Some Practical Statistical Problems in Small Area Estimation

William R. Bell

Research and Methodology Directorate

U.S. Census Bureau

JPSM session on Small Area Estimation, March 15, 2024

Disclaimer: Any views expressed here are those of the author and not those of the U.S. Census Bureau.

Connect with us
@ uscensusbureau

Outline

1. Introduction
2. Sampling variances aren't known, they are estimated
3. Random effect variance may be estimated to be zero
4. Dealing with outliers in the Fay-Herriot model
5. Conclusions

United States® Census Bureau

Area level model (Fay & Herriot 1979)

$$y_i = Y_i + e_i$$

$$= (\boldsymbol{x}_i'\beta + u_i) + e_i$$

Model Assumptions:

- $u_i \sim i.i.d.\ N(0, \sigma_u^2)$ (and independent of e_i)
- $e_i \sim ind.\ \ N(0, v_i)$
- **$\boldsymbol{v_i}$ are known**

Note: We really only have estimates $\hat{v}_i$ of v_i.

Best Linear Unbiased Prediction

Assuming σ_u^2 and the v_i are known:

$$\hat{Y}_i = h_i y_i + (1 - h_i)\boldsymbol{x}_i'\hat{\beta}$$

$$\text{var}\left(Y_i - \hat{Y}_i\right) = \sigma_u^2(1 - h_i) + (1 - h_i)^2 \boldsymbol{x}_i' \text{var}(\hat{\beta}) \boldsymbol{x}_i$$

where $h_i = \sigma_u^2/(\sigma_u^2 + v_i)$.

To account for uncertainty about σ_u^2:

- ML, REML – asymptotic results (Prasad and Rao 1990, Datta and Lahiri (2000))
- Jackknife approach (Jiang, Lahiri, and Wan 2002)
- Bayes – posterior variance

Uncertainty about sampling error variances v_i is generally ignored.

Application: Census Bureau's SAIPE (Small Area Income and Poverty Estimates) state poverty rate models for ages 0 – 4, 5 – 17, 18 – 64, and 65+ (developed by Bob Fay)

Direct state estimates y_i from Current Population Survey (CPS), 1989 – 2005

Regression variables in $\boldsymbol{x}_i$ include a constant term and, for each state,

- pseudo state poverty rate from tax return information and also tax "nonfiler" rate
- food stamp (SNAP) participation rate (age 0 – 4, 5 – 17, 18 – 64) or supplemental security income participation rate (age 65+)
- previous census long form estimated poverty rate, or residuals from regressing previous census estimate on other elements of $\boldsymbol{x}_i$ for the census year

Note: Model specifics have changed some over the years.

2. Sampling variances aren't known, they are estimated

What happens when $\hat{v}_i \neq v_i$ in the Fay-Herriot model?

- How much does $E[(Y_i - \hat{Y}_i)^2]$ increase?
- How much do we misstate $E[(Y_i - \hat{Y}_i)^2]$?
- Can we (partly) address these issues by modeling $\hat{v}_i$?

2. Rough calculations of consequences of $\hat{v}_i \neq v_i$

Consider simple case where β and σ_u^2 are known (m very large), but v_i are unknown, estimated by $\hat{v}_i$. Let

$$\begin{aligned}\tilde{Y}_i &= h_i y_i + (1 - h_i)\mathbf{x}_i'\beta \\ \hat{Y}_i &= \hat{h}_i y_i + (1 - \hat{h}_i)\mathbf{x}_i'\beta\end{aligned}$$

where

$$h_i = \frac{\sigma_u^2}{\sigma_u^2 + v_i} = \left(1 + \frac{v_i}{\sigma_u^2}\right)^{-1} \qquad \hat{h}_i = \frac{\sigma_u^2}{\sigma_u^2 + \hat{v}_i} = \left(1 + \frac{\hat{v}_i}{\sigma_u^2}\right)^{-1}.$$

Then the MSE of $\hat{Y}_i$ conditional on $\hat{v}_i$ is

$$E[(Y_i - \hat{Y}_i)^2 | \hat{v}_i] = E[(Y_i - \tilde{Y}_i)^2] + E[(\tilde{Y}_i - \hat{Y}_i)^2 | \hat{v}_i].$$

The MSE of $\tilde{Y}_i$ is $\sigma_u^2(1 - h_i)$. The reported MSE of $\hat{Y}_i$ is $\sigma_u^2(1 - \hat{h}_i)$.

After a little algebra, we have that

$$\text{MSE pct diff} \equiv 100 \times \frac{\text{MSE}(Y_i - \hat{Y}_i) - \text{MSE}(Y_i - \tilde{Y}_i)}{\text{MSE}(Y_i - \tilde{Y}_i)} = 100 \times \frac{(h_i - \hat{h}_i)^2}{h_i(1 - h_i)}.$$

$$\begin{aligned}\text{MSE relbias} &= 100 \times \frac{\text{reported MSE}(Y_i - \hat{Y}_i) - \text{actual MSE}(Y_i - \hat{Y}_i)}{\text{actual MSE}(Y_i - \hat{Y}_i)} \\ &= 100 \times \left\{ \frac{\sigma_u^2(1 - \hat{h}_i)}{\sigma_u^2(1 - h_i) + (h_i - \hat{h}_i)^2(\sigma_u^2 + v_i)} - 1 \right\} \\ &= 100 \times \left\{ \frac{h_i(1 - \hat{h}_i)}{h_i(1 - h_i) + (h_i - \hat{h}_i)^2} - 1 \right\}.\end{aligned}$$

We examine MSE pct diff and MSE relbias for multiplicative errors in $\hat{v}_i$ as an estimate of v_i:

underestimation factors: $\hat{v}_i/v_i = \frac{3}{4}, \frac{1}{2}, \frac{1}{4}$

overestimation factors: $\hat{v}_i/v_i = \frac{4}{3}$, 2, 4.

For each of the above values of $\hat{v}_i/v_i$, plot MSE pct diff and MSE relbias for values of v_i/σ_u^2 from $.02, \ldots, 1, \ldots, 50$ (on log scale).

Fig. 1. Percent difference in MSE and percent bias in reported MSE
from using estimated versus true sampling variance

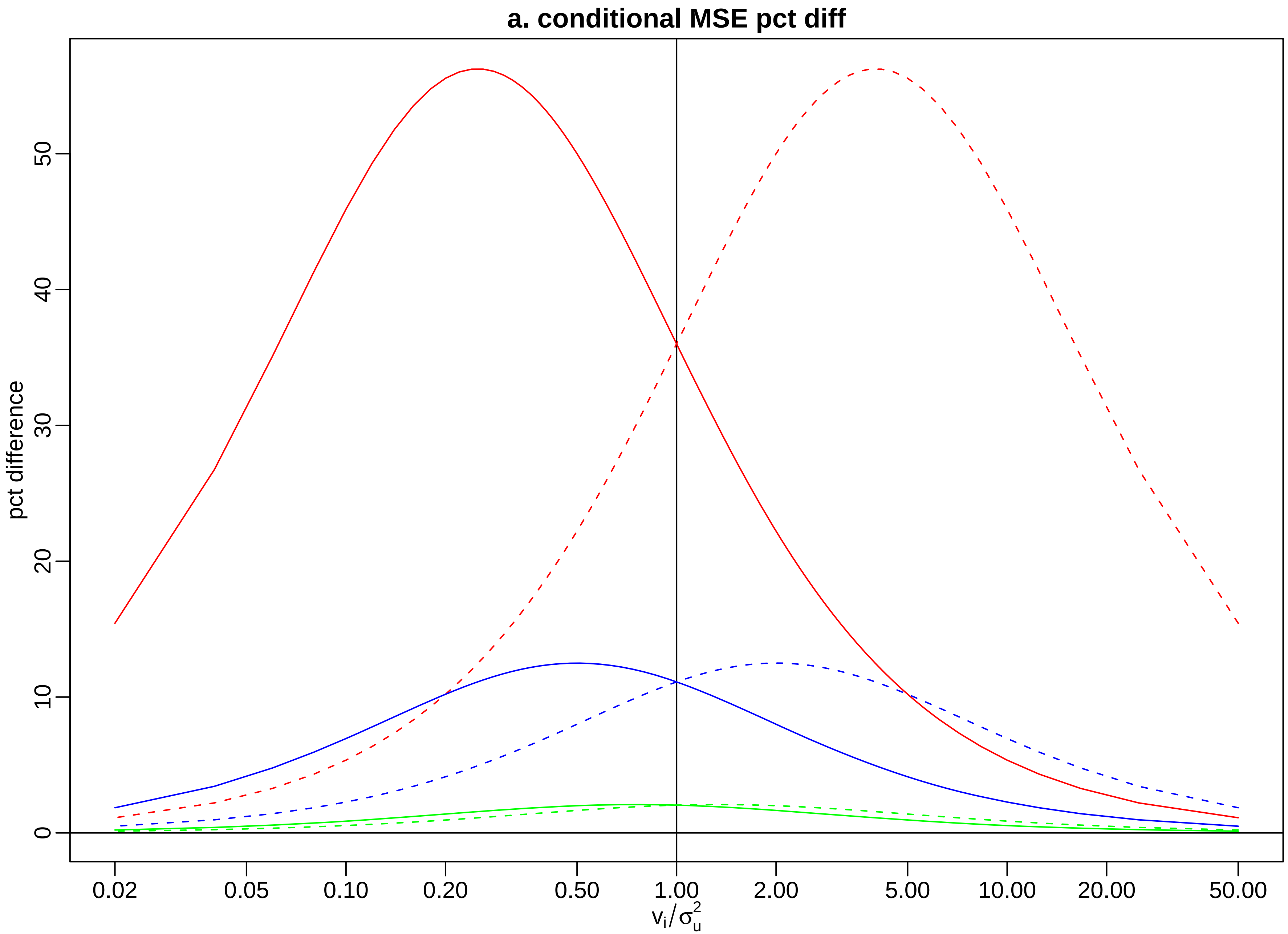
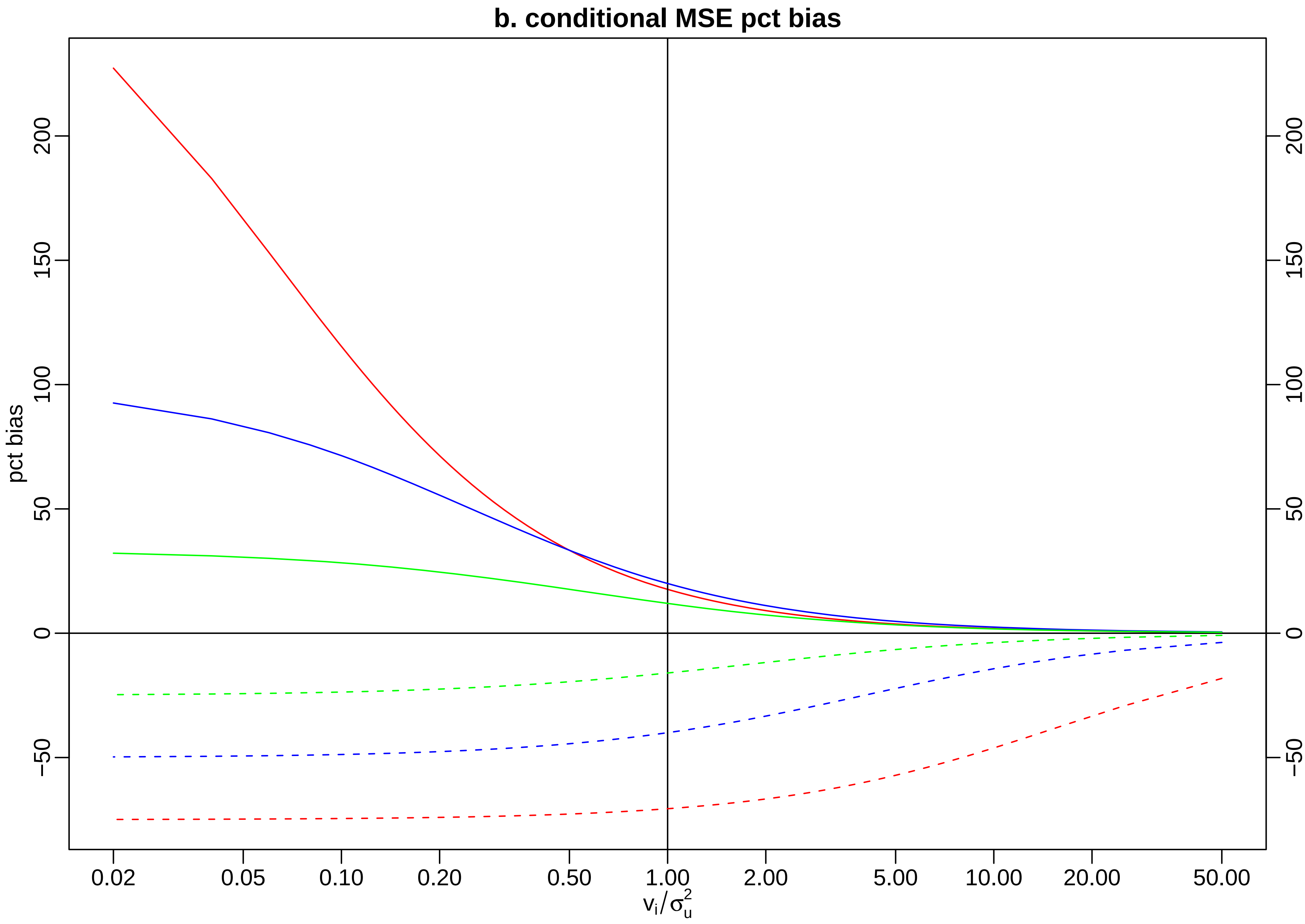


Fig. 1. Percent difference in MSE and percent bias in reported MSE
from using estimated versus true sampling variance



Conclusions for large v_i/σ_u^2:

- Underestimation of v_i is the more severe problem for both MSE pct diff and MSE relbias.

  MSE increase is due to $\hat{h}_i > h_i$, so too much weight given to y_i.

Conclusions for small v_i/σ_u^2:

- Overestimation of v_i is the more severe problem for MSE pct diff.

- MSE relbias is very severe from either severe under- or overestimation of v_i.

Since large errors in $\hat{v}_i$ seem more likely when v_i/σ_u^2 is large, our general conclusion is:

The largest potential problem comes from

severe underestimation of v_i when v_i/σ_u^2 is large.

Given an assumed distribution of $\hat{v}_i$, unconditional versions of MSE pct diff and MSE relbias can be computed as

$$\text{MSE pct diff } = 100 \times \frac{E[(h_i - \hat{h}_i)^2]}{h_i(1-h_i)}.$$

$$\text{MSE relbias } = 100 \times \left\{ \frac{h_i E(1-\hat{h}_i)}{h_i(1-h_i) + E[(h_i - \hat{h}_i)^2]} - 1 \right\}.$$

We do this (by numerical integration) assuming $\hat{v}_i \sim v_i \chi^2_d / d$ for three values of d (6, 16, 80):

Table 1. 5% and 95% points for the χ^2_d/d distribution

d	5% point	95% point
6	.27	2.10
16	.50	1.64
80	.75	1.27

Fig. 1. Percent difference in MSE and percent bias in reported MSE
from using estimated versus true sampling variance

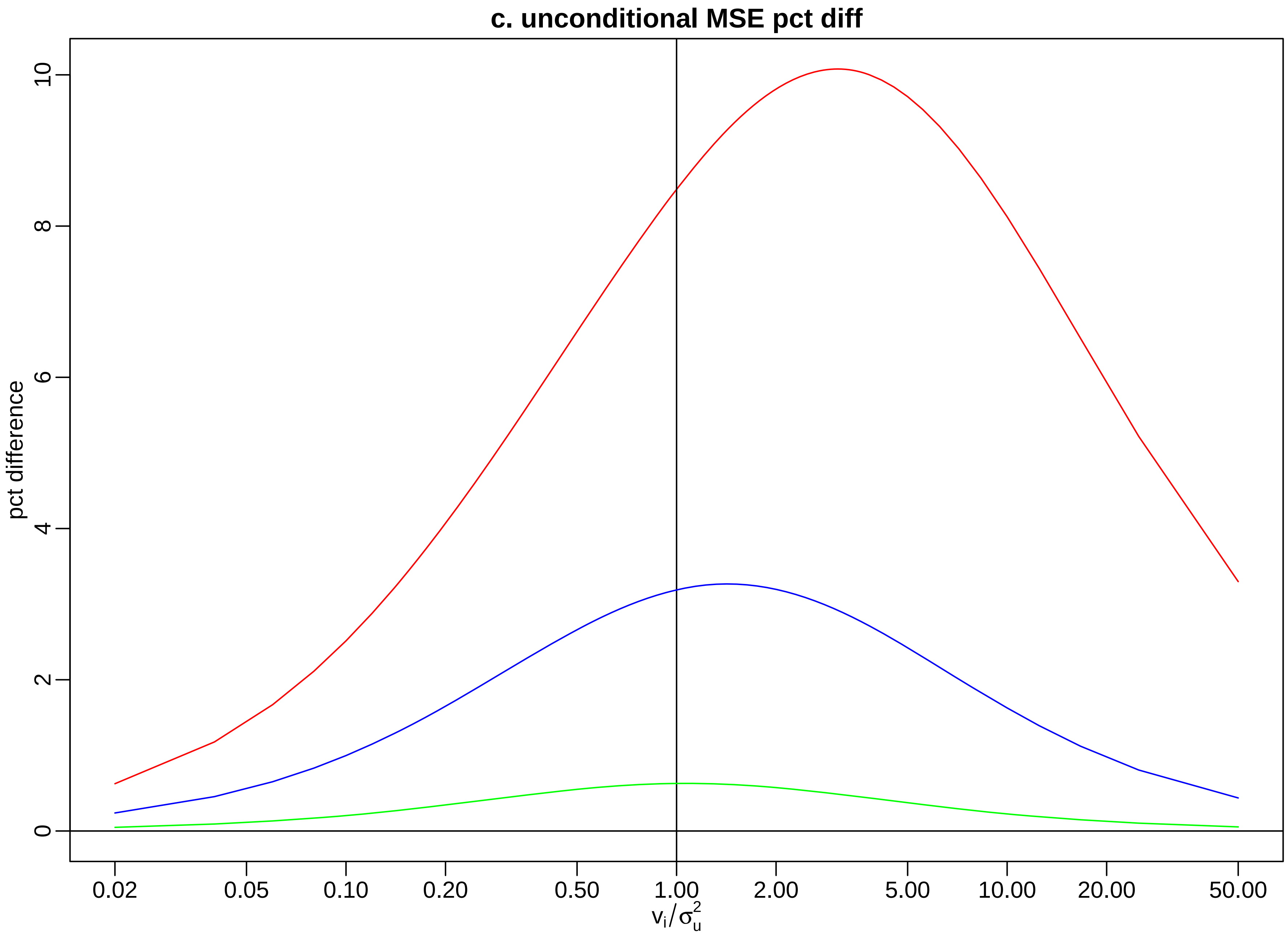
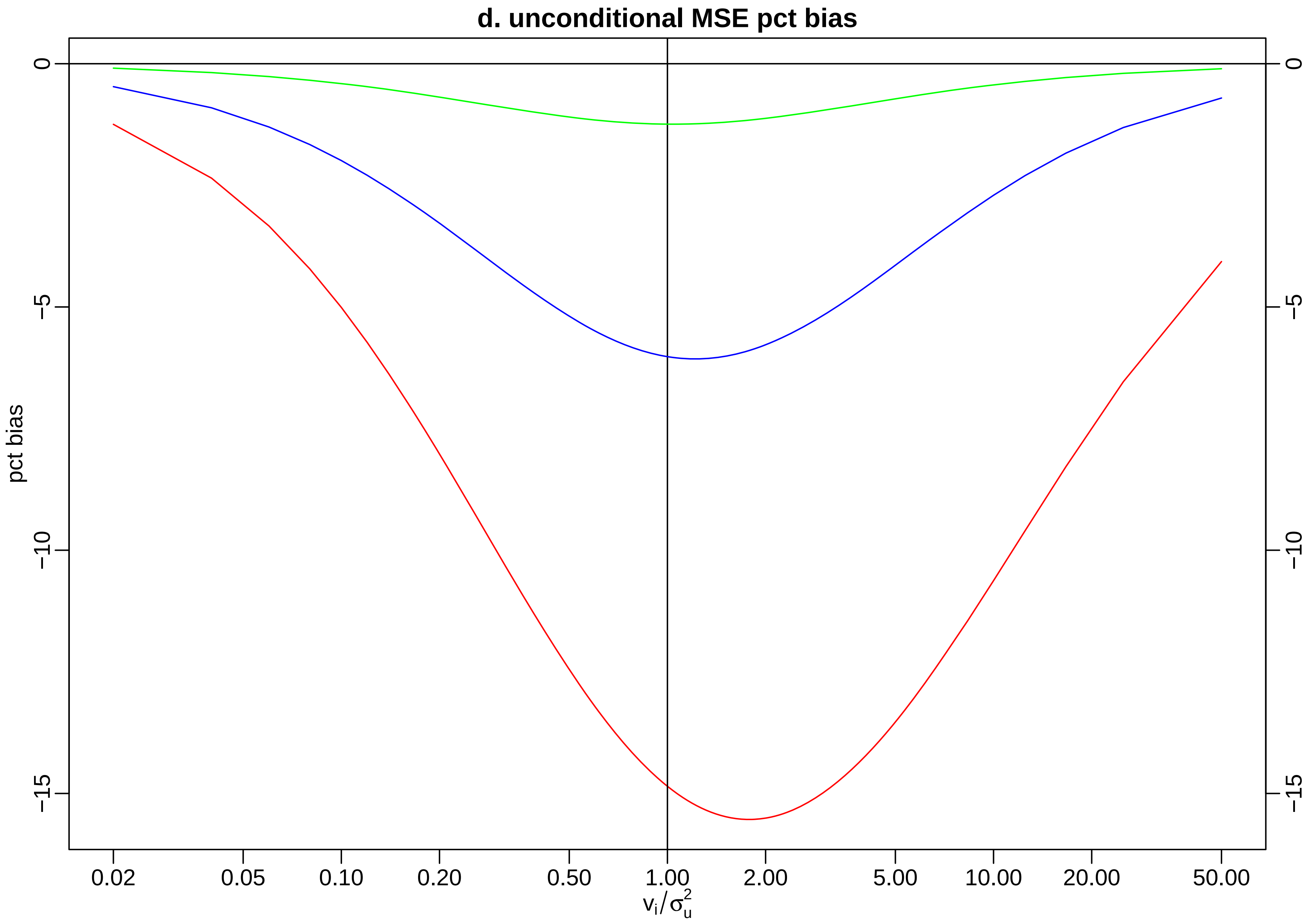


Fig. 1. Percent difference in MSE and percent bias in reported MSE
from using estimated versus true sampling variance



Dealing with uncertainty about v_i

- Frequentist approach – asymptotic MSE results
  - Wang and Fuller (2003); Rivest and Vandal (2003)

- Bayesian approach – model $\hat{v}_i$ as scaled χ_d^2 or Gamma
  - You and Chapman (2006); Sugasawa, Tamae, and Kubokawa (2017)
  - Issues:
    - Degrees of freedom, *d*, may be difficult to determine (Huang and Bell 2010).
    - Assumes independence of y_i and $\hat{v}_i$. This may be OK when estimating means but can fail badly for estimating proportions.

- Try to improve $\hat{v}_i$ by using, e.g., a generalized variance function (GVF)

3. Random effect variance may be estimated to be zero

Comments on alternative estimation approaches

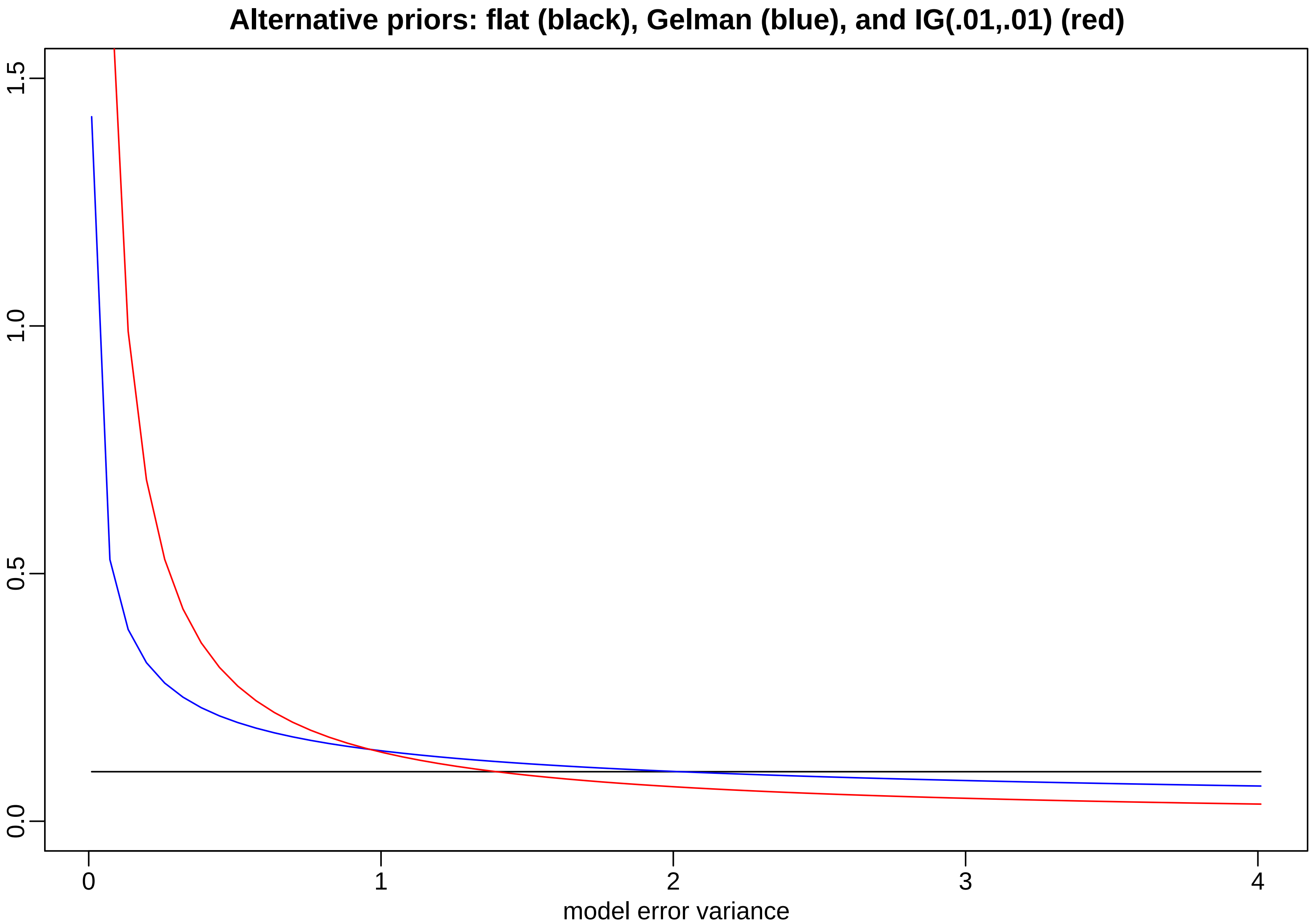
- Method of moments – inefficient, unstable, can produce $\hat{\sigma}^2_{MOM} < 0$.
  - Weighted version from Fay and Herriot (1979) should be better.
- ML, asymptotically efficient, but can produce $\hat{\sigma}^2_{ML} = 0$
- REML – removes some bias in ML, but can produce $\hat{\sigma}^2_{REML} = 0$
- Adjusted maximum likelihood approaches have been proposed to avoid $\hat{\sigma}^2_u = 0$ (e.g., Yoshimori and Lahiri 2012).
- Bayes – avoids problem of $\hat{\sigma}^2_u = 0$ (with appropriate prior on σ^2_u).

Alternative priors for σ_u^2 in the Bayesian approach:

- $p(\sigma_u^2) \propto$ constant on $(0, \infty)$ (SAIPE state model uses this)
- $p(\sigma_u) \propto$ constant on $(0, \infty) \Rightarrow p(\sigma_u^2) \propto 1/\sigma_u$ (Gelman, et al. 2004)
- $\sigma_u^2 \sim IG(\epsilon, \epsilon)$, that is, $1/\sigma_u^2 \sim$ Gamma(ϵ, ϵ), for small $\epsilon > 0$ such as $\epsilon = .01$ or $.001$ (WinBUGS)

CPS equation, 2002, age 5−17 Poverty Rates

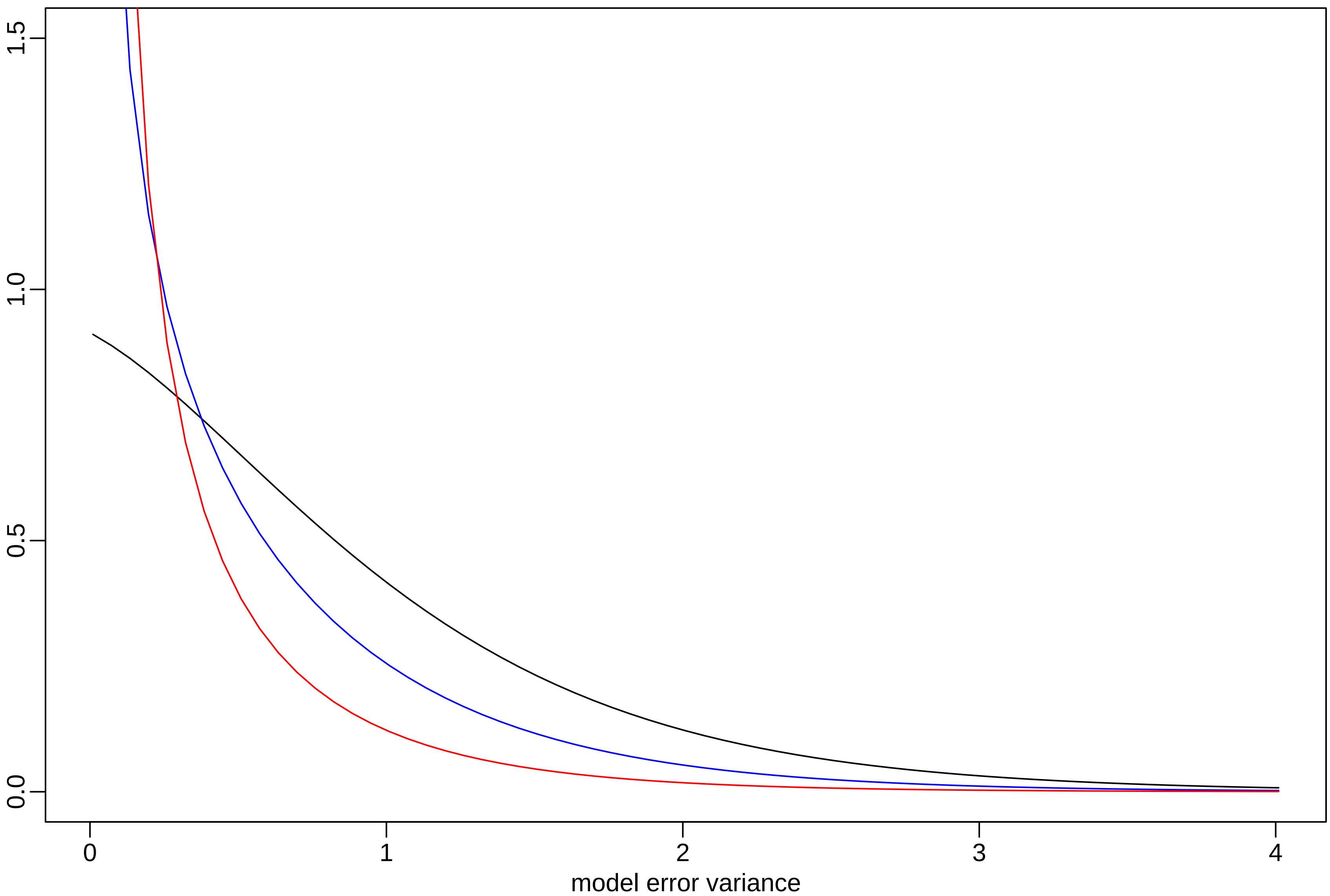
Prior densities for model error variance

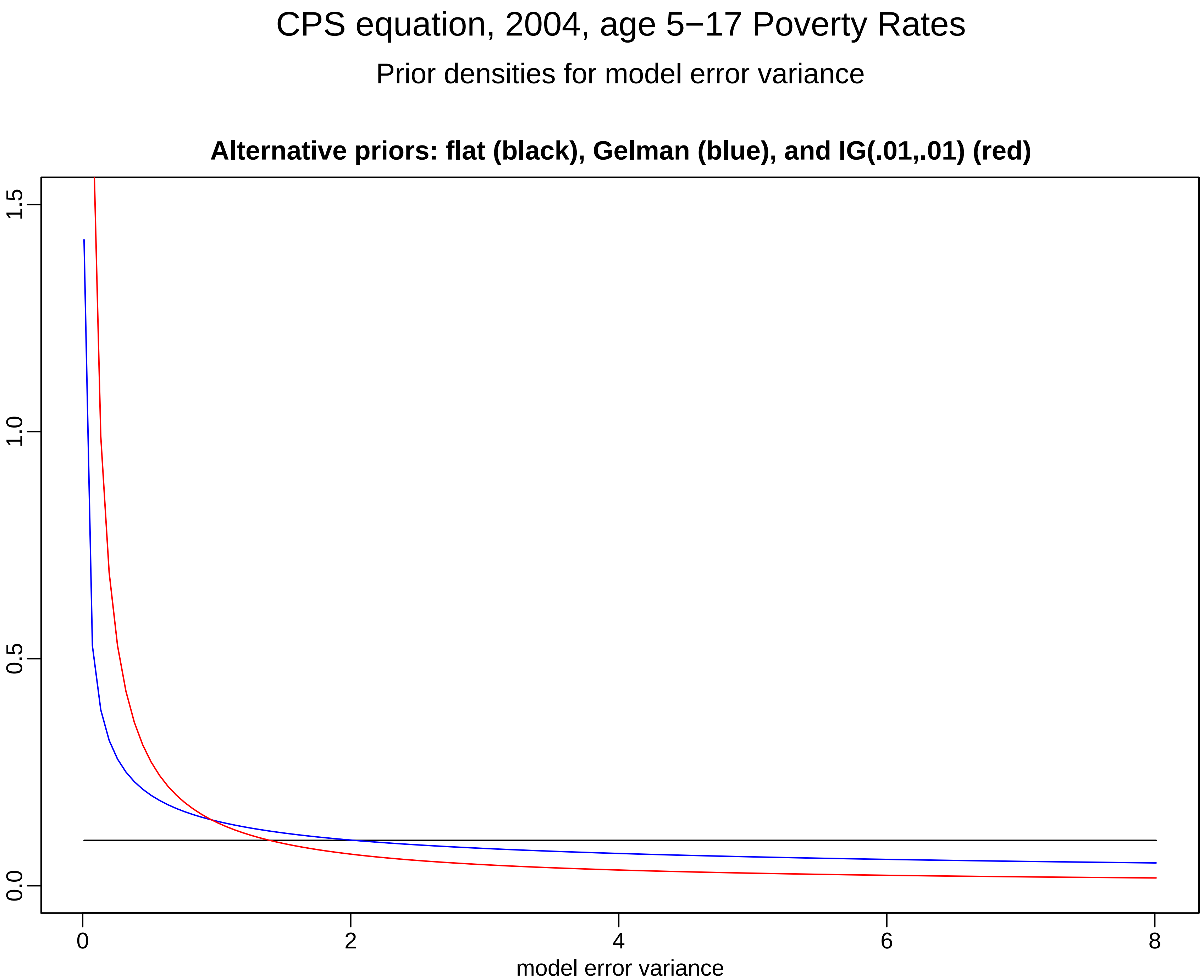


CPS equation, 2002, age 5−17 Poverty Rates

Posterior densities for model error variance from alternative priors

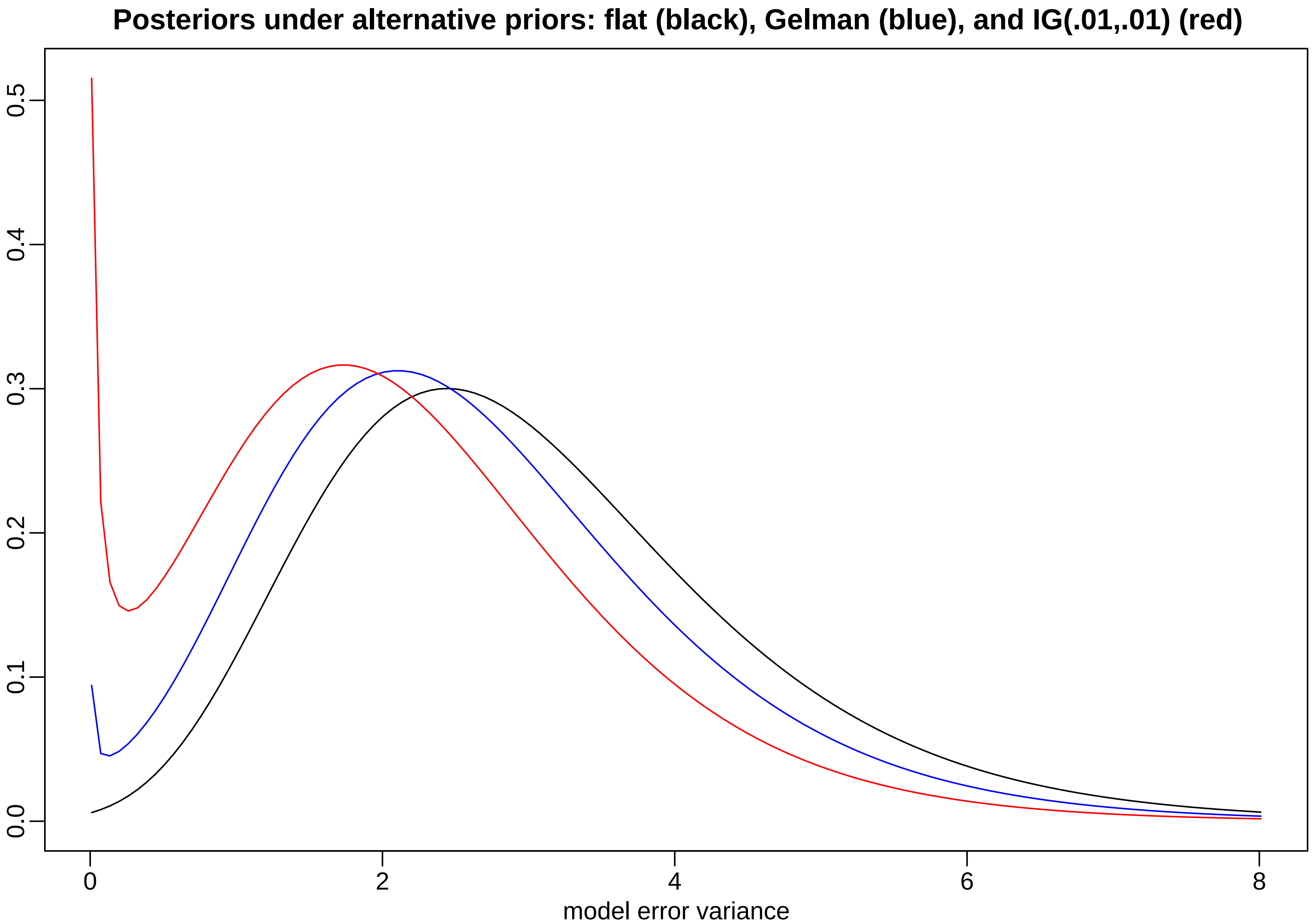
Posteriors under alternative priors: flat (black), Gelman (blue), and IG(.01,.01) (red)




CPS equation, 2004, age 5−17 Poverty Rates
Prior densities for model error variance
Alternative priors: flat (black), Gelman (blue), and IG(.01,.01) (red)
0.0
0.5
1.0
1.5
0
2
4
6
8
model error variance

CPS equation, 2004, age 5−17 Poverty Rates

Posterior densities for model error variance from alternative priors



Implications of $\sigma_u^2 = 0$:

1. If Y_i were observed (CPS a complete census) model would fit perfectly.

2. Weight given to direct estimate is $h_i = 0$ for all i, so $\widehat{Y}_i = \mathbf{x}_i'\widehat{\beta}$.

3. $\text{Var}(Y_i - \widehat{Y}_i) = \mathbf{x}_i'\text{Var}\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{x}_i$ comes only from error in estimating $\boldsymbol{\beta}$, and varies only with $\mathbf{x}_i$.

Illustrating the potential problems in prediction with $\hat{\sigma}_u^2 = 0$

SAIPE state 5-17 poverty rate model

Alternative Estimates of σ_u^2 for Five Years

year	ML	REML	Bayes
1989	0	0	1.7
1990	0	0	2.2
1991	0	0	1.6
1992	0	0	1.6
1993	.4	1.7	3.4

Why do we get $\hat{\sigma}_u^2 = 0$?

Sampling variances (v_i) for IY 1992

state	y_i	n_i	v_i
CA	20.9	4,927	1.9
NC	23.0	2,400	5.5
MS	29.6	796	12.4
DC	31.6	527	33.4

Compare the v_i to

year	1989	1990	1991	1992	1993
$\hat{\sigma}^2_{Bayes}$	1.7	2.2	1.6	1.6	3.4

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

MSE Estimates for 1992

$(\hat{\sigma}^2_{ML} = \hat{\sigma}^2_{REML} = 0,\ \hat{\sigma}^2_{Bayes} = 1.6)$

state	$naive$ ML	$enh.$ ML	$naive$ $REML$	$enh.$ $REML$	$naive$ $Bayes$	$enh.$ $Bayes$
CA	1.3	3.6	1.3	2.8	1.4	1.4
NC	.6	2.0	.6	1.2	1.4	2.0
MS	2.8	3.8	2.8	3.0	3.9	4.0

MSE estimates for 1993

$(\hat{\sigma}^2_{ML} = .4,\ \hat{\sigma}^2_{REML} = 1.7,\ \hat{\sigma}^2_{Bayes} = 3.4)$

state	*naive ML*	*enh. ML*	*naive REML*	*enh. REML*	*naive Bayes*	*enh. Bayes*
CA	1.5	3.2	1.6	2.2	1.7	1.7
NC	1.0	2.4	1.7	2.2	2.0	2.0
MS	3.2	4.3	4.2	4.5	5.0	5.1

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

Conclusions (for estimating prediction MSE):

- Problems with $\hat{\sigma}_u^2 = 0$ for ML; unreasonable implications for prediction MSE

- Problems not solved by REML

- MSE problems not solved by asymptotically allowing for uncertainty about σ_u^2

- Bayesian results look more reasonable

4. Dealing with outliers in the Fay-Herriot model

Regression residuals from the Fay-Herriot model:

$$y_i = (\boldsymbol{x}_i'\beta + u_i) + e_i \;\Rightarrow y_i - \boldsymbol{x}_i'\beta = u_i + e_i$$

We see outliers in $y_i - \boldsymbol{x}_i'\beta$ could arise from large values of either $|u_i|$ or $|e_i|$ (or both).

Implications:

1. u_i an outlier ⇒ regression model is no good for area i

   ⇒ give more weight to direct estimate y_i in the prediction of Y_i

   Unfortunately, not a good solution if sample size for y_i is small.

2. e_i an outlier ⇒ direct estimate y_i is no good for area i

   ⇒ give more weight to regression prediction $\boldsymbol{x}_i'\hat{\beta}$ in the prediction of Y_i

   A reasonable solution in this case.

Problem: It may be difficult to distinguish between these two cases.

Literature (e.g., Datta and Lahiri (1995); Ghosh, Maiti, and Roy (2006)) has tended to focus on Case 1 (u_i an outlier) probably for two reasons:

- Seems a natural extension of the Gaussian model
- Central limit theorem invoked for approximate normality of sampling error.

Why might e_i be subject to outliers?

- Approximate normality is a large sample result that may not hold for direct estimates for small areas.
- Nonsampling error

Example: SAIPE state poverty rate models for 1989

Standardized residuals for Connecticut

Age	std. res.
0-4	-2.5
5-17	-3.5
18-64	-2.9
65+	$-.03$

1989 poverty rate estimates for Connecticut

Age	std. res.	CPS	$\mathbf{x}_i'\hat{\beta}$
0-4	-2.5	2.6%	13.2%
5-17	-3.5	2.2%	10.5%
18-64	-2.9	2.2%	5.2%
65+	$-.03$	7.0%	7.1%

1989 poverty rate estimates for Connecticut

Age	std. res.	CPS	$\mathbf{x}_i'\hat{\beta}$	90 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%

Extending the Fay-Herriot (1979) model using the t-distribution for outliers – Case 1

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

1. Assume, as usual, that $e_i \sim$ ind. $N(0, v_i)$ with the v_i assumed known, but specify a t-distribution for the random effects u_i:

$$u_i | \theta_i, \sigma_u^2 \sim \text{ind. } N(0, \theta_i \sigma_u^2)$$

$$\frac{1}{\theta_i} \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu - 2}{2}\right)$$

Then (Gelman, et al. 2004)

- $u_i | \sigma_u^2 \sim i.i.d.\ t_\nu\left(0, \sigma_u^2\left(\frac{\nu-2}{\nu}\right)\right)$
- $E(\theta_i) = 1,\ var(u_i) = \sigma_u^2$

Extending the Fay-Herriot (1979) model using the t-distribution for outliers – Case 2

$$y_i = Y_i + e_i$$
$$= (\boldsymbol{x}_i'\beta + u_i) + e_i$$

2. Assume, as usual, that $u_i \sim i.i.d.\ N(0, \sigma_u^2)$, but specify a t-distribution for the survey errors e_i:

$$e_i | \theta_i, v_i \sim \text{ind. } N(0, \theta_i v_i) \quad \text{where } v_i \text{ is known}$$

$$\frac{1}{\theta_i} \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu - 2}{2}\right)$$

Then (Gelman, et al. 2004)

- $e_i | v_i \sim\ i.i.d.\ t_\nu\left(0, v_i\left(\frac{\nu-2}{\nu}\right)\right)$
- $E(\theta_i) = 1,\ var(e_i) = v_i$

Notes on the t-distribution models:

- We assume, for simplicity, that ν is fixed and known, and $\nu > 2$. Here we use $\nu = 3, 4, 5, 8, \infty$ (normal).

- Priors for Bayesian inference:

  – $\boldsymbol{\beta} \sim N(0, cI)$ with c large

  – $\sigma_u^2 \sim U(0, m)$ with m large.

- To make inferences under either of the above two models we used WinBUGS 1.4 with 10,000 simulations of the model parameters.

Example: SAIPE state 5-17 poverty rate model for 1989

Effects of assuming a t-distribution on the posterior mean of σ_u^2

t-distribution assumed for	degrees of freedom (ν) 3	4	5	8	∞ (normal)
u_i	3.4	2.8	2.6	2.3	2.2
e_i	3.3	2.4	2.1	1.9	2.2

United States Census Bureau

Example: SAIPE state 5-17 poverty rate model for 1989

Standardized regression residuals for Connecticut conditional on θ_i

t-distribution assumed for	degrees of freedom (ν)				
	3	4	5	8	∞ (normal)
u_i	-1.94	-2.26	-2.43	-2.65	-2.80
e_i	-1.45	-1.71	-1.86	-2.19	-2.80

(Note: Bayesian results above differ some from ML results given earlier.)

Example: SAIPE state 5-17 poverty rate model for 1989

Posterior means and variances for Connecticut

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

Note for Connecticut: $y_i = 2.2\%$, $\mathbf{x}_i'\hat{\beta} = 10.5\%$, Var$(e_i)^{.5} = 2.5$

United States Census Bureau

5. Conclusions

Sampling variances aren’t known, they are estimated.

- Lowers prediction accuracy
- Affects MSE estimates
- Try to improve estimated sampling variances $\hat{v}_i$ (e.g., with a GVF model)

Model estimation can produce $\hat{\sigma}_u^2 = 0$, which leads to awkward prediction results.

- Bayesian approach with an appropriate prior can address this

Outliers in Fay-Herriot models can arise from the random area effects (u_i) or from the sampling errors (e_i) in the direct survey estimates.

- The implications of these two sources for outliers are very different.
- Try to use additional information to determine the source.

References

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.

Datta, Gauri S. and Lahiri, Partha (2000), "A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small-Area Estimation Problems," Statistica Sinica, 10, 613–627.

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. (2004), *Bayesian Data Analysis, 2nd edition*, New York: Chapman and Hall.

Ghosh, Malay, Maiti, Tapabrata, and Roy, Ananya (2008), "Influence Functions and Robust Bayes and Empirical Bayes Small Area Estimation," *Biometrika*, **95**, 563–585.

Jiang, Jiming, Lahiri, Partha, and Wan, Shu-Mei (2002), "A Unified Jackknife Theory for Empirical Best Prediction with M-estimation," *The Annals of Statistics*, **30**, 1782–1810.

Rivest, Louis-Paul and Vandal, Nathalie (2003), "Mean Squared Error Estimation for Small Areas when the Small Area Variances are Estimated," in Proceedings of the International Conference on Recent Advances in Survey Sampling, eds. G. Roberts and D. Bellhouse, Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, Canada.

Sugasawa, Shonosuke , Tamae, Hiromasa, and Kubokawa, Tatsuya (2017), "Bayesian Estimators for Small Area Models Shrinking Both Means and Variances," *Scandinavian Journal of Statistics*, **44**, 150-167.

Yoshimori, Masayo and Lahiri, Partha (2014), "A New Adjusted Maximum Likelihood Method for the Fay-Herriot Small Area Model," *Journal of Multivariate Analysis*, **124**, 281–294.

You, Yong and Chapman, Beatrice (2006), "Small Area Estimation Using Area Level Models and Estimated Sampling Variances," *Survey Methodology*, **32**, 97–103.

Wang, Junyuan and Fuller, Wayne A. (2003), "The Mean Squared Error of Small Area Predictors Constructed with Estimated Error Variances," *Journal of the American Statistical Association*, **98**, 716–723.

Prasad, N.G.N. and Rao, J.N.K. (1990), "The Estimation of Mean Squared Errors of Small Area Estimators," *Journal of the American Statistical Association*, **78**, 47–59.