University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2008 *Paper* 236

Doubly Robust Ecological Inference

Daniel B. Rubin[*] Mark J. van der Laan[†]

[*] Division of Biostatistics, School of Public Health, University of California, Berkeley, daniel.rubin@fda.hhs.gov

[†] Division of Biostatistics and Department of Statistics, University of California, Berkeley, laan@berkeley.edu

Doubly Robust Ecological Inference

Daniel B. Rubin and Mark J. van der Laan

Abstract

The ecological inference problem is a famous longstanding puzzle that arises in many disciplines. The usual formulation in epidemiology is that we would like to quantify an exposure-disease association by obtaining disease rates among the exposed and unexposed, but only have access to exposure rates and disease rates for several regions. The problem is generally intractable, but can be attacked under the assumptions of King's (1997) extended technique if we can correctly specify a model for a certain conditional distribution. We introduce a procedure that it is a valid approach if either this original model is correct or if we can pose a correct model for a different conditional distribution. The new method is illustrated on data concerning risk factors for diabetes.

1 Introduction

Can air pollution lead to diabetes? Lockwood (2002) posed this question and suggested it warranted attention, because dioxin exposure was a possible mechanism, and diabetes prevalences for U.S. states strongly correlated with a measure of statewide toxic releases ($r = 0.54$, $p < .0001$). Others were skeptical, such as Nicolich (2002), who showed these prevalences also strongly correlated with irrelevant variables like state capital latitudes or state places in an alphabetical list. Lockwood's question appears open, as there does not seem to be convincing evidence for or against the hypothesis.

Why didn't the correlation settle that toxic emmisions cause diabetes? First, as Lockwood implied, confounding was a possibility. Maybe people in polluted states had different diets than those in other states, or different demographics (Marquez et al., 2004). Confounding could have been an issue even if there had been observational data on individuals instead of states, and people were asked about both diabetes and their exposure to pollution. A second problem was that the state-level correlation did not even determine an individual-level association, as there was no way to tell whether diabetics in polluted states were the ones actually breathing the polluted air.

Unlike confounding, this second difficulty arose because the study was ecological, meaning observational units were not individuals, but regions. To incorrectly assume that an exposure-disease association among aggregates of individuals implies the association holds among individuals themselves is to commit what is known as the ecological fallacy. Because ecological studies cannot always determine an individual-level association, they are considered weak evidence when investigating an exposure-disease relationship.

Nevertheless, ecological studies are common in epidemiology, as reviewed in Wakefield (2008). This is primarily due to the ease of data collection. Lockwood's pollution and diabetes data were found from the web pages of two government agencies, and it would have taken much longer to sample individuals from a population, determine if they were diabetic, and quantify their exposure to pollution. If two surveillance systems have released data on an exposure and a disease, ecological studies are often convenient. They can therefore serve as starting points for epidemiological inquiry, which was clearly Lockwood's aim. They have played supporting roles in many success stories, including John Snow's classical work on cholera, and arguments for conducting clinical trials that showed circumcision to be protective against HIV (Bongarrts et al., 1989; Moses et al., 1990; Bailey et al., 2001).

A great deal of work has gone into ecological inference, or the analysis of ecological studies. We give pointers to references in Section 14, but the impetus for this paper is the literature growing out of King's (1997) extended model. King's book is influential and highly cited, his approach has the distinction of being accepted in U.S. courts for legal testimony, and a report for the U.S. National Academy of Science concluded "it is not premature to note that this estimation procedure to date represents the most dramatic advance in researchers' ability to draw microlevel inferences from aggregate-level data" (Schuessler, 1999).

Under King's assumptions, it is possible to estimate an individual-level exposure-disease association from an ecological study if one can specify an approximately correct statistical model for a certain conditional distribution. To state our contribution convincingly but vaguely, we present a method allowing one to also specify a statistical model for a different conditional distribution, and obtain desirable overall performance if at least one of the two models is approximately valid. This property of having two chances for success is known as double robustness in the statistical literature.

2 The statistical problem

Suppose we have data for n regions. The data for region i are

W_i = vector of regional covariates, including regional population size N_i

A_i = proportion of people subject to an exposure

Y_i = proportion of people with a disease.

The air pollution and diabetes example does not fall exactly into this framework, because Lockwood's state-level exposure measurements are pounds of released toxins. However, it is easy to imagine there instead being state-level data on the proportion of people living in highly polluted cities. The covariate vector could include prevalences for potential confounders of the exposure-disease relationship. In the diabetes example, we could imagine obtaining state-level dietary information and demographic profiles.

The main problem for inferring an exposure-disease association is that we don't know regional exposure-specific disease rates

$$Y_{E,i} = \text{proportion diseased among those exposed}$$
$$Y_{U,i} = \text{proportion diseased among those unexposed.}$$

Instead, observed exposure and disease rates are related to the unobserved exposure-specific disease rates through

$$Y_i = A_i Y_{E,i} + (1 - A_i) Y_{U,i}.$$

If we had access to regional exposure-specific disease rates, we could compute the exposure-specific disease rates among all individuals, or

$$\begin{aligned}\mu_E &= \frac{\text{number of exposed with disease}}{\text{number of exposed}} = \frac{n^{-1}\sum_{i=1}^{n} N_i A_i Y_{E,i}}{n^{-1}\sum_{i=1}^{n} N_i A_i} \\ \mu_U &= \frac{\text{number of unexposed with disease}}{\text{number of unexposed}} = \frac{n^{-1}\sum_{i=1}^{n} N_i (1 - A_i) Y_{U,i}}{n^{-1}\sum_{i=1}^{n} N_i (1 - A_i)}.\end{aligned}$$

The two unknowns are building blocks for epidemiology, as they determine measures of association like the excess risk, relative risk, and odds ratio. Our goal will consequently be to estimate these total exposure-specific disease rates. To make denominators nonzero we ignore trivial scenarios where everyone is exposed or unexposed. In such cases one exposure-specific disease rate is the disease rate itself while the other is undefined. We are less ambitious than other analysts who estimate exposure-specific disease rates for each of the n regions. Rather than two unknowns the latter problem entails estimating $2n$ unknowns. In the given example, our two quantities of interest would be the total proportion of diabetics among those who live or do not live in heavily polluted cities.

3 Incorporating deterministic information

Before viewing the data, we can only know the total exposure-specific disease rates are somewhere in the unit square. The data tell us they live on a line segment in the unit square, and estimators can exploit this knowledge.

From Duncan and Davis (1953), we can bound regional disease rates under exposure through

$$Y_{E,i} \in [L_i, R_i] = \left[\max(0, (Y_i - 1 + A_i)/A_i),\ \min(1, Y_i/A_i)\right].$$

The implication for the total disease rate under exposure is the bound

$$\mu_E \in [L, R] = \left[\frac{n^{-1}\sum_{i=1}^{n} N_i A_i L_i}{n^{-1}\sum_{i=1}^{n} N_i A_i},\ \frac{n^{-1}\sum_{i=1}^{n} N_i A_i R_i}{n^{-1}\sum_{i=1}^{n} N_i A_i}\right].$$

The total exposure-specific disease rates (μ_E, μ_U) must also satisfy $\bar{Y} = \bar{A}\mu_E + (1-\bar{A})\mu_U$, where $\bar{A} = \sum_{i=1}^n N_i A_i / \sum_{i=1}^n N_i$ is the total proportion of people exposed, and $\bar{Y} = \sum_{i=1}^n N_i Y_i / \sum_{i=1}^n N_i$ is the total proportion of people with the disease. Together with the bound on μ_E, it follows that total exposure-specific disease rates (μ_E, μ_U) live on the line segment connecting $(L, (\bar{Y} - \bar{A}L)/(1-\bar{A}))$ and $(R, (\bar{Y} - \bar{A}R)/(1-\bar{A}))$ in the unit square. Following King (1997), this segment is often called the tomography line.

Even though our two proportions of interest must live on the tomography line, many estimates returned by common procedures do not. An example is Goodman's (1953, 1959) ecological regression estimate, which can be outside the unit square, even though the object is to find two rates. Our initial doubly robust estimate shares this disadvantage. The difficulty can be avoided by mapping an initial estimate to a point on the tomography line. For instance, we could take our final estimate $\tilde{\mu} = (\tilde{\mu}_E, \tilde{\mu}_U)$ to be the closest point on the line segment to the initial estimate $\hat{\mu} = (\hat{\mu}_E, \hat{\mu}_U)$ in terms of Euclidean distance. There is a simple solution to

$$\begin{aligned} \tilde{\mu} &= (\tilde{\mu}_E, \tilde{\mu}_U) = \text{argmin}_{(x,y)} \|(\hat{\mu}_E, \hat{\mu}_U) - (x,y)\|_2 \\ &\quad \text{such that } L \leq x \leq R \text{ and } \bar{Y} = \bar{A}x + (1-\bar{A})y. \end{aligned}$$

We project the initial $\hat{\mu}$ on the extension of the tomography line to the plane, set $\tilde{\mu}$ equal to this projection result if it is on the tomography line, and otherwise set $\tilde{\mu}$ to whichever tomography line endpoint is closer to $\hat{\mu}$.

It is thus relatively simple for any estimator to benefit from potentially valuable deterministic constraints, including our doubly robust estimator.

4 Assumptions of King's extended model

A consensus exists in the ecological inference literature that any point estimation procedure must depend on strong suppositions that cannot be verified from the data. Techniques related to King's extended model make the three assumptions described below, which we modify and relax in a new way.

On matters of notation, King supposes there are two covariates for each region, with one related to the exposed and the other to the unexposed, but we combine all covariates in one vector. He also presents our epidemiological problem as a mathematically equivalent one in political science. Finally, we have referenced King's "extended model," but his basic model is the special case with no covariates.

4.1 No spatial autocorrelation

We begin by thinking of regional exposure rates and exposure-specific disease rates as random variables. The variables corresponding to different regions are assumed to be independent. It makes little difference whether covariates are fixed or random. Although King's model was introduced in the fixed design setting, we take them to be random. Formally, the assumption is then that $(W_1, A_1, Y_{E,i}, Y_{U,i}), ..., (W_n, A_n, Y_{E,n}, Y_{U,n})$ are mutually independent. Although the assumption is not intuitive for many applications, King (1997), Cho (1998), and King et al. (2004) note that reasonable dependence across regions does not introduce major problems for standard estimators.

Numerators $n^{-1}\sum_{i=1}^{n} N_i A_i Y_{E,i}$ and $n^{-1}\sum_{i=1}^{n} N_i(1-A_i)Y_{U,i}$ of our desired quantities can be handled by extensions of the law of large numbers to the non-i.i.d. setting. With many regions and no handful of terms dominating variances of these numerators, they will with large chance be approximately

$$\theta_E = n^{-1}\sum_{i=1}^{n} E[N_i A_i Y_{E,i}]$$

$$\theta_U = n^{-1}\sum_{i=1}^{n} E[N_i(1-A_i)Y_{U,i}].$$

We will thus view the problem through the lens of parameter estimation, and estimate total exposure-specific disease rates with

$$\hat{\mu}_E = \frac{\hat{\theta}_E}{n^{-1}\sum_{i=1}^{n} N_i A_i}$$

$$\hat{\mu}_U = \frac{\hat{\theta}_U}{n^{-1}\sum_{i=1}^{n} N_i(1-A_i)}.$$

This $\hat{\mu} = (\hat{\mu}_E, \hat{\mu}_U)$ can then be mapped into a final $\tilde{\mu} = (\tilde{\mu}_E, \tilde{\mu}_U)$ obeying deterministic constraints, as discussed in Section 3.

4.2 No unmeasured aggregation bias

The second condition is that we collect enough covariate information to ensure a region's exposure rate can provide no additional insight into its exposure-specific disease rates, written as the conditional independence

$$\{(Y_{E,i}, Y_{U,i}) \perp A_i | W_i\}. \tag{1}$$

Unfortunately, this assumption cannot be tested from the data. It is also not entirely clear to what extent domain knowledge can help in collecting covariates to guarantee it holds.

Viewing observed regional data $(W_i, A_i, Y_i = A_i Y_{E,i} + (1 - A_i) Y_{U,i})$ as a coarsened version of unavailable $(W_i, Y_{E,i}, Y_{U,i})$, Imai et al. (2008) remark that the assumption corresponds to Heitjan and Rubin's (1991) coarsening at random, meaning the coarsening process only depends on data that is always available. This is generally the weakest condition allowing one to ignore the unavailable information when fitting statistical models with maximum likelihood in coarsened data structures, and is used by the most popular methods for missing data problems, causal inference, and survival analysis.

4.3 Correct exposure-specific disease rate model

The final assumption is that exposure-specific disease rates for different regions are drawn from a common conditional distribution given covariates, and that this distribution can be correctly modeled.

We write this as $\{Y_{E,i}, Y_{U,i} | W_i\} \sim f_0$, and suppose f_0 belongs to some parametric family. In King's model, exposure-specific disease rates $(Y_{E,i}, Y_{U,i})$ are bivariate normal with mean $[\alpha_0 + \alpha_1^T W_i, \beta_0 + \beta_1^T W_i]$ and unknown covariance matrix, but conditioned to fall in the unit square. Other models have been proposed. Imai and Lu (2005) take [logit $Y_{E,i}$, logit $Y_{U,i}$] to be bivariate normal with mean $[\alpha_0 + \alpha_1^T W_i, \beta_0 + \beta_1^T W_i]$ and unknown covariance matrix, and Wakefield (2004) discusses a related strategy. More complex models involve beta and binomial distributions (King et al., 1999).

With no spatial autocorrelation and no unmeasured aggregation bias, it is possible to factor the observed data likelihood into a part only depending on f_0, and fit the model with $\hat{f}$ through an MLE or a Bayes estimate. Due to no unmeasured aggregation bias, $E[Y_{E,i} | W_i, A_i, Y_i]$ and $E[Y_{U,i} | W_i, A_i, Y_i]$ do not depend on parts of the $(W_i, A_i, Y_{E,i}, Y_{U,i})$ distribution other than f_0. Our parameters of interest can then be estimated through

$$\hat{\theta}_E = n^{-1} \sum_{i=1}^{n} N_i A_i E_{\hat{f}}[Y_{E,i} | W_i, A_i, Y_i]$$
$$\hat{\theta}_U = n^{-1} \sum_{i=1}^{n} N_i (1 - A_i) E_{\hat{f}}[Y_{U,i} | W_i, A_i, Y_i], \qquad (2)$$

which would be unbiased if we could substitute the unknown f_0 for fit $\hat{f}$.

5 An exposure rate model

In addition to modeling the conditional distribution just discussed in Section 4.3, we need to model another conditional distribution. If the new model is accurate, we exploit information previous methods have ignored.

The second model supposes exposure rates for different regions are drawn from a common conditional distribution given covariates, so $\{A_i|W_i\} \sim g_0$, and we can form a fit $\hat{g}$ as g_0 belongs to a parametric family. We treat exposure rates as continuous, and hence take g_0 to be a conditional density, but our estimator could be formulated with g_0 a conditional mass function.

In our data analysis of Section 10 we formed $\hat{g}$ with the beta regression of Ferrari and Cribari-Neto (2004), available through the Comprehensive R Archive Network in the betareg package (Bustamante Simas, 2004). The model specifies that conditional on covariate W_i, exposure rate A_i follows a beta distribution with shapes $\phi\sigma(\beta_0 + \beta_1^T W_i)$ and $\phi(1 - \sigma(\beta_0 + \beta_1 W_i))$, where ϕ is a dispersion parameter and $\sigma(x) = (1 + \exp(-x))$ is the sigmoidal link.

Although not done in our data analysis, we recommend artificially bounding the fitted function $\hat{g}$ away from zero, because our estimator involves inverse density weights $1/\hat{g}(A_i|W_i)$ that can lead to instability if too large.

6 Doubly robust estimator

We are now ready to present a new method that should perform well if one of two models is approximately correct.

Our procedure depends on a user-specified weight function h on the unit interval that is nonnegative, bounded, and nonzero in that $\int_0^1 h(s)ds > 0$. We will have more to say about how to choose the weight function in the next section, but for now suppose we have made a decision. Define the matrix

$$H = \begin{bmatrix} \int_0^1 h(s)ds & \int_0^1 sh(s)ds \\ \int_0^1 sh(s)ds & \int_0^1 s^2 h(s)ds \end{bmatrix},$$

and note that it is nonsingular. We also define the functions

$$\begin{aligned} V_E(a;h) &= \left([1,1]^T H^{-1} \begin{bmatrix} 1 \\ a \end{bmatrix} \right) h(a) \\ V_U(a;h) &= \left([1,0]^T H^{-1} \begin{bmatrix} 1 \\ a \end{bmatrix} \right) h(a). \end{aligned}$$

For given f and g not necessarily related to the unknown data generating distribution, we additionally define functions of regional data according to

$$D_E(W_i, A_i, Y_i; h; f; g) = N_i A_i E_f[Y_{E,i}|W_i]$$
$$+N_i E_g[A_i|W_i] \frac{V_E(A_i; h)}{g(A_i|W_i)} (Y_i - A_i E_f[Y_{E,i}|W_i] - (1 - A_i) E_f[Y_{U,i}|W_i])$$

$$D_U(W_i, A_i, Y_i; h; f; g) = N_i(1 - A_i) E_f[Y_{U,i}|W_i]$$
$$+N_i(1 - E_g[A_i|W_i]) \frac{V_U(A_i; h)}{g(A_i|W_i)} (Y_i - A_i E_f[Y_{E,i}|W_i] - (1 - A_i) E_f[Y_{U,i}|W_i]) .$$

With fits $\hat{f}$ and $\hat{g}$ from models described in Section 4.3 and Section 5, the doubly robust parameter estimates are

$$\hat{\theta}_E = n^{-1} \sum_{i=1}^{n} D_E(W_i, A_i, Y_i; h; \hat{f}; \hat{g})$$
$$\hat{\theta}_U = n^{-1} \sum_{i=1}^{n} D_U(W_i, A_i, Y_i; h; \hat{f}; \hat{g}),$$

which induce total exposure-specific disease rate estimates as in Section 4.1.

The following theorem expresses that our parameter estimators should behave in large samples like unbiased empirical means if one of $\hat{f}$ or $\hat{g}$ is based on a correct model. The previously given estimators of (2) can behave similarly, but depend on $\hat{f}$ being built from a correct model.

Theorem 1. *Assume no unmeasured aggregation bias as in (1). Let $f_{0,i}$ and $g_{0,i}$ denote the true conditional distribution and conditional density for $\{Y_{E,i}, Y_{U,i}|W_i\}$ and $\{A_i|W_i\}$, and let f and g denote another conditional distribution and another conditional density. Assume N_i is bounded, and take g to be bounded away from zero on the support of weight function h in that $h(A_i)/g(A_i|W_i)$ is bounded. Suppose that either $f = f_{0,i}$ in the sense that*

$$E_f[Y_{E,i}|W_i] = E_{f_{0,i}}[Y_{E,i}|W_i] \text{ and } E_f[Y_{U,i}|W_i] = E_{f_{0,i}}[Y_{U,i}|W_i] \text{ almost surely,}$$

or $g = g_{0,i}$ in the sense that

$$\frac{h(A_i)}{g(A_i|W_i)} = \frac{h(A_i)}{g_{0,i}(A_i|W_i)} \text{ and } E_g[A_i|W_i] = E_{g_{0,i}}[A_i|W_i] \text{ almost surely.}$$

Then $D_E(W_i, A_i, Y_i; h; f; g)$ and $D_U(W_i, A_i, Y_i; h; f; g)$ have expected values $E[N_i A_i Y_{E,i}]$ and $E[N_i(1 - A_i)Y_{U,i}]$.

Proof. The conditions on N_i and $h(A_i)/g(A_i|W_i)$ being bounded ensure the expectations that follow will be well defined and finite. The latter implies $V_E(A_i;h)/g(A_i|W_i)$ and $V_U(A_i;h)/g(A_i|W_i)$ are also bounded random variables, almost surely equal to $V_E(A_i;h)/g_{0,i}(A_i|W_i)$ and $V_U(A_i;h)/g_{0,i}(A_i|W_i)$ if the theorem's $g = g_{0,i}$ condition holds. We also note that it is a simple calculation to show $\int_0^1 sV_E(s;h)ds = 1$ and $\int_0^1 (1-s)V_E(s;h)ds = 0$.

We find $E[D_E(W_i, A_i, Y_i; h; f; g)]$ by first conditioning on $(W_i, Y_{E,i}, Y_{U,i})$. As $\mathcal{L}(A_i|W_i, Y_{E,i}, Y_{U,i}) = \mathcal{L}(A_i|W_i)$ by no unmeasured aggregation bias, this conditional mean is an integral with respect to $g_{0,i}(\cdot|W_i)$, and is given by

$$\begin{aligned}
&N_i E_{g_{0,i}}[A_i|W_i] E_f[Y_{E,i}|W_i] \\
&+N_i E_g[A_i|W_i](Y_{E,i} - E_f[Y_{E,i}|W_i]) \int_0^1 \frac{g_{0,i}(s|W_i)}{g(s|W_i)} sV_E(s;h)ds \\
&+N_i E_g[A_i|W_i](Y_{U,i} - E_f[Y_{U,i}|W_i]) \int_0^1 \frac{g_{0,i}(s|W_i)}{g(s|W_i)} (1-s)V_E(s;h)ds.
\end{aligned}$$

Taking a further conditional expectation by conditioning on W, we obtain

$$\begin{aligned}
&N_i E_{g_{0,i}}[A_i|W_i] E_f[Y_{E,i}|W_i] \\
&+N_i E_g[A_i|W_i](E_{f_{0,i}}[Y_{E,i}|W_i] - E_f[Y_{E,i}|W_i]) \int_0^1 \frac{g_{0,i}(s|W_i)}{g(s|W_i)} sV_E(s;h)ds \\
&+N_i E_g[A_i|W_i](E_{f_{0,i}}[Y_{U,i}|W_i] - E_f[Y_{U,i}|W_i]) \int_0^1 \frac{g_{0,i}(s|W_i)}{g(s|W_i)} (1\text{-}s)V_E(s;h)ds.
\end{aligned}$$

If $f = f_{0,i}$ as in the theorem, the second and third terms vanish and the first is equal to $N_i E_{g_{0,i}}[A_i|W_i] E_{f_{0,i}}[Y_{E,i}|W_i]$. If $g = g_{0,i}$ as in the theorem, then $\int_0^1 sV_E(s;h)ds = 1$ and $\int_0^1 (1-s)V_E(s;h)ds = 0$ imply the integrals in the second and third terms respectively are one and zero, so the third term vanishes. With $E_g[A_i|W_i] = E_{g_{0,i}}[A_i|W_i]$, we combine the first and second terms to again obtain $N_i E_{g_{0,i}}[A_i|W_i] E_{f_{0,i}}[Y_{E,i}|W_i]$. We conclude from the law of iterated expectations that $D_E(W_i, A_i, Y_i; h; f; g)$ has mean

$$\begin{aligned}
&E[E[E[D_E(W_i, A_i, Y_i; h; f; g)|W_i, Y_{E,i}, Y_{U,i}]|W_i]] \\
&= E[N_i E_{g_{0,i}}[A_i|W_i] E_{f_{0,i}}[Y_{E,i}|W_i]] \\
&= E[E[N_i A_i Y_{E,i}|W_i]] \text{ by no unmeasured aggregation bias} \\
&= E[N_i A_i Y_{E,i}].
\end{aligned}$$

To show the result for $D_U(W_i, A_i, Y_i; h; g; f)$, we repeat the argument, only now use that $\int_0^1 sV_U(s;h)ds = 0$ and $\int_0^1 (1-s)V_U(s;h)ds = 1$. □

Remark 1. The theorem uses fixed f and g, but our estimator uses fits $\hat{f}$ and $\hat{g}$ depending on the data. It would require finer analysis to examine the bias and variance of actual estimators, or to establish consistency and asymptotic normality. To our knowledge, this project has also not been undertaken for the existing estimators of Section 4.3.

Remark 2. With trivial $\hat{f}$ such that $E_{\hat{f}}[Y_{E,i}|W_i] = E_{\hat{f}}[Y_{U,i}|W_i] = 0$, the theorem implies the resulting Horvitz-Thompson style estimator using inverse density weighting should be appropriate with correct modeling of the regional exposure rate. To our knowledge, such estimators depending entirely on this model are also new to ecological inference.

7 The weight function

Rescaling the weight function does not change the estimator, so we can without loss of generality take $\int_0^1 h(s)ds = 1$. The problem is in determining how much of the unit interval should be heavily weighted.

We recommend choosing the weight function to make $h(A_i)$ small whenever $\hat{g}(A_i|W_i)$ can be large, so extreme $h(A_i)/\hat{g}(A_i|W_i)$ do not destabilize results. However, if $h(A_i)$ is frequently close to zero then the estimator starts depending less on our exposure-rate model fit, and we lose protection if the model is correct. Thus, there is a tradeoff in choosing the weight function.

In our data analysis we used the estimator with $h(a) = n^{-1} \sum_{i=1}^n \hat{g}(a|W_i)$. Section 11 develops an analogy between our problem and one in causal inference, and our choice of h was analogous to what in the latter context is called a stabilizing weight function.

The matrix H in the estimator involves integrals with respect to h, and in our data analysis we computed these with Monte Carlo by viewing h as a density function and making $10,000$ draws. For a single draw, we drew a covariate at random from $W_1, ..., W_n$ and then drew a random exposure rate from our fitted conditional density $\hat{g}$.

8 Relaying uncertainty

The nonparametric bootstrap (Efron, 1979) is probably the simplest way to put a standard error on an exposure-specific disease rate estimate, although results could be sensitive to the assumption of no spatial autocorrelation.

9 Implementation with regression

Although we alluded to models that others have proposed for the conditional distribution of exposure-specific disease rates, they are few in number. On the other hand, there are many techniques for regressing an outcome on explanatory variables, and these can ease implementation.

Notice that the exposure-specific disease rate model fit $\hat{f}$ only enters our estimator through $E_{\hat{f}}[Y_{E,i}|W_i]$ and $E_{\hat{f}}[Y_{U,i}|W_i]$. We do not need to know the entire joint density of $(Y_{E,i}, Y_{U,i})$ given W_i to proceed. Also note that the no unmeasured aggregation bias assumption $\{(Y_{E,i}, Y_{U,i}) \perp A_i|W_i\}$ of (1) implies the regression of Y_i on (W_i, A_i) has the form

$$\begin{aligned}
m(w,a) &= E[Y_i|W_i = w, A_i = a] \\
&= E[A_iY_{E,i} + (1 - A_i)Y_{U,i}|W = w, A = a] \\
&= aE[Y_{E,i}|W_i = w, A_i = a] + (1 - a)E[Y_{U,i}|W_i = w, A_i = a] \\
&= aE[Y_{E,i}|W_i = w] + (1 - a)E[Y_{U,i}|W_i = w] \\
&= E[Y_{U,i}|W_i = w] + (E[Y_{E,i}|W_i = w] - E[Y_{U,i}|W_i = w])\, a.
\end{aligned}$$

The representation tells us that if we have an accurate fit $\hat{m}$ of the regression function, we should be able to approximate the two functions needed for the doubly robust estimator. Because $a \to m(w, a)$ is linear in a, the intercept and slope for the simple linear regression of $[\hat{m}(w, a_1), ..., \hat{m}(w, a_B)]$ on $[a_1, ..., a_B]$ can estimate $E[Y_{U,i}|W_i = w]$ and $E[Y_{E,i}|W_i = w] - E[Y_{U,i}|W_i = w]$. The number of values B and their place in the unit interval are up to the user, but matter less as the regression fit $\hat{m}$ becomes more accurate. The implication of no unmeasured aggregation bias is therefore that we can use any type of method to regress $(Y_1, ..., Y_n)$ on $(W_1, A_1), ..., (W_n, A_n)$. If we do a good job then the doubly robust estimator should perform well. Otherwise, performance will depend on the other of the two models. An even more direct approach to finding the two desired functions is to initially make the regression fit have the form $\hat{m}(w, a) = \hat{m}_0(w) + \hat{m}_1(w)a$.

In our data analysis we regressed disease rates on covariates and exposure rates with the same beta regression procedure discussed in Section 5. To make the linear approximation of $a \to \hat{m}(w, a)$, we drew $a_1, ..., a_{100}$ uniformly in the range of observed regional exposure rates, and as just discussed, performed a simple linear regression of $[\hat{m}(w, a_1), ..., \hat{m}(w, a_{100})]$ on $[a_1, ..., a_{100}]$, where the same 100 random points were used for different values of w.

10 Risk factors for diabetes

A nice feature of ecological inference is that estimator performance can be evaluated on datasets where the truth is known, by aggregating individual-level data and feeding aggregate data to estimators. We didn't know true diabetes rates for those exposed and unexposed to air pollution, but found that our doubly robust estimator could use ecological data to reliably capture other risk factor associations with diabetes.

The Centers for Disease Control and Prevention (CDC) operates a large telephone survey called the Behavioral Risk Factor Surveillance System, and their Web Enabled Analysis Tool allows cross tabulation of the 2005 data concerning individuals over 18 in the 50 U.S. States, Washington D.C., Puerto Rico, and the U.S. Virgin Islands. 2005 prevalence information for risk factors and diseases in the various regions are also given by the CDC at http://apps.nccd.cdc.gov/brfss/index.asp. We used this data to estimate risk factor-specific diabetes rates (type 1 or 2 diabetes, excluding pregnancy-related cases). Ecological inference would only truly have been needed for this problem if the information on risk factors and diabetes had been collected by two different surveillance systems, such as the CDC and the American Diabetes Association.

Risk factors were hypertension (has had high blood pressure?), age (65 or older?), cholesterol (has had high cholesterol), obesity (obese?), exercise (any exercise in the last month?), alcohol use (any drinks in the past 30 days?), income (annual household income over $50,000?), asthma (has ever had asthma?), race (white?), diet (five or more fruits and vegetables daily?), health care (any kind of coverage?), and smoking (a current smoker?) Hence, we had regional data on the prevalence of diabetes and 12 risk factors. When examining a single risk factor, data on other risk factors were considered covariates, and we thus had 11 covariates for each region concerning risk factor prevalences.

We ignored more available regional covariates, including finer categorizations of the risk factors, more information on demographics, more information on physical activity, and data on immunizations. There were also data on arthritis, cardiovascular disease, and disability that we could not figure out how to cross tabulate with diabetes on the CDC website. We did not attempt to obtain any more covariates from different sources.

The number of people over 18 living in each region in 2005 was found from census data at www.census.gov/popest/states/asrh/SC-EST2005-01.html.

As discussed in Section 9, we evaluated $E_{\hat{f}}[Y_{E,i}|W_i]$ and $E_{\hat{f}}[Y_{U,i}|W_i]$ in the doubly robust estimator by making a linear approximation to a beta regression of diabetes prevalences on covariates and exposure rates. We also formed the exposure rate model fit $\hat{g}$ with this beta regression technique, as discussed in Section 5, and chose the stabilizing weight function h as in Section 7. Even though we have considered the regional population size N_i to be part of the covariate vector, it was not included as an explanatory variable in the two beta regressions, because for some reason this led to termination errors with the betareg package. We then formed doubly robust estimates of risk factor-specific disease rates. Along the lines of Section 8, we used 1000 bootstrap resamples to obtain standard errors.

We also applied the popular ecological regression procedure of Goodman (1953, 1959). The method does not use covariates, and performs well when exposure-specific disease rates are nearly constant across regions, but otherwise can be unreliable. The technique fits a linear regression of disease rates on exposure rates, and estimates disease rates under no exposure and exposure with the fitted intercept and slope + intercept.

Both the doubly robust and ecological regression results were mapped to the tomography line as in Section 3 to incorporate deterministic constraints. In no cases did the initial doubly robust exposure-specific disease rate estimates fall outside Duncan-Davis bounds, but initial ecological regression estimates did when estimating diabetes rates with risk factors of high blood pressure, high cholesterol, and avoiding fruit.

King's extended model was not used for this problem, because to our knowledge no existing implementation allowed more than two covariates.

Results are shown in Table 1. Doubly robust point estimates were accurate, and would generally have provided the same interpretation as true values. In contrast, ecological regression overshot the diabetes association for hypertension, age, cholesterol, obesity, and lack of exercise, and falsely concluded that avoiding fruit and smoking were major risk factors.

The bootstrap suggested our method was somewhat prone to high standard errors. We did not follow our recommendation for artificially constraining inverse density weights $1/\hat{g}(A_i|W_i)$, and this may have hurt. It was also unclear whether injecting unnecessary noise had an impact, as we used the Monte Carlo method of Section 9 for the linear approximation to the initial regression fit, and the Monte Carlo integration of Section 7 for matrix H.

	truth	doubly robust	ecological regression
high blood pressure	19.2	17.2 (7.0)	29.7 (0.4)
no high blood pressure	3.7	4.4 (2.5)	0.0 (0.0)
65 or older	17.8	14.5 (9.0)	23.9 (7.8)
under 65	5.8	6.4 (1.8)	4.5 (1.6)
high cholesterol	15.8	8.4 (4.4)	21.7 (1.8)
no high cholesterol	6.0	7.5 (2.5)	0.0 (1.0)
obese	15.7	13.5 (5.5)	28.7 (3.1)
not obese	5.7	5.9 (1.8)	1.0 (1.0)
don't exercise	12.6	11.6 (4.4)	24.6 (1.4)
exercise	6.1	6.5 (1.5)	2.0 (0.5)
don't drink alcohol	11.4	8.9 (1.3)	12.7 (1.5)
drink alcohol	4.6	6.8 (1.1)	3.5 (1.2)
no high income	10.5	8.1 (0.9)	12.1 (0.9)
high income	4.9	7.4 (1.2)	2.0 (1.2)
asthma	10.1	7.6 (7.1)	7.9 (11.4)
no asthma	7.4	7.8 (1.0)	7.8 (1.6)
nonwhite	9.2	9.6 (1.5)	9.9 (0.9)
white	7.1	7.0 (0.6)	6.8 (0.4)
eat fruit	8.3	8.4 (3.2)	0.0 (0.6)
don't eat fruit	7.6	7.6 (1.0)	10.3 (0.3)
health coverage	8.3	7.6 (0.7)	7.1 (0.7)
no health coverage	5.3	8.9 (3.8)	11.5 (3.7)
nonsmoker	8.2	8.8 (1.2)	5.9 (1.8)
smoker	6.3	3.9 (4.6)	15.2 (6.7)

Table 1: Estimates and bootstrap standard errors for the percentage of the U.S. adult population with diabetes, conditional on certain risk factors.

11 Marginal structural model derivation

We have not explained our thought process in obtaining the doubly robust method of Section 6, and it may appear to have been pulled out of the air. Rather, it came from noticing that the ecological inference problem under our assumptions is similar to the problem of having observational individual-level data and wishing to make causal inferences under no unmeasured confounding. We were able to transfer our problem to the latter setting, and modify an existing doubly robust approach.

Perhaps we can clarify the derivation by explaining the transference. While assumptions in the two problems look mathematically similar, they require different types of domain knowledge (Greenland and Robins, 1994).

Temporarily suppose we have data on individuals instead of regions. The data on individual i are (W_i, A_i, Y_i), where W_i is a vector of covariates, A_i is the level of an exposure, and Y_i is an outcome. We can also consider a set of counterfactual outcomes $\{Y_i(a) : a\}$, where $Y_i(a)$ is the outcome that would have occurred if we had intervened and set exposure at $A_i = a$. The observed outcome is $Y_i = Y_i(A)$. Further, we assume the n subjects constitute and independent and identically distributed sample from a population, and no unmeasured confounding in that $\{(Y_i(a) : a) \perp A_i|W_i\}$. Finally, make the assumption that $E[N_i A_i Y_i(a)] = \beta_0 + \beta_1 a$, for N_i in covariate vector W_i.

We have also been writing the observed data in ecological studies as (W_i, A_i, Y_i). If we could strengthen our no spatial autocorrelation assumption so variables for different regions were not only independent but identically distributed, assumptions in the two problems would transfer as follows:

regional covariates $W_i \longleftrightarrow$ subject-level covariates W_i
regional exposure rate $A_i \longleftrightarrow$ subject-level exposure A_i
regional disease rate $Y_i \longleftrightarrow$ subject-level outcome Y_i
convex combination $Y_i(a) = aY_{E,i} + (1-a)Y_{U,i} \longleftrightarrow$ counterfactual $Y_i(a)$
i.i.d. sample of regions $\longleftrightarrow$ i.i.d. sample of subjects
no unmeasured aggregation bias $\longleftrightarrow$ no unmeasured confounding
$(\theta_U, \theta_E - \theta_U) = (E[N_i A_i Y_{U,i}], E[N_i A_i Y_{E,i}] - E[N_i A_i Y_{U,i}]) \longleftrightarrow (\beta_0, \beta_1)$.

Double robustness in the individual-level problem refers to correctly specifying either the regression function $m(w, a) = E[Y_i|W_i = w, A_i = a]$ or the conditional density function g_0 of A_i given W_i, which by Section 9 is analogous to double robustness in ecological inference.

For individual-level causal inference, the assumption $E[Y(a)] = \alpha_0 + \alpha_1 a$ that an individual's counterfactual mean is linear in the counterfactual index is a special type of marginal structural model (Robins, 1997), and doubly robust estimators are available (van der Laan and Robins, 2003).

Unfortunately, in transferring a casual inference method to an ecological inference method, the relevant parameter is not (α_0, α_1) of the marginal structural model, but (β_0, β_1). This is not exactly a standard parameter in causal inference problems, as it depends on the exposure distribution.

It requires only a slight tweaking of marginal structural model machinery to form doubly robust estimates of (γ_0, γ_1) if $E[NR(W)Y(a)] = \gamma_0 + \gamma_1 a$. Our initial thought was to use this machinery with $R(W_i) = E_{\hat{g}}[A_i|W_i]$, as in that case $(\gamma_0, \gamma_1) = (\beta_0, \beta_1)$ if $\hat{g} = g_0$. The method turned out to lack double robustness, but we obtained our procedure after altering it through inspection, replacing several instances of $E_{\hat{g}}[A_i|W_i]$ with A_i.

12 Extensions

Doubly robust estimation is not limited to ecological studies with binary exposure and disease variables. We describe two generalizations, but do not repeat the double robustness proof.

As in Section 3, the data can provide deterministic constraints on our quantities of interest for extensions of the standard ecological inference problem, and in applications we could again map an initial doubly robust estimate to one satisfying these restrictions.

12.1 Multiple exposures

Suppose $A_i^{(1)}, ..., A_i^{(p)}, A_i^{(p+1)} = 1 - \sum_{j=1}^{p} A_i^{(j)}$ are the ith region's rates for $p+1$ exposures, the disease rate is again Y_i, and $Y_i^{(1)}, ..., Y_i^{(p+1)}$ are unobserved exposure-specific disease rates. The assumption of no unmeasured aggregation bias becomes that $(A_i^{(1)}, ..., A_i^{(p)})$ and $(Y_i^{(1)}, ..., Y_i^{(p+1)})$ are independent given covariate W_i. We must estimate parameter $\theta_j = \frac{1}{n} \sum_{i=1}^{n} E[N_i A_i^{(j)} Y_i^{(j)}]$ to approximate the jth total exposure-specific disease rate. Define a weight function $h : [0,1]^p \to \mathbb{R}$. Also define

$$V_j(A_i^{(1)}, ..., A_i^{(p)}; h) = h(A_i^{(1)}, ..., A_i^{(p)}) \left(\beta_{j,0} + \sum_{k=1}^{p} \beta_{j,k} A_i^{(k)} \right),$$

where $(\beta_{j,0}, ..., \beta_{j,p})$ is found by solving a linear system of $p+1$ equations in $p+1$ unknowns given by

$$\int_0^1 ... \int_0^1 s_k V_j(s_1, ..., s_p; h) ds_1 ... ds_p = I(j = k) \text{ for } k = 1, ..., p$$
$$\int_0^1 ... \int_0^1 (1 - \sum_{k=1}^{p} s_k) V_j(s_1, ..., s_p) ds_1 ... ds_p = I(j = p+1).$$

If f_0 is the conditional distribution for $(Y_i^{(1)}, ..., Y_i^{(p+1)})$ given W_i and g_0 is the conditional density for $(A_i^{(1)}, ..., A_i^{(p)})$ given W_i, define

$$\begin{aligned}&D_j(W_i, A_i^{(1)}, ..., A_i^{(p)}, Y_i; h; f; g) =\\ &N_i E_g[A_i^{(j)}|W] \frac{V_j(A_i^{(1)}, ..., A_i^{(p)}; h)}{g(A_i^{(1)}, ..., A_i^{(p)}|W_i)} (Y_i - \sum_{k=1}^{p+1} A_i^{(k)} E_f[Y_i^{(k)}|W_i])\\ &+N_i A_i^{(j)} E_f[Y_i^{(j)}|W_i].\end{aligned}$$

The parameter estimate is $\hat{\theta}_j = n^{-1} \sum_{i=1}^n D_j(W_i, A_i^{(1)}, ..., A_i^{(p)}, Y_i; h; \hat{f}; \hat{g})$, and the doubly robust estimate of the total exposure-specific disease rate for exposure j is $\hat{\mu}_j = n\hat{\theta}_j / \sum_{i=1}^n N_i A_i^{(j)}$.

12.2 Continuous outcomes

In assessing the individual-level association between an exposure and an outcome, the outcome does not have to be a binary indicator such as diabetes status. With an outcome such as height, $Y_i = A_i Y_{E,i} + (1-A_i) Y_{U,i}$ is the observed average height in the ith region, while $Y_{E,i}$ and $Y_{U,i}$ are the unobserved average heights in the exposed and unexposed populations. Across all regions, the average height for the exposed and unexposed are $\sum_{i=1}^n N_i A_i Y_{E,i} / \sum_{i=1}^n N_i A_i$ and $\sum_{i=1}^n N_i(1-A_i) Y_{U,i} / \sum_{i=1}^n N_i(1-A_i)$. The doubly robust estimates can be constructed as before, using identical notation. Our theorem still applies, but we might require different types of models for the conditional distribution of regional exposure-specific heights given regional covariates. It is also straightforward to combine our two extensions and consider continuous outcomes with more than two exposure levels.

13 An open problem

Even though they depend on correctly specifying one model rather than one of two, the existing estimators of Section 4.3 are admirable in that they naturally satisfy the deterministic constraints of Section 3, and there is no need for post hoc adjustment. From this standpoint, our new approach takes steps both forward and backward. An open problem is whether there exist doubly robust estimators naturally meeting constraints. As we now describe, it appears a construction is theoretically possible but difficult to implement.

Recalling D_E and D_U in the doubly robust estimator, define functions of the observed regional data according to

$$\begin{aligned} S(W_i, A_i, Y_i; h; f; g) &= [S_E(W_i, A_i, Y_i; h; f; g), S_U(W_i, A_i, Y_i; h; f; g)]^T \\ S_E(W_i, A_i, Y_i; h; f; g) &= N_i A_i E_f[Y_{E,i}|W_i, A_i, Y_i] - D_E(W_i, A_i, Y_i; h; f, g) \\ S_U(W_i, A_i, Y_i; h; f; g) &= N_i(1 - A_i)E_f[Y_{U,i}|W_i, A_i, Y_i] \\ &- D_U(W_i, A_i, Y_i; h; f; g). \end{aligned}$$

For fits $\hat{f}$ and $\hat{g}$, note that the doubly robust estimate and existing estimate of the form (2) will be equal if $n^{-1}\sum_{i=1}^n S(W_i, A_i, Y_i; h; \hat{f}; \hat{g})$ is zero. It seems we can force this to occur by iteratively updating fit $\hat{f}$ by maximizing likelihood along well-chosen submodels, resembling van der Laan and Rubin's (2006) general targeted maximum likelihood algorithm.

We will not make a long digression into efficiency theory, but viewing observed data (W_i, A_i, Y_i) as a coarsened version of unavailable $(W_i, Y_{E,i}, Y_{U,i})$, the fact that $E_f[S(W_i, A_i, Y_i; h; f; g)|W_i, A_i] = [0, 0]^T$ implies S is orthogonal in a certain Hilbert space to what is known as the augmentation space or T_{CAR} (van der Laan and Robins, 2003; Tsiatis, 2006), and there is a function T such that $E_f[T(W_i, Y_{E,i}, Y_{U,i}; h; f; g)|W_i, A_i, Y_i] = S(W_i, A_i, Y_i; h; f; g)$, or at least such a sequence whose conditional means converge to S in the right Hilbert space. If we make a regular parametric submodel through conditional distribution $\hat{f}$ that has score $T(W_i, Y_{E,i}, Y_{U,i}; h; \hat{f}; \hat{g})$ at $\hat{f}$, such as $\hat{f}_\epsilon(y_e, y_u|w) = (1 + \epsilon^T T(w, y_e, y_u; h; \hat{f}; \hat{g}))\hat{f}(y_e, y_u|w)$, then the score at $\epsilon = [0, 0]^T$ for the corresponding submodel varying observed data distribution $\mathcal{L}(W_i, A_i, Y_i)$ can be made equal to $S(W_i, A_i, Y_i; h; \hat{f}; \hat{g})$. Consider iteratively fitting the submodel with maximum likelihood until there is convergence in that the MLE for ϵ is approximately $[0, 0]^T$. As the score is the gradient of the log likelihood, it has empirical mean zero at the likelihood maximizer. Hence, $n^{-1}\sum_{i=1}^n S(W_i, A_i, Y_i; h; \hat{f}; \hat{g})$ is zero at the updated $\hat{f}$, as desired.

The trouble with this story is that it's easier to describe the algorithm than implement it. For one, we don't have a useful representation for the needed function T, even though it exists, so we state the following problem:

$$\text{what T yields } E_f[T(W_i, Y_{E,i}, Y_{U,i}; h; f; g)|W_i, A_i, Y_i] = S(W_i, A_i, Y_i; h; f; g)?$$

Beyond finding the function, it remains to be seen whether there is practical method for updating $\hat{f}$ so the existing approach of (2) is made doubly robust.

14 Pointers to literature

Although the ecological inference problem has "fascinated scholars for nearly a century" (Cho and Manski, 2008), the difficulties became widely known following Robinson (1952). Isomorphic formulations arise across disciplines, including political science, sociology, economics, marketing, geography, and medical imaging. The best known application is probably in litigation related to the U.S. Voting Rights Act of 1965. Achen and Shively (1995), Schuessler (1999), Freedman (2001), and Cho and Manski (2008) provide overviews of ecological inference, while Wakefield (2008) summarizes ecological studies in epidemiology, and King et al. (2004) discuss recent developments. The deterministic bounds on regional exposure-specific disease rates of Section 3 are due to Duncan and Davis (1953). Methods for point estimation are proposed in Goodman (1953, 1959), Freedman et al. (1991), King (1997), King et al. (1999), Wakefield (2004), Imai et al. (2008), and other papers. The problem's canonical formulation does not have regional covariates, even though they are often available in practice. As we mentioned, many techniques attempt to find exposure-specific disease rates in each region rather than the total exposure-specific disease rates, although our quantities are of primary interest in many applications.

Doubly robust estimators, which fit two models but only need one to be correct, are used in problems involving missing data, causal inference, survival analysis, current status data, and other fields (Scharfstein et al., 1999; van der Laan and Robins, 2003; Tsiatis, 2006). They can be constructed quite generally in coarsened data structures under Heitjan and Rubin's (1991) coarsening at random. Neither of the two models will be exactly correct in practice, but the motivating intuition is that we should expect good performance if at least one model is approximately valid (Bang and Robins, 2005).

There is some disagreement over whether "approximately" is good enough for double robustness to be useful. In the fields where doubly robust procedures are most popular, such debate and estimator comparison typically involves simulations, while real data appear in methodological papers for illustration (Lunceford and Davidian, 2004; Bang and Robins, 2005; Neugebauer and van der Laan, 2005; Kang and Schafer, 2007; Freedman and Berk, 2008). This differs from standard practice in the ecological inference literature. We used the diabetes data mainly to illustrate our approach, but noted that performance could be evaluated by aggregating individual-level data where the truth was known. By developing our estimator in a problem conducive to performance assessment, we hope to shed more light on the phenomenon of double robustness and its practical benefits and limitations.

15 Conclusion

In a critique of King's method used as our starting point, Freedman et al. (1999) claimed "the models are just shots in the dark." The advantage of double robustness is getting to fire two shots instead of one, whether at the witching hour or high noon, and we expect this allows more people to hit their targets. While no procedure can overcome indeterminacies of the ecological inference problem without assumptions, we can now make progress by learning how relevant regional covariates relate to regional exposure rates.

References

[1] Achen, C.H. and Shively, W.P. (1995). *Cross-Level Inference*. University of Chicago Press.

[2] Bailey, R.C., Plummer, F.A., and Moses, F. (2001). Male circumcision and HIV prevention: current knowledge and future research directions. *Lancet Infectious Diseases*, 1:223-231.

[3] Bang, H. and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962-973.

[4] Bongarrts, J., Reining, P., Conant, F. (1989). The relationship between male circumcision and HIV infection in African populations. *AIDS*, 3:373-397.

[5] Bustamonte Simas, A. (2004). The betareg package. Available at ftp://ftp1.sinica.edu.tw/pub1/r-project/R/doc/packages/betareg.pdf.

[6] Centers for Disease Control and Prevention (2005). *Behavioral Risk Factor Surveillance System Survey Data.* U.S. Department of Health and Human Services. Atlanta, GA.

[7] Cho, W.K.T. (1998). Iff the assumption fits... :A Comment on the King Ecological Inference Solution. *Political Analysis*, 7:143-163.

[8] Cho, W.K.T. and Manski, C.F. (2008). Cross-level/ecological inference. *Oxford Handbook of Political Methodology,* Oxford University Press, 547-569. Janet Box-Steffensmeier, Henry Brady, and David Collier, eds.

[9] Duncan, O.D. and Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review*, 18:665-666.

[10] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1-26.

[11] Ferrari, S.L.P. and Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, 10:1-18.

[12] Freedman, D.A. (2001). Ecological inference and the ecological fallacy. International Encyclopedia of the Social & Behavioral Sciences. Elsevier, 6:402730. Neil J. Smelser and Paul B. Baltes, eds.

[13] Freedman, D.A. and Berk, R.A. (2008). On weighting regressions by propensity scores. *Evaluation Review*, 32:392409.

[14] Freedman, D.A., Klein, S.P., Sacks, J., Smyth, C.A., and Everett, C.G., (1991). Ecological regression and voting rights. *Evaluation Review* 15:673-711.

[15] Freedman, D.A., Klein, S.P., Ostland, M., and Roberts, M.R. (1999). Response to King's comment. *Journal of the American Statistical Association*, 94:35557.

[16] Goodman, L.A. (1953). Ecological regressions and the behavior of individuals. *American Sociological Review*, 18:663-664.

[17] Goodman, L.A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, 64:610-625.

[18] Greenland, S. and Robins, J.M. (1994). Invited commentary: Ecologic studies-biases, misconceptions, and counterexamples. *American Journal of Epidemiology,*, 139:747-760.

[19] Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics*, 19:2244-2253.

[20] Imai, L, and Lu, Y. (2005). An incomplete data approach to the ecological inference problem. *Working paper; Princeton University.* Available at http://imai.princeton.edu/research/files/coarse.pdf.

[21] Imai, K., Lu, Y., and Strauss, A. (2008). Bayesian and likelihood inference for 2 x 2 ecological tables: an incomplete data approach. *Political Analysis*, 16:41-69.

[22] Kang, J.D.Y. and Schafer, J.L. (2007). Demystifying double robustness: A comparison for alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:523-580.

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

[24] King, G., Rosen, R., and Tanner, M.A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods and Research*, Vol. 28, No. 1.

[25] King, G., Rosen, O., and Tanner, M.A, eds. (2004). *Ecological Inference: New Methodological Strategies.* Cambridge University Press, NY.

[26] Lockwood, A. Diabetes and air pollution (2002). *Diabetes Care*, 25:1487-1488.

[27] Lunceford, J.K. and Davidian, M. (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23:2937-2960.

[28] Marquez, E.B., Diaz, B.R., and Gurian, P.L. (2004). Understanding the associations between statewide diabetes prevalence and air pollution emissions. *Diabetes Care*, 27:1515-1517.

[29] Moses, S., Bradley, J., Nagelkerke, N., Ronald, A., Ndinya-Anchola, J., and Plummer, F (1990). Geographical patters of male circumcision practices in Africa: association with HIV seroprevalence. *International Journal of Epidemiology*, 19:693-697.

[30] Neugebauer, R. and van der Laan, M.J. (2005). Why prefer double robust estimators in causal inference? *Journal of statistical planning and inference*, 129:405-426.

[31] Nicolich, M.J. (2002). Diabetes and the state capital. *Diabetes Care*, 25:2367.

[32] Robins J.M. (1997). Marginal Structural Models. Proceedings of the American Statistical Association, Section on Bayesian Statistical Science, 1-10.

[33] Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15:351-357.

[34] Scharfstein D.O., Rotnitzky, A., and Robins, J.M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association*, 94:1096-1120.

[35] Schuessler, A.A. (1999). Ecological Inference. *Proceedings of the National Academy of Science USA*, 96:10578-10581.

[36] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data.* Springer Science + Business Media, LLC.

[37] van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* Springer-Verlag, NY.

[38] van der Laan, M.J. and Rubin, D.B. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics*, Vol. 2, Article 11.

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

[40] Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health*, 29:75-90.