Weighting in the regression analysis of survey data with a cross-national application

Chris Skinner* Ben Mason†

17 July 2012

Abstract

A class of survey weighting methods provides consistent estimation of regression coefficients under unequal probability sampling. The minimization of the variance of the estimated coefficients within this class is considered. A series of approximations leads to a simple modification of the usual design weight. One type of application where unequal probabilities of selection arise is in cross-national comparative surveys. In this case, our argument suggests the use of a certain kind of within-country weight. We investigate this idea in an application to data from the European Social Survey, where we fit a logistic regression model with vote in an election as the dependent variable and with various variables of political science interest included as explanatory variables. We show that the use of the modified weights leads to a considerable reduction in standard errors compared to design weighting.

1 INTRODUCTION

Survey weighting is often used when regression models are estimated from survey data. Unweighted estimators of regression coefficients may be biased

*Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, U.K.

†School of Social Sciences, University of Southampton, Southampton SO17 1BJ, U.K

if the inclusion of units in the sample is correlated with the outcome variable conditional on the explanatory variables. Weighting by the reciprocals of the unit inclusion probabilities, as in Horvitz-Thompson estimation, enables this bias to be corrected and regression coefficients to be estimated consistently (Fuller, 2009, Sect. 6.3). A potential disadvantage of weighting, however, is that it may inflate the variances of the coefficient estimators. This problem may be more acute the greater the variability of the sample inclusion probabilities.

A number of approaches exist to control the variance inflation while retaining consistency. A simple option takes the design variables, which account for variation in the sample inclusion probabilities, and includes these as additional explanatory variables in the model. This is only appropriate, however, when this respecified model remains of scientific interest.

In this paper we consider approaches which seek to improve estimation efficiency by modifying the survey weights. Such approaches can be straightforward to implement for practitioners. We develop such an approach within the general estimating function framework of Thompson (1997, Ch.6). Our approach extends a variance minimization approach proposed by Fuller (2009, Sect. 6.3.2) for linear regression and leads to a weight modification similar to the semi-parametric approach proposed by Pfeffermann and Sverchkov (1999).

We have presented the rationale for weighting as the removal of bias due to unequal selection probabilities. Another rationale that is sometimes given is that it addresses model misspecification, in the sense that the weighted estimator is consistent for a finite population regression coefficient, which

is meaningful even if the model fails (Godambe and Thompson, 1986). We shall discuss this issue too, but only as a secondary consideration.

We explore the application of our modified weighting method to a specific cross-national regression analysis of European Social Survey (ESS) data. The application exemplifies a particular problem of weighting arising in cross-national comparative surveys when data are pooled across countries (Thompson, 2008, Section 3). It is common in the design of such surveys for sample sizes in different countries to be much less variable than population sizes and for this to lead to very different sampling fractions across countries. This implies that data from large countries may dominate pooled analyses employing Horvitz-Thompson weighting, leading to inefficient use of sample data (Thompson, 2008, Section 3). In this paper we show how the modified weighting approach may address this problem.

The paper is organized as follows. We develop the theory behind our modified weighting approach in Section 2. We discuss the generic kind of cross-national application in Section 3. The specific application to ESS data is presented in Section 4 and some final discussion is provided in Section 5.

2 ESTIMATION THEORY

Following Thompson (1997, Ch. 6), let the units in a finite population U be labelled $j = 1, \ldots, N$ and let the row vector $(y_j, \boldsymbol{x}_j)$ denote the associated values of a pair of response and explanatory variables in a regression analysis, where y_j is the realized value of the random variable Y_j and the $1 \times k$ vector $\boldsymbol{x}_j$ is treated as fixed. Consider a regression model under which the Y_j are independent, with a distribution depending on a $k \times 1$ column vector $\boldsymbol{\theta}$ of

parameters, such that

$$E_m\{\boldsymbol{\phi}_j(Y_j;\boldsymbol{x}_j,\boldsymbol{\theta})\} = \mathbf{0} \quad \text{for} \quad j = 1,\ldots,N, \tag{1}$$

where $\boldsymbol{\phi}_j(Y_j;\boldsymbol{x}_j,\boldsymbol{\theta})$ is a $k \times 1$ vector estimating function and $E_m(.)$ denotes expectation under the model. The population-level equations

$$\sum_{j=1}^{N} \boldsymbol{\phi}_j(Y_j;\boldsymbol{x}_j,\boldsymbol{\theta}) = \mathbf{0}, \tag{2}$$

are unbiased estimating equations in the terminology of Godambe and Thompson (2009). Assuming that in some asymptotic framework a solution $\boldsymbol{\theta}_U$ to these equations eventually exists uniquely and under additional regularity conditions (Godambe and Thompson, 2009), $\boldsymbol{\theta}_U$ converges in probability to $\boldsymbol{\theta}$. A particular instance of such an estimating function is the unit-level score function given by

$$\boldsymbol{\phi}_j(Y_j;\boldsymbol{x}_j,\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} \log f_j(Y_j;\boldsymbol{x}_j,\boldsymbol{\theta}).$$

where $f_j(Y_j;\boldsymbol{x}_j,\boldsymbol{\theta})$ is the probability density or mass function for Y_j and $\boldsymbol{\theta}_U$ is the 'census' maximum likelihood estimator of $\boldsymbol{\theta}$ which would apply if all population values of $(y_j,\boldsymbol{x}_j)$ were observed. For illustration, if a binary variable y_j, taking values 0 or 1, obeys a logistic regression model, where $\boldsymbol{\theta}$ is the vector of regression coefficients, we have $\log\{f_j(1;\boldsymbol{x}_j,\boldsymbol{\theta})/f_j(0;\boldsymbol{x}_j,\boldsymbol{\theta})\} = \boldsymbol{x}_j\boldsymbol{\theta}$ and

$$\boldsymbol{\phi}_j(Y_j;\boldsymbol{x}_j,\boldsymbol{\theta}) = \{Y_j - f_j(1;\boldsymbol{x}_j,\boldsymbol{\theta}))\}\boldsymbol{x}_j^T, \tag{3}$$

where ^{T} denotes transpose.

Now suppose that the $(y_j,\boldsymbol{x}_j)$ are only observed for units j in a sample drawn by a probability sampling scheme from U and let I_j, $j = 1,\ldots,N$, be

the sample indicators, where $I_j = 1$ if unit j is sampled and $I_j = 0$ if not. We shall be interested in the weighted estimator $\hat{\boldsymbol{\theta}}_w$ which solves the sample estimating equations

$$\sum_{j=1}^{N} w_j I_j \boldsymbol{\phi}_j(Y_j; \boldsymbol{x}_j, \boldsymbol{\theta}) = \mathbf{0}, \tag{4}$$

where w_j is a survey weight. Corresponding to condition (1) for the consistency of $\boldsymbol{\theta}_U$, these estimating equations are unbiased under the joint distribution induced by the design and the model and $\hat{\boldsymbol{\theta}}_w$ is consistent for $\boldsymbol{\theta}$ if

$$E_m E_p[w_j I_j \boldsymbol{\phi}_j(Y_j; \boldsymbol{x}_j, \boldsymbol{\theta})] = \mathbf{0}, \quad \text{for} \quad j = 1, \ldots, N \tag{5}$$

where $E_p(.)$ denotes expectation with respect to the sampling scheme. Two basic cases when condition (5) holds are:

(i) the w_j are constant, so that $\hat{\boldsymbol{\theta}}_w$ is the unweighted estimator, and sampling is *noninformative*, that is I_j and Y_j are independent (conditional on $\boldsymbol{x}_j$) for each j. This arises, in particular, when sample inclusion depends just on a set of design variables which are included in the vector $\boldsymbol{x}_j$. Fuller (2009, Section 6.3.1) reviews tests of this noninformative condition, including a test proposed by DuMouchel and Duncan (1983).

(ii) $w_j = d_j$, the design (Horvitz-Thompson) weight given by $d_j = \pi_j^{-1}$, where $\pi_j = E_p(I_j)$ is the inclusion probability of unit j.

In each of these cases, the proof of (5) assumes model (1) holds. For example, to demonstrate that (5) holds in case (ii) we write $E_m E_p(d_j I_j \boldsymbol{\phi}_j) = E_m\{E_p(d_j I_j)\boldsymbol{\phi}_j\}$, where $\boldsymbol{\phi}_j = \boldsymbol{\phi}_j(Y_j; \boldsymbol{x}_j, \boldsymbol{\theta})$, and use $E_p[d_j I_j] = 1$ and (1).

Following a similar argument, (5) holds for the class of cases, generalizing (ii), defined by:

(iii) $w_j = d_j q_j$, where $q_j = q(\boldsymbol{x}_j)$ and $q(.)$ is an arbitrary function.

The class of weighted estimators defined by such weights is the one of primary interest in this paper and within which we consider minimising the variance of linear combinations of the elements of $\hat{\boldsymbol{\theta}}_w$. In regular cases (Thompson, 1997, p.212), the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_w$ is

$$var_{mp}(\hat{\boldsymbol{\theta}}_w) = \boldsymbol{J}(\boldsymbol{\theta})^{-1} var_{mp}\{\sum_{j=1}^{N} w_j I_j \boldsymbol{\phi}_j\} \boldsymbol{J}(\boldsymbol{\theta})^{-1}, \tag{6}$$

where $\boldsymbol{J}(\boldsymbol{\theta}) = E_{mp}\{\sum_{j=1}^{N} w_j I_j \frac{\partial \boldsymbol{\phi}_j}{\partial \boldsymbol{\theta}}\}$ and, when $w_j = d_j q(\boldsymbol{x}_j)$, we can write

$$\boldsymbol{J}(\boldsymbol{\theta}) = \sum_{j=1}^{N} q_j E_m\big(\frac{\partial \boldsymbol{\phi}_j}{\partial \boldsymbol{\theta}}\big).$$

We should like to choose q_j so that the variance in (6) is minimized. For practical purposes, we consider it sufficient to minimize an approximation to this variance, since any weighted estimator in the class defined by $w_j = d_j q_j$ is consistent. We shall make a series of approximations to enable us to specify q_j in a simple and practical way. Our first approximation is of independence between units, so that the covariances between the different $d_j q_j I_j \boldsymbol{\phi}_j$ terms can be ignored. This is similar to the simplification employed by Fuller (2009, p.359) of assuming Poisson sampling. Under this approximation, we may rewrite (6) as

$$var_{mp}(\hat{\boldsymbol{\theta}}_w) \approx \boldsymbol{J}(\boldsymbol{\theta})^{-1} \{\sum_{j=1}^{N} var_{mp}(d_j q_j I_j \boldsymbol{\phi}_j)\} \boldsymbol{J}(\boldsymbol{\theta})^{-1}.$$

Furthermore, we have

$$\begin{aligned} var_{mp}(d_j q_j I_j \boldsymbol{\phi}_j) &= E_m\{var_p(d_j q_j I_j \boldsymbol{\phi}_j)\} + var_m\{E_p(d_j q_j I_j \boldsymbol{\phi}_j)\} \\ &= E_m\{(d_j - 1) q_j^2 \boldsymbol{\phi}_j \boldsymbol{\phi}_j^T\} + var_m(q_j \boldsymbol{\phi}_j) \\ &= E_m(d_j q_j^2 \boldsymbol{\phi}_j \boldsymbol{\phi}_j^T). \end{aligned}$$

Hence, when $w_j = d_j q(\boldsymbol{x}_j)$, the asymptotic covariance matrix can be expressed as

$$var_{mp}(\hat{\boldsymbol{\theta}}_w) \approx \{ \sum_{j=1}^{N} q_j E_m\big(\frac{\partial \boldsymbol{\phi}_j}{\partial \boldsymbol{\theta}}\big) \}^{-1} \sum_{j=1}^{N} q_j^2 E_m(d_j \boldsymbol{\phi}_j \boldsymbol{\phi}_j^T) \{ \sum_{j=1}^{N} q_j E_m\big(\frac{\partial \boldsymbol{\phi}_j}{\partial \boldsymbol{\theta}}\big) \}^{-1}$$

As a second simplification, we assume that $\boldsymbol{\phi}_j(Y_j; \boldsymbol{x}_j, \boldsymbol{\theta})$ is a score function so that

$$E_m(\boldsymbol{\phi}_j \boldsymbol{\phi}_j^T) = -E_m\big(\frac{\partial \boldsymbol{\phi}_j}{\partial \boldsymbol{\theta}}\big) = \boldsymbol{H}_j, \quad \text{say,}$$

and also that we have a generalized linear model so that $\boldsymbol{\theta} = \boldsymbol{\beta}$ is the vector of regression coefficients and $\boldsymbol{\phi}_j(Y_j; \boldsymbol{x}_j, \boldsymbol{\beta}) = \lambda_j(Y_j; \boldsymbol{x}_j \boldsymbol{\beta}) \boldsymbol{x}_j^T$ where $\lambda_j(.)$ is a scalar function. Then we may write $\boldsymbol{H}_j = \tau_j^2 \boldsymbol{x}_j^T \boldsymbol{x}_j$, where $\tau_j^2 = E_m(\lambda_j^2)$, $\lambda_j = \lambda_j(Y_j; \boldsymbol{x}_j \boldsymbol{\beta})$ and

$$var_{mp}(\hat{\boldsymbol{\beta}}_w) \approx \big\{ \sum_{j=1}^{N} q_j \tau_j^2 \boldsymbol{x}_j^T \boldsymbol{x}_j \big\}^{-1} \sum_{j=1}^{N} q_j^2 \nu_j \boldsymbol{x}_j^T \boldsymbol{x}_j \big\{ \sum_{j=1}^{N} q_j \tau_j^2 \boldsymbol{x}_j^T \boldsymbol{x}_j \big\}^{-1} \qquad (7)$$

where $\nu_j = E_m(d_j \lambda_j^2)$. By analogy to the Gauss-Markov Theorem, the choice of q_j which minimises the variance given by (7) of any linear combination of the elements of $\hat{\boldsymbol{\beta}}_w$ is

$$q_j^{opt} = q^{opt}(\boldsymbol{x}_j) \propto \tau_j^2 / \nu_j = E_m(\lambda_j^2 | \boldsymbol{x}_j) / E_m(d_j \lambda_j^2 | \boldsymbol{x}_j). \qquad (8)$$

This generalizes an argument used by Fuller (2009, pp. 359, 360) for the special case of heteroskedastic normal error linear regression with $k = 1$. We

make the conditioning on $\boldsymbol{x}_j$ explicit on the right hand side of (8) to be clear that q_j^{opt} depends on $\boldsymbol{x}_j$. The quantity on the right hand side of (8) is not observed, but is estimable from auxiliary regressions of $\hat{\lambda}_j^2$ and $d_j\hat{\lambda}_j^2$ on $\boldsymbol{x}_j$, where $\hat{\lambda}_j = \lambda_j(Y_j; \boldsymbol{x}_j\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. These regressions and the estimation of $\boldsymbol{\beta}$ could, for example, employ design-weighted estimation. We do not pursue this idea further in this paper, however. Rather, we make the further approximation that d_j is uncorrelated with λ_j^2 (given $\boldsymbol{x}_j$) so that expression (8) simplifies to

$$q_j^{opt} \propto 1/E_m(d_j|\boldsymbol{x}_j). \tag{9}$$

Some justification for this approximation will be given in Section 4 in the context of our application. The form of weighting in (9) is similar to the semi-parametric approach of Pfeffermann and Sverchkov (1999), although they propose to take $q_j \propto 1/E_{mp}(d_j|\boldsymbol{x}_j, I_j = 1)$.

Expression (9) can be yet further simplified by replacing $E_m(d_j|\boldsymbol{x}_j)$ by the conditional expectation of d_j given a subset of the explanatory variables making up $\boldsymbol{x}_j$. In practice, there is often just a single explanatory factor which is the dominant source of variation in the π_j. In our cross-national application, this is the *country factor*, i.e. a categorical variable with categories corresponding to countries. In this case, we may simply set q_j to be equal within the categories of this factor and, for a given category, to be the reciprocal of the design-weighted mean of d_j for sample units in the category. In the more general setting $E_m(d_j|\boldsymbol{x}_j)$ in (9) may be estimated by design-weighted regression.

Turning to standard error estimation and assuming that the finite population correction can be ignored, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_w$ in (6)

may be estimated consistently (Fuller, 2009, Theorem 2.2.1) by $\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{V}}\hat{\boldsymbol{J}}^{-1}$, where $\hat{\boldsymbol{J}} = \sum_{j=1}^{N} w_j I_j \partial\boldsymbol{\phi}_j/\partial\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_w$ and $\hat{\boldsymbol{V}}$ is a consistent estimator of the covariance matrix of the Horvitz-Thompson estimator $\sum_{j=1}^{N} d_j I_j \boldsymbol{u_j}$, where $\boldsymbol{u_j} = d_j^{-1} w_j \boldsymbol{\phi}_j(y_j, \boldsymbol{x}_j, \hat{\boldsymbol{\theta}}_w)$ is treated as a fixed vector of variables. Thus, standard errors can be produced using a standard approach for fixed survey weights, ignoring the fact that the weights have been modified.

In all of this section it has been assumed that the model in (1) is correct. Godambe and Thompson (1986) argue that $\boldsymbol{\theta}_U$ defined by (2) may still be of interest even if the model fails. They note that $\hat{\boldsymbol{\theta}}_w$ is still design-consistent for $\boldsymbol{\theta}_U$ even under model misspecification when design weights are used and they also demonstrate a minimum variance property of design-weighting under the constraint that $\boldsymbol{\theta}_U$ is estimated consistently. However, if model (1) fails, then $\boldsymbol{\theta}_U$ is not the only finite population parameter which can be defined and may be of interest. An arbitrary finite population parameter can be defined as the value of the parameter $\boldsymbol{\theta}$ which indexes that version of the model which represents a good fit to the finite population values $\{y_j, \boldsymbol{x}_j; j\epsilon U\}$ according to a specified criterion, whatever the truth of the model. The criterion for $\boldsymbol{\theta}_U$ is that (2) holds. To consider an alternative finite population parameter, suppose the weights are of the form $w_j = d_j q_j$. Then $\hat{\boldsymbol{\theta}}_w$ is consistent for the solution of

$$\sum_{j=1}^{N} q_j \boldsymbol{\phi}_j(Y_j; \boldsymbol{x}_j, \boldsymbol{\theta}) = \mathbf{0}, \tag{10}$$

assumed to exist uniquely, and this will not in general be the same as $\boldsymbol{\theta}_U$. Nevertheless, it is a finite population parameter with (10) as the criterion and it is defined even if model (1) fails. We suggest that whether the solution of

(2) or (10) is of scientific interest depends on the application and we shall return to this matter at the end of Section 4.

3 CROSS-NATIONAL AND MULTIPOPULATION SURVEYS

In Section 4 we shall apply the modified weighting approach introduced in Section 2 to the analysis of data from a particular cross-national survey. In this section, we discuss the kinds of applications which this survey analysis is intended to exemplify.

The basic setting we consider is one where regression analysis is applied to data from several countries and where country is an explanatory variable in the model, that is binary indicators of the different countries form part of $\boldsymbol{x}_i$. Such analyses have various purposes. One is to enable a quasi-experimental evaluation of the relative impacts of different policies which are adopted in different countries, for example to compare the effects of different national tobacco control policies on smoker behaviour (Thompson et al., 2006). Another is to enable replication of some phenomenon of interest such as an election in our application in Section 4. Cross-national analyses may have broad comparative purposes, enabling the comparison of regression relationships across different national settings. See Thompson (2008) for further discussion of the purposes of cross-national surveys and Hox et al. (2010) for alternative methods for the analysis of cross-national data, such as multilevel modelling with country treated as a level.

Sampling designs for cross-national surveys are often subject to considerable variation in inclusion probabilities between countries. Since their prin-

cipal aim is often comparative, it is common to set a minimum sample size or effective sample size in each country in order to achieve adequate precision of each national estimate (Häder and Gabler, 2003; Lynn et al., 2007). Kish (1994, p.181) notes "population sizes of countries have a tremendous, thousand-fold range; whereas sample sizes tend to be made more constant in order to obtain similar errors for national means". As a consequence, sampling fractions can vary greatly between countries and the country factor (defined in the previous section) may be viewed as an important 'design variable'.

It should be noted that this source of variation in sampling fractions between countries may also arise in national surveys between subnational groups, such as regions or jurisdictions with policy differences. For example, estimates are required at the province level for many surveys of the Canadian population conducted by Statistics Canada. In order to achieve sufficient precision for each province estimate, it is common that sampling fractions vary considerably between provinces. Thus, the cross-national focus in this paper may be viewed as just a special case of what Kish (1994) refers to as 'multipopulation' surveys.

A further general feature of the sampling design of cross-national surveys of relevance to our application is that the design variables leading to unequal inclusion probabilities within countries will often differ between countries, since quite different kinds of sampling frames and field practices can be employed (Kish, 1994; Häder and Gabler, 2003). As a consequence, it will often be impractical as well as potentially scientifically inappropriate to include these design variables as explanatory variables in a pooled regression

analysis of data across countries. Yet, it is still feasible that these design variables may be associated with the outcome variable within the corresponding countries, after controlling for the explanatory variables which are included. Hence sample selection bias could occur if within-country design variables are excluded from the model. Some adjustment, such as weighting, may therefore often be needed.

To apply the weight modification method introduced in Section 2, we consider constructing a weight $d_j q_j$, where d_j is the design weight and q_j is a function of the country factor (defined in the previous section), assumed to be included as an explanatory variable. Following the discussion below equation (9), we set $q_j = 1/\bar{d}_{c(j)}$ where $\bar{d}_c$ denotes the design-weighted mean of the design weights within country c and $c(j)$ denotes the country to which unit j belongs. We refer to $d_{Wj} = d_j q_j = d_j/\bar{d}_{c(j)}$ as the within-country weight and $d_{Bj} = \bar{d}_{c(j)}$ as the between-country weight and note that the product of these two weights is the design weight.

4 EUROPEAN SOCIAL SURVEY APPLICATION: VOTER TURNOUT IN EUROPE

4.1 Overview of Analysis and its Objectives

We now investigate a logistic regression analysis of data from the European Social Survey (ESS) in order to illustrate the methodological issues discussed in Sections 2 and 3. We seek to reproduce an analysis presented in Fieldhouse et al. (2007). We refer to that paper for the political science context and justification of the models considered. The only deliberate difference in the

models considered here is that we treat the countries as fixed in our regression analysis, whereas Fieldhouse et al. (2007) employ a multilevel model in which countries are represented by random effects.

The outcome variable of interest y_j is whether the respondent voted in the last national election held in their country. Electoral turnout is in decline in Europe, as elsewhere, and political scientists are interested in factors associated with turnout. Low turnout is of particular concern amongst young people and we analyse data for those aged 18-24, as in Fieldhouse et al. (2007).

We consider what Fieldhouse et al. (2007) describe as a rational choice model and which they find accounts well for country-level variation in comparison with two other models. The model includes two sets of explanatory variables: variables of political science interest, which reflect voting behaviour as a rational choice, and basic demographic control variables. The variables of political science interest include five scales derived from principal component analysis of individual questions: the first two principal components of questions measuring the extent to which respondents think they can understand and influence politics, termed 'political efficacy 1 and 2'; the first two principal components of questions measuring respondents' feelings of civic duty, termed 'system benefits 1 and 2'; and the first principal component of questions relating to satisfaction with the economy, government competence, democracy, education and health services, termed 'collective benefits'. In addition there is a measure of partisanship and of the closeness of the contest (the difference in vote between the first and second placed parties in the election) and an interaction between these two variables. The five demographic

control variables consist of gender; whether the respondent belongs to an ethnic minority; whether the respondent was born in the country; whether the respondent has a partner; and whether the respondent has a child.

4.2 Data, Sampling and Design Weights

The data come from the first round of the ESS in 2002/03. Although the survey covered 22 countries, we use data from just 20 countries because of concern about the comparability of survey questions pertinent to our analysis. In France, the scale used to measure the extent to which the respondent believed that politicians are more interested in votes than in people's opinions differed from that used in other countries. Similarly, in Ireland, the question measuring the respondent's satisfaction with the national government referred instead to the Dáil (parliament). These variables were used in the generation of our political efficacy and collective benefits scale respectively and we therefore chose to omit data from France and Ireland from our analysis.

Of the 42,359 respondents to the first round of the ESS, 3549 from France and Ireland were removed leaving 38,310. Of these, 3787 were aged 18-24 and, of these, 3109 were eligible to vote. Among these 3109 respondents 488 had missing data for at least one of the variables used in the model. This left 2621 complete records which were used in our analysis. There was no obvious systematic reason for the item nonresponse. There are many variables underlying our analysis, since several variables in our model are principal components of other variables and all of the attitudinal variables had some amount of missing data. There was much less missing data in the socio-

demographic control variables and, perhaps most importantly, there were only 5 missing values for the outcome variable on whether the respondent voted. The variable with the most severe item nonresponse was satisfaction with government with 130 missing values. We did not attempt to take account of any potential biasing effects from this item nonresponse nor from unit nonresponse in the ESS.

Samples in the ESS were selected independently in different countries, each with a minimum effective sample size of 1500 (or 800 if the country had fewer than 2 million inhabitants). The sampling schemes varied according to the sampling frames available. In Denmark, Finland and Sweden population registers could be used to select individuals by single stage sampling with equal probability. In other countries different forms of stratified multi-stage sampling were employed with varying kinds of strata, multi-stage units and numbers of stages (European Social Survey, 2004). In these countries, the probabilities of inclusion of individuals could vary for a number of reasons. Just one individual was typically selected per sampled household or address so that inclusion probabilities would vary by the number of eligible individuals in the household or address. Other reasons for the inclusion probabilities to vary included: differential sampling fractions between strata; the use of probability proportional to size sampling in some countries to select multistage units; and different sampling procedures for individuals with and without listed telephone numbers.

As expected from the discussion in Section 3, the resulting design weights show considerable variation between countries. Standardized values of the between-country weights $d_{Bj} = \bar{d}_{c(j)}$ vary between 0.02 for Luxembourg to

4.07 for Italy, with the size of the weight being strongly related to population size. In addition, the distributions of within-country weights $d_{Wj} = d_j/\bar{d}_{c(j)}$ show considerable variation between countries. In those countries employing equal probability sampling, the distribution shows no dispersion, with all within-country weights equal to unity. In other countries there is appreciable dispersion. For example, in Austria, the within-country weights range from 0.34 to 4.00 with an inter-quartile range of 0.94.

4.3 Results of Analyses

The results of fitting the logistic regression model, referred to in Section 4.1, are presented in Table 1. Pseudo-maximum likelihood estimation is employed, solving (4) with $\boldsymbol{\phi}_j$ defined by (3), where the weights w_j are either constant (unweighted estimation) or are design weights. The unweighted estimates are broadly similar to those in Fieldhouse et al. (2007, Table 4), although there are some differences, as may be expected since: (i) data from two fewer countries were used; (ii) there may be some differences in the codings of the variables and the definitions of the principal components. Recoding of variables and computation of principal components were required for our analysis and precise details of how these steps were undertaken by Fieldhouse et al. (2007) were unavailable; and (iii) we treated the countries as fixed in the analysis whereas Fieldhouse et al. (2007) used a multilevel model.

Standard errors were calculated as described in Section 2, which corresponds to the approach of Roberts et al. (1987) for logistic regression. Because of the lack of availability of primary sampling unit identifiers, it

Table 1: Estimated coefficients of logistic regression with standard errors

Variable	unweighted	s.e.	design-weighted	s.e.
political efficacy 1	0.27	0.05	0.25	0.08
political efficacy 2	0.15	0.05	0.13	0.08
closeness of contest (%)	-0.03	0.01	-0.06	0.01
Partisanship	0.41	0.13	0.48	0.22
closeness*partnership	0.03	0.01	0.01	0.02
collective benefits	0.02	0.05	0.04	0.08
system benefits 1	0.31	0.04	0.35	0.07
system benefits 2	0.03	0.04	0.11	0.07
is female	0.16	0.09	0.14	0.14
belongs to ethnic minority	-0.65	0.21	-0.31	0.36
has partner	-0.07	0.12	-0.05	0.19
has dependent child	-0.28	0.16	-0.46	0.23
born in country	-0.74	0.17	1.20	0.30

was not possible to take account of the multistage sampling. This is not a problem for some countries, where single stage sampling was used. However, since there were some countries where multistage sampling was employed, there is likely to be some degree of underestimation in the reported standard errors.

For several variables, the differences between the weighted and unweighted estimates are not large enough to lead to different substantive interpretations. For example, political efficacy 1 and system benefits 1 play an important role in the rational choice interpretation of the model but weighting has little impact on the point estimates of the coefficients of these variables. Such limited impact of weighting on coefficient estimates is not unusual in such social survey analyses.

There are, nevertheless, some variables where weighting has a more notable impact. The coefficient of 'closeness of contest' is doubled and the coefficients of some of the socio-demographic control variables, including 'ethnic

minority', 'dependent child' and 'born in country' are changed substantially. To test the significance of these observed differences, we follow the approach of DuMouchel and Duncan (1983), including interactions between the weight and the explanatory variables and testing whether the coefficients of each interaction term is zero. The F test statistic for this null hypothesis is F (14, 2607)=2.0 with a p-value of 0.015. As in the case of standard errors, no account was taken of multistage sampling when calculating this statistic and this may make the p-value misleadingly small. Putting aside this possibility, this test appears to provide evidence of a difference in the expectations of the weighted and unweighted estimators. One might conclude from this finding that, given evidence of potential bias in the unweighted estimator, the weighted estimator (or a modified weighted estimator as discussed in Section 2) is to be preferred. However, it seems more sensible first to consider alternative possible explanations for the observed significant effect of weighting, specifically whether we can identify a design variable which has induced correlation between the sample selection indicator I_j and the outcome variable y_j and which it might be scientifically reasonable to include in the model as an explanatory variable.

In Table 2 we consider four variables where there were strong weighting effects in Table 1 and compare the unweighted coefficient estimates with estimates obtained through three alternative choices of weights, d_j, d_{Wj} and d_{Bj}. For each of the variables, the between-country weighting has a very similar effect to design weighting, whereas the within-country weighting has little effect. We did repeat the F test for the whole model using just the within-country weights and the result was non-significant.

Table 2: Estimated coefficients with alternative choices of weights for variables with strong weighting effects

Variable	unweighted	d_{Wj}	d_{Bj}	d_j
closeness of contest (%)	-0.03	-0.03	-0.06	-0.06
belongs to ethnic minority	-0.65	-0.50	-0.30	-0.31
has dependent child	-0.28	-0.36	-0.41	-0.46
born in country	0.74	0.76	1.22	1.20

The conclusion we draw from these findings is that the most plausible design variable to account for the significant weighting effects is a country-level variable which accounts for different sampling fractions in different countries. This conclusion is reinforced by the observation, when computing the F test above by including interaction terms, that the most significant interaction between the design weight and an explanatory variable is for 'closeness of contest', a country-level variable. One country-level variable which we might consider adding to the model is population size since, as we observed earlier, this is strongly related to the between-country weight. However, to avoid choosing between country-level variables, we simply include the country factor in the model as a series of 19 indicator variables. It seems scientifically reasonable to include such a factor in the model, since the aim is to study variations in turnout of young people in the context of country variation. After including this factor the F test for the impact of the design weights is no longer significant. The unweighted and weighted estimates of the coefficients in this new model are presented in Table 3 (other than for the country indicator variables). We observe that some of the differences between weighted and unweighted estimates have been reduced compared to Table 1, in particular the difference for the closeness of contest variable has disappeared.

Table 3: Estimates for revised logistic regression model which includes a country factor

Variable	unweighted	s.e.	design-weighted	s.e.
political efficacy 1	0.27	0.05	0.25	0.08
political efficacy 2	0.17	0.06	0.18	0.09
closeness of contest (%)	-0.05	0.02	-0.05	0.02
partisanship	0.59	0.14	0.78	0.25
closeness*partnership	0.00	0.02	0.03	0.03
collective benefits	0.06	0.06	0.10	0.09
system benefits 1	0.31	0.04	0.31	0.07
system benefits 2	0.06	0.05	0.06	0.07
is female	0.14	0.10	0.12	0.14
belongs to ethnic minority	-0.67	0.22	-0.34	0.37
has partner	-0.05	0.13	-0.02	0.22
has dependent child	-0.22	0.16	-0.57	0.26
born in country	-0.66	0.18	1.14	0.30

There remain some apparent differences for the demographic variables but, according to the overall F test, these could be attributable to chance.

Although the hypothesis of noninformative selection is not rejected by this F test, it is still possible that the test lacks sufficient power to detect informativeness. Hence, weighting may still have the advantage of protecting against possible selection bias. A disadvantage of weighting, however, is that it inflates standard errors. The average inflation in Table 3 is a substantial 63%, equivalent to reducing the effective sample size by a factor close to 3. This suggests using a modified weight, as introduced in Section 2. We present standard errors for four choices of weights in Table 4. We see that the variance inflation arises primarily because of the between-country variation in the weights. We propose therefore to use the within-country weights d_{Wj},

viewed as modified weights, as discussed in Section 3. These weights do not suffer from the variance inflation of the design weights. They do, however, provide protection against possible selection bias arising from the within-country variation in the sample inclusion probabilities. It makes little sense to attempt to protect against such bias by respecifying the model further. This would require including design variables which account for the within-country variation in inclusion probabilities, but different sampling designs are employed in different countries and there are no obvious common design variables which it makes sense to introduce into the model across the whole sample.

The fact that the standard errors for within-country weighting are so close to those without weighting in Table 4 suggests that the series of approximations made in section 2 resulted in little loss of optimality. In particular, consider the approximation that the d_j and λ_j^2 are uncorrelated, where $\lambda_j = Y_j - E(Y_j|\boldsymbol{x}_j)$ is a simple residual in the case of logistic regression. In this application, the main possible source of correlation between d_j and the squared residuals will be via variation in the country means of the squared residuals, but since $E(Y_j|\boldsymbol{x}_j)$ is around the middle of the [0,1] interval for all counties, the country means of the squared residuals should be fairly stable and little correlation between d_j and λ_j^2 is expected.

We did compare estimates based upon our within-country weights and those based on the semi-parametric approach of Pfeffermann and Sverchkov (1999) and found little difference. For example, only two of the estimated coefficients weighted by d_{Wj} in Table 4 change in the second decimal place. The coefficient 0.05 for political efficacy 1 becomes 0.06 for the Pfeffermann

Table 4: Standard errors for revised logistic regression model with alternative choices of weights

Variable	unweighted	d_{Wj}	d_{Bj}	d_j
political efficacy 1	0.05	0.05	0.08	0.08
political efficacy 2	0.06	0.06	0.09	0.09
closeness of contest (%)	0.02	0.02	0.02	0.02
Partisanship	0.14	0.15	0.23	0.25
closeness*partnership	0.02	0.02	0.02	0.03
collective benefits	0.06	0.06	0.09	0.09
system benefits 1	0.04	0.05	0.06	0.07
system benefits 2	0.05	0.05	0.07	0.07
is female	0.10	0.10	0.14	0.14
belongs to ethnic minority	0.22	0.25	0.35	0.37
has partner	0.13	0.14	0.20	0.22
has dependent child	0.16	0.18	0.24	0.26
born in country	0.18	0.20	0.26	0.30

and Sverchkov approach and the coefficient 0.25 of 'belongs to an ethnic minority' becomes 0.24. The Pfeffermann and Sverkov approach has the practical advantage that it corresponds to the way that ESS weights are released, that is they are standardized within country by the unweighted mean of the design weights so that they sum to the country sample sizes. To construct our within-country weights requires a further step in which the design-weighted means of the design weights are calculated.

The use of just the within-country weights does assume that the model controls for differential selection across countries and this raises the question of the effect of weighting under model misspecification, as discussed by Godambe and Thompson (1986). As formalized by the solutions of (2) and (10), the weighted estimators estimate the logistic regression models which provide one of two types of best fit. The design-weighted estimator estimates the model which provides the best fit across all individuals in the population,

that is the importance it attaches to the different countries is proportional to their population sizes, denoted N_c (c.f. Thompson, 2008, p.135). In contrast, the within-country weighted estimator estimates the model which attaches importance to the different countries in proportion to $N_c/\bar{d}_c$, in the notation of Section 3. To a first approximation, this is equivalent to attaching equal importance to different countries. Which of these approaches to best fit is more appropriate? This is debatable, but we suggest that attaching equal weight to different countries may be more appropriate in this study where the aim is to learn about patterns of turnout through the process of comparing the experiences of 20 elections. To attach importance to the countries in proportion to their populations would essentially imply that, for the purpose of such pooled cross-national analyses, it was a waste of time to collect data about elections in small countries like Luxembourg since their data would have a negligible effect on the results.

To explore the effect of the choice of weights on model fit, we divided the observations in each country into weighted quintile groups $k = 1, \ldots, 5$, according to the values of the probability $\hat{p}_{wj}$ that $Y_j = 1$, predicted using the model estimated with weighting method w, following the approach of Graubard et al. (1997). For each country c, weight w and quintile group k we computed p_{wck}, the weighted observed proportion with $y_j = 1$, and $\hat{p}_{wck}$, the weighted mean of the $\hat{p}_{wj}$. As a simple measure of fit in country c of the model estimated using weights w, we computed $A_{wc} = \sum_{k=1}^{5} |p_{wck} - \hat{p}_{wck}|/5$. Since sample size varies quite considerably between countries (see Table 5) and smaller sample sizes may tend to inflate A_{wc} irrespective of the validity

of the model, we also computed

$$X^2_{wc} = \sum_{k=1}^{5} \frac{(p_{wck} - \hat{p}_{wck})^2}{\hat{p}_{wck}(1 - \hat{p}_{wck})/n_{wck}},$$

based on the test statistic of Hosmer and Lemeshow (1980), where n_{wck} is the sample size in the quintile group. The statistics A_{wc} and X^2_{wc} are designed primarily for comparative purposes, but, as a very crude test of fit we may compare X^2_{wc} against critical values of a chi-squared distribution with 5 degrees of freedom. This is very crude because it takes no account of the sampling variation in the fitted values $\hat{p}_{wj}$ nor of the weighting. Alternative approaches which take account of the weighting have been proposed (Graubard et al., 1997; Archer et al., 2007) but these would require extension to assess fit in subpopulations, such as countries in our case.

Values of A_{wc} and X^2_{wc} are presented in Table 5 by country, for the design weights and the within-country weights. The rows are ordered by the value of A_{wc} for the design weights. From the earlier discussion about model misspecification, we might anticipate that the fit for countries with high between-country weights, such as Italy and Germany, might be better with the design weights. This is true for Germany when countries are ranked by either the A_{wc} and X^2_{wc} measures, but not for Italy. Conversely, we might expect the country with the smallest between-country weight, Luxembourg, to fit better with the within-country weights than the design weights. This is true for X^2_{wc} , but not for A_{wc}, with respect to which Luxembourg has the worst fit with either choice of weights. Another pattern which might be anticipated from the discussion of model misspecification is that goodness of fit may show greater dispersion between countries for the design weights

Table 5: Goodness-of-fit of Model in Different Countries for two Weighting Methods and two Measures of Fit with Rankings in Parentheses

		design weights		within-country weights	
Country	sample size	A_{wc}	X^2_{wc}	A_{wc}	X^2_{wc}
Poland	260	0.02 (1)	0.79 (1)	0.03 (3)	1.80 (4)
Denmark	97	0.04 (2)	1.51 (2)	0.03 (2)	1.34 (2)
Germany	208	0.04 (3)	2.42 (8)	0.05 (9)	3.81 (11)
United Kingdom	112	0.04 (4)	2.00 (4)	0.03 (1)	0.65 (1)
Slovenia	138	0.05 (5)	1.68 (3)	0.09 (16)	6.52 (15)
Sweden	149	0.05 (6)	4.51 (13)	0.05 (8)	4.86 (13)
Italy	84	0.05 (7)	3.55 (10)	0.04 (4)	2.08 (5)
Switzerland	58	0.06 (8)	2.15 (6)	0.06 (11)	2.64 (9)
Israel	224	0.06 (9)	4.89 (14)	0.08 (12)	9.94 (18)
Spain	90	0.06 (10)	2.31 (7)	0.05 (7)	1.38 (3)
Finland	102	0.06 (11)	2.52 (9)	0.08 (13)	4.47 (12)
Hungary	129	0.06 (12)	3.76 (12)	0.06 (10)	3.61 (10)
Netherlands	116	0.07 (13)	6.07 (15)	0.04 (5)	2.57 (7)
Portugal	92	0.08 (14)	3.58 (11)	0.08 (14)	5.27 (14)
Czech Republic	52	0.08 (15)	2.01 (5)	0.10 (17)	2.58 (8)
Greece	156	0.08 (16)	6.74 (16)	0.05 (6)	2.10 (6)
Belgium	149	0.08 (17)	9.62 (17)	0.10 (19)	14.57 (12)
Norway	169	0.09 (18)	10.20 (18)	0.09 (15)	8.12 (16)
Austria	160	0.10 (19)	12.63 (20)	0.10 (18)	11.04 (19)
Luxembourg	76	0.14 (20)	10.49 (19)	0.11 (20)	8.19 (17)

than for the within-country weights. There seems little evidence of this. The standard deviation of A_{wc} between countries is almost identical for the two types of weights. For X^2_{wc} it is in fact higher for the within-country weights. The fact that we do not observe clearly the effects expected from misspecification may simply mean that the model is adequately specified. The 95% critical value for a chi squared with 5 d.f. is 11.1 and, whilst recognizing that this is a very crude approximation to its null distribution, X^2_{wc} only exceeds this value for 1 of the 20 countries, just as would be anticipated if the model were true.

5 DISCUSSION

Standard statistical methods for regression analysis remain valid when sample units have been selected according to the values of $\boldsymbol{x}$. Such sample selection is common with survey data, as illustrated by the cross-national application in this paper, where $\boldsymbol{x}$ includes country identifiers and sample inclusion probabilities vary considerably by country. The problem is that, even though $\boldsymbol{x}$ may be the dominant source of variation in the inclusion probabilities, there may be some residual variation which is associated with y and thus could lead to bias if standard methods are employed. The option of applying full design weighting may be heavy-handed when the residual variation and resulting bias are likely to be small, especially given the potential serious inflation of standard errors. In this paper we have discussed the more modest option of applying modified weights which still correct for bias arising from such residual variation but which avoid such serious inflation of standard errors. The modification of the weights may also protect against

the model-fitting being dominated by a small number of large countries.

Survey weighting is not the only approach that has been proposed to correct for for bias from unequal selection probabilities. See, for example, Thompson (1997, Ch.6), Chambers (2003), Pfeffermann (2011) and Scott and Wild (2011) for some alternative approaches, including a variety of likelihood-based methods. We have not considered such alternatives in this paper, however. One justification is that survey weighting has appeal to practitioners, given the wide availability of software for data from complex survey designs. Moreover, the estimation of standard errors of the estimators obtained with modified weights can follow standard methods developed for use with design weights (Fuller, 2009, Section 6.3.2).

We end with Thompson's (2008, p.137) cautionary remark about cross-national analysis that "an analysis which pools data across countries should be adopted with caution. For such an analysis to be appropriate, the model structure (the regression equation and its variables) should be correct for all countries, and the assumption of common parameters should be supported by theory and observation". She discusses a number of alternative ways of proceeding with such an analysis, including the possibility, not considered in this paper, of representing country differences by random effects.

6 ACKNOWLEDGEMENTS

We are grateful to Mark Tranmer for supplying code used to obtain the results in Fieldhouse et al. (2007). Research was supported by the Economic and Social Research Council.

References

Archer, K.J., Lemeshow, S. and Hosmer, D.W. (2007). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics and Data Analysis*, **51**, 4450-4464.

Chambers, R.L. (2003). Introduction to Part A. In Chambers, R.L. and Skinner, C.J. (Eds.) *Analysis of Survey Data*, Wiley, Chichester, 13-28.

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

European Social Survey (2004). ESS Round 1 2002/2003. Technical Report Edition 2, http://ess.nsd.uib.no/ess/round1/.

Fieldhouse, E., Tranmer, M., & Russell, A. (2007). Something about young people or something about elections? Electoral participation of young people in Europe: evidence from a multilevel analysis of the European Social Survey. *European Journal of Political Research*, **46**, 797-822.

Fuller, W.A. (2009) *Sampling Statistics*, Wiley, Hoboken.

Godambe, V.P. & Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, **54**, 127-138.

Godambe, V.P. & Thompson, M.E. (2009) Estimating functions and survey sampling. In D. Pfeffermann & C.R.Rao (Eds.) *Sample Surveys: Inference*

and Analysis, Handbook in Statistics Volume 29B, Elsevier, Amsterdam, 83-101.

Graubard, B.I., Korn, E.L. and Midthune, D. (1997) Testing goodness-of-fit for logistic regression with survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 170-174.

Häder, S. & Gabler, S. (2003). Sampling and estimation. In Harkness, J.A., Van de Vijver, F.J.R & Mohler, P.Ph. (Eds.) *Cross-Cultural Survey Methods*, Wiley, Hoboken, 117-134.

Hosmer, D.W. and Lemeshow, S. (1980) Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics*, **A9**, 1043-1069.

Hox, J.J., de Leeuw, E.D. & Brinkhuis, M.J.S. (2010). Analysis models for comparative surveys. In Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.Ph., Pennell B-E. & Smith T.W. (Eds.) *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, Wiley, Hoboken, 395-418.

Kish, L. (1994). Multipopulation survey designs: five types with seven shared aspects. *International Statistical Review*, **62**, 167-186.

Lynn, P., Häder, S., Gabler, S. & Laaksonen, S. (2007). Methods for achieving equivalence of samples in cross-national surveys: the European Social Survey experience, *Journal of Official Statistics*, **23**, 107-124.

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. & Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhya B*, **61**, 166-186.

Roberts, G., Rao, J.N.K. & Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74**, 1-12.

Scott, A.J. & Wild, C.J. (2011) Fitting regression models with response-biased samples. *Canadian Journal of Statistics*, **39**, 519-536.

Thompson, M.E. (1997). *Theory of Sample Surveys*, Chapman and Hall, London.

Thompson, M.E. (2008). International surveys: motives and methodologies. *Survey Methodology*, **34**, 131-141.

Thompson, M.E., Fong, G.T., Hammond, D., Boudreau, C., Driezen, P., Hyland, A., Borland, R., Cummings, K.M., Hastings, G.B., Siahpush, M., Mackintosh, A.M. & Laux. F.L. (2006). Methods of the International Tobacco Control (ITC) four country survey, *Tobacco Control*, **15** (Suppl. III), iii12-iii18.