Methodological aspects of small area estimation from the National Electronic Health Records Survey (NEHRS).

Vladislav Beresovsky
National Center for Health Statistics
3311 Toledo Road Hyattsville, MD 20782

Janey Hsiao
National Center for Health Statistics
3311 Toledo Road Hyattsville, MD 20782

February 11, 2014

Abstract

Estimates of the proportions of physicians using electronic medical/health records (EMR/EHR) systems by state are important in monitoring the progress of adopting this technology in the US. The National Electronic Health Records Survey (NEHRS) collects information on EMR/EHR systems and has been used to report official design-based estimates of adoption at the state-level. More efficient estimates may be obtained using area-level models (Fay and Herriot, 1979), and are expected to be consistent and have smaller errors if a good set of model covariates is available. We demonstrate by simulations that relying on sample data for selecting a set of model covariates and finding an optimal vector of model parameters within this set results in overfitting of sample data and *increased* errors of estimates. Fit statistics provide a good measure of errors of estimates when model covariates are selected from information sources *independent* from sample data. Employing good independent information for covariate selection results in better model fit and smaller errors of estimates.

1 Introduction

Physicians' adoption of EMR/EHR systems has continued to climb as federal efforts have promoted their use in the context of implementation of the Patient Protection and Affordable Care Act (Hsiao et al., 2012). At the same time there are significant differences in adoption of EMR/EHR systems between states depending on local healthcare policies (Hsiao and Hing, 2012). It is important to have efficient in-state estimates of the proportion of physicians adopting such systems.

Prior to 2010 the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS) surveyed EMR/EHR systems adoption at the national level through the National Ambulatory Medical Care Survey (NAMCS) and its mail supplement (NEHRS). Starting in 2010, the NEHRS sample size was increased fivefold to allow for *state-level* estimates. Still, direct randomization-based estimates from NEHRS data have substantial sampling errors rendering many important analyses difficult or impossible.

Model-based methods are expected to produce more efficient estimates in small areas compared with direct estimates. There is extensive literature on such methods, beginning with the landmark paper by Fay and Herriot (1979). For many years these methods were successfully applied by the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program for estimating various poverty level indicators in small places. A methodology paper by Bell et al. (2007) describes application of area-level Fay-Herriot type models to single-year data from the American Community Survey (ACS) for estimating county- and state-level poverty of school-age (5-17 years) children.

Successful application of model-based methods largely depends on selecting a good set of model covariates (regressors). A large part of the paper by Bell et al. (2007) was devoted to a description of regressors and analysis of model fit criteria. The SAIPE program utilizes regressors available from administrative data sources, such as the number of IRS child tax exemptions, the number of Food Stamp Program participants and the number of children in poverty estimated from Census 2000. Common sense suggests that besides

being instrumental for fitting ACS *sample* data, these regressors should correlate with poverty indicators for the whole *population*.

However, the relationship of the data collected by healthcare surveys to the administrative records in general is not obvious. Two of the variables collected by NEHRS are the adoption of *any* and the adoption of a *basic* EMR/EHR system. It is assumed that a physician office is adopting *any* EMR/EHR system from the positive response to the survey question, "Does this practice use electronic medical records or electronic health records (not including billing records)?". Adoption of a *basic* EMR/EHR system requires implementation of *all* of the following functionalities: patient history and demographics, patient problem lists, physician clinical notes, comprehensive list of patients' medications and allergies, computerized orders for prescriptions, and ability to view laboratory and imaging results electronically (Hsiao and Hing, 2012).

One source of model covariates available to healthcare data analysts is the Area Resource File (ARF) distributed by the Health Resources and Services Administration (HRSA). It includes hundreds of county-level demographic, geographic, economic and healthcare covariates. Unfortunately, an *a priori* relationship between these covariates and local adoption of EMR/EHR systems is not as clear as the relationship between poverty data collected by the ACS and administrative regressors used by the SAIPE program.

Therefore, an analyst faces the problem of selecting an appropriate set of model covariates from the large number of ARF covariates based on the available sample. In our simulations we observed that finding the optimal parameter vector using exclusively sample data, results in overfitting of the sample data that translates into *increased* mean squared errors (MSE) of the estimates. Overfitting would not happen if the parameter space was based on the finite population data while the estimates of the model parameters within that space were found from fitting sample data. In the latter case, models with better fit criteria, such as AIC, provide for smaller MSE of the estimates. In fact, we found that a positive association between AIC and MSE takes place when the parameter space is defined using any source of information independent of sample data, such as prior information or a separate data set. This independently discovered parameter space may or may not provide for a good model fit, but overfitting does not happen in this case and AIC remains a reasonable measure of MSE.

In Section 2 we formulate a model with random effects at the area level used to estimate in-state proportions of adoption of any and adoption of a basic EMR/EHR system by physicians and present an EBLUP estimator for proportions. Ways of meaningful aggregation of county-level covariates from the ARF at the state level are also discussed. The design of our simulation experiment is explained in Section 3. Section 4 is devoted to presenting simulation results and comparing different trends between fit criteria and MSE. In the final section we further discuss the observed results and draw some conclusions.

2 Area-level model for estimating proportions

Some of the Fay-Herriot area-level models described by Bell et al. (2007) used ACS estimates of the (log) number of school-age children in poverty by county and state as the dependent variable. Other models had the logarithm of the direct survey estimate of poverty rate among school-age children as the dependent variable. We used the latter approach, defining the dependent variable as the logarithm of the direct NEHRS estimate of the proportion of office-based physicians adopting either any or basic EMR/EHR systems for state i:

$$\log(y_i) = \log(Y_i) + e_i \quad \text{where} \quad e_i \sim ind\, N(0, v_i) \tag{1a}$$

$$\log(Y_i) = x_i'\beta + u_i \quad \text{where} \quad u_i \sim iid\, N\left(0, \sigma_u^2\right) \tag{1b}$$

where, for state i,

y_i = direct survey estimate of dependent variable;
Y_i = true population value of dependent variable;
$e_i = \log(y_i) - \log(Y_i)$ = sampling error of estimating $\log(Y_i)$;
v_i = direct survey estimate of the variance of sampling error e_i;
$x_i' = 1 \times r$ vector of model covariates for state i;

$\beta = r \times 1$ vector of model parameters;
u_i = random effect on state level;
σ_u^2 = variance of the random effect.

The unknown parameters β and σ_u^2 defined by equations (1a) and (1b) can be estimated iteratively by the maximum likelihood method, or using the Carter and Rolph (1974) estimator described in Fay and Herriot (1979). Estimated parameters can be plugged into the standard formula (see Bell (1999) and Bell et al. (2007)) for an EBLUP of $\log(Y_i)$:

$$\widehat{\log(Y_i)} = (1 - w_i)\log(y_i) + w_i\left(x_i\hat{\beta}\right) \tag{2}$$

where $w_i = v_i / \left(\hat{\sigma}_u^2 + v_i\right)$.
The variance of the prediction error, ignoring the error of estimating variance of the random effect $\hat{\sigma}_u^2$, can be estimated as:

$$\mathrm{Var}\left[\widehat{\log(Y_i)}\right] = w_i\hat{\sigma}_u^2 + w_i^2\left(x_i\mathrm{Var}\left(\hat{\beta}\right)x_i\right) \tag{3}$$

On the original scale prediction of the proportion of physicians adopting either any or basic EMR/EHR systems by state is:

$$\hat{Y}_i = \exp\left(\widehat{\log(Y_i)}\right)\exp\left(\mathrm{Var}\left[\widehat{\log(Y_i)}\right]/2\right) \tag{4}$$

The area-level model (1b) employs covariates at the state level X_i^{st}. However, the ARF comprises demographic, economic and healthcare covariates at the county level $X_{ij}^c (j \in i)$, so they must be aggregated to the state level. The question is: what should be used as the basis for aggregation? The sample unit of both NAMCS and NEHRS is a physician. The numbers of physicians in counties $N_{ij}^{c,phys}$ are available from the sampling frame used in both surveys. We used them as weights in the weighted average of covariates at the state level:

$$X_i^{st} = \frac{\sum\limits_{j\in i} X_{ij}^c N_{ij}^{c,phys}}{\sum\limits_{j\in i} N_{ij}^{c,phys}} \tag{5a}$$

Some of the ARF covariates $X_{ij}^{c,pop}$ were proportional to the census counts in the counties $N_{ij}^{c,pop}$, for example the number of people who are black or older than 65 years. They were normalized by the county's census counts in order to minimize interdependence between overall census and the number of physicians in counties:

$$X_i^{st} = \frac{\sum\limits_{j\in i} X_{ij}^{c,pop}/N_{ij}^{c,pop}\ N_{ij}^{c,phys}}{\sum\limits_{j\in i} N_{ij}^{c,phys}} \tag{5b}$$

State covariates defined according to (5) maintain meaningful association with county covariates X_{ij}^c of the ARF for different possible distributions of physicians by counties j within state i. If, for instance, $N_{ij}^{c,phys}$ are the same for all counties, state covariates X_i^{st} are equal to the mean of corresponding county covariates $X_{ij}^c (j \in i)$. On the other hand, if all physicians reside in just one county j_0, state covariates are equal to covariates in that county $X_{ij_0}^c$.

3 Simulation experiment

We conducted non-parametric simulations using a two-step procedure that combines an inverse sampling process (Step 1) and a bootstrap resampling algorithm (Step 2) described by Sverchkov and Pfeffermann (2004). At Step 1 we generated a single"pseudo population" for each state i by selecting *with replacement* $N_i = \sum_{k\in i} w_{ik}$ physicians from the original sample with probabilities proportional to w_{ik}/N_i, where w_{ik} is the survey weight of physician k in state i after adjustment for non-response and post-stratification. At

Step 2 we drew $B = 1000$ bootstrap samples from the "pseudo population" generated in Step 1, by selecting without replacement within each state i the same number of physicians n_i as in the original sample with probability proportional to inverse survey weight $1/w_{ik}$.
We calculated "true" "pseudo population" proportions Y_i in small areas of physicians adopting *any* and adopting a *basic* EMR/EHR system. For every bootstrap sample we calculated design-based estimates of logarithms of these proportions $\log(y_i)$ and their sampling error variances v_i. The latter estimates were calculated using Taylor linearization. This provides all required data input for the area-level Fay-Herriot model (1).
Utilizing different methods for selecting model covariates and defining parameter space, we estimated parameters of model (1) and calculated corresponding model-based estimates (4) of proportions of physicians who adopted *any* and adopted a *basic* EMR/EHR system by state. In *Method 1*, model covariates were selected only once by fitting the model to "pseudo population" data. In *Method 2*, model covariates were selected by finding the best fit for *every* bootstrap sample. In *Method 3*, model covariates were also selected once from fitting a single arbitrary bootstrap sample. For all three methods, the optimal vector of model parameters within the defined parameter space was estimated from the data for *every* bootstrap sample.
Using"pseudo population" for defining model covariates (*Method 1*) corresponds to practical situations when analysts have good prior intelligence about meaningful correlation between the dependent variable and administrative data sources. The Census Bureau's SAIPE program would be a good example of such experience. However, in many cases analysts doing small area estimation have only sample data to rely upon for model selection (*Method 2*) and all other inferences. Sometimes an analyst may possess approximate knowledge about the association between the dependent variable and a predefined set of covariates. If an analyst chooses to rely on this knowledge (avoiding the temptation to "improve" the model fit using sample data for selecting "better" covariates), that would correspond to *Method 3*.
Using the methods described above for selecting covariates of model (1b), the following model-based estimates in small areas were calculated:

$\hat{Y}_i^0$– only-intercept model;

$\hat{Y}_i^2(1)$– intercept and 2 covariates defined from "pseudo population" (*Method 1*);

$\hat{Y}_i^{2+3}(1,1)$, $\hat{Y}_i^{2+3}(1,2)$, $\hat{Y}_i^{2+3}(1,3)$– same as for $\hat{Y}_i^2(1)$ plus ~ 3 more covariates defined by *Methods 1-3*;

$\hat{Y}_i^{2+8}(1,1)$, $\hat{Y}_i^{2+8}(1,2)$, $\hat{Y}_i^{2+8}(1,3)$– same as for $\hat{Y}_i^2(1)$ plus ~ 8 more covariates defined by *Methods 1-3*;

In the above notations, numbers in parentheses denote methods of defining a set of model covariates and numbers in superscript denote the number of covariates utilized by the model. For example, for the estimator $\hat{Y}_i^{2+8}(1,2)$ first two model covariates were selected using *Method 1* and then ~ 8 were added to improve fit of sample data (*Method 2*).
For every estimator $\hat{Y}_i$ listed above we calculated the root average MSE (RMSE) by averaging its squared deviation from "pseudo population" value Y_i over simulations $s = (1, ..., B)$ and states $i = (1, ..., N_s)$:

$$\text{RMSE}\ \ \hat{Y}_i\ \ = \sqrt{\frac{1}{N_s B}\sum_{i=1}^{N_s}\sum_{s=1}^{B}\ \ \hat{Y}_i - Y_i\ \ ^2} \tag{6}$$

4 Results and conclusions

Results of the simulations are presented in Figure 1 below. In agreement with the well-known James and Stein (1961) theorem proving inadmissibility of the direct estimator with dimension $k \geq 3$ under the quadratic loss function, even the estimator $\hat{Y}_i^0$ utilizing the only-intercept model has on average smaller RMSE than the direct estimator y_i in small areas.
Furthermore, when covariates of a model utilized by an estimator were selected using reliable prior information (*Method 1*), then adding more covariates resulted in better model fit (smaller AIC) and smaller RMSE of an estimator:

$$\text{AIC}\ \ \hat{Y}_i^2(1)\ \ > \text{AIC}\ \ \hat{Y}_i^{2+3}(1,1)\ \ > \text{AIC}\ \ \hat{Y}_i^{2+8}(1,1) \tag{7a}$$

and

$$\text{RMSE}\ \ \hat{Y}_i^2(1)\ \ > \text{RMSE}\ \ \hat{Y}_i^{2+3}(1,1)\ \ > \text{RMSE}\ \ \hat{Y}_i^{2+8}(1,1) \tag{7b}$$

The opposite happened when covariates were added to the model conditional on achieving better fit of sample data (*Method 2*). Models with better fit corresponded to estimators having larger RMSE:

$$\text{AIC}\ \ \hat{Y}_i^2(1)\ \ > \text{AIC}\ \ \hat{Y}_i^{2+3}(1,2)\ \ > \text{AIC}\ \ \hat{Y}_i^{2+8}(1,2) \tag{8a}$$

but

$$\text{RMSE}\ \ \hat{Y}_i^2(1)\ \ < \text{RMSE}\ \ \hat{Y}_i^{2+3}(1,2)\ \ < \text{RMSE}\ \ \hat{Y}_i^{2+8}(1,2) \tag{8b}$$

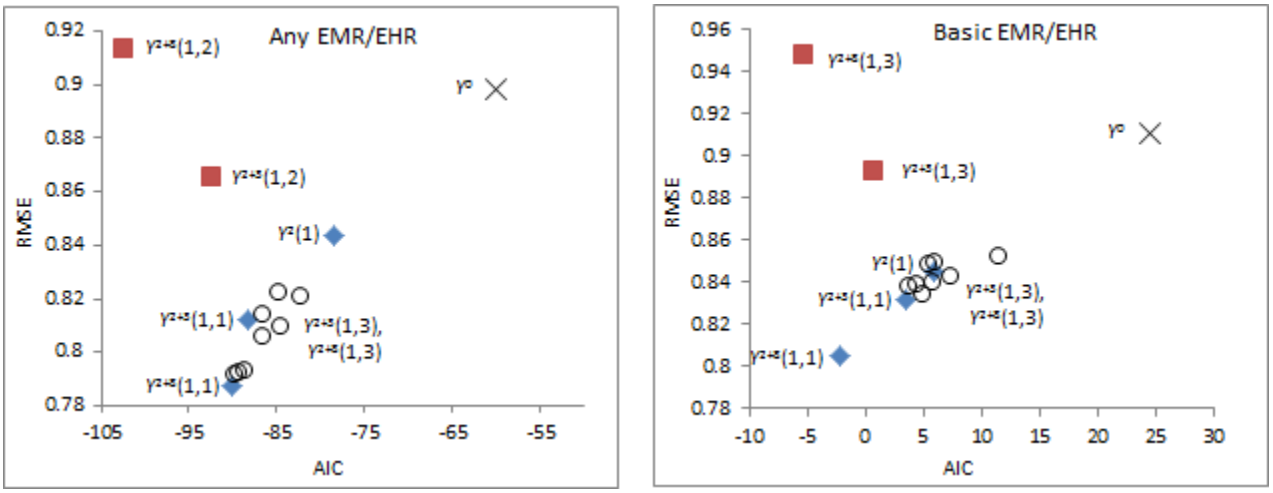

Figure 1: Dependence of the root average MSE (RMSE) on the model fit criteria AIC for different model-based estimators of proportions of adoption of any and basic EMR/EHR systems by physicians in states. RMSE of model-based estimators is measured relative to RMSE of direct estimator y_i. $\times-$ intercept-only model; $\blacklozenge-$ covariates defined from fitting "pseudo population"; $\blacksquare-$ covariates defined from fitting every bootstrap sample; $\bigcirc-$ covariates defined from fitting the first bootstrap sample were used for the rest.

Open circles on Figure 1 denote AIC of models and RMSE of corresponding estimators when model covariates were selected conditional on fitting the *first* bootstrap sample ($s = 1$) and then used for fitting the rest of bootstrap samples' data ($s = 2, ..., B$) (*Method 3*). We ran simulations under these conditions multiple times, changing the initial seed of the random number generator and thus generating different first bootstrap samples. Fitting models to these samples resulted in different sets of model covariates for each run of simulations. Average model fit (AIC) and error of estimators in small areas (RMSE) differed between runs of simulations, depending on proximity of the data distribution in the first sample to the "pseudo population" distribution. Increasing dimension of parameter space did not always result in decrease of RMSE, as happened when there was reliable prior information about model covariates (7). Neither did we observe an increase of RMSE, associated with overfitting of sample data (8). For different simulations, RMSE $\hat{Y}_i^{2+8}(1,3)$ could be either larger or smaller than RMSE $\hat{Y}_i^{2+3}(1,3)$. There was, however, a net positive effect from adding extra covariates for estimating adoption of *any* EMR/EHR system but no such effect for estimating adoption of a *basic* EMR/EHR system:

$$\text{RMSE}\ \ \hat{Y}_i^2(1)\ \ < \text{avg}\ \ \text{RMSE}\ \ \hat{Y}_i^{2+3}(1,3)\ \ , \text{RMSE}\ \ \hat{Y}_i^{2+8}(1,3)\ \ \text{, for any EMR/EHR} \tag{9a}$$

$$\text{RMSE}\ \ \hat{Y}_i^2(1)\ \ \sim \text{avg}\ \ \text{RMSE}\ \ \hat{Y}_i^{2+3}(1,3)\ \ , \text{RMSE}\ \ \hat{Y}_i^{2+8}(1,3)\ \ \text{, for basic EMR/EHR} \tag{9b}$$

In conclusion, overfitting can be characterized as a misleading situation when a model better fitting sample data results in estimators having larger RMSE. This happens when sample data are used for defining the set of model covariates (parameter space). If model covariates are set in advance using either reliable or partially reliable information and the optimal vector of model parameters is found within the predefined parameter

space, then overfitting does not occur and AIC of the model fit reliably predicts RMSE of estimators in small areas.

Analysts often repeat the rhetorical wisdom stating that model-based estimators are efficient only when the underlying model is "correct". The meaning of the word "correct" remains mysteriously unexplained. Let us define "correctness" of a model when there is proper correspondence between model fit criteria (AIC) and RMSE of the corresponding estimators. In the course of conducted simulations we found that the Fay-Herriot predictive model is always "correct" when parameter space is defined from external sources of information rather than being conditional on sample data.

In small area estimation problems dealing with health data, reliable external sources of information about explanatory powers of model covariates are not always available. In other applications, particularly in data mining, there is enough data to allocate for both model fitting and model validation. Ultimately, a working population model is selected based on its ability to fit data reserved for validation. However, in small area estimation problems there is usually barely enough data for estimating model parameters and making predictions with reasonably small errors. In such cases, other methods of avoiding overfitting of sample data must be considered.

One of these methods is Bayesian model averaging (BMA), which tackles the problem by estimating models for all possible combinations of covariates and constructing a weighted average over all of them. Hoeting et al. (1999) provided a thorough introduction to BMA. George and Foster (2000) advocated an "Empirical Bayes" approach by using information contained in the data $(y; X)$ to elicit weight of models via maximum likelihood. Application of model averaging to small area estimation problems will be a subject for future research.

5 Acknowledgment

The authors (V.B.) wish to express gratitude to Alan Dorfman for discussions and insights and to Susan Schappert for improving the use of language of this paper.

References

W. Bell, B. Wesley, C. Cruse, L. Dalzell, J. Maples, B. O'Hara, and D. Powers. *Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties*. U.S. Census Bureau, Methodology Paper, December 2007. http://www.census.gov/did/www/saipe/methods/index.html.

William R. Bell. *Accounting for Uncertainty About Variances in Small Area Estimation*. U.S. Census Bureau, Methodology Paper, 1999. http://www.census.gov/hhes/www/saipe/asapaper/Bell99.pdf.

Grace E. Carter and John E. Rolph. Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities. *Journal of the American Statistical Association*, 69:880–885, 1974.

Robert E. Fay and Roger A. Herriot. Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74:269–277, 1979.

E. George and D. Foster. Calibration and Empirical Bayes Variable Selection. *Biometrika*, 87:731–747, 2000.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14:382–417, 1999.

C.J. Hsiao and E. Hing. *Use and Characteristics of Electronic Health Record Systems Among Office-based Physician Practices: United States, 20012012*. National Center for Health Statistics, Data Brief, No. 111, December 2012. http://www.cdc.gov/nchs/data/databriefs/db111.htm.

C.J. Hsiao, E. Hing, T.C. Socey, and B. Cai. *Electronic health record systems and intent to apply for meaningful use incentives among office–based physician practices: United States, 2001 - 2011*. National Center for Health Statistics, Data Brief, No. 79, February 2012. http://www.cdc.gov/nchs/data/databriefs/DB79.pdf.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, 1961.

Michail Sverchkov and Danny Pfeffermann. Prediction of Finite Population Totals Based on the Sample Dstribution. *Survey Methodology*, 30:79–92, 2004.