The Use of Survey Weights in Regression Modelling

Chris Skinner

London School of Economics and Political Science

NIASRA workshop, Sydney, February 2017

Some History

1950s $\rightarrow$

- weights as general tool for point estimation
- whether to weight in regression - disciplinary division (survey statistics, biostatistics, econometrics)

1970s,1980s $\rightarrow$

- basic inferential methods, including e.g. logistic regression

1990s, 2000s $\rightarrow$

- wider availability of software and textbooks
- disciplinary divisions blurred

Drivers of Modern Developments

- increased nonresponse
- greater variability of weights and their impact
- weights for non-probability sampling
- new forms of auxiliary information
- new modelling settings

Pros and Cons of Weighting

Pros

- to avoid bias from informative sampling, when inclusion probabilities π_j unequal
- to protect against model misspecification
- to make efficient use of population-level information

Cons

- variance inflation from unequal inclusion probabilities

Modifying Weights

- to retain pros, while mitigating cons
- allow weights to depend on variables included in regression model

Motivating Application: Cross-National Comparative Survey Analysis

Common for π_j to vary between countries.

"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" Kish (1994)

Outline

1. First kind of weight modification - theory
2. Application to European Social Survey
3. Second kind of weight modification - theory

Model

Population $U = \{1, 2, ..., N\}$

Regression model $f(y_j; \boldsymbol{x}_j, \boldsymbol{\beta})$

Score $\boldsymbol{u}_j(y_j; \boldsymbol{x}_j, \boldsymbol{\beta}) = \frac{\partial \log f(y_j; \boldsymbol{x}_j, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$

write $\boldsymbol{u}_j(\boldsymbol{\beta}) = \boldsymbol{u}_j(y_j; \boldsymbol{x}_j, \boldsymbol{\beta})$

Mean score model: $E_m\{\boldsymbol{u}_j(\boldsymbol{\beta})\} = \mathbf{0}$, $j = 1, 2, ..., N$

Census Parameter

$\boldsymbol{\beta}_U$ solves

$$\sum_{j=1}^{N} \boldsymbol{u}_j(\boldsymbol{\beta}) = \mathbf{0}.$$

Some argue that $\boldsymbol{\beta}_U$ of interest even if model fails.

Sampling

Sample $s \subset U$

Sample indicator variable

$$I_j = 1 \quad \text{if } j \in s$$
$$I_j = 0 \quad \text{if } j \notin s,$$

(Unweighted) Sample Estimating Equations

$\widehat{\boldsymbol{\beta}}$ solves

$$\sum_{j=1}^{N} I_j \boldsymbol{u}_j(\boldsymbol{\beta}) = \mathbf{0}.$$

design-model consistent if $E_m E_p \{ I_j \boldsymbol{u}_j(\boldsymbol{\beta}) \} = \mathbf{0}$, $j = 1, 2, ..., N$

in particular, if I_j and Y_j independent (given $\boldsymbol{x}_j$) noninformative sampling

Weighted Sample Estimating equations

$\widehat{\boldsymbol{\beta}}_w$ solves

$$\sum_{j=1}^{N} I_j w_j \boldsymbol{u}_j(\beta) = \mathbf{0}.$$

design-model consistency condition:

$$E_m E_p \{ I_j w_j \boldsymbol{u}_j(\beta) \} = \mathbf{0} \ , \ j = 1, 2, ..., N$$

Design Consistency

design consistency condition $\Rightarrow$ design-model consistency condition

$\widehat{\boldsymbol{\beta}}$ design consistent for $\boldsymbol{\beta}_U$ if

$$E_p\Big\{\sum_{j=1}^{N} I_j w_j \boldsymbol{u}_j(\boldsymbol{\beta})\Big\} = \sum_{j=1}^{N} \boldsymbol{u}_j(\boldsymbol{\beta}) \quad \text{for arbitrary } y_j \text{ and } \boldsymbol{\beta}$$

so $E_p\big(I_j w_j\big) = 1$ design consistency condition

holds if $w_j = \pi_j^{-1}$, Horvitz-Thompson (design) weight

Widening Class of Weights

But design consistency condition may not be necessary on scientific grounds. We drop it to enable us to improve efficiency.

Now require only design-model consistency condition, but do assume mean score model holds.

Class of possible weights is widened

First Kind of Weight Modification

Modification by function of covariates.

Class of modified weights $w_j = d_j q_j$, where $d_j = \pi_j^{-1}$ and $q_j = q(\boldsymbol{x}_j)$ (not sample dependent)

- meets design-model consistency condition if mean score model holds.

- will generally not meet design-consistency condition unless $q_j \equiv constant$

Optimization Problem

to determine function $q(\cdot)$ which minimises $var_{mp}(\widehat{\boldsymbol{\beta}}_w)$

$$var_{mp}(\widehat{\boldsymbol{\beta}}_w) = \mathbf{J}(\boldsymbol{\beta})^{-1} var_{mp}\left\{ \sum_{j=1}^{N} w_j I_j \boldsymbol{u}_j(\boldsymbol{\beta}) \right\} \mathbf{J}(\boldsymbol{\beta})^{-1}$$

where

$$\mathbf{J}(\boldsymbol{\beta}) = E_{mp}\left\{ \sum_{j=1}^{N} I_j w_j \frac{\partial \boldsymbol{u}_j(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\}$$

Approximations/Assumptions

- observations for different units are approximately independent

- generalized linear model so that $\boldsymbol{u}_j(\boldsymbol{\beta}) = e_j \boldsymbol{x}_j$

$$var_{mp}(\widehat{\boldsymbol{\beta}}_w) \;\approx$$

$$\left\{ \sum_{j=1}^{N} q_j E_m(e_j^2) \boldsymbol{x}_j \boldsymbol{x}_j^T \right\}^{-1} \sum_{j=1}^{N} q_j^2 E_m(d_j e_j^2) \boldsymbol{x}_j \boldsymbol{x}_j^T \left\{ \sum_{j=1}^{N} q_j E_m(e_j^2) \boldsymbol{x}_j \boldsymbol{x}_j^T \right\}^{-1}$$

(Approximately) Optimal Solution

$$q_j \;\propto\; \frac{E_m\left(e_j^2 \mid \boldsymbol{x}_j\right)}{E_m\left(d_j e_j^2 \mid \boldsymbol{x}_j\right)}$$

Equivalent to Fuller (2009, Sect 6.3.2) for linear regression model

Requires fitting of model to $E_m\left(d_j e_j^2 \mid \boldsymbol{x}_j\right)$

Estimating $q(\cdot)$

e_j is scaled version of residual $y_j - E_m(Y_j \mid \boldsymbol{x}_j)$

As a first approximation, suppose d_j and e_j^2 are uncorrelated and set $q_j = 1/E_m(d_j \mid \boldsymbol{x}_j)$

And thus $w_j = d_j q_j = d_j / E_m(d_j \mid \boldsymbol{x}_j)$

Design weight standardized for its dependence on $\boldsymbol{x}_j$

Will discuss estimation of $E_m(d_j \mid \boldsymbol{x}_j)$ later

Application: Voter Turnout in Europe

- European Social Survey Round 1 - 2002

- subsample of 2621 people aged 18-24 in 19 European countries, providing data on variables relating to political interest and civic duty

- analysis similar to Fieldhouse, Tranmer and Russell (2007) *European J. Political Research*

Logistic Regression Analysis

$$
\begin{aligned}
y &= 1, \quad \text{if voted in last national election in country} \\
&= 0, \quad \text{otherwise}
\end{aligned}
$$

$\boldsymbol{x}$ variables for rational choice model, including

political efficacy - principal components of questions measuring extent to which respondents think they can understand and influence politics

system benefits - principal components of respondent's feelings of civic duty

Estimated Coefficients of Logistic Regression

variable	unweighted	s.e.	weighted	s.e
political efficacy 1	**0.27**	0.05	**0.25**	0.08
political efficacy 2	**0.15**	0.05	0.13	0.09
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

Estimated Coefficients of Logistic Regression

variable	unweighted	s.e.	weighted	s.e
political efficacy 1	**0.27**	0.05	**0.25**	0.08
political efficacy 2	**0.15**	0.05	0.13	0.09
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

Test for Informative Sampling

Augment model by adding interactions between weight and $\boldsymbol{x}$ variables.

Test whether coefficients of new variables are all 0

Wald test F (14, 2607)=2.0, p=0.015.

Conclude significant effect of weighting.

DuMouchel and Duncan (1983), Fuller (2009)

Test as Diagnostic for Misspecification

- informative sampling may indicate omitted variable in model
- are there variables which explain π_j which could be included in model?

Sample Selection

- separate sampling in different countries
- sampling schemes vary according to different sampling frames, e.g.
  - lists of residents
  - lists of households
  - lists of addresses

Respecified Model

- major variation in design weights between countries
- introduce country dummy variables in model
- scientifically reasonable, since aim is to study variations in turnout of young people in context of country variation

Variable	Initial model unweighted	Initial model weighted	New model unweighted	New model weighted
political efficacy 1	**0.27**	**0.25**	**0.27**	**0.28**
political efficacy 2	**0.15**	0.13	**0.17**	**0.18**
closeness of contest (%)	**0.03**	**0.06**	**0.05**	**0.05**
Partisanship	**0.41**	**0.48**	**0.59**	**0.78**
closeness*partnership	**0.03**	0.01	0.00	0.03
collective benefits	0.02	0.04	0.06	0.10
system benefits 1	**0.31**	**0.35**	**0.31**	**0.31**
system benefits 2	0.03	0.11	0.06	0.06
is female	0.16	0.14	0.14	0.12
belongs to ethnic minority	**-0.65**	-0.31	**-0.67**	-0.34
has partner	-0.07	-0.05	-0.05	0.02
has dependent child	-0.28	**-0.46**	-0.22	**-0.57**
born in country	**0.74**	**1.20**	**0.66**	**1.14**

Results for Respecified Model

- effect of weighting reduced
- no longer significant
- little sense in attempting to respecify model further to include within country design variables, since such variation in designs between countries

Standard Errors

Variable	Unweighted	Weighted
political efficacy 1	0.05	0.08
political efficacy 2	0.06	0.09
closeness of contest (%)	0.02	0.02
Partisanship	0.14	0.27
closeness*partnership	0.02	0.02
collective benefits	0.06	0.09
system benefits 1	0.04	0.07
system benefits 2	0.05	0.07
is female	0.10	0.14
belongs to ethnic minority	0.22	0.37
has partner	0.13	0.22
has dependent child	0.16	0.26
born in country	0.18	0.30

Within Country Weights as Modified Weights

- Recall modified weight: $w_j = d_j q_j = d_j / E_m(d_j \mid \boldsymbol{x}_j)$
- Let $\boldsymbol{x}^{(1)}$ be vector of country dummy variables, subvector of $\boldsymbol{x}$
- Simplify $E_m(d_j \mid \boldsymbol{x}_j)$ to $E_m(d_j \mid \boldsymbol{x}_j^{(1)})$
- Weighting with $d_j / E_m(d_j \mid \boldsymbol{x}_j^{(1)})$ still consistent - achieves most of efficiency gain?
- Estimate $E_m(d_j \mid \boldsymbol{x}_j^{(1)})$ by design-weighted mean of design weights $\bar{d}_{c(j)}$ for country $c(j)$
- $d_j / \bar{d}_{c(j)}$ is within country weight

Standard Errors

Variable	Unweighted	Within country	Design weights
political efficacy 1	0.05	0.05	0.08
political efficacy 2	0.06	0.06	0.09
closeness of contest (%)	0.02	0.02	0.02
Partisanship	0.14	0.15	0.27
closeness*partnership	0.02	0.02	0.02
collective benefits	0.06	0.06	0.09
system benefits 1	0.04	0.05	0.07
system benefits 2	0.05	0.05	0.07
is female	0.10	0.10	0.14
belongs to ethnic minority	0.22	0.24	0.37
has partner	0.13	0.14	0.22
has dependent child	0.16	0.18	0.26
born in country	0.18	0.20	0.30

Within Country Weighting

- consistent estimation, provided dependence of y on countries correctly specified in model

- protects against bias from informative selection within countries (non-significant test may lack power)

- avoids inflation of standard errors

- may give consistent estimation for more suitable pseudo-parameter under model misspecification?

Return to Theory

Widening Class of Weights Further

Still assume design-model consistency condition and mean score model holds. Want to minimize

$$var_{mp}(\widehat{\boldsymbol{\beta}}_w) = \mathbf{J}^{-1} var_{mp}\left\{ \sum_{j=1}^{N} w_j I_j \boldsymbol{u}_j(\boldsymbol{\beta}) \right\} \mathbf{J}^{-1}$$

Write

$$var\left\{ \sum_{j=1}^{N} w_j I_j \boldsymbol{u}_j(\boldsymbol{\beta}) \right\} =$$

$$var\left\{ \sum_{j=1}^{N} E(w_j \mid y_j, \boldsymbol{x}_j, I_j) I_j \boldsymbol{u}_j(\boldsymbol{\beta}) \right\} + E\left\{ var(\sum_{j=1}^{N} w_j I_j \boldsymbol{u}_j(\boldsymbol{\beta}) \mid y_j, \boldsymbol{x}_j, I_j) \right\}$$

Second Kind of Weight Modification

Consistency unaffected if $E(w_j \mid y_j, \boldsymbol{x}_j, I_j)I_j = E(d_j \mid y_j, \boldsymbol{x}_j, I_j)I_j$
Variance minimized if $var(w_j \mid y_j, \boldsymbol{x}_j, I_j) = 0$
Achieved by setting

$$
\begin{aligned}
w_j &= E(d_j \mid y_j, \boldsymbol{x}_j, I_j = 1)q(\boldsymbol{x}_j) \\
&\equiv \widetilde{d}_j q(\boldsymbol{x}_j)
\end{aligned}
$$

modify d_j to $\widetilde{d}_j$, smooths noise in weights unrelated to y_j given $\boldsymbol{x}_j$

c.f Beaumont (2008) weight smoothing
Chaudhuri et al. (2010), Pfeffermann (2011) (conditional)
empirical likelihood

Weight Smoothing in Descriptive Surveys

$$T = \sum_j y_j$$

Horvitz-Thompson estimator $\widehat{T}_{HT} = \sum_j I_j d_j y_j$
Smoothed Horvitz-Thompson estimator $\widehat{T}_{SHT} = \sum_j I_j \widetilde{d}_j y_j$
where $\widetilde{d}_j \equiv E_{mp}(d_j \mid y_j, I_j = 1)$

$$E_{mp}(\widehat{T}_{HT}) = E_{mp}(\widehat{T}_{SHT}) = T$$

$$V_{mp}(\widehat{T}_{HT}) \geq V_{mp}(\widehat{T}_{SHT})$$

Weight Smoothing in Regression

$$w_j = \widetilde{d}_j q(\boldsymbol{x}_j)$$

where $\widetilde{d}_j \equiv E(d_j \mid y_j, \boldsymbol{x}_j, I_j = 1)$

optimal choice

$$q_j \;\propto\; \frac{E_m\left(e_j^2 \mid \boldsymbol{x}_j\right)}{E_m\left(\widetilde{d}_j e_j^2 \mid \boldsymbol{x}_j\right)}$$

Auxiliary Weight Model(s)

$\widetilde{d}_j = E(d_j \mid y_j, \boldsymbol{x}_j, I_j = 1) = \widetilde{d}(y_j, \boldsymbol{x}_j; \phi)$

E.g. $\widetilde{d}(y_j, \boldsymbol{x}_j; \phi) = 1 + exp(-\phi_1 \boldsymbol{x}_j - \phi_2 y_j)$

q_j depends on $E(\widetilde{d}_j e_j^2 \mid \boldsymbol{x}_j)$

Iterative estimation of β and ϕ Kim and Skinner (2013)

See also Beaumont (2008), Pfeffermann (2011)

Variance Estimation

- first weight modification
  - replacement of $q(\boldsymbol{x}_j)$ by $q(\boldsymbol{x}_j; \widehat{\alpha})$ does not affect asymptotic variance
  - consistent variance estimation by treating weights $d_j q_j$ as fixed

- second weight modification
  - replacement of $\widetilde{d}_j$ by $\widetilde{d}(y_j, \boldsymbol{x}_j; \widehat{\phi})$ does affect asymptotic variance
  - variance estimation does need take account of error in estimating ϕ
  - Kim and Skinner (2013) give linearization variance estimator

Summary

Both weight modifications offer efficiency gains, assuming (mean score) model holds.

First weight modification offers:

- consistency under misspecification of $q(x)$
- no need to modify variance estimation approach

Smoothing requires:

- correct specification of $E(d_j \mid y_j, \boldsymbol{x}_j, I_j = 1)$ for consistency
- modification of variance estimation approach

In both cases, consider whether implied pseudoparameter under misspecification is scientifically reasonable.

References 1

Kim, J.K. and Skinner, C.J. (2013) Weighting in survey analysis under informative sampling. *Biometrika*, **100**, 385-398.

Skinner, C.J. and Mason, B. (2012) Weighting in the regression analysis of survey data with a cross-national application. *Canadian Journal of Statistics*, **39**, 519-536.

References 2

Beaumont, J.-F. (2008) *Biometrika*

Chaudhuri, S., Handcock, M. and Rendall, M. (2010) Working paper

DuMouchel, W. and Duncan, G. (1983) *JASA*

Fieldhouse, E. et al. (2007) *Eur. J. Political Research*

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

Godambe, V. and Thompson, M. (1986) *Int. Statist.Rev.*

Kish, L. (1994) *Int. Statist. Rev.*

Kish, L. and Frankel, M. (1974) *J.R.S.S.B*

Pfeffermann, D. (2011) *Survey Methodology*

Pfeffermann, D. and Sverchkov, M. (1999) *Sankhya, B*

Pfeffermann, D. and Sverchkov, M. (2003) in Chambers, R. and Skinner, C. eds. *Analysis of Survey Data*. Wiley.

Scott, A. and Wild, C. (2002) *J.R.S.S.B*

Scott, A. and Wild, C. (2003) in Chambers, R. and Skinner, C. eds. *Analysis of Survey Data*. Wiley.

Skinner, C. (2003) in Chambers, R. and Skinner, C. eds. *Analysis of Survey Data*. Wiley.