Survey Methodology, June 1995
Vol. 21, No. 1, pp. 45–54
Statistics Canada

Modelling Net Undercoverage in the 1991 Canadian Census

PETER DICK[1]

ABSTRACT

In 1991, Statistics Canada for the first time adjusted the Population Estimates Program for undercoverage in the 1991 Census. The Census coverage studies provided reliable estimates of undercoverage at the provincial level and for national estimates of large age - sex domains. However, the population series required estimates of undercoverage for age - sex domains within each province and territory. Since the direct survey estimates for some of these small domains had large standard errors due to the small sample size in the domain, small area modelling techniques were needed. In order to incorporate the varying degrees of reliability of the direct survey estimates, a regression model utilizing an Empirical Bayes methodology was used to estimate the undercoverage in small domains. A raking ratio procedure was then applied to the undercoverage estimates to preserve consistency with the marginal direct survey estimates. The results of this modelling process are shown along with the estimated reduction in standard errors.

KEY WORDS: Small area; Empirical Bayes; Undercoverage.

1. INTRODUCTION AND BACKGROUND

The Census of Canada is conducted every five years; one of its objectives is to provide the Population Estimates Program with accurate baseline counts of the number of persons by age and sex within each province and territory. Unfortunately, not all eligible persons are correctly enumerated by the Census. As part of the evaluation of the Census, Statistics Canada estimates, through two sample surveys, the net number of persons missed by the Census. The estimates are from the Reverse Record Check Study, which estimates the gross number of persons missed by the Census, and the Overcoverage Study, which estimates persons double counted or erroneously included in the final Census count. When combined the figures estimate the net number of people missed by the Census.

These surveys were designed to produce reliable direct estimates for large areas, such as provinces, and for large domains, such as age - sex combinations at the national level. However, the Population Estimates Program requires estimates of missed persons for single year of age for both sexes for each province. However using the direct survey estimate would result in individual estimates having unacceptably high standard errors due to insufficient sample in the small domain. One approach to reducing the variance of the small domain estimates would be to borrow strength from related domains. This approach leads to creating an explicit model for the small domain that can be used to predict the net missed persons in that domain.

The result of modelling the small domain estimates is to produce a series of estimates with a smaller Mean Square Error than the direct estimate. However, as opposed to the direct survey estimate which is design unbiased, the modelling approach will introduce a bias for each estimate. Thus modelling the small domain estimates implies that a trade off is required between reducing the variance of each estimate and the bias introduced through the modelling process. One approach to ensuring that the more reliable direct survey estimates are utilized is to introduce an Empirical Bayes model. This procedure creates an estimate that is a combination of a model estimate and the direct survey estimate weighted by their respective variances. It is an Empirical Bayes estimate instead of a Bayes estimate because underlying parameters are first estimated, then these estimated parameters are considered known in later calculations. Note that since the individual sampling variances are used in the estimation, a more precise direct estimate would contribute much more to the final Empirical Bayes estimate than a similar estimate with low precision. This ensures that the model does not dominate estimates that are already considered reliable. It is also possible to approach this estimation problem through a Hierarchical Bayes methodology: details on this method can be found in Datta, *et al.* (1992). Ghosh and Rao (1994) give an appraisal of both the Hierarchical Bayes and Empirical Bayes approaches to small area estimation.

Outside of Canada, two different approaches to smoothing the Census undercoverage have been described in the literature. In the United States, the net undercoverage in the 1990 American Census was evaluated by means of the Post Enumeration Survey (Hogan 1992). Initially, it was planned to multiply the US Census counts by adjustment factors (the ratio of true population over the enumerated population) for 1,392 *a priori* defined post strata.

[1] Peter Dick, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

These estimated adjustment factors would then be used to adjust the Census count for missed persons. Since some of these 1,392 estimated adjustment factors had high standard errors, it was proposed to smooth the direct estimates through an Empirical Bayes regression model, similar to one proposed by Ericksen and Kadane (1985), and then to rake the smoothed estimates to agree with direct estimates for large geographic regions. However, this approach was criticized by Freedman and Navidi (1992). Eventually, the United States Department of Commerce, the U.S. Census Bureau's parent agency, decided not to proceed with adjusting the Census counts for under-enumeration in July 1991. Consideration was also given in the United States to adjusting the post Censal population estimates for undercoverage in the Census, but the Department of Commerce also rejected this adjustment.

The Australians use a different method than the Americans for estimating the domain totals. Choi, Steel and Skinner (1988) describe a methodology that incorporates the estimates of net undercoverage of the Census into the population estimates but leaves the actual Census counts as enumerated. The under enumeration is estimated through a Post Enumeration Survey (PES) and demographic analysis. The small domain estimates are produced by raking the Census age counts for each sex to the PES estimates for national age/sex totals and part of State/Territory/sex totals.

The procedure proposed for the 1991 Canadian Census combines some of the elements of both the American and Australian approaches. As in the American procedure, a model is postulated for the underlying true adjustment factors and another model is postulated for relating the direct survey estimates to the true underlying adjustment factors. Through Empirical Bayes estimation, a new smoothed adjustment factor is estimated that will have a lower MSE then the direct survey estimate. These smoothed adjustment factors are then converted into estimates of missed persons. The Australian method for constraining the resulting estimates to known marginal totals is then adopted. These final raked estimates are used as the base for the small domain estimates of missed persons. In turn, these estimates are adjusted to account for known demographic principles (See Michalowski 1993). Details on the technical criteria for adjustment of the population estimates can be found in Royce (1992).

This paper is organized as follows. In Section 2, some background information on the two sample surveys is described and the basic Empirical Bayes model is presented. Assumptions and limitations of the model are also discussed and the estimation of the parameters is briefly discussed. In Section 3, the explanatory variables used in the regression model are presented and the model building process is described. The final model is presented and the results displayed. Section 4 presents a discussion on the rationale behind constraining the Empirical Bayes estimates to reliable marginal totals. The final adjusted estimates are then presented. Finally, Section 5 presents some conclusions and topics for further study.

2. MODEL FOR THE ADJUSTMENT FACTORS

2.1 Background and Notation

The model for the adjustment factors requires input data. The actual data originates with two coverage studies: the Reverse Record Check (RRC) and the Overcoverage Study (OCS). The RRC is used to estimate the number of persons missed by the Census while the OCS is used to estimate the number of persons erroneously included in the Census count. These surveys are designed to give reliable estimates of net undercoverage for all provinces, some of the larger metropolitan areas and for some large national domains, such as males aged 20 to 24. Since the surveys are independent, it can be assumed that the variance of net missed persons will be the sum of the two estimated variances from the RRC and the OCS. Further details on these studies can be found in Germain and Julien (1993) and the 1991 Census Technical Report - Coverage (Statistics Canada 1994).

The domains of interest can be defined by partitioning the sample into $p = 1, 2, \ldots, P$ provinces/territories and $a = 1, 2, \ldots, A$ age - sex groups, hence a total of $A \times P$ domains require estimates. Let C_i be the number of persons in the i-th province - age domain enumerated in the Census and T_i be the true population of the same domain. The net number of persons missed in the i-th cell is $M_i = T_i - C_i$. The adjustment factor, Θ_i, is the ratio of the true population in a domain over the Census count, while the undercoverage rate, U_i, the unit that is usually reported in the releases from the coverage studies, is the ratio of missed persons over the true population.

The true adjustment factors, Θ_i, which are the variables that we wish to estimate, can be written as:

$$\Theta_i = \frac{T_i}{C_i} = \frac{M_i + C_i}{C_i}.$$

Undercoverage rates (U_i) which are usually reported in the releases from Statistics Canada, are related to the adjustment factors through the relationship

$$U_i = M_i(M_i + C_i)^{-1} = 1 - \Theta_i^{-1}.$$

In the modelling of the adjustment factors, the creation of ultimate domains is required. These domains are those at which the actual direct survey estimates of the adjustment factors will be produced. There must be an estimate for each province (10) and territory (2), so immediately P is fixed at 12. The age groups were fixed at 4 to create

national estimates that have acceptably low standard errors. These age groups are defined for male and female as follows: 0 to 19 years of age; 20 to 29 years of age; 30 to 44 years of age; and 45 years and older. In total there are $12 \times 8 = 96$ direct survey estimates of adjustment factors that have to be fitted into the Empirical Bayes model. Each domain requires, besides the direct estimate of the adjustment factor, an associated estimate of the sampling variance.

2.2 Model and Assumptions

The basic model for the undercount is composed of two distinct parts. The first part describes how the direct survey estimates are related to the true underlying adjustment factors, while the second part models the relationship between the true adjustment factors and a set of explanatory variables. Since the parameters in the regression model are estimated by first estimating the parameters of an assumed underlying prior distribution and then assuming that these estimated parameters are known for any further calculation, this model is known as an Empirical Bayes model (Maritz and Lwin 1989).

The first part of the model, the sampling model, relates the observed adjustment factors to the true adjustment factors. This relationship is assumed true within each domain, and can be expressed as:

the observed adjustment factor =
the true adjustment factor + a random error.

The sampling model is written as follows:

$$F_i = \Theta_i + \epsilon_i : \epsilon_i \sim \text{Normal } (0,\sigma_i^2),$$
$$i = 1, 2, \ldots, n = A \times P,$$

where Θ_i is the true adjustment factor and ϵ_i is a random error component with a variance of σ_i^2. The assumptions underlying this model are:

(a) the sampling errors, ϵ_i, have mean zero;

(b) the sampling variances, σ_i^2, are known in each of the n domains;

(c) since the sample was selected independently within each domain, the covariance between the sampling errors ϵ_i in domain i and ϵ_j in domain j is zero; and

(d) the random errors ϵ_i are normally distributed in each domain.

Further discussion on the assumption of the known sampling variance in each domain is given below.

The second part of the model, the regression model, relates the true adjustment factors to a set of underlying explanatory variables. This model states that:

the true adjustment factor = a linear combination
of explanatory variables + a random error.

The regression model can be written as:

$$\Theta_i = X_i \beta + \delta_i : \delta_i \sim \text{Normal } (0, \tau^2),$$
$$i = 1, 2, \ldots, n = A \times P,$$

where X_i is the i-th row in X, a known $(n \times p)$ matrix of explanatory variables, β is a $(p \times 1)$ vector of unknown regression parameters and δ_i is (a different) random error with a model variance of τ^2. Underlying the system model are the following assumptions:

(a) the model errors, δ_i, have mean zero;

(b) the model variance, τ^2, is constant over all n domains;

(c) the model errors, δ_i, are normally distributed;

(d) the model errors, δ_i, are independent of sampling errors, ϵ_i;

(e) the covariance between different domains is zero (*i.e.*, $\text{Cov}(\delta_i, \delta_j) = 0$).

The problem is to use both the sampling model and the regression model to estimate Θ_i, the true adjustment factors. The conditional expectation for Θ_i given β, σ_i^2, τ^2, F_i can be determined for the joint model. Using standard arguments (Rao 1973), it can be shown that the conditional expectation of Θ_i is:

$$E(\Theta_i \mid \beta,\sigma_i^2,\tau^2,F_i) = (1 - \omega_i)X_i\beta + \omega_i F_i, \qquad (1)$$

where $\omega_i = \tau^2(\tau^2 + \sigma_i^2)^{-1}$.

Equation (1) is the basis for all the estimates that follow, although a few modifications need to be made before applying it to the data. Note that it is basically a weighted average of the direct survey estimate and the regression model estimate of the adjustment factor. Each estimate is weighted according to the precision with which it was estimated. If the sampling error, σ_i^2, is small compared to the model error, τ^2, implying that the direct survey estimate is relatively precise, then the final smoothed estimate will be mainly composed of the direct survey estimate. However, if the direct survey estimate has a large sampling variance relative to the model variance then the final smoothed estimate will be mainly constituted from the best linear unbiased predictor. The amount each estimate contributes to the final smoothed estimate is controlled by the weighting coefficient, ω_i.

Some limitations apply to interpretations that can be made about this model. First, it must be emphasized that this model is purely descriptive; it cannot be considered to be a causal model. Since the primary goal of this model is descriptive, the inferences on the regression parameters, β, while interesting are not of primary importance. Hence, the final regression model when it contains a term, say, on British Columbia renters and not Manitoba renters, is only saying that British Columbia renters explain a

significant portion of the variation in adjustment factors in British Columbia while Manitoba renters does not explain a significant portion of the variation in adjustment factors in Manitoba.

As mentioned above, the sampling variances associated with the direct survey estimates of the adjustment factors are considered known in the Empirical Bayes model. However, experience has shown that the directly estimated variances are, in fact, somewhat unstable. In order to create some stability with the estimation of these variances it is proposed to model them. If we consider the design of the two sample surveys, then, under relatively mild assumptions, Dick (1993) has shown that within each domain the variance of the estimate of missed persons is proportional to a power of the Census count. If we add in appropriate normalizing parameters, then this relationship can be written as:

$$\sigma_i^2 C_i^2 = V(M_i) = KC_i^\gamma,$$

or, as in the form of a regression equation,

$$\text{Log}(V(M_i)) = \alpha + \gamma \log(C_i) + \eta_i \quad \text{with}$$

$$\eta_i \sim N(0, \zeta^2).$$

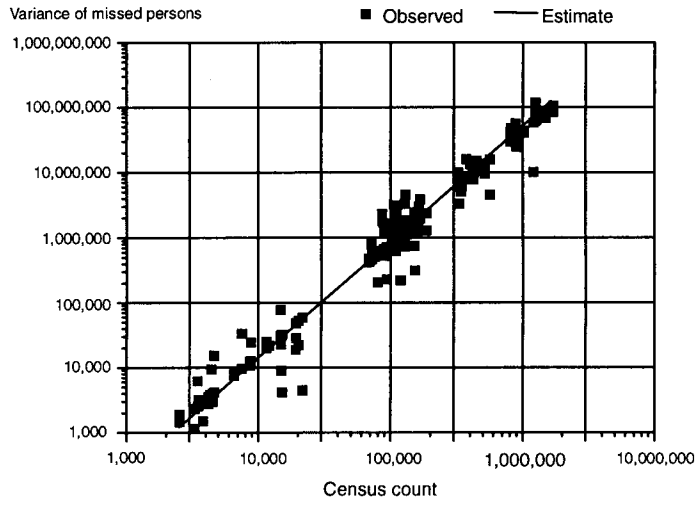

Figure 1. Observed variance *vs.* census.

This model for the sampling variance assumes that the product of the design effect and the undercoverage rate is constant within each domain. As discussed in Dick (1993), this assumption appears to be reasonable. Figure 1 shows the plot of the observed variance of missed persons calculated from the two coverage studies versus the Census count for the 96 domains. The least squares regression line was estimated as

$$\log(\hat{v}(M_i)) = -6.133 + 1.715 \log C_i$$

and is also plotted in Figure 1. A residual analysis (Dick 1993) did not detect any apparent violations of the underlying model assumptions. Since, in addition, the coefficient of determination, the R^2, is 0.943, this model was adopted for producing the sampling variances. The estimated survey variances were calculated for the adjustment factors through

$$\hat{v}(F_i) = \hat{v}(M_i)/C_i^2.$$

It will be assumed that these predicted values for the sampling variances are the actual 'known variances' required for the Empirical Bayes model.

2.3 Parameter Estimation

So far the model has been described in purely Bayesian terms: only the parameter Θ_i is considered unknown. Taking the usual Empirical Bayes approach (Maritz and Lwin 1989), we will assume that all the parameters except β, the regression parameter, are known. The conditional expectation of Θ_i with the regression parameter estimated can be written as

$$\hat{F}_i^{(\text{eb})} = E(\Theta_i \mid \hat{\beta}, \sigma_i^2, \tau^2, F_i).$$

However, in practice, the model variance, τ^2, is also unknown and must be estimated. The conditional expectation of Θ_i will now change to

$$\hat{F}_i^{(\text{eb})} = E(\Theta_i \mid \hat{\beta}, \sigma_i^2, \hat{\tau}^2, F_i),$$

where the sampling variance, σ_i^2, is still considered known.

To estimate the model variance and the regression coefficients in the Empirical Bayes model, the marginal distribution of the observed adjustment factors, $m(F_i) \sim N(x_i\beta, \tau^2 + \sigma_i^2)$, can be used. Three possible methods were examined for estimating the variance parameter, τ^2, : Method of Moments (MM) as in Fay and Herriot (1979), Maximum Likelihood (MLE) as in the PES in the United States (Hogan 1992) and Restricted Maximum Likelihood (REML).

It is well known that MLE estimation of variance components is biased downwards (Harville 1977). Underestimation of the model variance in the Empirical Bayes model would result in more reliance being placed upon the regression model instead of the direct survey estimate. This is a result we wished to avoid. In Dick (1993), it is shown that there is little difference between the estimates of the model variance from REML or MM. Since the REML has a well understood asymptotic theory, it was adopted for the estimation of the model variance in the Empirical Bayes model.

Harville gives a full account of REML estimation. The basic approach is to first estimate the regression parameter, and then to estimate the model variance from the resulting residuals instead of the actual data. If we let X^* be a matrix

of $(n - p)$ linear contrasts such that $E[X^{*t}F] = 0$, then Harville shows that the resulting (log) likelihood function, L_{reml}, when maximized with respect to the unknown model variance will give the restricted maximum likelihood estimates.

In the context of the Empirical Bayes model, Harville's approach can be described as follows. First, an initial estimate, usually zero, of the model variance, $\hat{\tau}^2_{(0)}$, is made and then the regression parameter, β, is estimated through weighted least squares:

$$\hat{\beta}_{(1)} = (X^t \hat{V}_0^{-1} X)^{-1} X^t \hat{V}_0^{-1} F, \qquad (2)$$

where $\hat{V}_0 = \text{diag}(\hat{\tau}^2_{(0)} + \sigma_i^2 : i = 1, \ldots, n)$. Using this estimate of $\hat{\beta}_{(1)}$, a new REML estimate of the model variance, $\hat{\tau}^2_{(1)}$, can be made through

$$\hat{\tau}^2_{\kappa+1} = \hat{\tau}^2_\kappa + \left(\frac{\partial L_{\text{reml}}}{\partial \tau^2}\right)[i(\tau^2)]^{-1}, \quad \kappa = 0, 1, \ldots, \qquad (3)$$

where, if we set $\hat{P}_\kappa = \hat{V}_\kappa^{-1} - \hat{V}_\kappa^{-1} X(X^t \hat{V}_\kappa^{-1} X)^{-1} X \hat{V}_\kappa^{-1}$, we have

$$\frac{\partial L_{\text{reml}}}{\partial \tau^2} = -\frac{1}{2} \text{ trace } \hat{P}_\kappa + \frac{1}{2}(F - X\hat{\beta})^t \hat{V}_\kappa^{-1} \hat{V}_\kappa^{-1} (F - X\hat{\beta})$$

and

$$i(\tau^2) = -E\left[\frac{\partial^2 L_{\text{reml}}}{\partial(\partial\tau^2)^2}\right] = \frac{1}{2} \text{ trace } (\hat{P}_\kappa^t \hat{P}_\kappa).$$

Note, upon convergence of τ^2 and β, $i(\tau^2)^{-1}$ will be the asymptotic variance of $\hat{\tau}^2$.

By iterating between (2) and (3), new estimates of τ^2 will be used to update the estimate of β, which in turn will be used to update the estimate of τ^2. The iterations then continue until a suitable convergence has been reached: in this case $((\hat{\tau}^2_{\kappa+1}/\hat{\tau}^2_\kappa) - 1) < 10^{-6}$ was used.

Once the estimates for $\hat{\beta}$, the regression parameters, and $\hat{\tau}^2$, the model variance, have been determined, then the final smoothed estimates can be found. Maritz and Lwin (1989) show that the Empirical Bayes, or smoothed, estimate can be written as

$$\hat{F}_i^{\text{eb}} = (1 - \hat{\omega}_i) X_i \hat{\beta} + \hat{\omega}_i F_i,$$

where $\hat{\omega}_i = \hat{\tau}^2(\hat{\tau}^2 + \sigma_i^2)^{-1}$. This is a combination of the original estimate and the regression estimate weighted by their respective variances.

The objective of the smoothing model is to create a series of estimates with smaller MSE than the original estimates. Prasad and Rao (1990), through asymptotic arguments, have suggested using the following estimator for the mean square error:

$$\text{MSE}[\hat{F}_i^{\text{eb}}] = \text{MSE}[\tilde{F}_i^{\text{eb}}] + \left[\left(\frac{\partial \omega_i}{\partial \tau^2}\right)^2 \omega_i E(\hat{\tau}^2 - \tau^2)^2\right].$$

The mean square error for the Empirical Bayes estimate, using restricted maximum likelihood estimation, has been conjectured by Cressie (1992) to be:

$$\widehat{\text{MSE}}[\hat{F}_i^{\text{eb}}] = \widehat{\text{MSE}}(\tilde{F}_i^{\text{eb}}) + 2\hat{g}_{3i}(\hat{\tau}^2) = \hat{g}_{1i}(\hat{\tau}^2) + \hat{g}_{2i}(\hat{\tau}^2) + 2\hat{g}_{3i}(\hat{\tau}^2),$$

where

$$\hat{g}_{1i}(\hat{\tau}^2) = \hat{\tau}^2(1 - \hat{\omega}_i)$$

$$\hat{g}_{2i}(\hat{\tau}^2) = (1 - \hat{\omega}_i)^2 X_i^t (X^t \hat{V}^{-1} X)^{-1} X_i$$

and

$$\hat{g}_{3i}(\hat{\tau}^2) = (1 - \hat{\omega}_i)^2 (\hat{\tau}^2 + \sigma_i^2)^{-1} [i(\tau^2)]^{-1}.$$

The assumed normality of ϵ_i and δ_i is an important assumption in the derivation. Note the value for the sampling variance, σ_i^2, is assumed known.

Prasad and Rao give the following interpretation to each of the three components: $\hat{g}_{1i}(\hat{\tau}^2)$ is the Bayes estimate of the variance, $\hat{g}_{2i}(\hat{\tau}^2)$ is the contribution from estimating the regression parameters and $\hat{g}_{3i}(\hat{\tau}^2)$ is the contribution from estimating the model variance τ^2. An estimate of the component due to the estimation of the sampling variance is *not* available: the additional variance this would add is not known but its absence clearly implies that the MSE is underestimated.

3. EMPIRICAL BAYES LINEAR MODEL

3.1 Explanatory Variables

The Empirical Bayes model described above was fitted to the 96 observed adjustment factors, with the sampling variances estimated using the method described in Section 2. The linear model that was fitted to this data included the following explanatory variables:

(a) An indicator variable for each province/territory.

(b) An indicator variable for each sex.

(c) An indicator variable for each age group.

(d) A variable indicating the percentage of people in the domain that are renters.

(e) A variable indicating the percentage of people in the domain that do not speak either official language.

(f) Various interaction variables including province by renters, province by non-official language, age and sex by renters.

In total, 42 variables were used in the initial regression.

These variables were selected for the initial regression model based, in part, on the experiences of previous RRC studies (Burgess 1988), partly on the results of the 1991 coverage studies (Germain and Julien 1993) and partly on the experiences of the PES in the United States as described in Hogan (1992) and Datta *et al.* (1992). The actual rationale for the variables to be included are as follows:

(a) The province indicator was included as an indication of the difficulty of Census collection within each province. Prior to the 1991 Census, it was assumed that collection would be more difficult in British Columbia and Ontario, and the anecdotal field evidence during collection seemed to support this conjecture.

(b) The age and sex variable were included because of the known differences in undercoverage rates between males and females. The undercoverage, in previous studies, has also shown a marked increase for individuals in their 20's.

(c) Tenure, in effect the percent of renters in each domain, was included because of the experiences in the United States PES, results of previous RRC studies and as a suggestion from the Statistics Canada Statistical Methods Advisory Committee.

(d) The use of non-official language was an attempt to locate the immigrant and minority groups that in the past have tended to have higher undercoverage rates.

(e) The interaction terms were included to further refine the predictive power of the model.

The mean encompasses all those variables that are not included in the model. Note that since indicator variables are used for province, sex and age-sex, one variable has to be excluded in order to avoid a singular design matrix. In effect, the missing variable, say the province indicator for Newfoundland, is included in the mean.

An operational constraint was also placed on the model. The SAS IML program written to estimate the parameters was limited to 4,095 numeric elements in the design matrix, hence with 96 domains, or observations, the model was limited to a maximum of 42 variables.

3.2 Model Building Process

After starting with the full regression model and 42 explanatory variables, a procedure was needed to remove those variables that were not statistically significant. The procedure chosen was to eliminate the least significant variable after each completed estimation cycle. This implies that for the 42 variable model, the variable Female Renters aged 0 to 19 would be eliminated since it has a t-value of 0.05. The regression model was then re-run with the remaining 41 variables. The least significant variable was then eliminated from that model. This procedure is equivalent to the Backward Stepwise Regression described in Draper and Smith (1966, page 167).

The Backward Stepwise Regression method was used to eliminate all variables until all remaining variables had a t-values greater than 2 (in absolute value). However when the final model was examined, it was noticed that a multicollinearity problem existed between the indicator variables for certain provinces and the interaction terms for renters within the same provinces. The implication of this problem is that there are some explanatory variables which are highly correlated with each other. This in turn implies that not all parameters in the model can be estimated precisely. As a rule of thumb Judge *et al.* (1984, page 459) suggest that this can be a problem when the simple correlation between variables is greater than R^2, the coefficient of determination. The final model had a $R^2 = 0.85$ and the simple correlation between the variables in question were all greater than 0.90 (in absolute value, since the correlations were negative).

A solution to this problem was to delete the variables with the lower t-values which turned out to be the provincial indicators. The final model is shown in Table 1 with the estimated coefficients and their t-values. The effect of removing the provincial indicators was to lower the final R^2 from 0.85 to 0.844, thus little predictive power has been lost.

Table 1
Final Estimates of Variables Used in Regression

Category	Variable	Final Estimate ($\hat{\beta}$)	T-Value (absolute value) ($H_0 : \beta = 0$)
Mean	Mean	1.0075	575.72
Age - Sex Combination	Male 20 to 29	0.0563	15.34
	Male 30 to 44	0.0208	5.81
	Female 20 to 29	0.0240	6.49
Sex by Age by Non-Official Language	Female Language 0 to 19	0.0797	2.75
Tenure by Province	British Columbia Renters	0.0449	3.96
	Ontario Renters	0.0804	7.35
	Quebec Renters	0.0255	2.66
	New Brunswick Renters	0.1064	5.61
	Yukon Renters	0.0639	3.80
	Northwest Territories Renters	0.0682	6.22

The final regression model then had various diagnostic tests performed on it. Since the regression is a weighted least squares with a random error term, Lange and Ryan (1989) have suggested using the following form to create standardized residuals:

$$z_i = \frac{\hat{F}_i^{(eb)} - X_i \hat{\beta}}{\sqrt{\sigma_i^2 + \hat{\tau}^2}}.$$

The residuals were analyzed using both Q-Q plots and outlier detections methods: no major departures from the assumed distribution of the residuals were detected. More details on the residual analysis can be found in Dick (1993).

Table 2
Direct, Smoothed and Raked Estimates of Adjustment Factors

Sex	Age	Estimate	B.C.	Alta	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld	Yukon	N.W.T.
Male	0-19	Direct	1.017	1.026	1.012	1.029	1.028	1.017	1.022	1.019	1.004	0.999	1.031	1.036
		Smooth	1.019	1.013	1.009	1.013	1.029	1.016	1.027	1.010	1.007	1.006	1.026	1.027
		Raked	1.020	1.016	1.011	1.015	1.031	1.018	1.027	1.013	1.005	1.007	1.029	1.031
	20-29	Direct	1.087	1.036	1.068	1.058	1.113	1.071	1.122	1.063	1.060	1.057	1.098	1.127
		Smooth	1.086	1.056	1.065	1.062	1.104	1.074	1.103	1.064	1.063	1.062	1.094	1.122
		Raked	1.083	1.061	1.073	1.067	1.101	1.079	1.096	1.073	1.041	1.074	1.096	1.127
	30-44	Direct	1.031	1.021	1.028	1.034	1.054	1.047	1.043	1.018	1.025	1.026	1.069	1.080
		Smooth	1.039	1.026	1.028	1.030	1.053	1.041	1.046	1.026	1.028	1.028	1.052	1.059
		Raked	1.038	1.028	1.032	1.032	1.051	1.043	1.043	1.029	1.018	1.033	1.053	1.059
	45+	Direct	1.019	1.018	1.002	1.014	1.013	1.011	1.014	1.016	1.018	1.016	0.992	1.076
		Smooth	1.017	1.011	1.006	1.009	1.019	1.013	1.019	1.010	1.009	1.009	1.021	1.039
		Raked	1.014	1.010	1.006	1.009	1.016	1.012	1.015	1.010	1.005	1.010	1.019	1.035
Female	0-19	Direct	1.034	1.018	1.017	1.012	1.037	1.029	1.029	1.014	0.995	1.016	1.026	1.054
		Smooth	1.030	1.015	1.013	1.015	1.038	1.023	1.030	1.010	1.006	1.010	1.028	1.061
		Raked	1.032	1.018	1.016	1.017	1.040	1.026	1.030	1.012	1.004	1.013	1.030	1.068
	20-29	Direct	1.068	1.047	1.028	1.020	1.072	1.043	1.070	1.030	1.004	1.041	1.068	1.072
		Smooth	1.058	1.036	1.031	1.029	1.070	1.044	1.071	1.031	1.027	1.033	1.069	1.092
		Raked	1.058	1.041	1.036	1.032	1.070	1.048	1.068	1.037	1.018	1.041	1.072	1.099
	30-44	Direct	1.013	1.009	1.004	1.006	1.027	1.017	1.031	1.019	1.004	1.024	1.031	1.020
		Smooth	1.018	1.008	1.007	1.007	1.030	1.017	1.029	1.010	1.007	1.011	1.028	1.026
		Raked	1.017	1.008	1.007	1.007	1.028	1.017	1.025	1.011	1.004	1.012	1.027	1.026
	45+	Direct	1.007	1.003	1.018	1.001	1.011	1.011	1.000	1.002	0.993	1.013	1.024	1.007
		Smooth	1.014	1.006	1.010	1.006	1.021	1.015	1.020	1.006	1.005	1.009	1.031	1.026
		Raked	1.008	1.004	1.007	1.004	1.012	1.009	1.011	1.004	1.002	1.006	1.019	1.016

3.3 Estimates of Adjustment Factors

Table 2 shows both the direct survey estimate and the smoothed Empirical Bayes estimate of the adjustment factors. An inspection of the table shows that these estimates are relatively close, reflecting the Empirical Bayes methodology of combining the direct survey estimate with the model estimate. Note that all of the domains that were originally estimated to have overcoverage - shown by an estimated adjustment factors being less than one - have been changed, by the Empirical Bayes estimates to being an estimate of undercoverage. The difference between the two sets of estimated adjustment factors - in absolute terms - differ by under 1% and in the larger provinces by less than 0.5%. However, for some of the smaller provinces and territories the difference between the two estimates can be substantially larger. In the Northwest Territories the change between the directly estimated adjustment factor and the Empirical Bayes estimate is about 2% for 3 age - sex groups and over 3% for another.

The objective of the Empirical Bayes model is to produce estimates with smaller MSE than the survey estimates. From Section 2.2 it can be shown that the variance for the direct survey estimates is calculated from

$$\log \hat{v}(F_i) = -6.133 - 0.285 \log C_i,$$

while the Prasad-Rao MSE, from Section 2.3, is calculated by

$$\widehat{\mathrm{MSE}}[\hat{F}_i^{\mathrm{eb}}] = \widehat{\mathrm{MSE}}(\hat{F}_i^{\mathrm{eb}}) + 2\hat{g}_{3i}(\hat{\tau}^2) = \hat{g}_{1i}(\hat{\tau}^2) + \hat{g}_{2i}(\hat{\tau}^2) + 2\hat{g}_{3i}(\hat{\tau}^2).$$

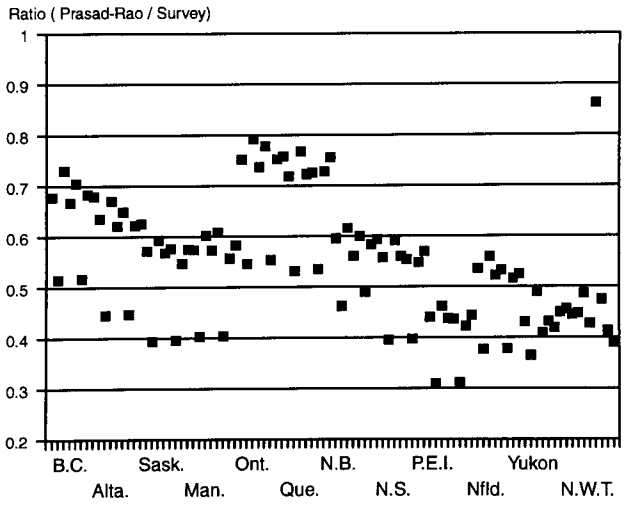

Figure 2. Ratio of root MSE, Prasad-Rao and survey.

Figure 2 plots, for each domain, $R = \sqrt{\widehat{\text{MSE}}[\hat{F}_i^{eb}]}/\hat{v}(F_i)$ the ratio of the root mean square errors for the Empirical Bayes model and the estimated survey variance (Note that within provinces the domains are ordered as Male aged 0-19, 20-29, 30-44 and 45 and over and Female aged 0-19, 20-29, 30-44 and 45 and over). Clearly, the Empirical Bayes MSE is smaller in all domains. However, in the larger provinces, Ontario and Quebec, the ratio of the root MSEs is only between 0.7 and 0.8. This relatively small gain is a reflection of the large sample sizes in these domains which in turn give a reliable estimate of the variance. The large gains are made in the smaller provinces and territories. For instance, in Prince Edward Island, the ratio of the root MSEs are all smaller than 0.5 showing the large improvement in the estimates. The one outlier is in the Northwest Territories (females aged 0 to 19): the Prasad-Rao MSE appears to have been overestimated in this domain.

4. ADJUSTMENTS MADE TO EMPIRICAL BAYES ESTIMATES

4.1 Rationale and Methodology

The advantage of the Empirical Bayes method is apparent from the above discussion. However, the Empirical Bayes methodology does not preserve the higher level (*i.e.*, the large domain) direct survey estimates that are reliable. By this it is meant that the provincial totals and the age - sex domain totals for the direct survey estimates and the Empirical Bayes estimates are not equal. Since the two surveys were designed to produce estimates at these levels, it is crucial that the Empirical Bayes be consistent with these reliable marginal totals.

To achieve consistency of estimates of missed persons between the reliable provincial and age - sex totals from the direct survey estimates and the final Empirical Bayes estimates, a raking ratio procedure was used. This is basically the method used in Australia to determine their small domain estimates (see Choi *et al.* 1988). This technique re-scales the individual Empirical Bayes estimates to conform to the known provincial and national age - sex totals. Once this procedure has converged, the final estimates will be consistent with the direct survey totals. In terms of a log-linear model, we are using as the main effects (province and age-sex) estimates the results from the two coverage studies and the interaction terms (province by age-sex) estimates from the Empirical Bayes modelling.

Details of the procedure can be described as follows. Assume that we have a matrix of estimated missed persons that has P columns (corresponding to the provinces) and A rows (corresponding to the age-sex groups). First set $F_{pa} = F_i$, then let $\hat{M}_{pa}^{s} = C_{pa}(F_{pa} - 1)$ be the direct survey estimate of the number of missed persons in province p and age - sex group a and let $\hat{M}_{pa}^{(eb)} = \hat{M}_{pa}^{(0)} = C_{pa}(\hat{F}_{pa}^{(eb)} - 1)$ be the Empirical Bayes estimate of missed persons from the Empirical Bayes model. If we let a plus sign (+) represent addition across the variable then the raking estimate can be written for cycles $\kappa = 0, 1, \ldots$ as;

$$\hat{M}_{pa}^{(2\kappa+1)} = \hat{M}_{pa}^{(2\kappa)}\left(\sum_{a=1}^{A}\hat{M}_{pa}^{s}\Big/\sum_{a=1}^{A}\hat{M}_{pa}^{(2\kappa)}\right)$$

and

$$\hat{M}_{pa}^{(2\kappa+2)} = \hat{M}_{pa}^{(2\kappa+1)}\left(\sum_{p=1}^{P}\hat{M}_{pa}^{s}\Big/\sum_{p=1}^{P}\hat{M}_{pa}^{(2\kappa+1)}\right).$$

This procedure will converge to a unique solution. Since this is basically a log-linear model, the underlying assumption is that the relationship determined by the Empirical Bayes model for the interaction between province and age - sex group is valid and will be preserved.

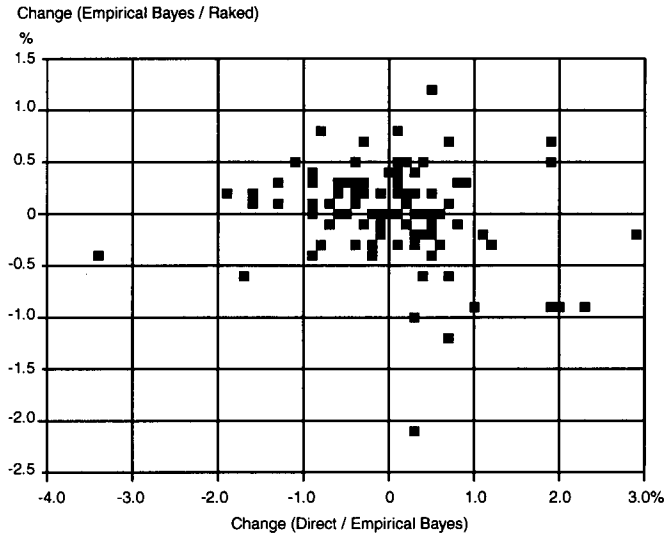

Figure 3. Percent change in estimates of adjustment factors.

Table 2 shows the final raked estimates of the adjustment factors along with both the original survey estimates and the Empirical Bayes estimates. Generally, the impact of raking is to shrink the Empirical Bayes estimate back towards the survey estimate. This is shown in Figure 3. Here two different percent changes in the estimated adjustment factors are plotted. The X-axis shows the percent change between the direct survey estimate and the Empirical Bayes estimate. The Y-axis shows the percent change between the Empirical Bayes estimate and final raked estimate. The plot shows that the two variables are negatively correlated: hence the raking tends to move the Empirical Bayes estimates closer to the original survey estimates.

One draw back of this procedure is that the MSEs of the raked adjustment factors are now very difficult to estimate. Due to the non-linear nature of the raking ratio procedure, a direct calculation is impossible. It is possible to use a Taylor series expansion; however this assumes a large sample size in each domain when in fact we know some domains have very small sample sizes. A possible procedure is to adjust the estimated MSE from the Empirical Bayes estimates and multiply these by the squared ratio of the raked Empirical Bayes estimate over the Empirical Bayes estimate. While this procedure is only a crude approximation, it can at least give some guidance as to the reliability of the individual estimates. This method will ensure that the coefficient of variations calculated for the Empirical Bayes estimates will be retained for the corresponding raked Empirical Bayes estimates. This is the procedure that was used to produce the final MSE estimates for the raked Empirical Bayes estimates of missed persons.

4.2 Detailed Domain Estimates

The Population Estimates Program requires even finer detail than that produced by the various models discussed above. In fact the program needs estimates for single years of age for each sex for each Census Division within each province. Since the Empirical Bayes methodology is limited somewhat by the direct survey results - an estimate with a non-zero standard error is required for each domain - synthetic methods must be used to generate the more detailed estimates.

For the Population Estimates Program, estimates for each province and sex were produced for 9 age groups instead of the 4 age groups used in the Empirical Bayes model. A straight synthetic model, using the raked Empirical Bayes estimates as initial values, was proposed for this stage of estimation. To produce these more detailed estimates, the raked Empirical Bayes estimate was allocated proportionally by Census count across all sub-age groups within each province and sex. Let the final raked estimate in the p-province and the a-th age-sex group be $\hat{M}_{\text{pa}}^{2\kappa+2} = \hat{M}_{\text{pa}}^{rf}$. Also if the a-th age - sex group is composed of Q exclusive sub-age groups then the estimate of the missed persons in the p-th province and the q-th sub-age group within the a-th age - sex group would be

$$\hat{M}_{\text{pa}_q} = \hat{M}_{\text{pa}}^{rf}\left(\frac{C_{\text{pa}_q}}{C_{\text{pa}_+}}\right),$$

where $C_{\text{pa}_+} = C_{\text{pa}} = \sum_{q=1}^{Q} C_{\text{pa}_q}$. This approach guarantees that the estimates from the earlier raked Empirical Bayes output are preserved for the original domain total. The further estimates that are required for the population estimates program use demographic methods. In fact, one of the objectives of the Empirical Bayes procedure is to provide initial estimates for the demographic methods. See Michalowski (1993) for further details.

5. SUMMARY AND CONCLUSIONS

The Empirical Bayes methodology was adopted because it preserves the more reliable estimates from the larger provinces and domains while permitting a model based estimate to dominate if the underlying direct estimate is unreliable. This is in accordance with standard survey methods of using the direct survey estimates as much as possible. The raking ratio procedure used for adjusting the estimates from the Empirical Bayes model was used to ensure consistency with the direct survey results that were known to be reliable.

As for the explicit model used to describe the underlying true adjustment factors, it must be noted that this model is purely descriptive. Its primary function is to use explanatory variables to describe the variation in adjustment factors, taking into account the sampling error associated with each adjustment factor. It would not be prudent to make far-reaching conclusions on the nature of undercoverage from the final set of parameters included in the model.

The main weakness of this approach is with the two variances that are estimated. The assumption of the regression model errors being approximately normally distributed is difficult to assess. In the absence of any real knowledge about the true underlying distributions any assumption about the model variance will be essentially unverifiable. The proposed model variance seems reasonable and diagnostic checks have not revealed any major problems.

The sampling variance model is more problematic. All Empirical Bayes methods assume that this variance is known, when in fact it has to be estimated. Efforts to extend the Prasad-Rao MSE calculation to include the contribution from this estimated parameter have not yielded any new results.

In the future, research will concentrate in working around the problem associated with estimating the sampling variances. Further work needs to be conducted on the Prasad-Rao MSE calculation. In addition, the possibility of using the micro level data from the coverage studies and estimating the undercoverage rates directly through logistic regressions as in Wong and Mason (1985) will be pursued.

Another project would be to examine the implications of recasting the Empirical Bayes model into the standard state space framework (Robinson 1991). Pfeffermann and Burck (1990) have suggested a method for calculating the MSE for a time series placed in a state space model that has to conform to certain periodic benchmarks. The state space formulation would also be useful in explicitly incorporating the demographic methods.

ACKNOWLEDGEMENTS

The author is grateful to D. Binder, M. Hidiroglou, R. Carter, M. Armstrong, J. Tourigny, J.N.K. Rao and especially D. Royce for commenting on an earlier version of this paper. An Associate Editor and two referees also provided comments that improved the final version.

REFERENCES

BURGESS, R.D. (1988). Evaluation of Reverse Record Check estimates of undercoverage in the Canadian Census of Population. *Survey Methodology*, 14, 137-156.

CHOI, C.Y., STEEL, D.G., and SKINNER, T.J. (1988). Adjusting the 1986 Australian Census count for under-enumeration. *Survey Methodology*, 14 173-189.

CRESSIE, N. (1992). REML estimation in Empirical Bayes smoothing of census undercount. *Survey Methodology*, 18, 75-94.

DATTA, G.S., GHOSH, M., HUANG, E.T., ISAKI, C.T., SCHULTZ, L.K., and TSAY, J.H. (1992). Hierarchical and Empirical Bayes methods for adjustment of census undercount: The 1988 Missouri Dress Rehearsal data. *Survey Methodology*, 18, 95-108.

DICK, J.P. (1993). Procedures used in modelling net undercoverage in the 1991 Census. Internal Statistics Canada memorandum.

DRAPER, N.R., and SMITH, H. (1966). *Applied Regression Analysis*. New York: John Wiley and Sons.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year (with discussion). *Journal of the American Statistical Association*, 84, 927-943.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 82, 269-277.

FREEDMAN, D., and NAVIDI, W. (1992). Should we have adjusted the U.S. Census of 1980? (with discussion). *Survey Methodology*, 18, 3-74.

GERMAIN, M.-F., and JULIEN, C. (1993). Results of the 1991 Census coverage error measurement program. *Proceedings of Seventh Annual Research Conference*. United States Bureau of the Census, 55-70.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.

HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-337.

HOGAN, H. (1992). The 1990 Post-Enumeration Survey: an overview. *The American Statistician*, 46, 261-269.

JUDGE, G.G., GRIFFITHS, W.E., CARTER HILL, R., and LEE, T-C. (1984). *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.

LANGE, N., and RYAN, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17, 624-642.

MARITZ, J.S., and LWIN, T. (1989). *Empirical Bayes Methods (2nd edition)*. London: Chapman and Hall.

MICHALOWSKI, M. (1993). Revised postcensal and intercensal estimates: Canada, provinces and territories, 1971 - 1991. Internal report, Population Estimates Section, Statistics Canada.

PFEFFERMANN, D., and BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*. New York: John Wiley and Sons.

ROBINSON, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Sciences*, 6, 15-51.

ROYCE, D. (1992). A comparison of some estimators of a set of population totals. *Survey Methodology*, 18, 109-125.

STATISTICS CANADA (1993). *1991 Census Technical Report: Coverage*. Ottawa: Supply and Services Canada, 1994. 1991 Census of Canada: Catalogue No. 92-341E.

WONG, G.Y., and MASON, W.M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80, 513-524.