Can Survey Bootstrap Replicates Be Used for Cross-Validation?

Geoff Rowe[1] and David Binder[2]
[1]Geoff Rowe, Statistics Canada, Tunney's Pasture, Ottawa, ON, K1A 0T6, Canada; geoff.rowe@statcan.gc.ca
[2]David Binder, Statistics Canada, Tunney's Pasture, Ottawa, ON, K1A 0T6, Canada; dbinder49@hotmail.com

Abstract
We propose an extension to bootstrap methods for evaluating regression models estimated with data from surveys with complex design. Such methods involve selection of replicate samples formed from simple random samples of sampled clusters within strata. Selection is carried out with replacement, so that about one third of clusters are typically left out of a given replicate sample. Our evaluation method exploits the excluded clusters, using them as cross-validation samples for assessment of a model's prediction error, and at the same time using the bootstrap samples to estimate the variance of regression coefficients. We also consider the use of a sample of the replicates as a cross-validation sample.

Key Words: Complex surveys, .632+ Bootstrap, Health Utility Index

1. Introduction

As is well known, regression residuals will give an overly optimistic view of the predictive value of an equation (Efron, 1986). It is also known that model-specification searches that consist simply of eliminating all of the "non-significant" terms from a trial specification can result in a selected equation with inferior predictive value (Hastie, Tibshirani and Friedman, 2001). Simply retaining all terms that have an intuitive appeal (whether "significant" or not) can also result in an equation with inferior predictive value.

Cross-validation methods attempt to directly facilitate the search for specifications that will produce accurate predictions. In this paper, we extend the scope of cross-validation methods to data from surveys with complex design. The paper is in two parts. Following the introduction, Section 2 outlines design-based properties of the bootstrap/cross-validation and establishes the validity of methods utilizing replicate samples when those methods depend only on first and second moments. Section 3 illustrates our method with a comparative assessment of selected models of health dynamics using Statistics Canada's longitudinal National Population Health Survey (1992-2004). Section 4 provides concluding comments.

2. Cross-validation applied to survey samples

The term 'cross-validation' generally refers to techniques that directly assess prediction error of a fitted equation by splitting the available sample and using one part to fit the equation (model construction) and reserving the other part for an assessment of predictions (model validation) (Picard and Cook, 1984). Model selection by cross-validation consists of proposing and fitting alternative models, assessing the out-of-sample prediction error of each, and choosing the one with the smallest prediction error.

In practice, some care needs to be exercised in applying the cross-validation method. This is because the size of the sample used in model construction will affect the bias in predictions in one way and affect the variance of prediction error assessments in the opposite direction. The larger the model-construction sample, the smaller the bias in predictions; but, the smaller the model-validation sample, the larger the variance of the assessment. Both Shao (1993) and Efron and Tibshirani (1997) have considered improvements on 'naïve' cross-validation, most of which have some of the features of the bootstrap.

A typical 'K-fold cross-validation' for samples assumed to have been generated directly from a model is obtained by partitioning the original sample into K subsamples, retaining one of the subsamples for validating the estimated model. The remaining K−1 subsamples are used as model-construction or training data. Normally the training and validation steps occur K times with each of the K subsamples making a contribution to the validation average. In a 'Leave-one-out cross-validation' only a single observation from the original sample is used to validate the model, and the remaining observations are the training data. This is repeated such that each observation in the sample is used once as the validation data. The usual assumption made for the validation sets are that they are independent from the training sets.

With complex survey data, however, without making strong assumptions about the non-informativeness of the sample design, the observations are not independent, so it would seem that cross-validation techniques that have been developed for non-survey data cannot be applied in a complex survey setting. However, an interesting property of the Rao-Wu-Yue bootstrap (see Rao et al, 1992) is that the bootstrap replicates can be uncorrelated. Samples that are uncorrelated can be used for cross-validation purposes when the methods depend on only the first and second moments.

2.1 Cross-validation using Rao-Wu-Yue Bootstrap Replicates

The Rao-Wu-Yue bootstrap (RWYB) is now used by many survey producers, including Statistics Canada, as a useful way to obtain design-based variance estimates for a large number of descriptive statistics that estimate finite population quantities. To obtain the RWYB, for a multi-stage survey, where it can be assumed that the primary sampling units (psu's) are selected with replacement, at least approximately, the survey producer selects bootstrap replicates by selecting within each of the H strata a sample of m_h psu's with replacement from the n_h psu's in the original sample. Letting $z_{hij}^{(b)}$ be an indicator variable taking the value one when the *i*th psu of the *h*th stratum is selected on the *j*th draw for the *b*th replicate, we define

$$\hat{\bar{Y}}^{(b)} = \frac{1}{N}\sum_{h=1}^{H}\left[\left(\frac{m_h}{n_h-1}\right)^{1/2}\frac{n_h}{m_h}\sum_{i=1}^{n_h}\hat{Y}_{hi}\sum_{j=1}^{m_h}z_{hij}^{(b)} + \left[1-\left(\frac{m_h}{n_h-1}\right)^{1/2}\right]\hat{Y}_h\right]$$

to be the *b*th bootstrap replicate estimating the finite population mean $\bar{Y}$. When $m_h = n_h - 1$, this simplifies to

$$\hat{\bar{Y}}^{(b)} = \frac{1}{N}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{h-1}}z_{hij}^{(b)}\hat{Y}_{hi}\ .$$

If we produce estimates given by

$$\hat{\bar{U}}^{(b)} = \hat{\bar{Y}}^{(b)} - \hat{\bar{Y}}\ ,$$

it turns out that under the design-based randomization, these replicates have means equal to zero, and that they are uncorrelated – details are available from the authors. Under a model-design based randomization framework, these replicates also have means equal to zero and are uncorrelated – see Binder and Roberts (2006) for details of the model-design-based randomization framework. Therefore, many methods in the standard literature for cross-validation are applicable to bootstrap replicates when the methods depend on only the first and second moments. A key to this technique is to define replicate estimates that have mean zero.

2.2 An Alternative Cross-validation Method Based on Unsampled PSU's

In each bootstrap replicate, there will be some psu's that are not included in the replicate sample. This is similar to the .632+ bootstrap used in non-survey settings. We consider the properties of estimates based on these unsampled psu's.

We let

$$\tilde{\bar{Y}}_h^{(b)} = \left(1-\frac{1}{n_h}\right)^{-m_h}\sum_{i=1}^{n_h}\tilde{z}_{hi}^{(b)}\hat{Y}_{hi}$$

where $\tilde{z}_{hi}^{(b)}$ is the indicator variable for whether the ith psu in the hth stratum is not in the bth bootstrap replicate. In this case, $\tilde{Y}_h^{(b)}$ is design-unbiased for $\hat{Y}_h$ - details are available from the authors. We refer to the first factor on the right hand side of the above expression as the adjustment factor for the full sample weights. Properties of this new cross-validation sample need to be studied; however, based on the example given below, the use of such samples for cross-validation purposes appears to hold much promise. The advantage of this method is that larger samples can be used as training sets. This concern in the non-survey setting is one that led to the 'Leave-one-out cross-validation' rather than the 'K-fold cross-validation', where a single subsample used for one validation step can be quite small – the sample size being only (1/K) of the original sample size (hence K is often limited to 5 or 10).

3. Illustrating Cross-validation Techniques

In order to illustrate our techniques, we present details of an analysis of longitudinal health data drawn from Statistics Canada's National Population Health Survey (NPHS) (Statistics Canada, 1999). The NPHS is a panel survey of self-reported health based on interviews conducted biannually over more than a decade. The initial sample comprised over 17,000 respondents, with more than 11,000 providing a full response in all of the six cycles available to us. NPHS data files are disseminated with 500 sets of bootstrap weights (Yeo, et.al., 1999).

Our analysis focuses on the health dynamics of individuals as measured by the Health Utility Index or HUI (Feeny, et.al., 2002: see also www.healthutilities.com/HUI.htm). The HUI provides a description of an individual's overall functional health using eight attributes: vision, hearing, speech, mobility, dexterity, cognition, emotion, and pain. Based on a standard set of questions, the HUI provides a summary health score between -.360 and 1.000. For instance, an individual who is nearsighted, yet fully healthy on the other seven attributes, receives a score of 0.973. On that scale, the most preferred health level (perfect health) is rated 1.000 and death is rated 0.000, while negative scores reflect health states considered worse than death.

Health dynamics can be complex: periods of stability might be followed by abrupt temporary changes in state (e.g., accidents) or by spells of gradual change. In this illustration, we will be concerned only with the conditions under which a change may or may not occur and do not consider the subsequent magnitudes of change. A key scientific question is whether accounting for observations from earlier time periods would reveal persistence or momentum/inertial effects on health change.

The analysis was conducted in two phases. The first phase focused on model selection by inclusion or exclusion of subsets of candidate predictors. Here, cross-validation serves as a means of ranking models in order of their predictive accuracy. The second phase focused on non-linearities in the association between predictors. In this case, cross-validation facilitates comparison of non-nested models that differ in the form of non-linear associations.

3.1 Population Health by Age Group

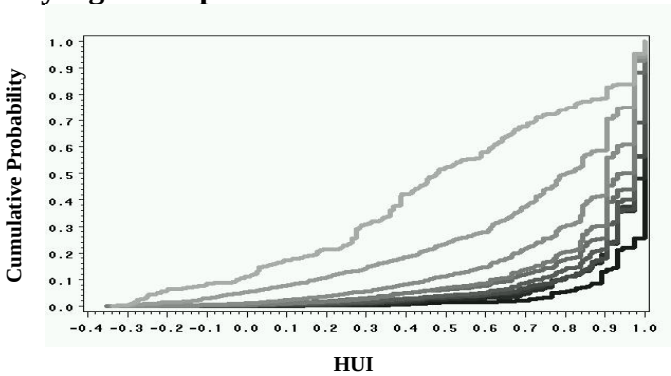

Figure 1: Empirical HUI Distribution Functions by Age Group: 10-year groups ordered youngest (black)-to-oldest (light grey) – based on six cycles of NPHS data.

The health of a majority of children, as assessed by the HUI, is characterized by perfect or near perfect health. At succeeding ages, the proportion at or near perfect health declines and the range of HUI over which the remainder of the population is distributed increases. These basic facts can be seen in the empirical distributions functions in Figure 1 which display empirical cumulative probability curves versus corresponding HUI values for each of ten 10-year age groups. In this chart, HUI appears to provide a plausible description of the affect of aging on population health.

3.2 Modeling Health Change of Individuals

We were interested in modelling whether or not the HUI changes in a two-year period. If $HUI_{i,t}$ is referred to as 'Current Health', a change was observed for individual *i* if

$$HUI_{i,t+2} \neq HUI_{i,t}.$$

Our model of health change was expressed as a logistic regression:

$$pr\left(HUI_{i,t+2} \neq HUI_{i,t}\right) = \left(1 + \exp\left(-X_{i,t}\hat{\theta}\right)\right)^{-1}$$

The original NPHS household sample that was in-scope for longitudinal follow-up comprised about 17,000 respondents. We divided the sample into overlapping sets of responses from each combination of three consecutive cycles. Including attrition, there were under 50,000 such sets of triplets. Reasoning that the transition from perfect to less-than-perfect health would require a special model on its own, we chose to exclude observations for response sets in which $HUI_{i,t}$ equaled 1.0. Similarly, working from the assumption that the health dynamics of men and women might differ in special ways, we chose to focus here exclusively on men (in anticipation of observing more changes occurring earlier in life). These two additional selections reduced our working sample to just over 14,000 sets of three consecutive cycles.

The matrix of candidate predictors ($X_{i,t}$) included terms representing Immigrant Status, Presence of a Spouse, and Broad Education Attainment; as well as (natural) cubic spline basis functions (Hastie et.al., 2001) representing non-linear effects of Age at period t, Current Health, $HUI_{i,t}$, and Lagged Health (given by $HUI_{i,t-2}$). The cubic splines involve two regression parameters each, and each pair of basis functions require that three knot locations be specified.

In Phase 1 of our analysis, spline knot locations for the age variable were chosen to broadly group responses into younger, mid and older age groups: positioning knots at ages 25, 50, and 75. For the HUI variables, the two upper knot locations (0.9 and 0.5) were those that have been used in the past to represent dividing lines between good/fair heath and fair/poor health, respectively. The third HUI knot was set at 0.0, the dividing line between worse than dead and better than dead.

3.3 Regression Estimates and Prediction Error

Our logistic regression equations were estimated using the SAS GENMOD procedure. Given that our data contained as many as four observations on each respondent, we chose to estimate an odds ratio, assumed to be constant over time, to account for the association between observations from the same respondent and adopted the Alternating Logistic Regressions variant of GEE estimation (Carey, et.al., 1993). However, since we had no intention of using the resulting estimates of coefficient standard errors, GEE estimation was not critical.

The cross-validation set-up employed here uses the 500 sets of bootstrap weights that are disseminated with NPHS data. Each model to be estimated and evaluated makes use of one set of bootstrap weights at a time. Our first step is estimation of a logistic regression using those responses with non-zero bootstrap weights. Our second step uses the estimated equation and responses from unsampled PSUs to perform an out-of-sample assessment of prediction error (for cross-validation purposes, the weights used were the full sample weights multiplied by the adjustment factor described in section 2.2).

We have used two measures of predictive accuracy: Deviance and mean-squared error (MSE). These two measures are defined in terms of the survey weights W, the binary dependent variable Y, and the probability $p(X_i\theta)$ which is predicted on the basis of covariate information X and estimates of the parameters θ. The terms Y**, X**, and W** are based on the cross-validation replicate sample. Θ^* identifies a parameter estimate based on the b-th bootstrap replicate

sample. The subscript t denoting time period and the subscript b denoting the bootstrap replicate have been suppressed for simplicity.

$$Deviance = 2 \sum_i W_i^{**} \left(Y_i^{**} \ln \frac{Y_i^{**}}{pr(X_i^{**} \hat{\theta}^*)} + (1\text{-}Y_i^{**}) \ln \frac{1-Y_i^{**}}{1-pr(X_i^{**} \hat{\theta}^*)} \right)$$

$$MSE = \sum_i W_i^{**} \left(Y_i^{**} - pr(X_i^{**} \hat{\theta}^*) \right)^2$$

Efron (1978) demonstrates that both of these are appropriate measures of the distance of observations from predictions. In addition, he shows that Deviance and MSE will be roughly proportional (Deviance ≈ 6 MSE). Thus, MSE, being the simpler measure, is likely sufficient for our purposes. However, Deviance provides a useful conceptual link to likelihood methods.

Another link to more conventional methods is provided by Akaike's Information Criterion (AIC), which has the following definition:

$$AIC = 2 \sum_i W_i \left(Y_i \ln \frac{Y_i}{pr(X_i \hat{\theta})} + (1\text{-}Y_i) \ln \frac{1-Y_i}{1-pr(X_i \hat{\theta})} \right) + 2\, Model\ df$$

where the first term is twice the negative weighted (pseudo) log-likelihood and the second term is a penalty varying with the number of parameters in the model. (Note that AIC is estimated using the full sample without bootstrapping or cross-validation.) Efron (1986) shows that, for logistic regression, the AIC penalty term will approximate the negative bias in the full-sample estimate of Deviance. Thus, expressed as error per observation, values of AIC and cross-validated Deviance should be of similar magnitude.

3.4 Phase 1 Results: Preliminary Model Selection

Using the candidate predictors identified in section 3.2, 33 models were estimated (using the full-sample and each of the 500 bootstrap replicate samples). These 33 models correspond to most of the interesting sub-sets of the candidate predictors. Figure 2 displays for each of the models the AIC, and the average value of each of the cross-validated Deviance and MSE statistics over the 500 replicates. These models are ordered in decreasing order of the AIC.

The out-of-sample criteria (Deviance and MSE) are in close agreement with the preference ordering of models provided by AIC. In only five cases, would re-ordering by cross-validated Deviance result in an exchange of positions between adjacent models. A corresponding re-ordering of seven neighbours would result if re-ordering were based on cross-validated MSE. Thus, the criteria appear to be largely mutually consistent.

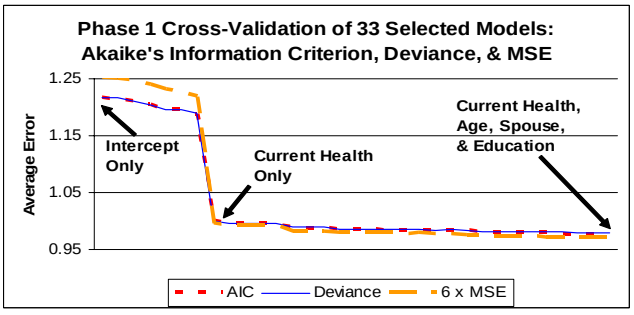

Figure 2: Phase 1 Cross-Validation – 33 models ordered by decreasing AIC

There is only one point where a marked reduction in prediction error is evident. The large jump seen in the chart distinguishes models that do not contain 'Current Health' terms (to the left of the jump) from models that contain 'Current Health' terms (to the right of the jump). Evidently, 'Current Health' terms are crucial to a good model.

Differences among succeeding models – all containing 'Current Health' – appear small. The reduction in MSE obtained by adding 'Current Health' is an order of magnitude greater than the incremental reduction in MSE provided

by the over-all best fitting model. Given the relatively small improvement in predictive power, it is appropriate to question whether the best fitting model has been established on a robust basis.

3.5 Confirming Phase 1 Model Selection

Another common model selection strategy is Backward Elimination. Here we estimated the model parameters using the full sample, and we estimated the standard errors from the first 250 bootstrap samples. At each stage of the elimination we dropped the least significant estimated regression coefficient, and we continued until of all remaining terms had a p-value less than 0.06. This resulted in eliminating the Lagged Health terms and the Immigrant indicator - all of the remaining terms were judged significant.

To confirm the Backward Elimination results, the joint significance of the three terms that had been eliminated was assessed. An independent bootstrap estimate of the covariance matrix of the terms in the full model was obtained using the second 250 bootstrap samples. A Wald test was performed on the two Lagged Health spline coefficients and on the Immigrant/Non-Immigration Indicator (Figure 3). Again, the three terms appeared jointly 'insignificant'.

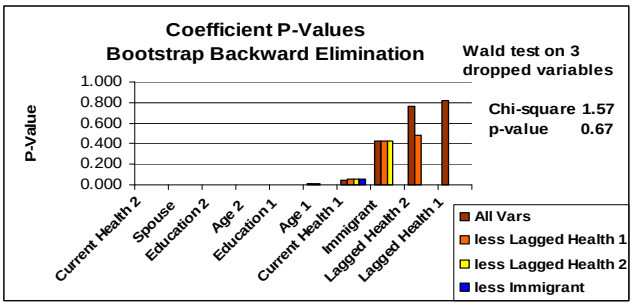

Figure 3: Model Selection by Backward Elimination

We see, therefore, that Cross-validation and Backward Elimination identify the same 'best' model among those examined in Phase 1: among the 33 models considered, the best predictions were obtained when the two Lagged Health terms and the Immigrant/Non-Immigration Indicator were dropped, while Age, Current Health, Spouse and Education terms were retained.

3.6 Phase 2 Results: Additional Variables and Empirical Placement of Spline Knots

The presence or absence of the Lagged Health terms in the preferred equation specification has scientific significance. The elimination of Lagged Health from the preferred model implies little or no inertia in the process of health change. Our concern that the role of Lagged Health had not been adequately assessed led to the 2nd Phase of the analysis.

Exploratory work involving graphical displays of residuals led to the observation that those who had been in perfect health in the previous period ($HUI_{i,t-2}$ equal to 1.0) seemed to have, all else being equal, markedly different chances of HUI change than others (recall that no men in this sample are 'currently' in perfect health). Correspondingly, an indicator of perfect lagged health was added to the set of candidate predictors.

There were, additionally, concerns about the specification of immigrant effects, because immigrants are generally selected for good health at the time of immigration. These concerns were addressed by adding age-at-immigration terms in the form of 2-parameter splines.

As a final step, Phase 2 provided an opportunity to explore alternative knot placements and hence a more flexible non-linear response specification (recall that knot placements used in Phase 1 were taken as given and were identified by an appeal to intuition).

In Figure 4 we display results for eight models: two Phase 1 models (i.e., that including one Current Health term only and the best among Phase 1 models which included only Age, Current Health, Spouse and Education terms), and six Phase 2 models which include various combinations involving extended immigrant (I+) and Lagged Health (L+) specifications. Estimation of each Phase 2 model also involved a random search for improved knot placements. (The random search was performed with an ad hoc SAS macro that randomly perturbed knot locations followed by repeated

calls to PROC GENMOD.) The first two points on the x-axis give the results for the two Phase 1 models; the remaining six points are for the Phase 2 model sorted in decreasing order of the AIC (Current-Health-only with improved knot placement being the one with highest AIC of the Phase 2 models).

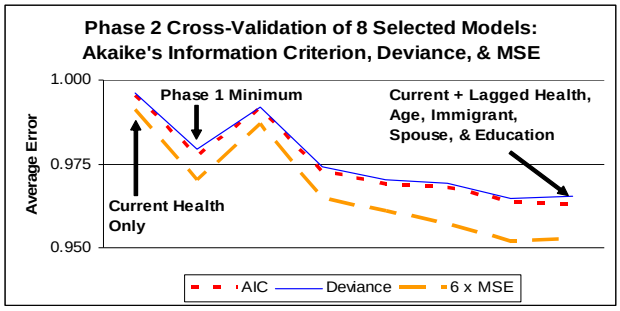

Figure 4: Phase 2 Cross-Validation – Two Phase 1 models compared with eight Phase 2 models

The best fitting Phase 2 model based on the AIC criterion contains all terms; Immigrant and Lagged Health terms included; however, the best fitting Phase 2 model based on both the MSE and Deviance criteria still excludes Immigrant terms, but not Lagged Health terms. The addition of new variables and the search for more appropriate knot locations led to definite, but modest improvements in the accuracy of the best model. Prediction error is still not that much smaller than a model with Current Health only and with knot placement based on intuition.

When we compared the functional form of predictions produced in Phase 1 and the best of the Phase 2 models, marked differences were revealed. Figure 5 shows the estimated contribution of Current and Lagged Health to the odds-ratio for change in health, based on the full Phase 1 model and on the best of the Phase 2 models.

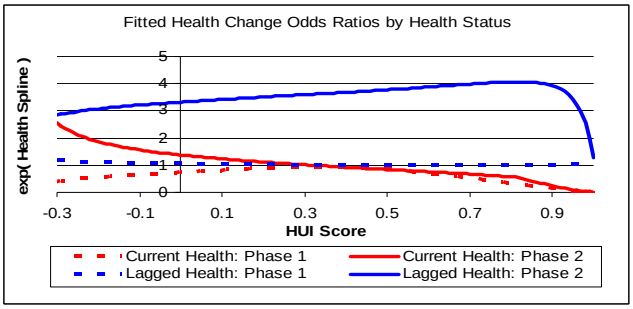

Figure 5: Comparing Phase 1 and 2 Fitted Values

The Phase 2 knot placement search has uncovered dramatically greater curvature in the fitted odds ratios by HUI (current and lagged) than were found in the corresponding model from Phase 1. Rather than being insignificant, Lagged Health plays a key role in accurately representing the health dynamics of men in less than perfect (current) health. Further digging showed that the high degree of non-linearity represented in the Lagged Health terms may result from two different types of health dynamics being confounded in this model. One type is characterized by progressive change in health status with moderate inertia and the other type is characterized by onset-recovery sequences that apply only to those at or near perfect health. The latter could arise from accidental injuries that lead to a complete recovery; a phenomenon that was demonstrated by the observation that about 40% of those with perfect health at t-2 had perfect health at t+2 regardless of what their health status was at t.

We had assumed that transitions directly from perfect health were special. It also appears are transitions from recently perfect health are special. Perhaps separate models would be more appropriate in order to differentiate transitions odds for those with perfect lagged health from those with less than perfect lagged health. And so, we should probably conclude that our model is still inadequate and that no single model is likely to be able to encompass both types of health dynamic simultaneously.

4. Conclusions

Our illustration of bootstrap/cross-validation methods represents an exploratory approach to data analysis. This is an approach that can be highly effective in uncovering inadvertent effects of simplistic modeling; but that, without care, also runs a high risk of over-fitting and over-optimistic evaluation. Moreover, the greater the extent of interaction with the data during the model selection phase of analysis, the less valid conventional (unconditional) significance tests will be. Cross-validation is a useful tool in the assessment of alternative exploratory models and deserves wider use by the analytical community. In our illustration of the techniques, cross-validation made a deep exploration of a key scientific question involving the dynamics of health relatively easy, where conventional approaches would have required technical virtuosity and/or greater expenditure of time and effort.

We have demonstrated that cross-validation techniques may be put in a design-based setting. In that setting, we can expect, using cross-validatory techniques, to identify preferred models that are similar, but not necessarily identical, to those that might be identified using conventional inferential procedures.

Given the increasing availability of sets of bootstrap weights to aid users accounting for complex survey designs; we would encourage further research into use of combined bootstrap/cross-validation techniques. In our view, a promising direction for further research may be the use of some bootstrap replicate samples for Training (trial model estimation) with simultaneous use of the unsampled PSU's used for Validation (model selection), while reserving some bootstrap replicate samples for Testing (final model assessment).

References

Binder, David and Roberts, Georgia (2006), "Approaches for Analyzing Survey Data: a Discussion," *2006 Joint Statistical Meetings – Section on Survey Research Methods*, 2771-2778.

Carey, V., Zeger, S.L., and Diggle, P. (1993), "Modelling Multivariate Binary Data with Alternating Logistic Regressions," *Biometrika*, 80, 517-526.

Efron, Bradley (1978). "Regression and ANOVA with Zero-One Data: Measures of Residual Variation?" *Journal of the American Statistical Association*, 73, pp.113-121.

Efron, Bradley (1986). "How Biased Is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association*, 81, 461-470.

Efron, Bradley and Tibshirani, Robert, (1997). "Improvements on Cross-Validation: The .632+ Bootstrap Method," *Journal of the American Statistical Association*, 92, 548-560.

Feeny,D., Furlong,W., Torrance, G., Goldsmithh, C.H., Zhu, Z., DePauw, S., Denton, M., and Boyle, M. (2002). "Multi-attribute and single-attribute utility functions for the Health Utilities Index Mark 3 system." *Medical Care*, 40(2), 113-128.

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Picard, R.R. and Cook, R.D. (1984). "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575-583.

Rao, J.N.K., Wu, C.F.J. and YUE, K. (1992). "Some resampling methods for complex surveys," *Survey Methodology*, 18, 209-217.

Shao, Jun (1993). "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486-494.

Statistics Canada (1999). *Information about the National Population Health Survey*, Catalogue No. 82F0068XIE, www.statcan.ca/english/concepts/nphs/index.htm

Yeo, Douglas, Mantel, Harold and Liu, Tzen-Ping (1999). "Bootstrap Variance Estimation for the National Population Health Survey," *1999 Joint Statistical Meetings – Section on Survey Research Methods*, 778-783.