sample unit $k$ in its $k$th element, denoted by $w_{ks}$. $\hat{\beta}_{ws}$ is a weighted estimator for $\beta$ and $\hat{\Delta}(s, w_s)$ is a weighted analogue to $\Delta(s)$ in that each sample unit $k$ is weighted by its survey weight $w_{ks}$ whereas there is no weighting with $\Delta(s)$. We thus have $\hat{\Delta}(s, 1) = \Delta(s)$, where $1_s$ is a sample vector of one’s. As a result, the statistic $\hat{t}(s, w_s; e)$ is also a weighted analogue to $t(s; e)$ and we have $\hat{t}(s, 1_s; e) = t(s; e)$. If the statistic $t(s; e)$ can be computed using some classical software package, not necessarily developed to handle survey data, the statistic $\hat{t}(s, w_s; e)$ can also be computed using the same software package provided that it can allow each observation to be weighted by its survey weight.

Typically, the survey weight $w_{ks}$, for a unit $k \in s$, is equal to the inverse of its selection probability, which may then be calibrated to account for known external information (e.g., Deville and Särndal 1992). We assume that the sampling design and the survey weights are constructed so that the following two assumptions hold:

**Assumption 1**: $\sqrt{n} (\hat{\beta}_{ws} - \beta) \rightarrow_n N(0, \Sigma)$, where $\rightarrow_n$ denotes convergence in distribution under the model and the sampling design, and $\Sigma$ is the asymptotic variance-covariance matrix of $\sqrt{n} \hat{\beta}_{ws}$ under the model and the sampling design. The notation “$\rightarrow_n$” stands for the model while the notation “$p$” stands for the probability sampling design.
Assumption 2: $n\hat{A}(s, w_i)$ is symmetric, positive definite and $mp$-consistent for some fixed symmetric positive definite scaling matrix $\hat{A}$.

Note that assumption 2 does not require $\hat{A}(s, w_i)$ to be $p$-consistent for $A(U)$. Indeed, $NA(U)$ will be typically $m$-consistent for $\hat{A}$. Other choices could replace the weighted scaling matrix $\hat{A}(s, w_i)$ in (2.3). For instance, it could be replaced by an estimator of the design variance of $H\hat{\beta}_{ws}$ under simple random sampling (e.g., Rao and Scott 1981). An alternative choice is the common Wald statistic. It is obtained by replacing $\hat{A}(s, w_i)$ in (2.3) by $V_{mp}(H\hat{\beta}_{ws})$, which is an $mp$-consistent estimator of $V_{mp}(H\beta_{ws})$; the variance of $H\beta_{ws}$ evaluated with respect to the model and the sampling design. As pointed out in the paragraph below (2.3), an advantage of using a scaling matrix $\hat{A}(s, w_i)$ such that $\hat{A}(s, 1_s) = A(s)$ is that the resulting test statistic $\hat{i}(s, w_i; c)$ can then be directly computed using classical software packages provided that they allow each observation to be weighted by its survey weight. It is thus more convenient for the users of survey data.

Continuing the above example, we may define our weighted test statistic as

$$\hat{i}(s, w_i; c) = \frac{(\hat{\beta}_{ws} - c)^2}{((\hat{N} - 1)/(n - 1))((\hat{\theta}_{ws}/\hat{N})},$$

where $\hat{N} = \sum_{k=1}^{n} w_k\hat{\theta}_{ws} = \sum_{k=1}^{n} w_k y_k/\sum_{k=1}^{n} w_k$ and $\hat{\theta}_{ws} = \sum_{k=1}^{n} w_k(y_k - \hat{\beta}_{ws})^2/(\hat{N} - 1)$. In (2.4), the underlying weighted scaling matrix is $\hat{A}(s, w_i) = ((\hat{N} - 1)/(n - 1))((\hat{\theta}_{ws}/\hat{N})$, which does not depend on the way the weights are scaled. If they are rescaled so that $\sum_{k=1}^{n} w_k = n$, which is typically done by analysts, then the factor $(\hat{N} - 1)/(n - 1)$ vanishes. The role of this factor, along with other regularity conditions, is to satisfy assumption 2. If the SAS® System is chosen, the test statistic (2.4) is obtained by using the WEIGHT statement in standard procedures. When the null hypothesis is true, it is well known that (2.4) unfortunately does not follow the distribution $\chi^2_1$ or $F_{1,n-1}$ under the model and the sampling design.

To obtain a valid test procedure, we need to approximate the distribution of $\hat{i}(s, w_i; c)$ under the null hypothesis. This can be achieved by using the following result:

**Result 1:** $\hat{i}(s, w_i; c) \xrightarrow{mp} \sum_{q=1}^{Q} \lambda_q \Omega_q$, where $\lambda_q$, for $q = 1, \ldots, Q$, are the eigenvalues of $\Lambda = (\hat{A}^{-1})(H \Sigma H')$ and $\Omega_q$ are independent chi-square random variables with one degree of freedom.

The proof of result 1 uses assumptions 1 and 2 and is given in the appendix. When the null hypothesis is true (i.e., $H\beta = c$), we thus have

$$\hat{i}(s, w_i; c) \xrightarrow{mp} \sum_{q=1}^{Q} \lambda_q \Omega_q.$$  

Rao and Scott (1981) used a similar result to construct their test procedures. They approximated a distribution like (2.5) by a scaled chi-square distribution that matches the estimated first two moments of the right-hand side of (2.5). Instead, we approximate the distribution of $\hat{i}(s, w_i; c)$ under the null hypothesis by using bootstrap weights. This is described in the next section.

Before giving details of our test procedure, it is useful to note that $\hat{i}(s, w_i; c)$ in (2.3) can be written as

$$\hat{i}(s, w_i; c) = \hat{i}(s, w_i; H\beta) + 2(H\hat{\beta}_{ws} - H\beta)'(\hat{A}(s, w_i))^{-1}(H\beta - c) + (H\beta - c)'(\hat{A}(s, w_i))^{-1}(H\beta - c).$$

Under the null hypothesis, the last two terms on the right-hand side of (2.6) vanish and we have $\hat{i}(s, w_i; c) = \hat{i}(s, w_i; H\beta)$. When the null hypothesis is false, the third term on the right-hand side of (2.6) dominates the others as the sample size increases since the first, second and third terms are $O_p(1), O_p(\sqrt{n})$ and $O_p(n)$ respectively, provided that assumptions 1 and 2 hold. Also, since $\hat{A}(s, w_i)$ is positive definite, the third term is always positive. Therefore, a large positive observed value of $\hat{i}(s, w_i; c)$ compared to a large percentile of the distribution of $\hat{i}(s, w_i; H\beta)$ is an indication that the null hypothesis may be wrong.

### 3. The proposed bootstrap method

Let $w_{ik}$ denote a random bootstrap weight for unit $k$, obtained using some bootstrap procedure such as that of Rao et al. (1992), and let $w^*_i$ be the $n$-vector that contains the random bootstrap weight $w_{ik}$ in its $k$th element. The bootstrap estimator $\hat{\beta}_{ws}^* \equiv \hat{\beta}_{ws}$ is obtained similarly to $\hat{\beta}_{ws}$ by replacing the survey weight $w_k$ by its bootstrap version $w_{ik}^*$ for each sample unit. We also denote by $w_{is}^*$, for $b = 1, \ldots, B$, the $B$ $n$-vectors containing the bootstrap weights $w_{is}^b$ in their $k$th element. These $B$ vectors are drawn independently and have the same distribution as $w_{is}^b$; this distribution is called the bootstrap distribution and is denoted by the symbol ‘*’. The $b$th bootstrap estimator $\hat{\beta}_{ws}^*_b$ is defined in an obvious manner.

Before describing our bootstrap test procedure, we first introduce three additional assumptions related to the construction of the bootstrap weights:

**Assumption 3:** $\sqrt{n}(\hat{\beta}_{ws}^* - \hat{\beta}_{ws}) \xrightarrow{**} N(0, \Sigma)$, where $\xrightarrow{**}$ denotes convergence in bootstrap distribution and
\( \hat{\Sigma} \) is the asymptotic bootstrap variance-covariance matrix of \( \sqrt{n} \hat{\theta}_{w^*_t} \).

**Assumption 4:** \( n\hat{A}(s, w^*_t) \) is \( \ast \)-consistent for \( n\hat{A}(s, w_t) \).

**Assumption 5:** \( \hat{\Sigma} \) is \( mp \)-consistent for \( \Sigma \).

Assumptions 3 and 4 are bootstrap analogues to assumptions 1 and 2 and should be satisfied with most bootstrap methods (e.g., those described in the review paper by Sitter 1992) and models (e.g., linear regression model, logistic regression model, etc.). The reader is referred to Shao and Tu (1995, Chapter 6; in particular section 6.4.4) for greater detail.

A comment is in order about assumption 5. This assumption is equivalent to requiring that the bootstrap variance \( V_m(\hat{\beta}_{w^*_t}) \) be \( mp \)-consistent for

\[
V_m(\hat{\beta}_{w^*_t}) = E_n V_p(\hat{\beta}_{w^*_t}) + V_n E_p(\hat{\beta}_{w^*_t}). \tag{3.1}
\]

This means that the bootstrap distribution must reflect the variability due to both the model and the sampling design. Unfortunately, standard design-based bootstrap methods reflect only the variability due to the sampling design so that they only track the first term of the right-hand side of (3.1). Thus, these bootstrap methods do not satisfy assumption 5 in general. However, when the overall sampling fraction \( n / N \) is negligible, the second term of the right-hand side of (3.1) becomes negligible (e.g., see Binder and Roberts 2003) so that the approximation \( V_m(\hat{\beta}_{w^*_t}) \approx E_n V_p(\hat{\beta}_{w^*_t}) \) is appropriate and design-based bootstrap methods can be used. In many household surveys, the overall sampling fraction is actually quite small. Indeed, bootstrap weights are often obtained under the assumption that the first-stage sampling fractions are small (e.g., Rao et al. 1992). Developing bootstrap procedures that capture both terms of (3.1) is an area for future research.

Under assumptions 3 and 4, we obtain our second result:

**Result 2:** \( \hat{i}(s, w^*_t; \boldsymbol{H}\hat{\beta}_{w^*_t}) \xrightarrow{mp} \sum_{q=1}^Q \hat{\lambda}_q \Omega_q \), where \( \hat{\lambda}_q \), for \( q = 1, ..., Q \), are the eigenvalues of \( \hat{\Lambda} = [n\hat{A}(s, w^*_t)]^{-1}(\hat{\Sigma}\hat{\Sigma}'H') \) and \( \Omega_q \) are again independent chi-square random variables with one degree of freedom.

The proof of result 2 is omitted as it is very similar to the proof of result 1 given in the appendix. From assumptions 2 and 5, \( \hat{\Lambda} \) is \( mp \)-consistent for \( \Lambda \). Thus, using results 1 and 2, the bootstrap distribution of \( \hat{i}(s, w^*_t; \boldsymbol{H}\hat{\beta}_{w^*_t}) \) is asymptotically the same as the \( mp \)-distribution of \( \hat{i}(s, w^*_t; \boldsymbol{H}\beta) \), which is itself the same as the \( mp \)-distribution of \( \hat{i}(s, w^*_t; \boldsymbol{e}) \) under the null hypothesis; the distribution that we want to approximate. This suggests the following bootstrap test procedure:

1. Obtain bootstrap weights, \( w^*_b \), for \( b = 1, ..., B \).

2. Compute \( \hat{i}(s, w^*_b; \boldsymbol{H}\hat{\beta}_{w^*_b}) \), for \( b = 1, ..., B \).

3. Since a large value of \( \hat{i}(s, w^*_b; \boldsymbol{e}) \) leads to rejecting the null hypothesis, compute the observed significance level (\( p \)-value) as

\[
\frac{\# \{\hat{i}(s, w^*_b; \boldsymbol{H}\hat{\beta}_{w^*_b}) > \hat{i}(s, w^*_b; \boldsymbol{e})\}}{B}.
\]

The null hypothesis is rejected if this value is lower than the significance level \( \alpha \) (e.g., 5%).

Note that the statistic to be bootstrapped is \( \hat{i}(s, w^*_b; \boldsymbol{H}\hat{\beta}_{w^*_b}) \) and not \( \hat{i}(s, w^*_b; \boldsymbol{e}) \). The use of the latter would not properly reflect the distribution under the null hypothesis and would thus violate the first guideline in Hall and Wilson (1991).

If \( \hat{i}(s, w^*_b; \boldsymbol{H}\hat{\beta}_{w^*_b}) \) is pivotal then the second guideline of Hall and Wilson (1991) is also satisfied. The fact that \( \hat{i}(U; \boldsymbol{e}) \) is asymptotically pivotal certainly helps in obtaining a better bootstrap test procedure. However, it does not unfortunately guarantee that \( \hat{i}(s, w^*_b; \boldsymbol{e}) \) is also asymptotically pivotal, particularly when sampling is informative. Nevertheless, failure to use a pivotal statistic does not invalidate the above test procedure and may not reduce its power. But, it may reduce the level accuracy of the test. As pointed out by Hall and Wilson (1991), it is sometimes appropriate to disregard the second guideline.

The main advantage of using the simple (possibly non-pivotal) statistic \( \hat{i}(s, w^*_b; \boldsymbol{e}) \) in (2.3) and the bootstrap statistic \( \hat{i}(s, w^*_b; \boldsymbol{H}\hat{\beta}_{w^*_b}) \) that, once bootstrap weights have been provided on the microdata file, these statistics are easily obtained using classical software packages that ignore sampling design features. Moreover, we show in section 5, through a simulation study, that our bootstrap test procedure performs similarly to the Rao-Scott procedure and better than the Wald and Bonferroni procedures.

### 4. A linear regression example

To better illustrate the theory in a practical context, let us now assume that, conditional on \( \boldsymbol{X}_U \), the random variables \( y_k \), for \( k \in U \), are independently distributed with mean \( E_n(y_k | \boldsymbol{X}_U) = \mathbf{x}_k' \boldsymbol{\beta} \) and variance \( V_n(y_k | \boldsymbol{X}_U) = 0 \), where \( \mathbf{x}_k \) is an \( r \)-vector of linearly independent variables for unit \( k \). Recall that we are interested in testing the null hypothesis \( \Lambda \): \( \boldsymbol{H}\beta = \boldsymbol{e} \) against the alternative hypothesis \( \Lambda \): \( \boldsymbol{H}\beta \neq \boldsymbol{e} \). If the entire population could be observed, the common statistic

\[
\hat{i}(U; \boldsymbol{e}) = \frac{\boldsymbol{H}(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' H')^{-1} (\boldsymbol{H}\hat{\beta}_U - \boldsymbol{e})}{Q \hat{\theta}_U} \tag{4.1}
\]
could be used, where
\[ \hat{\beta}_U = \left( \sum_{k \in U} x_k x_k' \right)^{-1} \sum_{k \in U} x_k y_k \]
and
\[ \hat{\theta}_U = \frac{\sum_{k \in U} (y_k - x_k' \hat{\beta}_U)^2}{N - r} . \]

The statistic \( t \) in (4.1) follows the distribution \( F_{q, N-r} \) under the null hypothesis. It reduces to (2.2) when \( Q = r = H = x_k = 1 \) in (4.1).

A weighted sample version of (4.1), which can be written in the form of (2.3), is
\[ \hat{i} (s, w_s; c) = \frac{(\hat{\beta}_{ws} - \Omega) \left( \left( \sum_{k \in s} w_k x_k x_k' \right)^{-1} - H \right) \left( \hat{\beta}_{ws} - \Omega \right)}{Q \hat{\theta}_{ws} \{(N - r)/(n - r)\}} , \]
where
\[ \hat{\beta}_{ws} = \left( \sum_{k \in s} w_k x_k x_k' \right)^{-1} \sum_{k \in s} w_k x_k y_k \]
and
\[ \hat{\theta}_{ws} = \frac{\sum_{k \in s} w_k (y_k - x_k' \hat{\beta}_{ws})^2}{N - r} . \]

For instance, the statistic \( \hat{i} (s, w_s; c) \) in (4.2) could be obtained by using the WEIGHT statement in the procedure REG of SAS as long as \( w_k > 0 \), for \( k \in s \). Note that it satisfies assumption 2 and does not depend on the way the weights are scaled. Again, if the weights are rescaled so that \( \sum_{k \in s} w_k = n \), the factor \((N - r)/(n - r)\) in (4.2) vanishes. The test statistic (4.2) reduces to (2.4) when \( Q = r = H = x_k = 1 \) in (4.2), (4.3) and (4.4). The bootstrap statistic \( \hat{i} (s, w_s; \hat{\beta}_{ws}) \) as well as \( \hat{\theta}_{ws} \) and \( \hat{\beta}_{ws} \) are obtained similarly to \( \hat{i} (s, w_s; c) \), \( \hat{\beta}_{ws} \) and \( \hat{\theta}_{ws} \) in (4.2), (4.3) and (4.4) respectively, except that \( w_k \) is replaced by \( w_k^* \) and \( c \) is replaced by \( \hat{\beta}_{ws} \).

Remark 1: Note that \( w_k^* \) is likely to be 0 for some units \( k \in s \) (see, for example, Rao et al. 1992). In some software packages such as SAS, the number of observations used in the analysis of the \( b^\text{th} \) bootstrap replicate, \( n^b \), is equal to the number of units \( k \in s \) for which \( w_k^b > 0 \). Such software packages may use \( n^b - r \) instead of \( n - r \) when computing the bootstrap statistic \( \hat{i} (s, w_s^b; \hat{\beta}_{ws}) \). One must thus make sure that \( n - r \) is used and, if not, that the bootstrap statistic computed from these packages is properly adjusted before applying the proposed bootstrap test procedure. One way of avoiding this problem is to add a very small positive value (e.g., \( 1 \times 10^{-10} \)) to each bootstrap weight \( w_k^b \), for \( k \in s \), so that no observation is excluded from the computation of \( \hat{i} (s, w_s^b; \hat{\beta}_{ws}) \).

Remark 2: Let us define the bootstrap statistic \( \hat{i} (s, w_s^b; 0) \) by replacing \( y_k \) by \( e_k = y_k - x_k' \hat{\beta}_{ws} \) in \( \hat{i} (s, w_s^b; 0) \), for each \( k \in s \). It is not difficult to show that \( \hat{i} (s, w_s^b; 0) = \hat{i} (s, w_s^b; \hat{\beta}_{ws}) \) so that our bootstrap procedure can be implemented using either \( \hat{i} (s, w_s^b; 0) \) or \( \hat{i} (s, w_s^b; \hat{\beta}_{ws}) \) when a linear regression model is used. The former may sometimes be more convenient with some software packages. This was the case in our simulation study since the use of \( \hat{i} (s, w_s^b; 0) \) allowed us to get rid of manually typing the values of \( \hat{\beta}_{ws} \) for each selected sample. An informal explanation for the equality \( \hat{i} (s, w_s^b; 0) = \hat{i} (s, w_s^b; \hat{\beta}_{ws}) \) can be obtained by treating \( \hat{\beta}_{ws} \) as a fixed quantity, which is actually the case under the bootstrap distribution. The bootstrap statistic \( \hat{i} (s, w_s^b; \hat{\beta}_{ws}) \) can thus be interpreted as a statistic aiming at testing the null hypothesis \( H_0^* : \beta = 0 \) or, alternatively, \( H_0^* : \beta = 0 \).

Remark 3: We have already mentioned that the WEIGHT statement is necessary to obtain a weighted statistic if the proposed bootstrap test procedure is implemented using the procedure REG of SAS. Also, the TEST statement is necessary to request the desired statistics to be produced and the “ODS OUTPUT TESTANNOVA =” statement to save these requested statistics in a SAS dataset specified by the user.

5. Some alternative procedures for linear regression

In this section, we briefly describe some test procedures in the context of linear regression exposed in section 4; namely, two naïve procedures that are sometimes used in practice as well as specific implementations of the Rao-Scott, Wald and Bonferroni procedures. They will all be evaluated in the simulation study in section 6.

The Bonferroni, Wald and Rao-Scott procedures, described in sections 5.2, 5.3 and 5.4 respectively, all need an mp-consistent estimator \( \hat{V}_{mp} (\hat{\beta}_{ws}) \) of \( V_{mp} (\beta_{ws}) \). In the simulation study in section 6, we have used the bootstrap variance estimator
\[ \hat{V}_{mp} (\hat{\beta}_{ws}) = \frac{\sum_{b=1}^B (\hat{\beta}_{ws}^{b*} - \hat{\beta}_{ws}) (\hat{\beta}_{ws}^{b*} - \hat{\beta}_{ws})'}{B} . \]
It is worth noting that the validity of assumption 5 is thus not only required for our proposed bootstrap method but also for the Bonferroni, Wald and Rao-Scott methods.

5.1 Two naïve procedures

The weighted version of the naïve procedure consists of using the statistic \( i(s, w; c) \) in (4.2), which is compared to the upper tail of the distribution \( F_{Q, n-r} \). The unweighted version uses the statistic \( i(s, 1; c) \), which is again compared to the upper tail of the distribution \( F_{Q, n-r} \). Both procedures are not expected to work well under informative sampling but are still often used in practice, especially the weighted version. Note that if sampling is not informative, the unweighted version, that ignores the sampling design, leads to a simple, valid and reasonably powerful test.

5.2 The Bonferroni procedure

The Bonferroni procedure was studied by Korn and Graubard (1990). It is simple to use and was shown to work well in their empirical study. To describe this procedure, let \( H^q \) represent the \( q^{th} \) row of \( H \) and \( c_q \) the \( q^{th} \) element of \( c \). Then, compute the \( Q \) weighted statistics

\[
i^\text{BON}_q(s; c_q) = \frac{(H^q \hat{\beta}_{ws} - c_q)^2}{H^q \hat{V}_{mp} (\hat{\beta}_{ws}) H_q}.
\]

(5.2)

The largest statistic \( i^\text{BON}_q(s; c_q) \), for \( q = 1, \ldots, Q \), is compared to the upper tail of the distribution \( F_{1, \alpha} \) with a revised significance level \( \alpha \). The number of degrees of freedom \( d \) is equal to the number of sampled primary sampling units minus the number of strata. Note that this procedure depends in general on the model parametrization used.

5.3 WALD F-procedure

An F-version of the standard Wald chi-square statistic, with adjusted denominator degrees of freedom as proposed by Fellegi (1980), can be defined as

\[
i^F(s; c) = \frac{d - Q + 1}{Qd} (H^T \hat{\beta}_{ws} - c)^T (H^T \hat{V} \hat{V}_{mp} (\hat{\beta}_{ws}) H^T) (H^T \hat{\beta}_{ws} - c). \]

(5.3)

The statistic \( i^F(s; c) \) is compared to the upper tail of the distribution \( F_{Q, d- Q_v+1} \). This procedure is implemented in the software package SUDAAN (Research Triangle Institute 2004).

5.4 Rao-Scott F-procedure

Another procedure consists of using an F-version (see Rao and Thomas 2003) of the second-order adjusted chi-square statistic of Rao and Scott (1981), which is based on Satterthwaite’s correction for the number of degrees of freedom. We use an adaptation of these authors’ method for linear regression, as implemented in the software package SUDAAN (Research Triangle Institute 2004). The statistic is defined as

\[
i^\text{RS}(s; c) = \frac{1}{\lambda(1 + a^2)} Q^* \]

(5.4)

where \( \hat{V}_{\text{SRS}}(\hat{\beta}_{ws}) \) is an estimator of the variance-covariance matrix of \( \hat{\beta}_{ws} \) under a simple random sampling design, \( \lambda \) is the average of the eigenvalues of the generalized design effect matrix \( \hat{V}_{\text{SRS}}(\hat{\beta}_{ws})^{-1} \hat{V}_{mp} (\hat{\beta}_{ws}) \), \( a \) is the coefficient of variation of these eigenvalues and \( Q^* = Q / (1 + a^2) \). The Rao-Scott F-statistic \( i^\text{RS}(s; c) \) is compared to the upper tail of the distribution \( F_{Q', d'} \).

6. Simulation study

We performed a simulation study to investigate the level and power of the above test procedures in the case of informative and non-informative sampling. In sections 6.1 and 6.2, we describe the population and sample creation respectively. We then define the null hypotheses to be tested in section 6.3, describe the methods evaluated in section 6.4 and present simulation results in section 6.5.

6.1 Generation of the populations

We generated four populations of \( N = 10,000 \) units. First, a categorical variable \( v_k \) was generated independently for each population unit \( k \) so that \( v_k = i \), for \( i = 1, \ldots, I \), with probability \( P(v_k = i) = 1 / I \), where \( I \) is the number of categories of \( v_k \), which was set equal to 5. The dependent variable \( y \) was generated as

\[
y_k = \alpha_o + \alpha_i \left( \frac{v_k - (I + 1) / 2}{2} \right) + \sigma \varphi_k, \]

(6.1)

where \( \varphi_k \sim N(0, 1) \), \( \alpha_o = 10 \) and \( \sigma = 3 \). The four populations that we generated only differ in the choice of \( \alpha_i \), which controls the correlation between \( y \) and \( v \). We considered \( \alpha_i = 0, 0.25, 0.50 \) and 0.75.

6.2 Generation of samples and bootstrap weights

From each of the above four populations, 5,000 stratified simple random samples of size 100 were selected without replacement under two different stratification scenarios aimed at simulating both informative and non-informative sampling. In the case of non-informative sampling, the strata correspond exactly to the five categories of variable \( v \) defined above. In the case of informative sampling, the
strata are defined by the cross-classification of variable \( v \) and another categorical variable \( z \) that depends on the random error term \( \sigma \phi_k \) in (6.1). For each population unit \( k \), variable \( z \) was created as follows: \( z_k = 1 \), if \( \sigma \phi_k > 0 \), and \( z_k = 2 \), otherwise. This leads to 10 strata in the informative case that are constructed by crossing the five categories of \( v \) with the two categories of \( z \). Each of the 10 informative strata contains about 1,000 population units while each of the 5 non-informative strata contains about 2,000 population units.

Furthermore, two different stratum allocation schemes were used. The scheme, SCHEME_UNEQUAL, allocates the 100 sample units among the strata in the following way:

<table>
<thead>
<tr>
<th>Informative</th>
<th>Non-informative</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v )</td>
<td>( v )</td>
</tr>
<tr>
<td>1  2  3  4  5</td>
<td>1  2  3  4  5</td>
</tr>
<tr>
<td>1  4  4  16 4</td>
<td>8  12 20 28 32</td>
</tr>
<tr>
<td>2  4  8  4  24</td>
<td>4</td>
</tr>
</tbody>
</table>

The second scheme, denoted SCHEME_EQUAL, assigns the same number of units in each stratum as follows:

<table>
<thead>
<tr>
<th>Informative</th>
<th>Non-informative</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v )</td>
<td>( v )</td>
</tr>
<tr>
<td>1  2  3  4  5</td>
<td>1  2  3  4  5</td>
</tr>
<tr>
<td>1  10 10 10 10</td>
<td>20 20 20 20 20</td>
</tr>
<tr>
<td>2  10 10 10 10</td>
<td>20 20 20 20 20</td>
</tr>
</tbody>
</table>

The two different schemes lead to very different sets of survey weights. The weights resulting from the SCHEME_UNEQUAL allocation are much more variable than those from SCHEME_EQUAL. Note that we simply defined the survey weight \( w_k \) as the inverse of the selection probability of unit \( k \).

Finally, for each selected sample, 500 design-based bootstrap weights were calculated for each sampled unit, as described in Rao et al. (1992), among others. In our implementation of this methodology, each bootstrap sample was selected with replacement by stratified simple random sampling with \( n_h - 1 \) draws from the \( n_h \) sample units in stratum \( h \). This methodology takes the sampling design variability into account (with a slight overestimation of the design variance due to assuming with-replacement sampling) but ignores the model variability. This is acceptable since the overall sampling fraction (1/100) is small.

6.3 Null hypotheses

For each selected sample, we modeled \( y_k \) as a function of \( v_k \) using an analysis of variance model. More specifically, we defined indicator variables

\[
x_{ik} = \begin{cases} 1, & \text{if } v_k = i, \\ 0, & \text{otherwise,}
\end{cases}
\]

for \( i = 1, \ldots, I \), and fitted the linear model

\[
y_k = \beta_0 + \sum_{i=1}^{I-1} \beta_i x_{ik} + \epsilon_k
\]

using the weighted least-squares technique, where \( \epsilon_k \) is a random error term with mean 0 and constant variance. We considered testing the following two null hypotheses:

**TEST1:** \( H_0: \beta_1 = 0 \)

**TEST2:** \( H_0: \beta_1 = \beta_2 = \ldots = \beta_{I-1} = 0 \).

Note that both null hypotheses are true for the population with \( v_k = 0 \) while they are false for the other populations. The latter three populations are used to assess the power of the different test procedures under study.

6.4 Test methods

For each selected sample, we tested the above two null hypotheses using five different methods: the proposed bootstrap method, the naïve method (both unweighted and weighted versions) described in section 5.1, the Bonferroni method described in section 5.2, the Wald F method described in section 5.3 and the Rao-Scott F method described in section 5.4. Results for the naïve method are standard output in the software SAS whereas the Wald and Rao-Scott F-statistics are standard output in the SUDAAN statistical software, version 9. The Bonferroni statistics (5.2) are also obtained through SUDAAN. The proposed method is programmed in the statistical software SAS, version 8.

In addition, we also performed the simulation study using a linearized variance estimator in the Wald, Rao-Scott and Bonferroni methods instead of the bootstrap variance estimator (5.1). Rejection rates obtained using the linearized variance estimator were slightly lower but quite similar to those obtained using (5.1). Given this observation and that our focus is on bootstrap methods, we neither show nor discuss these additional results in the next section.

6.5 Simulation results

For each population, stratification scenario, allocation scheme, null hypothesis and method, we calculated the rejection rate in percentage over the 5,000 selected samples (using a 5% significance level). Results are given below in tables 3A, 3B, 4A and 4B. The results are more striking and more interesting for the null hypothesis TEST2 than the null
hypothesis TEST1. We will thus focus our discussion of the results on the former.

Tables 3A and 3B contain the results in the case of informative sampling, which is of more interest to us. Let us discuss first results in table 3A for SCHEME_UNEQUAL. Both naïve methods perform poorly as they do not properly exploit sampling design information. On the one hand, the unweighted version is definitely too liberal as its rejection rate is far above 5% under the null hypothesis. On the other hand, the weighted version is too conservative and significantly lacks power when compared to other methods. The Wald method is too liberal with a rejection rate of 15.8% when \( H_0 \) is true. The simple Bonferroni method improves the situation although it is still too liberal with a rejection rate of 11.4% when \( H_0 \) is true. This result is somewhat surprising as the Bonferroni method is known to be (asymptotically) conservative. A referee suggested that we consider an improved Bonferroni method such as that developed by Benjamini and Hochberg (1995). In this simulation study, such a method would not help as it always rejects more often than the standard Bonferroni method. The Rao-Scott method significantly outperforms the Wald and Bonferroni methods under the null hypothesis with a rejection rate of 6.8%. The proposed bootstrap method is comparable to the proven but more complicated Rao-Scott method with perhaps even a slight improvement in the level with a rejection rate of 6.2% when \( H_0 \) is true. However, the Rao-Scott method is slightly more powerful than the proposed bootstrap method.

Table 3B contains results under SCHEME_EQUAL in the informative sampling scenario. Here, the weighted and unweighted versions of the naïve method yield similar results since the variability of the survey weights is quite small. Even in this case, the naïve method is definitely too conservative, which results in an extremely low power. All other methods are comparable both in terms of level (\( H_0 \) true) and power (\( H_0 \) false) although the Wald method is still slightly too liberal compared to the Bonferroni, Rao-Scott and proposed bootstrap methods with a rejection rate of 7.9% when \( H_0 \) is true.

Tables 4A and 4B contain the results in the case of non-informative sampling. Again, let us discuss first results in table 4A for SCHEME_UNEQUAL. As expected, the naïve unweighted method performs well here while the naïve weighted method becomes too liberal with a rejection rate of 12.8% when \( H_0 \) is true. In terms of the level, the proposed method is competitive to the naïve unweighted method and even slightly conservative. It outperforms the Wald and Bonferroni methods with a rejection rate of 6.8% when \( H_0 \) is true. However, its power is however slightly less than these latter two competitors but still acceptable.
Table 4A
Rejection rates at the 5% significance level under SCHEME_UNEQUAL and non-informative sampling

<table>
<thead>
<tr>
<th>SCHEME_UNEQUAL</th>
<th>Ho TRUE</th>
<th>Non-Informative Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>( \alpha_1 = 0 )</td>
<td>( \alpha_1 = 0.25 )</td>
</tr>
<tr>
<td>Naïve Unweighted</td>
<td>4.2</td>
<td>13.5</td>
</tr>
<tr>
<td>Naïve Weighted</td>
<td>11.4</td>
<td>24.6</td>
</tr>
<tr>
<td>Wald</td>
<td>7.6</td>
<td>16.8</td>
</tr>
<tr>
<td>Rao-Scott</td>
<td>7.6</td>
<td>6.4</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>7.6</td>
<td>7.1</td>
</tr>
<tr>
<td>Proposed Bootstrap</td>
<td>6.3</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Table 4B
Rejection rates at the 5% significance level under SCHEME_EQUAL and non-informative sampling

<table>
<thead>
<tr>
<th>SCHEME_EQUAL</th>
<th>Ho TRUE</th>
<th>Non-Informative Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>( \alpha_1 = 0 )</td>
<td>( \alpha_1 = 0.25 )</td>
</tr>
<tr>
<td>Naïve Unweighted</td>
<td>4.9</td>
<td>17.2</td>
</tr>
<tr>
<td>Naïve Weighted</td>
<td>5.0</td>
<td>17.4</td>
</tr>
<tr>
<td>Wald</td>
<td>5.7</td>
<td>18.8</td>
</tr>
<tr>
<td>Rao-Scott</td>
<td>5.7</td>
<td>18.3</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>5.7</td>
<td>18.8</td>
</tr>
<tr>
<td>Proposed Bootstrap</td>
<td>5.0</td>
<td>16.4</td>
</tr>
</tbody>
</table>

Table 4B contains results under SCHEME_EQUAL in the non-informative sampling scenario. In this table, the methods do not appear to differ drastically. As expected, the naïve method (both weighted and unweighted versions) performs well although it did not outperform the Rao-Scott and Bonferroni methods in this simulation study. The proposed method is still slightly conservative in this non-informative scenario and has slightly less power than the other methods.

To investigate the effect of large samples on the test procedures, we also performed some simulations with sample sizes that are ten times larger than in the original setup, as suggested by one reviewer. That is, we considered a population size of 100,000 and selected 1,000 samples of size 1,000 thus deliberately keeping the same small sampling fraction. From this setup, we obtained results when \( H_0 \) is true, shown in table 5, for both informative and non-informative sampling under unequal stratum allocation. As expected, all the methods other than the naïve ones have similar rejection rates that are indeed slightly lower than 5%. This illustrates that the differences between the methods become less important as the sample size increases.

Table 5
Rejection rates at the 5% significance level under SCHEME_UNEQUAL

<table>
<thead>
<tr>
<th>SCHEME_UNEQUAL</th>
<th>Informative</th>
<th>Non-informative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>( \alpha_1 = 0 )</td>
<td>( \alpha_1 = 0 )</td>
</tr>
<tr>
<td>Naïve Unweighted</td>
<td>100.0</td>
<td>3.7</td>
</tr>
<tr>
<td>Naïve Weighted</td>
<td>1.3</td>
<td>9.3</td>
</tr>
<tr>
<td>Wald</td>
<td>4.6</td>
<td>3.2</td>
</tr>
<tr>
<td>Rao-Scott</td>
<td>4.6</td>
<td>3.2</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>4.6</td>
<td>3.2</td>
</tr>
<tr>
<td>Proposed Bootstrap</td>
<td>4.4</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Overall, our proposed bootstrap method was the best in terms of the level, followed closely by the Rao-Scott method. It gave somewhat conservative results in the non-informative sampling scenarios. This was accompanied by a slight loss of power. The Rao-Scott method is a good alternative if users have access to an appropriate software package. The Bonferroni method is simple to use but may be too liberal and the Wald method is even worse. The naïve methods may have serious deficiencies, either in the level or in the power, although the naïve unweighted method is viable if one is reasonably sure that sampling is not informative.
7. Summary and discussion

We have proposed a general and simple bootstrap procedure for testing hypotheses from survey data, which could also be applied outside the survey sampling field. Our procedure uses classical model-based test statistics and is thus easy to implement for analysts using classical software packages. We have shown in a simulation study that it performed well in the context of a linear regression model. These good results are encouraging and may suggest that our proposed bootstrap procedure could be useful with other more complicated models and other statistics. The idea could also be easily adapted for the construction of bootstrap confidence intervals.

One could also consider bootstrapping an asymptotically pivotal statistic such as the Rao-Scott statistic (5.4). This would however involve double bootstrapping if is estimated using the bootstrap technique as in (5.1). Double bootstrapping requires generating another set of bootstrap replicates for each initial bootstrap replicate. Although better test procedures could potentially be obtained, double bootstrapping may not be convenient for analysts. By focusing on simpler statistics that do not involve the bootstrap technique, our test procedure avoids double bootstrapping and remains simple.

The properties of our method depend not only on the choice of the test statistic but also on the construction of the bootstrap weights. Typically, bootstrap weights capture the first two design moments of the sampling error, which should be sufficient in most cases to satisfy our bootstrap assumptions 3, 4 and 5. Bootstrap weights that also capture the third design moment could perhaps be useful for improving the level accuracy of the bootstrap test. This needs further investigation. Finally, as already pointed out in section 3, standard design-based bootstrap weights satisfy assumption 5 only when the overall sampling fraction is negligible so that the model portion of the total variance (3.1) is negligible. Research is needed to develop proper bootstrap weights, when a non-negligible sampling fraction is used, that capture both the model and the design portions of the total variance.

Acknowledgements

We sincerely thank the Associate Editor and two referees for their comments. We also thank J.N.K. Rao from Carleton University as well as David Binder and Yves Lafortune from Statistics Canada for their comments and stimulating discussions on this topic. All these comments and discussions were useful to improve the general quality of the paper and its clarity.

Appendix

Proof of result 1

From assumption 1, we can easily see that
\[ \sqrt{n} \left( \hat{H} \hat{w} - H \hat{\beta} \right) \overset{mp}{\longrightarrow} N(0, H \Sigma H'). \tag{A.1} \]

Using a standard result on quadratic forms (e.g., Seber 1984, page 540) and equation (A.1), we obtain
\[ n \left( \hat{H} \hat{w} - H \hat{\beta} \right) \overset{mp}{\longrightarrow} \sum_{q=1}^{Q} \lambda_q \Omega_q, \tag{A.2} \]

where \( \lambda_q \), for \( q = 1, \ldots, Q \), are the eigenvalues of \( \Lambda = \hat{\Lambda}^{-1} \left( \hat{H} \Sigma H' \right) \) and \( \Omega_q \) are independent chi-square random variables with one degree of freedom. Therefore, from (A.2) and assumption 2, we have
\[ \hat{i} (s, w; H \beta) = \left( \hat{H} \hat{w} - H \hat{\beta} \right) \hat{A} (s, w)^{-1} \left( \hat{H} \hat{w} - H \hat{\beta} \right) \overset{mp}{\longrightarrow} \sum_{q=1}^{Q} \lambda_q \Omega_q. \]

References


