Using the t-distribution in Small Area Estimation: An Application to SAIPE State Poverty Models*

Elizabeth T. Huang and William R. Bell
SRD, Room 3000-4, U.S. Census Bureau, Washington, DC 20233-9100

Keywords: Fay-Herriot model, robust models, Current Population Survey, American Community Survey

1. Introduction

The Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program produces state age-group (0-4, 5-17, 18-64, 65+) poverty ratio estimates from Bayesian treatment (Bell 1999) of a Fay-Herriot model (Fay and Herriot 1979) applied to direct state poverty ratio estimates from the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC, formerly known as the CPS March income supplement). The models borrow information from regression variables related to poverty that are constructed from administrative records data, as well as age group poverty ratio estimates from the previous decennial census. Estimates are identified by the "income year" (IY), which refers to the year for which income is reported in the ASEC. Since 2001, the CPS ASEC sample size has been about 100,000 households. Further information is available on the SAIPE web site at www.census.gov/hhes/www/saipe/index.html. For simplicity, in what follows we shorten references to "CPS ASEC" to just "CPS."

In recent years supplementary surveys for the American Community Survey (ACS) have also provided state poverty estimates. The ACS asks essentially the same questions as previous decennial census long form surveys, and is replacing the long form, but with the data collection spread continuously throughout the decade. The supplementary surveys have had sample sizes of about 800,000 addresses, significantly larger than the CPS. Further information on the ACS may be found at www.census.gov/acs/www/.

The ACS procedures for collecting income data differ from those of the CPS. ACS collects income data continuously with a reference period of the previous 12 months (at the time income is reported), whereas the CPS collects income data in February–April with a reference period of the previous calendar year. Annual ACS state estimates use data collected over a full year, and thus involve income reports that cover different 12 month time frames extending over a period of nearly two years. These data collection differences, and other differences, produce systematic differences between ACS and CPS poverty estimates (Bishaw and Stern 2006). Since the CPS provides the official direct poverty estimates at the national level, to date SAIPE has striven to estimate poverty as defined by the CPS.[1]

In Huang and Bell (2004) we reported results of an empirical study investigating the potential benefits to the SAIPE state poverty models of using data from the ACS supplementary surveys. (These results covered IYs 2000-2001. We have since extended this work to include results for IY 2002, and also revised the calculations. It is these revised and extended results that we report on here. For IY 2000-2001 the reported numbers changed slightly from those given in Huang and Bell (2004), but not enough to change any overall conclusions.) The results assess the potential benefits of using the ACS data by comparing prediction error variances from the current state poverty ratio models with variances from bivariate models that use data from both CPS and ACS for the given IY. The results show little benefit from using ACS data in an unrestricted bivariate model (discussed in Section 2), but more substantial benefits from using a restricted bivariate model that assumes common regression coefficients between the two equations (CPS and ACS). However, the results also show occasional instances of large posterior variance increases from use of either bivariate model, which turn out to correspond to large regression residuals from the ACS equation. If these few data points are rejected as "outliers,"

*Disclaimer: This report is released to inform interested parties of 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.

[1] Full production ACS estimates are being released starting this year, including direct poverty estimates for all states. In future years, ACS will produce direct poverty estimates for all U.S. counties and school districts (based on 3- or 5-year data collections for smaller counties and school districts). In light of this, SAIPE is considering switching to basing its estimation on ACS results in future years.

we could fall back on the (CPS equation) univariate model results for these particular observations to avoid the higher posterior variances of the bivariate model. However, empirically detecting and rejecting outliers, and then reporting lower posterior variances calculated as if these certain data points were known in advance to be outliers, is a statistically unprincipled approach (absent concrete explanations as to why the points in question indeed should be regarded as outliers). As a more principled approach to dealing with this issue, it was suggested to us by Alan Zaslavsky that we try using the t-distribution for the model errors (state random effects) in the ACS equation. Lange, Little, and Taylor (1989) explored use of the t-distribution as a useful general extension of the normal distribution for statistical modeling of data sets with longer-than-normal tails. They applied models with t-distributed errors to a variety of problems.

In this paper, we apply the idea of assuming errors follow a t-distribution to the unrestricted bivariate model developed in Huang and Bell (2004) for the CPS and ACS state poverty ratio estimates. This model has four error components – state random effects (model errors) and sampling errors in each of the two equations – any of which could be assumed to follow a t- rather than a normal distribution. We consider all four possibilities here (individually, i.e., only one of the four errors in any given model is t-distributed), examining both the interpretation of these models and their empirical results, particularly the effects on posterior means and variances of various quantities, including posterior variances of the state poverty ratios in the CPS equation.

Section 2 reviews the bivariate models in Huang and Bell (2004) for the CPS and ACS state poverty ratios, and the empirical results motivating the consideration here of the t-distribution. Section 3 discusses the extension of these models to allow one of the error terms to follow a t-distribution, and offers some comments on the interpretation of the resulting models. Section 4 presents empirical results from applying the models to CPS and ACS state poverty ratio estimates of school-age children for IY 2002. We find that, overall, the models behave somewhat as expected, and the models assuming a t-distribution for one of the error components in the ACS equation do something to address the motivating issue – the occasional large posterior variance increases that arise with the bivariate models.

It should be noted that in using the t-distribution we fix the degrees of freedom parameter at various values. This contrasts with the paper of Lange, Little, and Taylor (1989), who used maximum likelihood estimation of the degrees of freedom in models with multivariate t errors. Here we fix the degrees of freedom because of the difficulties in making inferences about this parameter in models with multiple error components.

2. Bivariate Models for State Poverty Ratios using CPS and ACS Data

The general (unrestricted) bivariate model of Huang and Bell (2004), labeled "Model A," is as follows. We illustrate using the model for the age 5-17 poverty ratios. For any given IY and age group, let Y_{1i} and Y_{2i} be the "true poverty ratios" (number poor / population) for state i that are being estimated by the CPS and ACS, respectively, for $i = 1, \ldots, 51$ (including the 50 states and the District of Columbia). Note that due to data collection and other differences between the CPS and ACS, this model assumes that $Y_{1i} \neq Y_{2i}$, in general. Let y_{1i} and y_{2i} be the direct sample estimated poverty ratios for state i from the CPS and ACS, respectively. Then we have

$$y_{1i} = Y_{1i} + e_{1i} \quad (1)$$

$$y_{2i} = Y_{2i} + e_{2i}, \quad (2)$$

where the survey errors e_{1i} and e_{2i} are assumed to be independently distributed as $N(0, v_{ji})$, $j = 1, 2$. Typically the survey errors are assumed to be sampling errors (ignoring possible nonsampling errors) with the v_{ji} assumed known, though the v_{ji} are actually estimates of the true sampling variances. In the case of CPS, the direct sampling variance estimates are smoothed using a sampling error model (Otto and Bell 1995) to get the v_{1i}. In the case of ACS, we use the direct sampling variance estimates as the v_{2i}. Finally, we assume $\mathrm{Cov}(e_{1i}, e_{2i}) = 0$, because the CPS and ACS use independent samples.

The models for the true poverty ratios are:

$$Y_{1i} = x_i'\beta_1 + u_{1i} \quad (3)$$

$$Y_{2i} = x_i'\beta_2 + u_{2i}, \quad (4)$$

where the β_js are vectors of regression parameters, x_i' is a row vector of regression variables for state i, and $(u_{1i}, u_{2i})'$ are independently and identically bivariately normally distributed with zero means. Note the same vector of regression variables x_i' is used in both the CPS and ACS equations since the two estimates are assumed to refer to the same IY (though only approximately, due to data collection differences). We write

$$\mathrm{Var}(u_{1i}) = s_{11},\ \mathrm{Var}(u_{2i}) = s_{22},\ \mathrm{Corr}(u_{1i}, u_{2i}) = \rho.$$

The regression variables in x_i' for IYs 2000-2002 include a pseudo state poverty rate constructed from Internal Revenue Service (IRS) tax data, a tax nonfiler proportion constructed from IRS data and state population estimates, and the Census 2000 estimated state age 5-17 poverty ratios. For more information see the SAIPE web site mentioned earlier.

Noninformative independent prior distributions for the model parameters are assumed as follows:

$$\begin{aligned}\boldsymbol{\beta} &= (\beta_1', \beta_2')' \sim N(0, cI) \\ s_{11} &\sim U(0, m_1) \text{ and } s_{22} \sim U(0, m_2) \\ \rho &\sim U(-1, 1).\end{aligned}$$

The values of c, m_1, and m_2 were chosen to be sufficiently large so that the priors could effectively be regarded as flat on $(-\infty, +\infty)$ and $(0, +\infty)$ as appropriate. We used $c = 1,000$ for all age groups and chose values for m_1 and m_2 separately for each age group so that the likelihoods (for the univariate models discussed below) were effectively zero beyond m_1 and m_2. (E.g., for age 5-17, $m_1 = m_2 = 20$.)

Notice that if $\rho = 0$, then Model A reduces to univariate models (Model U) that can be fit separately. This univariate model for the CPS data is the current SAIPE production model. The restricted bivariate model mentioned earlier is a special case of Model A where the regression coefficients (except the intercept) are assumed the same in both equations. We shall not consider this restricted bivariate model here, however.

To obtain posterior means and variances of the true state poverty ratios (Y_{1i} and Y_{2i}) from Model A we proceeded as follows. We used Gibbs sampling via WinBUGs (Spiegelhalter, et al. 2003) to simulate 10,500 (first 500 discarded as burn in) sets of model parameters from their joint posterior distribution, $p(\boldsymbol{\beta}, \rho, s_{11}, s_{22}|\mathbf{y})$ where $\mathbf{y} = \{(y_{1i}, y_{2i}),\ i = 1, \ldots, 51\}$ is the observed data. The posterior means and covariance matrices of $Y_i = (Y_{1i}, Y_{2i})'$ were then approximated by averaging results over the simulations of $(\rho, s_{11}, s_{22}|\mathbf{y})$ to approximate the following formulas (5) and (6). (Note that we don't actually use the simulations of $\boldsymbol{\beta}$.) For brevity we let $\hat{Y}_i = E(Y_i|\mathbf{y}, \rho, s_{11}, s_{22})$. Then we have

$$\begin{aligned}E(Y_i|\mathbf{y}) &= E_{\rho, s_{11}, s_{22}|\mathbf{y}}[\hat{Y}_i] &(5)\\ \mathrm{Var}(Y_i|\mathbf{y}) &= E_{\rho, s_{11}, s_{22}|\mathbf{y}}[\mathrm{Var}(Y_i|\mathbf{y}, \rho, s_{11}, s_{22})] \\ &\quad + \mathrm{Var}_{\rho, s_{11}, s_{22}|\mathbf{y}}[\hat{Y}_i] &(6)\end{aligned}$$

In equations (5) and (6) $\hat{Y}_i$ and $\mathrm{Var}(Y_i|\mathbf{y}, \rho, s_{11}, s_{22})$ can be readily calculated from standard formulas that account for the effects of uncertainty about $\boldsymbol{\beta}$. (See, e.g., Bell (1999), for univariate results that generalize in a straightforward way to the bivariate case.) The same approach was used to produce posterior means and variances under the univariate model (Model U) by holding $\rho = 0$ fixed through the whole process, and just using the simulations of (s_{11}, s_{22}).

As noted in the Introduction, in Huang and Bell (2004) we compared posterior variances for the Y_{1i} from the bivariate and univariate models over IYs 2000-2001 and the four age groups. (We also did the same comparisons for the Y_{2i}, but that is not of interest here. Both sets of results were later extended, as noted earlier, to include IY 2002.) We examined the percent differences between these variances, which can be expressed as

$$100 \times \frac{\mathrm{Var}(Y_{1i}|\mathbf{y}, \text{ Model A}) - \mathrm{Var}(Y_{1i}|\mathbf{y}_1, \text{ Model U})}{\mathrm{Var}(Y_{1i}|\mathbf{y}_1, \text{ Model U})} \qquad (7)$$

where $\mathbf{y}_1 = (y_{1,1}, \ldots, y_{1,51})'$. We found that the average percent difference was only a few percent, and though it was in favor of the bivariate model, this was not enough to represent a worthwhile improvement. Larger percent differences were obtained with the resricted bivariate model (the model that assumed $\beta_1 = \beta_2$ apart from the intercepts). However, both bivariate models led to occasional large values for (7). Table 1 shows the instances of posterior variance increases for the unrestricted bivariate model A that were 25 percent or greater. Of particular concern is the 52 percent increase for age 5-17 for Alaska (AK) in IY 2002. (There were no increases of over 25 percent for the 65+ age group.)

Table 1. Large ($\geq$ 25 percent) posterior variance increases from bivariate Model A compared to the univariate model for the CPS equation

Age	IY 2000	IY 2001	IY 2002
0-4	HI(29%) WA(30%)	AK(32%)	
5-17			AK(52%)
18-64		OR(32%) AK(40%) KS(40%) NC(41%)	

The reason for these occasional posterior variance increases lies in the second term in (6), $\mathrm{Var}_{\rho, s_{11}, s_{22}|\mathbf{y}}[\hat{Y}_i]$. With the generalization of the results in Bell (1999) to the bivariate case we can write $\hat{Y}_{1i}$ as

$$\hat{Y}_{1i} = x_i'\hat{\beta}_1 + h_{11,i} \times r_{1i} + h_{12,i} \times r_{2i}$$

where $r_{1i} = y_{1i} - x_i'\hat{\beta}_1$ and $r_{2i} = y_{2i} - x_i'\hat{\beta}_2$ are the regression residuals in the CPS and ACS equations, $\hat{\beta}_1$ is the generalized least squares estimate of β_1, and $h_{11,i}$ and $h_{12,i}$ are functions of (ρ, s_{11}, s_{22}) as well as v_{1i} and v_{2i}. As a crude approximation, if we ignore the dependence of $x_i'\hat{\beta}_1$, r_{1i}, and r_{2i} on (ρ, s_{11}, s_{22}), then we would have

$$\begin{aligned}\mathrm{Var}_{\rho,s_{11},s_{22}|\mathbf{y}}[\hat{Y}_{1i}] \quad &\approx \quad \mathrm{Var}_{\rho,s_{11},s_{22}|\mathbf{y}}(h_{11,i}) \times r_{1i}^2 \\ &\quad +\mathrm{Var}_{\rho,s_{11},s_{22}|\mathbf{y}}(h_{12,i}) \times r_{2i}^2.\end{aligned}$$

This suggests that $\mathrm{Var}_{\rho,s_{11},s_{22}|\mathbf{y}}[\hat{Y}_{1i}]$ could be large whenever r_{1i} or r_{2i} is large in magnitude. The analogous formula for the univariate model (obtained by fixing $\rho = 0$) would have a term analogous to $\mathrm{Var}_{\rho,s_{11},s_{22}|\mathbf{y}}(h_{11,i}) \times r_{1i}^2$, so a large $|r_{1i}|$ would boost the posterior variance of both the univariate and bivariate models. But the formula for the univariate model would not include a term like $\mathrm{Var}_{\rho,s_{11},s_{22}|\mathbf{y}}(h_{12,i}) \times r_{2i}^2$, and indeed the value of r_{2i} would not matter to the CPS equation results in the univariate model. So we conjecture that large values of $|r_{2i}|$ could lead to large posterior variance increases for the bivariate relative to the univariate model.

Empirical results bear out the last conclusion. To illustrate, Figure 1 plots the percent differences in posterior variances from (7) against corresponding values of r_{2i} (actually, against $r_{2i}/\sqrt{\mathrm{Var}(r_{2i})}$, which is a rescaling that facilitates interpretation) for age 5-17 poverty ratios in IY 2002. The point to the extreme left is the 52 percent posterior variance difference for Alaska, which corresponds to a standardized residual in the ACS equation of about -3.1. The point to the extreme right is the 18 percent posterior variance increase for North Carolina, which corresponds to a standardized residual in the ACS equation of about 2.1. We can also discern evidence of a quadratic shape in the remaining points, with the minimum occurring around an ACS equation standardized residual of zero. Similar plots for other age groups and/or other IYs also show quadratic shapes with the minima near zero and with the variance differences increasing as one moves to the right and left of zero.

As noted in the Introduction, a large value of r_{2i} could be taken as an indication that the corresponding ACS estimate, y_{2i}, is an "outlier." If y_{2i} is then "rejected" (dropped from our observations), the prediction of Y_{1i} would fall back on the univariate model, which would likely have a lower posterior variance. But absent an explanation of why y_{2i} should be rejected as an outlier, this is a statistically unprincipled approach that would misstate statistical uncertainty. In the following section we examine whether using a t-distribution for error terms in the bivariate model can accomplish the same thing without the somewhat arbitrary nature (and the all-or-nothing aspect) of outlier detection and rejection.

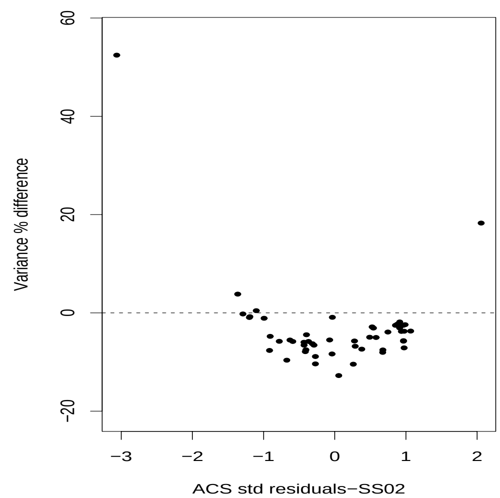

Figure 1: State posterior variance percent differences (for CPS equation poverty ratios for age 5-17, IY 2002) of the unconstrained bivariate model versus the univariate model plotted against corresponding ACS equation standardized residuals

3. Bivariate Models Using the t-distribution

Combining equation (1) with equation (3), and equation (2) with equation (4), we can write the bivariate model of Section 2 as

$$y_{1i} \;=\; Y_{1i} + e_{1i} = (x_i'\beta_1 + u_{1i}) + e_{1i} \qquad (8)$$
$$y_{2i} \;=\; Y_{2i} + e_{2i} = (x_i'\beta_2 + u_{2i}) + e_{2i}. \qquad (9)$$

In this section we examine what happens when one of the error terms above is assumed to follow a t-distribution. This can be assumed for u_{1i}, e_{1i}, u_{2i}, or e_{2i}, leading to four different possible models, all of which we will examine. For simplicity we assume only one of the error terms in a given model is t-distributed; the other errors remain normally distributed with the same assumptions as before.

We will assume that the degrees of freedom, ν, of the t-distribution is known, though we will examine results for different values of ν ($\nu = 3, 4, 5, 8, \infty,$

where $\nu = \infty$ puts us back at the normal distribution). Estimating, or making inferences about, unknown degrees of freedom can be very difficult in models such as ours that involve multiple error components. From the Bayesian perspective, given a mildly informative prior about ν the posterior for ν may not be much more informative than the prior. (Note that some form of informative prior for ν would be needed to deal with the fact that t-distributions for "large" values of ν, e.g., $\nu = 1,000$, $10,000$, $100,000$, and so on, are all approximately normal). More worrisome, experience we've had with similar models suggests that, when ν is unknown, MCMC (Markov Chain Monte Carlo) simulations (e.g., from WinBUGS) may have difficulty converging.

To specify a t-distribution for u_{1i} it is convenient for our purposes to do so in the following way[2]:

$$u_{1i}|s_{11},\theta_i \sim N(0, s_{11}\theta_i) \quad (10)$$

$$1/\theta_i \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu-2}{2}). \quad (11)$$

where these distributions are independent over i. As before, we assume a flat prior on s_{11} ($s_{11} \sim U(0, m_1)$). With the specification in (10) and (11), the distribution of u_{1i} conditional on just s_{11} is $t_\nu(0, \frac{\nu-2}{\nu}s_{11})$, a t-distribution with mean zero, scale $\frac{\nu-2}{\nu}s_{11}$, and ν degrees of freedom (Gelman et al. 2000). The factor $\frac{\nu-2}{\nu}$ in the scale arises from using $\frac{\nu-2}{2}$ in the Gamma distribution in (11) instead of $\frac{\nu}{2}$, and we do this because the variance of the scaled t-distribution is $\frac{\nu}{\nu-2}$ times the scale, so that from (10) and (11) $\text{Var}(u_{1i}|s_{11}) = (\frac{\nu}{\nu-2})(\frac{\nu-2}{\nu}s_{11}) = s_{11}$. So with this specification s_{11} remains the variance of u_{1i} for any value of ν, facilitating comparisons across models. Note that (11) requires that $\nu > 2$, ruling out the t-distributions with 1 (Cauchy) and 2 degrees of freedom, for which the variance does not exist.

When ν is small the distribution of u_{1i} will have a long tail, allowing for possible outliers. Since $E(\theta_i) = 1$ (which leads to the result for $\text{Var}(u_{1i}|s_{11})$ just noted), the θ_i can be regarded as distributed around 1, and so possible outliers can be thought of as corresponding to large values of θ_i. Thus, posterior means of θ_i from the MCMC simulations can be examined to look for large values as evidence of outliers in u_{1i}.

[2] The specification in equation (11) is equivalent to $\theta_i \sim (\nu-2)/\chi^2_\nu$. Specifying models this way in version 1.4 of WinBUGS was found, however, to lead to incorrect results apparently due to some program bug. More plausible results were obtained using the inverse Gamma specification in (11).

To instead specify a t-distribution for u_{2i}, we obviously set $u_{2i}|\theta_i \sim N(0, s_{22}\theta_i)$ and make the same assumption about θ_i as in (11). To specify a t-distribution for e_{1i} we assume $e_{1i}|\theta_i \sim N(0, v_{1i}\theta_i)$, where we still take the survey error variances v_{1i} as known. Similarly, to specify a t-distribution for e_{2i} we assume $e_{2i}|\theta_i \sim N(0, v_{2i}\theta_i)$, where we still take the survey error variances v_{2i} as known. Note that for the latter two cases the (unconditional) survey error variances remain fixed at $\text{Var}(e_{1i}) = v_{1i}$ and $\text{Var}(e_{2i}) = v_{2i}$, respectively.

The motivation for assuming one of the survey error components follows a t-distribution with low degrees of freedom, allowing for possible outliers in the survey errors, may seem unclear given that conventional survey estimation typically bases inference on some form of central limit theorem that would suggest the sampling error should be approximately normally distributed. This, however, ignores nonsampling error, and if nonsampling errors are significant and differential across states this could potentially lead to outliers in e_{1i} or e_{2i}. (Note that the models for Y_{1i} and Y_{2i} take no explicit account of nonsampling error.) Thus, for the model (8), if y_{1i} contains a large nonsampling error for a specific state i, corresponding to a large e_{1i}, this could produce a large regression residual, $y_{1i} - x_i'\beta_1 = u_{1i} + e_{1i}$. Similarly for the model (9). This is not to say that any large regression residuals should simply be assumed to arise from large nonsampling errors in y_{1i} or y_{2i}. But it is one possible explanation.

Another model assumption worth considering involves the correlation (or covariance) between the model errors u_{1i} and u_{2i}. Section 2 assumed that $\text{Corr}(u_{1i}, u_{2i}) = \rho$ had a uniform prior on $(-1, 1)$. We maintain this assumption here, conditional on θ_i, for the cases where either u_{1i} follows (10)-(11), or where we make the analogous assumptions for u_{2i}. In these two cases we then have that

$$\text{Cov}(u_{1i}, u_{2i}|\theta_i) = \rho\sqrt{s_{11}s_{22}\theta_i}.$$

It can then be shown that, under (11), the unconditional covariance of u_{1i} and u_{2i} is $\sqrt{(\nu-2)/2}\{\Gamma[(\nu-1)/2]/\Gamma(\nu/2)\}\rho\sqrt{s_{11}s_{22}}$. The conditional covariance depends on θ_i and the unconditional covariance depends on ν, as does the unconditional correlation. An alternative would be to specify that $\text{Cov}(u_{1i}, u_{2i}|\theta_i) = s_{12} \equiv \rho\sqrt{s_{11}s_{22}}$ so that $\text{Corr}(u_{1i}, u_{2i}|\theta_i) = \rho/\sqrt{\theta_i}$. The latter approach could lead to complications with MCMC simulations, though, as small values of θ_i could easily produce values of the conditional correlation exceeding 1 in magnitude.

Allowing any of the error terms to follow a t-distribution as discussed above, and with a low value of ν, will allow for possible outliers in that error component. Whether an outlier in, say, u_{1i}, produces a detectable outlier in the corresponding observation y_{1i} will depend on the magnitude of u_{1i} relative to e_{1i}. Similar comments apply to using a t-distribution to account for possible outliers in e_{1i}, u_{2i}, or e_{2i}. But whether or not these component outliers can be detected in the observations, we can consider the implications of component outliers in general terms:

- An outlier in u_{1i} signifies that $x_i'\beta_1$ is not a very good regression predictor of Y_{1i}. This presents a difficult situation because the objective is indeed to predict Y_{1i}, and since the motivation for using the small area model is generally high levels of sampling error in the direct estimates, falling back more heavily on the direct estimate y_{1i} for the prediction may not be a very attractive option.

- An outlier in e_{1i} does not affect Y_{1i}. If it makes y_{1i} appear to be an outlier, we can rely more heavily on the regression predictor $x_i'\beta_1$, discounting the direct estimate y_{1i}. This is a reasonable thing to do.

- An outlier in u_{2i} signifies that $x_i'\beta_2$ is not a very good regression predictor of Y_{2i}. If we assume that an outlier in Y_{2i} does not imply a corresponding outlier in Y_{1i} (and, since our models allow only one error component to have a t-distribution, they do make this assumption), then y_{2i} may not provide very useful information for predicting Y_{1i}, and could be downweighted in the prediction. On the other hand, we may ask if $x_i'\beta_2$ is not a good predictor of Y_{2i}, should we still assume (as does our model) that $x_i'\beta_1$ is a reasonable predictor of Y_{1i}?

- An outlier in e_{2i} distorts y_{2i} and should make it less useful in predicting Y_{1i}, so that discounting y_{2i} in forming the prediction is appropriate.

4. Empirical Results

We now examine empirical results to see how using a t-distribution for one of the error components in (8) or (9) affects inferences about various quantities, focusing (for brevity) on posterior means and variances or standard deviations. This includes examining the posterior variances of Y_{1i} (for a few states) under these models to see whether the models address the motivating issue of dealing with the occasional large posterior variance increases with the bivariate model.

We first examine how assumption of a t-distribution affects posterior means and standard deviations of the model error variances, s_{11} and s_{22}. These are given in Table 2 at the end of the paper. We see that when a t-distribution is assumed for either u_{1i} or e_{1i}, that the posterior mean of s_{11} increases with decreasing degrees of freedom ν. Also, the posterior distribution of s_{11} becomes more spread out (the standard deviation of s_{11} increases) with decreasing ν. On the other hand, assuming a t-distribution for either u_{1i} or e_{1i} has very little effect on the posterior mean or standard deviation of s_{22} (results not shown). Analogous results hold when assuming a t-distribution for either u_{2i} or e_{2i}: this results in the posterior mean and standard deviation of s_{22} increasing with decreasing ν, but has very little effect on the posterior mean or standard deviation of s_{11} (latter results not shown).

Table 3 (at end of paper) shows results on the posterior means of the θ_is – their average across states and the one or two maximum values of their posterior means (and the corresponding posterior standard deviations) – when a t-distribution is assumed for one of the error components.[3] We see that the average of the posterior means of the θ_is always decreases with decreasing ν. For the case of assuming a t-distribution for u_{1i} or u_{2i} this may help explain the reason for the increase in the posterior means of s_{11} and s_{22} with decreasing ν, since the product of θ_i and either s_{11} or s_{22} becomes, conditional on θ_i, the state specific model error variance. So in these cases if the average value of the θ_is decreases, then the posterior mean of s_{11} or s_{22} would have to increase to maintain the same average level of model error variance. This reasoning does not seem to apply, however, to the cases of assuming a t-distribution for e_{1i} or e_{2i}.

Turning now to the maximum values of the posterior means of the θ_is, Table 3 shows that these are for Maine (ME) and Arkansas (AR) in the CPS equation, and for Alaska (AK) in the ACS equation. (In the ACS equation only the values for Alaska are shown because they are much larger than those for any other state.) A large posterior mean for a θ_i shows that the data suggest the associated variance for the given state needs to be much larger than in the Gaussian model to account for the observed value of y_{1i} or y_{2i}, so these are the states that might be flagged as "outliers." Notice also, though, that

[3] Note that the θ_is in Table 3 always refer to the component for which the t-distribution is assumed.

the corresponding standard deviations are large, reflecting considerable uncertainty about how large the θ_i value actually is, and so there is considerable uncertainty about how likely it is that y_{1i} or y_{2i} is an outlier. The posterior means of the max θ_i increase with decreasing ν, simply reflecting the fact that the longer is the tail of the t-distribution of the given error component the further out in the tail will be the maximum value of θ_i. Finally, notice that the posterior means of θ_i for Alaska in the ACS equation are considerably larger than those for Maine or Arkansas in the CPS equation, so y_{2i} for Alaska is more of an outlier than y_{1i} for Maine or Arkansas.

Finally, we return to the issue that motivated this investigation, the effect of apparent outliers on posterior variances of the true poverty rates (in the CPS equation), and how these are affected by assuming a t-distribution. Table 4 (at end of paper) shows posterior variances of Y_{1i} for Maine and Arkansas when t-distributions are assumed for either u_{1i} or e_{1i}. We see that this leads to larger posterior variances for Y_{1i} for Maine and Arkansas, the two states closest to being outliers in the CPS equation. This is somewhat consistent with what was conjectured earlier – a long-tailed distribution for u_{1i} or e_{1i} leads to y_{1i} or the regression fit, $x_i'\hat{\beta}_1$, providing less information about Y_{1i}. However, for Maine assuming a t-distribution for e_{1i} has the greater effect, while for Arkansas assuming a t-distribution for u_{1i} has the greater effect. The reason for this difference is unclear.

Table 4 also shows how posterior variances of Y_{1i} (again, CPS equation) for Alaska are affected by assuming a t-distribution for either u_{2i} or e_{2i}. In this case smaller values of ν lead to lower values of the posterior variance of Y_{1i}, and assuming a t-distribution for e_{2i} has the greater effect. These effects are in the desired direction of reducing the posterior variance of Y_{1i} for Alaska by discounting its ACS observation y_{2i}. However, these results do not go as far as rejecting y_{2i} for Alaska as an outlier and then behaving as if this were known in advance to be the case. This (statistically unprincipled) procedure essentially falls back on the CPS equation univariate model results, which yield a posterior variance of Y_{1i} for Alaska of .82. The additional uncertainty about Y_{1i} for Alaska reflected in the higher posterior variances in Table 4 is probably appropriate.

References

Bell, William R (1999), "Accounting for Uncertainty About Variances in Small Area Estimation," *Bulletin of the International Statistical Institute*, 52nd Session, Helsinki, available at www.census.gov/hhes/www/saipe under "Publications."

Bishaw, Alemayehu and Stern, Sharon (2006), "Evaluation of Poverty Estimates: A Comparison of the American Community Survey and the Current Population Survey," Housing and Household Economic Statistics Division, U.S. Census Bureau, available at www.census.gov/hhes/www/poverty/acs_cpspovcompreport.pdf.

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.(2000), *Bayesian Data Analysis*, New York: Chapman and Hall.

Huang, Elizabeth T. and Bell, William R. (2004), "An Empirical Study of 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.

Lange, Kenneth L., Little, Roderick J.A., and Taylor, Jeremy M.G. (1989), " Robust Statistical Modeling Using the t- Distribution", *Journal of the American Statistical Association*, **84**, 881-896.

Otto, Mark C. and Bell, William R. (1995), "Sampling Error Modelling of Poverty and Income Statistics for States," *1995 Proceedings of the American Statistical Association, Section on Government Statistics*, 160-165, Alexandria, VA: American Statistical Association.

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

Table 2. Effects of assuming a t-distribution on posterior means (and standard deviations) of s_{11} and s_{22}

t-distribution assumed for	effect on	degrees of freedom 3	4	5	8	∞ (normal)
u_{1i}	s_{11}	1.31	1.07	.99	.89	.84
		(1.34)	(1.08)	(.94)	(.82)	(.80)
e_{1i}	s_{11}	1.68	1.24	1.08	.94	.84
		(.99)	(.92)	(.88)	(.81)	(.80)
u_{2i}	s_{22}	1.08	.89	.83	.81	.77
		(.69)	(.54)	(.48)	(.46)	(.43)
e_{2i}	s_{22}	1.05	.88	.80	.73	.77
		(.43)	(.43)	(.42)	(.42)	(.43)

Table 3. Average and maximum posterior means (and corresponding standard deviations) of θ_is

t-distribution assumed for		degrees of freedom 3	4	5	8	∞ (normal)
u_{1i}	avg. θ_i	.88	.96	.98	.99	1.00
	max θ_i (ME)	1.71	1.39	1.28	1.10	1.00
		(7.52)	(3.00)	(2.01)	(.81)	—
	θ_i (AR)	1.55	1.25	1.17	1.05	1.00
		(5.51)	(2.86)	(1.87)	(.81)	—
e_{1i}	avg. θ_i	.85	.94	.96	.98	1.00
	θ_i (ME)	2.45	2.18	1.91	1.54	1.00
		(5.44)	(2.72)	(2.10)	(1.06)	—
	max θ_i (AR)	3.52	2.72	2.26	1.71	1.00
		(6.77)	(3.26)	(2.10)	(1.12)	—
u_{2i}	avg. θ_i	.93	.99	.99	.99	1.00
	max θ_i (AK)	7.23	4.69	3.35	1.93	1.00
		(26.27)	(12.47)	(5.19)	(1.70)	—
e_{2i}	avg. θ_i	.85	.96	.98	1.00	1.00
	max θ_i (AK)	5.37	4.17	3.41	2.27	1.00
		(11.28)	(6.77)	(3.91)	(1.75)	—

Table 4. Posterior variances of Y_{1i} (CPS equation) for states with the largest θ_i values assuming a t-distribution for u_{1i}, e_{1i}, u_{2i}, or e_{2i}

t-distribution assumed for	state	degrees of freedom 3	4	5	8	∞ (normal)
u_{1i}	ME	2.04	1.71	1.53	1.27	1.04
	AR	2.69	2.13	1.91	1.60	1.40
e_{1i}	ME	2.32	1.75	1.49	1.26	1.04
	AR	2.59	1.96	1.78	1.58	1.40
u_{2i}	AK	1.13	1.17	1.19	1.22	1.24
e_{2i}	AK	.99	1.02	1.01	1.04	1.24