6-2012

Perils and Prospects of Using Aggregate Area Level Socioeconomic Information as a Proxy for Individual Level Socioeconomic Confounders in Instrumental Variables Regression

Jesse Yenchih Hsu
University of Pennsylvania

Scott A. Lorch
University of Pennsylvania

Dylan S. Small
University of Pennsylvania

Recommended Citation

Hsu, J., Lorch, S. A., & Small, D. S. (2012). Perils and Prospects of Using Aggregate Area Level Socioeconomic Information as a Proxy for Individual Level Socioeconomic Confounders in Instrumental Variables Regression. *Health Services and Outcomes Research Methodology, 12* (2), 119-140. http://dx.doi.org/10.1007/s10742-012-0095-9

Perils and Prospects of Using Aggregate Area Level Socioeconomic Information as a Proxy for Individual Level Socioeconomic Confounders in Instrumental Variables Regression

Abstract
A frequent concern in making statistical inference for causal effects of a policy or treatment based on observational studies is that there are unmeasured confounding variables. The instrumental variable method is an approach to estimating a causal relationship in the presence of unmeasured confounding variables. A valid instrumental variable needs to be independent of the unmeasured confounding variables. It is important to control for the confounding variable if it is correlated with the instrument. In health services research, socioeconomic status variables are often considered as confounding variables. In recent studies, distance to a specialty care center has been used as an instrument for the effect of specialty care vs. general care. Because the instrument may be correlated with socioeconomic status variables, it is important that socioeconomic status variables are controlled for in the instrumental variables regression. However, health data sets often lack individual socioeconomic information but contain area average socioeconomic information from the US Census, e.g., average income or education level in a county. We study the effects on the bias of the two stage least squares estimates in instrumental variables regression when using an area-level variable as a controlled confounding variable that may be correlated with the instrument. We propose the aggregated instrumental variables regression using the concept of Wald's method of grouping, provided the assumption that the grouping is independent of the errors. We present simulation results and an application to a study of perinatal care for premature infants.

Keywords
aggregation, casual inference, instrumental variables, proxy variables, Wald's grouping method

Disciplines
Statistics and Probability

Perils and Prospects of Using Aggregate Area Level Socioeconomic Information As a Proxy for Individual Level Socioeconomic Confounders in Instrumental Variables Regression

Jesse Yenchih Hsu · **Scott A. Lorch** · **Dylan S. Small**

Received: date / Accepted: date

Abstract A frequent concern in making statistical inference for causal effects of a policy or treatment based on observational studies is that there are unmeasured confounding variables. The instrumental variable method is an approach to estimating a causal relationship in the presence of unmeasured confounding variables. A valid instrumental variable needs to be independent of the unmeasured confounding variables. It is important to control for the confounding variable if it is correlated with the instrument. In health services research, socioeconomic status variables are often considered as confounding variables. In recent studies, distance to a specialty care center has been used as an instrument for the effect of specialty care vs. general care. Because the instrument may be correlated with socioeconomic status variables, it is important that socioeconomic status variables are controlled for in the instrumental variables regression. However, health data sets often lack individual

J. Y. Hsu (✉) · D. S. Small

Department of Statistics, Wharton School, University of Pennsylvania,

400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6302

Tel.: (215) 746-8565 / Fax: (215) 898-1280

E-mail: hsu9@wharton.upenn.edu

J. Y. Hsu · S. A. Lorch

Center for Outcomes Research, The Children's Hospital of Philadelphia

S. A. Lorch

Department of Pediatrics, School of Medicine, University of Pennsylvania,

Division of Neonatology, The Children's Hospital of Philadelphia,

E-mail: LORCH@email.chop.edu

D. S. Small

E-mail: dsmall@wharton.upenn.edu

socioeconomic information but contain area average socioeconomic information from the US Census, e.g., average income or education level in a county. We study the effects on the bias of the two stage least squares estimates in instrumental variables regression when using an area-level variable as a controlled confounding variable that may be correlated with the instrument. We propose the aggregated instrumental variables regression using the concept of Wald's method of grouping, provided the assumption that the grouping is independent of the errors. We present simulation results and an application to a study of perinatal care for premature infants.

Keywords Aggregation · Causal inference · Instrumental variables · Proxy variables · Wald's grouping method

1 Introduction

In health science, researchers and policy makers are often interested in the causal effects of a treatment or a policy on a health outcome. In a randomized trial — coin flipping decides whether the next subject is assigned to treatment or not, the causal effects of a treatment can be correctly estimated from a simple comparison between the treated group and control group. However, a randomized trial is not always feasible or ethical to conduct in most circumstances. Instead, researchers can only obtain data from an observational study; i.e., treatment assignment is not decided by experimental control. In an observational study, the causal relationship between treatment and outcome is usually confounded by a set of covariates called confounding variables. Controlling for all the confounding variables in a conventional analysis such as ordinary least squares regression would provide a valid estimation for the causal effect of treatment. With the presence of unmeasured confounding variables, the ordinary least squares regression of the outcome on the treatment controlling for measured confounding variables cannot provide correct estimation for the causal effects of the treatment. To obtain the correct estimation from an observational study with the presence of unmeasured confounding variables, an alternative method is needed; e.g. instrumental variables regression.

1.1 Instrumental variables and confounding

The instrumental variables (IV) regression is an approach to overcoming the problem of unmeasured confounding variables in observational studies. A valid IV is a variable that (1) is associated with the received treatment; (2) has no direct effect on the outcome of interest other than through its effect on the received treatment; and (3) is independent of unmeasured confounding variables given measured confounding variables. The idea of the IV method is to extract variation in the treatment

that is independent of the unmeasured confounding variables and use this bias-free variation to estimate the effect of the treatment on the outcome. As compared to the conventional analysis that requires researchers to control for all confounding variables, the IV method allows researchers to obtain consistent estimates for the causal parameters after only controlling for the confounding variables that are correlated with the instrumental variable. In the presence of unmeasured confounding variables that are not correlated with IV, the IV method provides consistent estimation for the causal parameters. For more discussions of IV, see Angrist, Imbens and Rubin (1996), Angrist and Krueger (2001), Abadie (2003), and Hernán and Robins (2006).

Suppose we have an IV regression that would meet all the assumptions; specifically, we have controlled for all variables that are correlated with both the IV and the outcome, i.e., confounding variables for the IV-outcome relationship. There may be confounding variables for the treatment-outcome relationship that we have not controlled for. The question we consider in this paper is the following. Suppose that an individual measurement of a confounding variable that is supposed to be controlled for in the IV regression is not available but an aggregated form of the variable is available, can we use its aggregate form in the IV regression and still obtain consistent estimates for the causal parameters? We will show that using the aggregated confounding variable in the usual IV regression would violate the IV criteria (3) that the IV must be independent of unmeasured confounding variables. On the contrary, we will show that a consistent estimate of the causal parameters can be obtained by using an aggregated IV regression with aggregate variables completely replacing all individual level variables. More importantly, the estimate of the causal parameters obtained from the aggregated IV regression will have the same interpretation as in the individual IV regression (without aggregation), provided the aggregation/grouping does not depend on errors.

1.2 A regionalization of perinatal care study

In studies of perinatal health and outcomes, hospitals vary in their ability to care for premature infants. ‘Regionalization of perinatal care’ is the concept of transferring a mother or infant within a geographic area or region based on an infant’s potential illness, the need for resuscitation immediately after delivery, and the capabilities of the hospital to manage the expected needs of the mother and infant. The system tends to transfer higher-risk mothers and babies to specialized perinatal centers to optimize the outcome of these infants. Because these infants have a higher baseline risk of poor outcomes, there are differences in the casemix between hosepitals with greater capabilities to manage sick mothers and infants and those hospitals with fewer capabilities. The difference in outcomes between delivering in a facility with greater capabilities compared to

one with lesser capabilities determines the value of perinatal care for a given geographic area [Lorch et al., 2010]. A crude comparison of health outcomes by levels of hospital does not provide clear evidence whether regionalization is effective or not, because populations, or casemix, are not comparable among hospitals with different levels of capability [Pearl, 2000]. A recent study suggested that the protective effect of delivering at a high-level hospital on rates of mortality were under-reported in studies that failed to control for unmeasured differences in casemix [Lorch et al., 2012]. Controlling for this difference in casemix is important to obtain accurate estimates of the value of perinatal care in a given area.

Our motivating neonatology study contains the data that describe all premature births in the State of Missouri in the years 1993–2003. The data combine information from birth and death certificates, which include information on the mother's sociodemographic factors such as education and age and the infant's birth weight and gestational age, and the UB-92 form, which hospitals provide when submitting bills to Medicare or third-party payers for reimbursement for health services provided. The UB-92 form contains information on the diagnoses and procedures experienced by the infant or mother during the hospital stay and are used to define many of the complications of pregnancy used as confounding variables in this study, as well as the neonatal diagnoses used as outcomes in this study. According to the American Academy of Pediatrics [American Academy of Pediatrics, Committee on Fetus and Newborn, 2004], hospitals are classified by six levels of neonatal intensive care units (NICUs) of increasing technical expertise and capability: 1, 2, 3A, 3B, 3C, and 3D. For example, level 1 hospitals can only provide basic neonatal care and level 3D hospitals can provide all surgeries without restrictions. Following Rogowski et al. (2004), we define a NICU as high-level if it delivers an average of at least 50 preterm babies per year and if the NICU has a level of 3A–3D, while a low-level NICU is any unit that fails to meet both criteria. The outcomes of infants were assigned to their delivery hospital, regardless of future transfer of care to other hospitals as in other previous research [Phibbs et al., 2007, Cifuentes et al., 2002]. While data exist on the effect of delivering at a high-level NICU on lowering the risk of mortality for premature infants, we are particularly interested in studying the causal effect on the number of complications of a premature baby being delivered at a high-level NICU compared to a low-level NICU. A major difficulty in studying the effect of being delivered at a high-level NICU on infants' health outcome is that infants with higher risk of poor outcomes are most likely to be sent to high-level NICUs; i.e., being delivered at a high-level NICU is not a random assignment. There are variables that physicians use to determine where mothers deliver their babies, and these variables are likely to affect infants' health outcome as well. These variables that occur before birth or at the time of birth include sociodemographic factors, such as mothers age, race, insurance status, and educational status; complications of pregnancy, including diabetes and hypertension; congenital anomalies diagnosed prior to birth; and the infants birth weight

and gestational age. Although some variables describing the infants' health prior to being delivered are measured in the data, we are still missing other important variables that are known to the physicians caring for the expectant mother such as fetal heart tracing results, the severity of specific comorbidities such as mother's hypertension, the compliance of the mother to medical treatment and mother's medical history with the physicians. Because we are unable to adjust for these variables and physicians likely use these variables in determining whether to suggest to a mother that she delivers at a high-level NICU (i.e., there exists unmeasured confounding variables), even if high-level NICUs are saving lives, the number of infant complications at these hospitals might be higher (not lower) than that at low-level NICUs because their patient populations are sicker.

Instrumental variable regression with a valid IV provides consistent estimation for the causal effects of being delivered at a high-level NICU on the number of infant complications. In IV regression, controlling for individual level confounding variables that are correlated with the IV is important to ensure that the extracted variation in the treatment is independent of unmeasured confounding variables. We believe that a mother's socioeconomic status (SES) variables are confounding variables for the effects of being delivered at a high-level NICU on an infant's health outcome, and we also believe that a mother's SES variables are correlated with the infant's risk of a complication. For example, lower educational status may be associated with how well a complication of pregnancy such as diabetes is controlled, which may result in a sicker infant at higher risk of a complication after delivery. Meanwhile, SES variables are often considered as confounding variables that are correlated with the IV when the IV is based on where the person lives or obtains medical services [Brookhart and Schneeweiss, 2007]. Therefore, it is necessary to control for SES variables in using IV regression. Often, SES variables, such as family income, education, occupation, etc., at individual level are not available in hospital records or medical charts, or it is difficult to collect these variables from individuals. One commonly used strategy is to replace individual SES variables with aggregate census-based SES variables using a person's residential zip code or county. Using aggregate proxies to replace unavailable individual data is frequently adapted by economists and education researchers [Card and Krueger, 1992]. There is a growing tendency of health researchers to use this approach [Geronimus and Bound, 1998, Krieger et al., 2003a, Krieger et al., 2003b].

For our neonatology example, we used excess travel time that the difference in travel times from a mother's residential zip code to the nearest high-level NICU and the mother's residence to the nearest low-level NICU as an IV [Phibbs and Robinson, 1993, McClellan et al., 1994, Baiocchi et al., 2010]. Excess travel time is negative if the closest hospital has a high-level NICU. For excess travel time to be a valid IV conditional on the measured confounding variables, it must (1) be correlated with whether a mother delivers at a high-level NICU, which because a mother typically obtains

prenatal care from and would prefer to deliver at a close by hospital [Phibbs et al., 1993]; (2) not have a direct effect on the outcome through ways other than the pathway between level of the NICU and the outcome, which is reasonable because a nearby high-level NICU presumably only affects a baby's outcome if the baby receives the care and is delivered at that hospital; and (3) be independent of unmeasured confounding variables, which is plausible after controlling for measured characteristics that predict where people live.

1.3 Perils and Prospects

In this paper, we raise the concern of aggregating a measured confounding variable that is also correlated with IV in making inference about the causal effect of a treatment in an observational study using IV regression. Suppose there are two kinds of confounding variables: one is associated with IV; the other is not. It is important to control for all those confounding variables that are correlated with IV. For example, in our motivating neonatology study, a mother's SES variable is an example of a confounding variable that is correlated with IV (e.g., X in Figure 1). The SES variable needs to be controlled in IV regression because of its correlation with IV. When the SES variable is aggregated or its aggregate form is used, it creates an error term — aggregation error (not observed). If the individual SES variable is correlated with IV, most likely, both the aggregate SES variable and the aggregation error would be correlated with the IV. When we use the aggregate variable in IV regression, we leave the aggregation error unobserved. This aggregation error is unmeasured and is correlated with the IV. Thus, we may fail to control for all confounding variables that are correlated with the IV. Therefore, the use of aggregate confounding variables violates the IV assumption that the IV must be uncorrelated with unmeasured confounding variables. The IV method will result in inconsistent estimation for the causal parameters in the presence of such aggregation error; specifically, a violation of the IV assumption that an IV must be independent of unmeasured confounding variables given measured confounding variables.

We propose the use of Wald's method of grouping [Wald, 1940] to obtain consistent estimates for the causal parameters in IV regression when a confounding variable is only available in aggregate form. Wald's method involves dividing the observations into groups, averaging the outcomes and covariates within the groups, and conducting the analysis on the group averages. In the regionalization of perinatal care study, for example, zip code or county could be used for grouping. When only the aggregate confounding variable is available (note that we specifically focus on the confounding variable that is correlated with the IV), we can replace both those variables where individual patient level data are available and those variables where only aggregate data are available with mean values from geographic group, and fit an IV regression using

these mean values. The grouped-data estimators are consistent in the presence of aggregation errors [Angrist, 1991]. The variance of the grouped estimators are greater than (or equal to) the variance of the estimators obtained from ungrouped data, because grouping will not gain any information about parameters [Prais and Aitchison, 1954].

This paper is organized as follows. In Section 2, using potential outcomes, we introduce a model framework for IV and its inference. We closely examine the problem of aggregation in IV regression in Section 3, when the aggregate variable is a confounding variable in an observational study. In Section 4, we present our idea of using Wald's method of grouping to solve the problem of aggregation in IV regression in observational studies. We investigate the effect of aggregation in IV regression using simulations in Section 5, and illustrate our method using a study of regionalization of perinatal care in Section 6. Section 7 concludes with the discussion of our findings.

2 Framework

In this section we introduce notation and assumptions used in this paper and describe how a valid IV enables identification of the model.

2.1 Notation

To define the causal effect of a policy/treatment, we use the potential outcomes approach [Neyman, 1990, Rubin, 1974]. A potential outcome is the outcome that would have been observed had a subject been assigned to a treatment action; e.g., for a binary treatment, treated or not treated, a subject has two potential outcomes even though only one would be observed. Let Y denote an outcome of interest and D denote a policy/treatment variable. For instance in the regionalization of perinatal care study, Y is the infant's number of complications and D a binary variable for whether (1) or not (0) the baby was delivered at a high-level NICU. Let $y_{ij}^{(d^*)}$ denote the outcome that would be observed for unit i in area j if such unit's level of D was set to d^*. This notation explicitly assumes the restriction that will be described in Section 2.2, Assumption 3 below. Let $y_{ij}^{obs} := y_{ij}$ and $d_{ij}^{obs} := d_{ij}$ be the observed values of Y and D for unit i in area j. Each unit has a vector of potential outcomes, $\{y_{ij}^{(1)}, y_{ij}^{(0)}\}$, but we observe only one potential outcome, $y_{ij} = y_{ij}^{(d_{ij})}$.

Ordinary least squares (OLS) regression provides a way to estimate β_1 by regressing y_{ij} on d_{ij}. The estimator $\hat{\beta}_1^{OLS}$ is consistent if d_{ij} were randomly assigned. This is usually not true in an observational study, and $\hat{\beta}_1^{OLS}$ is not a consistent estimator for β_1. One strategy to overcome this is to collect data on all confounding variables X and regress y_{ij} on d_{ij} controlling for all confounding variables. However, it is not guaranteed that all confounding variables will be available in

hospital records or medical charts; i.e., there exists unmeasured confounding variables U that are associated with both D and Y. Another strategy is to use IV regression. Let Z denote the IV and z_{ij} be the observed IV for unit i in area j. The idea of IV regression is to use instrument Z to extract variation in treatment D that is independent of unmeasured confounding U and to use only this part of the variation in D to estimate the causal relationship between D and Y. In the next section, we provide assumptions that enable a valid IV to provide a consistent estimator of β_1.

2.2 Assumptions

In this paper, we use some of the assumptions in Angrist, Imbens and Rubin (1996) and Holland (1988). To describe the assumptions, we introduce the following additional potential variables: $y_{ij}^{(d,z)}$ which is the outcome unit i in area j would experience if she were assigned level d for the treatment and level z for the instrument, and $d_{ij}^{(z)}$ is the level of d that unit i in area j would have if he/she were assigned level z of the instrument.

Assumption 1 *Stable Unit Treatment Value Assumption (SUTVA) by Rubin (1986).*

The SUTVA assumption states that the potential outcomes for unit i in area j depend only on the level of D for unit i in area j and not on the levels of D for other units.

Assumption 2 *Ignorability of the Instrument.*

Conditional on the measured confounding variables x_{ij}, the observed value of the instrument Z is independent of the set of all potential variables $\{y_{ij}^{(1,z)}, y_{ij}^{(0,z)}, d_{ij}^{(z)}\}$, $z \in \mathscr{Z}$ where $\mathscr{Z}$ denotes the set of all possible values of the instrument Z. This assumption means that the instrument is "as good as randomly assigned" once we condition on the measured confounding variables X and will be satisfied if the instrument is independent of all unmeasured confounding variables given the measured confounding variables X.

Assumption 3 *Exclusion Restriction.*

This assumption is that $y_{ij}^{(1,z)} = y_{ij}^{(1,z')}$ and $y_{ij}^{(0,z)} = y_{ij}^{(0,z')}$ for all z, z'. The assumption allows us to write $y_{ij}^{(d,z)}$ as $y_{ij}^{(d)}$. In words the exclusion restriction assumes that any effect of Z on Y must be through an effect of Z on D; i.e., no direct effect on the outcome of interest other than through its effect on the received treatment.

Assumption 4 *Nonzero Average Causal Effect of Z on D.*

This assumption requires Z to have some effect on the average probability of treatment. In other words, the instrument Z is associated with the treatment D.

Assumption 5 *An Additive, Linear Constant-Effect Causal Model for Potential Outcomes [Holland, 1988]*

This model assumption we make implies that $y_{ij}^{(d^*)} = y_{ij}^{(0)} + \beta_1 d^*$. Therefore, the causal effect of the treatment D — our parameter of interest — is $\beta_1 = y_{ij}^{(1)} - y_{ij}^{(0)}$.

In Section 3, we will examine the effect of violation of Assumption 2 due to aggregation, provided Assumptions 1, 3, 4 and 5 hold.

2.3 Instrumental variables regression models

In an observational study of the regionalization of perinatal care, let X denote an observed confounding variable; e.g., mother's family income, and U denote an unmeasured confounding variable that a physician used to determine if a mother would deliver her baby at a high-level NICU but is not available in the hospital records; e.g., a variable describing a baby's health prior to being delivered such as fetal heart tracing results. Let Z be a valid IV; e.g., the excess travel time given by the difference in travel times from a mother's residential zip code to the nearest high-level NICU and the mother's residence to the nearest low-level NICU. Figure 1 depicts the idea of causal diagram in the context of regionalization of perinatal care study. We consider a linear model as follows

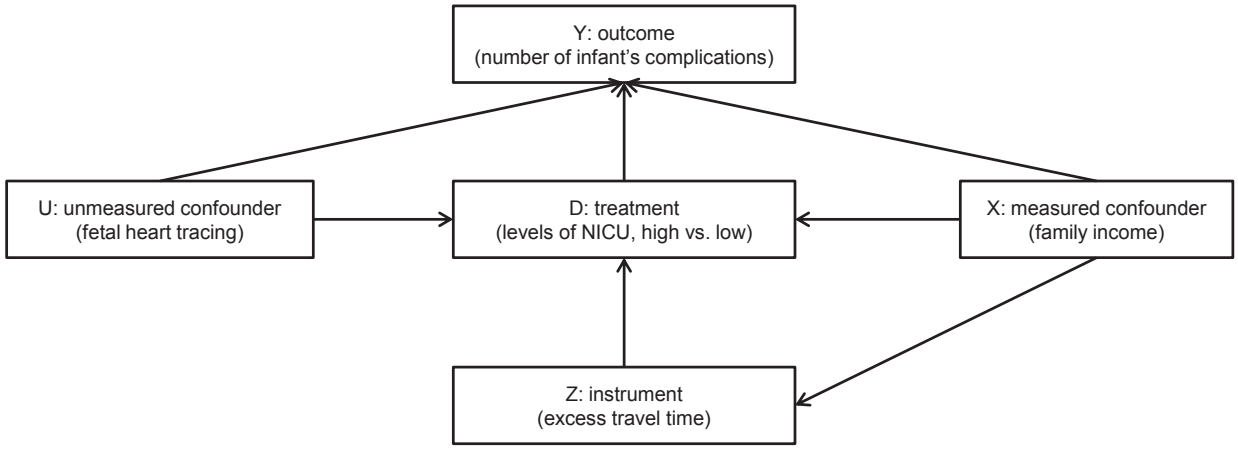

Fig. 1 Causal diagram in regionalization of perinatal care study

$$y_{ij}^{(d^*)} = \beta_1 d^* + \beta_2 x_{ij} + \varepsilon_{ij}, \quad \text{and} \quad E(\varepsilon_{ij} \mid x_{ij}, z_{ij}) = 0, \tag{1}$$

where i indexes subjects and j indexes areas. The error term ε_{ij} can be considered as a composite of random disturbance and unmeasured confounding U. The model for the observed data is

$$y_{ij} = \beta_1 d_{ij} + \beta_2 x_{ij} + \varepsilon_{ij}, \tag{2}$$

where $E(\varepsilon_{ij} \mid x_{ij}, z_{ij}) = 0$ and $Var(\varepsilon_{ij} \mid x_{ij}, z_{ij}) = \sigma_\varepsilon^2$. Random variable ε_{ij} is independently and identically distributed (i.i.d.) for ith subject in jth area. As we mentioned earlier, an OLS estimator $\hat{\beta}_1^{OLS}$ obtained from (2) will not be consistent in the presence of unmeasured confounding U in an observational study. In other words, the treatment d_{ij} is endogenous in (2), or d_{ij} is correlated with ε_{ij} (i.e., $E(d_{ij}\varepsilon_{ij}) \neq 0$). One solution for the problem of unmeasured confounding variables is to use the IV method.

In our motivating neonatology study in Section 1.2, a mother's choice of delivering her baby at a high-level or low-level NICU is not randomly assigned. Researchers are aware of missing important variables that describe the risk of the baby; these are major confounding variables. To overcome the problem of unmeasured sickness of the baby, we could use IV regression with excess travel time as an IV. To assure that excess travel time is a valid IV, we assume that (1) excess travel time is correlated with whether a mother delivers at a high-level NICU (i.e., a mother who lives closer to a high-level NICU will have higher chance to deliver her baby there); (2) excess travel time does not have direct effect on outcome (i.e., no effect other than through the pathway between the level of the NICU and the outcome); and (3) excess travel time is independent of unmeasured confounding variables.

A valid IV Z provides a way to extract exogeneity in treatment D when unmeasured confounding U exists. The two stage least squares (TSLS) method [Theil, 1971] is a common approach to making inference about the treatment effect β_1 in (2) using IV. In TSLS, we first regress d_{ij} on (z_{ij}, x_{ij}) using OLS to obtain $\hat{E}(d_{ij} \mid z_{ij}, x_{ij}) = \hat{d}_{ij}$, and then regress y_{ij} on $(\hat{d}_{ij}, x_{ij})$ using OLS to estimate β_1. From (2), we can write

$$\begin{aligned} y_{ij} &= \beta_1 d_{ij} + \beta_2 x_{ij} + \varepsilon_{ij} \\ &= \beta_1 \hat{d}_{ij} - \beta_1 \hat{d}_{ij} + \beta_2 x_{ij} + \beta_1 d_{ij} + \varepsilon_{ij} \\ &= \beta_1 \hat{d}_{ij} + \beta_2 x_{ij} + \varepsilon_{ij}^*, \text{ where } \varepsilon_{ij}^* = \varepsilon_{ij} + \beta_1 (d_{ij} - \hat{d}_{ij}). \end{aligned} \tag{3}$$

Thus, by regressing d_{ij} on z_{ij} and x_{ij} first, we can create a regressor $\hat{d}_{ij}$ that is exogenous to ε_{ij}^* in (3), or

$$E(\hat{d}_{ij}\varepsilon_{ij}^*) = E\{E(\hat{d}_{ij}\varepsilon_{ij}^* \mid x_{ij}, z_{ij})\} = E\{\hat{d}_{ij} E(\varepsilon_{ij}^* \mid x_{ij}, z_{ij})\} = 0.$$

We can also obtain an equivalent estimator for the causal effect β_1 through the estimating equation based on $E(z_{ij}\varepsilon_{ij}) = 0$ that is implied by z_{ij} being a valid IV. Let $\mathbf{Y} = (y_{11}, \ldots, y_{n_j j})^T$, $\mathbf{X} = (x_{11}, \ldots, x_{n_j j})^T$, $\mathbf{D} = (d_{11}, \ldots, d_{n_j j})^T$, $\mathbf{Z} = (z_{11}, \ldots, z_{n_j j})^T$,

$\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{n_j j})^T$, $\mathbf{W} = [\mathbf{D}, \mathbf{X}]$, $\mathbf{A} = [\mathbf{Z}, \mathbf{X}]$, and $\beta = (\beta_1, \beta_2)^T$. Model (3) can be written as $\mathbf{Y} = \hat{\mathbf{W}}\beta + \varepsilon^*$, where $\hat{\mathbf{W}} = [\hat{\mathbf{D}}, \mathbf{X}]$ and $\varepsilon^* = (\varepsilon^*_{11}, \ldots, \varepsilon^*_{n_j j})^T$. TSLS estimators $\hat{\beta}$ from (3), therefore, are equivalent to the solution of the following estimating equations,

$$\mathbf{A}^T(\mathbf{Y} - \mathbf{W}\hat{\beta}) = \mathbf{0}. \tag{4}$$

The estimator $\hat{\beta} = (\mathbf{A}^T\mathbf{W})^{-1}\mathbf{A}^T\mathbf{Y}$ from (4) will be a consistent estimator as long as $\mathbf{Z}$ is a valid instrument, and

$$\begin{aligned} \hat{\beta} &= (\mathbf{A}^T\mathbf{W})^{-1}\mathbf{A}^T\mathbf{Y} \\ &= (\mathbf{A}^T\mathbf{W})^{-1}\mathbf{A}^T(\mathbf{W}\beta + \varepsilon) \\ &= \beta + (\mathbf{A}^T\mathbf{W})^{-1}\mathbf{A}^T\varepsilon. \end{aligned} \tag{5}$$

The consistency of $\hat{\beta}$ follows the fact that the term $\mathbf{A}^T\varepsilon$ in (5) converges in probability to zero, since $\mathbf{A}$ is uncorrelated (exogenous) with unmeasured error ε. The asymptotic variance of $\hat{\beta}$ is $\sigma^2_\varepsilon(\mathbf{A}^T\mathbf{W})^{-1}\mathbf{A}^T\mathbf{A}(\mathbf{W}^T\mathbf{A})^{-1}$.

3 Problem with aggregation

In observational studies, it is important that the confounding variable is measured and controlled for. Often, SES variables (e.g., family income, education, occupation, etc.) play an important role as measured confounding variables in public health research, such as X in Figure 1. However, one of the biggest challenges is that most health related databases lack individual level SES data. The census-based SES data from the US census are suggested to overcome this difficulty [Krieger, 1992, Krieger et al., 2003a, Krieger et al., 2003b]. The validity of using census-based SES data to proxy individual SES data has been discussed in literature [Geronimus et al., 1996, Geronimus and Bound, 1998]. In this section, we discuss the bias that arises from using aggregate SES variables as controlled confounding variables in IV regression when the interest is on estimating the causal effect of a treatment. In Section 4, we propose a analytical method that could reduce the bias due to aggregation and provide consistent estimation for the causal effect of a treatment.

Following the notation used in Section 2, let x_j^{agg} be an aggregate measurement of the observed confounding variable x_{ij} in the jth area; e.g., average family income within a county. Let h_{ij} be the aggregation error, where $h_{ij} = x_j^{agg} - x_{ij}$; i.e., the difference between subject ij's income and aggregated income in the jth county. To illustrate the problem of aggregating

measured confounding variable X into X^{agg}, we can rewrite model (2) as

$$\begin{aligned} y_{ij} &= \beta_1 d_{ij} + \beta_2 x_{ij} + \varepsilon_{ij} \\ &= \beta_1 \hat{d}'_{ij} + \beta_2 (x_j^{agg} + h_{ij}) + \beta_1 (d_{ij} - \hat{d}'_{ij}) + \varepsilon_{ij} \\ &= \beta_1 \hat{d}'_{ij} + \beta_2 x_j^{agg} + \varepsilon'_{ij}, \quad \text{where } \varepsilon'_{ij} = \varepsilon_{ij} + \beta_2 h_{ij} + \beta_1 (d_{ij} - \hat{d}'_{ij}). \end{aligned} \tag{6}$$

When the aggregated confounding variable x_j^{agg} is used, $\hat{\beta}'_1$, the TSLS estimator for the causal effect β_1, is obtained by (6), where $\hat{d}'_{ij}$ is obtained by regressing d_{ij} on z_{ij} and x_j^{agg} using OLS. This TSLS estimator could give misleading inferences when the aggregated variable is a confounding variable of the IV-outcome relationship. The TSLS method uses part of the variability in treatment that is uncorrelated with unmeasured confounding to consistently estimate the treatment effect. When X is aggregated into X^{agg}, the part that we use in the second stage of the TSLS method is no longer uncorrelated with unmeasured confounding U because of the existence of aggregation error H. Figure 2 shows the problem of aggregation graphically. The correlation between instrument Z and aggregation error H violates the assumption that an instrument has to be uncorrelated with unmeasured confounding. The problem of aggregation can also be shown through estimating equations.

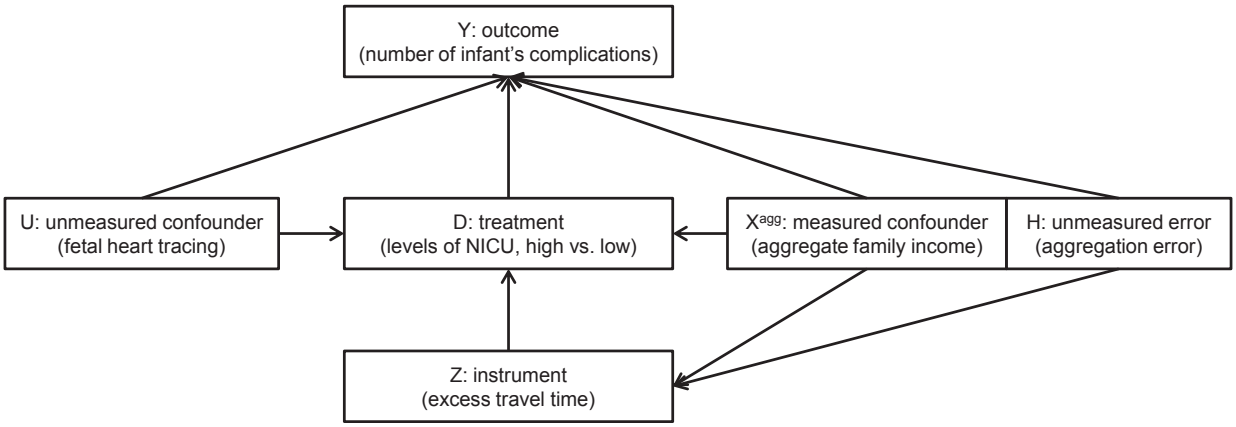

Fig. 2 Causal diagram in regionalization of perinatal care study when only aggregate confounding variable is available

Let $\mathbf{W}^{agg} = [\mathbf{D}, \mathbf{X}^{agg}]$ and $\mathbf{A}^{agg} = [\mathbf{Z}, \mathbf{X}^{agg}]$ be $\mathbf{W}$ and $\mathbf{A}$ but replacing x_{ij} with x_j^{agg}, and $\mathbf{W} = \mathbf{W}^{agg} + \mathbf{H}$ (also, $\mathbf{A} = \mathbf{A}^{agg} + \mathbf{H}$), where $\mathbf{H} = [\mathbf{0}, (\mathbf{X} - \mathbf{X}^{agg})]$ is a matrix for aggregation error due to the replacement of x_{ij} with x_j^{agg}. The estimator $\hat{\beta}'$ — the

solution of $\mathbf{A}^{agg^T}(\mathbf{Y} - \mathbf{W}^{agg^T}\hat{\beta}') = \mathbf{0}$ — is then

$$\begin{aligned}\hat{\beta}' &= (\mathbf{A}^{agg^T}\mathbf{W}^{agg})^{-1}\mathbf{A}^{agg^T}\mathbf{Y} \\ &= (\mathbf{A}^{agg^T}\mathbf{W}^{agg})^{-1}\mathbf{A}^{agg^T}(\mathbf{W}^{agg}\beta + \mathbf{H}\beta + \varepsilon) \\ &= \beta + (\mathbf{A}^{agg^T}\mathbf{W}^{agg})^{-1}\mathbf{A}^{agg^T}\mathbf{H}\beta + (\mathbf{A}^{agg^T}\mathbf{W}^{agg})^{-1}\mathbf{A}^{agg^T}\varepsilon. \end{aligned} \tag{7}$$

We will refer to the method of obtaining estimators from (7) as the naïve method in the rest of the paper. The exogeneity between $\mathbf{A}$ and ε implies that $\mathbf{A}^{agg^T}\varepsilon$ will converge in probability to zero. The other term $\mathbf{A}^{agg^T}\mathbf{H}$, however, will not be zero or converge in probability to zero, because

$$\mathbf{A}^{agg^T}\mathbf{H} = \mathbf{A}^{agg^T}(\mathbf{A} - \mathbf{A}^{agg}) = \begin{bmatrix} \mathbf{Z}^T \\ \mathbf{X}^{agg} \end{bmatrix}\begin{bmatrix} \mathbf{0} & (\mathbf{X} - \mathbf{X}^{agg}) \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{Z}^T(\mathbf{X} - \mathbf{X}^{agg}) \\ \mathbf{0} & \mathbf{X}^{agg^T}(\mathbf{X} - \mathbf{X}^{agg}) \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \mathbf{0} & \operatorname{plim}\mathbf{Z}^T(\mathbf{X} - \mathbf{X}^{agg}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{Z}^T(\mathbf{X} - \mathbf{X}^{agg})$ will not converge in probability to zero unless $\mathbf{Z}$ and $\mathbf{X}$ are uncorrelated in each area. In other words, aggregation of measured confounding variable X makes IV invalid.

In this section, we have shown that if a confounding variable is correlated with the IV, it is necessary to control for this confounding variable when estimating the causal effect of a treatment in IV regression using the TSLS method. In the next section, we provide an analytical method based on Wald's method of grouping and its corresponding assumptions that allows us to obtain consistent estimates for the causal parameters from IV regression using the TSLS method when the confounding variable is correlated with the IV and is only available in aggregate form.

4 Proposed method

We introduce Wald's method of grouping [Wald, 1940] to overcome the problem of aggregation discussed in Section 3. We first explain Wald's original method of grouping.

4.1 Wald's method of grouping

The problem Wald originally addressed is that we would like to fit a simple linear regression but there is measurement error in the explanatory variable X. Suppose we have two variables, Y and X, and they are connected by a linear model, $Y = \alpha + \beta X + \varepsilon$, where $E(\varepsilon) = 0$. Among n i.i.d. observation pairs (y_i, x_i), we divide them into two groups such that one group will have larger x-values than the other. Then, we compute the means of the two groups, which are $(\bar{y_1}, \bar{x_1})$ and $(\bar{y_2}, \bar{x_2})$.

Wald's estimator for β is $b = (\bar{y_2} - \bar{y_1})/(\bar{x_2} - \bar{x_1})$. This is equivalent to calculating the slope of the straight line from the means of these two groups. Wald has shown that the estimator b is a consistent estimator for β, even there exists an unmeasured error that may deteriorate the estimation of β, provided the method of grouping is not affected by the error. One advantage of Wald's method is simplicity. The only thing we have to do is to make sure the method of grouping does not depend on errors. Next, we introduce the concept of Wald's method of grouping to solve the problem of aggregation to obtain a consistent estimator for the causal effect of a treatment in IV regression using the TSLS method.

4.2 Grouped two-stage least squares

In Section 3, we have pointed out that aggregating a confounding variable that is correlated with the instrument in IV regression will result in creating an unmeasured error that makes IV invalid. In the regionalization of perinatal care study, we use excess travel time to a high-level NICU (Z) as an IV to estimate the effect of being delivered at a high-level NICU to a low-level NICU (D) in the infant's number of complications (Y), controlling for measured confounding variables (X); e.g., mother's family income. Because we assume that mother's family income (X) is correlated with excess travel time (Z), it is necessary to control for mother's family income in IV regression using the TSLS method. For each mother i in county j, the ideal observed data are $(y_{ij}, d_{ij}, z_{ij}, x_{ij})$ that all variables are measured at individual level. In hospital records, however, mother's family income is not available to us. The only data available are the aggregate family income in counties linked from the US census database; therefore, we only have x_j^{agg}, not x_{ij}.

We use the concept of Wald's method of grouping to overcome the problem of unmeasured error due to aggregation, provided the assumption that the grouping is independent of the errors ε. Suppose we have $\sum_{j=1}^{J} n_j$ mothers from J counties. We divide mothers into J groups based on their residential address (e.g., county), and calculate means of all variables — aggregate all variables — within each county. Each mother's observed data are then $(y_j^{agg}, d_j^{agg}, z_j^{agg}, x_j^{agg})$. For mothers from the same county, they share the same values of the observed data. These aggregated values are used in IV regression. Note that our proposed method — fitting at individual level with aggregate values replacing all individual values — is different from the ecological regression model, which is fitted at the aggregate level (e.g., the county level with one observation per county). The estimators $\hat{\beta}^{agg}$ will be a consistent estimator for β. The derivation of the consistent estimator for β and its variance estimate is provided in Appendix A.

5 Simulation studies

In this section, we conducted Monte Carlo simulations to investigate the problem of aggregation in observational studies. We simulated 2^K groups based on a simple exponential growth hierarchical structure inspired by the U.S. Census Bureau geographic hierarchy (see Figure 3). We randomly drew n_j subjects from the jth group for a total of $n = \sum_{j=1}^{2^K} n_j$ subjects. We followed the casual diagram we defined in Section 2 (see Figure 1) to create the observed confounding variable x_{ij}, the unmeasured confounding variable u_{ij}, the IV z_{ij}, and the binary treatment assignment d_{ij} as follows. For the ith subject in the jth group, we defined

$$x_{ij} = \left(10 + 90 \times \frac{j-1}{2^K - 1}\right) + \left(1 + 9 \times \frac{j-1}{2^K - 1}\right) \times e_{1,ij} \quad \text{for } j \in \{1, \ldots, 2^K\}$$

$$u_{ij} = 2 \times e_{2,ij},$$

$$z_{ij} = \pi_Z \times x_{ij} + 5 + 1.5 \times e_{3,ij}$$

$$d_{ij} = I(x_{ij} + z_{ij} + u_{ij} + e_{4,ij} > \Delta),$$

where $\mathbf{e}_{ij} = (e_{1,ij}, e_{2,ij}, e_{3,ij}, e_{4,ij})^T \sim MVN_4(\mathbf{0}, \mathbf{I}_4)$ and $I(\cdot)$ is an indicator function, such that $I(\cdot) = 1$ if the expression is true; $I(\cdot) = 0$ if otherwise. The parameter π_Z was used to determine $\rho_{Z,X}$; that is the correlation between z_{ij} and x_{ij}. The choices of $\pi_Z = (0.02, 0.03, 0.06, 0.1, 0.5)$ created approximate values of correlation $\rho_{Z,X} = (0.34, 0.5, 0.75, 0.87, 0.99)$. A constant Δ was chosen, such that the treatment allocation was $50/50$; i.e., the value of Δ depended on the choice of π_Z. For each subject, we generated two potential outcomes $y_{ij}^{(d^*)}$ for $d^* \in \{0, 1\}$, such that $y_{ij}^{(d^*)} = \beta_1 \times d^* + \beta_2 \times x_{ij} + 0.5 \times u_{ij} + e_{ij}$, where $e_{ij} \sim N(0, 1)$. The observed outcome was $y_{ij} = d_{ij} y_{ij}^{(1)} + (1 - d_{ij}) y_{ij}^{(0)}$.

Two thousand Monte Carlo samples were generated from the above population for each scenario. We considered ten scenarios with two sizes of samples ($n = 128,000$ for $K = 7$ and $n_j = 1,000$, and $n = 102,400$ for $K = 10$ and $n_j = 100$) and five different values of correlation $\rho_{Z,X}$. In the presence of unmeasured confounding variable u_{ij}, we used IV regression using the TSLS method to estimate parameters of interest $\beta = (\beta_1, \beta_2)^T = (-2, -0.5)^T$, especially β_1, the causal effect of the treatment. We first obtained estimated $\hat{d}_{ij} = \hat{\gamma}_0 + \hat{\gamma}_1 z_{ij} + \hat{\gamma}_2 x_{ij}$ using OLS, and then regressed y_{ij} on $\hat{d}_{ij}$ and x_{ij} to obtain estimators for β. Table 1 shows the results of IV regression using the TSLS method when the confounding variable x_{ij} is available for each observation. The estimators are consistent and the coverage probabilities for 95% Wald confidence intervals are close to 95%, regardless of the sizes of samples and $\rho_{Z,X}$.

In Table 2 ($n = 128,000$), we compare results of estimating β_1 — the causal effect of the treatment — controlling for the aggregate confounding variable x_j^{agg} by using the naïve and the proposed methods. In general, the estimators obtained by

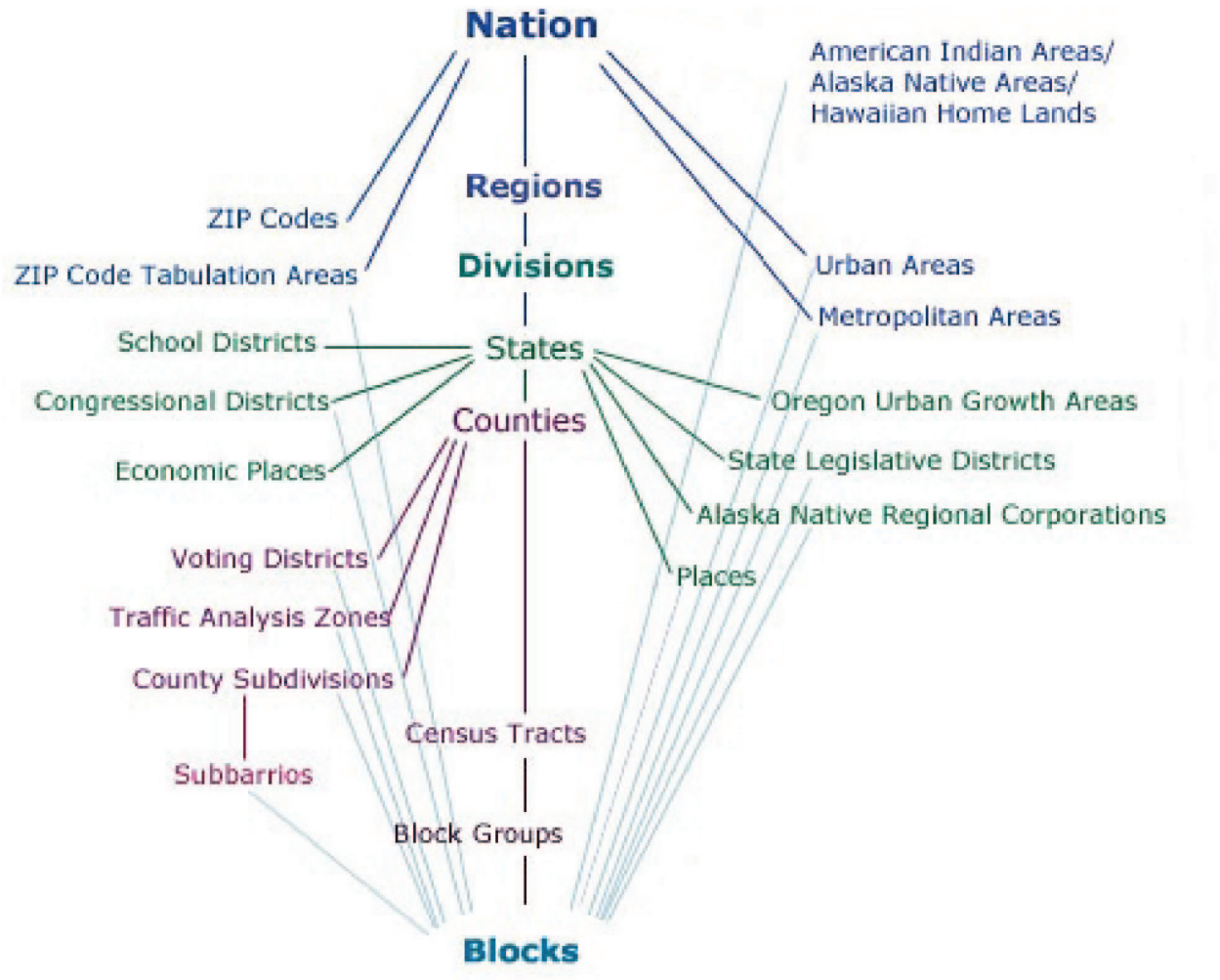

Fig. 3 U.S. Census Bureau geographic hierarchy (Source from www.census.gov)

the naïve method, $\hat{\beta}_1'$, are biased in all scenarios. The bias increases as $\rho_{Z,X}$ increases given the number of aggregate groups. The bias increases as the number of aggregate groups decreases given $\rho_{Z,X}$. For our proposed method, the estimators, $\hat{\beta}_1^{agg}$, are consistent, regardless of the number of aggregate groups and $\rho_{Z,X}$. In terms of estimating the standard deviation of our proposed estimator, the mean sandwich estimator based on (11) is similar to the Monte Carlo standard deviation when the number of aggregate groups is 128. When the numbers of aggregate groups are 32 and 64, the mean sandwich estimator based on (11) still provides reasonable estimated standard error for the estimator. However, when the number of aggregate groups is 2, 4, 8, or 16, the standard error based on (11) underestimates the standard deviation substantially. This is not surprising because the standard error base on (11) is a sandwich variance estimator, and sandwich variance estimator can

perform poorly from a small number of groups [Lipsitz and Fitzmaurice, 2009]; i.e., the robustness property of the sandwich variance estimator is an asymptotic property as the number of groups goes to infinity. The 95% Wald confidence intervals are close to nominal level of 95% for 32, 64, and 128 aggregate groups. The intervals drop below 95% when the number of aggregate groups is less than 32. For scenarios with $n = 102,400$ for $K = 10$ and $n_j = 100$, we have found the similar results as scenarios with $n = 128,000$ for $K = 7$ and $n_j = 1,000$ (data not shown).

Table 1 Simulation results of estimating β ($\beta_1 = -2$ and $\beta_2 = -0.5$) based on 2000 Monte Carlo samples ($\{\hat{\beta}_1, \hat{\beta}_2\}$: mean of the Monte Carlo estimates; $SD(\cdot)$: standard deviation of the Monte Carlo estimates; $\overline{SE}(\cdot)$: mean of the standard error of the Monte Carlo estimates; CP%: coverage probabilities for 95% Wald confidence intervals)

n ($=\sum_{j=1}^{2^K} n_j$)	$\rho_{Z,X}$	$\hat{\beta}_1$	$SD(\hat{\beta}_1)$	$\overline{SE}(\hat{\beta}_1)$	CP%	$\hat{\beta}_2$	$SD(\hat{\beta}_2)$	$\overline{SE}(\hat{\beta}_2)$	CP%
128,000 ($K = 7, n_j = 1,000$)	0.34	-2.0010	0.0234	0.0236	95.3	-0.5000	0.0002	0.0003	95.2
	0.50	-2.0000	0.0233	0.0236	95.8	-0.5000	0.0003	0.0003	95.4
	0.75	-2.0006	0.0238	0.0236	95.2	-0.5000	0.0003	0.0003	95.6
	0.87	-2.0004	0.0244	0.0236	94.0	-0.5000	0.0003	0.0003	93.9
	0.99	-1.9993	0.0237	0.0236	94.8	-0.5000	0.0003	0.0003	95.8
102,400 ($K = 10, n_j = 100$)	0.34	-2.0001	0.0256	0.0262	95.2	-0.5000	0.0003	0.0003	95.5
	0.50	-1.9995	0.0263	0.0262	95.4	-0.5000	0.0003	0.0003	95.4
	0.75	-1.9996	0.0263	0.0262	95.4	-0.5000	0.0003	0.0003	95.2
	0.87	-1.9992	0.0265	0.0261	94.2	-0.5000	0.0003	0.0003	94.3
	0.99	-2.0014	0.0265	0.0261	95.0	-0.5000	0.0003	0.0003	94.5

6 An empirical example: A regionalization of perinatal care study

As an illustrative example, we consider an IV regression model for modeling the causal effect of being delivered at a high-level NICU compared to a low-level NICU on the infant health outcome controlling for measured confounding variables in a study of regionalization of perinatal care.

6.1 Data

The data describe all premature infants born within 23–37 weeks gestational age in the State of Missouri in the years 1993–2003; that is 150,532 births, excluding fetal and infant deaths. The health outcome of interest is the infant's number of complications during the premature delivery. Nine complications of prematurity collected from birth to the time of discharge

Table 2 Comparisons of simulation results of estimating the causal effect of the treatment, $\beta_1 = -2$, obtained by the naïve and the proposed methods based on 2000 Monte Carlo samples of size 128,000 ($\{\hat{\beta}_1', \hat{\beta}_1^{agg}\}$: mean of the Monte Carlo estimates; $SD(\cdot)$: standard deviation of the Monte Carlo estimates; $\overline{SE}(\cdot)$: mean of the standard error of the Monte Carlo estimates; CP%: coverage probabilities for 95% Wald confidence intervals)

		Naïve method				Proposed method			
$\rho_{Z,X}$	Groups	$\hat{\beta}_1'$	$SD(\hat{\beta}_1')$	$\overline{SE}(\hat{\beta}_1')$	CP%	$\hat{\beta}_1^{agg}$	$SD(\hat{\beta}_1^{agg})$	$\overline{SE}(\hat{\beta}_1^{agg})$	CP%
0.34	2	-2.9105	0.0329	0.0900	0	-2.0005	0.0176	0.0000	0
	4	-2.4346	0.0290	0.0727	0	-2.0005	0.0191	0.0141	74.2
	8	-2.2398	0.0261	0.0600	0	-2.0006	0.0215	0.0189	87.7
	16	-2.1911	0.0254	0.0558	0	-2.0007	0.0221	0.0211	92.0
	32	-2.1794	0.0253	0.0547	0.3	-2.0008	0.0225	0.0221	93.6
	64	-2.1765	0.0253	0.0544	0.4	-2.0007	0.0227	0.0226	94.4
	128	-2.1757	0.0253	0.0543	0.4	-2.0008	0.0228	0.0229	94.6
0.50	2	-3.7777	0.0446	0.0885	0	-2.0001	0.0174	0.0000	0
	4	-2.8192	0.0333	0.0719	0	-2.0003	0.0190	0.0142	73.4
	8	-2.4337	0.0272	0.0597	0	-2.0000	0.0212	0.0188	86.8
	16	-2.3395	0.0259	0.0556	0	-2.0000	0.0220	0.0210	92.1
	32	-2.3170	0.0256	0.0545	0	-2.0000	0.0224	0.0219	94.2
	64	-2.3114	0.0255	0.0542	0	-2.0001	0.0226	0.0225	94.9
	128	-2.3098	0.0255	0.0541	0	-2.0000	0.0227	0.0229	95.9
0.75	2	-6.3288	0.0700	0.0853	0	-2.0000	0.0178	0.0000	0
	4	-4.0845	0.0470	0.0700	0	-1.9998	0.0193	0.0143	75.6
	8	-3.0713	0.0349	0.0588	0	-2.0004	0.0217	0.0187	87.0
	16	-2.8200	0.0319	0.0550	0	-2.0006	0.0226	0.0211	91.6
	32	-2.7608	0.0312	0.0539	0	-2.0005	0.0230	0.0221	92.4
	64	-2.7461	0.0311	0.0537	0	-2.0005	0.0232	0.0226	93.8
	128	-2.7420	0.0310	0.0536	0	-2.0005	0.0233	0.0229	94.0
0.87	2	-8.8155	0.0897	0.0833	0	-2.0002	0.0179	0.0000	0
	4	-5.7097	0.0595	0.0684	0	-2.0003	0.0197	0.0144	73.5
	8	-3.9565	0.0434	0.0579	0	-2.0003	0.0222	0.0192	87.5
	16	-3.4930	0.0391	0.0543	0	-2.0003	0.0231	0.0212	92.2
	32	-3.3830	0.0380	0.0534	0	-2.0004	0.0234	0.0222	92.3
	64	-3.3556	0.0378	0.0531	0	-2.0004	0.0236	0.0227	93.3
	128	-3.3478	0.0377	0.0530	0	-2.0004	0.0237	0.0229	93.6
0.99	2	-12.9530	0.1215	0.0824	0	-1.9990	0.0175	0.0000	0
	4	-11.0521	0.0844	0.0695	0	-1.9990	0.0193	0.0142	73.5
	8	-8.2100	0.0617	0.0585	0	-1.9988	0.0214	0.0189	86.8
	16	-7.1364	0.0570	0.0547	0	-1.9988	0.0221	0.0211	91.9
	32	-6.8575	0.0561	0.0537	0	-1.9990	0.0227	0.0219	93.2
	64	-6.7861	0.0558	0.0535	0	-1.9991	0.0228	0.0225	94.3
	128	-6.7664	0.0556	0.0534	0	-1.9991	0.0229	0.0228	95.2

are pneumothorax, bronchopulmonary dysplasis (BPD), intraventricular hemorrhage (IVH), neonatal seizures, necrotizing enterocolitis (NEC), fungal sepsis, bacterial sepsis, urinary tract infection, and surgery retinopathy of prematurity (ROP). None of these complications are expected to occur after discharge from the neonatal intensive care. Complications were derived from ICD-9CM codes contained in the UB-92 forms of infants submitted to the state. Table 3 depicts the raw

comparison of outcome (the infant's number of complications) and covariates (mother's age, mother's education, months when prenatal care began, number of prenatal visits and infant's birth weight and gestational age) between high-level and low-level NICUs in the sample. High-level NICUs have higher number of complications per infant than low-level NICUs (standardized difference = 0.19). However, this does not necessarily mean high-level NICUs cause more complications as babies delivered at high-level NICUs may be at higher risk for complications: their mothers are less educated and started prenatal care later, and the babies have lighter birth weight and shorter gestational age.

We next examined the association between excess travel time and the measured confounding variables. Travel time was determined using ArcView software (ESRI) as the time from the centroid of mother's zip code to the closest low- and high-level NICUs. Excess travel time is the additional travel time that a mother needs to the nearest high-level NICU compared to the travel time to the nearest low-level NICU. Excess travel time is negative if the closest hospital has a high-level NICU, and is positive if otherwise. Excess travel time is used as an IV to account for unmeasured confounding variables in IV regression using the TSLS method. Measured confounding variables used in this example are mother's education; month prenatal care began; and mother's age. Table 4 provides the evidence that excess travel time is correlated with measured confounding variables. Mothers whose travel time is less to a high-level NICU (e.g., those women whose excess travel time to a high-level NICU was less than 30 minutes) are elder, less educated, having more prenatal care visits, begin prenatal care later and deliver lighter babies with shorter gestational ages. A mother's education is part of a mother's SES. The mother's education is highly correlated with her excess travel time at individual and aggregate levels (i.e., the absolute standardized difference is greater than 0.2 standard deviations in Table 4).

Table 3 Raw comparison of number of infant complications and covariates between high-level and low-level NICUs

Variable (n=150,532)	High-level NICU (n=42,764)	Low-level NICU (n=107,768)	|Std. Dif.*|
Outcome			
Number of complications	0.145	0.068	0.19
Covariate			
Mother's age (years)	27.2	26.1	0.17
Mother graduated from high school (proportion)	0.73	0.82	0.22
Prenatal care began (months)	2.06	2.34	0.20
Prenatal visits (times)	11.8	11.0	0.16
Birth weight (grams)	2727	2908	0.26
Gestational age (weeks)	35.0	35.5	0.22

* Std. Dif. (standardized difference) = $\frac{\text{mean difference}}{\text{pooled standard deviation}}$

Table 4 Standardized differences between babies delivered at near to and far from high-level NICUs

	Excess travel time		
Variable	Below 30 minutes	Above 30 minutes	|Std. Dif.*|
Individual			
Mother's age (years)	26.8	25.3	0.24
Mother graduated from high school (proportion)	0.77	0.86	0.24
Prenatal visits (times)	11.4	10.8	0.11
Prenatal care began (months)	2.19	2.48	0.20
Birth weight (grams)	2850	2873	0.03
Gestational age (weeks)	35.3	35.5	0.06
Aggregate			
Mother graduated from high school (proportion)	0.48	0.81	0.67

* Std. Dif. (standardized difference) = $\frac{\text{mean difference}}{\text{pooled standard deviation}}$

6.2 Results

Recall that we defined a naïve method in Section 3 as using an aggregate area-level confounding variable to replace individual confounding variable in an IV regression. The aggregate confounding variable is calculated from each county; there are 116 counties in the data. We compare estimators for the causal effect of being delivered at a high-level NICU obtained from the naïve method to those obtained from the proposed method. Additionally, estimators obtained from a model with all individual level variables are referenced in comparisons.

For the purpose of illustration, we consider a linear model:

$$y_{ij}^{(d^*)} = \beta_0 + \beta_1 d^* + \beta_2 x_{1,ij} + \beta_3 x_{2,ij} + \beta_4 x_{3,ij} + u_{ij} \quad \text{for} \quad d^* \in \{0,1\}, \tag{8}$$

where $\{y_{ij}^{(0)}, y_{ij}^{(1)}\}$ is a vector of potential infant's numbers of complications that would be observed whether an infant i in county j is delivered at a high-level NICU, $x_{1,ij}$ is a SES variable — mother's education, $x_{2,ij}$ is the month prenatal care began, $x_{3,ij}$ is mother's age, and u_{ij} can be viewed as a composite of unmeasured confounding variables such as fetal heart

tracing and the severity of mother's comorbidities. The model for observed data is

$$y_{ij} = \beta_0 + \beta_1 d_{ij} + \beta_2 x_{1,ij} + \beta_3 x_{2,ij} + \beta_4 x_{3,ij} + u_{ij}, \quad E(u_{ij} \mid x_{1,ij}, x_{2,ij}, x_{3,ij}, z_{ij}) = 0, \tag{9}$$

where d_{ij} is indicator for observed treatment — being delivered at a high-level NICU and z_{ij} is excess travel time.

To illustrate the problem of aggregation and our proposed method, we fit three different settings of models based on (9) as follows: (i) reference model: using individual mother's education, $x_{1,ij}$; (ii) naïve method: assuming individual mother's education, $x_{1,ij}$, is not available, and replacing it with aggregate mean value of mother's education from each county, $x_{1,j}^{agg}$; (iii) proposed method: replacing all variables $(y_{ij}, d_{ij}, x_{1,ij}, x_{2,ij}, x_{3,ij}, z_{ij})$ with aggregate mean values from each county, $(y_j^{agg}, d_j^{agg}, x_{1,j}^{agg}, x_{2,j}^{agg}, x_{3,j}^{agg}, z_j^{agg})$. Table 5 depicts the results of comparisons from three IV regressions using TSLS described previously. In the first-stage regression, the significant partial correlation coefficient between being delivered at a high-level NICU and excess travel time controlling for mother's education, month prenatal care began and mother's age confirms that the IV is correlated with the treatment after controlling for confounding variables. In addition, a large first-stage partial F statistic for excess travel time indicates that excess travel time is a strong enough instrument that TSLS method will produce reliable inferences. The partial F statistics in three models are 6633, 4699 and 22915, respectively, which are relatively larger than a usual critical value of F exceeding 10 for a reliable result from IV regression using the TSLS method [Stock et al., 2002]. In the second-stage regression, we focus on comparing estimators for β_1, the causal effect of being delivered at a high-level NICU, controlling for mother's education, month prenatal care began and mother's age. Estimated from reference model, the causal effect is approximately 0.0213 reduction in infant's number of complications. Assuming that individual mother's education is not available and replacing it with aggregate mean value at the county level, the estimated causal effect from the naïve method model is approximately 0.0276 reduction — a 30% difference from the reference model. For the proposed method model, the estimated causal effect is about 0.0212 reduction which is similar to the results from the reference model.

7 Discussion

In this paper, we have developed a method of IV regression for estimating the causal effect of a treatment in observational studies when the individual level confounding variable — the one that is also correlated with the IV — is not available in the data. Our proposed method can provide consistent causal estimators compared to the commonly used analytical strategy in practice that replaces the unavailable confounding variable with its aggregate form in the model. Instead of analyzing data

Table 5 Comparisons of IV regression estimators among three settings of models: (i) reference model — individual mother's education is available; (ii) naïve method model — individual mother's education is not available and replace it with aggregate mother's education at the county level; (iii) proposed method model — aggregate all other variables along with mother's education at the county level

	Reference model		Naïve method model		Proposed method model	
First-stage regression						
Partial correlation for the IV	-0.21		-0.17		-0.36	
Partial F statistic* for the IV	6633		4699		22915	
Second-stage regression						
Covariates	Estimates	SE	Estimates	SE	Estimates	SE
Intercept (β_0)	0.1507	0.0057	0.1342	0.0122	0.0957	0.1070
Being delivered at a high-level NICU† (β_1)	-0.0213	0.0112	-0.0276	0.0133	-0.0212	0.0604
Mother's education‡ (β_2)	-0.0066	0.0019	0.0065	0.0069	0.0254	0.0496
Month prenatal care began‡ (β_3)	-0.0084	0.0008	-0.0080	0.0008	0.0053	0.0144
Mother's age‡ (β_4)	-0.0009	0.0002	-0.0012	0.0002	-0.0023	0.0048

* the first-stage F statistic must be evidently large, typically exceeding 10, for TSLS inference to be reliable

† causal effect of interest

‡ measured confounding variables

with the aggregate confounding variable, we propose an analytical strategy that is to aggregate all variables involved in the analysis at the level of the aggregate confounding variable.

In simulation studies, we have shown that the proposed method can provide good estimation for the parameters of the causal effect in IV regression. Our proposed method can also provide consistent estimators for the variance of estimators as long as the numbers of aggregate groups is large. We not only can estimate the causal effect consistently, but also can calculate confidence intervals and perform statistical inference for the causal effect as long as the number of aggregate groups is large.

An observational study of the regionalization of perinatal care is used as a motivating example to raise the concern of aggregating measured confounding variables in IV regression, when the confounding variable may be correlated with the IV. We have found that the estimated causal effects of the treatment from the reference model (no aggregation) and the naïve method model (aggregating confounding variable that is correlated with the IV) are different whereas results from our proposed method (aggregating all variables) are close to that from the reference model. The standard errors for the aggregate variables are higher than those for the individual variables, because aggregation makes the standard errors larger; i.e., a trade-off between consistency and efficiency. Figure 4 shows the mean squared error (MSE) for both the individual model

with aggregate confounding variable and the aggregate model using one scenario from Section 5 ($K = 7, \beta_1 = 2, \rho_{Z,X} = 0.5$) for different number of samples per group (n_j). As number of samples increases, the MSE from the individual model with aggregate confounding variable converges to 0.1, and the MSE from the aggregate model converges to 0. When number of samples per group is greater than 50, most of the MSE of the individual model with aggregate confounding variable is contributed by bias, because the variance of the estimator has asymptotic property such that it goes toward zero when sample size goes toward infinity. However, the bias of the estimator does not have such asymptotic property.

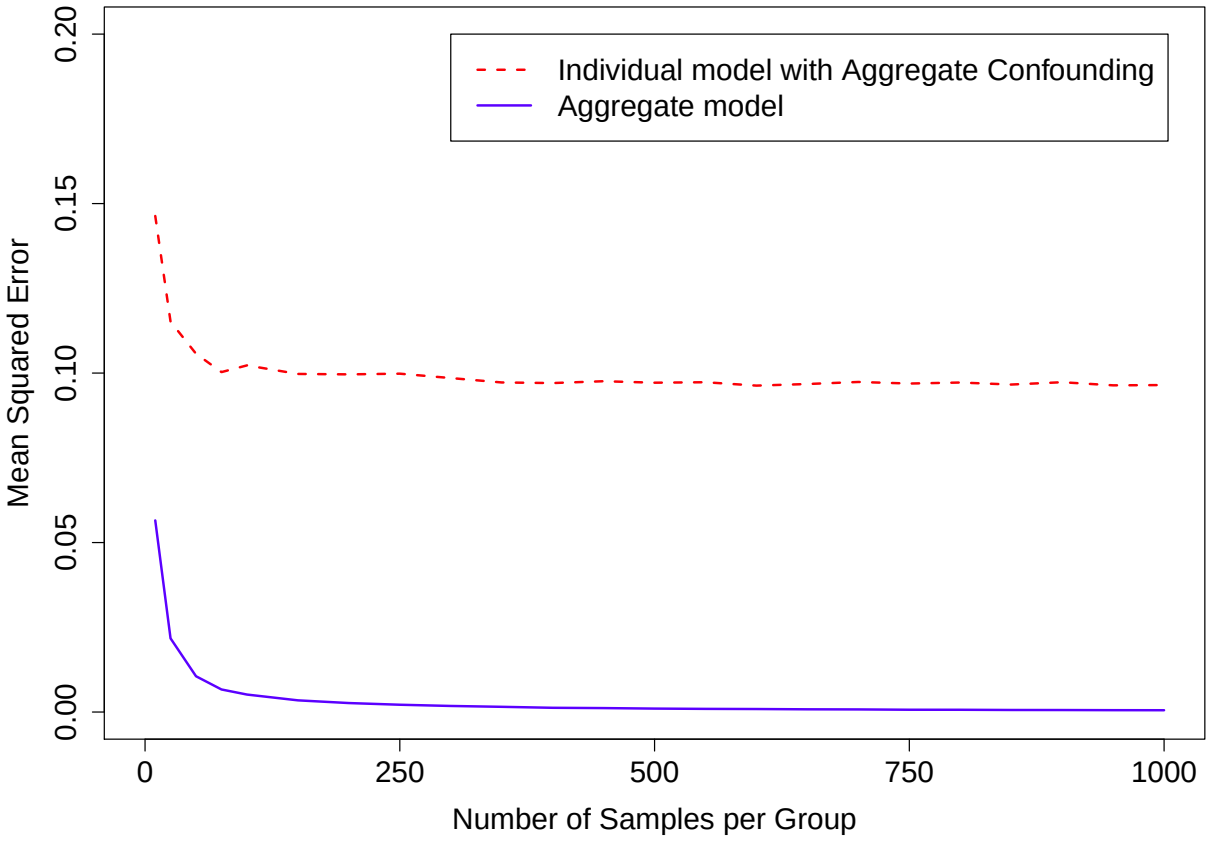

Fig. 4 Plot of the mean squared errors on number of samples in each group for a total of 128 groups

For the purpose of illustration in this paper, we assume that the aggregate confounding variable is calculated from the individual confounding variable in the samples. In practice, the individual confounding variable is not available and the US census database could be a source for the aggregate variable as long as the population of interest is the general population. Even though the population of interest is not the general population, our proposed method still provides consistent estimates if the samples in each group can be assumed to be random samples from the general population of that group and

the number of groups becomes large. In the cases that random samples assumption is not plausible, there is 'inconsistent aggregation', meaning that the aggregate variable is obtained from the population that is different from the population of interest [Rosenbaum and Rubin, 1985]. In Rosenbaum and Rubin (1985), they raised the problems that consistent estimates of regression coefficients may not be obtained from inconsistently aggregated data in linear regression models. In our motivating neonatology study, the population of interest is mothers. Mothers in a county may differ in their socioeconomic characteristics from the general population in the same county. We are currently developing a method of sensitivity analysis for IV regression when there is inconsistent aggregation.

In this paper, we have maintained the exclusion restriction assumption (Assumption 3 in Section 2.2) that the IV has no direct effect on the outcome except through treatment. One way in which the exclusion restriction could be violated is that there is a neighborhood effect, meaning for example that one's neighborhood SES directly influence one's outcomes even after controlling for one's own SES because of within-neighborhood social interactions [Mayer and Jencks, 1989]. If the exclusion restriction is violated because of a neighborhood effect, then the aggregation IV method presented in this paper would produce biased estimates. It would be of future research interest to examine whether the aggregation IV method in this paper could be combined with the extended IV method discussed in Joffe et al. (2008) to address the problem of neighborhood effects.

This paper warns of the possibility of aggregation error in observational studies using IV regression and provides a solution to the problem. Note that the proposed solution requires sufficient size of samples and certain assumptions such as SUTVA, ignorability of the instrument, the exclusion restriction, nonzero average causal effect of the instrument on the treatment, independence of grouping and errors.

8 Acknowledgments

The authors thank the Editors and the referees for helpful comments. The work was supported by the National Science Foundation (Measurement, Methodology and Statistics program), grant # NSF 0961971.

References

[Abadie, 2003] Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113:231–263.

[American Academy of Pediatrics, Committee on Fetus and Newborn, 2004] American Academy of Pediatrics, Committee on Fetus and Newborn (2004). Levels of neonatal care. *Pediatrics*, 114(5):1341–1347.

[Angrist, 1991] Angrist, J. D. (1991). Grouped-data estimation and testing in simple labor-supply models. *Journal of Econometrics*, 47:243–266.

[Angrist et al., 1996] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

[Angrist and Krueger, 2001] Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: from supply and demand to natural experiments. Working paper 8456, National Bureau of Economic Research.

[Baiocchi et al., 2010] Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296.

[Brookhart and Schneeweiss, 2007] Brookhart, M. A. and Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *The International Journal of Biostatistics*, 3(1):Article 14.

[Card and Krueger, 1992] Card, D. and Krueger, A. B. (1992). Does school quality matter? returns to education and the characteristics of public schools in the united states. *Journal of Political Economy*, 100(1):1–40.

[Cifuentes et al., 2002] Cifuentes, J., Bronstein, J., Phibbs, C. S., Phibbs, R. H., Schmitt, S. K., and Carlo, W. A. (2002). Mortality in low birth weight infants according to level of neonatal care at hospital of birth. *Pediatrics*, 109(5):745–751.

[Geronimus and Bound, 1998] Geronimus, A. T. and Bound, J. (1998). Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples. *American Journal of Epidemiology*, 148(5):475–486.

[Geronimus et al., 1996] Geronimus, A. T., Bound, J., and Neidert, L. J. (1996). On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics. *Journal of the American Statistical Association*, 91(434):529–537.

[Hernán and Robins, 2006] Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372.

[Holland, 1988] Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18:449–484.

[Joffe et al., 2008] Joffe, M. M., Small, D., Ten Have, T., Brunelli, S., and Feldman, H. I. (2008). Extended instrumental varialbes estimation for overall effects. *International Journal of Biostatistics*, 4(1):Article 4.

[Krieger, 1992] Krieger, N. (1992). Overcoming the absence of socioeconomic data in medical records: Validation and application of a census-based methodology. *American Journal of Public Health*, 82(5):703–710.

[Krieger et al., 2003a] Krieger, N., Chen, J. T., Waterman, P. D., Rehkopf, D. H., and Subramanian, S. V. (2003a). Race/ethnicity, gender, and monitoring socioeconomic gradients in health: A comparison of area-based socioeconomic measures – the public health disparities geocoding project. *American Journal of Public Health*, 93(10):1655–1671.

[Krieger et al., 2003b] Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M.-J., Subramanian, S. V., and Carson, R. (2003b). Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The public health disparities geocoding project (us). *Journal of Epidemiology & Community Health*, 57:186–199.

[Lipsitz and Fitzmaurice, 2009] Lipsitz, S. and Fitzmaurice, G. (2009). Generalized estimating equations for longitudinal data analysis. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal data analysis.*, pages 43–78. CRC/Chapman & Hall, Boca Raton, FL.

[Lorch et al., 2012] Lorch, S. A., Baiocchi, M., Ahlberg, C. E., and Small, D. S. (2012). The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics*. in press.

[Lorch et al., 2010] Lorch, S. A., Myers, S., and Carr, B. (2010). The regionalization of pediatric health care. *Pediatrics*, 126(6):1182–1190.

[Mayer and Jencks, 1989] Mayer, S. E. and Jencks, C. (1989). Growing up in poor neighborhoods: How much does it matter? *Science*, 243(4897):1441–1445.

[McClellan et al., 1994] McClellan, M., McNeil, B. J., and Newhouse, J. P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? *Journal of the American Medical Association*, 272(1):859–866.

[Neyman, 1990] Neyman, J. (1990). On the application of probability theory to agricultural experiments (translated and edited by D.M. Dabrowska and T.P. Speed). *Statistical Science*, 5(4):465–480.

[Pearl, 2000] Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press, New York.

[Phibbs et al., 2007] Phibbs, C. S., Baker, L. C., Caughey, A. B., Danielsen, B., Schmitt, S. K., and Phibbs, R. H. (2007). Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *The New England Journal of Medicine*, 356:2165–2175.

[Phibbs et al., 1993] Phibbs, C. S., Mark, D. H., Luft, H. S., Peltzman-Rennie, D. J., Garnick, D. W., Lichtenberg, E., and McPhee, S. J. (1993). Choice of hospital for delivery: A comparison of high-risk and low-risk women. *Health Services Research*, 28(2):201–222.

[Phibbs and Robinson, 1993] Phibbs, C. S. and Robinson, J. C. (1993). A variable-radius measure of local hospital market structure. *Health Services Research*, 28(3):313–324.

[Prais and Aitchison, 1954] Prais, S. J. and Aitchison, J. (1954). The grouping of observations in regression analysis. *Review of the International Statistical Institute*, 22(1/3):1–22.

[Rogowski et al., 2004] Rogowski, J. A., Horbar, J. D., Staiger, D. O., Kenny, M., Carpenter, J., and Geppert, J. (2004). Indirect vs direct hospital quality indicators for very-low-birth-weight infants. *Journal of the American Medical Association*, 291(2):202–209.

[Rosenbaum and Rubin, 1985] Rosenbaum, P. R. and Rubin, D. B. (1985). Discussion of "on state education statistics": A difficulty with regression analyses of regional test score averages. *Journal of Educational Statistics*, 10(4):326–333.

[Rubin, 1974] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

[Rubin, 1986] Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

[Stock et al., 2002] Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

[Theil, 1971] Theil, H. (1971). *Principles of Econometrics*. Wiley, New York.

[Wald, 1940] Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.

A Consistency of the TSLS Estimator from the Aggregate IV Regression and Its Variance Estimate

In this section, we will show the show the consistency of $\hat{\beta}^{agg}$ obtained from the aggregate IV regression in Section 4. We will provide an estimate for the variance of $\hat{\beta}^{agg}$.

Let $(\mathbf{Y}^{agg}, \mathbf{X}^{agg}, \mathbf{D}^{agg}, \mathbf{Z}^{agg})$ denote vectors of aggregated $(\mathbf{Y}, \mathbf{X}, \mathbf{D}, \mathbf{Z})$, and $\underline{\mathbf{W}}^{agg} = [\mathbf{D}^{agg}, \mathbf{X}^{agg}]$ and $\underline{\mathbf{A}}^{agg} = [\mathbf{Z}^{agg}, \mathbf{X}^{agg}]$. We use $(\underline{\mathbf{W}}^{agg}, \underline{\mathbf{A}}^{agg})$ to distinguish $(\mathbf{W}^{agg}, \mathbf{A}^{agg})$ used in Section 3 in which $\mathbf{W}^{agg} = [\mathbf{D}, \mathbf{X}^{agg}]$ and $\mathbf{A}^{agg} = [\mathbf{Z}, \mathbf{X}^{agg}]$. Let $(\mathbf{H}_Y, \mathbf{H}_A, \mathbf{H}_W)$ be aggregation errors for $(\mathbf{Y}, \mathbf{A}, \mathbf{W})$,

where $\mathbf{H}_Y = \mathbf{Y} - \mathbf{Y}^{agg}$, $\mathbf{H}_A = \mathbf{A} - \underline{\mathbf{A}}^{agg}$, and $\mathbf{H}_W = \mathbf{W} - \underline{\mathbf{W}}^{agg}$. The TSLS estimator $\hat{\beta}^{agg}$ obtained from all aggregate variables is

$$\begin{aligned}\hat{\beta}^{agg} &= (\underline{\mathbf{A}}^{agg^T}\underline{\mathbf{W}}^{agg})^{-1}\underline{\mathbf{A}}^{agg^T}\mathbf{Y}^{agg} \\ &= (\underline{\mathbf{A}}^{agg^T}\underline{\mathbf{W}}^{agg})^{-1}\underline{\mathbf{A}}^{agg^T}\{(\underline{\mathbf{W}}^{agg}+\mathbf{H}_W)\beta - \mathbf{H}_Y + \varepsilon\} \\ &= \beta + (\underline{\mathbf{A}}^{agg^T}\underline{\mathbf{W}}^{agg})^{-1}\underline{\mathbf{A}}^{agg^T}(\mathbf{H}_W\beta - \mathbf{H}_Y + \varepsilon) \end{aligned} \tag{10}$$

If we can show that $\underline{\mathbf{A}}^{agg^T}(\mathbf{H}_W\beta - \mathbf{H}_Y + \varepsilon) \xrightarrow{p} \mathbf{0}$ in (10), then $\hat{\beta}^{agg}$ is a consistent estimator for β. We could write the aggregate matrices $(\mathbf{Y}^{agg}, \underline{\mathbf{A}}^{agg}, \underline{\mathbf{W}}^{agg})$ as $(\mathbf{GY}, \mathbf{GA}, \mathbf{GW})$. The matrix $\mathbf{G}$ is a diagonal grouping matrix, where

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_j \end{bmatrix} \quad \text{and} \quad \mathbf{G}_j = \begin{bmatrix} 1/n_j & \cdots & 1/n_j \\ \vdots & \ddots & \vdots \\ 1/n_j & \cdots & 1/n_j \end{bmatrix}.$$

Thus, $\underline{\mathbf{A}}^{agg^T}(\mathbf{H}_W\beta - \mathbf{H}_Y + \varepsilon)$ can be written as $\mathbf{A}^T\mathbf{G}^T\{(\mathbf{W}-\mathbf{GW})\beta - (\mathbf{Y}-\mathbf{GY}) + \varepsilon\}$. Since $\mathbf{G}$ is a symmetric and idempotent matrix, $\mathbf{G}^T(\mathbf{W}-\mathbf{GW})$ and $\mathbf{G}^T(\mathbf{Y}-\mathbf{GY})$ are zero. Also, $\mathbf{A}^T\mathbf{G}^T\varepsilon$ converges in probability to zero because of the assumption of independence between $\mathbf{G}$ and ε. The variance of $\hat{\beta}^{agg}$ is

$$\begin{aligned} Var\left(\hat{\beta}^{agg}\right) &= Var\left\{(\underline{\mathbf{A}}^{agg^T}\underline{\mathbf{W}}^{agg})^{-1}\underline{\mathbf{A}}^{agg^T}(\mathbf{H}_W\beta - \mathbf{H}_Y + \varepsilon)\right\} \\ &= (\mathbf{A}^T\mathbf{GW})^{-1}\mathbf{A}^T\mathbf{G} \times Var\{(\mathbf{W}-\mathbf{GW})\beta - (\mathbf{Y}-\mathbf{GY}) + \varepsilon\} \times \mathbf{GA}(\mathbf{W}^T\mathbf{GA})^{-1} \\ &= (\mathbf{A}^T\mathbf{GW})^{-1}\mathbf{A}^T\mathbf{G} \times Var\{(\mathbf{GY}-\mathbf{GW}\beta)\} \times \mathbf{GA}(\mathbf{W}^T\mathbf{GA})^{-1} \\ &= (\mathbf{A}^T\mathbf{GW})^{-1}\mathbf{A}^T\mathbf{G} \times Var(\mathbf{G}\varepsilon) \times \mathbf{GA}(\mathbf{W}^T\mathbf{GA})^{-1}, \end{aligned} \tag{11}$$

where $Var(\mathbf{G}\varepsilon)$ can be estimated by

$$\widehat{Var}(\mathbf{G}\varepsilon) = \begin{bmatrix} \hat{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \hat{\Sigma}_j \end{bmatrix} \quad \text{and}$$

$$\hat{\Sigma}_j = \begin{bmatrix} (n_j-1)^{-1}\sum_{i=1}^{n_j}(y_j^{agg} - [z_j^{agg}, x_j^{agg}]\hat{\beta}^{agg})^2 & \cdots & (n_j-1)^{-1}\sum_{i=1}^{n_j}(y_j^{agg} - [z_j^{agg}, x_j^{agg}]\hat{\beta}^{agg})^2 \\ \vdots & \ddots & \vdots \\ (n_j-1)^{-1}\sum_{i=1}^{n_j}(y_j^{agg} - [z_j^{agg}, x_j^{agg}]\hat{\beta}^{agg})^2 & \cdots & (n_j-1)^{-1}\sum_{i=1}^{n_j}(y_j^{agg} - [z_j^{agg}, x_j^{agg}]\hat{\beta}^{agg})^2 \end{bmatrix}.$$