CHAPTER FOURTEEN

FIXED EFFECTS AND DIFFERENCE-IN-DIFFERENCES

Erin C. Strumpf, Sam Harper, and Jay S. Kaufman

Would increasing mothers' education reduce infant mortality? Does giving new parents additional time off from work affect their newborns' health? Do tobacco taxes reduce or exacerbate social inequalities in smoking? Questions about the impact of social and economic exposures on health are a chief concern of social epidemiology. Our aim in this chapter is to provide an introduction to analytic techniques and study designs—namely fixed effects and difference-in-differences—that are useful for answering questions about the causal impact of social exposures on health.

Understanding the role that social factors play in population health has become an important aim for public health scientists in recent decades, and most governments have now made reducing social inequalities an important consideration of public health policy (Wanless 2007; WHO Commission on Social Determinants of Health 2008; Xavier *et al.* 2009; Koh 2010). Population health interventions are expensive, and if we decide to spend our resources on interventions to improve population health and reduce health inequalities, we want that money to be spent on effective solutions. As other chapters in this volume make clear, strong associations exist between health and socioeconomic, ethnic, geographic, and other social factors over the life course. Much of this work is strictly descriptive, and there are good reasons to continue measuring and monitoring health inequalities. However, as the sheer volume of studies documenting the existence of social inequalities in health has increased, the

question of how to design policies to reduce inequalities has become more pressing, and the demand for reliable and valid evidence on how to do so has increased (Petticrew *et al.* 2004; Petticrew 2007; O'Campo 2012).

Observational studies in epidemiology have come under substantial scrutiny in recent years, as scientists have sought explanations for discrepancies between randomized and non-randomized studies of similar exposure contrasts (Davey Smith and Ebrahim 2001; Lawlor *et al.* 2004; Ioannidis 2005; Hernán *et al.* 2008). Because randomized interventions are often impractical or unethical for social exposures, social epidemiology must instead rely mostly on observational data, where unmeasured confounding is a constant threat. This is particularly true for social exposures, since it is often easy to hypothesize non-causal explanations for observed differences in health between social groups (Cutler *et al.* 2011). Translating observed relationships between social exposures and health outcomes into effective policies and interventions requires identifying which correlations reflect underlying causal effects (Harper and Strumpf 2012).

Causal effects necessarily involve a comparison of individuals under (at least) two different exposure/treatment regimes, but we can only observe one (factual) treatment and must therefore try and obtain a suitable counterfactual substitute for what would have happened to individuals under the alternative treatment (Maldonado and Greenland 2002). Randomization greatly helps to solve this problem since in expectation the treated and untreated groups are exchangeable, but if we cannot randomize then how should we make comparisons between different levels of exposure? Much of the literature in social epidemiology has focused on comparisons across individuals with different levels of, for example, income or education. Knowing that there are many reasons why individuals with high versus low education differ apart from their health, researchers often use regression to adjust for these measured differences between groups. The validity of this strategy for making causal inferences rests on the strong assumption that we have measured all of the relevant confounders (in addition to the usual assumptions about absence of measurement error and selection bias, and correct model specification).

To make the assumption of no residual confounding more credible, we could use other comparisons that may be closer to our desired counterfactual. One potential strategy could be to use *changes* in exposure status *within* individuals as a way to control for hard-to-measure individual-level factors that do not vary over time (e.g., innate ability, stable personality characteristics) that are likely to be correlated with both exposure and outcome. This is the logic behind the fixed effects models we describe below: that we improve causal inference by comparing within individuals over time rather than across

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 343

individuals. However, when changes in exposure status are under the control of the individuals we are studying, concerns about unobserved factors that are correlated with changes in exposure status remain.

Another option to generate more credible counterfactual inference is to utilize changes in exposure status that result from changes in policy or other decisions made at other levels beyond the individual (e.g., safety legislation, antipoverty programs, taxation changes). The resulting changes in exposure status are less likely to be related to unmeasured individual-level factors that also affect health outcomes. Policy changes generate what are often referred to as "natural experiments" (Craig *et al.* 2012) and form the typical setup for difference-in-differences analyses. In what follows, we describe the rationale for considering quasi-experimental research designs that help to reduce confounding from (some) unobserved factors, introduce basic methods for how to implement these strategies, highlight important methodological considerations that are common to both, and provide examples and applications from published studies.

Methods

Intuition

Researchers interested in measuring causal effects might first prefer to inhabit a universe in which they are able to observe both potential outcomes for each individual, that is, both $Y_{\text{SET}[X=1]}$ and $Y_{\text{SET}[X=0]}$ for each individual in the population (where the notation SET$[X = x]$ refers to assigning treatment status). Barring the possibility of this metaphysical miracle, we appeal to the analogy of the randomized controlled trial where each individual is randomly assigned to treatment or control and in expectation both observed and unobserved confounders are balanced across the two treatment groups. In other words, on average these potential confounders are no longer correlated with treatment assignment and therefore will not bias the estimated treatment effect.

Because it is typically not feasible to learn about the effects of most social exposures by randomly assigning them (notable exceptions include housing (Ludwig *et al.* 2011; Thomson *et al.* 2013), income (Forget 2011), early childhood education (Heckman 2006), and health insurance (Brook *et al.* 1983; Baicker *et al.* 2013)), we turn to quasi-experimental methods that aim to mimic the design of a randomized controlled trial as closely as possible. More specifically, we seek quasi-experimental study designs in which the only difference between exposed and unexposed units, or between individuals in

their exposed and unexposed states, is the exposure itself. If this is indeed the case, then we find ourselves back in the setting of a randomized controlled trial where potential confounders are in fact not correlated with exposure and thus will not bias the treatment effect. In such a study design, we can claim that the exposure is "as good as random," that is, independent of unobserved confounders, either unconditionally or conditional on observed confounders.

To illustrate the strengths of the quasi-experimental approach, suppose our question of interest is whether a mother's education affects child health. We specify a naïve model where the health of child i at time t is regressed on the mother's education:

$$Y_{it} = \alpha_0 + \alpha_1 Exposure_{it} + \alpha_2 X_{it} + \alpha_3 Z_i + \varphi_{it} \quad (1)$$

where Y_{it} is the outcome of interest (child's health), $Exposure_{it}$ is the mother's education level (e.g., primary, secondary, university), X_{it} are observed time-varying covariates (e.g., age), Z_i are observed time-invariant covariates (e.g., race/ethnicity), and φ_{it} is the error term. Our naïve estimate of α_1, the "effect" of education on health, is almost certainly biased by omitted variables and/or reverse causality, and this would be equally true in any other regression model form such as logistic or Poisson. Estimating the causal effect requires isolating variation in the exposure that is both uncorrelated with unobserved confounders and not affected by the outcome.

The fundamental challenge is therefore selecting a control group that meets this criterion: it does not differ from the treatment group in any systematic way that would bias the estimated treatment effect. The intuition is the same as controlling for observed confounders in regression analysis: what is the effect of maternal education on child health holding factors like maternal age, race, and health status constant? The quality of inference in this case depends on the credibility of the assumption that, conditional on these measured confounders, maternal education is as good as randomly assigned. While this is certainly a step in the right direction relative to a crude model, it is still unsatisfactory if important potential confounders like motivation, paternal health status, or social class remain unobserved. Some of these confounders could be measured with more and better data, but others are likely to remain poorly measured or unmeasured no matter which dataset is used. The added value of fixed effects (FE) models is their ability to control for both observed and stable unobserved confounders, which lends greater credibility to the assumption necessary to estimate unbiased causal effects. The difference-in-differences (DD) model goes one step further to leverage external factors that lead to changes in an individual's exposure status.

Trim Size: 7in x 9.25in
Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 345

Generally speaking, the class of unobserved confounders that FE models can account for are those characteristics that remain fixed over time within observed units of analysis, whether individuals or states (we use "states" throughout this chapter as an example of a jurisdictional unit, which could easily be countries, provinces, cities, schools, hospitals, etc.). While this will not capture all potential unobserved confounders, it will surely include a large set of important factors that we expect to be correlated with both the exposure and the outcome. For example, women who get more education may also take fewer risks and invest more for their futures; states or countries with universal public health insurance may also provide a broader range of other social programs. Rather than try to measure each of these factors and include them in the regression model, FE and DD models will control for all factors—both observed and unobserved—that are constant over time within individuals or states. The variation in the exposure that we use to estimate the causal effect must therefore stem only from *change* in the exposure over time.

DD models can be considered a special case of FE models. Both FE and DD models include "fixed effects" for individuals or states that control for time-invariant ("fixed") confounders. FE models can be estimated using longitudinal/panel individual-level data (where exposure changes for at least some individuals) or on aggregate longitudinal/panel state-level data (where exposure changes for at least some states). A distinction to note is that in an FE model, the change in exposure may be under the control of the individual or state, depending on the level of observation in the analysis. In DD models, changes in exposure are a function of decisions made outside of the unit of observation (e.g., a policy change happens at the state level, but the analysis uses individual-level data). For both FE and DD models, the most important determinant of our confidence that the effect estimate is indeed causal is whether the exposure change is plausibly unconfounded. This must be evaluated using both data and knowledge about the specific context in order to make a credible claim that the analysis estimates a causal effect with little or no confounding bias.

The potential sources of confounding and the types of fixed effects we include to address it inform the distinction between these types of models. In a FE model with individual-level data, individuals make decisions that lead to the change in their exposure status (they drop out of school after primary or decide to go to university), so the source of confounding is unobserved individual-level factors and we include individual-level fixed effects to address it. In an FE model with state-level data, states make decisions that lead to the change in their exposure status (they change educational policy), so the source of confounding is unobserved state-level factors and we include state-level fixed

effects to address it. In a DD model with data on individuals within states, the states make decisions that lead to the change in individuals' exposure status (similarly to an RCT with imperfect compliance, states change their educational policy, which should affect the amount of education individuals attain). The source of potential confounding is unobserved state-level factors, so we include state-level fixed effects to address it.

Fixed Effects Models

Setup. Fixed effects models are useful when we have panel data, that is, repeated observations on the same unit over time. While we often think of individual-level panels with multiple observations for the exact same people at different points in time, we can also consider aggregate data that contain repeated observations of geographic units (cities, states, countries), families, and so on. If we consider using a 10-year panel of individual women to estimate the impact of their education on their children's health, the FE model will estimate this effect based on women who experience a change in their level of education over those 10 years. Note the key difference between using individual-level changes in education rather than comparing outcomes for women with more education with those for women with less education, as the naïve regression model would. The FE model effectively compares each woman with herself at an earlier time as her control. Moreover, the FE estimate of maternal education on child's health controls for all time-constant observed and unobserved confounders including the mother's ability, genetics, and, if they remain constant, her parents' education and social class. In parallel, an aggregate FE model with state-level data will estimate the effect of average mothers' education on average child health outcomes based on states that experience a change in the average level of mothers' education over those 10 years. Each state serves as its own control, and the FE estimate holds all state-level, time-constant observed, and unobserved confounders constant. The question of why education changed for some women and not others, or in some states and not others, and whether this variation is unconfounded, should be addressed qualitatively and quantitatively as part of the analysis and interpretation.

Basic FE Regression Framework. The basic unit-level FE regression is as follows:

$$Y_{it} = \gamma_0 + \rho_i + \gamma_1 Exposure_{it} + \gamma_2 X_{it} + \gamma_3 T_t + \omega_{it} \quad (2)$$

where all variables are defined as above and T_t denotes the time period capturing common secular trends; ρ_i is a unit-specific intercept ("fixed effect"), capturing all time-invariant factors that are correlated with the outcome. Because

ρ_i controls for all time-invariant characteristics of the woman (or state), the estimated effect of education γ_1 is based on individuals who change their exposure status over time. All individuals contribute to the coefficient estimates on the other time-varying covariates X_{it}, but only women whose exposure status changes contribute to the estimate of γ_1 (Gunasekara *et al.* 2014). As should be clear by examining Equations (1) and (2), the error term in Equation (1), φ_{it}, includes the time-invariant characteristics, ρ_i, therefore biasing the estimate of α_1 in Equation (1). In contrast, in Equation (2) γ_1 estimates the effect of an actual *change* in education in the same individual, thus holding all other time-invariant individual characteristics constant.

The FE model can be estimated via unconditional or conditional maximum likelihood. When FE is estimated with conditional logistic regression for a categorical outcome, individuals with no change in the outcome will be dropped from the regression, and therefore do not contribute to any of the estimated coefficients (Gunasekara *et al.* 2014). The FE estimator with two time periods is equivalent to a model where changes in the outcome are regressed on changes in the exposure ("first-differenced estimator"), since in both cases confounders that are constant over time are controlled for. Further exposition of this point, as well as a discussion of differencing with more than two time periods, can be found in Angrist and Pischke (2009) and Wooldridge (2013, Chapter 13).

Similar to a case-crossover design (Maclure and Mittleman 2000), in the FE model each woman whose exposure status changes serves as her own control. Therefore the assumption is that her outcome before the change is a reasonable counterfactual for what her outcome would have been at a later time in the absence of an increase in educational attainment. Since fixed characteristics that vary across women are controlled for, particularly unobserved confounders, this assumption may be more plausible than assuming that a control group of different women provides a valid counterfactual. However, if other confounders are changing concurrently with the exposure change, a causal estimate based on only those who change exposure will not be valid. This is illustrated by the fact that in the simplest model with no X_{it} or T_t, the estimate of γ_1 will be the same whether the sample includes only individuals whose exposure status changes, or both changers and non-changers—the estimate of γ_1 is informed by only those who change exposure. However, if secular time trends or other time-varying confounders are important, the effect estimate will vary depending on the sample. Once γ_1 is conditional on other covariates, X_{it} or T_t, it will vary depending on whether non-changers are included in the estimation sample. This desire for a counterfactual based on an external control group to account for time-varying confounding is part of the motivation behind the DD models presented below.

Key Assumptions for FE Models. The FE model is useful for causal inference because it controls for all fixed characteristics, both observed and unobserved, that may confound the estimate of the effect of education on health. The remaining assumption necessary to preclude confounding is therefore that all time-varying factors that are correlated with both the outcome and the exposure are included in the regression model. Care is warranted in interpreting the coefficients on time-varying variables (Kaufman 2013). For example, since educational achievement can only increase, a woman who changes her educational level must also be older at the higher educational level, and her child must also be older. As usual, the form of the model must also be correctly specified, and so flexible specifications (e.g., splines or interactions of time-varying covariate terms) may enhance credibility of inferences.

Interpretation of Estimates from FE Models. The FE estimate can be considered a "treatment on the treated" (TOT) or "average treatment effect among the treated" (ATT), since it is based on those individuals whose exposure status changed (i.e., we are asking about the counterfactual of what would have happened to the treated group had they been untreated). To inform policy decisions, an estimate of the average treatment effect among the untreated (ATU) is also valuable. Whether the average treatment effect is constant (ATT = ATU = ATE), and therefore whether the ATT should be generalized to other groups in the population, is an important question common to all study designs. In addition, even if a state-level FE model provides an unbiased estimate of a causal effect at the aggregate level, the interpretation should not extend to the individual level to avoid committing the ecological fallacy (Greenland 2001).

Another question of interpretation involves whether the FE estimate is indeed causal. To address this point we must return to the idea of variation in the exposure that is "as good as random," and why certain individuals (or states) changed their education, neighborhood, or their employment status while others did not. Was this change plausibly unconfounded? Is the causal effect of the exposure on the outcome free of time-varying confounding and reverse causality? We argue that seeking to identify the causal effects of such exposures using individual-level changes will generally be more productive than comparing across individuals with different levels of exposure, though there may be tradeoffs between reducing bias and increasing imprecision (Kaufman 2008). A large number of potential unobserved confounders—those that are constant over time—are already controlled for by individual- or state-level fixed effects, which lends more credibility to this approach. However, if unobserved time-varying confounders remain

a concern, caution should be exercised in interpreting the FE estimates as causal. If a woman did not complete her schooling because of a financial crisis, her exposure status is a function of this time-varying confounder that can also affect the child health outcome. In that case, it must be measured and adjusted for to obtain a causal estimate. In an FE model, we observe how outcomes change as individuals change their education, move to a better neighborhood, marry or divorce, or retire or move from part-time to full-time work. However, the chance that these exposure changes are "as good as random" in the lives of individuals is slim. For example, if our outcome of interest is health status, it is probable that health concerns have some impact on the decision to retire, leading to concerns about reverse causality (Disney *et al.* 2006). Similarly, states' levels of average educational attainment could be affected by improvements in health.

For these reasons, changes in exposure that are driven at least in part by changes in factors external to the individual (policies, laws, public programs, area-level conditions, etc.) can often be especially useful to estimate causal effects reliably. Individuals' changes in educational achievement in the presence of new mandatory minimum education laws (e.g., Glymour *et al.* 2008; Mazumder 2008) or subsidies to offset schooling costs are more plausibly unconfounded than changes in the absence of such incentives that could be due to a wide range of factors. The situation is similar for changes in marital status in the presence of new laws (such as same-sex marriage laws) and for changes in employment status as retirement incentives, area-level unemployment rates, or other policy changes. Any FE analysis should incorporate an examination of the underlying reasons for variation in the exposure and address whether it generates estimates that can indeed be interpreted as plausibly causal. This interpretation ultimately rests on the convincing story that, within units, exposure is plausibly as good as random with respect to the outcome. In this respect, FE models are not a recipe for automatically producing estimates of causal effects. They provide some clear advantages over naïve models with respect to time-invariant confounders, but if time-varying confounders or reverse causality are a relevant concern, FE models can still provide biased estimates of the effect of interest.

FE Extensions and Considerations. Fixed effects analyses are less commonly used than random effects analyses in much of social epidemiology. Additional comparison between fixed effects and random effects models is covered in Chapter 13 by Cerdá and Keyes and Chapter 15 by Hirai and Kaufman in this volume. In the context of trying to estimate a causal effect, the key difference is what we assume about the correlation between the exposure and

the individual-specific intercepts ρ_i in Equation (2). In order for the random effects model to provide consistent estimates, these must be independent. In many cases, however, we expect both observed and unobserved individual characteristics to be correlated with the exposure (which is why we want to control for them in the first place), rendering such an assumption invalid. In such cases, FE models are generally preferred to random effects models (Allison 2009; Kravdal 2011; Harper *et al.* 2012; Gunasekara *et al.* 2014). While the Hausman test can be used to test whether the FE and RE estimates are statistically different, and therefore whether the RE assumption is likely to hold, a substantive argument based on the causal model and the context at hand is also strongly recommended to justify using random effects models for causal inference. This is because the Hausman test, like all statistical tests, requires consideration of power and the substantive importance of the magnitude of difference detected (Wooldridge 2013, Chapter 14; Clark and Linzer 2015).

In considering whether FE estimation can be a useful tool for social epidemiologists, it must also be noted that, in its simple form, the model does not allow one to examine the impacts of exposures that are themselves time-invariant. These include such factors as race/ethnicity, gender, and birth cohort, exposures about which there is a large accumulated literature and continuing interest. Because these factors generally remain constant over time within individuals, their effects are combined in the unique individual-specific intercept, which is a composite effect of all fixed factors—both observed and unobserved—for that individual. While an entire argument exists regarding whether causal effects of these exposures can actually be measured (Kaufman and Cooper 1999, 2001; Krieger and Smith 2000), we will limit our remarks here to noting that their associations cannot be estimated with standard FE models. Nonetheless, a "hybrid" model can be specified that offers the advantage of fixed effects inference while also allowing estimation for the time-invariant covariates (Hirai and Kaufman, Chapter 15 in this volume). Allison (2005) also discusses including interactions of fixed characteristics with time-varying characteristics (including time itself) as control variables.

Difference-in-Differences

Setup. Fixed effects models use within-unit changes over time to estimate the causal effect, with units serving as their own controls. If time-varying confounding remains a concern, an external control group may help provide a counterfactual for what would have happened to the units with exposure changes in the absence of that change. Difference-in-differences models

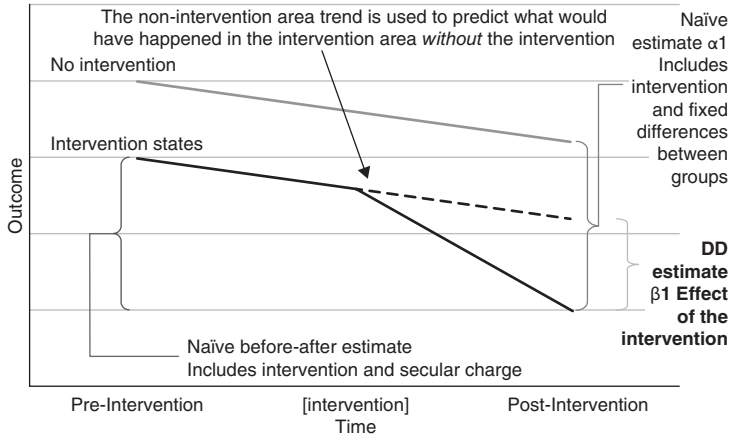
estimate the effect of exposure by using changes over time in a treatment group relative to a control group (Meyer 1995; Angrist and Pischke 2009; Dimick and Ryan 2014). DD models also differ in that the treatment group's exposure status changes over time due to changes at a more aggregate level (policy or administrative rule change), while the control group experiences no change in the policy or the rule governing exposure (remains exposed or unexposed). This design mimics a controlled trial, but one with no randomized assignment to the two groups. In a randomized trial, the exchangeability of the groups is ensured by the randomization, which balances all characteristics in expectation. In a DD model, on the other hand, exchangeability is asserted based on examination of time trends before the policy change. It is also similar to a case–time–control design (Suissa 1995).

Repeated cross-sectional datasets (e.g., the Behavioral Risk Factor Surveillance Surveys, Demographic and Health Surveys) are commonly used by social epidemiologists. Large, nationally representative surveys, for example, usually involve stratified random samples and are conducted year after year. Because the same individuals are not surveyed each year, individual-level FE models are not feasible and aggregate state-level FE models sacrifice valuable information at the individual level, particularly for considering effect heterogeneity in subgroups (Petticrew *et al.* 2012). DD models that use individual-level data and leverage exposure contrasts driven by aggregate-level policy changes can be particularly useful in this context.

Population-level changes in exposure commonly arise due to policy changes at some level of governance. As such, individual exposure status is plausibly unconfounded—neither driven by outcomes nor unobserved confounders at the individual level. However, confounding at the aggregate level may still exist since states or countries that enact policies may be compositionally different from those that do not (e.g., Macinko and Silver 2015). Potential state-level confounders are very commonly differences across states that do not vary much over time (at least over the time frame of most analyses), some of which are observed (e.g., socioeconomic composition) and others not (e.g., social norms). Secular trends that affect both the treatment and control groups may also confound naïve estimates.

The DD design therefore utilizes policy *changes* rather than time-invariant policies that differ across jurisdictions. By controlling for all fixed differences between states and shared changes over time, the DD model focuses on changes in the exposure of interest that occur in some states but not others and can thereby estimate the unbiased causal effect of the exposure. Returning to the previous example, rather than comparing child health between areas with lower versus higher maternal education or before versus after the policy change in

FIGURE 14.1. GRAPHICAL EXAMPLE OF DD ESTIMATE



affected jurisdictions, the DD model compares *changes* in child health in areas that experienced a change in their level of education due to a policy change (one difference) relative to changes in child health in areas that do not change their exposure status (a second difference) (see Figure 14.1).

Basic DD Regression Framework. The basic DD regression with two groups, exposed ($j = 1$) and unexposed ($j = 0$), and two time periods representing pre- ($t = 0$) and post-policy change ($t = 1$) is as follows:

$$Y_{ijt} = \beta_0 + \beta_1 E_j + \beta_2 Post_t + \beta_3 E_j \times Post_t + \beta_4 X_{ijt} + \varepsilon_{ijt} \tag{3}$$

where Y_{ijt} is the outcome for individual i in group j at time t, E_j is an indicator variable for exposure group j, $Post_t$ is an indicator variable for time t being after the policy change, X_{ijt} are individual-level covariates, and ε_{ijt} is the error term. E_j is equal to one if the observation is in a state that changes its policy, regardless of the value of t, and equal to zero in a state that does not change its policy. $Post_t$ is equal to one if the observation occurs after the policy change, regardless of the value of j. The interaction term therefore equals one only for observations that are in the exposed group after the policy change. The estimated coefficient β_3 reveals any change in outcome Y from the pre-policy time to the post-policy time that occurs in the exposed group and not in the unexposed group.

To see this point, consider Tables 14.1 and 14.2, representing the potential outcomes $Y_{E,Post}$ in each exposure group and time period, as well as the regression coefficients that estimate these quantities.

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 353

TABLE 14.1. DIFFERENCE-IN-DIFFERENCES IN POTENTIAL OUTCOMES

	Pre	Post
No change in exposure	Y_{00}	Y_{01}
Change in exposure	Y_{10}	Y_{11}

TABLE 14.2. DIFFERENCE-IN-DIFFERENCES IN REGRESSION COEFFICIENTS

	Pre	Post
No change in exposure	β_0	$\beta_0 + \beta_2$
Change in exposure	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

The linear "difference-in-differences" estimate is therefore

$$(Y_{11} - Y_{01}) - (Y_{10} - Y_{00}) = [(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2)] - [(\beta_0 + \beta_1) - \beta_0] = \beta_3 \quad (4)$$

While this two-group, two-period case illustrates transparently how this estimator identifies the causal effect, many applications of this method involve policy changes that affect multiple states using data over multiple time periods. Equation (3) can be easily expanded to include multiple groups and multiple periods (Imbens and Wooldridge 2009; Angrist and Pischke 2015). One could replace E_j with multiple indicator variables for different states (some exposed and some not), $Post_t$ with multiple indicators for different years (some pre- and some post-), and the $E_j \times Post_t$ product term with a time-varying treatment indicator such as $Treat_{jt}$, which reflects interactions of the state and year indicator variables (i.e., exposure to the policy varies over time within at least some units). It is also fairly common that the policy change does not happen at the same time for all exposed states. In this case, the data can be centered for each state at "time-zero," the time of the policy change (or at a later time to accommodate any desired etiologic lag). The fact that the pre- and post-periods vary across years across states can be addressed with controls for calendar time, in the case that there are secular time trends in the mean outcome. In some studies, all states are eventually exposed to the policy change. Thus the model estimates the effect of the policy by comparing exposed states whose policy changes at a given time to control states who do not experience a change at that same time (Strumpf 2011).

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 354

Key Assumptions for DD Models. Much as in randomized controlled trials, the DD model makes the identifying assumption that the control group serves as an adequate proxy for the counterfactual outcome we would have observed in the treatment group had they not been treated (Ryan *et al.* 2015). The choice of control group is therefore of fundamental importance for the validity of the DD estimate. The identifying assumption rests on parallel trends in the outcome in the pre-period, which suggests that the trend in the outcome among the unexposed in the post-period provides a good counterfactual for what would have happened to the exposed group in the absence of exposure. Baseline equivalence of the outcome is not necessary, since any time-invariant difference between the two groups (including at baseline) is subtracted out (β_1 in Equations (3) and (4)). Establishing parallel pre-period trends of course requires that there be more than one time point observed before the advent of the exposure change. In the simple two time-period scenario, similarity of outcome values in the pre-exposure period between the exposed and unexposed groups should provide some reassurance. However, DD models using only two time points will always be somewhat suspect, since in this case the causal identification rests on the specification of the additive model form, and cannot be investigated by confirming that the two groups change in parallel over time before the policy takes effect (Angrist and Pischke 2009; Imbens and Wooldridge 2009).

In order to interpret the DD estimate as a causal effect, the analyst must make a convincing qualitative argument that these two groups are indeed exchangeable, relying on observed data and other available evidence that can speak to (the lack of) unobserved confounders. The first step is usually a "Table 1" that compares observed characteristics and outcomes in the pre-exposure period between exposed and unexposed states. Although they are unlikely to be as evenly matched as in a randomized controlled trial with large samples, the two groups should be reasonably comparable in terms of both substantive and statistical differences. Because the DD estimate controls for all fixed differences between groups, some differences in the levels of measured confounders or outcomes that can be expected to persist over time (one group is consistently older, has higher average income, etc.) need not be a threat to validity. Such differences may, of course, raise concerns about potential unmeasured differences between the groups that are not controlled for with group fixed effects (i.e., confounding factors that vary over time in some states but not in others), concerns any analyst should be prepared to address. Stuart and colleagues (2014) recently suggested a weighted DD regression model, with weights chosen as inverse propensity scores. This would serve to minimize any imbalances between the two groups based on measured baseline confounders.

Trim Size: 7in x 9.25in
Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 355

Given a series of observations before the intervention, the two groups may then be examined with respect to trends in the outcome. The degree to which pre-intervention trends in the outcome are parallel in the exposed and unexposed groups is usually assessed graphically, and sometimes also with statistical tests in a regression framework (Ionescu-Ittu *et al.* 2015). This assessment requires attention to statistical precision and whether the magnitude of any differences is substantively important. The potential for differential compositional changes in the two groups should also be considered. This can be done by examining trends in characteristics that should not change differentially over time, particularly those that might indicate selective migration in response to the policy change (Levine *et al.* 1999; Joyce and Kaestner 2001). These trends can be examined in the pre-exposure period, or over the entire period if the characteristics in question should not plausibly be affected by the policy change being studied.

A second assumption required to interpret the DD estimate causally is that the policy change is (conditionally) exogenous. This means that the policy change is not driven by pre-policy outcomes (no reverse causality), nor by any unmeasured time-varying common cause of the policy and the outcome (no confounding). This assumption can be supported empirically by checking whether pre-policy outcomes predict the policy change. Thus the key assumption is no unmeasured changes over the study period that affect outcomes in the two groups differentially. Any correlation between the policy change and time-invariant factors will be controlled for by the group fixed effects. While ultimately no assumption regarding unmeasured confounders is empirically testable, substantive knowledge regarding the reasons for policy changes and the conditions under which they occurred can help establish the reasonableness of the casual inference (see Cohen and Einav 2003 for an example).

Interpretation of Estimates from DD Models. The DD model estimates an average causal effect of the treatment on the treated group—a new law making primary education mandatory, for example. The contrast is between a factually treated group at some specified time and a stand-in for what would have been observed in that same group had they, counter to fact, not been treated at that time. However, remember that the use of the DD approach was also motivated by finding unconfounded variation in the exposure of interest, education in our running example. In reference to the effect of education on health, the DD approach estimates the effect of treatment assignment to the policy, but it is likely that not all people living in an "exposed" state are subject to or compliant with the policy. This is analogous to the intent-to-treat estimate (ITT) in a randomized trial with non-compliance. Alternatively, consider that an increase

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 356

in cigarette taxes may reduce smoking intensity on average, but any individual can opt to maintain or even increase their smoking if they wish. The effect estimated by DD is therefore not the effect of smoking per se, but rather the effect at the population level of an increase in cigarette taxes: a very specific mechanism to discourage smoking. If the degree of compliance—the extent to which the policy change results in changes in smoking behavior—is known, the causal effect on the treated group can be recovered from the ITT by scaling the DD estimate by the compliance rate. This is in contrast to a policy such as banning an environmental contaminant which, if effective, eliminates exposure for everyone without requiring any active compliance by individuals. Here, 100% compliance means that the ITT and causal effect on the treated population are equivalent.

Care must also be taken in generalizing DD estimates. They are "local" treatment effects in the sense that they refer to subpopulations. The ITT estimates are based on the states that implemented policy changes, and the effect of treatment on the treated estimates are based on individuals who change their behavior in response to the policy change (compliers). Whether these groups are relevant and interesting in and of themselves and/or whether these estimates can be reasonably applied to other subpopulations or to the entire target population is an important part of interpreting the results from studies using the DD approach.

DD Estimation and Extensions.

Model Specification. Recall that the DD strategy relies on interaction between time (pre/post) and treatment status. As with the use of any interaction term, the main group and time fixed effects must be included in the regression in order to correctly interpret the DD estimate as the additional contribution of being in the exposed group after the policy relative to the unexposed group. If a higher-order interaction term is used (see the discussion of triple-difference models below), all main effects and lower-level interaction terms must also be included so that the model remains hierarchically well specified. Moreover, interaction terms always have reduced power in statistical tests compared to main effect terms, meaning that tests on the "significance" of β_3 as the causal estimate may lead to high rates of Type II error (Greenland 1983).

Given the preference for non-linear models in epidemiology, including logistic, Poisson, Cox, and binomial regressions, the interpretation of the product interaction term of a DD model as an estimate of causal effect requires special care. In the case of a binary outcome, the DD model could be estimated using a linear probability model (OLS), a generalized linear model with a binomial distribution and an identity or log link, or a logistic regression

model. The first two models make the implicit assumption that the joint effect pattern between exposure and time is additive, while the latter two assume that it is multiplicative. With only two time periods, this modeling choice can only be based on an outright assumption about the functional form, whereas with more pre-period data points the appropriate functional form can be estimated from the data (VanderWeele and Knol 2014). The DD literature from economics and public policy relies almost entirely on the additive scale as the default null and, indeed, this scale has attractive theoretical and interpretative properties in epidemiology as well (Kaufman 2010). If analysts prefer to fit non-linear models because of their statistical advantages, interpretation can be enhanced by simply manipulating estimated coefficients to form absolute or marginal probability contrasts (Carpenter 2009; Harper *et al.* 2014; Muller and MacLehose 2014).

State-specific trends are sometimes added to control for potential confounders that may be changing over time in a linear way (Ryan *et al.* 2015). For the addition of these state-specific trends to address omitted variables bias, they must be correlated with the timing of the intervention and therefore should be added only if there is reason to believe this is the case (Angrist and Pischke 2015). Moreover, it should be emphasized that the standard DD model (as in Equation (3)) tests whether there is a contemporaneous change in the *level* of Y in the post-treatment period. However, if the true effect affects the *trend* of Y in the treated group in the post-period, then including unit-specific trends will generate bias (Meer and West, 2016). If used, only trends based on the pre-intervention period should be included in order to control for factors that change differently across states *before* the advent of the intervention. Including post-period state-specific trends can generate bias, since these may be influenced by any effect of the policy itself. Caution has also been noted about adding state-specific trends to a model with a single post-policy indicator when the effect of interest may be dynamic and could change over time (Wolfers 2006; Angrist and Pischke 2015). In this case, state-specific trends will capture differential pre-existing trends as well as differences in the evolution of the outcome between exposed and unexposed states in the post-policy period, again potentially leading to bias. With multiple time periods, a better alternative is to first investigate whether the policy effects are indeed dynamic (interact the exposed indicator with an indicator for each year) and then consider whether state-specific trends are needed to address confounding in the pre-period. In sum, substantial consideration of the specific context at hand is warranted before adding state-specific trends to address unobserved confounding.

It is widely recognized that observations in DD analyses are generally not independent. Concerns arise about both correlation between individuals in a state at a point in time and serial correlation for the same state over time (Bertrand *et al.* 2004; Donald and Lang 2007). While it has become standard practice to use the cluster robust variance estimator in DD models to account for the cross-sectional correlation within states (Williams 2000), this solution is probably not entirely adequate. Other alternatives include aggregating data, block bootstrapping if the number of states is large (Bertrand *et al.* 2004; Kolstad and Kowalski 2012) and permutation tests for inference (Abadie *et al.* 2010; Buchmueller and Marko 2014). While there may be no best practice across all scenarios involving different numbers of states and time points, it seems clear that formal statistical inference becomes questionable as the number of states becomes small (Bertrand et al. 2004; Cameron and Miller 2015). Therefore inference from a DD model based on 50 US states is more reliable than one based on 10 Canadian provinces, although it will also depend on the division of states between exposed and unexposed groups and the number of time points.

Validity and Robustness. A common approach to assessing the DD model's validity and robustness is the use of so-called "placebo" tests, in which the analyst tests to see whether the DD model detects an "effect" when it logically should not, or for an outcome that should logically be unrelated to the policy intervention (Lipsitch *et al.* 2010). This requires contextual knowledge about when and how the policy was implemented, and substantive knowledge about outcomes that should not have been affected by the policy. For example, McKinnon and colleagues performed a DD analysis for the effect of antismoking legislation in Quebec on adverse birth outcomes, finding consistent effects on birthweight and pre-maturity. As a sensitivity analysis, they repeated the model using an arbitrarily chosen false date for the intervention, for which no intervention effects were detected (McKinnon *et al.* 2015a). Likewise, Riddell *et al.* considered the effect of a hospital's previous uterine rupture on subsequent rates of vaginal delivery in women with a previous cesarean, revealing a transient dip in the willingness of the doctors to allow women to continued attempted labor. As a robustness check, the authors then considered the effect of the hospital's rupture history on diabetes diagnoses as a placebo condition that could not plausibly have affected the exposure. Reassuringly, there was no effect of the exposure on the placebo outcome (Riddell *et al.* 2014).

Policy exposures are naturally aggregate, applied to whole states, provinces, or countries, and therefore concerns sometimes arise around the potential for the ecological fallacy (Greenland 2001). This fallacy is to

attribute individual-level causation to an association observed only at the aggregate level. However, when the DD model estimates the effect of an aggregate exposure on individual-level outcomes, it is precisely the question of interest to understand the effects of policy changes. As is true more generally, care should be taken so that inferences drawn are consistent with the levels of observation in the analysis.

Extensions. A number of extensions to the basic DD model exist. In the context of a continuous or multicategory instead of binary exposure, the DD model can be modified accordingly (Duflo 2001; Angrist and Pischke 2009; Dunkley-Hickin 2014). Effect measure modification can be investigated by interacting the DD interaction term with a third variable, for example, education or income (e.g., Harper *et al.* 2014). Effect heterogeneity over time or across states can be allowed by including indicator variables for each time period (e.g., year) or each treated state.

When it is known based on the policy design that a subgroup within treated state × time strata is ineligible or should be unaffected by the policy, a triple-difference or DDD estimator may be used. For example, in Strumpf (2011) the Medicaid program was implemented in different states in different years, and women with children were eligible while women without children were not. As stated previously, all three main effects and three two-way interaction terms must be included for the model to be well specified. Therefore, in addition to controlling for fixed differences across states and shared changes over time, group-specific time trends and group-specific differences across states are also controlled for, significantly strengthening the case for causal inference. The hope is that the remaining list of potential unobserved confounders that vary by time, state, and subgroup diminishes accordingly. DD models may also be used in concert with other methods to improve the exchangeability of the exposed and unexposed groups, including propensity scores and matching (Stuart *et al.* 2014) and synthetic controls (Abadie *et al.* 2010; Bauhoff 2014).

General Considerations for Both FE and DD

It may be the case that the number of individuals or states that change their exposure, those that contribute to the estimate of the effects (γ_1 in the FE model and β_1 and β_3 in the DD model), are relatively few in number or are quite different from the general population of interest. If relatively few individuals change their educational attainment, for example, then the effective sample size can be quite limited, leading to imprecise intervals and low power

Trim Size: 7in x 9.25in
Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 360

for tests against the null (Bell and Jones 2015). The small effective sample size may also have implications for the generalizability of the estimate, so care must be taken with interpretation on that front as well.

With longitudinal data the question of lags and dynamic effects are important considerations that must be resolved statistically or substantively. For example, in a fixed-effects case-control study of flooding in relation to gastrointestinal emergencies, Wade *et al.* (2014) determined that there was an acute effect, but by five days after the flood the exposure effect fell to the null. In contrast, in a 2014 DD study on the relation between tobacco taxes and overall mortality, the authors postulated that a change in tax policy could not produce an acute effect on mortality. Rather, it would have to impact on smoking behavior over several years to begin to change mortality patterns. Therefore, these authors imposed a five-year lag after the policy change to begin looking for the impact on mortality (Bowser *et al.* 2016).

Another consideration regarding logistic regression is the incidental parameters problem, which epidemiologists know as the small sample bias of the odds ratio (Greene 2004). The coefficient parameters of the logistic regression model are not consistently estimated by unconditional maximum likelihood when the size of each stratum is small, as it often is in fixed-effects panel models that have only a few time periods per unit. Conditional logistic regression is the standard tool in epidemiology to overcome this problem, although this too can produce important small-sample bias in the presence of many covariates or small numbers of units (Greenland *et al.* 2000). Researchers may also consider linear probability models, which yield similar results to logistics when outcomes are not rare (Long 1997).

Finally, like all regression models, the promise that fixed effects remove confounding from all time-invariant factors is only valid when the form of the model is correctly specified. Statistical inferences require the correct error term distribution and variables are typically assumed to be measured without error.

Applications

By way of illustration, we briefly describe two papers that have rigorously implemented the FE and DD methods to answer causal questions in social epidemiology.

Blakely and colleagues (2014) studied the causal relationship between improving social circumstances, such as finding a job or moving into a good neighborhood, on tobacco use. To avoid confounding and reverse causation,

they used an FE design in which approximately 15 000 individuals were followed longitudinally and the within-person changes in social circumstances were considered in relation to within-person changes in tobacco use. Only about 2% of participants changed their smoking status between waves, but exposure changes were somewhat more common: 10% of respondents experienced a change in labor force status between waves, nearly a third had a decrease or increase in log income of half a standard deviation or more, and one in five changed neighborhood. The authors observed both important increases in smoking with income gains in young participants, as well as increased odds of smoking for those who relocated to neighborhoods with poorer conditions. In a related paper using the same within-person design, Ivory and colleagues (2015) adjusted for additional time-varying neighborhood characteristics such as neighborhood smoking prevalence. They found that a one decile increase in neighborhood deprivation between waves was associated with an 8% increase in the odds of smoking. This neighborhood effect was modest compared to the influence of the home environment. For example, moving in with a smoker more than doubled the odds of smoking compared to moving in with a non-smoker.

In another recently published example, McKinnon and colleagues (2015b) considered the effects of a policy that removed user fees for facility-based deliveries on the proportion of births occurring in facilities within low-income countries. Moreover, the authors also considered whether the removal of fees affected socioeconomic inequalities in facility-based births (McKinnon *et al.* 2015c). Using Demographic and Health Survey data from nine sub-Saharan African countries, three of which had eliminated user fees during the study period, the authors applied DD models to control for secular trends and time-invariant differences among countries, and they allowed for differential effects of the policy by the socioeconomic position of the mother. The analysis is premised on the assumption that changes in the proportion of facility deliveries by socioeconomic position that are due to factors other than the policy do not differ between the intervention and control countries. The authors checked this assumption by ensuring that trends in the proportion of facility-based deliveries by socioeconomic position were similar for the intervention and control areas prior to introduction of the policy. They reported weak evidence of differential effects of removing user fees across wealth quartiles, but results suggested that educated women benefited more from removing user fees compared to women with no education: a difference of 8.6 facility deliveries per hundred live births (95% CI: 5.4, 11.9) among women with secondary education versus a difference of 4.6 per hundred (95% CI: 2.2, 7.0) for women with no education. Thus, the intervention appears to

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 362

benefit all social groups while at the same time disproportionately benefiting the most advantaged women, potentially exacerbating inequality between educational strata.

Conclusion

All causal inference based on observational data rests on finding an adequate substitute population for the unobservable counterfactual of interest. For example, for the average effect of the treatment on the treated, we observe the treated group under treatment, but we need a group to stand in for the outcome that would have been observed in these same individuals if, counter to fact, they had not been treated (Hernán and Robins 2006). When the outcomes in the chosen substitute population differ systematically from what would have been observed counterfactually in the treatment group, epidemiologists refer to the bias in the estimated causal effect as "confounding." The epidemiologic tradition, following from a larger biomedical culture, has traditionally approached this problem using either randomization or statistical adjustment for measured confounders. These tools have been less persuasive for social epidemiology, however, because in most cases it is difficult to randomize social exposures and to enumerate and successfully measure all of the important confounders. This has led the field to consider other more convincing approaches to causal inference, two of which we reviewed in this chapter.

One approach is the fixed effects model, which uses changes in exposure at the level of observation (i.e., individual, state). Each individual's counterfactual is its own outcome before the exposure change. This effectively matches on all time-fixed characteristics that might generate confounding in other designs. The second approach is the difference-in-differences model, in which a control group is chosen and the model uses changes in exposure determined outside of the unit of observation. The appropriateness of the chosen control group as a counterfactual for the treatment group is evaluated in the pre-exposure period. Then the trend across time is compared in the exposed and the unexposed groups, and any deviation from the same time trend in the post-exposure period is attributed to the exposure.

Both of these methods have the potential to reveal causal effects if their assumptions are satisfied, and so careful examination of these assumptions is always necessary. In most cases it is possible to observe that an assumption is violated, but impossible to prove that it is not. This motivates an approach in which the sincere investigator tries diligently to find evidence of a falsified assumption. If all attempts at falsification fail, then the causal story told by

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 363

the investigator becomes increasingly persuasive. The demand for reliable and valid evidence that informs how to design policies to reduce health inequalities continues to mount. Multiple investigators using multiple methods that provide this kind of robust evidence of a consistent, causal relationship between exposure and outcome will help inform policies that have more rational and secure bases.

Key Readings and Resources

Here we provide just a few key references for interested readers. Meyer (1995) provides a nice overview of quasi-experimental designs, which get a more in-depth treatment in Shadish, Cook, and Campbell (2001) and are nicely summarized and illustrated in a non-technical way by Gertler *et al.* (2011). Angrist and Pischke (2009, Chapter 5 and 2015, Chapter 5) give excellent applied overviews of fixed effects and difference-in-differences designs, and Wooldridge (2013, Chapters 13 and 14) provides a more formal textbook treatment. FE and DD models are easily implemented using any standard regression software (e.g., SAS, Stata, R), but some example code is useful. For applied researchers, Allison (2005) gives fixed effects models a full treatment using SAS, and there are user-written Stata packages that attempt to make estimating DD models easier (Allison 2009; Villa 2014; Linden 2015). The supplemental appendix to Ryan *et al.* (2015) provides Stata code for estimating DD models and testing key assumptions, and the online supplement to Harper *et al.* (2012) contains data and Stata code. For R users, Kim and Imai (2014) have recently published a package that includes DD estimation.

Acknowledgments

The authors acknowledge funding support from the Fonds de la Recherche du Québec–Santé (FRQS) and Ministère de la Santé et des Services Sociaux du Québec (ECS), FRQS (SH), and the Canada Research Chairs Program (JK). They would like to thank Jacob Bor, Ashley Hirai, Corinne Riddell, and Sahar Saeed for helpful comments and suggestions.

References

Abadie, A., Diamond, A., *et al.* (2010) Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105 (490), 493–505.

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 364

Allison, P.D. (2005) *Fixed Effects Regression Methods for Longitudinal Data Using SAS*, SAS Institute, Cary, NC.

Allison, P.D. (2009) *Fixed Effects Regression Models*, Sage, Los Angeles, CA.

Angrist, J.D. and Pischke, J.-S. (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton, NJ.

Angrist, J.D. and Pischke, J.R.S. (2015) *Mastering Metrics: The Path from Cause to Effect*, Princeton University Press, Princeton, NJ.

Baicker, K., Taubman, S.L., *et al.* (2013) The Oregon Experiment—Effects of Medicaid on clinical outcomes. *New England Journal of Medicine*, 368 (18), 1713–1722.

Bauhoff, S. (2014) The effect of school district nutrition policies on dietary intake and overweight: a synthetic control approach. *Economics and Human Biology*, 12, 45–55.

Bell, A. and Jones, K. (2015) Explaining fixed effects: random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3,(1), 133–153.

Bertrand, M., Duflo, E., *et al.* (2004) How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119 (1), 249–275.

Blakely, T., van der Deen, F.S., Woodward, A., *et al.* (2014) Do changes in income, deprivation, labour force status and family status influence smoking behaviour over the short run? Panel study of 15 000 adults. *Tobacco Control*, 23 (e2), e106–e113.

Bowser, D., Canning, D., and Okunogbe, A. (2016). The impact of tobacco taxes on mortality in the USA, 1970–2005. *Tobacco Control*, 25 (1), 52–59.

Brook, R.H., Ware, J.E., *et al.* (1983) Does free care improve adults health—results from a randomized controlled trial. *New England Journal of Medicine*, 309 (23), 1426–1434.

Buchmueller, T.C.M. and Marko, S.V. (2014) How do Providers Respond to Public Health Insurance Expansions? Evidence from Adult Medicaid Dental Benefits. NBER Working Paper Series, National Bureau of Economic Research, Cambridge, MA.

Cameron, A.C. and Miller, D.L. (2015) A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50 (2), 317–372.

Carpenter, C.S. (2009) The effects of local workplace smoking laws on smoking restrictions and exposure to smoke at work. *Journal of Human Resources*. 44 (4), 1023–1046.

Clark, T.S. and D.A. Linzer, D.A. (2015) Should I use fixed or random effects? *Political Science Research and Methods*, 3 (2), 399–408.

Cohen, A. and Einav, L. (2003) The effects of mandatory seat belt laws on driving behavior and traffic fatalities. *Review of Economics and Statistics*, 85 (4), 828–843.

Craig, P., Cooper, C., *et al.* (2012) Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *Journal of Epidemiology and Community Health*, 66 (12), 1182–1186.

Cutler, D.M., Lleras-Muney, A., *et al.* (2011) Socioeconomic status and health: dimensions and mechanisms, in *The Oxford Handbook of Health Economics* (eds S. Glied and P.C. Smith), Oxford University Press, New York, pp. 124–163.

Davey Smith, G. and Ebrahim, S. (2001) Epidemiology—it is time to call it a day? *International Journal of Epidemiology*, 30, 1–11.

Dimick, J.B. and Ryan, A.M. (2014) Methods for evaluating changes in health care policy: the difference-in-differences approach. *Journal of the American Medical Association*, 312 (22), 2401–2402.

Disney, R., Emmerson, C., *et al.* (2006) Ill health and retirement in Britain: a panel data-based analysis. *Journal of Health Economics*, 25 (4), 621–649.

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 365

Donald, S.G. and Lang, K. (2007) Inference with difference-in-differences and other panel data. *Review of Economics and Statistics*, 89 (2), 221–233.

Duflo, E. (2001) Schooling and labor market consequences of school construction in Indonesia: evidence from an unusual policy experiment. *American Economic Review*, 91 (4), 795–813.

Dunkley-Hickin, C. (2014) Effects of primary care reform in Quebec on access to primary health care services. Master of Science, McGill University.

Forget, E.L. (2011) The town with no poverty: the health effects of a Canadian guaranteed annual income field experiment. *Canadian Public Policy*, XXXVII (3), 283–305.

Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B., and Vermeersch, C.M.J. (2011) *Impact Evaluation in Practice*, World Bank, Washington, DC.

Glymour, M.M., Kawachi, I., *et al.* (2008) Does childhood schooling affect old age memory or mental status? Using state schooling laws as natural experiments. *Journal of Epidemiology and Community Health*, 62 (6), 532–537.

Greene, W. (2004) The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *Econometrics Journal*, 7 (1), 98–119.

Greenland, S. (1983) Tests for interaction in epidemiologic studies: a review and a study of power. *Statistics and Medicine*, 2 (2), 243–251.

Greenland, S. (2001) Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, 30 (6), 1343–1350.

Greenland, S., Schwartzbaum, J.A., *et al.* (2000) Problems due to small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology*, 151 (5), 531–539.

Gunasekara, F.I., Richardson, K., *et al.* (2014) Fixed effects analysis of repeated measures data. *International Journal of Epidemiology*, 43 (1), 264–269.

Harper, S. and Strumpf, E.C. (2012) Social epidemiology: questionable answers and answerable questions. *Epidemiology*, 23 (6), 795–798.

Harper, S., Strumpf, E.C., *et al.* (2012) Do medical marijuana laws increase marijuana use? Replication study and extension. *Annals of Epidemiology*, 22 (3), 207–212.

Harper, S., Strumpf, E.C., *et al.* (2014) The effect of mandatory seat belt laws on seat belt use by socioeconomic position. *Journal of Policy Analysis and Management*, 33 (1), 141–161.

Heckman, J.J. (2006) Skill formation and the economics of investing in disadvantaged children. *Science*, 312 (5782), 1900–1902.

Hernán, M.A. and Robins, J.M. (2006) Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7), 578–586.

Hernán, M.A., Alonso, A., *et al.* (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, 19 (6), 766–779.

Imbens, G.W. and Wooldridge, J.M. (2009) Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47 (1), 5–86.

Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLoS Medicine*, 2 (8), 696–701.

Ionescu-Ittu, R., Glymour, M.M., *et al.* (2015) A difference-in-differences approach to estimate the effect of income-supplementation on food insecurity. *Preventive Medicine*, 70, 108–116.

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 366

Ivory, V.C., Blakely,T., *et al.* (2015) Do changes in neighborhood and household levels of smoking and deprivation result in changes in individual smoking behavior? A large-scale longitudinal study of New Zealand adults. *American Journal of Epidemiology*, 182 (5), 431–440.

Joyce, T. and Kaestner, R. (2001) The impact of mandatory waiting periods and parental consent laws on the timing of abortion and state of occurrence among adolescents in Mississippi and South Carolina. *Journal of Policy Analysis and Management*, 20 (2), 263–282.

Kaufman J.S. (2008). Why are we biased against bias? *International Journal of Epidemiology*, 37 (3), 624–626.

Kaufman, J.S. (2010) Toward a more disproportionate epidemiology. *Epidemiology*, 21 (1), 1–2.

Kaufman, J.S. (2013) Some models just can't be fixed. A commentary on Mortensen. *Social Science and Medicine*, 76 (1), 8–11.

Kaufman, J.S. and Cooper, R.S. (1999) Seeking causal explanations in social epidemiology. *American Journal of Epidemiology*, 150 (2), 113–120.

Kaufman, J.S. and Cooper, R.S. (2001) Commentary: considerations for use of racial/ethnic classification in etiologic research. *American Journal of Epidemiology*, 154 (4), 291–298.

Kim, I.S. and Imai, K. (2014) wfe: weighted linear fixed effects regression models for causal inference, Version 1.3. Available from: https://cran.r-project.org/web/packages/wfe/index.html.

Koh, H.K. (2010) A 2020 vision for healthy people. *The New England Journal of Medicine*, 362 (18), 1653–1656.

Kolstad, J.T. and Kowalski, A.E. (2012) The impact of health care reform on hospital and preventive care: evidence from Massachusetts. *Journal of Public Economics*, 96 (11–12), 909–929.

Kravdal, O. (2011) The fixed-effects model admittedly no quick fix, but still a step in the right direction and better than the suggested alternative. *Journal of Epidemiology and Community Health*, 65 (4), 291–292.

Krieger, N. and Smith, G.D. (2000) Re: "Seeking causal explanations in social epidemiology." *American Journal of Epidemiology*, 151 (8), 831–833.

Lawlor, D.A., Davey Smith, G., *et al.* (2004) Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *The Lancet*, 363 (9422), 1724–1727.

Levine, P.B., Staiger, D., *et al.* (1999) Roe v. Wade and American fertility. *American Journal of Public Health*, 89 (2), 199–203.

Linden, A. (2015) Conducting interrupted time-series analysis for single- and multiple-group comparisons. *The Stata Journal*, 15 (2), 480–500.

Lipsitch, M., Tchetgen, E., *et al.* (2010) Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21 (3), 383–388.

Long J.S. (1997). Regression models for categorical and limited dependent variables. *Advanced Quantitative Techniques in the Social Sciences, 7*, Sage Publications, Thousand Oaks, CA.

Ludwig, J., Sanbonmatsu, L., *et al.* (2011) Neighborhoods, obesity, and diabetes—a randomized social experiment. *New England Journal of Medicine*, 365 (16), 1509–1519.

Macinko, J. and Silver, D. (2015) Diffusion of impaired driving laws among US states. *American Journal of Public Health*, 105 (9), 1893–1900.

Trim Size: 7in x 9.25in
Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 367

Maclure, M. and Mittleman, M.A. (2000) Should we use a case-crossover design? *Annual Review of Public Health*, 21, 193–221.

Maldonado, G. and Greenland, S. (2002) Estimating causal effects. *International Journal of Epidemiology*, 31 (2), 422–429.

Mazumder, B. (2008) Does education improve health? A reexamination of the evidence from compulsory schooling laws. *Federal Reserve Bank of Chicago Economic Perspectives*, Q2, 2–16.

McKinnon, B., Auger, N., Kaufman, J.S. (2015a) The impact of smoke-free legislation on educational differences in birth outcomes. *Journal of Epidemiology and Community Health*, 69 (10), 937–943.

McKinnon, B., Harper, S., *et al.* (2015b). Who benefits from removing user fees for facility-based delivery services? Evidence on socioeconomic differences from Ghana, Senegal and Sierra Leone. *Social Science and Medicine*, 135, 117–123.

McKinnon, B., Harper, S., *et al.* (2015c) Removing user fees for facility-based delivery services: a difference-in-differences evaluation from ten sub-Saharan African countries. *Health Policy Plan*, 30 (4), 432–441.

Meer, J. and West, J. (2016) Effects of the minimum wage on employment dynamics. *Journal of Human Resources*, 51 (2), 500–522.

Meyer, B.D. (1995) Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics*, 13 (2), 151–161.

Muller, C.J. and MacLehose, R.F. (2014) Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. *International Journal of Epidemiology*, 43 (3), 962–970.

O'Campo, P. (2012) Are we producing the right kind of actionable evidence for the social determinants of health? *Journal of Urban Health*, 89 (6), 881–893.

Petticrew, M. (2007) "More research needed": plugging gaps in the evidence base on health inequalities. *European Journal of Public Health*, 17 (5), 411–413.

Petticrew, M., Whitehead, M., *et al.* (2004) Evidence for public health policy on inequalities: 1: The reality according to policymakers. *Journal of Epidemiology and Community Health*, 58 (10), 811–816.

Petticrew, M., Tugwell, P., *et al.* (2012) Damned if you do, damned if you don't: subgroup analysis and equity. *Journal of Epidemiology and Community Health*, 66 (1), 95–98.

Riddell, C.A., Kaufman, J.S., *et al.* (2014) Effect of uterine rupture on a hospital's future rate of vaginal birth after cesarean delivery. *Obstetrics and Gynecology*, 124 (6), 1175–1181.

Ryan, A.M., Burgess, J.F., Jr, *et al.* (2015) Why we should not be indifferent to specification choices for difference-in-differences. *Health Service Research*, 50 (4), 1211–1235.

Shadish, W.R., Cook, T.D., *et al.* (2001) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston, MA.

Strumpf, E. (2011) Medicaid's effect on single women's labor supply: evidence from the introduction of Medicaid. *Journal of Health Economics*, 30 (3), 531–548.

Stuart, E.A., Huskamp, H.A., *et al.* (2014) Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Service Outcomes Research Methodology*, 14 (4), 166–182.

Suissa, S. (1995) The case–time–control design. *Epidemiology*, 6 (3), 248–253.

Thomson, H., Thomas, S., *et al.* (2013) Housing improvements for health and associated socio-economic outcomes. *Cochrane Database System Reviews*, 2, CD008657.

Trim Size: 7in x 9.25in

Oakes c14.tex V1 - 01/11/2017 7:23 P.M. Page 368

VanderWeele, T.J. and Knol, M.J. (2014) A tutorial on interaction. *Epidemiology Methods*, 3 (1), 33–72.

Villa, J.M. (2014) DIFF: Stata module to perform differences in differences estimation, Stata 10 Version. Available from https://ideas.repec.org/c/boc/bocode/s457083.html.

Wade, T.J., Lin, C.J., *et al.* (2014) Flooding and emergency room visits for gastrointestinal illness in Massachusetts: a case-crossover study. *PLoS One*, 9 (10), e110474.

Wanless, D. (2007) *Our Future Health Secured? A Review of NHS Funding and Performance*, King's Fund, London.

WHO Commission on Social Determinants of Health (2008) Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health. Final report of the Commission on the Social Determinants of Health, World Health Organization, Commission on Social Determinants of Health, Geneva, Switzerland.

Williams, R.L. (2000) A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56 (2), 645–646.

Wolfers, J. (2006) Did unilateral divorce laws raise divorce rates? A reconciliation and new results. *American Economic Review*, 96 (5), 1802–1820.

Wooldridge, J.M. (2013) *Introductory Econometrics: A Modern Approach*, South-Western Cengage Learning, Mason, OH.

Xavier, A., Price, C., *et al.* (2009) Solidarity in health: The European Commission sets out new actions on health inequalities. *Eurohealth*, 15 (3), 1–4.