Supplementary Material: Additional Real-data Example

Jiming Jiang, P. Lahiri and Thuan Nguyen
University of California, Davis, University of Maryland, College Park, and Oregon Health & Science University

Throughout this Supplementary Material, all of the equation numbers, section numbers, and references refer to the paper, “A Unified Monte-Carlo Jackknife for Small Area Estimation after Model Selection”, by J. Jiang, P. Lahiri, and T. Nguyen.

Datta, Lahiri, and Maiti (2002) considered a data set on median income of four-person families for the fifty states of U.S. and the District of Columbia using cross-sectional and time series modeling. The primary source of data is the annual supplement to the March Sample of the Current Population Survey (CPS), which provides individual annual income data categorized into intervals of $2500. The direct survey estimates were obtained from the CPS using linear interpolation. Two secondary sources of data were also available. The first source is the U.S. decennial census (Census) which produces median incomes for the 50 states and D.C. based on the “long form” filled out by approximately one-sixth of the U.S. population. These census median income estimates are believed to have negligible sampling errors. The second source is the Bureau of Economic Analysis (BEA) division of the U.S. Department of Commerce, which produces per-capita income estimates. Since the per-capita income estimates are not based on any sampling techniques, they do not have any sampling errors associated with them. From the Census and BEA data, an adjusted census median income (adjusted Census) is obtained by multiplying the preceding census median income by the ratio of BEA per-capita income for the current year to that of the preceding census year.

Following Datta *et al.* (2002), we consider the four-person families data for the years 1979, and 1981–1989, that is, a total of 10 years. The direct survey estimates are denoted by y_{it}, where $i = 1, \dots, 51$ corresponding to the 50 states and D.C., and $t = 1, \dots, 10$, corresponding to the 10 years. The Census variable is denoted by $x_{1,i}$ (note that this variable is at the state level only, i.e., does not change with the year), and the adjusted Census variable is denoted by $x_{2,it}$ [note that this variable is at both state and time (year) levels]. The goal is to estimate the median income of four-person families for all 50 states of U.S. and the D.C. for the year 1989. Datta *et al.* (2002) used x_2 as the only covariate in their modeling for the mean. As for the variance-covariance structure, the authors proposed a random walk model for the state-level (vector-valued) random effects, in addition to the sampling errors whose variances are known. We refer this model as DLM model. In a similar context, Rao and Yu (1994) proposed a cross-sectional/time series model, which is the same as the DLM model except that the variance-covariance structure of the random effects follows that of a stationary AR(1) model. This model is referred to as R-Y model. We intend to compare the DLM model and R-Y model in this particular application. In addition, we are interested in entertaining an additional covariate variable, namely, x_1. Thus, we set up a model-selection framework as follows.

Let $y_i = (y_{it})_{1\le t\le 10}$ denote the vector of direct survey estimators for state i (including D.C.). Let $X_i = (x'_{it})_{1\le t\le 10}$ denote the matrix of covariates for state i, where the first component of x_{it} is 1, corresponding to the intercept, and the rest of the components are subject to selection from $x_{1,it} = x_{1,i}$, $x_{2,it}$, or both. A candidate model can be expressed as

$$y_i = X_i\beta + u_i + e_i, \quad i = 1, \dots, 51, \tag{A.1}$$

where the vectors u_i, e_i are independent with $u_i \sim N(0, \Sigma)$ and $e_i \sim N(0, \Psi_i)$. Here Ψ_i is a diagonal matrix whose diagonal elements are known sampling variances (strictly speaking, those are estimated using within area observations). The covariance matrix Σ is either $\Sigma_{(1)}$, corresponding to the DLM model, or $\Sigma_{(2)}$, corresponding to the R-Y model. More specifically, $\Sigma_{(1)} = \sigma_v^2 J_{10} + \sigma_\epsilon^2 \Gamma\Gamma'$, with J_{10} being the 10×10 matrix of 1's, and Γ being the 10×10 lower triangular matrix with diagonal and non-zero off-diagonal elements equal to 1, and $\sigma_v^2, \sigma_\epsilon^2$ are unknown variance components. $\Sigma_{(2)} = \sigma_v^2 J_{10} + \sigma_\epsilon^2 \Sigma(\rho)$, with $\Sigma(\rho) = (1-\rho^2)^{-1}[\rho^{|t-s|}]_{1\le s,t\le 10}$, and $\sigma_v^2, \sigma_\epsilon^2, \rho$ are unknown variance components with $|\rho| < 1$. In all, (A.1) includes six candidate models: x_1, x_2, or both, for the covariates, and $\Sigma_{(1)}$ or $\Sigma_{(2)}$ for Σ. Also, to apply McJack we need to have a full model that covers all of the candidate models as special cases, and there is no such a model among the candidate models. Note that, although x_1, x_2 is a full model for the covariates, the DLM and R-Y covariance models do not cover each other as special cases. Therefore, we consider the following full model which is not a candidate model–it is only used for the McJack computation: A convex linear combination of $\Sigma_{(1)}$ and $\Sigma_{(2)}$:

$$\Sigma_{\rm f} = \sigma_v^2 J_{10} + \sigma_\epsilon^2\{\lambda\Gamma\Gamma' + (1-\lambda)\Sigma(\rho)\},$$

where $\lambda \in [0, 1]$ is an additional variance component. It is clear that, $\Sigma_{\rm f}$ includes $\Sigma_{(1)}$ and $\Sigma_{(2)}$ as special cases. The full model is the one with x_1, x_2 as the covariates and $\Sigma_{\rm f}$ as Σ.

The BIC procedure is used to select the optimal model among the candidates. The selected model has both x_1, x_2 as the covariates, and $\Sigma_{(1)}$ as Σ. This model is denoted by M^*. Based on the selected model, maximum likelihood estimation (MLE) is used to obtain estimates of the model parameters, and the EBLUPs for the 1989 median income of 4-person families are computed. The results are presented in Table A.1. Also presented are the corresponding square roots of the MSPE estimates using McJack, taking into account of the model selection, denoted by $\widehat{\rm MSPE}_1^{1/2}$. As a comparison, we also computed the McJack MSPE estimates based on M^* (assumed known), denoted by $\widehat{\rm MSPE}_2^{1/2}$. All numbers are rounded to the nearest integers. The Monte-Carlo sample size for computing the McJack estimates is $K = 2000$. It is seen that the two MSPE estimates are very close; for some states one MSPE estimate is slightly larger while for the other states it is slightly smaller. At first, this might seem a little surprising, as one would expect that $\widehat{\text{MSPE}}_1$, which takes into account the potentially additional variation in model selection, to be larger than $\widehat{\text{MSPE}}_2$. However, MSPE is not just the variation. This can be seen from the equation

$$\text{MSPE} = (\text{prediction bias})^2 + \text{prediction variance}. \qquad (A.2)$$

Although model selection may increase the second term on the right side of (A.2), it may reduce the first term for selecting the optimal model, which fits the data better. Note that, although M^* is the optimal model for this particular data set, it may not always be the optimal model, if the data are repeatedly generated, even under the same (estimated) parameters. For example, out of the 2,000 Monte-Carlo samples generated under the M-estimate based on the full data set, $\hat{\psi}$, 85.3% selected M^*; another 14.1% selected the model with only x_2 as the covariate, and the same $\Sigma = \Sigma_{(1)}$. Note that this is the same model as the one used by Datta *et al.* (2002), denoted by M_{dlm}. Take a look at another example by considering the M-estimate based on the data with the jth state deleted, $\hat{\psi}_{-j}$, where $j = 20$ (the number 20 is randomly chosen from $1, \dots, 51$ by a computer). This time, out of the 2,000 Monte-Carlo samples, 81.8% selected M^*, and another 17.8% selected M_{dlm}. To the end, it all depends on the relative contributions of the two terms on the right side of (A.2), when it comes to the MSPE measure. It appears that, for this application, the overall MSPE is about the same for the two McJack estimates. This may also be explained by the fact that there is not much variation in model selection after all. Once again, using the above examples, for the majority of the Monte-Carlo samples (85.3% and 81.8%) one has M^* as the selected model, and almost all of the rest (14.1% and 17.8%) M_{dlm} as the selected model. Note that the two models, M^* and M_{dlm}, are actually very close, especially in terms of the prediction performance, and the two models combined counted for 99.4% and 99.6% of the selected models, respectively, for those two examples.

For a similar reason, the MSPE estimates are not necessarily smaller than the corresponding $\Psi_{i,10}$, which is the (estimated) sampling variance of the direct estimator. As noted earlier, MSPE is not just a measure of variation. For the state of FL, for example, bias appears to be a more significant factor than variation. One standard practice in SAE has been to compare standard second order MSE estimates of EBLUP with the corresponding sampling variance estimates of the direct estimator. This is not a fair comparison because the MSE estimates are derived using the same model that generates the estimates, so there is an issue of “double-dipping”. By including variations due to model selection, we are at least making an effort to make the comparison more fair than the standard practice.

Table A1: **Estimation of 1989 Median Income of 4-Person Families**

State	EBLUP	$\widehat{\text{MSPE}}_1^{1/2}$	$\widehat{\text{MSPE}}_2^{1/2}$	State	EBLUP	$\widehat{\text{MSPE}}_1^{1/2}$	$\widehat{\text{MSPE}}_2^{1/2}$
ME	38360	2554	2729	NC	38482	682	663
NH	48938	360	370	SC	36039	1092	1130
VT	39056	1892	1841	GA	40113	1487	1326
MA	52018	851	885	FL	37294	3148	3102
RI	44918	337	360	KY	33156	579	586
CT	55118	687	678	TN	34231	1577	1637
NY	44199	240	238	AL	33394	3116	3058
NJ	54902	438	430	MS	30196	2350	2412
PA	39871	1224	1316	AR	29874	1529	1423
OH	40629	418	435	LA	34659	1606	1606
IN	38085	1616	1548	OK	35330	695	669
IL	43540	7236	7194	TX	36216	722	771
MI	43070	975	957	MT	32835	5667	5892
WI	41108	756	757	ID	32335	1900	1970
MN	42940	273	265	WY	36368	1084	1093
IA	36765	457	471	CO	40329	124	126
MO	38425	952	946	NM	29449	3581	3543
ND	34019	1400	1167	AZ	37962	1306	1330
SD	32323	1685	1723	UT	36856	1039	1007
NE	35819	207	218	NV	40622	876	926
KS	38774	1273	1306	WA	42001	1489	1526
DE	42791	463	470	OR	39566	1748	1699
MD	52395	468	489	CA	42586	717	737
DC	40232	7093	7304	AK	47359	990	1004
VA	46561	2156	2115	HI	45780	2360	2273
WV	29813	829	874				