

# Can Survey Bootstrap Replicates be used for Cross-validation?

**David Binder**

**(joint work with Geoff Rowe)**

**SSO 2008 Fall Seminar**

**November 21, 2008**



**Statistics  
Canada**

**Statistique  
Canada**

# Outline of Presentation

---

1. Some background theory using survey bootstraps
2. Illustration of exploratory data analysis utilizing cross-validation for model evaluation and selection

# What is Cross-Validation?

---

## **K-fold cross-validation**

Partitioned original sample into  $K$  subsamples. Of the  $K$  subsamples, a single subsample is retained as the validation data for testing the model, and the remaining  $K - 1$  subsamples are used as training data.

# What is Cross-Validation?

---

## **Leave-one-out cross-validation**

Using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data.

# What is Cross-Validation?

---

## Usual Assumption

The validation set is independent from the training set.

# The Rao-Wu-Yue Bootstrap

We select a bootstrap replicate 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.

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

When  $m_h = n_h - 1$ , this simplifies to

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

# The Rao-Wu-Yue Bootstrap

---

Considering

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

it turns out under the design-based randomization, these replicates have mean equal to zero, and they are uncorrelated.

**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.**

# Cross-validation based on Unsampled PSU's

---

In each bootstrap replicate, there will be some psu's that are not included in the replicate sample. We consider the properties of estimates based on these unsampled psu's.

# Cross-validation based on Unsampled PSU's

---

We let

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

where  $\tilde{z}_{hi}^{(b)}$  is the indicator variable for whether the  $i$ th psu in the  $h$ th stratum is not in the  $b$ th bootstrap replicate. Then  $\tilde{Y}_h^{(b)}$  is design-unbiased for  $\hat{Y}_h$ . Properties of this new cross-validation sample need to be studied.

# Applying the methods

---

Application of these bootstrap/cross-validation methods will be demonstrated with an analysis from a Canadian health survey.

- Akaike's Criterion (AIC) using full sample
- Deviance or MSE using unsampled PSUs as the validation sample
- p-values – significance tests using unsampled replicates (2-fold cross-validation)

# National Population Health Survey (NPHS)

---

Panel survey of self-reported health and health dynamics in the Canadian population with interviews conducted at 2 year intervals.

NPHS files come with 500 sets of survey weights - allowing 500 bootstrap regression estimates matched with 500 cross-validation assessments of prediction error.

# Health Status Measure: HUI

---

Health Utility Index (HUI): a generic index of health status:

*[www.healthutilities.com/hui3.htm](http://www.healthutilities.com/hui3.htm)*

Eight attributes: vision, hearing, speech, mobility, dexterity, cognition, emotion, and pain.

1	=>	Perfect Health
0	=>	As Good As Dead
<0	=>	Worse Than Dead

# Logistic Regression: HUI Change Specification

---

$$\text{Prob}(HUI_{i,t+2} \neq HUI_{i,t}) = \left( 1 + \exp(-X_{i,t}\beta) \right)^{-1}$$

## Predictors

<b>Age</b>	<b>spline basis functions: <math>\ln(\text{age}+1)</math></b>
<b>Current Health</b>	<b>spline basis functions: <math>HUI_t</math></b>
<b>Lagged Health</b>	<b>spline basis functions: <math>HUI_{t-2}</math></b>
<b>Immigrant</b>	<b>indicator</b>
<b>Spouse Present</b>	<b>indicator</b>
<b>Education</b>	<b>Secondary or Some Post-Secondary</b>

# Measures of Model Error

## Akaike's Criterion (AIC) – full sample

$$\text{AIC} = -2 \left( \sum_i W_i \ln \left( f \left[ Y_i \mid \hat{\theta} \right] \right) \right) + 2 \text{ model df}$$

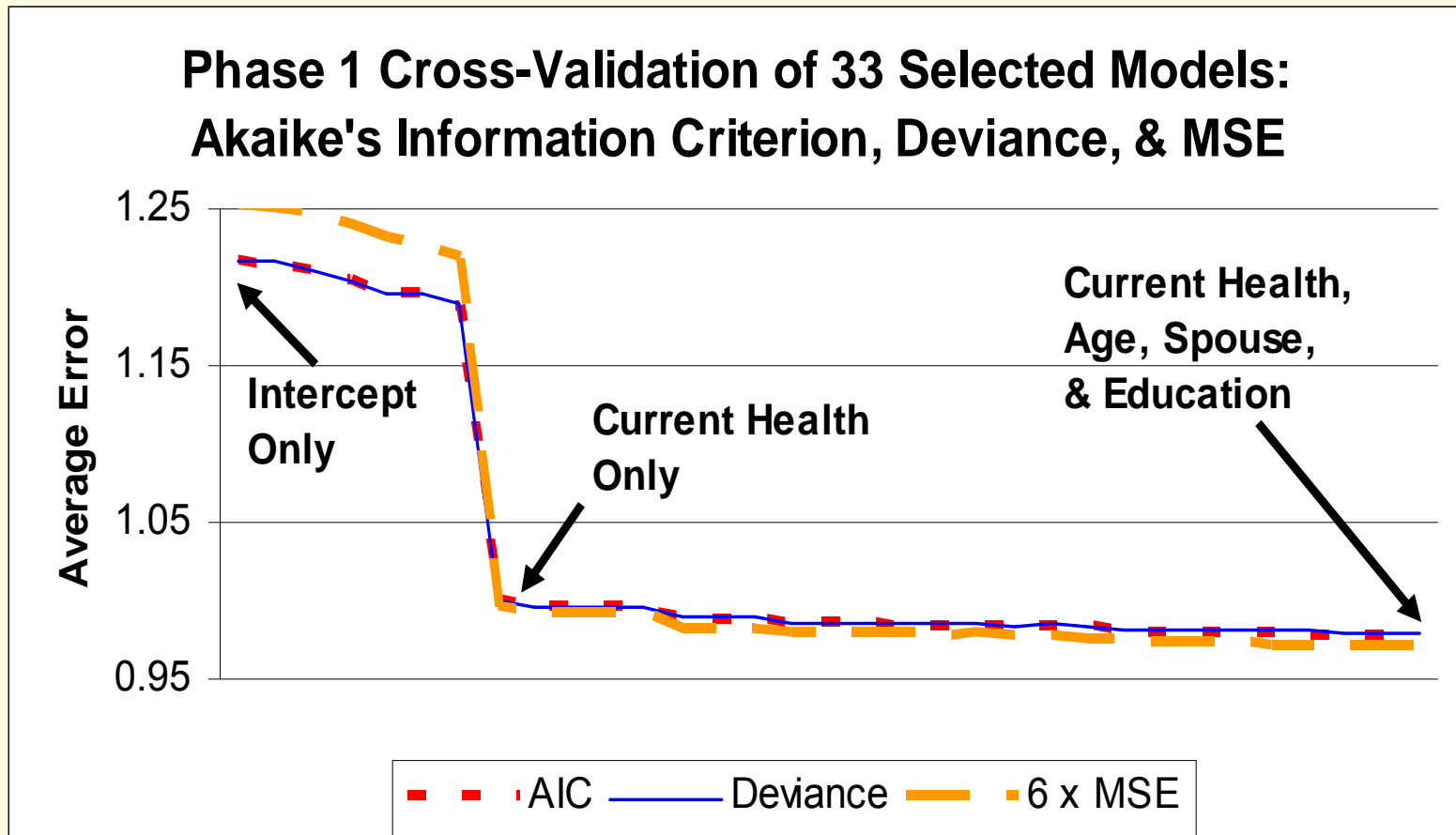
## Deviance – bootstrap replicates/unsampled PSUs

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

## MSE – bootstrap replicates/unsampled PSUs

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

# Specification Search: 1<sup>st</sup> Phase Results



# Specification Search: 2<sup>nd</sup> Phase

---

## 1) Additional Predictors

Perfect Lagged Health

indicator ( $HUI_{t-2} = 1$ )

Age at immigration

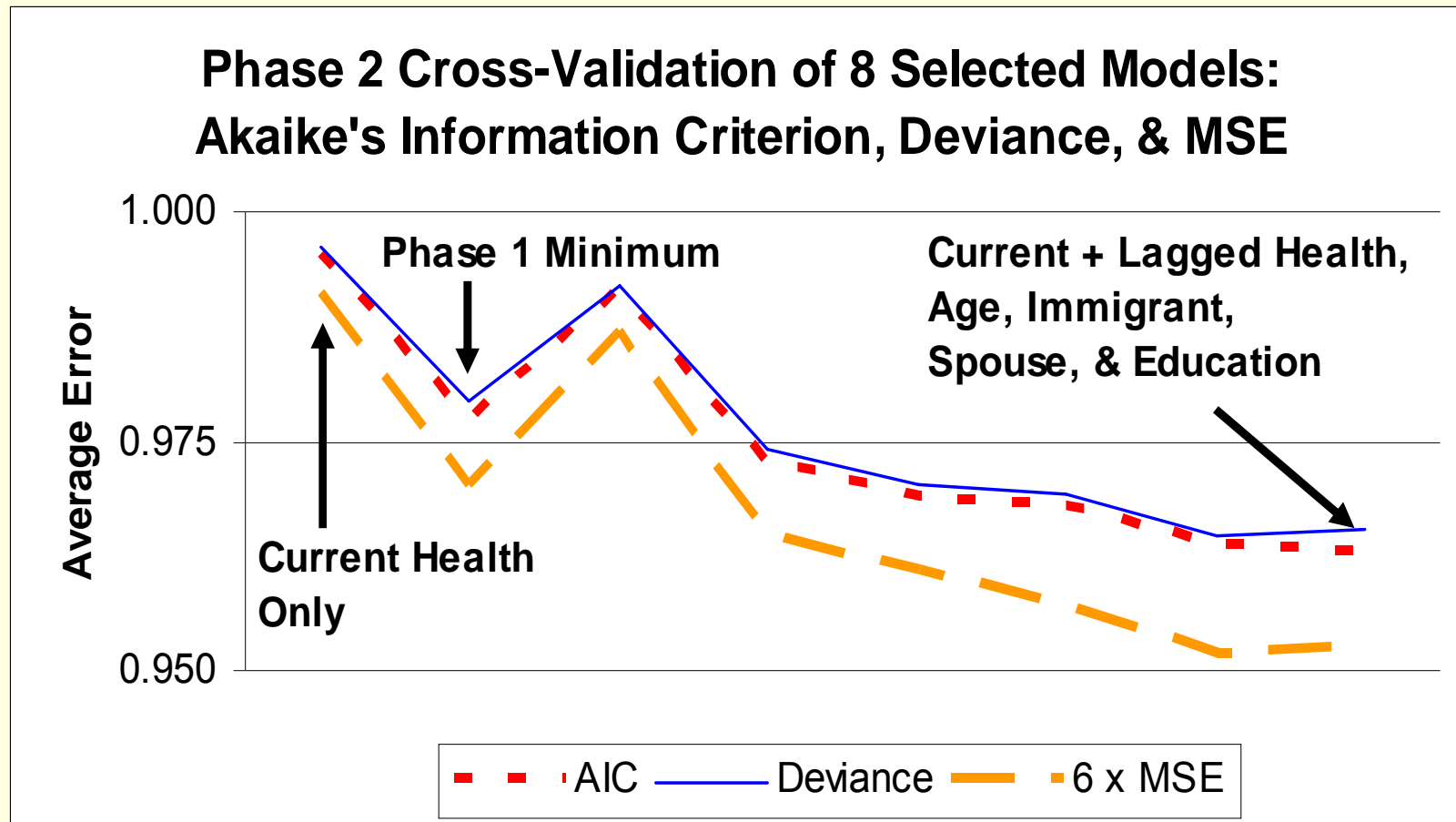
spline basis functions

## 2) Refining the Splines: Knot Placement Searches

Random searches using an ad hoc SAS macro.

No significance tests are directly available to assess knot location parameters, but cross-validation will measure resulting improvement in prediction.

# Specification Search: 2<sup>nd</sup> Phase Results



# Conclusions

---

- Exploratory data analysis inevitably requires interaction with the data and hence the risk of over-fitting and overly optimistic model assessment.
- The combination of bootstrap with cross-validation methods provides a powerful tool for modeling survey data.
- Some cross-validation methods can be shown to be valid for complex survey data.
- We regard these methods as an important area for further research.



Statistics  
Canada

Statistique  
Canada

---

For more information  
please contact

Pour plus d'information  
veuillez contacter

**Geoff Rowe** [Geoff.Rowe@statcan.gc.ca](mailto:Geoff.Rowe@statcan.gc.ca)

**David Binder** [dbinder49@hotmail.com](mailto:dbinder49@hotmail.com)