APPROXIMATIONS TO *b** IN THE PREDICTION OF DESIGN EFFECTS DUE TO CLUSTERING

Peter Lynn and Siegfried Gabler

ISER Working Papers
Number 2004-07

Institute for Social and Economic Research

The Institute for Social and Economic Research (ISER) specialises in the production and analysis of longitudinal data. ISER incorporates the following centres:

- ESRC Research Centre on Micro-social Change. Established in 1989 to identify, explain, model and forecast social change in Britain at the individual and household level, the Centre specialises in research using longitudinal data.

- ESRC UK Longitudinal Studies Centre. This national resource centre was established in October 1999 to promote the use of longitudinal data and to develop a strategy for the future of large-scale longitudinal surveys. It is responsible for the British Household Panel Survey (BHPS) and for the ESRC's interest in the National Child Development Study and the 1970 British Cohort Study

- European Centre for Analysis in the Social Sciences. ECASS is an interdisciplinary research centre which hosts major research programmes and helps researchers from the EU gain access to longitudinal data and cross-national data sets from all over Europe.

The British Household Panel Survey is one of the main instruments for measuring social change in Britain. The BHPS comprises a nationally representative sample of around 5,500 households and over 10,000 individuals who are reinterviewed each year. The questionnaire includes a constant core of items accompanied by a variable component in order to provide for the collection of initial conditions data and to allow for the subsequent inclusion of emerging research and policy concerns.

Among the main projects in ISER's research programme are: the labour market and the division of domestic responsibilities; changes in families and households; modelling households' labour force behaviour; wealth, well-being and socio-economic structure; resource distribution in the household; and modelling techniques and survey methodology.

BHPS data provide the academic community, policymakers and private sector with a unique national resource and allow for comparative research with similar studies in Europe, the United States and Canada.

BHPS data are available from the Data Archive at the University of Essex
http://www.data-archive.ac.uk

Further information about the BHPS and other longitudinal surveys can be obtained by telephoning +44 (0) 1206 873543.

The support of both the Economic and Social Research Council (ESRC) and the University of Essex is gratefully acknowledged. The work reported in this paper is part of the scientific programme of the Institute for Social and Economic Research.

Acknowledgement: This research was carried out while the first author was a Guest Professor at the Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim, Germany. The second author is director of the Statistics Department at ZUMA.

Readers wishing to cite this document are asked to use the following form of words:

> **Lynn, Peter and Gabler, Siegfried (June 2004) 'Approximations to *b** in the estimation of design effects due to clustering, *Working Papers of the Institute for Social and Economic Research*, paper 2004-07. Colchester: University of Essex.**

For an on-line version of this working paper and others in the series, please visit the Institute's website at: http://www.iser.essex.ac.uk/pubs/workpaps/

Institute for Social and Economic Research
University of Essex
Wivenhoe Park
Colchester
Essex
CO4 3SQ UK
Telephone: +44 (0) 1206 872957
Fax: +44 (0) 1206 873151
E-mail: iser@essex.ac.uk
Website: http://www.iser.essex.ac.uk

ABSTRACT

Kish's well-known expression for the design effect due to clustering is often used to inform sample design, using an approximation such as $\bar{b}$ in place of b. If the design involves either weighting or variation in cluster sample sizes, this can be a poor approximation. In this article we discuss the sensitivity of the approximation to departures from the implicit assumptions and propose an alternative approximation.

Key words: Complex sample design, design effect, intracluster correlation coefficient, selection probabilities, weighting

1. Introduction: Alternative Functions of Cluster Size

Kish (1965) used an expression for the design effect (variance inflation factor) due to sample clustering, $deff = 1 + (b-1)\rho$, where b is the number of observations in each cluster (primary sampling unit) and ρ is the intracluster correlation coefficient. This expression is well-known, is taught on courses on sampling theory, and is used by survey practitioners in designing and evaluating samples.

The expression holds when there is no variation in cluster sample size and the design is equal-probability (self-weighting). We can express these two criteria formally:

$$b_c = b \quad \forall c \tag{1}$$

where c = 1, … , C denote the clusters, and

$$w_i = w \quad \forall i \tag{2}$$

where i = 1, …, I denote the weighting classes, with w_i the associated design weights.

However, most surveys involve departures from (1) and (2). In the general case, i.e. removing restrictions (1) and (2), Gabler, Häder and Lahiri (1999) showed that under an appropriate model, $deff_c = 1 + (b^* - 1)\rho$, where

$$b^* = \sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_i b_{ci}\right)^2 \Bigg/ \sum_{i=1}^{I} w_i^2 b_i = \sum_{c=1}^{C}\left(\sum_{j=1}^{b_c} w_{cj}\right)^2 \Bigg/ \sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj}^2 \tag{3}$$

and b_{ci} is the number of observations in weighting class *i* in cluster *c*, $b_i = \sum_{c=1}^{C} b_{ci}$ (we have changed the notation from that of Gabler, Häder and Lahiri (1999), to provide consistency) and w_{cj} is the weight associated with the j^{th} observation in cluster *c*, *j* = 1, …., b_c.

The quantity b^* can be calculated from survey microdata, provided the design weight and cluster membership is known for each observation. However, at the sample design stage it is not clear how b^* can be predicted. Gabler, Häder and Lahiri (1999) interpreted Kish's b as a form of weighted average cluster size:

$$\bar{b}_w = \sum_{c=1}^{C} b_c \left(\sum_{i=1}^{I} w_i^2 b_{ci} \right) \Bigg/ \sum_{c=1}^{C} \sum_{i=1}^{I} w_i^2 b_{ci} = \sum_{c=1}^{C} \left(b_c \sum_{j=1}^{b_c} w_{cj}^2 \right) \Bigg/ \sum_{c=1}^{C} \sum_{j=1}^{b_c} w_{cj}^2 \qquad (4)$$

where b_c is the number of observations in cluster *c*, $b_c = \sum_{i=1}^{I} b_{ci}$. However, (4) is no easier than (3) to predict at the sample design stage. A simpler interpretation, perhaps commonly used in sample design, is the unweighted mean cluster size:

$$\bar{b} = \sum_{c=1}^{C} b_c \Big/ C = m/C \qquad (5)$$

It is much easier to predict $\bar{b}$ at the sample design stage than either $\bar{b}_w$ or b^*, as it requires knowledge only of the total number of observations, *m*, and total number of clusters, *C*.

2. Relationship Between b^*, $\overline{b}_w$ and $\overline{b}$ under Alternative Assumptions

Let

$$\overline{w}_c = \frac{1}{b_c}\sum_{j=1}^{b_c} w_{cj} = \sum_{i=1}^{I} w_i \frac{b_{ci}}{b_c}, \quad Cov\left(b_c, b_c \overline{w}_c^2\right) = \frac{1}{C}\sum_{c=1}^{C} b_c^2 \overline{w}_c^2 - \frac{m}{C^2}\sum_{c=1}^{C} b_c \overline{w}_c^2 \text{ and}$$

$$Var\left(w_{cj}\right) = \frac{1}{b_c}\sum_{j=1}^{b_c}\left(w_{cj} - \overline{w}_c\right)^2 = \sum_{i=1}^{I}\frac{b_{ci}}{b_c}\left(w_i - \overline{w}_c\right)^2 \quad \forall c\,.$$

Then

$$b^* = \frac{C \times Cov\left(b_c, b_c \overline{w}_c^2\right) + \overline{b}\sum_{c=1}^{C} b_c \overline{w}_c^2}{\sum_{c=1}^{C} b_c Var\left(w_{cj}\right) + \sum_{c=1}^{C} b_c \overline{w}_c^2}\,. \tag{6}$$

If (1) holds, then (6) becomes:

$$b^* = \overline{b}\left(\frac{\sum_{c=1}^{C} \overline{w}_c^2}{\sum_{c=1}^{C} Var\left(w_{cj}\right) + \sum_{c=1}^{C} \overline{w}_c^2}\right). \tag{7}$$

So, in that circumstance, $b^* \le \overline{b}$. If, additionally, weights are equal within clusters, viz:

$$w_{cj} = w_c \quad \forall j \in c \tag{8}$$

then $b^* = \overline{b}$.

If (8) holds, but not (1), then

$$b^* \geq \bar{b} \text{ if and only if } Cov\left(b_c, b_c \overline{w}_c^2\right) \geq 0 \text{ since } b^* - \bar{b} = \frac{C \times Cov\left(b_c, b_c \overline{w}_c^2\right)}{\sum_{c=1}^{C} b_c \overline{w}_c^2}.$$

The covariance would be negative only if small cluster sizes coincide with large average weights within the clusters and *vice versa*. In section 4 below, we observe that this did not occur in any country on round 1 of the European Social Survey. Furthermore, from (3) and (4), we have:

$$b^* = \bar{b}_w = \sum_{c=1}^{C} \left(w_c b_c\right)^2 \Big/ \sum_{c=1}^{C} w_c^2 b_c \ . \qquad (9)$$

If we additionally impose the restriction (1), then we have the obvious result $b^* = \bar{b}_w = \bar{b} = b_c \ \ \forall c$.

The result in (9) would apply to surveys where the only variation in selection probabilities was due to disproportionate sampling between domains that did not cross-cut clusters. A common example would involve disproportionate stratification by region, with PSUs consisting of geographical areas hierarchical to regions.

A practical relaxation of the restriction on the variation in weights is:

$$b_{ci} = b_c \left(\frac{b_i}{m}\right) \ \ \forall i, c \, . \qquad (10)$$

In other words, we allow variation in weights within clusters, but we constrain the weights to have the same relative frequency distribution in each cluster, i.e. the

means and the variances of the weights within clusters do not depend on the clusters.

Now, (3) simplifies as follows:

$$b^* = \sum_{c=1}^{C}\left(\sum_{i=1}^{I} w_i b_c \frac{b_i}{m}\right)^2 \Bigg/ \sum_{i=1}^{I} w_i^2 b_i = \sum_{c=1}^{C}\left(b_c^2\left(\sum_{i=1}^{I} w_i b_i\right)^2\right) \Bigg/ m^2 \sum_{i=1}^{I} w_i^2 b_i$$

$$= \frac{\left(\sum_{i=1}^{I} w_i b_i\right)^2}{\sum_{i=1}^{I} w_i^2 b_i} \frac{\sum_{c=1}^{C} b_c^2}{m^2}. \qquad (11)$$

Note that $\frac{\left(\sum_{i=1}^{I} w_i b_i\right)^2}{\sum_{i=1}^{I} w_i^2 b_i} = \frac{m}{\left(1+c_w^2\right)}$, where c_w^2 is the squared coefficient of variation, across all observations, of the weights. Also, $\frac{\sum_{c=1}^{C} b_c^2}{m^2} = \frac{1+c_b^2}{C}$, where c_b^2 is the squared coefficient of variation, across all clusters, of the cluster sample sizes. Thus, (11) becomes:

$$b^* = \frac{m}{\left(1+c_w^2\right)} \frac{\left(1+c_b^2\right)}{C} = \bar{b} \frac{\left(1+c_b^2\right)}{\left(1+c_w^2\right)} = \tilde{b}, \text{ say.} \qquad (12)$$

So, $\bar{b}$ will underestimate b^* if $c_b^2 > c_w^2$ and *vice versa*. In particular, if $w_{cj} = w \ \forall c, j$ and $c_b^2 > 0$, then $b^* > \bar{b}$. The greater the variation in b_c, the greater the extent to which $\bar{b}$ will under-estimate b^*.

Assumption (10) will rarely hold exactly, but this result might be useful in situations where the distribution of weights is expected to be similar across clusters. An example might be address-based samples where one person is selected per address. If the distribution of the number of persons per address is approximately constant across PSUs (in the population), then the distribution of weights will vary across clusters in the sample only due to sampling variation and disproportionate nonresponse. (The effect of this could, of course, be substantial if cluster sample sizes are small.)

If no restriction is imposed on the variation in weights, but $Var(w_{cj}) > 0$ for at least one *c*, then, from (6),

$$b^* \geq \bar{b} \text{ if and only if } \zeta = \frac{C^2 Cov(b_c, b_c \bar{w}_c^2)}{m \sum_{c=1}^{C} b_c Var(w_{cj})} \geq 1. \quad (13)$$

If (10) holds, then $\zeta = \frac{c_b^2}{c_w^2}$.

3. Implications for Sample Design

Expression (12) suggests that b^* may be predicted by predicting the relative magnitudes of c_b^2 and c_w^2. However, this result applies to a special situation, where

$$Cov(w_{cj}, b_c) = \frac{1}{m} \sum_{c=1}^{C} \sum_{j=1}^{b_c} (w_{cj} - \bar{w})(b_c - \bar{\bar{b}})$$

$$= \frac{1}{m}\sum_{c=1}^{C}\left(b_c - \bar{\bar{b}}\right)\left(\sum_{i=1}^{I} w_i b_{ci} - b_c \bar{w}\right)$$

$$= \overset{from(10)}{} \frac{1}{m^2}\sum_{c=1}^{C}\left(b_c - \bar{\bar{b}}\right) b_c \left(\sum_{i=1}^{I} w_i b_i - m\bar{w}\right)$$

$$= 0$$

where

$$\bar{w} = \frac{1}{m}\sum_{c=1}^{C}\sum_{j=1}^{b_c} w_{cj} = \frac{1}{m}\sum_{c=1}^{C} b_c \bar{w}_c \text{ and}$$

$$\bar{\bar{b}} = \frac{1}{m}\sum_{c=1}^{C}\sum_{j=1}^{b_c} b_c = \frac{1}{m}\sum_{c=1}^{C} b_c^2 = \frac{m}{C}\left(1 + c_b^2\right)$$

When this covariance is expected to be small, it may be appropriate to predict b^* thus:

$$\hat{b}^* = \hat{\bar{b}} = \bar{b}\,\frac{\left(1 + c_b^2\right)}{\left(1 + c_w^2\right)}. \qquad (14)$$

Both coefficients of variation can be estimated from knowledge of the proposed sample design. In the following section, we investigate sensitivity of predictions obtained in this way to assumption (10) using real data from different sample designs with $Cov\left(w_{cj}, b_c\right) > 0$.

4. Example: European Social Survey

The European Social Survey (ESS) is a cross-national survey for which great efforts have been made to achieve approximate functional equivalence in sample design between participating nations (Lynn *et al* 2004). Nevertheless, there is considerable variety in the types of design used, primarily due to variation in the nature of available frames and in local objectives, such as a desire for sub-national analysis which may lead to disproportionate stratification by domain. We use here data from the first round of the ESS, for which fieldwork was carried out in 2002-2003. Of the 22 participating nations, 17 had a clustered sample design. Of these, two had not yet provided useable sample data at the time of writing. In Table 1 we present the sample values of b^*, $\bar{b}$, c_b^2, c_w^2, $\tilde{\tilde{b}}$, $\left|\tilde{\tilde{b}} - b^*\right|$, $\left|\bar{b} - b^*\right|$, $Corr\left(w_{cj}, b_c\right)$ and ζ for the remaining 15. Note that the United Kingdom and Poland both had a 2-domain design with the sample clustered only in one domain, namely Great Britain (i.e. excluding Northern Ireland) and less densely-populated areas (i.e. all except the largest 42 towns) respectively. Figures presented in table 1 relate only to the clustered domain.

Table 1: Sample values of b^*, $\bar{b}$, c_b^2, c_w^2, $\tilde{\tilde{b}}$, $\left|\tilde{\tilde{b}} - b^*\right|$, $\left|\bar{b} - b^*\right|$, $Corr\left(w_{cj}, b_c\right)$ and ζ, for 15 surveys

Country		b^*	$\bar{b}$	c_b^2	c_w^2	$\tilde{\tilde{b}}$	$\left\|\tilde{\tilde{b}} - b^*\right\|$	$\left\|\bar{b} - b^*\right\|$	$Corr\left(w_{cj}, b_c\right)$	ζ
Austria	AT	6.49	7.08	0.08	0.25	6.15	0.34	0.58	0.0036	0.4549
Belgium	BE	6.56	5.79	0.13	0.00	6.56	0.00	0.77	.	.
Switzerland	CH	8.83	9.23	0.12	0.21	8.50	0.34	0.40	0.0223	0.7060
Czech Republic	CZ	2.94	2.70	0.24	0.25	2.68	0.26	0.24	0.0225	1.7350
Germany	DE	18.85	18.13	0.07	0.11	17.42	1.43	0.72	-0.2287	.
Spain	ES	4.96	5.04	0.17	0.22	4.80	0.15	0.08	-0.0767	0.8757
Great Britain	GB	11.11	12.27	0.08	0.22	10.90	0.21	1.16	0.0114	0.4198
Greece	GR	5.47	5.86	0.09	0.22	5.25	0.22	0.39	-0.0280	0.5207
Hungary	HU	8.68	8.18	0.06	0.00	8.68	0.00	0.50	.	.
Ireland	IE	12.09	11.18	0.13	0.04	12.05	0.05	0.91	0.0006	3.1054
Israel	IL	11.79	12.82	0.12	0.56	9.27	2.53	1.02	-0.1271	0.4401
Italy	IT	10.98	10.87	0.26	0.16	11.80	0.83	0.10	-0.5589	1.3018
Norway	NO	30.03	18.85	1.32	0.44	30.44	0.42	11.18	-0.1146	.
Poland (rural)	PL	10.07	9.45	0.06	0.01	9.88	0.19	0.62	0.2923	.
Slovenia	SI	10.76	10.13	0.06	0.00	10.76	0.00	0.63	.	.

From (12), we would expect to observe $\bar{b} > b^*$ when $c_w^2 > c_b^2$. A common sample design for which this inequality can be anticipated is one where, a) the selected cluster sample size is constant, so variation in b_c will be limited to that caused by differential non-response; and b) the samples are equal-probability samples of

addresses, with subsequent random selection of one person per address, leading to variation in design weights reflecting the variation in household size. There are six nations with sample designs of this type (AT, CH, ES, GB, GR, IL). It is indeed the case that for all of these nations, $\zeta < 1$ and $\bar{b} > b^*$. Furthermore, for 5 of these 6 nations (AT, CH, ES, GB, GR, h = 1, …, 5) we might expect (10) to be a reasonable approximation as the only variation in weights is that due to selection within a household/address. For these, we might expect $\hat{\hat{b}}$ to perform better than $\bar{b}$. Indeed, $\left|\hat{\hat{b}} - b^*\right| < \left|\bar{b} - b^*\right|$ for 4 of the 5, and $\dfrac{\sum_{h=1}^{5}\left|\hat{\hat{b}} - b^*\right|}{\sum_{h=1}^{5}\left|\bar{b} - b^*\right|} = 0.48$. The one nation where $\hat{\hat{b}}$ would not provide an improvement is Spain and this is to be expected as $\bar{b}$ is small. Small cluster sample sizes leave them relatively more susceptible to the effects of nonresponse and also sampling variance, which will lead to violation of (10). In Israel, there was a further source of variation in design weights as there was disproportionate stratification by geographical areas. This too causes violation of (10), so we would not expect $\hat{\hat{b}}$ necessarily to provide an improvement on $\bar{b}$ as a predictor of b^*.

Of the nations where $c_b^2 < c_w^2$, there is only one (CZ) for which $\bar{b} < b^*$ and $\zeta > 1$. This is also the nation with the smallest value of $\bar{b}$. When cluster sample sizes are particularly small, *deff* will be small and the choice between estimators of b^* may be less important.

There are five nations where sample units were individuals selected with equal probabilities (within clusters) from population registers (BE, DE, HU, PL, SI). In this case (8) (and, therefore, (10)) holds strictly, so we have $\bar{b} < b^*$. For three of these nations (BE, HU, SI) the sample is equal-probability, so we observe $\overleftrightarrow{b} = b^*$. It is clear that $\hat{\overleftrightarrow{b}}$ is superior to $\bar{b}$ for equal-probability samples. For Germany and Poland, there is some variation in design weights between clusters (but not within). This variation is modest in Poland, and $\left|\overleftrightarrow{b} - b^*\right| < \left|\bar{b} - b^*\right|$, but the same is not true in Germany, where the ex-East Germany was sampled at a considerably higher rate than the ex-West Germany.

The Norwegian sample design was the only one that resulted in considerable variation in cluster sample sizes at the selection stage. The dramatic impact of this on $\bar{b} - b^*$ can clearly be seen. Again, this is a situation in which $\hat{\overleftrightarrow{b}}$ is likely to be preferable to $\bar{b}$ as a predictor of b^*.

The designs in Ireland and Italy both involved selecting addresses from the electoral registers with probability proportional to number of electors and then selecting one resident at random from each selected address. Such designs are not equal-probability, but are likely to result in considerably less variation in design weights than the address-based sample designs discussed earlier (Lynn and Pisati, 2004). In both these cases, $c_w^2 < c_b^2$, the difference being greater in the case of Italy where some cluster sample sizes (in the largest municipalities) were considerably larger than the others (in Ireland, all were equal at the selection stage). Aside from the Czech Republic, these are the only two nations with $\zeta > 1$.

5. Conclusion

To aid prediction of the design effect due to clustering, we believe that $\hat{\tilde{b}}$ is likely to be a better choice than $\hat{\bar{b}}$ as a predictor of b^* in situations where it can reasonably be expected that (10) will approximately hold. This includes, but is not restricted to, the following common types of sample design:

- Equal-probability designs where cluster sample sizes vary by design;

- Equal-probability designs where clusters do not vary by design but are likely to vary due to nonresponse;

- Address-based samples where one person is selected at each address, there is no other significant source of variation in selection probabilities, and cluster sizes do not vary by design.

References

GABLER, S., HÄDER, S., and LAHIRI, P. (1999). A Model Based Justification of Kish's Formula for Design Effects for Weighting and Clustering. *Survey Methodology* 25, 105-106.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

LYNN, P., HÄDER, S., GABLER, S., and LAAKSONEN, S. (2004, forthcoming) Methods for Achieving Equivalence of Samples in Cross-National Surveys: the European Social Survey Experience.

LYNN, P. and PISATI, M. (2004, forthcoming) Improving the quality of sample design for social surveys in Italy: lessons from the European Social Survey.