Biometrics xx, 1–20 DOI: 10.1111/j.1541-0420.2005.00454.x
mm yyyy

Calibrating Sensitivity Analyses to Observed Covariates in Observational Studies

Jesse Y. Hsu

Department of Statistics, The Wharton School, University of Pennsylvania

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

Philadelphia, Pennsylvania, U.S.A.

email: hsu9@wharton.upenn.edu

and

Dylan S. Small

Department of Statistics, The Wharton School, University of Pennsylvania

Philadelphia, Pennsylvania, U.S.A.

email: dsmall@wharton.upenn.edu

Summary: In medical sciences, statistical analyses based on observational studies are common phenomena. One peril of drawing inferences about the effect of a treatment on subjects using observational studies is the lack of randomized assignment of subjects to the treatment. After adjusting for measured pretreatment covariates, perhaps by matching, a sensitivity analysis examines the impact of an unobserved covariate, u, in an observational study. One type of sensitivity analysis uses two sensitivity parameters to measure the degree of departure of an observational study from randomized assignment. One sensitivity parameter relates u to treatment and the other relates u to response. For subject matter experts, it may be difficult to specify plausible ranges of values for the sensitivity parameters on their absolute scales. We propose an approach that calibrates the values of the sensitivity parameters to the observed covariates and is more interpretable to subject matter experts. We will illustrate our method using data from the U.S. National Health and Nutrition Examination Survey regarding the relationship between cigarette smoking and blood lead levels.

Key words: Causal inference; Hidden bias; Simultaneous sensitivity analysis.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

1.1 *Observational Studies and Sensitivity Analyses*

An observational study estimates treatment effects on subjects, when subjects are not assigned to treatments at random. Because subjects in treated and control groups may not be comparable, differences in their responses could be observed even if there were no treatment effect. A sensitivity analysis examines the impact of hidden bias due to an unobserved covariate in an observational study; see §4, Rosenbaum (2002) for a review. Methods and applications of sensitivity analyses for general observational studies include Cornfield et al. (1959), Rosenbaum and Rubin (1983), Rosenbaum (1986), Gastwirth (1992), Marcus (1997), Pan and Frank (2003), Imbens (2003), Shepherd et al. (2007), Small (2007), and Hosman et al. (2010). These authors consider the effect of unobserved covariates to conduct sensitivity analyses to hidden bias in observational studies. In general, sensitivity parameters quantify the impact of the unobserved covariates on treatment or on response and measure the degree of departure of observational studies from randomized assignments. Sensitivity parameters can be difficult for some subject matter experts to interpret. In this paper, we develop an approach for calibrating sensitivity parameters to observed covariates to make sensitivity parameters more interpretable.

To illustrate the challenges of interpreting sensitivity parameters, consider Bingenheimer et al. (2005)'s study of the effect of exposure to firearm violence on subsequent perpetration of serious violence. Using data from Chicago and controlling for 153 pre-exposure covariates through propensity score sub-classification, Bingenheimer et al. estimate that exposure to firearm violence approximately doubles the probability that an adolescent will perpetrate violence over the next two years (p-value for no effect < 0.001). In a news article that accompanies Bingenheimer et al.'s paper, Holden (2005) reports that a number of scholars are not convinced by the findings of the study. For example, Holden says, "Economist

Stephen Durlauf of the University of Wisconsin, Madison, calls the study an "implausible modeling of violence exposure." The authors assume that the individuals with the same propensity rankings are equally likely to encounter violence, he says. But such exposure may not be random; rather it probably stems from "something that has not been measured" — such as recklessness, says Durlauf." Bingenheimer et al. in fact conducts a sensitivity analysis that considers what the effect of an unobserved confounder like recklessness would be. Suppose the unobserved confounder is standardized to have mean 0 and variance 1; e.g., a standardized measure of recklessness. Bingenheimer et al. consider various effects this unobserved confounder might have, for example: (A) the unobserved confounder has a 0.43 larger mean for exposed subjects and a 0.6 larger mean for perpetrators; or (B) the unobserved confounder has a 0.49 larger mean for exposed subjects and a 0.69 larger mean for perpetrators. An unobserved confounder with properties (A) would still leave the p-value for testing no effect below 0.05 but an unobserved confounder with properties (B) would bring the p-value above 0.05. Thus, the key issue for assessing what the impact of not having measured the confounder recklessness on the conclusions of the study is whether standardized recklessness has properties more like (A) or (B). It may be difficult for subject matter experts to think about this in terms of absolute numbers; e.g., it may be difficult to think about whether standardized recklessness is more likely to have a 0.43 larger mean for exposed subjects (A) or a 0.49 larger mean for exposed subjects (B). In this paper, we re-frame the question in terms of comparing unobserved covariates to observed covariates. For example, we would look for an observed covariate among the 153 observed covariates that is similar in its properties to hypothetical unobserved confounder (A) and another observed covariate that is similar in its properties to hypothetical unobserved confounder (B), and ask the question, do we think the unobserved confounder is more similar in its properties to the first observed covariate or to the second covariate. The key output of our method is a graph

like Figure 2 that tells readers if an unobserved covariate had a strength of confounding similar to a given observed covariate, would there still be strong evidence for the exposure having a causal effect on the outcome.

The paper is organized as follows. In the rest of Section 1 we describe a motivating example from the U.S. National Health and Nutrition Examination Survey regarding the relationship between cigarette smoking and blood lead levels followed by a sensitivity analysis for matched pairs. In Section 2, we explain notation and reviews of methodology. The proposed method of calibrating sensitivity analysis to observed covariates is discussed in Section 3. In Section 4, empirical results from the motivating example are given to illustrate our proposed method. Finally, in Section 5, we give a brief discussion.

1.2 *A Motivating Example: Lead in the Blood of Smokers*

Using data from the 2007–2008 U.S. National Health and Nutrition Examination Survey (NHANES), we study the effect of smoking on blood lead levels among 679 daily smokers and 2661 non-smokers. A daily smoker reported smoking every day for the previous 30 days and smoking an average of at least 10 cigarettes per day on these days. A non-smoker reported smoking fewer than 100 cigarettes in his or her life and smoking no cigarettes in the previous 30 days. All subjects were at least 20 years old and had no tobacco use besides cigarette smoking in the previous 5 days.

Does smoking increase blood lead levels? To answer this question, we first matched 679 daily smokers with 2661 candidates of non-smokers using the `pairmatch` function in the `optmatch` package in `R` (Hansen, 2007) applied to a distance matrix that combined a caliper on an estimated propensity score with a rank based Mahalanobis distance (Rosenbaum, 2002). Pairs were matched for age, gender, education, income and race. Table 1 shows means and standardized differences in means (i.e., the difference in means divided by a measure of the average within treatment group standard deviation) before and after matching. Before

matching, the covariates income-to-poverty level, gender, 9–11th grade education, high school education, college education, White race, Mexican American race, and other Hispanic are out of balance; the magnitude of their standardized differences are all greater than 0.2. The matching approximately balances all of the observed covariates. After matching, the magnitude of standardized difference is less than 0.1 for all of the covariates except 9–11th grade education, for which the standardized difference is 0.11. The outcome is a binary response of whether blood lead levels is greater than or equal to 5 ug/dl, where 5 ug/dl is the reference level at which Centers for Disease Control and Prevention recommends public health actions to be initiated (`www.cdc.gov/nceh/lead/`). We use McNemar's test statistic to test for the effect of smoking on high blood lead. Of 679 pairs, there are 68 pairs in which exactly one person has high blood lead. Of these, there are 46 pairs in which the daily smokers have high blood lead and 22 pairs in which the non-smokers have high blood lead. If the study were free of hidden bias, McNemar's test statistic would yield a significance level of 0.0025, which would suggest strong evidence that smoking causes high blood lead if the study were free of hidden bias.

[Table 1 about here.]

1.3 *Simultaneous Sensitivity Analyses to Hidden Bias*

How sensitive to hidden bias from an unobserved covariate is the result that smoking causes high blood lead? One type of sensitivity analyses, simultaneous sensitivity analyses, use two sensitivity parameters, Γ and Δ, to measure the degree of hidden bias due to the unobserved covariate in an observational study (Gastwirth et al., 1998). Suppose there is an unobserved covariate u that lies between 0 and 1. One sensitivity parameter, Γ, relates u to treatment; namely, the odds ratio of receiving treatment for two subjects with different values of u is at most Γ, and the other parameter, Δ, relates u to response; namely, the odds ratio of having higher response for two subjects with different values of u is at most Δ. The simultaneous

sensitivity analysis finds the maximum p-value over all distributions of u for given values of Γ and Δ. If u is either irrelevant to smoking (i.e., $\Gamma = 1$) or irrelevant to high blood lead (i.e., $\Delta = 1$), the maximum one-sided p-value = 0.0025; therefore, there is strong evidence that smoking causes high blood lead. If $\Gamma > 1$ and $\Delta > 1$, the maximum p-value is changed and this may or may not alter the result that there is strong evidence smoking causes high blood lead. For example, for $\Gamma = \Delta = 2$, one person in a pair may be twice as likely to smoke and twice as likely to have high blood lead as the other because they have different values of u, but there is still strong evidence that smoking causes high blood lead because the maximum one-sided p-value = 0.0285. On the other hand, when $\Gamma = \Delta = 2.5$, there is no longer strong evidence that smoking causes high blood lead because the maximum one-sided p-value = 0.0963. Web Appendix B.1 gives the simultaneous sensitivity analysis for the NHANES data.

In this paper, we propose a method to provide a way to interpret sensitivity parameters in terms of the observed covariates that subject matter experts are familiar with. Specifically, we calibrate the values of Γ and Δ in a simultaneous sensitivity analysis to the observed covariates in an observational study.

2. Notation and Reviews

2.1 *Notation*

Suppose that there are I matched sets, $i = 1, \ldots, I$, matched for observed covariates, $\mathbf{x}$. Within a set i, there are $n_i \geqslant 2$ subjects, $j = 1, \ldots, n_i$, and $N = \sum n_i$ subjects in total. If subject ij receives treatment, then $Z_{ij} = 1$. If otherwise, then $Z_{ij} = 0$. Let m_i be the number of treated subjects in set i such that $m_i = \sum_{j=1}^{n_i} Z_{ij}$. In a full matching, each set contains either one treated subject and $n_i - 1$ controls or $n_i - 1$ treated subjects and one control such that $m_i = 1$ or $m_i = n_i - 1$ for $i = 1, \ldots, I$. In a matching with multiple controls, $m_i = 1$

and $n_i \geqslant 2$ for $i = 1, \ldots, I$. In a matching with a fixed number of controls, k, $m_i = 1$ and $n_i = k+1$ for $i = 1, \ldots, I$. A pair matching is a special case of matching with a fixed number of controls with $k = 1$.

Let $\mathbf{x}_{ij}$ be the vector of P observed covariates for subject ij. If a study is free of hidden bias, the probability that subject ij receives the treatment is a function of the observed covariates $\mathbf{x}_{ij}$ describing subject ij, say $p_{Z_{ij}} = g(\mathbf{x}_{ij})$. If two subjects, ij and ij', with the same observed covariates have different probabilities of receiving the treatment, there is hidden bias. That is $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ but $p_{Z_{ij}} \neq p_{Z_{ij'}}$. We will assume that this hidden bias comes from an unobserved covariate u_{ij} such that if u_{ij} were observed, there would be no hidden bias. Following Neyman (1923) and Rubin (1974), each subject ij has two potential outcomes, $(y_{ij}^{(0)}, y_{ij}^{(1)})$, and $y_{ij}^{(z)}$ denotes the response that would be observed for subject ij if subject ij's level of Z were set to z. The response actually observed from subject ij is $y_{ij} = Z_{ij}y_{ij}^{(1)} + (1 - Z_{ij})y_{ij}^{(0)}$.

2.2 *Models for Treatment and Response in the Population Before Matching*

Following Gastwirth et al. (1998) and Small et al. (2009), in the population before matching, the following models will be used:

$$Z_{ij} \perp\!\!\!\perp (y_{ij}^{(0)}, y_{ij}^{(1)}) \mid \mathbf{x}_{ij}, u_{ij}, \tag{1}$$

$$\Pr(Z_{ij} = 1 \mid \mathbf{x}_{ij}, u_{ij}) = \frac{\exp\{\beta(z = 1, \mathbf{x}_{ij}) + \gamma u_{ij}\}}{\exp\{\beta(z = 0, \mathbf{x}_{ij})\} + \exp\{\beta(z = 1, \mathbf{x}_{ij}) + \gamma u_{ij}\}}, \text{ and} \tag{2}$$

$$\Pr(y_{ij}^{(0)} = y \mid \mathbf{x}_{ij}, u_{ij}) = \exp\{\zeta(\mathbf{x}_{ij}, u_{ij}) + \kappa(y, \mathbf{x}_{ij}) + \delta y u_{ij}\}, \tag{3}$$

where $A \perp\!\!\!\perp B \mid C$ is Dawid's (1979) notation for conditional independence of A and B given C, $\beta(\cdot)$ and $\kappa(\cdot)$ are unknown functions, $\zeta(\mathbf{x}_{ij}, u_{ij})$ are normalizing constants, and γ and δ are unknown parameters. Assumption (1) ensures that u_{ij} is the only relevant unobserved covariate; i.e., treatment Z_{ij} and response under control $y_{ij}^{(0)}$ are dependent only because of their dependence on u_{ij}. Model (2) is a logit model for the treatment assignment. Model (3) allows for $y_{ij}^{(0)}$ to take on several common distributions such as binary, Poisson, normal,

or gamma. For example, if $y_{ij}^{(0)}$ is binary, then $\zeta(\mathbf{x}_{ij}, u_{ij}) = -\log[\exp\{\kappa(y = 0, \mathbf{x}_{ij})\} + \exp\{\kappa(y = 1, \mathbf{x}_{ij}) + \delta u_{ij}\}]$ and $\Pr(y_{ij}^{(0)} = 1 \mid \mathbf{x}_{ij}, u_{ij}) = \exp\{\kappa(y = 1, \mathbf{x}_{ij}) + \delta u_{ij}\}/[\exp\{\kappa(y = 0, \mathbf{x}_{ij})\} + \exp\{\kappa(y = 1, \mathbf{x}_{ij}) + \delta u_{ij}\}]$ is a logit model. If $y_{ij}^{(0)}$ is continuous and follows a normal distribution with mean $\mu = \tau(\mathbf{x}_{ij}) + \delta u_{ij}$ where $\tau(\cdot)$ is a known function of $\mathbf{x}_{ij}$, and variance σ^2, then $\zeta(\mathbf{x}_{ij}, u_{ij}) = -1/2 \log(2\pi\sigma^2) - \mu^2/2\sigma^2$, $\kappa(y, \mathbf{x}_{ij}) = -\{y^2 - 2\tau(\mathbf{x}_{ij})y\}/2\sigma^2$ and $\Pr(y_{ij}^{(0)} = y \mid \mathbf{x}_{ij}, u_{ij}) = (2\pi\sigma^2)^{-1/2} \exp\{-(y - \mu)^2/2\sigma^2\}$. Model (3) only specifies the form of the relationship between u_{ij} and $y_{ij}^{(0)}$, and allows for any possible relationship between $\mathbf{x}_{ij}$ and $y_{ij}^{(0)}$ since $\kappa(\cdot)$ and $\zeta(\cdot)$ are not restricted.

The parameters γ and δ in models (2) and (3) are sensitivity parameters which determine the strength of the relationship between the unobserved covariate u_{ij} and the treatment Z_{ij} and the response under control $y_{ij}^{(0)}$, respectively. If u_{ij} were irrelevant to Z_{ij} and $y_{ij}^{(0)}$, then $\gamma = \delta = 0$. To make the parameters γ and δ meaningful, u_{ij} needs to be scaled in some way. We assume a binary unobserved covariate such that

$$\Pr(u_{ij} = 1 \mid \mathbf{x}_{ij}) = \Pr(u_{ij} = 0 \mid \mathbf{x}_{ij}) = \frac{1}{2}, \tag{4}$$

which also implies that u_{ij} and $\mathbf{x}_{ij}$ are independent. Wang and Krieger (2006) show that among all distributions for u_{ij} with mean 1/2 and variance 1/4, u_{ij} being Bernoulli with probability 1/2 maximizes the upper bound on the p-value. In practical terms, this means that the conclusions we reach assuming the unobserved covariate is binary with probability 1/2 are at worst conservative in terms of being able to reject the null hypothesis of no treatment effect.

2.3 *Test Statistics and Sensitivity Analyses of Hidden Bias*

For all matched sets, let $\mathbf{Z}$ be the vector of treatments, $\mathbf{Z} = (Z_{11}, \ldots, Z_{I,n_I})^T$, $\mathbf{y}$ be the vector of observed responses, $\mathbf{y} = (y_{11}, \ldots, y_{I,n_I})^T$, $\mathbf{X}$ be the matrix of observed covariates, $\mathbf{X} = \{\mathbf{x}_{11}, \ldots, \mathbf{x}_{I,n_I}\}^T$, and $\mathbf{u}$ be the vector of unobserved covariates, $\mathbf{u} = (u_{11}, \ldots, u_{I,n_I})^T$. Let $\mathbf{m}$ denote the vector of the number of treated units in set i, $\mathbf{m} = (m_1, \ldots, m_I)$. Consider

Fisher's sharp null hypothesis of no treatment effect (Fisher, 1935), $H_0 : \mathbf{y}^{(0)} = \mathbf{y}^{(1)}$, against the alternative hypothesis that more treatment causes higher responses. Common tests for matched studies such as McNemar's test, Wilcoxon's signed rank test, Mantel-Haenszel test, and aligned rank test of Hodeges and Lehmann have test statistics in the form of

$$T = t(\mathbf{Z}, \mathbf{y}) = \sum_{i=1}^{I} d_i \sum_{j=1}^{n_i} c_{ij} Z_{ij}, \tag{5}$$

where d_i and c_{ij} are any function of $\mathbf{y}$; see Rosenbaum (2002), §2 for details.

Let $y_{i(1)} \leqslant \ldots \leqslant y_{i(n_i)}$ and $Z_{i(1)} \leqslant \ldots \leqslant Z_{i(n_i)}$ be the order statistics for match set i, and $\widetilde{\mathbf{y}} = (y_{1(1)}, \ldots, y_{I(n_I)})^T$ and $\widetilde{\mathbf{Z}} = (Z_{1(1)}, \ldots, Z_{I(n_I)})^T$ denote the vectors of ordered y_{ij} and Z_{ij}. Following Gastwirth et al. (1998), we consider inference conditional on these order statistics $\widetilde{\mathbf{y}}$. For Fisher's sharp null hypothesis of no treatment effect, the upper tail area of the test statistic T is bounded by,

$$\Pr_{H_0}(T \geqslant k \mid \mathbf{m}, \widetilde{\mathbf{y}}, \mathbf{X}, \mathbf{u}) \leqslant \max_{\mathbf{u} \in \mathcal{U}} \left\{ \Pr_{H_0}(T \geqslant k \mid \mathbf{m}, \widetilde{\mathbf{y}}, \mathbf{X}, \mathbf{u}) \right\}, \tag{6}$$

where $\mathcal{U}$ is the set of possible values of $\mathbf{u}$; see Gastwirth et al. (1998) and Gastwirth et al. (2000) for calculation of bounds. For instance, in pair matching, the chance that the treated subject in pair i has the higher response under the null hypothesis is

$$p_i(\mathbf{u}) = \frac{\exp\{\gamma(u_{i2} - u_{i1})\} \exp\{\delta(y_{i(2)} - y_{i(1)})(u_{i2} - u_{i1})\} + 1}{[1 + \exp\{\gamma(u_{i2} - u_{i1})\}][1 + \exp\{\delta(y_{i(2)} - y_{i(1)})(u_{i2} - u_{i1})\}]}.$$

The right-hand side of (6) is then

$$\max_{\mathbf{u} \in \mathcal{U}} \left\{ \Pr_{H_0}(T \geqslant k \mid \mathbf{m}, \widetilde{\mathbf{y}}, \mathbf{X}, \mathbf{u}) \right\} = \sum_{\mathbf{b} \in \mathbf{B}} \chi \left(\sum_{i=1}^{I} b_i d_i \geqslant k \right) \prod (p_i^+)^{b_i} (1 - p_i^+)^{1 - b_i},$$

where $\mathbf{B}$ is the set containing 2^I distinct vectors $\mathbf{b}$ of dimension I with coordinates equal to one or zero, $\chi(\text{event}) = 1$ or 0 if the event occurs or does not, d_i is any function of $\mathbf{y}$ defined in (5), and p_i^+ is the maximum value of $p_i(\mathbf{u})$, which is

$$p_i^+ = \frac{\exp(\gamma) \exp\{\delta(y_{i(2)} - y_{i(1)})\} + 1}{\{1 + \exp(\gamma)\}[1 + \exp\{\delta(y_{i(2)} - y_{i(1)})\}]};$$

see Web Appendix A for details on computation of $p_i(\mathbf{u})$ and p_i^+ in the case of pair matching. In the case of matching with multiple controls or full matching, it may be computationally

difficult to find the exact upper bound in (6). However, a computationally quick asymptotic separability approach provides an approximation for the bounds that converges to the exact upper bound as the number of matched sets converges to infinity (Gastwirth et al., 2000; Small et al., 2009). The upper bound, $\max_{\mathbf{u}\in\mathcal{U}}\{\Pr_{H_0}(T \geqslant k \mid \mathbf{m}, \widetilde{\mathbf{y}}, \mathbf{X}, \mathbf{u})\}$, is the maximum one-sided p-value for a test statistic T. Hereafter we refer this upper bound as 'the maximum p-value' in the paper (i.e., dropping terms 'one-sided' and 'for a test statistic T').

3. Calibrating Sensitivity Analyses to Observed Covariates

3.1 *Sensitivity Models for Treatment and Response*

To calibrate the effects of unobserved covariates relative to observed covariates, we specify parametric models for the effects of observed covariates $\mathbf{x}_{ij}$ in models (2) and (3) in which $\mathbf{x}_{ij}$ are linear predictors in the link functions for generalized linear models. Specifically, we specify model (2) so that

$$\Pr(Z_{ij} = 1 \mid \mathbf{x}_{ij}, u_{ij}) = \frac{\exp(\boldsymbol{\theta}^T\mathbf{x}_{ij} + \gamma u_{ij})}{1 + \exp(\boldsymbol{\theta}^T\mathbf{x}_{ij} + \gamma u_{ij})}; \tag{7}$$

i.e., the treatment Z_{ij} follows a logistic regression model that is linear in $\mathbf{x}_{ij}$ and u_{ij}, where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_P\}^T$ are P parameters for effects of $\mathbf{x}_{ij}$ on Z_{ij} and the sensitivity parameter, $\Gamma = \exp(\gamma)$, is the odds ratio for receiving treatment for $u_{ij} = 1$ vs. $u_{ij} = 0$. For a binary response, we specify model (3) so that

$$\Pr(y_{ij}^{(0)} = 1 \mid \mathbf{x}_{ij}, u_{ij}) = \frac{\exp(\boldsymbol{\phi}^T\mathbf{x}_{ij} + \delta u_{ij})}{1 + \exp(\boldsymbol{\phi}^T\mathbf{x}_{ij} + \delta u_{ij})}, \tag{8}$$

where $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_P\}^T$ are P parameters for effects of $\mathbf{x}_{ij}$ on $y_{ij}^{(0)}$ and the sensitivity parameter, $\Delta = \exp(\delta)$, is the odds ratio for $y_{ij}^{(0)} = 1$ for $u_{ij} = 1$ vs. $u_{ij} = 0$. For a normal response, we specify model (3) so that

$$\Pr(y_{ij}^{(0)} = y \mid \mathbf{x}_{ij}, u_{ij}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y - \boldsymbol{\phi}^T\mathbf{x}_{ij} - \delta u_{ij})^2}{2\sigma^2}\right\}; \tag{9}$$

i.e., $y_{ij}^{(0)}$ is normal with mean $\boldsymbol{\phi}^T\mathbf{x}_{ij} + \delta u_{ij}$ and the sensitivity parameter, $\Delta = \exp(\delta)$, reflects the effect of the unobserved covariate u_{ij} on the response; see Gastwirth et al. (1998)

for further discussion of interpreting Δ. Similar models as (8) or (9) can be specified for other types of responses. Note that the models in (7)–(8) may be misspecified in that the effects of the covariates $\mathbf{x}_{ij}$ on the logit of the probability may not be linear. Such potential misspecification may be mitigated by including quadratics, interactions and other functions of the covariates in $\mathbf{x}_{ij}$.

Suppose we are considering possible effects (γ, δ) of the unobserved covariate u_{ij}. To calibrate how the effects of the unobserved covariate compare to the effects of an observed covariate x_{ijp}, we would like to compare γ to θ_p in model (7) and δ to ϕ_p in model (8) or (9), and hence need to estimate $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. We can do so by the method of maximum likelihood, but because u_{ij} is unobserved, we need to marginalize over u_{ij} to find the likelihood of the observed data. Note that, even though u_{ij} is independent of $\mathbf{x}_{ij}$, the coefficients on $\mathbf{x}_{ij}$ in model (7) or (8) cannot be consistently estimated by logistic regression Z_{ij} or $y_{ij}^{(0)}$ on $\mathbf{x}_{ij}$, because the logistic regression model is generally not collapsible (Guo et al., 1995) and estimates that ignore u_{ij} are biased (Gail et al., 1984). To marginalize over u_{ij}, we use models and assumptions (1)–(4) to obtain

$$p_{Z_{ij}} = \Pr(Z_{ij} = 1 \mid \mathbf{x}_{ij}, \gamma) = \frac{1}{2} \sum_{c=0}^{1} \Pr(Z_{ij} = 1 \mid \mathbf{x}_{ij}, u_{ij} = c, \gamma), \tag{10}$$

and

$$p_{y_{ij}^{(0)}} = \Pr(y_{ij}^{(0)} = y \mid \mathbf{x}_{ij}, \delta) = \frac{1}{2} \sum_{c=0}^{1} \Pr(y_{ij}^{(0)} = y \mid \mathbf{x}_{ij}, u_{ij} = c, \delta). \tag{11}$$

From the marginalized density in (10), the log-likelihood function for $\boldsymbol{\theta}$ given $\mathbf{Z}$, $\mathbf{X}$ and γ is

$$\log L(\boldsymbol{\theta}; \mathbf{Z}, \mathbf{X}, \gamma) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left\{ Z_{ij} \log \left(\frac{p_{Z_{ij}}}{1 - p_{Z_{ij}}} \right) + \log \left(1 - p_{Z_{ij}} \right) \right\}, \tag{12}$$

and from the marginalized density in (11), the log-likelihood function of $\boldsymbol{\phi}$ given $\mathbf{y}^{(0)}$, $\mathbf{X}$ and

δ is

$$\log L(\boldsymbol{\phi}; \mathbf{y}^{(0)}, \mathbf{X}, \delta)$$
$$= \begin{cases} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left\{ y_{ij}^{(0)} \log \left(\frac{p_{y_{ij}^{(0)}}}{1-p_{y_{ij}^{(0)}}} \right) + \log \left(1 - p_{y_{ij}^{(0)}} \right) \right\} & \text{if } y_{ij}^{(0)} \text{ is binary} \\ \sum_{i=1}^{I} \sum_{j=1}^{n_i} \log \left(p_{y_{ij}^{(0)}} \right) & \text{if } y_{ij}^{(0)} \text{ is continuous.} \end{cases} \tag{13}$$

The log-likelihood (12) can be maximized using the observed data to estimate $\boldsymbol{\theta}$. The log-likelihood (13) contains partially unobserved data because $y_{ij}^{(0)}$ is only observed if $Z_{ij} = 0$. One method is to replace $y_{ij}^{(0)}$ with $y_{ij}^{(1)}$ for those who have $Z_{ij} = 1$, because under Fisher's sharp null hypothesis, $y_{ij}^{(0)}$ is equal to $y_{ij}^{(1)}$. Another method is to use only data from subjects whose $Z_{ij} = 0$. Assumption (1) ensures that model (3) continues to hold if we condition on $Z_{ij} = 0$. The log-likelihood (13) restricted to subjects whose $Z_{ij} = 0$ is

$$\log L(\boldsymbol{\phi}; \underline{\mathbf{y}}^{(0)}, \underline{\mathbf{X}}, \delta) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (1 - Z_{ij}) \times \log L(\boldsymbol{\phi}; y_{ij}^{(0)}, \mathbf{x}_{ij}, \delta), \tag{14}$$

where $\underline{\mathbf{y}}^{(0)}$ and $\underline{\mathbf{X}}$ are rows of $\mathbf{y}^{(0)}$ and $\mathbf{X}$ when $Z_{ij} = 0$. If Fisher's sharp null hypothesis is true, both methods provide good estimates. Estimates based on (14) are less efficient than estimates based on (13) due to the use of partial data. If the alternative hypothesis is true (i.e., treatment effect exists), estimates based on (14) are still good estimates while estimates based on (13) may be biased depending on the magnitude of treatment effect. We conduct a simulation study to examine the performance of estimates from both methods under null and various alternative hypotheses. The details of the simulation study are provided in Web Appendix C.1. The results show that under the null, both methods are consistent with estimates based on (13) being more efficient. Note that the amount of lost efficiency depends on the proportion of controls. When the alternative hypothesis is true, estimates based on (14) remain consistent, and estimates based on (13) deviate from true parameters as the magnitude of treatment effect increases. In this paper, we use estimates based on the log-

likelihood (14), using only data from subjects whose $Z_{ij} = 0$, and obtain estimates for $\boldsymbol{\phi}$ from maximizing (14).

3.2 *Graphical Calibration to Observed Covariates*

In Section 3.1, we have shown how to compare the effect of an unobserved covariate with effect Γ on treatment and with effect Δ on response to that of observed covariates with $\boldsymbol{\Theta} = \exp(\boldsymbol{\theta})$ and $\boldsymbol{\Phi} = \exp(\boldsymbol{\phi})$. To summarize these comparisons in a digestible way, we would like to take a graphical approach. Using the simultaneous sensitivity analysis method described in Section 2, we can draw the curve in Figure 1(a) that shows for what values of Γ and Δ the maximum p-value for testing Fisher's sharp null hypothesis of no treatment effect in (6) is less than α (to the left of the curve) or greater than α (to the right of the curve). Figure 1(a) only shows $\Gamma > 1$ and $\Delta > 1$. Figure 1(b) expands Figure 1(a) to show all possible values of Γ and Δ. In Figure 1(b), quadrants I and III are mirror sides to each other and so are quadrants II and IV; e.g., both (Γ, Δ) and $(1/\Gamma, 1/\Delta)$ yield the same maximum p-value in (6). We denote two solid curves in quadrants I and III by $\boldsymbol{\Omega}_\alpha$, the shaded area by $\boldsymbol{\Omega}_\alpha^-$ and the white area by $\boldsymbol{\Omega}_\alpha^+$. On the graph, we would like to plot the effects of observed covariates so that if the effect of an observed covariate is in $\boldsymbol{\Omega}_\alpha^-$ and the unobserved covariate has similar effect as the observed covariate, then the maximum p-value would be less than α. Vice versa, if an observed covariate is in $\boldsymbol{\Omega}_\alpha^+$ and the unobserved covariate has similar effect as the observed covariate, then the maximum p-value would be greater than α. A complication in making this graph is that the estimated effects of the observed covariates from Section 3.1 depend on the values of γ and δ, or equivalently the values of Γ and Δ. We particularly care about the effects of the observed covariates relative to $(\Gamma, \Delta) \in \boldsymbol{\Omega}_\alpha$, since $\boldsymbol{\Omega}_\alpha$ is the borderline between having a significant effect vs. non-significant effect. In Web Appendix C.2, we conduct a simulation study to study the sensitivity of estimates for $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ along these curves. The results show that the estimates are not sensitive to the choice

of $(\Gamma, \Delta) \in \mathbf{\Omega}_\alpha$. We suggest using as a default choice for the estimated effects of observed covariates, the estimated effects for the $(\Gamma, \Delta) \in \mathbf{\Omega}_\alpha$ that intersect the line $\Gamma = \Delta$.

[Figure 1 about here.]

When estimating $\mathbf{\Theta}$ and $\mathbf{\Phi}$ from the log-likelihoods (12) and (14), we could consider (γ, δ) as unknown parameters and estimate them. Copas and Li (1997) discuss that such an approach may not be robust to small changes in the model; e.g., changing model (2) from a logistic regression to a probit regression. In this paper, we adopt Copas and Li (1997)'s suggestion and conduct a sensitivity analysis for the effect of different $(\Gamma, \Delta) \in \mathbf{\Omega}_\alpha$ rather than try to estimate (Γ, Δ).

4. Empirical Results: Smoking and High Blood Lead

To illustrate our proposed method, we consider the study from Section 1.2 on the effect of smoking on high blood lead. We provide the `R` codes for our proposed method in Web Section D.

In Section 1.2, assuming no hidden bias, McNemar's test statistic for no smoking effect on high blood lead yields a p-value of 0.0025. A simultaneous sensitivity analysis that examines the effect of associations between an unobserved covariate and, respectively, smoking and high blood lead is carried out to investigate the impact of the unobserved covariate on the results that smoking causes high blood lead; see Web Appendix B.1; Web Table 1 shows the maximum p-value for McNemar's test statistic for a given (Γ, Δ). To calibrate the simultaneous sensitivity analysis to observed covariates, we use a conventional level of significance, that is $\alpha = 0.05$, and therefore, $\mathbf{\Omega}_{.05}$ is a set of possible values of (Γ, Δ) for which the maximum p-value for McNemar's test statistic is 0.05. The set $\mathbf{\Omega}_{.05}$ does not have a closed form, but we can approximate it by the set $\overline{\overline{\mathbf{\Omega}}}_{.05}$ as follows. We expand the values of (Γ, Δ) from 1.01 to 16 with an increment of 0.01 and create a 1500×1500 grid. For each Γ,

we look for a Δ such that the maximum p-value is close to 0.05. For example, $\overline{\overline{\boldsymbol{\Omega}}}_{.05}$ includes $(\Gamma, \Delta) = (1.43,9.25)$, $(1.60,4.22)$, $(2.20,2.22)$ and $(6.95,1.47)$, where the maximum p-values for McNemar's test statistics are 0.0500, 0.0499, 0.0497 and 0.0500, respectively. In Web Appendix B.2, we list 100 randomly selected values of (Γ, Δ) and their maximum p-values from the collection of $(\Gamma, \Delta) \in \overline{\overline{\boldsymbol{\Omega}}}_{.05}$.

Our goal is to compare the effects of observed covariates on smoking and high blood lead to the possible effects of an unobserved covariate on smoking and high blood lead. The observed covariates include age, income-to-poverty level, gender, education and race. To make the effects of continuous observed covariates comparable with the effects of a binary unobserved covariate, we adopt Gelman (2008)'s suggestion to standardize the continuous covariates (age and income-to-poverty level in our case) to mean 0 and standard deviation (SD) 0.5. The coefficients for the scaled covariates correspond to 2-SD changes; this is roughly comparable to the coefficient on a binary covariate, which corresponds to exactly a 2-SD change if the binary covariate has probability 1/2. Following our suggestion in Section 3, we use $\Gamma = \Delta = 2.21$ as a default choice in this example, where the maximum p-value is approximately equal 0.0497. Using the `optim` function in `R`, we maximize two log-likelihood functions, $\log L\{\boldsymbol{\theta}; \mathbf{Z}, \mathbf{X}, \gamma = \log(2.21)\}$ and $\log L\{\boldsymbol{\phi}; \underline{\mathbf{y}}^{(0)}, \underline{\mathbf{X}}, \delta = \log(2.21)\}$, in (12) and (14) to obtain $\widehat{\boldsymbol{\theta}}_\gamma$ and $\widehat{\boldsymbol{\phi}}_\delta$. We then calculate $\widehat{\boldsymbol{\Theta}}_\gamma = \exp(\widehat{\boldsymbol{\theta}}_\gamma)$ and $\widehat{\boldsymbol{\Phi}}_\delta = \exp(\widehat{\boldsymbol{\phi}}_\delta)$ which are the estimated effects of observed covariates on smoking and high blood lead. See Web Appendix B.3 for more details.

Figure 2 depicts calibration of the simultaneous sensitivity analysis. The solid curves represent values of $(\Gamma, \Delta) \in \overline{\overline{\boldsymbol{\Omega}}}_{.05}$. The shaded area represents values of $(\Gamma, \Delta) \in \overline{\overline{\boldsymbol{\Omega}}}^{-}_{.05}$, where the maximum p-value is less than 0.05. Because covariates of age of 2-SD difference, income-to-poverty level of 2-SD difference, 9–11th grade vs. college education, high school vs. college education, some college vs. college education, Black vs. White race, Mexican

American vs. White race and other races vs. White race fall into $\overline{\overline{\mathbf{\Omega}}}^{-}_{.05}$, if the effect of an unobserved covariate is similar to the effect of any of these observed covariates, then there is strong evidence that smoking causes high blood lead. The white area represents values of $(\Gamma, \Delta) \in \overline{\overline{\mathbf{\Omega}}}^{+}_{.05}$, where the maximum p-value is greater than 0.05. Because covariates of male vs. female gender, less than 9th grade vs. college education and other Hispanic vs. White race fall into $\overline{\overline{\mathbf{\Omega}}}^{+}_{.05}$, if the effect of an unobserved covariate is similar to the effect of any of these observed covariates, then there is not strong evidence that smoking causes high blood lead.

[Figure 2 about here.]

In addition to individual covariate comparisons discussed previously, we can also provide calibration to the effects of multiple observed covariates by comparing an unobserved covariate to a binary covariate constructed from several observed covariates. For example, if a binary unobserved covariate is similar in its effects to level 1 being male gender and Black race vs. level 0 being female gender and White race, then the effects of this observed covariate constructed from gender and race are $(e^{0.86-0.77}, e^{1.67+0.56}) = (1.09, 9.33) \in \overline{\overline{\mathbf{\Omega}}}^{-}_{.05}$. Thus, if the effect of an unobserved covariate is similar to that of a binary covariate of male gender and Black race vs. female gender and White race, there would be strong evidence that smoking causes high blood lead; i.e., the maximum p-value for McNemar's test statistic would be less than 0.05. Another example is if an unobserved covariate is similar in its effects to level 1 being age of 2-SD older, 9–11th grade education and Black race vs. level 0 being age of 2-SD younger, college education and White race, then the effects of this observed covariate constructed from age, gender, education and race are $(e^{-0.63+2.24-0.77}, e^{0.81-0.09+0.56}) = (2.32, 3.63) \in \overline{\overline{\mathbf{\Omega}}}^{+}_{.05}$. Hence, if an unobserved covariate has similar effect as a binary covariate of age of 2-SD older/9–11th grade education/Black race vs. age of 2-SD younger/college education/White race, there would not be strong evidence

that smoking causes high blood lead; i.e., the maximum p-value for McNemar's test statistic would be greater than 0.05.

5. Discussion

In this paper, we have proposed a method to calibrate a simultaneous sensitivity analysis to observed covariates for a matched observational study. We believe that comparing effects of an unobserved covariate with effects of observed covariates on treatment and on response can aid subject matter experts in specifying plausible ranges of values for the sensitivity parameters on their absolute scales. Our method is a tool to help researchers interpret sensitivity parameters but still requires researchers' judgment to decide whether a study's conclusions made under the assumption of no unobserved confounding are likely to still be valid even if there is a plausible amount of unobserved confounding or should be taken with a grain of salt because of concern about unobserved confounding. For example, in one study, a researcher might think the unobserved covariate could only have an effect as strong as that of the most important observed covariates, but in another study, a researcher might think that the unobserved covariate could only have an effect half as strong as that of the most important observed covariate. The calibration, e.g., Figure 2, needs to be combined with such researcher's judgment to decide whether a study's conclusions made under the assumption of no unobserved confounding are likely to still be valid even if there is a plausible amount of unobserved confounding or should be taken with a grain of salt.

One setting in which our method is useful is when a proxy is used for the true confounder that we would like to control for. In medical studies, income, education and occupation are often used as proxies for socioeconomic status (SES) (Geronimus and Bound, 1998 and Hsu et al., 2012). For example, suppose college education is used as a proxy for SES and suppose it is thought that the unobserved part of SES is no more important than the observed part (i.e., college education). If a significant treatment effect is found, our method allows a researcher

to determine if the unobserved SES could explain the significant finding, or instead there is strong evidence for a treatment effect in spite of unobserved confounding of the specified magnitude.

Our method of calibrating a sensitivity analysis to observed covariates works for general matching designs (e.g., pair matching, matching with multiple controls, full matching, etc.) once $\mathbf{\Omega}_\alpha$ is determined. As discussed in Section 2, an approach using asymptotic separability is useful for finding the approximation bounds in (6) for more complicated designs.

Though we consider a binary unobserved covariate with probability 1/2 which leads to the most bias (Wang and Krieger, 2006), our method is able to handle other distributions of unobserved covariate (e.g., normal) by marginalizing out $\mathbf{u}$ in (10) and (11).

A potential limitation to our method and a suggestion for mitigating it is the following. Our method is motivated by the fact that the sensitivity parameters Γ and Δ may be difficult to interpret and we seek to make them easier to interpret by calibrating them to the observed covariates. This calibration does however require the choice of (Γ, Δ) at which to estimate the effects of the observed covariates. This may create a loop. We have suggested using the $(\Gamma, \Delta) \in \overline{\overline{\mathbf{\Omega}}}_{.05}$ that intersects with the line $\Gamma = \Delta$ as a default choice at which to estimate the effects of the observed covariates. In Web Appendix B.4, we examine how sensitive our proposed method is to different choices of $(\Gamma, \Delta) \in \overline{\overline{\mathbf{\Omega}}}_{.05}$ for our motivating example and find that it is not very sensitive. A different and interesting approach to calibration suggested by a referee is as follows. Find the smallest α_0 such that for all $(\Gamma, \Delta) \in \overline{\overline{\mathbf{\Omega}}}_{\alpha_0}$, the estimates for the observed covariates given (Γ, Δ) are all to the southwest of (Γ, Δ). This α_0 is the smallest significance level at which we can reject the null hypothesis of no treatment effect under the assumption that the effect of an unobserved covariate on the treatment and the outcome is less than that of all the observed covariates.

Supplementary Materials

Web Appendices A, B, C and D and data from the motivating example are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

We are grateful to the Associate Editor and two anonymous referees for their comments and suggestions. Support from the United States National Science Foundation Measurement, Methodology and Statistics Program is acknowledged.

References

Bingenheimer, J. B., Brennan, R. T., and Earls, F. J. (2005). Firearm violence exposure and serious violent behavior. *Science* **308,** 1323–1326.

Copas, J. B. and Li, H. G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society. Series B (Methodological)* **59,** 55–95.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer. *Journal of the National Cancer Institute* **22,** 173–203.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* **41,** 1–31.

Fisher, R. A. (1935). *Design of Experiments.* Oliver and Boyd, Edinburgh.

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71,** 431–444.

Gastwirth, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables. *Jurimetrics* **33,** 19–34.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85,** 907–920.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (2000). Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62,** 545–555.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* **27,** 2865–2873.

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,** 475–486.

Guo, J. and Geng, Z. (1995). Collasibility of logistic regression coefficients. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 263–267.

Hansen, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News* **7,** 18–24.

Holden, C. (2005). Controversial study suggests seeing gun violence promotes it. *Science* **308,** 1239–1240.

Hosman, C. A., Hansen, B. B., and Holland, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics* **4,** 849–870.

Hsu, J. Y., Lorch, S. A., and 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,** 119–140.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review* **93,** 126–132.

Marcus, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *Journal of Educational and Behavioral Statistics* **22,** 193–201.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Reprint in *Statistical Science* (5:465–480, 1990) .

Pan, W. and Frank, K. A. (2003). A probability index of the robustness of a causal inference. *Journal of Educational and Behavioral Statistics* **28,** 315–337.

Rosenbaum, P. R. (1986). Dropping out of high school in the united states: An observational study. *Journal of Educational Statistics* **11,** 207–224.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York, 2nd edition.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

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

Shepherd, B. E., Gilbert, P. B., and Mehrotra, D. V. (2007). Eliciting a counterfactual sensitivity parameter. *The American Statistician* **61,** 56–63.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restructions. *Journal of the American Statistical Association* **102,** 1049–1058.

Small, D. S., Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (2009). Simultaneous sensitivity analysis for observational studies using full matching or matching with multiple controls. *Statistics and Its Interface* **2,** 203–211.

Wang, L. and Krieger, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Statistics in Medicine* **25,** 2257–2271.

Received mm yyyy. *Revised mm* yyyy. *Accepted mm* yyyy.

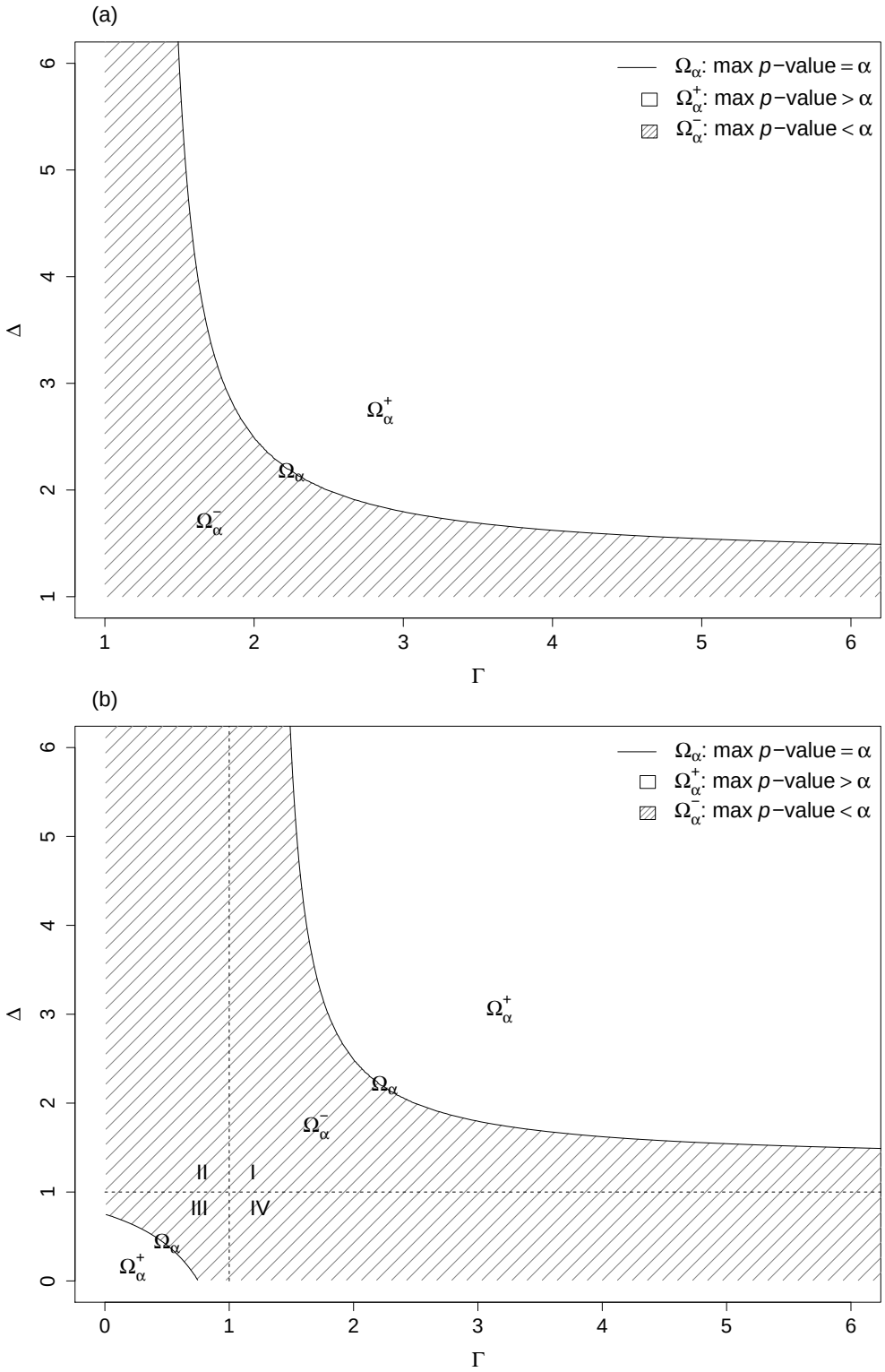

Figure 1. A simultaneous sensitivity analysis for the maximum p-value for a significance level of α, when (a) $\Gamma > 1$ and $\Delta > 1$; and (b) $\Gamma > 0$ and $\Delta > 0$. Panel (a) is equivalent to quadrant I in panel (b). In panel (b), quadrant III is the mirror side of quadrant I and quadrant II is the mirror side of quadrant IV. The solid curves represent values of $(\Gamma, \Delta) \in \mathbf{\Omega}_\alpha$ where the maximum p-value equals to α. The shaded area represents values of $(\Gamma, \Delta) \in \mathbf{\Omega}_\alpha^-$ where the maximum p-value is less than α. The white area represents values of $(\Gamma, \Delta) \in \mathbf{\Omega}_\alpha^+$ where the maximum p-value is greater than α.

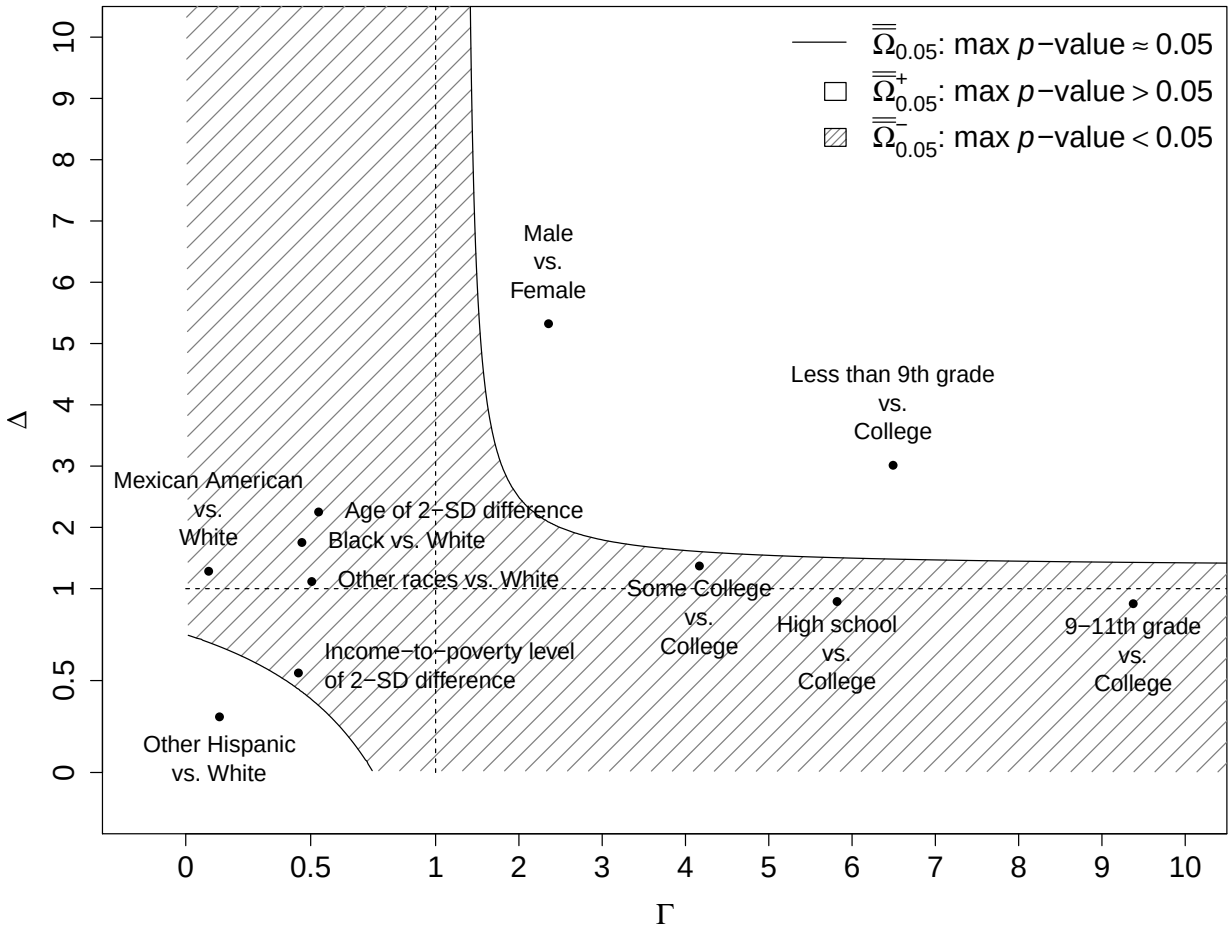

Figure 2. Calibration of the simultaneous sensitivity analysis to observed covariates in the NHANES data given $(\Gamma, \Delta) = (2.21, 2.21) \in \overline{\overline{\mathbf{\Omega}}}_{.05}$. The solid curves represent values of $(\Gamma, \Delta) \in \overline{\overline{\mathbf{\Omega}}}_{.05}$ where the maximum p-value is approximately equal to 0.05. The shaded area represents values of $(\Gamma, \Delta) \in \overline{\overline{\mathbf{\Omega}}}^{-}_{.05}$ where the maximum p-value is less than 0.05. The white area represents values of $(\Gamma, \Delta) \in \overline{\overline{\mathbf{\Omega}}}^{+}_{.05}$ where the maximum p-value is greater than 0.05. The areas $(\Gamma, \Delta) \in \{[0,1] \times [0,1], [0,1] \times [1,10], [1,10] \times [0,1]\}$ are magnified to enable clear display of covariates in these areas.

Table 1
Summary statistics and covariates balance before and after matching for pairs of a daily smoker and a non-smoker.

Matching	Before			After	
Variable	n=679 Smoker	n=2661 Non-smoker	Std. Diff.*	n=679 Non-smoker	Std. Diff.*
Covariate					
Age	46.7	49.8	-0.19	48.2	-0.09
Income-to-poverty level	2.0	2.6	-0.42	2.1	-0.09
Missing, %	6.0	9.2	-0.12	6.0	0.00
Male, %	56.8	37.8	0.39	55.4	0.03
Education, %					
Less than 9th grade	11.2	13.8	-0.08	10.3	0.03
9–11th grade	26.5	13.9	0.32	22.1	0.11
High school graduate	32.3	23.3	0.20	33.7	-0.03
Some college	24.7	25.6	-0.02	28.6	-0.09
College	5.3	23.3	-0.53	5.3	0.00
Unknown	0.0	0.1	-0.04	0.0	0.00
Race, %					
White	65.2	40.8	0.50	63.9	0.03
Black	19.1	19.1	0.00	20.6	-0.04
Mexican American	6.5	21.3	-0.44	6.5	0.00
Other Hispanic	4.7	13.4	-0.31	4.7	0.00
Other races	4.4	5.3	-0.04	4.3	0.01
Outcome					
High blood lead ($\geqslant 5\ ug/dl$), %	6.8	2.4	†	3.2	†

* Std. Diff.: Standardized differences $= (\bar{x}_1 - \bar{x}_2)/\sqrt{(s_1^2 + s_2^2)/2}$, where $\bar{x}_m$ and s_m^2 are sample mean and variance for smokers ($m = 1$) and non-smokers ($m = 2$)
† Covariate balance will not be checked for response