Biostatistics (2016), **0**, 0, *pp.* 1–8
doi:10.1093/biostatistics/Supplementary

The positive effects of population based preferential sampling in environmental epidemiology. Supplementary Materials

JOSEPH ANTONELLI*
Department of Biostatistics, Harvard University, Boston, MA, 02115
jantonelli@fas.harvard.edu

MATTHEW CEFALU
RAND Corporation, Santa Monica, CA 90401

LUKE BORNN
Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

APPENDIX

A. Details of bias calculation from Section 4

We define the vector (Y, X, X^*, C, C^*) to be jointly normal, and write this distribution as

$$\begin{pmatrix} Y \\ X \\ X^* \\ C \\ C^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_x \\ \mu_{x^*} \\ \mu_c \\ \mu_{c^*} \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{yx} & \Sigma_{yx^*} & \Sigma_{yc} & \Sigma_{yc^*} \\ & \sigma_x^2 & \Sigma_{xx^*} & \Sigma_{xc} & \Sigma_{xc^*} \\ & & \Sigma_{x^*} & \Sigma_{x^*c} & \Sigma_{x^*c^*} \\ & & & \Sigma_c & \Sigma_{cc^*} \\ & & & & \Sigma_{c^*} \end{pmatrix} \right)$$

We impose the following models:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$X = C\alpha + \epsilon_x$$

$$X^* = C^*\alpha + \epsilon_x^*$$

Recall that we also defined our exposure, W as

*To whom correspondence should be addressed.

$$W = \mu_X(\hat{\alpha}) + \Sigma_{X,X^*}(\hat{\phi})\Sigma_{X^*,X^*}(\hat{\phi})^{-1}(X^* - \mu_{X^*}(\hat{\alpha}))$$
$$= C\hat{\alpha} + \Sigma_{X,X^*}(\hat{\phi})\Sigma_{X^*,X^*}(\hat{\phi})^{-1}(X^* - C^*\hat{\alpha})$$

Where the random variables, C, C^*, and X^* are normally distributed and therefore W is normally distributed. We are interested in the coefficients of the model that regresses Y on W, i.e the conditional distribution of Y given W, which can now be written as

$$Y|W \sim N\left(\mu_y + \frac{\sigma_{yw}}{\sigma_w^2}(W - \mu_w), \sigma_y^2 - \frac{\sigma_{yw}^2}{\sigma_w^2}\right)$$

and the coefficient of interest is the one that lies in front of W in the mean component of the above conditional distribution. Using this we can derive the bias conditional on a given set of monitors and estimates of the first stage model parameters, $\hat{\theta}$

$$\begin{aligned}
E(\hat{\beta}_1|\hat{\theta}) &= f(\hat{\theta}) \\
&= \frac{\sigma_{yw}}{\sigma_w^2} \\
&= \frac{cov(Y,W)}{cov(W,W)} \\
&= \frac{cov(X\beta + \epsilon, W)}{cov(W,W)} \\
&= \beta_1 \frac{cov(X,W)}{cov(W,W)} \\
&= \beta_1 \left\{\frac{A}{B}\right\}
\end{aligned}$$

Where

$$A = \alpha\Sigma_c\hat{\alpha} + \alpha\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*c}\alpha - \alpha\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*c}\hat{\alpha} + \hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{x^*x}$$

$$\begin{aligned}
B = {} & \hat{\alpha}\Sigma_c\hat{\alpha} + \alpha\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*}\hat{\Sigma}_{x^*}^{-1}\hat{\Sigma}_{x^*x}\alpha + \hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{x^*}\hat{\Sigma}_{x^*}^{-1}\hat{\Sigma}_{x^*x} \\
& + \hat{\alpha}\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*}\hat{\Sigma}_{x^*}^{-1}\hat{\Sigma}_{x^*x}\hat{\alpha} + 2\hat{\alpha}\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*c}\alpha - 2\hat{\alpha}\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*c}\hat{\alpha} \\
& - 2\alpha\hat{\Sigma}_{xx^*}\hat{\Sigma}_{x^*}^{-1}\Sigma_{c^*}\hat{\Sigma}_{x^*}^{-1}\hat{\Sigma}_{x^*x}\hat{\alpha}
\end{aligned}$$

Noting that all estimated covariance matrices are estimated because we need to estimate the vector of parameters, ϕ, that represent the covariance function parameters. One example is $\hat{\Sigma}_{x^*}$, where we have suppressed the dependence on $\hat{\phi}$. We could have alternatively written $\hat{\Sigma}_{x^*}(\hat{\phi})$, though we shorten it for brevity.

To gain more intuition into this bias we can perform a taylor series expansion of $f(\hat{\theta})$ around $f(\theta)$.

$$f(\hat{\theta}) - f(\theta) \approx \frac{\partial f(\theta)}{\partial \theta}(\hat{\theta} - \theta) + \frac{1}{2}(\hat{\theta} - \theta)^T \frac{\partial^2 f(\theta)}{\partial \theta \partial \theta^T}(\hat{\theta} - \theta) \tag{A.1}$$

and now we can take the expectation on both sides with respect to the distribution governing the monitoring locations. Denoting these expectations by $E_{S^*}()$ we see that

$$\begin{aligned} E_{S^*}\left(f(\hat{\theta}) - f(\theta)\right) &= E_{S^*}(\hat{\beta}_1 - \beta_1) \\ &\approx \frac{\partial f(\theta)}{\partial \theta} E_{S^*}(\hat{\theta} - \theta) + \frac{1}{2} Tr\left(\frac{\partial^2 f(\theta)}{\partial \theta \partial \theta^T} Var_{S^*}(\hat{\theta} - \theta)\right) \\ &+ \frac{1}{2} E_{S^*}(\hat{\theta} - \theta)^T \frac{\partial^2 f(\theta)}{\partial \theta \partial \theta^T} E_{S^*}(\hat{\theta} - \theta) \end{aligned}$$

So the marginal bias (no longer conditional on an estimate of θ from the first stage model) is a function of the bias and variance of the first stage model parameters.

B. Trade-off for variance of $\hat{\beta}_1$

In the main text, we illustrated the trade-off that comes with preferential sampling for the measurement error variance, $var(X - W)$. We used this to show how preferential sampling could lead to less measurement error variance and therefore less variance in estimating β_1. Here we illustrate directly how this trade-off manifests in the estimation of $\hat{\beta}_1$ by making simplifying assumptions and approximations. Let's assume that our exposure surface follows:

$$\begin{pmatrix} X \\ X^* \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_X(\alpha) \\ \mu_{X^*}(\alpha) \end{pmatrix}, \begin{pmatrix} \Sigma_{X,X}(\phi) & \Sigma_{X,X^*}(\phi) \\ \Sigma_{X^*,X}(\phi) & \Sigma_{X^*,X^*}(\phi) \end{pmatrix} \right\}$$

and that we estimate exposure W via

$$W_i = C_i\hat{\alpha}$$

where C_i represents a covariate and there is no intercept, because it is centered. The exposure model parameter, $\hat{\alpha}$ is estimated using least squares as

$$\hat{\alpha} = \frac{\sum_{j=1}^{n^*} C_j^* X_j^*}{\sum_{j=1}^{n^*} C_j^{*2}}$$

Then, conditional on our estimates, W, we estimate the parameter of our outcome model, which again for simplification we assume is centered with no intercept and estimated via least squares

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i^2} \\
&= \frac{\sum_{i=1}^{n} \hat{\alpha} C_i Y_i}{\sum_{i=1}^{n} \hat{\alpha}^2 C_i^2} \\
&= \frac{1}{\hat{\alpha}} \frac{\sum_{i=1}^{n} C_i Y_i}{\sum_{i=1}^{n} C_i^2} \\
&= \frac{\hat{\eta}}{\hat{\alpha}}
\end{aligned}$$

where now we have written the estimate of β_1 as a ratio of two random variables, one of which involves the monitor locations and the other involving the subject locations. Now we take the variance of this ratio and apply a taylor series approximation to the variance of a ratio

$$\begin{aligned}
var(\hat{\beta}_1) &= var(\frac{\hat{\eta}}{\hat{\alpha}}) \\
&\approx \left(\frac{E(\hat{\eta})}{E(\hat{\alpha})}\right)^2 \left[\frac{var(\hat{\eta})}{E(\hat{\eta})^2} + \frac{var(\hat{\alpha})}{E(\hat{\alpha})^2} - 2\frac{cov(\hat{\eta}, \hat{\alpha})}{E(\hat{\alpha})E(\hat{\eta})}\right]
\end{aligned}$$

We also assume that $E(\hat{\alpha}) \approx \alpha$ regardless of the sampling scheme for the location of the monitors, so that this term does contribute to the $var(\hat{\beta}_1)$. Since $\hat{\eta}$ is not dependent on the monitor locations,

we can now see that only two terms in the expression for the variance of $\hat{\beta}_1$ depend on the locations of the monitors. $cov(\hat{\eta}, \hat{\alpha})$ and $var(\hat{\alpha})$ will both change as a function of the monitors. Writing these terms out we see that

$$\begin{aligned} cov(\hat{\eta}, \hat{\alpha}) &= cov\left(\frac{\sum_{i=1}^{n} C_i Y_i}{\sum_{i=1}^{n} C_i^2}, \frac{\sum_{j=1}^{n^*} C_j^* X_j^*}{\sum_{j=1}^{n^*} C_j^{*2}}\right) \\ &= \frac{\beta_1}{(\sum_{i=1}^{n} C_i^2)(\sum_{j=1}^{n^*} C_j^{*2})} \sum_{i=1}^{n}\sum_{j=1}^{n^*} C_i C_j^* cov\left(\epsilon_{x_i}, \epsilon_{x_j^*}\right) \end{aligned}$$

Which will go up on average under preferential sampling because the locations of the monitors will be closer to the locations of the subjects. Now we can look at

$$\begin{aligned} var(\hat{\alpha}) &= cov(\hat{\alpha}, \hat{\alpha}) \\ &= cov\left(\frac{\sum_{j=1}^{n^*} C_j^* X_j^*}{\sum_{j=1}^{n^*} C_j^{*2}}, \frac{\sum_{j=1}^{n^*} C_j^{*2} X_j^*}{\sum_{j=1}^{n^*} C_j^{*2}}\right) \\ &= \frac{1}{(\sum_{j=1}^{n^*} C_j^{*2})^2} \sum_{j=1}^{n^*}\sum_{k=1}^{n^*} C_j^* C_k^* cov\left(\epsilon_{x_j^*}, \epsilon_{x_k^*}\right) \end{aligned}$$

Which will also go up on average under preferential sampling because the monitors will be located more closely to each other. Now we have illustrated the trade-off that comes with preferential sampling. On one hand the variance of $\hat{\beta}_1$ will go down under preferential sampling, since the monitors are closer to the subjects and $cov(\hat{\eta}, \hat{\alpha})$ goes up leading the overall variance to go down. On the other hand preferential sampling makes $var(\hat{\alpha})$ go up, which increases the variance of $\hat{\beta}_1$ as monitors get closer together.

C. Two dimensional simulation study

The two-dimensional study presented here will be very similar in structure to the one-dimensional study seen in the original manuscript. We again can define our matrix of covariates that prediction exposure to be C. In this case C consists of an intercept, and two covariates that repre-

sent elevated levels of population level. We will simulate our locations s to lie in the uniform grid, [0,1] by [0,1]. The first covariate is an indicator that s_i lies in the circle of radius 0.03 around the center of the grid. The second covariate is set to be $\left(1 - 10 * \| s_i - \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} \|\right) * 1(s_i$ is greater than 0.03 and less than 0.08 away from the center), which effectively produces a concentric circle where the value decreases as s_i moves away from the center of the grid. We set $\alpha = (5, 3, 3)$, which produces a similar effect to what was seen in the paper that the highest population area has the highest exposure, and the exposure steadily decreases as you move away form the highest population area. We simulate under this scenario as it is the one that is most likely to occur in Environmental Epidemiology. We use the same simple linear regression model to simulate our outcome

$$Y = \beta_0 + X\beta_1 + \epsilon$$

We restrict attention to the case where we have 30 monitors, though we do not expect the results to drastically change if we increase the number of monitors. Figure C.1 shows the measurement error variance, absolute bias of β_1, and variance of β_1 across 10,000 simulations. We see very similar results as those seen in the one-dimensional setting as all metrics point to preferential sampling improving inference, particularly near $p = 1$.

We can also look at the estimates of the exposure model parameters. Table C.1 shows the mean and standard errors of the estimated exposure model parameters across 10,000 simulations. We again see very similar results to those seen in the one-dimensional setting as there is very little difference in the bias, but differences in the variances. As before there is a slight increase in standard error of the intercept when switching from $p = 0$ to $p = 1$. Similarly to the one dimensional case we see a decrease in the standard errors of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ when going from $p = 0$ to $p = 1$. If we preferentially sample too far, by setting $p = 2$, then the standard error for $\hat{\alpha}_2$ increases back to a similar magnitude as when $p = 0$.

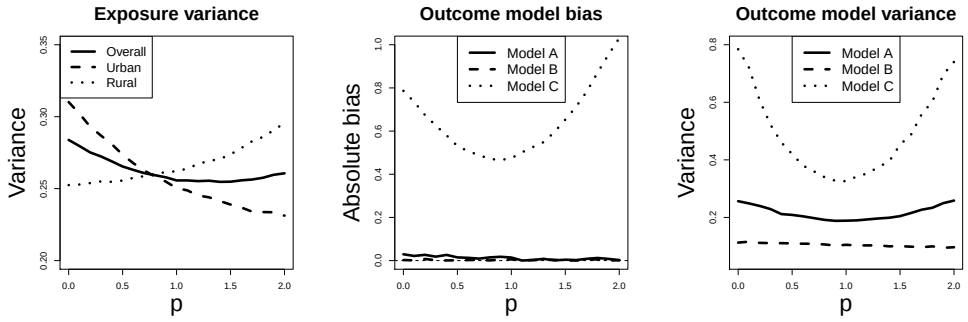

Fig. C.1. Left panel shows the measurement error variance of the predicted exposure, the middle panel shows the absolute bias in the estimation of β_1 under the three different models, and the right panel shows the variance of β_1 under the three different models.

p	α_0	α_1	α_2	ϕ
0.00	5.00 (0.12)	3.00 (0.38)	3.00 (0.36)	0.03 (0.08)
0.50	5.00 (0.13)	3.00 (0.32)	3.00 (0.31)	0.03 (0.08)
1.00	5.00 (0.15)	3.00 (0.29)	3.00 (0.3)	0.02 (0.08)
1.50	5.00 (0.18)	3.01 (0.28)	3.01 (0.31)	0.02 (0.07)
2.00	4.99 (0.23)	3.01 (0.31)	3.01 (0.35)	0.02 (0.07)

Table C.1. Mean and standard errors of exposure model parameters across 10,000 simulations. True values are $\alpha_0 = 5, \alpha_1 = 3, \alpha 2 = 3, \phi = 0.05$

D. Simulation study with confounders in health outcome model

This simulation will follow the exact same structure as in the manuscript, though we change the outcome model to be the following:

$$Y = \beta_0 + X\beta_1 + Z\beta_z$$

where Z is a matrix representing two confounders. We simulated one confounder to come from a standard Normal distribution, and another to be bernoulli with probabilty 0.3. Our exposure model now takes the following form

$$X = C\alpha + Z\alpha_z$$

where as before, $\alpha = (5, 3, 3, -3)$, and now $\alpha_z = (1, 1)$. In our outcome model we again set

$\beta_0 = 100$, $\beta_1 = 5$, and $\beta_z = (1, -1)$. For brevity we will restrict attention to the scenario where we have 30 monitors, though we don't expect the results to substantively differ for different numbers of monitors. Figure D.1 shows results from this simulation. Each panel represents one of the key figures from the simulation study in the main manuscript (figures 2-4) in the manuscript. The left panel shows the measurement error variance across 10000 simulations, the middle panel shows the absolute bias of β_1, and the right panel shows the variance of β_1. We see essentially identical patterns as what we saw in the manuscript when we did not include a vector of confounders. All three figures point to the fact that preferential sampling, particularly when $p = 1$, leads to improved inference overall.

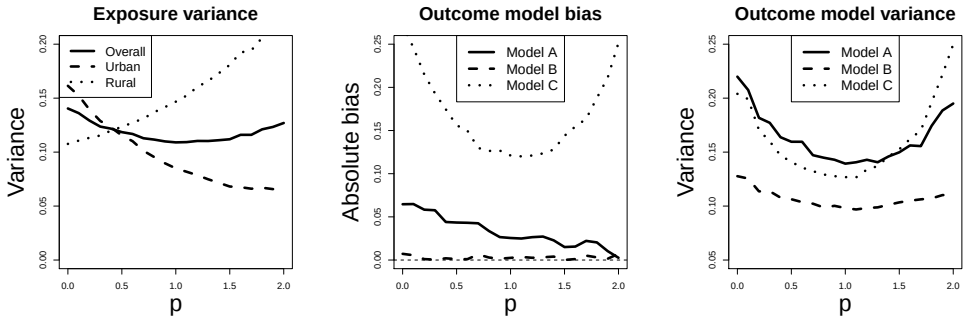

Fig. D.1. Left panel shows the measurement error variance of the predicted exposure, the middle panel shows the absolute bias in the estimation of β_1 under the three different models, and the right panel shows the variance of β_1 under the three different models.