Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2006
Methodological Issues in Measuring Population Health

Using the *t*-distribution to Deal with Outliers in Small Area Estimation

William R. Bell and Elizabeth T. Huang[1]

Abstract

Small area estimation using linear area level models typically assumes normality of the area level random effects (model errors) and of the survey errors of the direct survey estimates. Outlying observations can be a concern, and can arise from outliers in either the model errors or the survey errors, two possibilities with very different implications. We consider both possibilities here and investigate empirically how use of a Bayesian approach with a *t*-distribution assumed for one of the error components can address potential outliers. The empirical examples use models for U.S. state poverty ratios from the U.S. Census Bureau's Small Area Income and Poverty Estimates program, extending the usual Gaussian models to assume a *t*-distribution for the model error or survey error. Results are examined to see how they are affected by varying the number of degrees of freedom (assumed known) of the *t*-distribution. We find that using a *t*-distribution with low degrees of freedom can diminish the effects of outliers, but in the examples discussed the results do not go as far as approaching outright rejection of observations.

KEY WORDS: Fay-Herriot model; robustness; bivariate model.

1. Introduction

Small area estimation often applies linear models, such as the Fay-Herriot (1979) model, to direct survey estimates for the small areas. An important example is the U.S. Census Bureau's SAIPE (Small Area Income and Poverty Estimates) program, which produces state and county poverty estimates from Fay-Herriot models applied to direct poverty estimates from the U.S. Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC, formerly known as the March income supplement). The models borrow information from regression variables related to poverty that are constructed from administrative records data, as well as from age group poverty estimates from the previous decennial census. For further information see the SAIPE web site at www.census.gov/hhes/www/saipe/. For simplicity, in what follows we shorten references to "CPS ASEC" to just "CPS."

Models such as the Fay-Herriot model typically assume normality of the error components in the model, namely, the area-level random effects (model errors) and the survey errors of the direct survey estimates. Some references have considered use of non-normal distributions for the model errors as a means of dealing with outliers. Most relevant to our work are the papers of Datta and Lahiri (1995) and Xie, Raghunathan, and Lepkowski (2005). Datta and Lahiri considered Bayesian treatment of a small area model where the model error follows a general mixture of normal distributions that includes the *t*-distribution as a special case. They derived prediction results for this model and showed that as a particular observation (which they assume is known) becomes successively more of an outlier (approaches $\pm\infty$) the Bayesian results effectively reject the outlier. Datta and Lahiri also review some other papers that provide related results for the simpler case of a simple mean model, that is, a model without covariates.

Xie, Raghunathan, and Lepkowski used a Bayesian treatment of a small area model with *t*-distributed model errors to estimate overweight prevalence using public-use data from the 2003 Behavioral Risk Factor Surveillance System. They derived a Markov Chain Monte Carlo (MCMC) approach to obtain simulations from the posterior distribution. In their model the degrees of freedom of the *t*-distribution is treated as an unknown parameter and is included in the

[1] William R. Bell, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA, e-mail: William.R.Bell@census.gov; Elizabeth T. Huang, same mailing address, e-mail: Elizabeth.T.Huang@census.gov. **Disclaimer**: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

MCMC calculations. In Section 3 we use a generically similar model but do the calculations for various fixed values of the degrees of freedom, because simulation studies we have done using the WinBUGS package (Spiegelhalter, et al. 2003) suggest to us that in this type of model the data typically provide limited information about the degrees of freedom, making inferences about this parameter difficult and imprecise.[2]

In this paper we consider extending the Fay-Herriot model by assuming either the model errors or the survey errors (but not both) follow a *t*-distribution. We apply such models to SAIPE data used for modeling U.S. state poverty ratios for school-age (5-17) children, taking a Bayesian approach and obtaining prediction results for various fixed values of the degrees of freedom, k. Setting $k = \infty$ takes us back to the usual Fay-Herriot model with the normal distribution. Our goal is to see how the use of the *t*-distribution affects results for particular states that might be suspected to be outliers. Section 2 reviews the basic SAIPE Fay-Herriot state model and data used, and considers the extension of the model to assume a *t*-distribution for either the model or survey errors. Section 3 applies the *t*-distribution model to the case of 5-17 poverty ratios in 1989, for which the direct survey estimate for Connecticut is a possible outlier. Section 4 reviews work of Huang and Bell (2006) on a *t*-distribution extension to a bivariate model that combines the CPS data with corresponding poverty rate estimates from the Census Bureau's American Community Survey (ACS). Here the *t*-distribution is assumed for the model or survey errors in the ACS equation to see how this affects results for a state (Alaska in 2002) for which the ACS estimate is a possible outlier. Finally, Section 5 summarizes the conclusions.

2. The SAIPE Fay-Herriot State Model and Extensions that Use the *t*-distribution

2.1 SAIPE production models

The general Fay-Herriot model, as applied in SAIPE to state poverty ratios for a particular year and age-group (0-4, 5-17, 18-64, or 65 and over), is as follows:

$$\begin{aligned} y_i &= Y_i + e_i \\ &= (\mathbf{x}_i'\boldsymbol{\beta} + u_i) + e_i \end{aligned} \tag{1}$$

where

- y_i = CPS direct estimated poverty ratio for state i
- Y_i = true population poverty ratio for state i
- $\mathbf{x}_i$ = vector of regression variables for state i
- $\boldsymbol{\beta}$ = vector of regression parameters
- u_i = state i random effect (model error) $\sim$ *i.i.d.* $N(0,\sigma^2)$ and independent of e_i
- e_i = survey error for state i $\sim$ independent $N(0,v_i)$ with v_i assumed known.

The regression variables in $\mathbf{x}_i$ include a constant term and, for each state i,

- a "pseudo state poverty rate" computed from tabulations by state of federal tax return information,
- the "nonfiler rate" computed as one minus the ratio of federal tax exemptions for the state divided by the state population estimate,
- the state food stamp participation rate (for age 0-4, 5-17, and 18-64) or supplemental security income participation rate (for age 65+), and
- the previous census estimated poverty ratio (for the age group) or the residuals from regressing the previous census estimate on the other elements of $\mathbf{x}_i$ for the census income year (1999 for Census 2000).

[2] The results of Xie, Raghunathan, and Lepkowski show a quite concentrated posterior distribution for the degrees of freedom in their model. Given our own experiments on this subject, however, the level of precision in estimation of this parameter implied by their results strikes us as unreasonable in the context of their model and application.

The model specifics have changed some over time. For specific details on the models for any given year, see the SAIPE web site at www.census.gov/hhes/www/saipe/.

In production the SAIPE state models are given a Bayesian treatment as discussed in Bell (1999). Flat priors are assumed for $\boldsymbol{\beta}$ and σ^2, the marginal posterior density of σ^2 is then computed analytically, and one-dimensional numerical integration is performed with respect to this posterior to obtain posterior means and variances of the true state poverty ratios, Y_i. The posterior means are provided as point estimates, and corresponding 90 percent prediction intervals (computed assuming normality) are provided to indicate statistical uncertainty about the Y_i.

2.2 Implications of outliers in the Fay-Herriot model

Notice that the theoretical regression residuals from the model (1) are:

$$y_i - \mathbf{x}_i'\boldsymbol{\beta} = u_i + e_i \,. \tag{2}$$

We see that outliers in the regression residuals could arise from large values of either $|u_i|$ or $|e_i|$. The implications of these two possibilities are quite different, however.

- If u_j is an outlier for some state j, this says that the regression model ($Y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$) is no good for state j. If we knew this to be true then, for state j, we might have to fall back on the direct estimate, y_j. Datta and Lahiri (1995) showed that this result is actually achieved mathematically under their model as $|u_j| \to \infty$. Notice, however, that this is not likely to be a good solution to the problem. We are using a small area model precisely because the direct survey estimates, y_i, are generally unreliable due to being based on small sample sizes. Unless state j happens to have a reasonably large sample size, resulting in a statistically reliable direct estimate, then under this scenario neither the direct estimate nor the regression fit can be considered reliable for state j.

- On the other hand, if e_j is an outlier for state j, this says that state j's direct estimate, y_j, is no good. If we knew this to be true we could fall back on the regression prediction, $\mathbf{x}_j'\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ perhaps estimated by fitting the model excluding the data for state j. Unless we had some reason to suspect that the regression model might not apply for state j either, then this would be a reasonable thing to do.

A significant problem is that, in practice, it may often be difficult to distinguish between these two scenarios since (2) shows that the data used in the modeling cannot readily disentangle outlier effects due to u_i versus those due to e_i. Sometimes outside information may help, as in the case of the example presented in Section 3.

So far the literature – e.g., Datta and Lahiri (1995); Xie, Raghunathan, and Lepkowski (2005); and a recent paper of Ghosh, Maiti, and Roy (2006) – has tended to focus on the first of these two scenarios (u_i an outlier). There are presumably two reasons for this: (*i*) this scenario seems a natural extension of the Gaussian Fay-Herriot model, and (*ii*) a central limit theorem is typically invoked to assert approximate normality of the sampling errors. Given (*ii*) why do we raise the possibility that e_i might be subject to outliers? The answer is "nonsampling error." Notice that e_i (= $y_i - Y_i$), which we've labelled "survey error," contains both sampling and nonsampling error. Typically, nonsampling error is ignored in the modeling. If, however, some form of nonsampling error becomes large for a particular area j, this could distort the direct estimate y_j and make it an outlier.

2.3 Extending the Fay-Herriot model using the *t*-distribution

We now extend the model (1) by assuming a *t*-distribution for either u_i or e_i. Other model assumptions remain the same. There are several ways the *t*-distribution could be implemented. A convenient way for our purposes is to make one of the following two assumptions:

Assumption 1: $u_i \mid \theta_i \,, \sigma^2 \sim$ independent $N(0, \theta_i\sigma^2)$

$1 / \theta_i \sim i.i.d.$ Gamma$(k/2, (k-2)/2)$

which implies that $u_i \mid \sigma^2 \sim i.i.d.\ t_k(0, \sigma^2(k-2)/k)$ (Gelman et al. 1995)

$E(\theta_i) = 1$, $\text{Var}(u_i \mid \sigma^2) = \sigma^2$.

Assumption 2: $e_i \mid \theta_i, v_i \sim$ independent $N(0, \theta_i v_i)$

$1 / \theta_i \sim i.i.d.\ \text{Gamma}(k/2, (k-2)/2)$

which implies that $e_i \mid v_i \sim$ independent $t_k(0, v_i(k-2)/k)$ (Gelman et al. 1995)

$E(\theta_i) = 1$, $\text{Var}(e_i \mid v_i) = v_i$.

The θ_i can be thought of as multiplicative random effects, distributed around 1, that inflate or deflate the variances of the relevant error component for all the states. Thus, for any state j that is potentially an outlier, we would expect θ_j to be much larger than 1. Note, though, that the above two assumptions both maintain the variances of the error terms (unconditional on the θ_i) as $\text{Var}(u_i \mid \sigma^2) = \sigma^2$ and $\text{Var}(e_i \mid v_i) = v_i$.

As mentioned earlier we assume that the degrees of freedom parameter, k, is fixed and known. In the examples of Sections 3 and 4 we will present results for the values $k = 3, 4, 5, 8$, and ∞, the last of these corresponding to results assuming a normal distribution. Restricting the t-distribution to have $k > 2$ guarantees that the variance exists and is finite. The models could be extended to values of $k \leq 2$ by specifying the t-distributions directly, rather than by introducing the θ_i. Not having finite variances would seem to pose a problem under Assumption 2, however, since it then seems unclear how to relate the scale parameters of the t-distributions to the survey "variance estimates" v_i.

We implemented the t-distribution models in WinBUGS 1.4 (Spiegelhalter, et al. 2003) to produce simulations from the joint posterior distribution of $(\boldsymbol{\beta}, \sigma^2, \theta_1, \ldots, \theta_{51})$. Following some experimentation to assess convergence, we decided that for each model we would use single chains of 10,000 simulations following burn-in periods of 500 simulations. For each of the 10,000 simulations we computed $E(Y_i \mid \mathbf{y}, \sigma^2, \theta_1, \ldots, \theta_{51})$ and $\text{Var}(Y_i \mid \mathbf{y}, \sigma^2, \theta_1, \ldots, \theta_{51})$ ($\mathbf{y} = (y_1, \ldots, y_{51})'$) from standard formulas (Bell 1999) that analytically integrate out $\boldsymbol{\beta}$. These formulas apply since, conditional on the θ_i, the new model is essentially similar to model (1). We then appropriately averaged these results over the simulations of $(\sigma^2, \theta_1, \ldots, \theta_{51})$ to calculate the posterior means and variances of the true poverty ratios via

$$\begin{aligned} E(Y_i \mid \mathbf{y}) &= E_{\sigma^2, \theta_1, \ldots, \theta_{51} \mid \mathbf{y}}[E(Y_i \mid \mathbf{y}, \sigma^2, \theta_1, \ldots, \theta_{51})] \\ \text{Var}(Y_i \mid \mathbf{y}) &= E_{\sigma^2, \theta_1, \ldots, \theta_{51} \mid \mathbf{y}}[\text{Var}(Y_i \mid \mathbf{y}, \sigma^2, \theta_1, \ldots, \theta_{51})] + \text{Var}_{\sigma^2, \theta_1, \ldots, \theta_{51} \mid \mathbf{y}}[E(Y_i \mid \mathbf{y}, \sigma^2, \theta_1, \ldots, \theta_{51})] \end{aligned} \quad (3)$$

In the example presented in Section 3 we are interested in how $E(Y_i \mid \mathbf{y})$ and $\text{Var}(Y_i \mid \mathbf{y})$ change when a t-distribution is assumed in place of the normal distribution, how this depends on which component (u_i or e_i) is assumed to follow a t-distribution, and how these results vary as the degrees of freedom parameter, k, varies. Similar considerations motivate the discussion of the bivariate model in Section 4. We have particular interest in the results for states whose direct estimates appear to be potential outliers.

3. Application: Univariate *t*-distribution Models for State 5-17 Poverty Ratios in 1989

To illustrate application of the t-distribution models we examine the case of the 5-17 state poverty ratio model for 1989. Table 1 shows various results for all four age-groups for the state of Connecticut (CT). Notice first the standardized residuals from the four models assuming normality; these are defined as $(y_i - \mathbf{x}_i'\boldsymbol{\beta}) / [\text{Var}(y_i - \mathbf{x}_i'\boldsymbol{\beta})]^{.5}$, with model parameters estimated here by maximum likelihood (ML).[3] The standardized residuals are all negative and, except for age 65+, are rather large in magnitude, suggesting that these estimates for CT might be regarded as

[3] ML is used here only for convenience in getting the standardized residuals. Similar results could be obtained with a little more effort from the Bayesian approach that provides the results for subsequent tables.

outliers. The fact that this occurs for all three of the younger age groups makes this even more suspicious. We can thus ask what is leading to these large standardized residuals, and whether it might be better characterized, in the context of our models, as arising from outliers in the model errors (u_i) or survey errors (e_i) for CT? Since the standardized residual for age 65+ is not large in magnitude, we focus on results for the three younger age groups.

Table 1. 1989 poverty ratio estimates and standardized model residuals for Connecticut

Age	Std. Res.	CPS	$\mathbf{x}_i'\boldsymbol{\beta}$	1990 Census
0-4	−2.5	2.6%	13.2%	11.6%
5-17	−3.5	2.2%	10.5%	9.7%
18-64	−2.9	2.2%	5.2%	5.3%
65+	−.03	7.0%	7.1%	7.2%

The direct CPS estimates,[4] given in the third column of Table 1, are very low for the first three age groups – few other of these single-year direct CPS estimates from 1989 through 2004 even approach these values[5]. The regression fitted values from the models ($\mathbf{x}_i'\boldsymbol{\beta}$ with $\boldsymbol{\beta}$ estimated by ML), which are given in the fourth column, are much higher than the CPS direct estimates, are more consistent with poverty ratio estimates overall, and are more consistent with estimates for CT for other years. (E.g., for 1990 the CPS estimates for CT were 12.1, 7.6, and 5.1 percent, respectively, for the three younger age groups.) Finally, for 1989 we have available poverty ratio estimates from another source: the 1990 census (which collected income reports for 1989 from the long form sample). These estimates, given in the last column, are reasonably consistent with the regression fits, and thus inconsistent with the CPS direct estimates for CT. While there are differences between the CPS and census data collections that lead to some systematic differences between the two, these are not nearly so large as to explain the differences for 1989 between the CPS and census direct estimates.

In summary, these results suggest that, (*i*) for the three younger age groups, CT in 1989 could be regarded as a potential outlier, and (*ii*) this seems due to the very low CPS direct estimates, which points in the direction of an outlier in the survey errors rather than in the model errors. In the rest of this section we examine, for the 1989 5-17 poverty ratios, how the models of Section 2 that involve the *t*-distribution address this potential outlier. We examine results for both the model with *t*-distributed model errors and the model with *t*-distributed survey errors.

Table 2 shows posterior means of the model error variance, σ^2, under the Gaussian model and the various *t*-distribution extensions. For both general models (u_i or e_i following a *t*-distribution) the posterior mean of σ^2 increases, for the most part, as the degrees of freedom, *k*, decreases. This is despite the fact that the *t*-distributions were set up to maintain σ^2 as the model error variance and the v_i as the sampling error variances. This behavior of the posterior means of σ^2 affects the behavior of the posterior variances of the poverty ratios, as will be seen shortly.

Table 2. State 5-17 poverty models for 1989
Effects of assuming a *t*-distribution on the posterior means of σ^2

t-distribution assumed for	degrees of freedom (*k*)				
	3	4	5	8	∞ (normal)
u_i	3.3	2.8	2.6	2.3	2.2
e_i	3.3	2.4	2.1	1.9	2.2

[4] The "direct CPS estimates" presented here are *not* the official Census Bureau CPS state poverty estimates. For modeling purposes SAIPE uses direct CPS estimates based on CPS income data collected for a single year, whereas the official CPS state estimates are averages of three such successive CPS single-year estimates centered on the year of interest.

[5] The next lowest direct CPS estimates for ages 5-17 and 18-64, which occurred in various later years for the state of New Hampshire, were 4.9 and 4.1 percent, respectively. New Hampshire also had the lowest estimate overall for age 0-4, of 1.4 percent, but the next lowest after this was a 3.9 percent in one year for Minnesota. We examined estimates for 1989-2004 to cover the range of years for which, as of this writing, the poverty ratio models have been fitted.

Table 3 shows some results on the posterior means of the variance inflation factors, θ_i. The prior means of the θ_i are all 1, and the table shows that the average (across states) of the posterior means does not differ much from 1 except for $k = 3$. The table also shows that the posterior means of θ_i for CT from the various t-distribution models are quite large, indicating that substantial variance inflation is needed to make the observation for CT consistent with the model. We also see that the posterior means of θ_i for CT grow as k decreases, and are larger for the model where e_i follows a t-distribution. For comparison, we also show posterior means of θ_i for Rhode Island (RI). While RI has the second largest posterior means of θ_i, the values are much smaller than are those for CT. Finally, note that the posterior standard deviations of θ_i for CT (given in parentheses) are quite large in relation to the posterior means, showing that there is a large amount of uncertainty about the value of θ_i for CT.

Table 3. State 5-17 poverty models for 1989
Average and maximum posterior means (and standard deviation for CT) of θ_i

t-distribution assumed for		degrees of freedom (k)			
		3	4	5	8
u_i	average θ_i	.91	.98	.99	.99
	max θ_i (CT)	4.79	2.94	2.18	1.40
		(20.02)	(8.39)	(4.34)	(1.45)
	2nd max θ_i (RI)	1.67	1.54	1.33	1.12
e_i	average θ_i	.90	.96	.98	.99
	max θ_i (CT)	5.65	3.99	3.30	2.23
		(16.47)	(4.91)	(3.43)	(1.62)
	2nd max θ_i (RI)	2.92	2.46	2.10	1.64

Table 4 shows the posterior means and variances of the true 5-17 poverty ratios, Y_i, for CT under the two models with various degrees of freedom for the t-distribution. When the model errors, u_i, follow a t-distribution, we see that the posterior mean, $E(Y_i \mid \mathbf{y})$, for CT decreases as the degrees of freedom decreases. This is heading away from the regression fit (10.5%) and towards the direct CPS estimate (2.2%) for CT, though u_i does not appear to be a large enough outlier to actually push the posterior mean close to the direct estimate (a la Datta and Lahiri's (1995) asymptotic result). In contrast, when the survey errors, e_i, follow a t-distribution, the posterior mean of Y_i for CT is not very different over the finite values of k shown in the table, but at these fairly low values of k the posterior means are higher than for the normal model, more in the neighborhood of the regression fit. Thus, these models are behaving somewhat as anticipated, giving less weight to the direct estimate, though e_i does not appear to be a large enough outlier to outright reject CT's direct estimate. Finally, we notice that under both models the posterior variances, $\mathrm{Var}(Y_i \mid \mathbf{y})$, increase with decreasing degrees of freedom, and are somewhat larger for the model where u_i follows a t-distribution. These results are reminiscent of the results for the posterior means of σ^2 given in Table 2.

Table 4. State 5-17 poverty models for 1989
Posterior means and variances of the true poverty ratios, Y_i, for Connecticut

t-distribution assumed for		degrees of freedom (k)				
		3	4	5	8	∞ (normal)
u_i	$E(Y_i \mid \mathbf{y})$	7.2%	7.7%	7.9%	8.3%	8.6%
	$\mathrm{Var}(Y_i \mid \mathbf{y})$	8.4	7.0	6.1	4.7	3.5
e_i	$E(Y_i \mid \mathbf{y})$	9.7%	10.0%	10.0%	9.8%	8.6%
	$\mathrm{Var}(Y_i \mid \mathbf{y})$	7.5	5.7	4.9	4.2	3.5

Finally, we briefly summarize how the posterior means and variances of the Y_i for the other states are affected by using either of the t-distribution models. The largest changes from the Gaussian results occur for $k = 3$. When $u_i \sim t_3$, the changes in $E(Y_i \mid \mathbf{y})$ for states other than CT are mostly quite small in magnitude (e.g., less than 0.1%), the largest change being a decrease of 0.26% for RI, the state with the second largest posterior mean of θ_i (Table 3). When $e_i \sim t_3$, the changes in $E(Y_i \mid \mathbf{y})$ for states other than CT are larger in magnitude, the largest being a decrease of

1.3% for Mississippi. This is actually larger in magnitude than the increase of 1.1% for CT shown in Table 4, and four other states have changes in $E(Y_i \mid \mathbf{y})$ exceeding 1.0% in magnitude. (For RI, though, the change is a miniscule increase of 0.07%.) In regard to the posterior variances, $\text{Var}(Y_i \mid \mathbf{y})$, for some states they increase while for others they decrease. The changes are mostly small for $u_i \sim t_3$, but are larger for $e_i \sim t_3$. For both models the largest changes in $\text{Var}(Y_i \mid \mathbf{y})$ for a state other than CT occur for RI, with increases of 2.6 when $u_i \sim t_3$ and of 2.7 when $e_i \sim t_3$.

4. Application: Bivariate *t*-distribution Models for State 5-17 Poverty Ratios in 2002

Huang and Bell (2004) investigated use of a bivariate Gaussian model for state poverty ratio estimates from both CPS and the Census Bureau's American Community Survey (ACS). The ACS estimates used were from supplemental surveys conducted as part of ACS testing and development, not from the full production ACS sample (which first produced poverty estimates for 2005). Differences between CPS and ACS data collection and procedures can lead to some differences in the poverty estimates, so the focus in using the bivariate model was to see if borrowing information from the ACS estimates could improve the predictions of Y_{1i}, the true poverty ratios in the CPS equation, beyond the improvements obtained by using information from the covariates $\mathbf{x}_i$.

The bivariate model used was a direct generalization of (1) that can be expressed by (*i*) adding the subscript 1 (denoting CPS) to every quantity in (1) except $\mathbf{x}_i$, and (*ii*) copying the equations from (1) down again and adding the subscript 2 (denoting ACS) to every quantity there except $\mathbf{x}_i$. Thus, the only quantity in common between the two sets of equations is the covariate vector, $\mathbf{x}_i$. The same assumptions (e.g., flat prior distributions for parameters) were made for the CPS and ACS equations, and sampling variances for the direct ACS estimates were treated as known. One additional parameter is required; this can be defined as $\rho = \text{Corr}(u_{1i}, u_{2i})$. (The survey errors were assumed uncorrelated between the two surveys, since their samples are essentially drawn independently.) It is nonzero values of ρ that create potential gains for borrowing information from the ACS estimates. Of course, ρ is an unknown parameter that must be treated in the Bayesian calculations. Huang and Bell (2004) assumed a uniform prior for ρ on [−1,1], and used WinBUGS to generate simulations of the model parameters, which then drove calculations of the posterior means and variances of the true poverty ratios in the CPS equation (Y_{1i}) in the same manner as in (3).

Huang and Bell (2004) obtained bivariate model results for all four age groups for the years 2000 and 2001. We have since extended the analysis to include results for 2002. We found generally small reductions in posterior variances, $\text{Var}(Y_{1i} \mid \mathbf{y}_1, \mathbf{y}_2)$, from use of the bivariate model relative to the univariate model (1). However, we also found a few instances of large posterior variance *increases* from use of the bivariate model. Use of a restricted bivariate model that assumed the regression coefficients, apart from the intercept, to be the same in the CPS and ACS equations yielded more substantial improvements overall, but the occasional large posterior variance increases persisted. The large posterior variance increases corresponded to large standardized regression residuals in the ACS equation, i.e., to possible ACS equation outliers. In Huang and Bell (2006) we tried using a *t*-distribution for the model errors or survey errors in the ACS equation to deal with this situation; we briefly summarize the results here.

We focus here on posterior variances of Y_{1i} for the particular case of the 5-17 poverty ratios[6] in 2002, specifically on results for the state of Alaska (AK). For this particular case Huang and Bell (2004) found use of a bivariate Gaussian model produced a posterior variance of $Y_{1,AK}$ of 1.24, a 52% increase over the posterior variance from the univariate model (1), which was .82. The corresponding ACS estimate for AK had a standardized residual in the model of −3.1, suggesting a possible outlier. Table 5, an extract from a table given in Huang and Bell (2006), reports posterior variances for this case from various bivariate models, both the Gaussian model and models with a *t*-distribution assumed for either the ACS equation model errors (u_{2i}) or survey errors (e_{2i}). We see that when the ACS equation model errors are assumed to follow a *t*-distribution there is little effect on the posterior variances of $Y_{1,AK}$. When the ACS survey errors are assumed to follow a *t*-distribution, however, the posterior variance of $Y_{1,AK}$ is substantially diminished, and by increasing amounts as the degrees of freedom decreases. These effects do not, however, lower the bivariate model posterior variance for AK to the value from the univariate model, which is approximately the result that would be achieved by using the bivariate Gaussian model but rejecting the ACS estimate for AK as an outlier.

[6] In this example the CPS estimated poverty ratios are for related children age 5-17 in families, which differ slightly from the CPS poverty ratios for total children age 5-17, as were used in the example of Section 3.

Table 5. Bivariate (CPS and ACS) state 5-17 poverty ratio models for 2002
Posterior variances of the CPS equation true poverty ratio, $\mathrm{Var}(Y_{1i} \mid \mathbf{y}_1,\mathbf{y}_2)$, for Alaska

t-distribution assumed for		degrees of freedom (k)				
		3	4	5	8	∞ (normal)
u_i	$\mathrm{Var}(Y_{1,AK} \mid \mathbf{y}_1,\mathbf{y}_2)$	1.13	1.17	1.19	1.22	1.24
e_i	$\mathrm{Var}(Y_{1,AK} \mid \mathbf{y}_1,\mathbf{y}_2)$	.99	1.02	1.01	1.04	1.24

4. Conclusions

The examples of Sections 3 and 4 show that assuming a t-distribution with low degrees of freedom for the model error or survey error component in a Fay-Herriot model (1), or in the generalization of (1) to a bivariate model, can diminish the effects of possible outliers. In neither of the examples was a potential outlier large enough so that assumption of a t-distribution came close to outright rejection of the potential outlier, something that should occur asymptotically (e.g., as in Datta and Lahiri (1995)). In the univariate model example assuming the model errors had a t-distribution pushed posterior means towards the direct estimate, while assuming the survey errors had a t-distribution pushed them towards the regression fit. For a given application we need to consider which assumption may be more appropriate. Collateral information may be useful in making this determination, as seen in Section 3.

References

Bell, William R. (1999) "Accounting for Uncertainty About Variances in Small Area Estimation," *Bulletin of the International Statistical Institute*, 52nd Session, Helsinki, 1999.

Datta, Gauri S. and Lahiri, Partha (1995) "Robust Hierarchical Bayes Estimation of Small Area Characteristics in the Presence of Covariates and Outliers," *Journal of Multivariate Analysis*, **54**, 310-328.

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

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. (1995), *Bayesian Data Analysis*, New York: Chapman and Hall.

Ghosh, Malay, Maiti, Tapabrata, and Roy, Ananya (2006), "Influence Functions and Robust Bayes and Empirical Bayes Small Area Estimation," unpublished paper presented at the 2006 meeting of the American Statistical Association, Seattle, Washington.

Huang, Elizabeth T. and Bell, William R. (2004), "An Empirical Study on Using ACS Supplementary Survey Data in SAIPE State Poverty Models," *2004 Proceedings of the American Statistical Association*, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association, 3677-3684.

Huang, Elizabeth T. and Bell, William R. (2006), "Using the t-Distribution in Small Area Estimation: An Application to SAIPE State Poverty Models," *2006 Proceedings of the American Statistical Association*, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association, to appear.

Spiegelhalter, David, Thomas, Andrew, Best, Nicky, and Lunn, Dave (2003), "WinBUGS User Manual, Version 1.4," Cambridge: MRC Biostatistics Unit, available at http://www.mrc-bsu.cam.ac.uk/bugs.

Xie, Dawei, Raghunathan, Trivellore E., and Lepkowski, James M. (2005), "Estimation of Prevalence of Overweight in Small Areas – A Robust Extension of Fay-Herriot Model," *2005 Proceedings of the American Statistical Association*, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association, 3701-3712.