Correcting Ecological Bias from Correlated Continuous Predictors with Individual Sampling

Sophie Woodward

May 15, 2021

Abstract

In environmental and health studies, it is often the case that the primary source of data exists at the level of the group, and not at the level of the individual. Although group analyses may be performed, associations do not translate to the individual level. Estimates of associations obtained from the group level analyses often exhibit "ecological bias", due to within-area variability, unmeasured confounding, contextual effects and other sources. Within recent decades, multiple solutions have been developed to correct ecological bias. One of those solutions is hierarchical modeling, which incorporates small amounts of individual covariate and response data into the group model. In the following simulation study, I evaluate the performance of the hierarchical modeling in addressing ecological bias resulting from correlated covariates. The simulated data is roughly based off a 2020 study relating COVID-19 to air pollution [1]. I find that hierarchical modeling in this context exhibits slight improvement over the group model, and stratified sampling can provide covariance estimates that improve the model.

Introduction

Ecological bias, also known as ecological fallacy, refers to an erroneous inference about individuals based on findings on the group level. An analysis solely using group data may detect associations that do not exist or are in the opposite direction of the true relationship at the individual level. Unfortunately it is often the case that data at the individual level is not available, so a group analysis may be the only choice. Such an analysis allows one to make conclusions at the area level, which is important for policy-making [1]. On the other hand, if small amounts of individual data are available, ecological bias can be corrected. A hierarchical modeling framework incorporating both group and individual data can be used to infer individual association.

Sources of ecological bias include within-area variability, unmeasured confounding, and contextual effects, among others. Studies within the last two decades have studied these sources through both simulation and real data application [2][3][4]. In this simulation study, I focus on the effects of confounding on ecological bias. Although the effect of dependence relations between binary and continuous predictors on ecological bias have been extensively explored, dependence relations between two continuous predictors are not so well understood. In a simulation by Salway and Wakefield [3] they conclude that there is "little consistency in the size of bias compared to the extent of correlation misspecification". I aim to challenge their result in this simulation study, as well as explore possible sampling techniques to best estimate correlation.

In particular, I examine the incorporation of individual data in a disease model with two correlated continuous predictors. I am motivated by epidemiological studies, especially within the last year [1], that perform analyses with continuous predictors on the group level and now beg extension to analyses on the individual level. I seek to answer the following questions.

Research Questions

- How is ecological bias affected by correlation between continuous, normally-distributed predictors?
- What types of sampling are most effective at alleviating ecological bias in this situation?

Hierarchical Model Framework

The hierarchical model framework outlined by Jackson [2], Salway [3], Richardson [5], and others outlines a method to incorporate samples of individual data with group data to better estimate the true individual relationship between predictors and a binary response. We are interested in the association between cases of a disease and two continuous predictors, such as two different air pollution exposures. Specifically, we assume an underlying model at the individual level as follows

$$y_{ij} \sim \text{Bern}(p_{ij}), \tag{1}$$

$$\text{logit}(p_{ij}) = \mu_i + \alpha x_{ij} + \beta z_{ij} \tag{2}$$

where μ_i is the baseline risk of group i, x_{ij} and z_{ij} are continuous predictors, p_{ij} is the probability of disease, and y_{ij} is the binary disease outcome for individual j of group i. Given access to the predictor and response data of all individuals in our population, the log likelihood would be simple to write down and maximize. Thus it would be easy to obtain maximum likelihood estimates for α and β. However, few datasets have individual information available. Instead assume most of our data exists at the group level; that is, we are given the average of each continuous predictor in each group, and cases of disease in each group. Let y_i be the number of cases in group i, n_i be the population of group i, and g_i be the joint distribution of the continuous predictors. On the group level, we model

$$y_i \sim \mathrm{Bin}(n_i, p_i), \tag{3}$$

$$p_i = \int\int \Big(\mathrm{expit}(\mu_i + \alpha x + \beta z) g_i(x, z)\Big) dx dz. \tag{4}$$

where $\mathrm{expit}(x) = \frac{e^x}{1+e^x}$, the inverse logit function. If we observe that our continuous predictors are constant within group i then it is straightforward to simplify the integral above. Unfortunately this is not a realistic assumption to maintain. It has been shown that ignoring the within-group variance of predictors leads to ecological bias [3]. Instead, let us assume the predictors are distributed as

$$\begin{pmatrix} X_i \\ Z_i \end{pmatrix} \sim \mathcal{MVN}\Bigg(\begin{pmatrix} m_i^X \\ m_i^Z, \end{pmatrix}, \Sigma_i \Bigg). \tag{5}$$

Then we can rewrite the integral as

$$p_i \approx \mathrm{expit}\left(\frac{\mu_i + \alpha m_i^X + \beta m_i^Z}{\sqrt{(1 + (\frac{16\sqrt{3}}{15\pi})^2 \left(\begin{smallmatrix} \alpha & \beta \end{smallmatrix}\right) \Sigma_i \left(\begin{smallmatrix} \alpha \\ \beta \end{smallmatrix}\right))}} \right) \tag{6}$$

using the approximation $\mathrm{expit}(x) \approx \Phi(\frac{16\sqrt{3}}{15\pi} x)$ [3]. In this simulation we estimate entries of Σ_i with sample variances and covariances, under different sampling techniques.

Now we are able to extend this model to combine group and individual data. Observe that the coefficients μ_i, α, β, are shared by both the individual and group models. The likelihood for the group data alone and individual data alone equal

$$L_{\text{group}}(\mu_i, \alpha, \beta | \{y_i\}) = \prod_i p_i^{y_i} (1 - p_i)^{N_i - y_i}, \tag{7}$$

$$L_{\text{indiv}}(\mu_i, \alpha, \beta | \{y_{ij}\}) = \prod_{i,j} \mathrm{expit}(\mu_i + \alpha x_{ij} + \beta z_{ij})^{y_{ij}} (1 - \mathrm{expit}(\mu_i + \alpha x_{ij} + \beta z_{ij}))^{1 - y_{ij}}. \tag{8}$$

The combined likelihood for both group and individual data is the product of the two likelihoods above [2],

$$L_{\text{comb}}(\mu_i, \alpha, \beta | \{y_i\}, \{y_{ij}\}) = L_{\text{group}}(\mu_i, \alpha, \beta | \{y_i\}) \times L_{\text{indiv}}(\mu_i, \alpha, \beta | \{y_{ij}\}). \tag{9}$$

Maximum likelihood estimates for μ_i, α, and β obtained from this combined likelihood are theoretically less biased than those obtained from maximizing the group likelihood or individual likelihood alone. We compute and maximize the combined likelihood using the ecoreg package [6] in R. While this package also has the ability to compute and maximize the group and individual likelihoods, we instead use the GLM function in R for these likelihoods, which has faster computational speed. There is little difference in maximum likelihood estimates between the two methods.

Simulation Setup

As described above, we are interested in modeling the prevalence of a common disease against correlated covariates X and Z, given that our main source of data exists at the county level. For instance, the disease may be COVID-19, and the continuous covariates may be air pollution exposure to $PM_{2.5}$ and O_3 [1]. We assume that the continuous covariates are truly normally distributed. In particular, we let

$$\begin{pmatrix} X_i \\ Z_i \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} m_i^X \\ m_i^Z, \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{10}$$

so that $\rho = \text{Cov}(X_i, Z_i) = \text{Corr}(X_i, Z_i)$. We investigate the effect of ρ on the bias of $\hat{\alpha}$ and $\hat{\beta}$. Assume

- $\alpha = 1$, $\beta = 2$, $\mu_i = 0.05\ \forall i$, so that both air pollution exposures are positively associated with the disease,
- number of counties = 50, number of people per county = 100. Total population = 5000.
- $m_i^X, m_i^Z \overset{\text{i.i.d}}{\sim} \mathcal{N}(1, 0.01)$, so that mean exposures do not differ largely across counties, which leads to another source of ecological bias.

In the following simulations, we vary $\rho \in \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$, that is, increments of $1/4$ between -1 and 1. This simulation study consists of three parts. In the first part, bias of coefficient estimates are compared between individual, group, and hierarchical models. In the second part bias of coefficient estimates are compared for the hierarchical model for differing amounts of data sampled. In the third part, we estimate covariance matrices Σ_i using individual sampling and incorporate that information into our hierarchical model.

Results

Preliminary Comparison of Hierarchical Model with Individual and Group Models

Below we compare bias of coefficient estimates between three models:

1. a model based on all individual data (the "individual" model)
2. a model based on solely county data (the "group" model)
3. a model based on county data and a simple random sample of 500 individual datapoints (the "hierarchical" model).

1000 simulations were performed for each value of ρ. In each simulation, data was generated using the parameters specified in the simulation setup. The group model, individual model, and hierarchical model were fit to the generated data, and coefficient estimates were recorded. Table 1 displays coefficient estimates, percent biases, and standard errors of $\hat{\alpha}$, $\hat{\beta}$ for $\rho = -1, 0$ and 1.

ρ	Model	$\overline{\hat{\alpha}}$	% bias	SE($\hat{\alpha}$)	$\overline{\hat{\beta}}$	% bias	SE($\hat{\beta}$)
-1	Individual	0.984	-1.57	0.431	1.99	-0.646	0.432
-1	Group	0.916	-8.35	0.477	1.84	-7.76	0.469
-1	Hierarchical	0.992	-0.848	0.431	1.8	-9.81	0.426
0	Individual	1	0.121	0.0584	2	0.0817	0.0743
0	Group	0.652	-34.8	0.259	1.28	-35.9	0.257
0	Hierarchical	0.634	-36.6	0.0985	1.27	-36.6	0.109
1	Individual	1.01	0.86	0.384	2	0.0986	0.383
1	Group	0.517	-48.3	0.248	1.01	-49.6	0.25
1	Hierarchical	0.599	-40.1	0.238	1.11	-44.4	0.238

Table 1: Estimates, % Biases, Standard Errors of $\hat{\alpha}$, $\hat{\beta}$ for $\rho = -1, 0, 1$

Overall, we find that the individual model incorporating all individual data results in zero bias across all correlation values, for both $\hat{\alpha}$ and $\hat{\beta}$. For both the group and hierarchical models, biases of $\hat{\alpha}, \hat{\beta}$ are negative, and the magnitude of biases increase with correlation. Observe the negative slopes of the red and blue curves in Figure 1. These results suggest that positive correlation between predictors increases ecological bias, when these predictors are normally-distributed and have positive associations with the response. The smallest magnitude of bias of the group and hierarchical models is achieved at correlation of -1, and the largest at a correlation of 1. Because the covariates both have positive associations with the response, their effects on the response are less distinguishable for positive correlations. On the other hand, when the covariates are negatively correlated, their effects on the response are easy to distinguish.

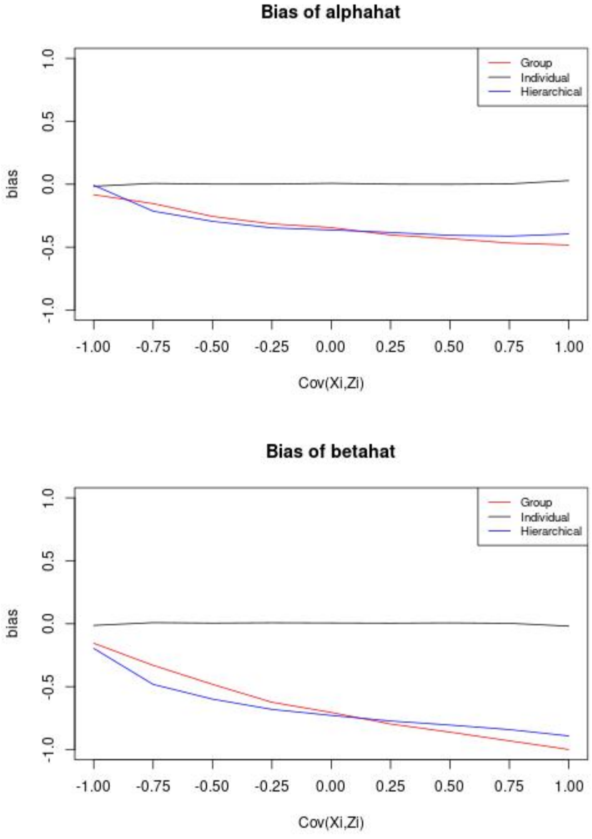


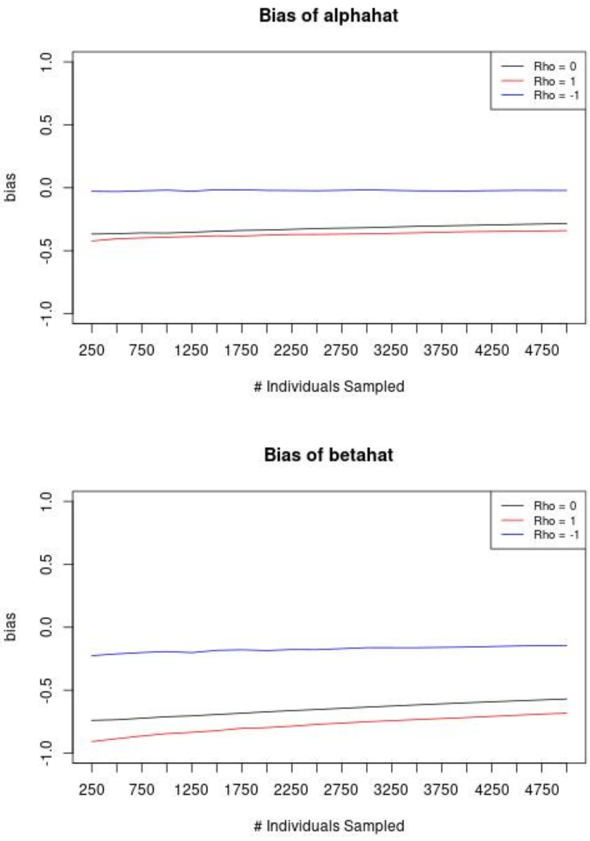
Figure 1: Bias of $\hat{\beta}, \hat{\alpha}$, varying correlation.

There is little difference between the group model and hierarchical model bias. Perhaps this is because sampling 500 individuals at random is not sufficient to distinguish the hierarchical model from the group model. However, the hierarchical model exhibits greater improvement in bias relative to the group model as correlation increases. In Figure 1, the slope of the blue curve is less steep than the slope of the red curve. This may suggest that it is valuable to incorporate small amounts of individual data into group analyses especially when suspecting high positive correlation between predictors.

Amount of Individual Sampling

We may also ask how much individual data in a simple random sample is sufficient to alleviate bias in the hierarchical model. In this part of the simulation, we vary the number of individuals incorporated from $k = 250$ (5 % of population) to $k = 5000$ (100 % of population). 1000 simulations were performed for each value of k. In each simulation, data was generated with covariances $\rho = -1, 0, 1$ and a hierarchical model was fit. As expected, we find that incorporating more individuals results in less extreme bias. In the first plot in Figure 2, the slopes of the red and black lines are positive, so bias of $\hat{\alpha}$ becomes less extreme as more individuals are sampled for $\rho = 0, 1$. Bias of $\hat{\alpha}$ equals 0 for $\rho = -1$. In the second plot in Figure 2, we find that the slopes of all lines are also positive, so biases of $\hat{\beta}$ becomes less extreme as more individuals are sampled for $\rho = -1, 0, 1$. The magnitudes of these positive slopes are not large, suggesting that incorporating many individuals is not a large improvement over incorporating a few, a conclusion validated in the literature [2].

Figure 2: Bias of $\hat{\beta}, \hat{\alpha}$, varying number of individuals incorporated into the hierarchical model.



Note that the hierarchical model is inaccurate when individual data consists of a large proportion of the population, since the model over-accounts for the individual data provided. The individual data shows up in the combined likelihood both through the individual likelihood and group likelihood. When all individual data is provided, the likelihood for the hierarchical model should equal the individual model likelihood, ie equation (9) should equal equation (8).

Variance and Sampling

The source of ecological bias in this simulation study comes from correlation between X_i and Z_i and their variances. In this last part of the simulation, we incorporate variance estimation into the hierarchical modeling. We compare three specifications of Σ_i: the default estimate of Σ_i as the zero matrix, an estimate of Σ_i from stratified sampling, and the true Σ_i.

1. If no estimate of Σ_i is provided, Σ_i is assumed to equal the zero matrix, so that X_i and Z_i are assumed to be constant within county i at their respective means.

2. Alternatively, we can estimate Σ_i from stratified sampling. Instead of sampling 500 individual datapoints at random, we sample 10 individuals from each county. Let $\vec{x_i}^{10}$, $\vec{z_i}^{10}$ denote the covariate data of these individuals. Then an estimate for Σ_i is

$$\hat{\Sigma}_i = \begin{pmatrix} \text{Var}(\vec{x_i}^{10}) & \text{Cov}(\vec{x_i}^{10}, \vec{z_i}^{10}) \\ \text{Cov}(\vec{x_i}^{10}, \vec{z_i}^{10}) & \text{Var}(\vec{z_i}^{10}) \end{pmatrix}. \tag{11}$$

   We substitute $\hat{\Sigma}_i$ for Σ_i in the expression for p_i in equation (5) and update the likelihoods.

3. When the true Σ_i is used in equation 5, we have

$$\hat{\Sigma}_i = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

   Theoretically this specification of Σ_i should result in the smallest bias.

Below we vary ρ in increments of 1/4 from -1 to 1. 1000 simulations were carried out for each value of ρ. In each simulation, the three estimates of Σ_i were calculated and incorporated into three different hierarchical models. As expected, specifying Σ_i as its true value results in approximately 0 bias. The stratified sampling estimate performs reasonably well. For correlation values greater than -0.75, the stratified sampling estimate performs better than the zero estimate, which performs the worst. We also observe that bias does not noticeably worsen as correlation increases when Σ_i is estimated with stratified sampling. Further exploration should investigate alternative sampling techniques to best estimate Σ_i.

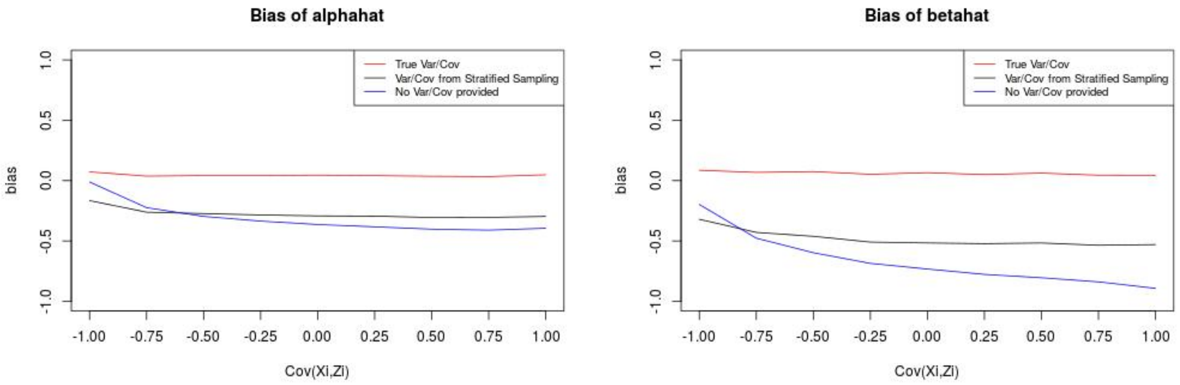


Figure 3: Bias of $\hat{\beta}, \hat{\alpha}$, comparing estimates of covariance matrices.

Discussion

To summarize the conclusions above, we find that ecological bias in this setting generally increases as correlation increases, given that the covariates are positively associated with the response. The hierarchical model consistently performs better than the group model. Incorporating large amounts of individual data into the hierarchical model improves bias, but it is powerful that small amounts of individual data can also alleviate bias substantially. Through stratified sampling, estimates of covariance of the predictors can be incorporated into the hierarchical model to greatly improve bias.

This simulation study faces several limitations. Overall, the study is simplistic and requires further extensions. First, the conclusions above are sensitive to parameter values and distributional assumptions. The values of α, β, μ will affect results. Setting $\alpha, \beta > 0$, we found that negative correlations produced the least ecological bias and positive correlations produced the most, since the effect of the covariates became inseparable. This study should be repeated with α, β having opposite signs and differing magnitudes.

Second, this study assumes that covariates are truly normally-distributed. Large magnitudes of bias may result from incorrect distributional assumptions, depending on the values of α, β [3]. This simulation study relies on a correct distributional assumption.

Third, an inadvertent source of bias in this study originates from $m_i^X, m_i^Z \overset{\text{i.i.d}}{\sim} \mathcal{N}(1, 0.01)$. This results in random effects across counties and contributes to additional ecological bias. There were singularity errors in Hessian matrices when m_i^X and m_i^Z were made constant, and these errors should be solved if this study were to be repeated.

Fourth, before addressing these simplistic assumptions and additional sources of bias, perhaps a larger challenge should be solved. The numeric optimization of the combined likelihood was extremely time-costly. Maximizing the combined likelihood in the "ecoreg" package implemented Gaussian-Hermite integration. As a result, I was limited in the number of simulations (1000).

Overall, I conclude that ecological bias is a deeply complex problem, but it can be corrected. It is always ideal to collect individual data if it is available, but if only small amounts of data are available, this information is still valuable. For the future, it is worth exploring sampling techniques and variance estimation further, in order to determine the most efficient ways to incorporate individual data into a group model.

Acknowledgements

I'd like to thank Kevin for guiding me throughout the semester, Danielle and Xiao from the Dominici lab for introducing me to ecological bias, Christopher Jackson for his package "ecoreg", and my fellow Stat 98 juniors for their constructive feedback.

References

[1] X. Wu, R. C. Nethery, M. Sabath, D. Braun, and F. Dominici, "Air pollution and covid-19 mortality in the united states: Strengths and limitations of an ecological regression analysis," *Science advances*, vol. 6, no. 45, p. eabd4049, 2020.

[2] C. Jackson, N. Best, and S. Richardson, "Improving ecological inference using individual-level data," *Statistics in medicine*, vol. 25, no. 12, pp. 2136–2159, 2006.

[3] R. Salway and J. Wakefield, "Sources of bias in ecological studies of non-rare events," *Environmental and Ecological Statistics*, vol. 12, no. 3, pp. 321–347, 2005.

[4] J. Wakefield, "Ecological inference for 2× 2 tables," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 167, no. 3, pp. 385–425, 2004.

[5] S. Richardson, I. Stücker, and D. Hémon, "Comparison of relative risks obtained in ecological and individual studies: some methodological considerations," *International journal of epidemiology*, vol. 16, no. 1, pp. 111–120, 1987.

[6] C. Jackson, N. Best, and S. Richardson, "Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 171, no. 1, pp. 159–178, 2008.