

Alleviating Linear Ecological Bias and Optimal Design with Subsample Data

Adam Glynn, Jon Wakefield, Mark S. Handcock, and Thomas S. Richardson

Department of Statistics, University of Washington, Seattle, USA

Summary. In this paper, we illustrate that combining ecological data with subsample data in situations in which a linear model is appropriate provides two main benefits. First, by including the individual level subsample data, the biases associated with linear ecological inference can be eliminated. Second, we can use readily available ecological data to design optimal subsampling schemes, so as to maximize information about parameters. We present an application of this methodology to the classic problem of estimating the effect of a college degree on wages. We show that combining ecological data with subsample data provides precise estimates of this value, and that optimal subsampling schemes (conditional on the ecological data) can provide good precision with only a fraction of the observations.

Keywords: Ecological bias; Combining information; Within-area confounding; Returns to education; Sample design.

1. Introduction

In its most inclusive definition, ecological inference is usually an attempt to estimate parameters of individual relationships with data that have been aggregated above the individual level (ecological data). Not surprisingly, this endeavor is fraught with peril, and Robinson (1950) is an early reference to some of the potential biases that may result when ecological data are used to estimate individual level parameters. Since the publication of that paper, the research community has roughly divided into two camps: those who disdain any ecological inference and advocate inference based on the sampling of individuals, e.g. Freedman et al. (1998), and those who attempt ecological inference through model assumptions, e.g. King (1997). Recent work has shown that inference can be improved by combining small samples of individual level data with ecological level data, gaining identification from the former and precision of estimates from the latter. In the case of 2×2 tables, Wakefield (2004) describes the joint likelihood for the ecological data and subsample data and shows that a combined approach reduces ecological bias. Steel et al. (2004) develops the observed information for this same case, but with the data sources treated as independent. Haneuse and Wakefield (2004) show that ecological data combined with case-control data can improve inference, and that rare case observations have the largest effect on observed information. In hierarchical linear models, Raghunathan et al. (2003) show that moment and maximum likelihood estimates of a common within group correlation coefficient will improve when aggregate data are combined with new individuals from within each group. In a similar linear hierarchical setting, Steel et al. (2003) develop the properties of moment estimators in a number of aggregate and individual data combinations.

In this paper, we assume that ecological data are available and that the researcher requires a design for subsampling individuals. This situation resembles many real world problems where ecological data are available through government agencies. In this type of application, subsample design will be of utmost importance, since data collection may be expensive, and therefore our goal is to maximize the information in our subsample, conditional on the ecological data. We will address this goal within the framework of linear models, focusing on sources of linear ecological bias, and answering the design question in terms of these sources.

The outline of this paper is as follows. In Section 2, we decompose linear ecological bias into three sources, using an approach close in spirit to Greenland and Morgenstern (1989) and Richardson (1992), and demonstrate how individual level data can correct this bias. In Section 3, we introduce a motivating application to measure the effect of a college degree on individual wages. In Section 4 we discuss maximum likelihood estimation with ecological and subsample data. Section 5 provides information comparisons between the different data sources and shows the magnitude of information gained using the combined data approach. In Section 6 we examine optimal subsampling design conditional on the ecological data. In Section 7 we apply the methodology to the college/wage example. We show that the combined data approach provides an

improvement over both the purely ecological and the purely individual approach, and that optimal sampling can further increase precision. Finally, Section 8 presents a discussion of unresolved issues and extensions for future research.

2. Sources of Ecological Bias

We first define the data at the individual and at the ecological level. We assume that we could potentially observe the triples $(x_{ij}, y_{ij}, z_{ij}^c)$ for individuals $j = 1, \dots, n_i$ in groups $i = 1, \dots, m$, where $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ is the vector of responses from group i , $\mathbf{x}_i = (x_{i1}, \dots, x_{in_i})$ is the vector of univariate exposure/covariates from group i , $\mathbf{z}_i^c = (z_{i1}^c, \dots, z_{in_i}^c)$ is the vector of confounders, and $n = \sum_{i=1}^m n_i$ represents the “full data” sample size. We will assume that we observe the ecological data that consists of the group means (\bar{x}_i, \bar{y}_i) for groups $i = 1, \dots, m$, and we may observe \bar{z}_i^c for groups $i = 1, \dots, m$. Furthermore, we assume that the n_i observed triples in group i represent i.i.d. values produced by some process, and we are interested in the parameters of this process. Within this framework, we will assume one of three models:

$$E[y_{ij} | \mathbf{x}_i, \mathbf{z}_i^c] = \beta_{0i} + \beta_w x_{ij} \quad (1)$$

$$E[y_{ij} | \mathbf{x}_i, \mathbf{z}_i^c] = \beta_{0i} + \beta_{wi} x_{ij} \quad (2)$$

$$E[y_{ij} | \mathbf{x}_i, \mathbf{z}_i^c] = \beta_{0i} + \beta_{wi} x_{ij} + z_{ij}^c \quad (3)$$

In (1), we assume that each group has a different intercept, but a common within-group slope. In (2), we assume that each group may have distinct intercepts and slopes. In (3), we assume that in addition to distinct intercepts and slopes, z_{ij}^c acts as a confounder so that $E[z_{ij}^c | x_{ij}] \neq E[z_{ij}^c]$. We do not parametrize the final term as it represents the combination of all possible confounding variables and their effects, so we could have written $z_{ij}^c = \sum_{k=1}^K \beta_k z_{ijk}$. These three models are nested, in that (1) is a special case of (2), which is a special case of (3).

The linearity of these models allows the derivation of their ecological counterparts:

$$E[\bar{y}_i | \bar{x}_i] = \beta_{0i} + \beta_w \bar{x}_i \quad (4)$$

$$E[\bar{y}_i | \bar{x}_i] = \beta_{0i} + \beta_{wi} \bar{x}_i \quad (5)$$

$$E[\bar{y}_i | \bar{x}_i, \bar{z}_i^c] = \beta_{0i} + \beta_{wi} \bar{x}_i + \bar{z}_i^c \quad (6)$$

If we are interested in only the m observed groups, then we interpret the slopes, $\beta_w = (\beta_{w1}, \dots, \beta_{wm})$, as fixed and unknown quantities. If we are interested in a superpopulation of groups, then the slopes are viewed as random quantities. For this paper, we assume that the m groups comprise the entire population of interest at the group level, and therefore β_w is fixed and unknown. However, an ecological regression based on (\bar{x}_i, \bar{y}_i) for groups $i = 1, \dots, m$ cannot identify all m of the slopes in β_w , so we often resign ourselves to estimating a convex combination of these slopes. In this paper, we assume that the weights of this combination are determined by n_i for each group, and therefore the parameter of interest is $\bar{\beta}_w \equiv \frac{1}{n} \sum_{i=1}^m n_i \beta_{wi}$. These

weights are appropriate in a number of situations, but a full discussion of this topic would be outside the scope of this paper. The ecological estimator for $\bar{\beta}_w$ is

$$\widehat{\beta}_w^{eco} = \frac{\sum_{i=1}^m n_i (\bar{y}_i - \bar{y})(\bar{x}_i - \bar{x})}{\sum_{i=1}^m n_i (\bar{x}_i - \bar{x})^2} \quad (7)$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^m n_i \bar{y}_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^m n_i \bar{x}_i$. Notice that this estimate will be undefined if the group specific covariate means are all equal to the grand covariate mean. In this case, ecological inference is not possible.

If we further define $\bar{\beta}_0 \equiv \frac{1}{n} \sum_{i=1}^m n_i \beta_{0i}$ and $\bar{z}^c \equiv \frac{1}{n} \sum_{i=1}^m n_i \bar{z}_i^c$, and we assume the most general model, (3), then the expectation of $\widehat{\beta}_w^{eco}$ conditional on $\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_m)$ can be written as the following (see Appendix A for details):

$$E[\widehat{\beta}_w^{eco} | \bar{\mathbf{x}}] = \bar{\beta}_w \quad (8)$$

$$+ \frac{\sum_{i=1}^m \{n_i (\bar{x}_i - \bar{x})(\beta_{0i} - \bar{\beta}_0)\}}{\sum_{i=1}^m n_i (\bar{x}_i - \bar{x})^2} \quad (9)$$

$$+ \frac{\bar{x} \sum_{i=1}^m \{n_i (\bar{x}_i - \bar{x})(\beta_{wi} - \bar{\beta}_w)\}}{\sum_{i=1}^m n_i (\bar{x}_i - \bar{x})^2} + \frac{\sum_{i=1}^m \{n_i (\bar{x}_i - \bar{x})^2 (\beta_{wi} - \bar{\beta}_w)\}}{\sum_{i=1}^m n_i (\bar{x}_i - \bar{x})^2} \quad (10)$$

$$+ \frac{\sum_{i=1}^m \{n_i (\bar{x}_i - \bar{x})(E[\bar{z}_i^c | \bar{\mathbf{x}}] - \bar{z}^c)\}}{\sum_{i=1}^m n_i (\bar{x}_i - \bar{x})^2} \quad (11)$$

The first term of (8) is the parameter of interest, while the remaining terms represent a decomposition of ecological bias. The numerator of (9) will be zero when the weighted sample covariance between the ecological covariate averages and the group specific intercepts is zero, and therefore, this term represents the bias due to the intercepts. Figure 1(a) shows an example of this bias. The numerator of the first term of (10) is \bar{x} times the weighted sample covariance of ecological covariate averages and the group specific slopes and therefore represents the bias due to correlation between the slopes and the ecological covariate averages. Figure 1(b) shows an example of this bias. The numerator of the second term of (10) represents the bias due to a quadratic relationship between the group specific slopes and the ecological covariate averages. Figure 1(c) shows an example of this bias. Both terms of (10) will be zero if there's no relationship between the slopes and the ecological covariate averages ($\beta_{wi} = \beta_w$ for all i is a special case), and we refer to the bias in this term as slope bias. The numerator of (11) will be zero when the weighted sample covariance between the ecological covariate averages and the projection of \bar{z}_i^c onto \bar{x}_i is zero. If the ecological covariate averages are uncorrelated with the ecological confounder averages in the joint distribution between these variables, then on average each term of (11) will be zero. Therefore, $E[\bar{z}_i^c | \bar{\mathbf{x}}] = E[\bar{z}_i^c]$ is a sufficient and almost necessary condition for (11) to be zero, and this term represents bias due to an unmeasured confounder. Our decomposition is similar to the decomposition in Equation (3) of Greenland and Morgenstern (1989) or Equation (8) of Richardson (1992), except that they assume the parameter of interest is a superpopulation average of slopes instead of a weighted average of the slopes from the m observed groups. Additionally, we have explicitly included a confounding term and taken expectations conditional on the ecological covariate vector. We now examine in detail the three sources of ecological bias that we have defined.

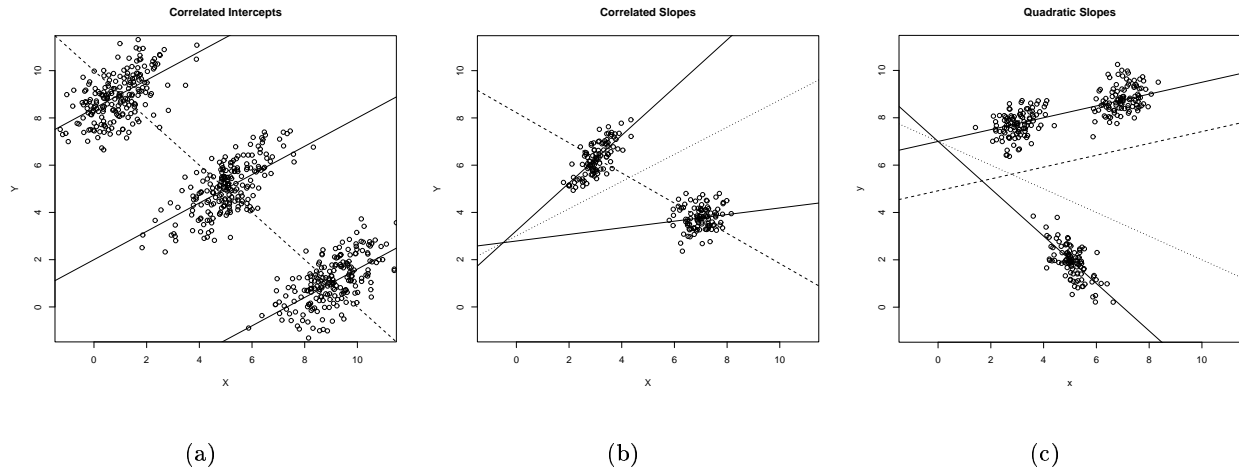


Fig. 1. Sources of ecological bias: (a) Intercept bias, group intercepts correlated with covariate group means. (b) Slope bias, group slopes correlated with covariate group means. (c) Slope bias, group slopes quadratically related to covariate group means. Solid lines are the within group regression lines. Dashed lines are the ecological regression lines. Dotted lines represent the average within group regression line, the slope of which is the parameter of interest.

2.1. Intercept Bias

If the linear expectation is given by (1), then correlated intercepts are the only possible source of ecological bias, because there is a common within group slope, and there is no confounder. Therefore, the estimate will be unbiased when the group specific intercepts are uncorrelated with the covariate group means. Figure 1(a) shows an example where this condition does not hold, and clearly illustrates that the ecological regression estimate is biased. The dashed line represents the ecological regression line, and we see that the slope of this line is negative, while the within group slopes are all positive.

Model (1) represents a causal model in that β_w is the average effect of changing x_{ij} by one and holding “everything else” constant. However, this parameterization does not illustrate when the β_{0i} s will be correlated with the \bar{x}_i s. In order to demonstrate the presence or lack of this correlation, we need a more explicit parameterization of the data generating process. There are a number of causal (data generating) models that lead to the correlated intercepts model, and we will address two: the contextual effects model and the group level confounder model. Within this context, we will use γ s to signify fixed, unknown parameters from the explicit data generating model and continue to use β s to signify fixed, unknown parameters from a less explicit causal model that only represents the causal effect of x_{ij} on y_{ij} . Furthermore, we will illustrate the relationship between the parameters from these two models.

In the contextual effects model, the covariate is assumed to have a within group effect (γ_w) and a between

group effect (γ_b):

$$\begin{aligned} E[y_{ij}|\mathbf{x}_i] &= \gamma_0 + \gamma_b \bar{x}_i + \gamma_w (x_{ij} - \bar{x}_i) \\ &= \gamma_0 + (\gamma_b - \gamma_w) \bar{x}_i + \gamma_w x_{ij} \\ &= \beta_{0i} + \beta_w x_{ij}, \end{aligned} \tag{12}$$

where $\beta_{0i} = \gamma_0 + (\gamma_b - \gamma_w) \bar{x}_i$ and $\beta_w = \gamma_w$. In the corresponding ecological model,

$$\begin{aligned} E[\bar{y}_i|\bar{x}_i] &= E[\gamma_0 + (\gamma_b - \gamma_w) \bar{x}_i|\bar{x}_i] + \gamma_w \bar{x}_i \\ &= \beta_{0i} + \beta_w \bar{x}_i, \end{aligned}$$

the intercept term, $\beta_{0i} = \gamma_0 + (\gamma_b - \gamma_w) \bar{x}_i$, is a linear function of \bar{x}_i and is therefore perfectly correlated with \bar{x}_i . Therefore, if we attempt to estimate $\beta_w = \gamma_w$ with $\hat{\beta}_w^{eco}$ we will get a biased estimate. Instead, we get an unbiased estimate of the between parameter, γ_b .

In the group level confounding model, we assume a single confounder, z_i , which only varies by group and affects both the covariate and the response.

$$\begin{aligned} E[y_{ij}|\mathbf{x}_i, z_i] &= \gamma_0 + \gamma_w x_{ij} + \gamma_c z_i \\ E[y_{ij}|\mathbf{x}_i] &= \gamma_0 + \gamma_c E[z_i|\mathbf{x}_i] + \gamma_w x_{ij} \\ &= \beta_{0i} + \beta_w x_{ij}. \end{aligned} \tag{13}$$

The intercept term is $\beta_{0i} = \gamma_0 + \gamma_c E[z_i|\mathbf{x}_i]$ and the slope term is $\beta_w = \gamma_w$. If we further assume that $E[z_i|\mathbf{x}_i] = E[z_i|\bar{x}_i]$, then the corresponding ecological model,

$$\begin{aligned} E[\bar{y}_i|\bar{x}_i] &= \gamma_0 + E[\gamma_c z_i|\bar{x}_i] + \gamma_w \bar{x}_i \\ &= \beta_{0i} + \beta_w \bar{x}_i, \end{aligned}$$

has $\beta_{0i} = \gamma_0 + E[\gamma_c z_i|\bar{x}_i]$ and $\beta_w = \gamma_w$. Failing to condition on z_i will lead to a β_{0i} that will be correlated with \bar{x}_i , unless \bar{x}_i and z_i are uncorrelated. Therefore, if we attempt to estimate $\beta_w = \gamma_w$ with $\hat{\beta}_w^{eco}$ we will get a biased estimate.

While intercept bias is a problem for ecological inference, we can remove the bias with observations of z_i , or with individual level data on \mathbf{x} and \mathbf{y} . If the linear expectation is given by (1), and we observe (x_{ij}, y_{ij}) for some individuals within each group, we can always fit a model with different intercepts for each group. This ‘‘fixed effects’’ estimation approach is well known to correct for group level confounding (Chamberlain (1984)), and it will also remove bias in any other intercept problem.

2.2. Slope Bias

If the linear expectation is given by (2), then ecological bias can arise from intercept bias or slope bias. Varying slopes is sometimes called ‘‘effect modification’’ (Greenland and Morgenstern (1989)). Figure 1(b)

shows an example of effect modification, and again, the ecological regression estimate is biased. The dashed line represents the ecological regression line, and we see that the slope of this line is negative, while the within group slopes are both positive. The slope of the dotted line represents the average within group slope. In order to motivate the correlated slopes model, we show that an explicitly parameterized data generating model gives rise to correlated slopes.

In a group level confounding model with an interaction between the confounder and the covariate, the interaction term is combined with the group specific slope when we fail to condition on the interaction ($\gamma_{int}z_i + \gamma_w$). For simplicity, we are assuming that either x_{ij} or z_i is binary.

$$\begin{aligned}
 E[y_{ij}|\mathbf{x}_i, z_i] &= \gamma_0 + \gamma_w x_{ij} + \gamma_c z_i + \gamma_{int} x_{ij} z_i \\
 E[y_{ij}|\mathbf{x}_i] &= \gamma_0 + \gamma_c E[z_i|\mathbf{x}_i] + (\gamma_{int} E[z_i|\mathbf{x}_i] + \gamma_w) x_{ij} \\
 &= \beta_{0i} + \beta_{wi} x_{ij}.
 \end{aligned} \tag{14}$$

The intercept term is $\beta_{0i} = \gamma_0 + \gamma_c E[z_i|\mathbf{x}_i]$ and the slope term is $\beta_{wi} = \gamma_{int} E[z_i|\mathbf{x}_i] + \gamma_w$. If we further assume that $E[z_i|\mathbf{x}_i] = E[z_i|\bar{\mathbf{x}}_i]$, then the corresponding ecological model,

$$\begin{aligned}
 E[\bar{y}_i|\bar{\mathbf{x}}_i] &= \gamma_0 + \gamma_c E[z_i|\bar{\mathbf{x}}_i] + (\gamma_{int} E[z_i|\bar{\mathbf{x}}_i] + \gamma_w) \bar{\mathbf{x}}_i \\
 &= \beta_{0i} + \beta_{wi} \bar{\mathbf{x}}_i,
 \end{aligned}$$

has the intercept term $\beta_{0i} = \gamma_0 + \gamma_c E[z_i|\bar{\mathbf{x}}_i]$ and the slope term $\beta_{wi} = \gamma_{int} E[z_i|\bar{\mathbf{x}}_i] + \gamma_w$, where β_{0i} and β_{wi} will be correlated with $\bar{\mathbf{x}}_i$. Therefore, if we attempt to estimate $\bar{\beta}_w$ with $\widehat{\bar{\beta}}_w^{eco}$ we will get a biased estimate. In (14) the correlated slopes are accompanied by correlated intercepts. This will be true for most models with correlated slopes, and hence we will often need to simultaneously correct both sources of bias. If the linear expectation is given by (2), and we observe (x_{ij}, y_{ij}) for some individuals within each group, we can always fit a model with different intercepts and slopes for each group. Therefore, our estimate for $\beta_{wi} = \gamma_{int} E[z_i|\bar{\mathbf{x}}_i] + \gamma_w$ within each group will be unbiased, and our estimate of the average slope will also be unbiased. Notice that β_{wi} includes both direct and indirect effects of x_{ij} on y_{ij} . In order to estimate the direct effect γ_w , we need to observe z_i .

2.3. Within Group Confounding

If the linear expectation is given by (3), then ecological bias can arise from intercepts, slopes, or an unmeasured confounder. However, the bias from a confounder cannot be corrected as easily as the previous two sources of bias because it cannot be remedied with individual level data on \mathbf{x} and \mathbf{y} only. Additionally, unmeasured confounding leads to different types of bias for individual level inference and ecological inference. In this section, we discuss these differences.

For a single within group confounder, the explicit data generating model is nearly identical to (3). We parameterize the linear expectation of y_{ij} as $\gamma_{0i} + \gamma_{wi} x_{ij} + \gamma_{ci} z_{ij}$, where γ_{ci} is the causal confounding

effect. If we don't measure z_{ij} , the bias due to unmeasured confounding can be expressed by algebraically decomposing the confounding variable into three terms: an intercept term, a slope term, and a residual term, i.e. $z_{ij} = a_i + b_i x_{ij} + u_{ij}$, where a_i and b_i are the OLS estimates from a regression of \mathbf{z}_i on \mathbf{x}_i within each group, and u_{ij} are the residuals from this regression. The individual model can then be re-written as,

$$\begin{aligned} E[y_{ij}|\mathbf{x}_i, \mathbf{z}_i] &= \gamma_{0i} + \gamma_{wi}x_{ij} + \gamma_{ci}z_{ij} \\ &= \gamma_{0i} + \gamma_{wi}x_{ij} + \gamma_{ci}(a_i + b_i x_{ij} + u_{ij}) \\ E[y_{ij}|\mathbf{x}_i] &= \gamma_{0i} + \gamma_{ci}E[a_i|\mathbf{x}_i] + (\gamma_{wi} + \gamma_{ci}E[b_i|\mathbf{x}_i])x_{ij}. \end{aligned} \quad (15)$$

Therefore, we can identify $\gamma_{0i} + \gamma_{ci}E[a_i|\mathbf{x}_i]$ and $\gamma_{wi} + \gamma_{ci}E[b_i|\mathbf{x}_i]$ with individual level data on \mathbf{y} and \mathbf{x} , but we cannot identify γ_{wi} . If we further assume that $E[a_i|\mathbf{x}_i] = E[a_i|\bar{\mathbf{x}}_i]$ and $E[b_i|\mathbf{x}_i] = E[b_i|\bar{\mathbf{x}}_i]$, then the ecological model can be re-written as,

$$E[\bar{y}_i|\bar{\mathbf{x}}_i] = \gamma_{0i} + \gamma_{ci}E[a_i|\bar{\mathbf{x}}_i] + (\gamma_{wi} + \gamma_{ci}E[b_i|\bar{\mathbf{x}}_i])\bar{\mathbf{x}}_i \quad (16)$$

where $\gamma_{0i} + \gamma_{ci}E[a_i|\bar{\mathbf{x}}_i]$ and $\gamma_{wi} + \gamma_{ci}E[b_i|\bar{\mathbf{x}}_i]$ will be correlated with $\bar{\mathbf{x}}_i$. Therefore, $\widehat{\beta}_w^{eco}$ will be a biased estimate for a parameter we are not interested in ($\frac{1}{n} \sum_{i=1}^m n_i(\gamma_{wi} + \gamma_{ci}E[b_i|\bar{\mathbf{x}}_i])$) and will also be biased for the parameter of interest ($\bar{\beta}_w = \bar{\gamma}_w$).

In summary, if we assume models (1) or (2) then we need individual level data on \mathbf{x} and \mathbf{y} to identify the parameters of the model. If we assume (3), then we need individual level data on \mathbf{x} , \mathbf{y} , and \mathbf{z} to identify the parameters.

3. Motivating Example: The Wage Value of a College Degree

In order to illustrate the problem of linear ecological bias, we will present data on wages and college degrees for individuals in the State of Washington, USA. The underlying scientific question concerning the economic value of a college degree has been well studied by labor economists. Estimating the value of a college degree is important both to members of the general public, who must decide whether to attend college, and to the government, which may seek to achieve social goals through the use of financial aid. There are a variety of definitions and estimators for the returns to education. For a comprehensive review see Card (1999, 2001), which compare different estimates of the causal effect of education on earnings in the context of the the British National Child Development Survey (NCDS). Our goal here is to demonstrate the dangers and relevance of ecological bias. The ecological data are available through the Public Use Microdata Survey (PUMS), Ruggles et al. (2004). These data represent male full time workers (35+ hours per week and 48+ weeks per year) in Washington State, aged 18 to 65 in the 2000 Census who earned between 0 and 175,000 dollars during the previous calendar year. We used this selection criteria because by convention the census recodes all yearly wages greater than 175,000 dollars to the state average of people with wages greater than 175,000. This group of high earners represented 1.7% of the data.

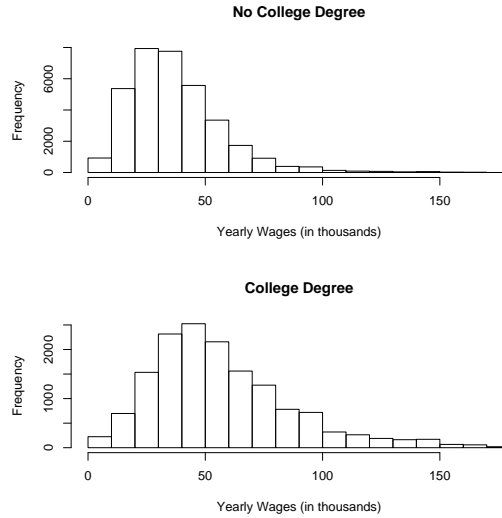


Fig. 2. Wage histograms for individuals with/without a college degree

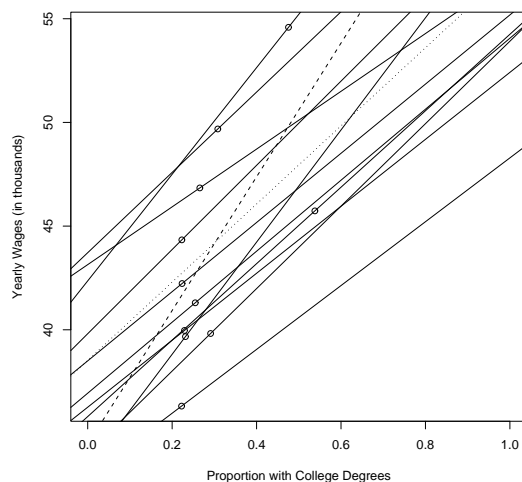
We initially examine two variables for each individual: the response, y_{ij} , is the yearly wage (in thousands) for individual j in group i , and the covariate, x_{ij} , is a college degree indicator, which takes the value 1 if individual j in group i has obtained a college degree and zero otherwise. These data are divided into eleven groups ($i = 1, \dots, 11$), where each group represents a geographical area known as a super-PUMA. Super-PUMAs are contiguous geographic areas that contain roughly 400,000 people. Populous counties are split into multiple super-PUMAs, while less populous counties may be grouped together into a single super-PUMA.

The histograms in Figure 2 show the distribution of yearly wages across all areas for individuals with and without college degrees. The skewness of these distributions is not surprising because distributions of incomes and wages are frequently known to show this shape. Usually, we would transform these data with the \log function to symmetrize the distribution. However, in most applications combining ecological and subsample data, we cannot make this transformation, because we do not have access to the original n_i observations from each group. Therefore, even though we do have access to these observations in our example, we will proceed as if we didn't, and use the untransformed data.

The linear models in Section 2 do not make explicit distributional assumptions, but in order to simplify the discussion, we will assume constant variances across groups and constant variances within each group. For our application, these assumptions appear to be somewhat problematic. The sample variances of yearly wages are moderately different across groups, and within each group the sample variance of yearly wages is larger for college graduates than for non-college graduates. Therefore our estimates under the current assumptions will be inefficient and the associated standard errors will be inaccurate. However, our estimates will still be unbiased, and our variance assumptions are reasonable from a design perspective. If

Table 1. The Ecological Data:

Area	1	2	3	4	5	6	7	8	9	10	11
n_i	4900	4273	3181	2855	5234	4188	4544	5963	4180	6433	4032
\bar{y}_i	41.3	36.3	39.8	39.7	46.8	40.0	45.7	54.6	49.7	42.2	44.3
\bar{x}_i	0.255	0.222	0.291	0.232	0.266	0.229	0.538	0.476	0.308	0.224	0.223

**Fig. 3.** Ecological Regression vs. Within Group Regressions

we initially observe only the ecological data, we cannot estimate separate variances, and therefore without other information, our subsample design must be based on some variance assumption. The constant variance assumption seems reasonable in this context.

The ecological data were constructed by aggregating the individual level data up to the super-PUMA level. Table 1 shows the ecological data and the within-group sample sizes for all eleven areas in WA state. The average yearly wage (in thousands) for area i (\bar{y}_i) is the ecological response, while the proportion of college degrees in area i (\bar{x}_i) is the ecological covariate.

In Figure 3, we see the effects of aggregating the data. The circles represent the ecological data from Table 1, and the dashed line is the ecological regression line. The solid lines represent the within group regression lines for each of the super-PUMAs. The dotted line represents the weighted average of the solid lines. Table 2 shows that the ecological slope (32.2) is biased upward compared to the weighted average within group slope (18.8), a difference of 13.4. In fact, the ecological regression is so biased that the ecological slope estimate is larger than the maximum of the within group slopes (27.0). Therefore, inference based solely on the ecological data would lead us to greatly overestimate the value of a college degree. Since we have the individual level data, we can determine that this ecological bias is due to both intercept bias and slope bias. Using the estimated parameters from the full data as the true parameter values, 7.0/13.4 of the bias comes from the intercepts (9) and 6.4/13.4 of the bias comes from the slopes (10). Additionally, the slope bias

Table 2. Ecological Bias. Intercept and slope bias defined in (9),(10)

Data	Slope Estimate	Standard Error	Bias	Intercept Bias	Slope Bias
Full Data	18.8	0.235	0	NA	NA
Ecological Data	32.2	11.90	13.4	7.0	4.7 1.7

can be broken into its linear component (4.7/13.4) and its quadratic component (1.7/13.4). Of course, we have ignored the possibility of confounders in this analysis, and therefore we cannot assume that the slope of the dotted line in Figure 3 is an unbiased estimate of the true college degree effect. For example, if a person's race has an effect on the likelihood of obtaining a college degree, and if race also has an effect on a person's wages, then the slopes in Figure 3 will not represent the true effect of a college degree because they will capture a race effect as well as the college degree effect. We will postpone the discussion of confounding within this application until Section 7, where we re-analyze the data.

4. Estimation with Ecological and Subsample Data

To perform estimation with combined ecological and subsample data, we assume the following:

$$\begin{aligned}
 y_{ij} &= E[y_{ij}|x_{ij}, z_{ij}] + \epsilon_{ij}, \\
 E[y_{ij}|x_{ij}, z_{ij}] &= \beta_{0i} + \beta_{wi}x_{ij} + \beta_{ci}z_{ij}, \\
 \epsilon_{ij}|x_{ij}, z_{ij} &\sim_{i.i.d.} N(0, \sigma_e^2),
 \end{aligned} \tag{17}$$

where the ecological model is given by,

$$\begin{aligned}
 \bar{y}_i &= E[\bar{y}_i|\bar{x}_i, \bar{z}_i] + \bar{\epsilon}_i, \\
 E[\bar{y}_i|\bar{x}_i, \bar{z}_i] &= \beta_{0i} + \beta_{wi}\bar{x}_i + \beta_{ci}\bar{z}_i, \\
 \bar{\epsilon}_i|\bar{x}_i, \bar{z}_i &\sim_{ind} N\left(0, \frac{\sigma_e^2}{n_i}\right).
 \end{aligned}$$

Suppose that we have a subsample of the individual level data (y_{ij}, x_{ij}, z_{ij}) for individuals $j = 1, \dots, k_i$ in groups $i = 1, \dots, m$, where $k_i < n_i$, and n_i represents the total number of individuals in group i . We will denote this subsample data $(\mathbf{y}_i^s, \mathbf{x}_i^s, \mathbf{z}_i^s)$. Without loss of information, these data can be transformed into $(y_{ij} - \bar{y}_i, x_{ij} - \bar{x}_i, z_{ij} - \bar{z}_i)$ and $(\bar{y}_i, \bar{x}_i, \bar{z}_i)$ for $j = 1, \dots, k_i$ and $i = 1, \dots, m$. Notice that the centering here is done around the ecological means and not the sub-sample means, which we denote $(\bar{y}_i^s, \bar{x}_i^s, \bar{z}_i^s)$ for $j = 1, \dots, k_i$ and $i = 1, \dots, m$. The model for the combined ecological and subsample data within each group can be written as:

$$\left(\left[\begin{array}{c} (\mathbf{y}_i^s - \bar{\mathbf{y}}_i) \\ \bar{\mathbf{y}}_i \end{array} \right] \middle| \begin{array}{c} (\mathbf{x}_i^s - \bar{\mathbf{x}}_i), (\mathbf{z}_i^s - \bar{\mathbf{z}}_i), \bar{\mathbf{x}}_i, \bar{\mathbf{z}}_i \\ (\beta_{0i}, \beta_{wi}, \beta_{ci}) \end{array} \right) \sim_{ind} N_{k_i+1} \left(\boldsymbol{\mu}_i, \left[\begin{array}{cc} \Sigma_{11i} & \Sigma_{12i} \\ \Sigma_{21i} & \Sigma_{22i} \end{array} \right] \right) \tag{18}$$

for $i = 1, \dots, m$ where

$$\begin{aligned}\mu_i &= \begin{bmatrix} \beta_{wi}(x_{i1} - \bar{x}_i) + \beta_{ci}(z_{i1} - \bar{z}_i) \\ \vdots \\ \beta_{wi}(x_{ik_i} - \bar{x}_i) + \beta_{ci}(z_{ik_i} - \bar{z}_i) \\ \beta_{0i} + \beta_{wi}\bar{x}_i + \beta_{ci}\bar{z}_i \end{bmatrix} \\ \Sigma_{11i} &= \sigma_e^2 \left(I_{k_i} - \frac{1}{n_i} J_{k_i} \right) \\ \Sigma_{12i} &= \mathbf{0}_{k_i} \\ \Sigma_{21i} &= \mathbf{0}_{k_i}^T \\ \Sigma_{22i} &= \frac{\sigma_e^2}{n_i}\end{aligned}$$

and I_{k_i} is an identity matrix of size k_i , J_{k_i} is a $k_i \times k_i$ matrix of ones, and $\mathbf{0}_{k_i}$ is a $k_i \times 1$ vector of zeros. We will refer to the first k_i equations in (18) as the centered model, and the last equation as the ecological model. This combined data model represents the basis for a likelihood estimation approach, and we emphasize that the centered data are independent of the ecological data.

The combined estimator for the β_{wi} and β_{ci} parameters depends only on the centered data, and has the standard GLS form:

$$\begin{aligned}\begin{bmatrix} \hat{\beta}_{wi}^{comb} \\ \hat{\beta}_{ci}^{comb} \end{bmatrix} &= \left(\mathbf{X}^{*T} \Sigma_{11i}^{-1} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*T} \Sigma_{11i}^{-1} \mathbf{y}^* \\ \mathbf{X}^* &= \begin{bmatrix} (x_i^s - \bar{x}_i) & (z_i^s - \bar{z}_i) \end{bmatrix} \\ \mathbf{y}^* &= \begin{bmatrix} (y_i^s - \bar{y}_i) \end{bmatrix}\end{aligned}\tag{19}$$

We note three important properties of this estimator. First, it is clearly unbiased for the group specific parameters, and therefore an unbiased estimator for $\bar{\beta}_w$ can be formed by combining the group specific estimates. Second, σ_e^2 cancels from this equation, and therefore we do not need to estimate this variance parameter in order to obtain an estimate for β_{wi} . An estimate of σ_e^2 would be necessary for interval estimates about $\bar{\beta}_w$ (here we focus on point estimation). Third, this combined estimator corresponds to the MLE.

5. Information Comparisons for the Subsample and Combined Data

In the previous section, we showed that the combined estimator (19) is unbiased for the slope parameters. However, an MLE based solely on the subsample data will also be unbiased, so we may wonder how much advantage we gain by adding the ecological data to the subsample data. As usual, adding any data can only increase our information about the parameters. However, adding the ecological data is not the same

as adding one additional observation, so we may question the nature of this information gain. Specifically, we want to know how much additional information the ecological data provide for the slope parameters. In this section, we show that the extra information provided by the ecological data about the slope parameters can be small under simple random subsampling. This result motivates the need for optimal design, which we discuss in Section 6.

Let E_i denote the ecological data and S_i the subsample data for group i . The Fisher information from S_i will be written as $I_{S_i}(\beta_{0i}, \beta_{wi}, \beta_{ci})$, and that for the combined data, $\{S_i, E_i\}$, as $I_{S_i, E_i}(\beta_{0i}, \beta_{wi}, \beta_{ci})$. In many cases, we will need to discuss the information for a single parameter treating the others as nuisance parameters. For example, if we had two parameters θ_1 and θ_2 with the information matrix,

$$I(\theta_1, \theta_2) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

then the information about θ_1 taking into account the uncertainty about θ_2 can be written:

$$I(\theta_1) = I_{11} - I_{12}I_{22}^{-1}I_{21}.$$

In the context of ecological and subsample data, we will often discuss the information about β_{wi} in the combined data, while taking into account the information lost due to uncertainty about β_{0i} and β_{ci} . We will write this as $I_{S_i, E_i}(\beta_{wi})$.

5.1. Information in the Varying Intercepts and Slopes Model

If we assume the varying intercepts and slopes model (2), then we only have two parameters per group, β_{0i} and β_{wi} . The information in the subsample is given by

$$I_{S_i}(\beta_{0i}, \beta_{wi}) = \frac{1}{\sigma_e^2} \begin{bmatrix} k_i & \sum_{j=1}^{k_i} x_{ij} \\ \sum_{j=1}^{k_i} x_{ij} & \sum_{j=1}^{k_i} x_{ij}^2 \end{bmatrix}. \quad (20)$$

We define $s_{x_i}^2$ to be the sample variance of x for the subsample in group i , $a_i = \bar{x}_i^e - \bar{x}_i$ to be the difference between the ecological covariate mean and the subsample covariate mean, $c_i = \frac{1}{1 - k_i/n_i}$ as the reciprocal of one minus the sampling fraction. Then the information from the combined data can be written as

$$I_{S_i, E_i}(\beta_{0i}, \beta_{wi}) = \frac{1}{\sigma_e^2} \begin{bmatrix} 0 & 0 \\ 0 & k_i(s_{x_i}^2 + c_i a_i^2) \end{bmatrix} + \frac{n_i}{\sigma_e^2} \begin{bmatrix} 1 & \bar{x}_i \\ \bar{x}_i & \bar{x}_i^2 \end{bmatrix}. \quad (21)$$

The first term of (21) corresponds to the information in the first k_i elements of (18), (the $y_{ij} - \bar{y}_i$ terms), and we observe that the information about β_{0i} in this term is zero and there is only information about β_{wi} . Also notice that this term is undefined if $n_i = k_i$. The second term of (21) corresponds to the information in the ecological data. We can add these information matrices together, because the centered values are independent of the ecological values as shown in (18). See Appendix B for a derivation.

Utilizing the result from the beginning of this section, the information about the intercepts from the combined data,

$$I_{S_i, E_i}(\beta_{0i}) = \frac{1}{\sigma_e^2} \left(n_i - \frac{\bar{x}_i^2}{k_i(s_{x_i}^2 + c_i a_i^2) + \bar{x}_i^2} \right) \quad (22)$$

will usually be much greater than the information from the subsample data:

$$I_{S_i}(\beta_{0i}) = \frac{1}{\sigma_e^2} \left(k_i - \frac{\left(\sum_{j=1}^{k_i} x_{ij} \right)^2}{\sum_{j=1}^{k_i} x_{ij}^2} \right). \quad (23)$$

The second term of (22) will be positive and less than one and the second term of (23) will be positive, so $I_{S_i, E_i}(\beta_{0i}) > I_{S_i}(\beta_{0i})$, when $n_i > k_i + 1$. Since we usually expect the ecological data to be aggregated from a large number of individuals, n_i will often be much larger than k_i and the information gain will be significant.

The magnitude of the increase in information about β_{wi} is not as transparent as it is for β_{0i} . The information in the subsample for β_{wi} is given by

$$I_{S_i}(\beta_{wi}) = \frac{1}{\sigma_e^2} k_i s_{x_i}^2. \quad (24)$$

The information in the combined data is given by

$$\begin{aligned} I_{S_i, E_i}(\beta_{wi}) &= \frac{1}{\sigma_e^2} \left(k_i(s_{x_i}^2 + c_i a_i^2) + n_i \bar{x}_i^2 - \frac{n_i^2 \bar{x}_i^2}{n_i} \right) \\ &= \frac{1}{\sigma_e^2} k_i (s_{x_i}^2 + c_i a_i^2), \end{aligned} \quad (25)$$

showing that the information gain hinges on a_i , the difference between the ecological and subsample covariate averages, and on c_i , a function of the sampling fraction. Since the increase in information about β_{wi} depends on a_i^2 , the increase can be quite small under subsampling schemes which produce $\bar{x}_i^* \approx \bar{x}_i$. In particular, if we view the ecological data as fixed, and we subsample from the n_i observations within group i , the expectation of $c_i a_i^2$ under this simple random subsample (SRS) is $\frac{1}{k_i} \sigma_{x_i}^2$ where $\sigma_{x_i}^2$ is the finite population variance (see Cochran (1977)). Therefore, the relative increase in information under SRS diminishes as k_i increases.

Notice also that (25) is identical to the lower right hand element of the first term in (21). In this sense, the ecological data only has information about the slope parameter through its inclusion in the first k_i elements of (18), or the $y_{ij} - \bar{y}_i$ terms. If we are only interested in the slope parameters, we can make inference based solely on these k_i data differences. This is a well known technique in the econometrics literature (Chamberlain (1984)), which we will adopt for the rest of this paper, effectively ignoring β_{0i} .

In some cases we may gain more information about β_{wi} from the ecological data if we model the β_{0i} terms. However, there are cases where modeling the β_{0i} terms will not help in the estimation of β_{wi} . For example, in the contextual effects model of Section 2.1, $\beta_{0i} = \gamma_0 + (\gamma_b - \gamma_w) \bar{x}_i$. Therefore, $\bar{y}_i = \gamma_0 + \gamma_b \bar{x}_i + \bar{\epsilon}_i$, and again, the ecological data provide no information about β_w outside of the difference data.

5.2. Information in the Within-Group Confounding Model

In model (17), we need only estimate β_{wi} and β_{ci} since we can ignore β_{0i} if we use the centered data equations. Let $s_{z_i}^2$ be the sample variance of z for the subsample in group i , let $s_{x_i z_i}$ be the sample covariance of x and z , and $b_i = \bar{z}_i^g - \bar{z}_i$ be the differences between the ecological confounder means and the subsample confounder means. When z_i^g and \bar{z}_i are observed for each group and included in the estimation, then the information from the subsample and combined data can be written as

$$I_{S_i}(\beta_{wi}, \beta_{ci}) = \frac{1}{\sigma_e^2} \begin{bmatrix} k_i s_{x_i}^2 & k_i s_{x_i z_i} \\ k_i s_{x_i z_i} & k_i s_{z_i}^2 \end{bmatrix}, \quad (26)$$

$$I_{S_i, E_i}(\beta_{wi}, \beta_{ci}) = \frac{1}{\sigma_e^2} \begin{bmatrix} k_i (s_{x_i}^2 + c_i a_i^2) & k_i (s_{x_i z_i} + c_i a_i b_i) \\ k_i (s_{x_i z_i} + c_i a_i b_i) & k_i (s_{z_i}^2 + c_i b_i^2) \end{bmatrix}. \quad (27)$$

Hence the diagonal elements of (27) are at least as large as the diagonal elements of (26), but the effect of this increase on the information about β_{wi} can be diminished by the effects of the off diagonal terms.

Accounting for the estimation of β_{ci} gives

$$I_{S_i}(\beta_{wi}) = \frac{1}{\sigma_e^2} \left(k_i s_{x_i}^2 - \frac{(k_i s_{x_i z_i})^2}{k_i s_{z_i}^2} \right) \quad (28)$$

$$I_{S_i, E_i}(\beta_{wi}) = \frac{1}{\sigma_e^2} \left(k_i (s_{x_i}^2 + c_i a_i^2) - \frac{k_i^2 (s_{x_i z_i} + c_i a_i b_i)^2}{k_i (s_{z_i}^2 + c_i b_i^2)} \right) \quad (29)$$

The second terms of (28) and (29) correspond to the amount of information lost due to uncertainty about the β_{ci} parameter. For different data observations, the information lost may be greater for the subsample or the combined approach. If the subsample covariance between x and z is zero ($s_{x_i z_i} = 0$), then the combined approach will lose at least as much information as the subsample approach due to the nuisance parameter. However, if the subsample covariance is large in absolute value between x and z , then the combined approach may lose less information due to nuisance parameter estimation.

The gain in information about β_{wi} now depends on a_i^2 , b_i^2 , and c_i , and this gain can be small in subsampling schemes where $\bar{x}_i^g \approx \bar{x}_i$ and $\bar{z}_i^g \approx \bar{z}_i$. In particular, the expectation of (27) under SRS will approach the expectation of (26) under SRS as k_i increases (see Cochran (1977)). We will show in Section 6, that the information gained through the utilization of the ecological data will greatly increase when we use the ecological data in the sampling design.

6. Optimal Subsampling Design conditional on the Ecological Data

When the ecological data are known, subsampling design will depend on the distribution of the subsample data conditional on the ecological data. Since the k_i centered data equations of (18) are independent of the ecological data, the information about β_{wi} from the subsample conditional on the ecological data, $I_{S_i|E_i}(\beta_{wi})$, will equal the information about β_{wi} from the combined data, $I_{S_i, E_i}(\beta_{wi})$. Therefore, we can use the information equations of the previous section to inform our subsampling procedure.

6.1. Varying Intercepts and Slopes

Within each group, the information about β_{wi} in model (2), conditional on the ecological data, is given by

$$I_{S_i|E_i}(\beta_{wi}) = I_{S_i,E_i}(\beta_{wi}) = \frac{k_i}{\sigma_e^2}(s_{x_i}^2 + c_i a_i^2). \quad (30)$$

Recall that $a_i = \bar{x}_i^s - \bar{x}_i$ and $c_i = \frac{1}{1 - k_i/n_i}$. Since the ecological data are observed, we can use (30) to design a subsampling scheme which maximizes information. The ecological covariate averages (\bar{x}_i) cannot be changed by our subsampling procedure, but we can increase the information by picking a subsample that will maximize a_i^2 , and/or maximize s_{x_i} , the variability of the subsample.

In our college degree/wage example, x_{ij} is a binary college degree indicator, and therefore (30) simplifies to $\frac{k_i}{\sigma_e^2}(\bar{x}_i^s(1 - \bar{x}_i^s) + c_i a_i^2)$. When $n_i > k_i > 0$, this expression is a convex function of \bar{x}_i^s , and will be maximized when we sample all ones (college degree) or all zeros (no college degree) within each group. Therefore, the maximum information available from the combined approach using optimal design is $\frac{k_i}{\sigma_e^2} \times \max\{c_i \bar{x}_i^s, c_i(1 - \bar{x}_i^s)\}$. The maximum information available in the subsample alone using an optimal design is only $\frac{k_i}{4\sigma_e^2}$, which is achieved when $\bar{x}_i^s = \frac{1}{2}$.

We know from Table 1 that \bar{x}_i , the percentage of college graduates in group i is less than 50% for all groups except group seven. Therefore, to maximize information we should only sample people without college degrees from group seven and only sample people with college degrees from all other groups. Such a sampling scheme has a familiar interpretation in that we will maximize information by sampling rare events (e.g. case based sampling). Of course, we would always want to sample some individuals with and without college degrees in each area for the purpose of model checking. However, even under this more robust sampling scheme, (30) will still be useful, because it describes the information lost when sampling “non-optimal” individuals.

Additionally, if we assume (1), then $\beta_{wi} = \beta_w$ for all $i = 1, \dots, m$, and we may want to know which group provides the most information about β_w . For example, we may only have the time and money to sample individuals from one group (super-PUMA). Again, (30) provides a basis for answering this question. In general, we can maximize information by selecting a group with an extreme \bar{x}_i and a large sampling fraction (k_i/n_i). Intuitively, we are rewarded for sampling rare events, and we should select the group which contains individuals who are rare in comparison to the rest of the group. In our example, if we had resources to subsample $k_i = 50$ observations from a single group, we would select group two because it has the smallest college degree proportion of 0.222 and a relatively small $n_i = 4273$. Therefore, if we sampled from this group, we would sample people with college degrees from an area that doesn’t have many people with college degrees. Of course, we would always want to sample some individuals from other groups, so we could check the model assumptions.

6.2. Within Group Confounding

In the within group confounding model, (17), the information can be written as (29). Recall that $s_{x_i}^2$ is the sample variance of x for the subsample in group i , $s_{z_i}^2$ is the sample variance of z for the subsample in group i , $s_{x_i z_i}$ is the sample covariance of x and z for the subsample in group i , and that $a_i = \bar{x}_i^s - \bar{x}_i$ and $b_i = \bar{z}_i^s - \bar{z}_i$. Then

$$I_{S_i|E_i}(\beta_{wi}) = I_{S_i, E_i}(\beta_{wi}) = \frac{1}{\sigma_e^2} \left(k_i(s_{x_i}^2 + c_i a_i^2) - \frac{k_i^2(s_{x_i z_i} + c_i a_i b_i)^2}{k_i(s_{z_i}^2 + c_i b_i^2)} \right) \quad (31)$$

and there is no easy rule for maximizing (31). The first term will be maximized as in the previous section, but minimization of the second term will require a case by case analysis.

In some cases, we can sample so as to make the second term go away entirely, hence we will lose no information due to uncertainty about β_{ci} . In our college degree/wage example, suppose we believe that race (white vs. non-white) is a confounder. As discussed before, we can maximize the first term of (31) with our college degree sampling scheme. Since x_{ij} is constant for the subsample within each group, the subsample covariance between x_{ij} and z_{ij} in each group is zero ($s_{x_i z_i} = 0$). Therefore, we need only force $b_i = 0$ in order to cancel the second term in (31). To achieve this cancellation in our example, we need to sample college graduates in racial proportions that match the population racial proportions. Whites will tend to be overrepresented in the population of college graduates, and we can maximize information by reducing the number of whites in our sample to match the proportion of whites overall.

7. Application: Subsampling to Estimate the Wage Value of a College Degree

Until now, we have argued for the superiority of the combined data approach over the subsample approach by showing that the information from the former will be greater than the information from the latter. When estimating the intercepts, the benefit of the combined approach was obvious, and (22) shows a significant information gain. However, when estimating the slope parameters, the increase in information can be negligible. In this section, we will study the benefits of the combined approach in the context of the PUMS data presented in Section 3 with yearly wages as the response, a college degree indicator as the covariate of interest, and a racial indicator (white/non-white) as a potential confounder.

We will show two main results. First, when the subsample is a simple random sample from the full data, the increased precision of the slope parameters derived from using the ecological data will decrease as the within-group subsample sizes, k_i , increase. Second, we show that optimal design in the combined approach will provide substantially increased precision about the slope parameters.

Table 3. Mean and standard deviation for the distribution of $|\hat{\beta}_w^{sub} - \hat{\beta}_w^{full}| - |\hat{\beta}_w^{comb} - \hat{\beta}_w^{full}|$ (“improvement” under the combined approach) based on 1000 simple random subsamples for three different within group subsample sizes in the varying intercepts model, varying intercepts and slopes model, and within group confounding model

	Varying Intercepts			Varying Intercepts and Slopes			Within Group Confounding		
	$k_i = 5$	$k_i = 10$	$k_i = 50$	$k_i = 30$	$k_i = 50$	$k_i = 100$	$k_i = 30$	$k_i = 50$	$k_i = 100$
Mean	.651	.210	.031	.077	.054	.010	.083	.061	.012
SD	3.290	1.667	.315	.665	.350	.156	.660	.358	.152

7.1. Simple Random Subsampling

If we believe model (1), then the causal parameter of interest is the common within group slope, β_w . We do not know the true value of this parameter, but we can calculate the full data MLE ($\hat{\beta}_w^{full} = 19.10$) which is likely to be accurate given the large sample size. Therefore, if (1) is the true model, a college degree is worth about \$19,100 a year to a randomly selected individual from the population.

We can compare the performance of the combined and subsample estimators by repeatedly subsampling from the full data to generate a pseudo sampling distribution based on k_i observations from each group. To simplify matters, we will choose k_i to be constant across groups for 1000 simulations (random subsamples from the full data). For each simulation/subsample, we will compare the combined estimator of β_w (the GLS estimate from the subsample data centered on the ecological means) to the subsample estimator for β_w (the OLS estimate from the subsample data) by comparing their absolute deviations from the full data MLE. Therefore, the 1000 subsamples will generate a pseudo sampling distribution for $|\hat{\beta}_w^{sub} - \hat{\beta}_w^{full}| - |\hat{\beta}_w^{comb} - \hat{\beta}_w^{full}|$, which we will call the “improvement” in the estimator. We will repeat this experiment with three different within-group subsample sizes: $k_i = 5, 10$, and 50 . Table 3 summarizes the simulation results. The average improvement from the subsampling distributions is positive for all models and sample sizes, but the average improvement gets closer to zero as k_i increases.

Figure 4(a) shows the subsampling distributions of “improvement” ($|\hat{\beta}_w^{sub} - \hat{\beta}_w^{full}| - |\hat{\beta}_w^{comb} - \hat{\beta}_w^{full}|$) for the three different within-group subsample sizes. We notice two things. First, the combined approach gives positive improvements for a majority of the subsamples, but the negative values in the boxplots show that for some subsamples, $\hat{\beta}_w^{sub}$ is preferable. Second, the improvement gets closer to zero as the within group sample sizes increase. We would expect this because the subsample group averages will get closer to the ecological group averages as the within group sample sizes increase. Additionally, this confirms the theoretical development of Section 5.1, where we showed that the information gained from using the ecological data decreases as k_i increases.

If we believe model (2), then the causal parameter of interest is the average within group slope, and the full data MLE is $\hat{\beta}_w^{full} = 18.84$. Notice that there is little change in the full data estimator when we allow

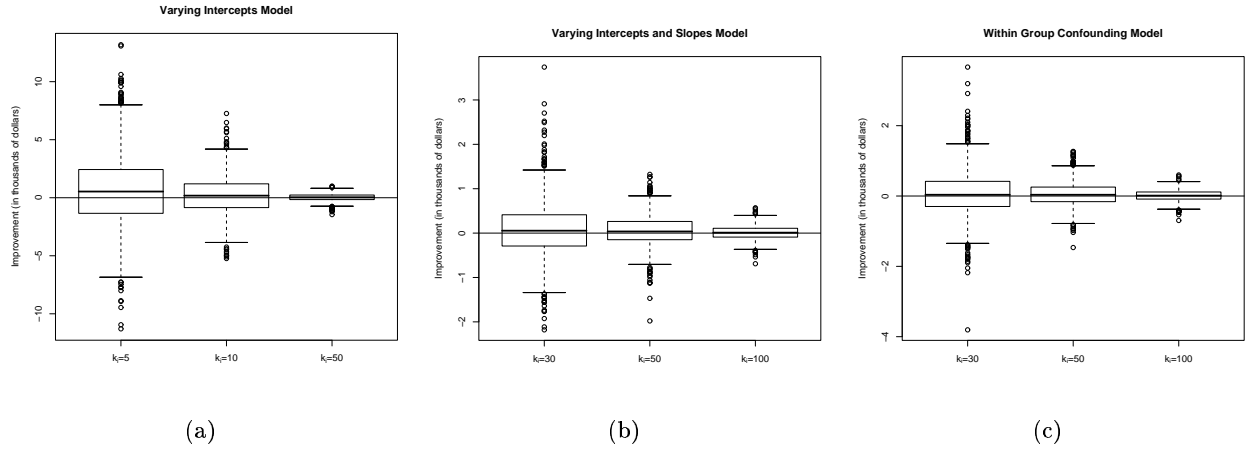


Fig. 4. Subsampling Distributions for $|\hat{\beta}_w^{sub} - \hat{\beta}_w^{full}| - |\hat{\beta}_w^{comb} - \hat{\beta}_w^{full}|$ (“improvement” under the combined approach) based on 1000 simple random subsamples for three different within group subsample sizes: (a) Varying intercepts model, (b) Varying intercepts and slopes model, (c) Within group confounding model

for different slopes. We can compare the performance of the combined and subsample estimators for (2) by repeatedly subsampling from the full data. However, in this model, we must separately estimate the m different within group slopes in order to calculate the average within group slope, and in order to ensure identification for the subsample approach (recall we have a binary covariate and we need observations in both groups), we need to sample larger within-group sample sizes: we choose $k_i = 30, 50,$ and 100 . Even with the larger subsample sizes, two of the 1000 simple random subsamples with $k_i = 30$ contained groups that had all non-college graduates. These two subsamples were removed from all analysis.

Figure 4(b) shows the pseudo sampling distribution of $|\hat{\beta}_w^{sub} - \hat{\beta}_w^{full}| - |\hat{\beta}_w^{comb} - \hat{\beta}_w^{full}|$ based on 1000 simulations (random subsamples from the full data) for the three different within-group subsample sizes. For this model, $\hat{\beta}_w^{sub}$ is a weighted average of OLS estimates from the subsamples within each group, and $\hat{\beta}_w^{comb}$ is a weighted average of the GLS estimates from the subsample data centered on the ecological means within each group. We see that the combined approach provides improvement over the subsample approach for a majority of the 1000 subsamples, but the median is now closer to zero for all three subsample sizes. However, when the within group sample sizes are 30, the positive outliers are more plentiful and extreme than the negative outliers. Therefore, the combined approach is less likely to produce a really bad result than the subsample only approach.

If we now add the white/non-white confounder to the model (17), then the causal parameter of interest is the average within group slope after adjusting for the effect of the confounder, and the full data MLE is $\hat{\beta}_w^{full} = 18.36$. We should notice that the full data estimator is slightly smaller for this model. Therefore,

if (3) is the true model, a college degree is worth about \$18,360 a year to a randomly selected individual from the population. We again compare the performance of the combined and subsample estimators by repeatedly subsampling from the full data with $k_i = 30, 50, \text{ and } 100$. From Figure 4(c), we see that the combined approach again provides insurance against “unlucky” subsamples.

7.2. *Optimal Subsampling Design*

In this section, we will investigate the benefit to be gained from subsampling design conditional on the ecological data. As discussed in Section 5, the information about $\bar{\beta}_w$ doesn’t depend on the ecological response data, and hence we only need to consider the ecological data for the covariate and the confounder. Additionally, the binary covariate and confounder in our example allow a simple solution to the problem of optimal design. In the following, we show that for the three models, optimal design in the combined approach will provide substantially increased precision about the slope parameters.

7.2.1. *Design in the Varying Intercepts Model*

We showed in Section 5 that we can maximize our information about the within group slopes by carefully subsampling based on covariate values. In the context of our application, the covariate is binary, and the percentage of individuals with college degrees is less than 50% in all groups but group seven (see Table 2). Therefore, we can maximize information by sampling only college graduates within most groups and non-college graduates in group seven.

In order to compare the combined estimator under optimal design to the combined and subsample estimators under simple random sampling, we generated 1000 simple random subsamples (SRS) and 1000 college random samples (CRS, i.e. random subsamples of non-college graduates from group seven, and college graduates for all other groups). To simplify things, we sampled equal numbers from within each group, and the process was repeated for three different within-group sample sizes: $k_i = 5, 10, 50$. We then used these subsamples to create three sampling distributions: $\hat{\beta}_w^{sub}$ under SRS, $\hat{\beta}_w^{comb}$ under SRS, $\hat{\beta}_w^{comb}$ under CRS.

Figure 5(a) shows the comparison between these sampling distributions. The solid line represents the full data MLE ($\hat{\beta}_w^{full}$), and the dashed line represents the ecological regression estimator. When the within group samples are small ($k_i = 5$), $\hat{\beta}_w^{comb}$ under CRS has more precision than $\hat{\beta}_w^{comb}$ under SRS, which has more precision than $\hat{\beta}_w^{sub}$ under SRS. However, all three approaches occasionally produce “bad” estimates for sample sizes this small. $\hat{\beta}_w^{sub}$ under SRS and $\hat{\beta}_w^{comb}$ under SRS produce negative estimates in these sampling distributions, and all three approaches can produce estimates that have more error than the ecological estimate. When $k_i = 10$, $\hat{\beta}_w^{comb}$ under SRS is only slightly more precise than $\hat{\beta}_w^{sub}$ under SRS, but $\hat{\beta}_w^{comb}$ under CRS maintains an advantage in precision. Additionally, the two estimators under SRS can still

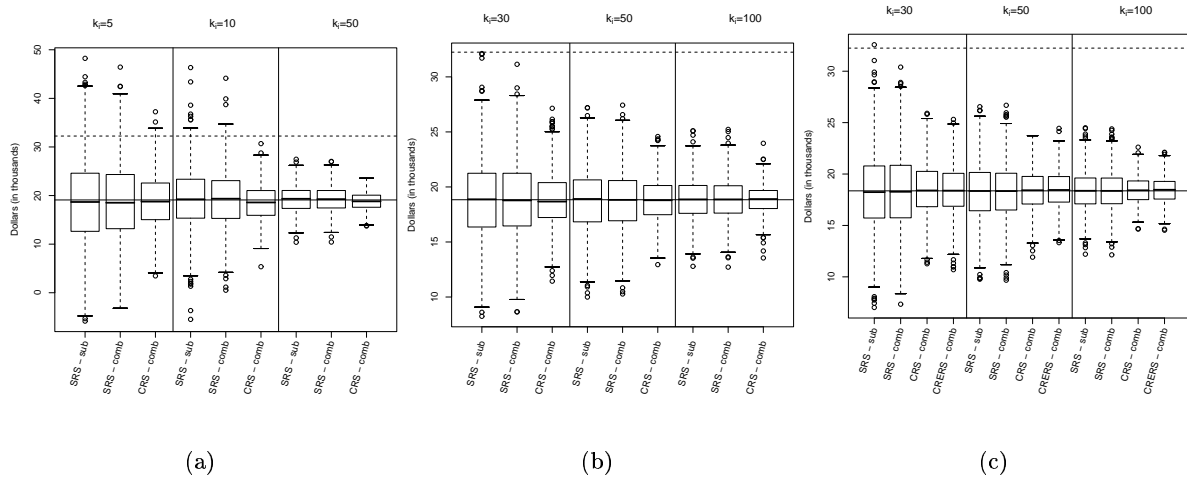


Fig. 5. Subsampling distributions for $\hat{\beta}_w$ in the varying intercepts model and $\hat{\beta}_w$ in the varying intercepts and slopes model and the within group confounding model. The first two models employ three estimation approaches: subsample data based on simple random subsamples (SRS-sub), combined data based on simple random subsamples (SRS-comb), combined data based on college random subsamples (CRS-comb). The within group confounding model also employs an estimation approach based on combined data with college random subsamples and ecological racial proportions (CRERS-comb). The solid horizontal line represents the full data MLE, and the dashed line represents the ecological estimate: (a) Varying intercepts model, (b) Varying intercepts and slopes model, (c) Within group confounding model

Table 4. Variance ratios for $\hat{\beta}_w$ in the varying intercepts model, and $\hat{\beta}_w$ in the varying intercepts and slopes model and the within group confounding model based on subsampling distributions for different estimation approaches: subsample data based on simple random subsamples (SRS-sub), combined data based on simple random subsamples (SRS-comb), combined data based on college random subsamples (CRS-comb), and college random subsamples with ecological racial proportions (CRERS-comb) for the within group confounding model.

Data	Varying Intercepts			Varying Intercepts and Slopes			Within Group Confounding		
	$k_i = 5$	$k_i = 10$	$k_i = 50$	$k_i = 30$	$k_i = 50$	$k_i = 100$	$k_i = 30$	$k_i = 50$	$k_i = 100$
SRS - Subsample	1	1	1	1	1	1	1	1	1
SRS - Combined	.831	.897	.973	.929	.951	.992	.940	.951	.991
CRS - Combined	.357	.352	.458	.451	.478	.444	.450	.507	.446
CRERS - Combined	NA	NA	NA	NA	NA	NA	.414	.397	.439

produce estimates worse than the ecological regression estimate, while $\hat{\beta}_w^{comb}$ under CRS is virtually assured of doing better. When $k_i = 50$, the sampling distributions for $\hat{\beta}_w^{sub}$ under SRS and $\hat{\beta}_w^{comb}$ under SRS are virtually identical, while the sampling distribution for $\hat{\beta}_w^{comb}$ under CRS preserves efficiency.

Table 4 reinforces the importance of optimal design in the varying intercepts model. The combined estimators have smaller variance than the subsample estimator, but under SRS, this advantage dissipates as the within group sample size increases. Under CRS, the combined estimator seems to maintain its advantage over the SRS subsample estimator.

7.2.2. Design in the Varying Intercepts and Slopes Model

In 7.2.1, we subsampled equal numbers of individuals from each group. Therefore, the optimal design for the varying intercepts and slopes model will be the same as the design from the previous section. We should sample only college graduates within most groups and non-college graduates in group seven.

In order to compare the combined estimator under optimal design to the combined and subsample estimators under simple random sampling, we generated 1000 simple random subsamples (SRS) and 1000 college random samples (CRS, i.e. random subsamples of non-college graduates from group seven, and college graduates for all other groups). In this model, we must separately estimate the m different within group slopes in order to calculate the average within group slope, so in order to ensure identification for the subsample approach (recall we have a binary covariate and we need observations in both groups), we sampled larger within-group sample sizes: $k_i = 30, 50$, and 100 . We then used these subsamples to create three sampling distributions: $\hat{\beta}_w^{sub}$ under SRS, $\hat{\beta}_w^{comb}$ under SRS, $\hat{\beta}_w^{comb}$ under CRS.

Figure 5(b) shows the comparison between these sampling distributions, for within group sample sizes of 30, 50, and 100. The solid line represents the full data MLE ($\hat{\beta}_w^{full}$), and the dashed line represents the

ecological regression estimator. Under all three subsample sizes, $\hat{\beta}_w^{comb}$ under CRS has more precision than $\hat{\beta}_w^{comb}$ under SRS, which has more precision than $\hat{\beta}_w^{sub}$ under SRS. However, because the smallest within group samples are relatively large ($k_i = 30$), only $\hat{\beta}_w^{sub}$ under SRS produces estimates that have more error than the ecological estimate. When $k_i = 50$ or $k_i = 100$, the sampling distributions for $\hat{\beta}_w^{sub}$ under SRS and $\hat{\beta}_w^{comb}$ under SRS are virtually identical, while the sampling distribution for $\hat{\beta}_w^{comb}$ under CRS preserves efficiency.

Table 4 reinforces the impressions from Figure 5(b). The combined estimators have smaller variance than the subsample estimator, but under SRS, this advantage dissipates as the within group sample size increases. Under CRS, the combined estimator seems to maintain its advantage over the SRS subsample estimator.

7.2.3. Design in the Within Group Confounding Model

In Section 5, we showed that an optimal subsampling design can be derived in the within group confounding model for binary covariates and confounders. In this application, we can maximize our information about $\hat{\beta}_w$ when using the combined approach by utilizing the college sampling scheme, and by sampling these college graduates so that the racial proportions in the sample match the racial proportions in the ecological data. When fitting a model with a confounder, it is important to use the ecological racial proportions in the college sampling scheme. When subsampling college graduates within each group, you occasionally get a sample with only white individuals. The combined estimator that controls for confounding will not be identified by a subsample of all white individuals with college degrees. Therefore, in our CRS subsampling scheme, we discarded college random samples with only white individuals. In Figure 5(c), we present sampling distributions based on three types of subsampling: simple random sampling (SRS), college random sampling (CRS), and college random sampling with ecological racial proportions (CRERS).

Table 4 shows that even with the introduction of the confounder, the CRS and CRERS combined estimators have greater precision than the SRS estimators (the precision of the CRS estimator is overstated due to the discarded samples). And again, this improvement is apparent as the subsample size gets larger. Additionally, the CRERS-comb estimator seems to perform better than the CRS-comb estimator, although the improvement in precision can be small.

8. Discussion

In this paper, we have discussed linear ecological bias, and have provided an approach to combining ecological and subsample data in order to correct this bias. We have also shown that while the increase in precision from the combined approach over a subsample approach can be small under simple random subsampling,

conditioning on the ecological data allows us to maximize information through optimal subsampling design. This result should inform future studies where ecological data are already available and individual subsample data are expensive to collect.

Our choice of assumptions throughout this paper has been guided by the problem of subsample design given ecological data, and three particular assumptions merit further discussion. First, we have assumed constant variances across groups and within each group. The constant variance assumptions seem reasonable in the design framework, and Section 7 shows that the design results of this paper can yield an improvement in precision, even when the model doesn't fit the data perfectly. Second, we have assumed that a stratified sample is possible on the covariate and the confounder. This will be more or less true depending on the application, but even an approximate sampling frame for the covariate and the confounder can be used to improve information. Third, we have assumed that the subsample has no missing data and that the subsample frame matches the sampling frame for the ecological data. The ecological data becomes quite useful if either of these assumptions does not hold. We can often use the ecological data to inform the correction of non-response in the subsample, and we can test for sampling frame bias by comparing the results from the subsample and combined data approaches to see if the differences are reasonable given the theoretical variability.

It is natural to extend the results of this paper to generalized linear models (GLMs). Ecological bias is often a larger problem in non-linear models than in linear models (Greenland (1992)), and therefore, a combined data approach would be beneficial. Additionally, many of the applications in which researchers resort to ecological inference have response variables which are discrete at the individual level and hence are ill-suited to the linear model. Unfortunately, derivation of the information from the combined data will be much more difficult in non-linear models, and it may not be possible to write down simple analytical formulas which will be interpretable in the same manner as (30) and (31). Therefore, answering the optimal design question will be more difficult in the GLM framework.

9. Acknowledgments

The work of the first and third authors was funded in part by the National Institute of Child Health and Human Development grant R01-HD043472-01. The work of the second author was supported by grant R01 CA095994 from the National Institutes of Health. The fourth author acknowledges support by NSF grant DMS 0505865. The authors are grateful to the editor, associate editor, and two reviewers whose helpful comments and suggestions greatly improved both the presentation and the content of this paper. The authors are also grateful to Ryan Admiraal and Anton Westveld for their helpful comments and discussion.

A. Appendix: Decomposition of Bias

$$\begin{aligned}
 E[\widehat{\beta}_w^{eco} | \bar{\mathbf{x}}] &= \frac{\sum_{i=1}^m \{n_i(E[\bar{y}_i - \bar{y} | \bar{\mathbf{x}}])(\bar{x}_i - \bar{x})\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &= \frac{\sum_{i=1}^m [n_i(\bar{x}_i - \bar{x}) \{\beta_{0i} - \bar{\beta}_0 + \beta_{wi}\bar{x}_i - \frac{1}{n} \sum_{k=1}^m (n_k \beta_{wk} \bar{x}_k) + E[\bar{z}_i - \bar{z} | \bar{\mathbf{x}}]\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &= \frac{\sum_{i=1}^m \{n_i(\bar{x}_i - \bar{x})(\beta_{0i} - \bar{\beta}_0)\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &\quad + \frac{\sum_{i=1}^m [n_i(\bar{x}_i - \bar{x}) \{\beta_{wi}\bar{x}_i - \frac{1}{n} \sum_{k=1}^m (n_k \beta_{wk} \bar{x}_k)\}]}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &\quad + \frac{\sum_{i=1}^m \{n_i(\bar{x}_i - \bar{x})(E[\bar{z}_i - \bar{z} | \bar{\mathbf{x}}])\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &= \bar{\beta}_w \\
 &\quad + \frac{\sum_{i=1}^m \{n_i(\bar{x}_i - \bar{x})(\beta_{0i} - \bar{\beta}_0)\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &\quad + \frac{\sum_{i=1}^m [n_i(\bar{x}_i - \bar{x}) \{(\beta_{wi} - \bar{\beta}_w)\bar{x}_i - \frac{1}{n} \sum_{k=1}^m (n_k (\beta_{wk} - \bar{\beta}_w) \bar{x}_k)\}]}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &\quad + \frac{\sum_{i=1}^m \{n_i(\bar{x}_i - \bar{x})(E[\bar{z}_i - \bar{z} | \bar{\mathbf{x}}])\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &= \bar{\beta}_w \\
 &\quad + \frac{\sum_{i=1}^m \{n_i(\bar{x}_i - \bar{x})(\beta_{0i} - \bar{\beta}_0)\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &\quad + \bar{x} \frac{\sum_{i=1}^m \{n_i(\bar{x}_i - \bar{x})(\beta_{wi} - \bar{\beta}_w)\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} + \frac{\sum_{i=1}^m \{n_i(\bar{x}_i - \bar{x})^2(\beta_{wi} - \bar{\beta}_w)\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2} \\
 &\quad + \frac{\sum_{i=1}^m \{n_i(\bar{x}_i - \bar{x})(E[\bar{z}_i - \bar{z} | \bar{\mathbf{x}}])\}}{\sum_{i=1}^m n_i(\bar{x}_i - \bar{x})^2}
 \end{aligned} \tag{32}$$

B. Appendix: Derivation of (21) and (25)

Since Σ_{11i} has the form $\sigma_e^2 (\mathbf{I}_{k_i} + \mathbf{b}\mathbf{J}_{k_i})$ where $b = \frac{-1}{n_i}$, its inverse will have the form $\frac{1}{\sigma_e^2} (\mathbf{I}_{k_i} - \frac{\mathbf{b}}{1+\mathbf{k}_i b} \mathbf{J}_{k_i})$, which simplifies to $\frac{1}{\sigma_e^2} (\mathbf{I}_{k_i} + \frac{1}{n_i - k_i} \mathbf{J}_{k_i})$. Therefore, we can derive (21) in the usual manner.

$$\begin{aligned}
 I_{S_i, E_i}(\beta_{0i}, \beta_{wi}) &= \begin{bmatrix} \mathbf{0}_{k_i} & (\mathbf{x}_i^s - \bar{\mathbf{x}}_i) \end{bmatrix}^T \Sigma_{11i}^{-1} \begin{bmatrix} \mathbf{0}_{k_i} & (\mathbf{x}_i^s - \bar{\mathbf{x}}_i) \end{bmatrix} + \begin{bmatrix} 1 & \bar{x}_i \end{bmatrix}^T \Sigma_{22i}^{-1} \begin{bmatrix} 1 & \bar{x}_i \end{bmatrix} \\
 &= \frac{1}{\sigma_e^2} \begin{bmatrix} 0 & 0 \\ 0 & [(\mathbf{x}_i^s - \bar{\mathbf{x}}_i)]^T \frac{1}{\sigma_e^2} (\mathbf{I}_{k_i} + \frac{1}{n_i - k_i} \mathbf{J}_{k_i}) [(\mathbf{x}_i^s - \bar{\mathbf{x}}_i)] \end{bmatrix} + \begin{bmatrix} 1 & \bar{x}_i \end{bmatrix}^T \frac{n_i}{\sigma_e^2} \begin{bmatrix} 1 & \bar{x}_i \end{bmatrix} \\
 &= \frac{1}{\sigma_e^2} \begin{bmatrix} 0 & 0 \\ 0 & \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_i)^2 + \frac{1}{n_i - k_i} \left(\sum_{j=1}^{k_i} (x_{ij} - \bar{x}_i) \right)^2 \end{bmatrix} + \frac{n_i}{\sigma_e^2} \begin{bmatrix} 1 & \bar{x}_i \\ \bar{x}_i & \bar{x}_i^2 \end{bmatrix} \\
 &= \frac{1}{\sigma_e^2} \begin{bmatrix} 0 & 0 \\ 0 & \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_i^s)^2 + k_i (\bar{x}_i^s - \bar{x}_i)^2 + \frac{k_i^2}{n_i - k_i} (\bar{x}_i^s - \bar{x}_i)^2 \end{bmatrix} + \frac{n_i}{\sigma_e^2} \begin{bmatrix} 1 & \bar{x}_i \\ \bar{x}_i & \bar{x}_i^2 \end{bmatrix} \\
 &= \frac{1}{\sigma_e^2} \begin{bmatrix} 0 & 0 \\ 0 & \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_i^s)^2 + \frac{n_i k_i}{n_i - k_i} (\bar{x}_i^s - \bar{x}_i)^2 \end{bmatrix} + \frac{n_i}{\sigma_e^2} \begin{bmatrix} 1 & \bar{x}_i \\ \bar{x}_i & \bar{x}_i^2 \end{bmatrix}
 \end{aligned}$$

If we let $s_{x_i}^2 = \frac{1}{k_i} \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_i^s)^2$, $a_i = \bar{x}_i^s - \bar{x}_i$ and $c_i = \frac{1}{1 - k_i/n_i}$, then we get (21).

References

- Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, vol. 3. Amsterdam: Elsevier.
- Card, D. (2001). Estimating the returns to schooling: Progress on some persistent econometric problems. *Econometrica* 69(5), 1127–1160.
- Chamberlain, G. (1984). Panel data. In Z. Griliches and M. Intriligator (Eds.), *Handbook of Econometrics*, Volume II, pp. Ch. 22. Elsevier B.V.
- Cochran, W. (1977). *Sampling Techniques*. New York: John Wiley & Sons.
- Freedman, D., S. Klein, M. Ostland, and M. Roberts (1998). Review of a solution to the ecological inference problem. *Journal of the American Statistical Association* 93, 1518–1522.
- Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine* 11, 1209–1223.
- Greenland, S. and H. Morgenstern (1989). Ecological bias, confounding, and effect modification. *International Journal of Epidemiology* 18 (1), 269–274.
- Haneuse, S. and J. Wakefield (2004). The combination of ecological and case-control data. *Submitted for publication*.
- King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton: Princeton.
- Raghunathan, T., P. Diehr, and A. Cheadle (2003). Combining aggregate and individual level data to estimate an individual level correlation model. *Journal of Educational and Behavioral Statistics* 28, 1–19.
- Richardson, S. (1992). Statistical methods for geographical correlation studies. In P. Elliott, J. Cuzick, D. English, and R. Stern (Eds.), *Analysis of Survey Data*, pp. 181–204. New York: Oxford University Press.
- Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* 15, 351–357.
- Ruggles, S., M. Sobek, T. Alexander, C. Fitch, R. Goeken, P. Hall, M. King, and C. Ronnander (2004). Integrated public use microdata series: Version 3.0 [machine-readable database].
- Steel, D., E. Beh, and R. Chambers (2004). The information in aggregate data. In G. King, O. Rosen, and M. Tanner (Eds.), *Ecological Inference: New Methodological Strategies*. Cambridge: Cambridge University Press.

Steel, D., M. Tranmer, and D. Holt (2003). Analysis combining survey and geographically aggregated data. In R. Chambers and C. Skinner (Eds.), *Analysis of Survey Data*. New York: Wiley.

Wakefield, J. (2004). Ecological inference for 2x2 tables (with discussion). *Journal of the Royal Statistical Society - A* 167, 385–445.