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 endeavour 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 Robinson (1950), 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. Within the econometric literature, ecological data are usually treated as population level information and included in a generalized method-of-moments approach with the ecological data providing extra moment conditions. Imbens and Lancaster (1994) and Hellerstein and Imbens (1999) are important papers in this area. Handcock *et al.* (2005) and Chaudhuri *et al.* (2005) adopted a similar approach in the likelihood framework, assuming that the ecological data provide population level information and utilizing this information as a constraint in the maximization of the likelihood. Other work has relaxed the assumption of population level information. In the case of 2×2 tables, Wakefield (2004) described the joint likelihood for the ecological data

Address for correspondence: Adam N. Glynn, Department of Government, Harvard University, CGIS Knafel Building, Cambridge, MA 02138, USA.
E-mail: aglynn@gov.harvard.edu

 0964–1998/08/171179

and subsample data and showed that a combined approach reduces ecological bias. Steel *et al.* (2004) developed the observed information for this same case, but with the data sources treated as independent. Jackson *et al.* (2006) also utilized this independence assumption to reduce bias in a combined model that allows for continuous and discrete covariates when logistic regression is appropriate at the individual level. Haneuse and Wakefield (2008) showed 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) showed 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) developed the properties of moment estimators in various 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 shall 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 that is 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, using as an example the measurement of 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 comparisons of information between the various sources of data and shows the magnitude of information that is gained by 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 eliminates the bias that can result from ecological regression, and that optimal subsampling design conditional on the ecological data provides an improvement over both simple random subsampling (SRS) and the purely ecological approach. 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 $(x_{ij}, y_{ij}, z^{\mathrm{c}}_{ij})$ for individuals $j = 1, \ldots, n_i$ in groups $i = 1, \ldots, m$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$ is the vector of responses from group i, $\mathbf{x}_i = (x_{i1}, \ldots, x_{in_i})$ is the vector of univariate exposure or covariates from group i, $\mathbf{z}^{\mathrm{c}}_i = (z^{\mathrm{c}}_{i1}, \ldots, z^{\mathrm{c}}_{in_i})$ is the vector of confounders, and $n = \Sigma_{i=1}^{m} n_i$ represents the 'full data' sample size. We shall assume that we observe ecological data that consist of the group means $(\bar{x}_i, \bar{y}_i)$ for groups $i = 1, \ldots, m$, and we may observe $\bar{z}^{\mathrm{c}}_i$ for groups $i = 1, \ldots, m$. Furthermore, we assume that the n_i observed triples in group i represent independently and identically distributed vectors that are produced by some process, and we are interested in the parameters of this process. Within this framework, we shall assume one of three models:

$$E[y_{ij}|\mathbf{x}_i, \mathbf{z}^{\mathrm{c}}_i] = \beta_{0i} + \beta_{\mathrm{w}} x_{ij}, \tag{1}$$

$$E[y_{ij}|\mathbf{x}_i, \mathbf{z}^{\mathrm{c}}_i] = \beta_{0i} + \beta_{\mathrm{w}i} x_{ij}, \tag{2}$$

$$E[y_{ij}|\mathbf{x}_i, \mathbf{z}^{\mathrm{c}}_i] = \beta_{0i} + \beta_{\mathrm{w}i} x_{ij} + z^{\mathrm{c}}_{ij}. \tag{3}$$

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

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

$$E[\bar{y}_i|\bar{x}_i] = \beta_{0i} + \beta_{\mathrm{w}}\bar{x}_i, \tag{4}$$

$$E[\bar{y}_i|\bar{x}_i] = \beta_{0i} + \beta_{\mathrm{w}i}\bar{x}_i, \tag{5}$$

$$E[\bar{y}_i|\bar{x}_i, \bar{z}^{\mathrm{c}}_i] = \beta_{0i} + \beta_{\mathrm{w}i}\bar{x}_i + \bar{z}^{\mathrm{c}}_i. \tag{6}$$

If we are interested in only the m observed groups, then we interpret the slopes $\boldsymbol{\beta}_{\mathrm{w}} = (\beta_{\mathrm{w}1}, \ldots, \beta_{\mathrm{w}m})$ as fixed 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 $\boldsymbol{\beta}_{\mathrm{w}}$ is fixed. However, an ecological regression that is based on $(\bar{x}_i, \bar{y}_i)$ for groups $i = 1, \ldots, m$ cannot identify all m of the slopes in $\boldsymbol{\beta}_{\mathrm{w}}$, 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

$$\bar{\beta}_{\mathrm{w}} \equiv \frac{1}{n}\sum_{i=1}^{m} n_i \beta_{\mathrm{w}i}.$$

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

$$\hat{\bar{\beta}}^{\mathrm{eco}}_{\mathrm{w}} = \frac{\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})(\bar{x}_i - \bar{x})}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \tag{7}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{m} n_i \bar{y}_i$$

and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} n_i \bar{x}_i.$$

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

$$\bar{z}^{\mathrm{c}} \equiv \frac{1}{n}\sum_{i=1}^{m} n_i \bar{z}_i^{\mathrm{c}},$$

we assume the most general model (3) and we restrict $\Sigma_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2 \neq 0$, then the expectation of $\hat{\bar{\beta}}_{\mathrm{w}}^{\mathrm{eco}}$ conditional on $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_m)$ can be written as (see Appendix A for details)

$$E[\hat{\bar{\beta}}_{\mathrm{w}}^{\mathrm{eco}} | \bar{\mathbf{x}}] = \bar{\beta}_{\mathrm{w}} \tag{8}$$

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

$$+ \frac{\bar{x}\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})(\beta_{\mathrm{w}i} - \bar{\beta}_{\mathrm{w}})}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} + \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2(\beta_{\mathrm{w}i} - \bar{\beta}_{\mathrm{w}})}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \tag{10}$$

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

The first term (8) is the parameter of interest, whereas the remaining terms represent a decomposition of ecological bias. The numerator of term (9) will be 0 when the weighted sample covariance between the ecological covariate averages and the group-specific intercepts is 0 and, therefore, this term represents the bias due to the intercepts. Fig. 1(a) shows an example of this bias. The numerator of the first term of expression (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. Fig. 1(b) shows an example of this bias. The numerator of the second term of expression (10) represents the bias due to a quadratic relationship between the group-specific slopes and the ecological covariate averages. Fig. 1 (c) shows an example of this bias. Both terms of expression (10) will be 0 if there is no relationship between the slopes and the ecological covariate averages ($\beta_{\mathrm{w}i} = \beta_{\mathrm{w}}$ for all i is a special case), and we refer to the bias in these two terms as slope bias. The numerator of term (11) will be 0 when the weighted sample covariance between the ecological covariate averages and the projection of $\bar{z}_i^{\mathrm{c}}$ onto $\bar{x}_i$ is 0. 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 expression (11) will be 0. Therefore, $E[\bar{z}_i^{\mathrm{c}} | \bar{\mathbf{x}}] = E[\bar{z}_i^{\mathrm{c}}]$ is a sufficient and almost necessary condition for expression (11) to be 0, and this term represents bias due to an unmeasured confounder. Our decomposition is similar to the decomposition in equation (3) of Greenland and Mogenstern (1989) or equation (8) of Richardson (1992), except that they assumed that 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.

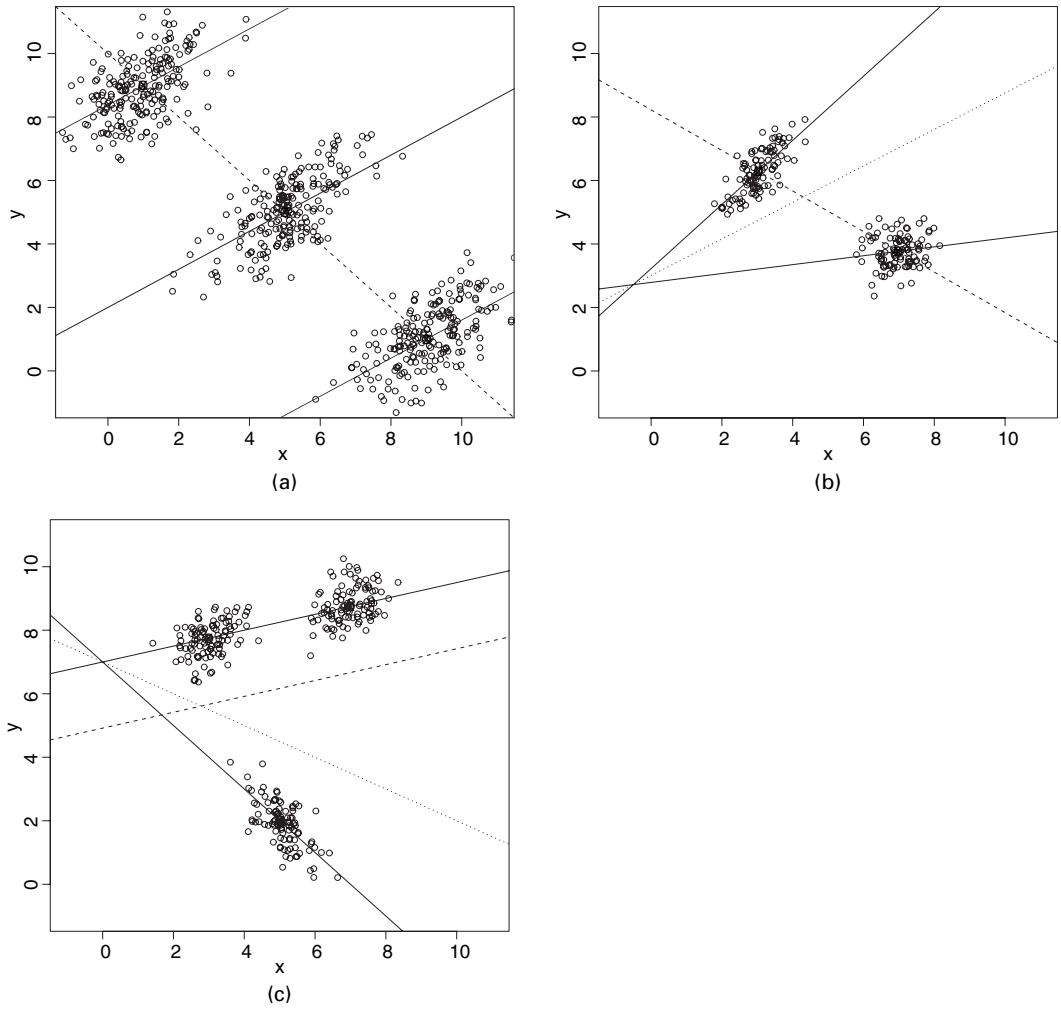

Fig. 1. Sources of ecological bias (———, within-group regression; -------, ecological regression; · · · · · · ·, average within-group regression, the slope of which is the parameter of interest): (a) example of intercept bias, group intercepts negatively correlated with covariate group means; (b) example of slope bias, group slopes negatively correlated with covariate group means; (c) example of slope bias, group slopes quadratically related to covariate group means

2.1. Intercept bias

If the linear expectation is given by model (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 estimator will be unbiased when the group-specific intercepts are uncorrelated with the covariate group means. Fig. 1(a) shows an example where this condition does not hold and clearly illustrates that the ecological regression estimate has the wrong sign. The broken line represents the ecological regression line, and we see that the slope of this line is negative, whereas 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 1 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. To demonstrate the presence or lack of this correlation, we

need a more explicit parameterization of the data-generating process. There are various causal (data-generating) models that lead to the correlated intercepts model, and we shall address two: the contextual effects model and the group level confounder model. Within this context, we shall use γs to signify fixed unknown parameters from the explicit data-generating model and continue to use βs to signify fixed unknown parameters from a less explicit causal model that represents only the causal effect of x_{ij} on y_{ij}. Furthermore, we shall 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 shall obtain a biased estimate. Instead, we obtain an unbiased estimate of the between-group parameter γ_b.

In the group level confounding model, we assume a single confounder z_i, which varies only 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 $\beta_{0i} = \gamma_0 + \gamma_c E[z_i|\bar{x}_i]$ 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 shall obtain a biased estimate.

Although 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 model (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 correct other intercept bias problems.

2.2. Slope bias

If the linear expectation is given by model (2), then ecological bias can arise from intercept bias or slope bias. Varying slopes is sometimes called 'effect modification' (Greenland and Morgenstern, 1989). Fig. 1(b) shows an example where the group-specific slopes are correl-

ated with the covariate group means, and again the ecological regression estimate has the wrong sign. The broken line represents the ecological regression line, and we see that the slope of this line is negative, whereas the within-group slopes are both positive. The slope of the dotted line represents the average within-group slope. 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 include the interaction term in the model. For simplicity, we are assuming that either x_{ij} or z_i is binary:

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

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

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

has the intercept term $\beta_{0i} = \gamma_0 + \gamma_{\mathrm{c}}\, E[z_i|\bar{x}_i]$ and the slope term $\beta_{\mathrm{w}i} = \gamma_{\mathrm{int}}\, E[z_i|\bar{x}_i] + \gamma_{\mathrm{w}}$, where β_{0i} and $\beta_{\mathrm{w}i}$ will be correlated with $\bar{x}_i$. Therefore, if we attempt to estimate $\bar{\beta}_{\mathrm{w}}$ with $\hat{\bar{\beta}}_{\mathrm{w}}^{\mathrm{eco}}$ we shall obtain a biased estimate. In model (14) the correlated slopes are accompanied by correlated intercepts. This will be true for most models with correlated slopes, and hence we shall often need to correct both sources of bias simultaneously. If the linear expectation is given by model (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 $\beta_{\mathrm{w}i} = \gamma_{\mathrm{int}}\, E[z_i|\bar{x}_i] + \gamma_{\mathrm{w}}$ within each group will be unbiased, and our estimate of the average slope will also be unbiased.

2.3. *Within-group confounding*

If the linear expectation is given by model (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 **x** and **y** only (unless there is only group level confounding). 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 model (3). We parameterize the linear expectation of y_{ij} as $\gamma_{0i} + \gamma_{\mathrm{w}i} x_{ij} + \gamma_{\mathrm{c}i} z_{ij}$, where $\gamma_{\mathrm{c}i}$ is the causal confounding effect. If we do not 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 ordinary least squares estimates from a regression of $\mathbf{z}_i$ on $\mathbf{x}_i$ within each group, and u_{ij}s are the residuals from this regression. The individual model can then be rewritten as

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

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 $E[a_i|\mathbf{x}_i] = E[a_i|\bar{x}_i]$ and $E[b_i|\mathbf{x}_i] = E[b_i|\bar{x}_i]$, then the ecological model can be rewritten as

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

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

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 model (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

To illustrate the problem of linear ecological bias, we shall 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 labour 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), who compared different estimates of the causal effect of education on earnings in the context of the British National Child Development Survey. Our goal here is to demonstrate the dangers and relevance of ecological bias. The ecological data are available through the Public Use Microdata Survey; Ruggles *et al.* (2004). These data represent male full-time workers (35 or more hours per week and 48 or more weeks per year) in Washington State, aged 18–65 years in the 2000 census who earned between \$0 and \$175000 during the previous calendar year. We used this selection criterion because by convention the census recodes all yearly wages that are greater than \$175000 to the state average of people with wages that are greater than \$175000. This group of high earners represented 1.7% of the data.

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 0 otherwise. These data are divided into 11 groups ($i = 1, \ldots, 11$), where each group represents a geographical area known as a super public use microdata area (super-PUMA). Super-PUMAs are contiguous geographic areas that contain roughly 400000 people. Populous counties are split into multiple super-PUMAs, whereas less populous counties may be grouped into a single super-PUMA.

The histograms in Fig. 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 logarithmic 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 shall proceed as if we did not and use the untransformed data.

The linear models in Section 2.2 do not make explicit distributional assumptions, but to simplify the discussion we shall assume constant variances across groups and constant variances within each group. For our application, these assumptions appear to be somewhat problematic.

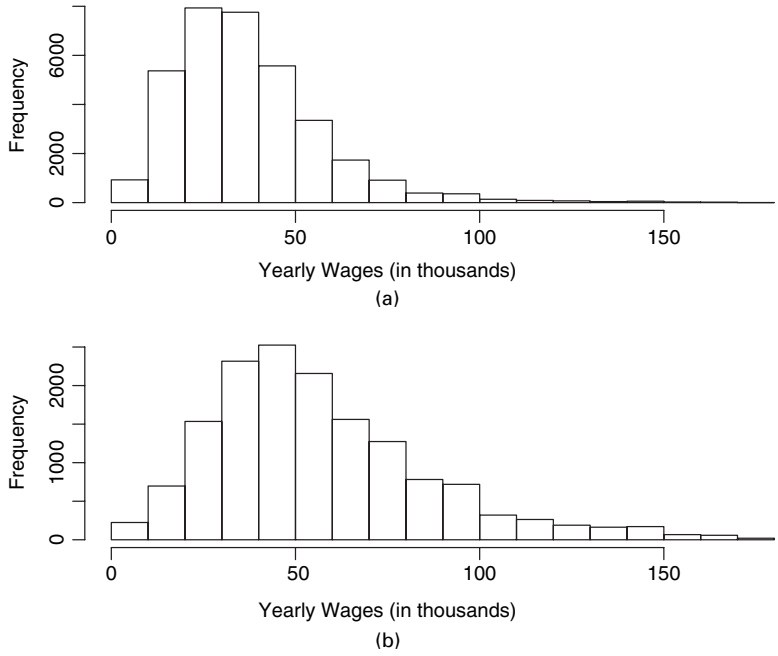

Fig. 2. Wage histograms for individuals (a) without and (b) with a college degree

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 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 11 areas in Washington state. The average yearly wage (in thousands) for area i ($\bar{y}_i$) is the ecological response, whereas the proportion of college degrees in area i ($\bar{x}_i$) is the ecological covariate.

In Fig. 3, we see the effects of aggregating the data. The circles represent the ecological data from Table 1, and the broken line is the ecological regression line. The full lines represent the

Table 1. Ecological data

	Results for the following areas:										
	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

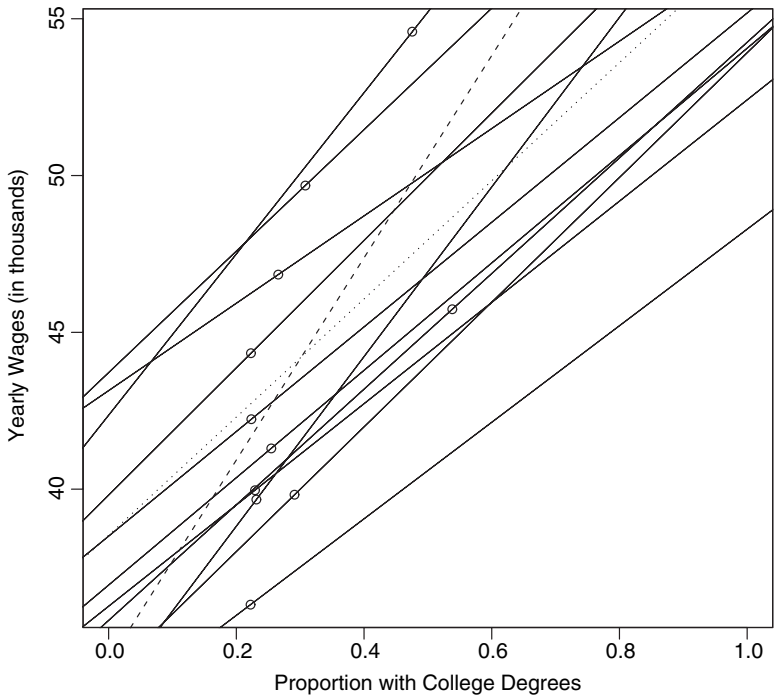

Fig. 3. Ecological regression *versus* within-group regressions

Table 2. Ecological bias (using the full data as the population): intercept and slope bias defined in expressions (9) and (10)

Data	*Slope estimate*	*Standard error*	*Bias*	*Intercept bias*	*Slope bias*
Full	18.8	0.235	0	—†	—†
Ecological	32.2	11.91	13.4	7.0	4.7 1.7

†Not applicable.

within-group regression lines for each of the super-PUMAs. The dotted line represents the weighted average of the full lines. Table 2 shows that the ecological slope (32.2) is too high compared with the weighted average within-group slope (18.8): a difference of 13.4. In fact, the ecological regression is so bad 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 overestimate the value of a college degree greatly. If we treat the individual level data as the population, then the difference between the ecological estimate and the weighted average within-group slope can be treated as bias, and we can decompose this bias by using expressions (9) and (10). Using the estimated parameters from the full data as the true parameter values, a proportion $7.0/13.4 = 0.52$ of the bias comes from the intercepts (9) and a proportion $6.4/13.4 = 0.48$ of the bias comes from the slopes (10). Additionally, the slope bias can be broken into its linear component proportion ($4.7/13.4 = 0.35$) and its quadratic component proportion ($1.7/13.4 = 0.13$). 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 Fig. 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 Fig. 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 shall postpone the discussion of confounding within this application until Section 7, where we reanalyse the data.

4. Estimation with ecological and subsample data

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

$$\begin{gathered} y_{ij} = E[y_{ij}|x_{ij}, z_{ij}] + \varepsilon_{ij}, \\ E[y_{ij}|x_{ij}, z_{ij}] = \beta_{0i} + \beta_{\mathrm{w}i} x_{ij} + \beta_{\mathrm{c}i} z_{ij}, \\ \varepsilon_{ij}|x_{ij}, z_{ij} \underset{\mathrm{IID}}{\sim} N(0, \sigma_e^2), \end{gathered} \tag{17}$$

where the ecological model is given by

$$\begin{gathered} \bar{y}_i = E[\bar{y}_i|\bar{x}_i, \bar{z}_i] + \bar{\varepsilon}_i, \\ E[\bar{y}_i|\bar{x}_i, \bar{z}_i] = \beta_{0i} + \beta_{\mathrm{w}i}\bar{x}_i + \beta_{\mathrm{c}i}\bar{z}_i, \\ \bar{\varepsilon}_i|\bar{x}_i, \bar{z}_i \underset{\mathrm{ind}}{\sim} N(0, \sigma_e^2/n_i). \end{gathered}$$

Suppose that we have a subsample of the individual level data (y_{ij}, x_{ij}, z_{ij}) for individuals $j = 1, \ldots, k_i$ in groups $i = 1, \ldots, m$, where $k_i < n_i$, and n_i represents the total number of individuals in group i. We shall denote this subsample data $(\mathbf{y}_i^{\mathrm{s}}, \mathbf{x}_i^{\mathrm{s}}, \mathbf{z}_i^{\mathrm{s}})$. 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 $j = 1, \ldots, k_i$ and $i = 1, \ldots, m$. Note that the centring here is done around the ecological means and not the subsample means, which we denote $(\bar{y}_i^{\mathrm{s}}, \bar{x}_i^{\mathrm{s}}, \bar{z}_i^{\mathrm{s}})$. The model for the combined ecological and subsample data within each group can be written as

$$\left(\begin{pmatrix} (\mathbf{y}_i^{\mathrm{s}} - \bar{\mathbf{y}}_i) \\ \bar{y}_{i\cdot} \end{pmatrix} \middle| \begin{matrix} (\mathbf{x}_i^{\mathrm{s}} - \bar{\mathbf{x}}_i), (\mathbf{z}_i^{\mathrm{s}} - \bar{\mathbf{z}}_i), \bar{x}_i, \bar{z}_i \\ (\beta_{0i}, \beta_{\mathrm{w}i}, \beta_{\mathrm{c}i}) \end{matrix} \right) \underset{\mathrm{ind}}{\sim} N_{k_i+1}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{18}$$

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

$$\mu_i = \begin{pmatrix} \beta_{\mathrm{w}i}(x_{i1} - \bar{x}_i) + \beta_{\mathrm{c}i}(z_{i1} - \bar{z}_i) \\ \vdots \\ \beta_{\mathrm{w}i}(x_{ik_i} - \bar{x}_i) + \beta_{\mathrm{c}i}(z_{ik_i} - \bar{z}_i) \\ \beta_{0i} + \beta_{\mathrm{w}i}\bar{x}_i + \beta_{\mathrm{c}i}\bar{z}_i \end{pmatrix},$$

$$\Sigma_i = \begin{pmatrix} \Sigma_{11i} & \Sigma_{12i} \\ \Sigma_{21i} & \Sigma_{22i} \end{pmatrix} = \begin{pmatrix} \sigma_e^2\{I_{k_i} - (1/n_i)J_{k_i}\} & \mathbf{0}_{k_i} \\ \mathbf{0}_{k_i}^{\mathrm{T}} & \sigma_e^2/n_i \end{pmatrix}$$

and I_{k_i} is an identity matrix of size k_i, J_{k_i} is a $k_i \times k_i$ matrix of 1s and $\mathbf{0}_{k_i}$ is a $k_i \times 1$ vector of 0s. We shall refer to the first k_i equations in expression (18) as the centred 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 centred data are independent of the ecological data.

An unbiased combined estimator for the $\beta_{\mathrm{w}i}$- and $\beta_{\mathrm{c}i}$-parameters can be derived from the centred data and has the standard generalized least squares form

$$\begin{pmatrix} \hat{\beta}_{\mathrm{w}i}^{\mathrm{comb}} \\ \hat{\beta}_{\mathrm{c}i}^{\mathrm{comb}} \end{pmatrix} = (\mathbf{X}_i^{*\mathrm{T}} \Sigma_{11i}^{-1} \mathbf{X}_i^*)^{-1} \mathbf{X}_i^{*\mathrm{T}} \Sigma_{11i}^{-1} \mathbf{y}^*, \tag{19}$$

$$\mathbf{X}_i^* = ((\mathbf{x}_i^s - \bar{\mathbf{x}}_i)(\mathbf{z}_i^s - \bar{\mathbf{z}}_i)),$$

$$\mathbf{y}_i^* = (\mathbf{y}_i^s - \bar{\mathbf{y}}_i).$$

We note two important properties of this estimator. First, it is clearly unbiased for the group-specific parameters, and therefore an unbiased estimator for $\bar{\beta}_{\mathrm{w}}$ 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 to obtain an estimate for $\beta_{\mathrm{w}i}$. An estimate of σ_e^2 would be necessary for interval estimates about $\bar{\beta}_{\mathrm{w}}$ (here we focus on point estimation).

Since the ecological estimator can be severely biased and the combined estimator eliminates this bias, we rule out the use of the ecological estimator for comparison. In the remaining sections of this paper, we compare the combined estimator under different subsample designs with an estimator that is based solely on a simple random subsample.

5. Comparisons of information for the subsample and combined data

In the previous section, we showed that the combined estimator (19) is unbiased for the slope parameters. However, a maximum likelihood estimator that is 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 gain in information. 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 that is provided by the ecological data about the slope parameters can be small under SRS. 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 $I_{S_i}(\beta_{0i}, \beta_{\mathrm{w}i}, \beta_{\mathrm{c}i})$, and that for the combined data, $\{S_i, E_i\}$, as $I_{S_i,E_i}(\beta_{0i}, \beta_{\mathrm{w}i}, \beta_{\mathrm{c}i})$. In many cases, we shall 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

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

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 shall often discuss the information about $\beta_{\mathrm{w}i}$ in the combined data, while taking into account the information that is lost owing to uncertainty about β_{0i} and $\beta_{\mathrm{c}i}$. We shall write this as $I_{S_i,E_i}(\beta_{\mathrm{w}i})$.

5.1. Information in the varying intercepts and slopes model

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

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

We define

$$\bar{x}_i^s = \frac{1}{k_i} \sum_{j=1}^{k_i} x_{ij}$$

and

$$s_{x_i}^2 = \frac{1}{k_i} \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_i^s)^2$$

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

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

The first term of equation (21) corresponds to the information in the first k_i elements of expression (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 information only about β_{wi}. Also note that this term is undefined if $n_i = k_i$. The second term of equation (21) corresponds to the information in the ecological data. We can add these information matrices together, because the centred values are independent of the ecological values as shown in expression (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,

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

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

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

The second term of equation (22) will be positive and less than 1 and the second term of equation (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 gain in information 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

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

The information in the combined data is given by

$$I_{S_i,E_i}(\beta_{\mathrm{w}i}) = \frac{1}{\sigma_e^2}\left\{k_i(s_{x_i}^2 + c_i a_i^2) + n_i \bar{x}_i^2 - \frac{n_i^2 \bar{x}_i^2}{n_i}\right\}$$
$$= \frac{1}{\sigma_e^2} k_i (s_{x_i}^2 + c_i a_i^2), \tag{25}$$

showing that the gain in information 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 $\beta_{\mathrm{w}i}$ depends on a_i^2, the increase can be quite small under subsampling schemes which produce $\bar{x}_i^{\mathrm{s}} \approx \bar{x}_i$. In particular, if we view the ecological data as fixed, and we subsample at random from the n_i observations within group i, the expectation of $c_i a_i^2$ under this SRS is $(1/k_i)\sigma_{x_i}^2$ where

$$\sigma_{x_i}^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

is the finite population variance (see Cochran (1977), pages 23–24). Therefore, the relative increase in information under SRS diminishes as k_i increases.

Note also that equation (25) is identical to the lower right-hand element of the first term in equation (21). In this sense, the ecological data have information only about the slope parameter through its inclusion in the first k_i elements of expression (18), or the $(y_{ij} - \bar{y}_i)$-terms. If we are interested only 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 (see Chamberlain (1984), pages 1254–1256), which we shall adopt for the rest of this section, effectively ignoring β_{0i}.

In some cases we may gain more information about $\beta_{\mathrm{w}i}$ from the ecological data if we model the β_{0i}-terms. However, there are cases where modelling the β_{0i}-terms will not help in the estimation of $\beta_{\mathrm{w}i}$. For example, in the contextual effects model of Section 2, $\beta_{0i} = \gamma_0 + (\gamma_{\mathrm{b}} - \gamma_{\mathrm{w}})\bar{x}_i$ and $\bar{y}_i = \gamma_0 + \gamma_{\mathrm{b}}\bar{x}_i + \bar{\varepsilon}_i$, so the ecological data provide no information about β_{w} outside the difference data.

5.2. Information in the within-group confounding model

In model (17), we need only to estimate $\beta_{\mathrm{w}i}$ and $\beta_{\mathrm{c}i}$ since we can ignore β_{0i} if we use the centred 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 $b_i = \bar{z}_i^{\mathrm{s}} - \bar{z}_i$ be the differences between the ecological confounder means and the subsample confounder means. When $\mathbf{z}_i^{\mathrm{s}}$ 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

$$I_{S_i}(\beta_{\mathrm{w}i}, \beta_{\mathrm{c}i}) = \frac{1}{\sigma_e^2}\begin{pmatrix} k_i s_{x_i}^2 & k_i s_{x_i z_i} \\ k_i s_{x_i z_i} & k_i s_{z_i}^2 \end{pmatrix}, \tag{26}$$

$$I_{S_i,E_i}(\beta_{\mathrm{w}i}, \beta_{\mathrm{c}i}) = \frac{1}{\sigma_e^2}\begin{pmatrix} k_i(s_{x_i}^2 + c_i a_i^2) & k_i(s_{x_i z_i} + c_i a_i b_i) \\ k_i(s_{x_i z_i} + c_i a_i b_i) & k_i(s_{z_i}^2 + c_i b_i^2) \end{pmatrix}. \tag{27}$$

Hence the leading diagonal elements of matrix (27) are at least as large as the diagonal elements of matrix (26), but the effect of this increase on the information about $\beta_{\mathrm{w}i}$ can be diminished by

the effects of the off-diagonal terms. Accounting for the estimation of β_{ci} gives

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

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

The second terms of equations (28) and (29) correspond to the amount of information that is lost owing to uncertainty about the β_{ci}-parameter. For different data realizations, the information that is lost may be greater for the subsample or the combined approach. If the subsample covariance between x and z is 0 ($s_{x_i z_i} = 0$), then the combined approach will lose at least as much information as the subsample approach owing to the nuisance parameter. However, if there is strong subsample correlation between x and z, then the combined approach may lose less information owing 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 $\bar{x}_i^s \approx \bar{x}_i$ and $\bar{z}_i^s \approx \bar{z}_i$. In particular, the expectation of equation (27) under SRS will approach the expectation of equation (26) under SRS as k_i increases (see Cochran (1977)). We shall show in Section 6 that the information which is 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, the subsampling design should depend on the distribution of the subsample data conditional on the ecological data. Since the k_i centred data equations of expression (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. In this section, we consider the optimal design under the various models.

6.1. Varying intercepts and slopes

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

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

Recall that $a_i = \bar{x}_i^s - \bar{x}_i$ and $c_i = 1/(1 - k_i/n_i)$. Since the ecological data are observed, we can use equation (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 $s_{x_i}^2$, the variability of the subsample.

In our college degree–wage example, x_{ij} is a binary college degree indicator, and therefore equation (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 1s (college degree) or all 0s (no college degree) within each group. Therefore, the maximum information that is available from the combined approach using optimal design is

$$\frac{k_i}{\sigma_e^2} \max\{c_i \bar{x}_i^2, c_i(1-\bar{x}_i)^2\}.$$

The maximum information that is available in the subsample alone by using an optimal design is only $k_i/4\sigma_e^2$, 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 7. Therefore, to maximize information we should sample only people who are without college degrees from group 7 and sample only people with college degrees from all other groups. Such a sampling scheme has a familiar interpretation in that we shall 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 model checking. However, even under this more robust sampling scheme, equation (30) will still be useful, because it describes the information that is lost when sampling 'non-optimal' individuals.

Additionally, if we assume model (1), then $\beta_{wi} = \beta_w$ for all $i = 1, \ldots, m$, 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, equation (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 with 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 2 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 does not have many people with college degrees. Of course, in practice 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 equation (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

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

and there is no easy rule for maximizing equation (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 shall lose no information due to uncertainty about β_{ci}. In our college degree–wage example, suppose that we believe that race (white *versus* non-white) is a confounder. As discussed before, we can maximize the first term of equation (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 0 ($s_{x_i z_i} = 0$). Therefore, we need only to force $b_i = 0$ to cancel the second term in equation (31). To achieve this cancellation in our example, we need to sample college

graduates in racial proportions that match the population racial proportions within each area, i.e. $\bar{z}_i^s = \bar{z}_i$. 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-only 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 clear, and equation (22) shows a significant gain in information. However, when estimating the slope parameters, the increase in information can be negligible. In this section, we shall study the benefits of the combined approach in the context of the Public Use Microdata Survey data that were presented in Section 3 with yearly wages (in thousands) as the response, a college degree indicator as the covariate of interest and a racial indicator (white–non-white) as a potential confounder.

As in Section 3, we treat the full individual level data as if they were the population of interest. This allows us to assess the efficacy of the methodology that is developed in this paper. Specifically, we investigate the benefit to be gained from a subsampling design conditional on the ecological data. As discussed in Section 5, the information about $\bar{\beta}_w$ does not depend on the ecological response data, and hence we need to consider only 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 what follows, we show that, for the three models, optimal design in the combined approach produces substantially more precise estimates of the slope.

7.1. 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 on the basis of 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 except group 7 (see Table 2). Therefore, we can maximize information by sampling only non-college graduates in group 7 and college graduates within all other groups.

To compare the combined estimator under optimal design with the combined and subsample estimators under SRS, we generated 1000 simple random subsamples and 1000 samples under the optimal design. These optimal design subsamples include random subsamples of non-college graduates from group 7, and college graduates for all other groups. From here on, we shall refer to this type of subsample as college random subsampling (CRS). 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 and $\hat{\beta}_w^{comb}$ under CRS.

Fig. 4(a) shows the comparison between these sampling distributions. The full line represents the full data maximum likelihood estimator ($\hat{\beta}_w^{full}$), and the broken line represents the ecological regression estimator. When the within-group samples are small ($k_i = 5$), $\hat{\beta}_w^{comb}$ under CRS has less variability than $\hat{\beta}_w^{comb}$ under SRS, which has less variability than $\hat{\beta}_w^{sub}$ under SRS. Note that lower variability of the estimates indicates greater precision in this case, because the subsampling distributions are centred at the full data maximum likelihood estimator for all three estimators. Also note that 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

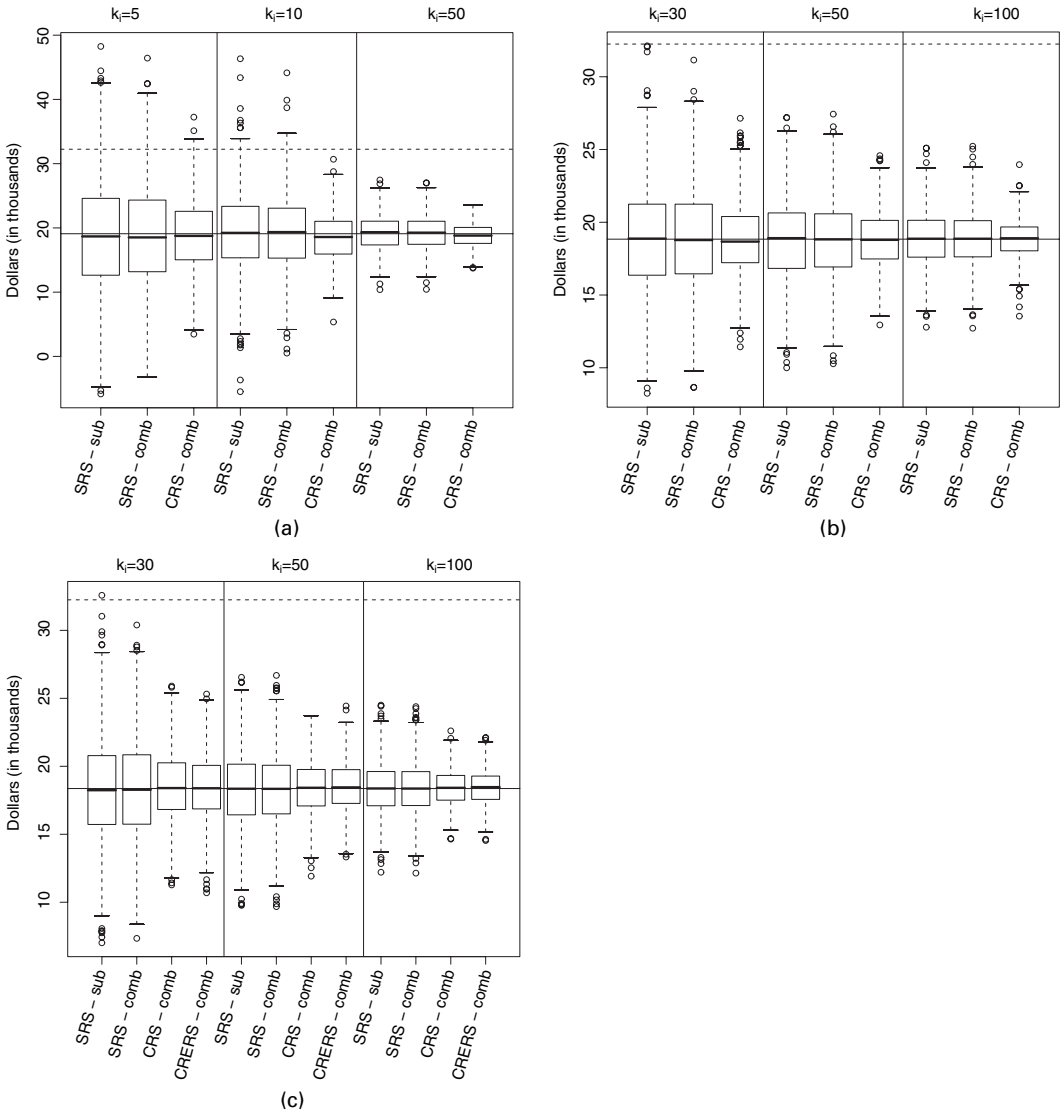

Fig. 4. Subsampling distributions for (a) $\hat{\beta}_{\mathrm{w}}$ in the varying intercepts model and $\hat{\beta}_{\mathrm{w}}$ in (b) the varying intercepts and slopes model and (c) the within-group confounding model: (a) and (b) employ three estimation approaches (subsample data based on simple random subsamples (SRS-sub), combined data based on simple random subsamples (SRS-comb) and combined data based on college random subsamples (CRS-comb)); (c) also employs an estimation approach based on combined data with college random subsamples and ecological racial proportions (CRERS-comb) (———, full data maximum likelihood estimate; -------, ecological estimate)

sampling distributions, and all three approaches can produce estimates that have more error than the ecological estimate. When $k_i = 10$, $\hat{\beta}_{\mathrm{w}}^{\mathrm{comb}}$ under SRS is only slightly more precise than $\hat{\beta}_{\mathrm{w}}^{\mathrm{sub}}$ under SRS, but $\hat{\beta}_{\mathrm{w}}^{\mathrm{comb}}$ under CRS is far more precise than either of the SRS estimators. Additionally, the two estimators under SRS can still produce estimates that are worse than the ecological regression estimate, whereas $\hat{\beta}_{\mathrm{w}}^{\mathrm{comb}}$ under CRS is virtually assured of doing better. When $k_i = 50$, the subsampling distributions for $\hat{\beta}_{\mathrm{w}}^{\mathrm{sub}}$ under SRS and $\hat{\beta}_{\mathrm{w}}^{\mathrm{comb}}$ under SRS are

Table 3. Estimated variance ratios for $\hat{\beta}_w$ in the varying intercepts model, and $\hat{\tilde{\beta}}_w$ in the varying intercepts and slopes model and the within-group confounding model based on subsampling distributions for various estimation approaches: subsample data based on simple random subsamples, combined data based on simple random subsamples, combined data based on college random subsamples and college random subsamples with ecological racial proportions for the within-group confounding model

Data	*Results for varying intercepts*			*Results for varying intercepts and slopes*			*Results for 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	0.831	0.897	0.973	0.929	0.951	0.992	0.940	0.951	0.991
CRS—combined	0.357	0.352	0.458	0.451	0.478	0.444	0.450	0.507	0.446
CRSERP—combined	—†	—†	—†	—†	—†	—†	0.414	0.397	0.439

†Not applicable.

virtually identical, whereas the subsampling distribution for $\hat{\beta}_w^{comb}$ under CRS is more precise.

Table 3 reinforces the importance of optimal design in the varying intercepts model by reporting the estimated variance ratios that are calculated from the subsampling distributions. Although all three estimators are unbiased, the combined estimators have less variability than the subsample estimator. Additionally, 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. Design in the varying intercepts and slopes model

In the varying intercepts and slopes model, we must subsample individuals from all groups to eliminate bias. Therefore, we shall concentrate on optimal design within each group, and again the results from the previous section hold. We should sample only non-college graduates in group 7 and only college graduates in all other groups.

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

Fig. 4(b) shows the comparison between these sampling distributions, for within-group sample sizes of 30, 50 and 100. The full line represents the full data maximum likelihood estimator ($\hat{\tilde{\beta}}_w^{full}$), and the broken line represents the ecological regression estimator. Under all three subsample sizes, $\hat{\tilde{\beta}}_w^{comb}$ under CRS has less variability than $\hat{\tilde{\beta}}_w^{comb}$ under SRS, which has less variability than $\hat{\tilde{\beta}}_w^{sub}$ under SRS. However, because the smallest within-group samples are relatively large ($k_i = 30$), only $\hat{\tilde{\beta}}_w^{sub}$ under SRS produces estimates that have more error than the ecological estimate. When $k_i = 50$ or $k_i = 100$, the sampling distributions for $\hat{\tilde{\beta}}_w^{sub}$ under SRS and $\hat{\tilde{\beta}}_w^{comb}$ under SRS are virtually identical, whereas the sampling distribution for $\hat{\tilde{\beta}}_w^{comb}$ under CRS is more precise.

Table 3 reinforces the impressions from Fig. 4(b). The combined estimators have less variability 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.

7.3. 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 β_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. From here on, we refer to this type of sample as college random sampling with ecological racial proportions (CRSERP). When fitting a model with a confounder, it is important to use the ecological racial proportions in the college sampling scheme because, when using CRS, you occasionally have a sample with only white individuals. The combined estimator that controls for confounding will not be identified by a subsample that consists solely of white individuals with college degrees. Therefore, in our CRS scheme, we discarded college random samples with only white individuals. In Fig. 4(c), we present sampling distributions that are based on three types of subsampling: SRS, CRS and CRSERP.

Table 3 shows that, even with the introduction of the confounder, the CRS and CRSERP combined estimators are less variable than the SRS estimators (the variability of the CRS estimator is understated owing to the discarded samples). And, again, this improvement is apparent as the subsample size becomes larger. Additionally, the CRSERP combined estimator seems to perform better than the CRS combined estimator, although the decrease in variability can be small.

8. Discussion

In this paper, we have discussed linear ecological bias and provided an approach to combining ecological and subsample data to correct this bias. We have also shown that, although the increase in information from the combined approach over a subsample approach can be small under SRS, conditioning on the ecological data allows us to maximize information through optimal subsampling design. Furthermore, the application shows that, even when the ecological estimator is severely biased, estimates that are based on small simple random subsamples can have more error. In contrast, combined estimates that are based on ecological data and small optimally chosen subsamples almost always have less error than an estimate that is based solely on the ecological data. These results 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 does not fit the data perfectly. Second, we have assumed that it is possible to stratify 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

become quite useful if these assumptions do not hold. We can often use the ecological data to inform the correction of non-response in the subsample, and we can informally test for sampling frame bias by comparing the results from the subsample and combined data approaches to see whether the differences are reasonable given the theoretical variability.

It is natural to extend the results of this paper to generalized linear models. Ecological bias is often a larger problem in non-linear models than in linear models (Greenland, 1992), and therefore combined data estimation and design conditional on the ecological data would be beneficial. Some work has already been done on estimation in this area (Wakefield, 2004; Haneuse and Wakefield, 2008; Jackson *et al.*, 2006), and this work hints at sampling design heuristics without providing a formal framework for design. It would be useful to extend the results of this paper and to develop an information-based framework for design in generalized linear models conditional on ecological data.

Acknowledgements

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 National Science Foundation grant DMS 0505865 and grant R01 AI032475 from the National Institutes of Health. The authors are grateful to the Joint 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.

Any errors are the sole responsibility of the authors.

Appendix A: Decomposition of bias

By definition $\Sigma_{i=1}^{m} n_i \bar{x}_i = n\bar{x} = \Sigma_{i=1}^{m} n_i \bar{x}$, and therefore $\Sigma_{i=1}^{m} n_i \bar{y}(\bar{x}_i - \bar{x}) = 0$, and $\hat{\tilde{\beta}}_{\mathrm{w}}^{\mathrm{eco}}$ can be simplified as

$$\begin{aligned}\hat{\tilde{\beta}}_{\mathrm{w}}^{\mathrm{eco}} &= \frac{\sum_{i=1}^{m} n_i (\bar{y}_i - \bar{y})(\bar{x}_i - \bar{x})}{\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{m} n_i \bar{y}_i (\bar{x}_i - \bar{x})}{\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2}. \end{aligned} \tag{32}$$

Additionally, $E[\bar{y}_i|\tilde{\mathbf{x}}] = \beta_{0i} + \beta_{\mathrm{w}i}\bar{x}_i + E[\bar{z}_i^{\mathrm{c}}|\tilde{\mathbf{x}}]$, and, using the same algebraic method as above, we include $\tilde{\beta}_0$ in expression (33) and $E[\bar{z}^{\mathrm{c}}|\tilde{\mathbf{x}}]$ in expression (35) of the following equation:

$$\begin{aligned} E[\hat{\tilde{\beta}}_{\mathrm{w}}^{\mathrm{eco}}|\tilde{\mathbf{x}}] &= \frac{\sum_{i=1}^{m} n_i E[\bar{y}_i|\tilde{\mathbf{x}}](\bar{x}_i - \bar{x})}{\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})(\beta_{0i} + \beta_{\mathrm{w}i}\bar{x}_i + E[\bar{z}_i^{\mathrm{c}}|\tilde{\mathbf{x}}])}{\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2} \end{aligned}$$

$$= \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})(\beta_{0i} - \bar{\beta}_0)}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \tag{33}$$

$$+ \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})\beta_{wi}\bar{x}_i}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \tag{34}$$

$$+ \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})\, E[\bar{z}_i^c - z^c|\bar{\mathbf{x}}]}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2}. \tag{35}$$

Furthermore, $\beta_{wi}\bar{x}_i$ in expression (34) can be written as $\beta_{wi}\bar{x}_i = \bar{\beta}_w\bar{x}_i + \bar{x}(\beta_{wi} - \bar{\beta}_w) + (\bar{x}_i - \bar{x})(\beta_{wi} - \bar{\beta}_w)$ and, since

$$\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})\bar{\beta}_w\bar{x}_i = \sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})\bar{\beta}_w(\bar{x}_i - \bar{x}),$$

we can separate expression (34) into the parameter of interest, the correlated slope bias term, and the quadratic slope bias term. This allows us to write $E[\hat{\bar{\beta}}_w^{eco}|\bar{\mathbf{x}}]$ in the form of expressions (8)–(11):

$$\begin{aligned}
E[\hat{\bar{\beta}}_w^{eco}|\bar{\mathbf{x}}] &= \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})(\beta_{0i} - \bar{\beta}_0)}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \\
&\quad + \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})\{\bar{\beta}_w\bar{x}_i + \bar{x}(\beta_{wi} - \bar{\beta}_w) + (\bar{x}_i - \bar{x})(\beta_{wi} - \bar{\beta}_w)\}}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \\
&\quad + \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})\, E[\bar{z}_i^c - z^c|\bar{\mathbf{x}}]}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \\
&= \bar{\beta}_w + \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})(\beta_{0i} - \bar{\beta}_0)}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \\
&\quad + \bar{x}\frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})(\beta_{wi} - \bar{\beta}_w)}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} + \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2(\beta_{wi} - \bar{\beta}_w)}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2} \\
&\quad + \frac{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})\, E[\bar{z}_i^c - z^c|\bar{\mathbf{x}}]}{\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2}.
\end{aligned} \tag{36}$$

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

Since Σ_{11i} has the form $\sigma_e^2(\mathbf{I}_{k_i} + \mathbf{b}\mathbf{J}_{k_i})$ where $b = -1/n_i$, its inverse will have the form

$$\frac{1}{\sigma_e^2}\left(\mathbf{I}_{k_i}-\frac{\mathbf{b}}{1+k_i\mathbf{b}}\mathbf{J}_{k_i}\right),$$

which simplifies to

$$\frac{1}{\sigma_e^2}\left(\mathbf{I}_{k_i}+\frac{1}{n_i-k_i}\mathbf{J}_{k_i}\right).$$

Therefore, we can derive equation (21) in the usual manner:

$$\begin{aligned}
I_{S_i,E_i}(\beta_{0i},\beta_{\mathrm{w}i}) &= (\mathbf{0}_{k_i} \quad (\mathbf{x}_i^{\mathrm{s}}-\bar{\mathbf{x}}_i))^{\mathrm{T}}\,\Sigma_{11i}^{-1}\,(\mathbf{0}_{k_i} \quad (\mathbf{x}_i^{\mathrm{s}}-\bar{\mathbf{x}}_i)) + (1 \quad \bar{x}_i)^{\mathrm{T}}\,\Sigma_{22i}^{-1}\,(1 \quad \bar{x}_i)\\
&= \frac{1}{\sigma_e^2}\begin{pmatrix} 0 & 0 \\ 0 & (\mathbf{x}_i^{\mathrm{s}}-\bar{\mathbf{x}}_i)^{\mathrm{T}}\dfrac{1}{\sigma_e^2}\left(\mathbf{I}_{k_i}+\dfrac{1}{n_i-k_i}\mathbf{J}_{k_i}\right)(\mathbf{x}_i^{\mathrm{s}}-\bar{\mathbf{x}}_i)\end{pmatrix} + (1 \quad \bar{x}_i)^{\mathrm{T}}\frac{n_i}{\sigma_e^2}(1 \quad \bar{x}_i)\\
&= \frac{1}{\sigma_e^2}\begin{pmatrix} 0 & 0 \\ 0 & \displaystyle\sum_{j=1}^{k_i}(x_{ij}-\bar{x}_i)^2+\frac{1}{n_i-k_i}\left\{\sum_{j=1}^{k_i}(x_{ij}-\bar{x}_i)\right\}^2\end{pmatrix} + \frac{n_i}{\sigma_e^2}\begin{pmatrix}1 & \bar{x}_i\\ \bar{x}_i & \bar{x}_i^2\end{pmatrix}\\
&= \frac{1}{\sigma_e^2}\begin{pmatrix} 0 & 0 \\ 0 & \displaystyle\sum_{j=1}^{k_i}(x_{ij}-\bar{x}_i^{\mathrm{s}})^2+k_i(\bar{x}_i^{\mathrm{s}}-\bar{x}_i)^2+\frac{k_i^2}{n_i-k_i}(\bar{x}_i^{\mathrm{s}}-\bar{x}_i)^2\end{pmatrix} + \frac{n_i}{\sigma_e^2}\begin{pmatrix}1 & \bar{x}_i\\ \bar{x}_i & \bar{x}_i^2\end{pmatrix}\\
&= \frac{1}{\sigma_e^2}\begin{pmatrix} 0 & 0 \\ 0 & \displaystyle\sum_{j=1}^{k_i}(x_{ij}-\bar{x}_i^{\mathrm{s}})^2+\frac{n_ik_i}{n_i-k_i}(\bar{x}_i^{\mathrm{s}}-\bar{x}_i)^2\end{pmatrix} + \frac{n_i}{\sigma_e^2}\begin{pmatrix}1 & \bar{x}_i\\ \bar{x}_i & \bar{x}_i^2\end{pmatrix}.
\end{aligned}$$

If we let

$$s_{x_i}^2=\frac{1}{k_i}\sum_{j=1}^{k_i}(x_{ij}-\bar{x}_i^{\mathrm{s}})^2,$$

$a_i=\bar{x}_i^{\mathrm{s}}-\bar{x}_i$ and $c_i=1/(1-k_i/n_i)$, then we obtain equation (21).

References

Card, D. (1999) The causal effect of education on earnings. In *Handbook of Labor Economics*, vol. 3 (eds O. Ashenfelter and D. Card). Amsterdam: Elsevier.

Card, D. (2001) Estimating the returns to schooling: progress on some persistent econometric problems. *Econometrica*, **69**, 1127–1160.

Chamberlain, G. (1984) Panel data. In *Handbook of Econometrics*, vol. II (eds Z. Griliches and M. Intriligator), pp. 1247–1318. Amsterdam: Elsevier.

Chaudhuri, S., Handcock, M. and Rendall, M. (2005) An empirical likelihood based approach to incorporate population information in sample based inference. *Working Paper 48*. Center for Statistics and the Social Sciences, University of Washington, Seattle.

Cochran, W. (1977) *Sampling Techniques*. New York: Wiley.

Freedman, D., Klein, S., Ostland, M. and Roberts, M. (1998) Review of a solution to the ecological inference problem. *J. Am. Statist. Ass.*, **93**, 1518–1522.

Greenland, S. (1992) Divergent biases in ecologic and individual-level studies. *Statist. Med.*, **11**, 1209–1223.

Greenland, S. and Morgenstern, H. (1989) Ecological bias, confounding, and effect modification. *Int. J. Epidem.*, **18**, 269–274.

Handcock, M., Rendall, M. and Cheadle, J. (2005) Improved regression estimation of a multivariate relationship with population data on the bivariate relationship. *Sociol. Methodol.*, **35**, 303–346.

Haneuse, S. and Wakefield, J. (2008) The combination of ecological and case-control data. *J. R. Statist. Soc.* B, **70**, in the press.

Hellerstein, J. and Imbens, G. (1999) Imposing moment restrictions from auxiliary data by weighting. *Rev. Econ. Statist.*, **81**, 1–14.

Imbens, G. and Lancaster, T. (1994) Combining micro and macro data in microeconomic models. *Rev. Econ. Stud.*, **61**, 655–680.

Jackson, C., Best, N. and Richardson, S. (2006) Improving ecological inference using individual-level data. *Statist. Med.*, **25**, 2136–2159.

King, G. (1997) *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.

Raghunathan, T., Diehr, P. and Cheadle, A. (2003) Combining aggregate and individual level data to estimate an individual level correlation model. *J. Educ. Behav. Statist.*, **28**, 1–19.

Richardson, S. (1992) Statistical methods for geographical correlation studies. In *Analysis of Survey Data* (eds P. Elliott, J. Cuzick, D. English and R. Stern), pp. 181–204. New York: Oxford University Press.

Robinson, W. (1950) Ecological correlations and the behavior of individuals. *Am. Sociol. Rev.*, **15**, 351–357.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C., Goeken, R., Hall, P., King, M. and Ronnander, C. (2004) *Integrated Public Use Microdata Series: Version 3.0* (machine readable database). Minneapolis: Minnesota Population Center.

Steel, D., Beh, E. and Chambers, R. (2004) The information in aggregate data. In *Ecological Inference: New Methodological Strategies* (eds G. King, O. Rosen and M. Tanner). Cambridge: Cambridge University Press.

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

Wakefield, J. (2004) Ecological inference for 2×2 tables (with discussion). *J. R. Statist. Soc.* A, **167**, 385–445.