DISCUSSION PAPER SERIES

IZA DP No. 6682

Sharpening the Effectiveness of Natural Experiments as an Analytical Tool

Harriet Orcutt Duleep

June 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Sharpening the Effectiveness of Natural Experiments as an Analytical Tool

Harriet Orcutt Duleep
College of William and Mary
and IZA

Discussion Paper No. 6682
June 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 6682
June 2012

ABSTRACT

Sharpening the Effectiveness of Natural Experiments as an Analytical Tool

The importance of using natural experiments in economic research has long been recognized. Yet, it is only in recent years that natural experiments have become an integral part of the economist's analytical toolbox, thanks to the efforts of Meyer, Card, Peters, Krueger, Gruber, and others. This use promises to shed new light on a variety of public policy issues and has already caused a major challenge to some tightly held beliefs in economics, most vividly illustrated by the finding of a positive effect of a minimum wage increase on the employment of low-wage workers. Although currently in vogue in economic research, the analysis of natural experiments could be substantially strengthened. This paper discusses several methodological approaches that would increase the precision and reliability of the results stemming from the analysis of natural experiments. A theme underlying all of these proposals is how best to measure the effect of a treatment on a variable, as opposed to explaining a level or change in a variable.

JEL Classification: J08, J18, J38, J48

Keywords: natural experiments, regression format, treatment effect, precision, extraneous group-specific effects, minimum wage, divorce laws, military service, Julian Simon

Corresponding author:

Harriet Orcutt Duleep
Thomas Jefferson Program in Public Policy
College of William and Mary
Williamsburg, VA 23187-8795
USA
Email: hduleep@wm.edu

Sharpening the Effectiveness of Natural Experiments as an Analytical Tool

Harriet Orcutt Duleep

The importance of using natural experiments in economic research has long been recognized (Campbell and Cook, Simon, 1966, Orcutt, 1970). Yet, it is only in recent years that natural experiments have become an integral part of the economist's analytical toolbox, thanks to the efforts of Meyer, Card, Peters, Krueger, Gruber, and others. The increased use of natural experiments promises new understanding and has already caused a major challenge to one of economics' most tightly held beliefs: an increase in the minimum wage should have a negative effect on the employment of low-wage workers.

Although currently in vogue in economic research, the analysis of natural experiments could be substantially strengthened. This paper describes methodological approaches that would increase the precision and reliability of the results from the analysis of natural experiments. Precision can be increased by measuring the mean of the individual differences rather than the difference in means between treatment and control group observations. Precision can also be increased by matching before and after observations in the treatment and control groups, and matching observations across the treatment and control groups. These methods are part and parcel of the literature on experimental methods, with foundations in statistics so deep that it is difficult to cite their origins. Nevertheless, despite their long history, they are overlooked in many analyses of natural experiments by economists. The contribution here is to highlight these issues within the context of the predominant method of analysis of natural experiments by economists. I also discuss how natural experiments are not a panacea; depending on the analytical design, the estimates of an effect of a treatment in a natural experiment can be quite fragile, even when the researcher is armed with numerous observations. I discuss steps that researchers could take to surmount this latent fragility and increase the reliability of their results. The underlying theme throughout this paper is how best to measure the effect of a treatment on a variable, as opposed to how best to explain a level or change in a variable.

I. The Analysis of Data from Natural Experiments: The Difference in Averages versus the Average of the Differences

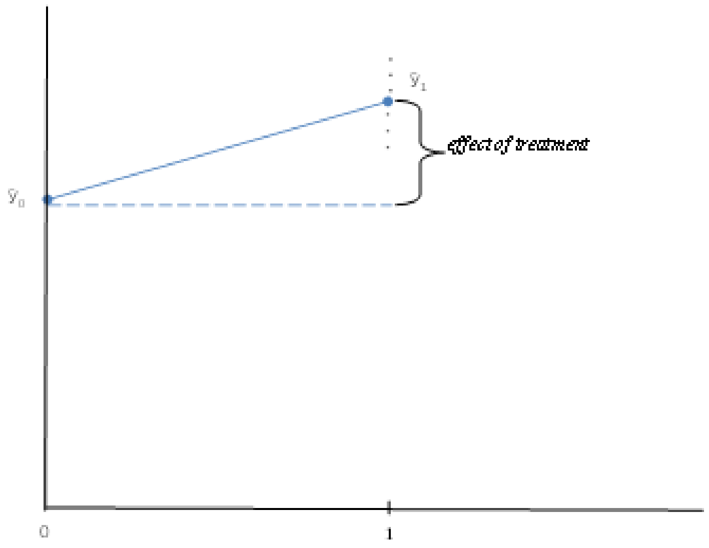
In recent studies by economists, the analysis of natural experiments has often been put into a regression format. This is not surprising given the predominant focus in economic research on explaining the level of a variable, or a change in a variable, as opposed to precisely estimating the effect of a treatment on a variable.

Meyer (1995) provides a comprehensive presentation and review of the analysis of natural experiments using a regression format. For a before-treatment/after-treatment analysis, we may estimate the following regression:

$$y_{it} = \alpha + \beta d_t + \varepsilon_{it}$$

Where i refers to the individual and t refers to the time period (t=0 for the initial period, t=1 for the post-treatment period). The dummy variable, d, equals 1 if the observation is after the treatment, and equals 0 if the observation is for the initial period, before the treatment. Estimating this equation, we get $\hat{\alpha} = \bar{y}_0$, or the average value of the before-treatment observations. As shown in Figure 1, the estimated effect of the treatment is $\hat{\beta} = \bar{y}_1 - \bar{y}_0$, or the average value of the after-treatment observations minus the average value of the before-treatment observations.

Figure 1



Similarly, Meyer shows that the effect of the treatment from an over time natural experiment that includes a control group can be estimated as

$$y_{it}^{j} = \alpha + \beta_1 d_t + \beta_2 d_j + \beta_3 d_t^{j} + \varepsilon_{it}^{j}$$

where i refers to the individual; t refers to the time period, with 0 for the initial period and 1 for the post-treatment period; and j refers to the group membership of the individual, with j=1 if individual i is in the group that receives the treatment in time period 1, and j=0 if the individual i is in the control group.

The dummy variables are defined as follows:

d_t = 1, if the time period is the post-treatment period (e.g. if t=1)
d_j = 1 if the group is the treatment group (e.g. if j=1)
d_t^j = 1 if the group is the treatment group and the time period is the post-treatment period (e.g. if

t=1 and if j=1)

Then,

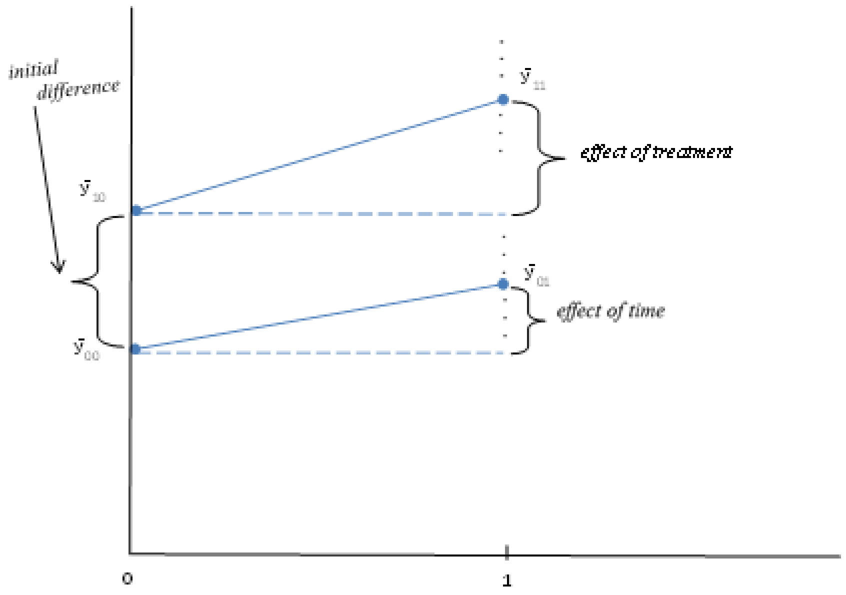
$\hat{\alpha} = \bar{y}_{00}$, the average value of the control observations in time period 0;
$\hat{\alpha} + \hat{\beta}_1 = \bar{y}_{01}$, the average value of the control observations in time period 1;
$\hat{\alpha} + \hat{\beta}_2 = \bar{y}_{10}$, the average value of the treatment observations in time period 0; and
$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = \bar{y}_{11}$, the average value of the treatment observations in time period 1.

These points are displayed in Figure 2.

As shown in Figure 2, $\hat{\beta}_2$, or ($\bar{y}_{10}$ - $\bar{y}_{00}$), is the difference between the average values of the treatment and control group in the initial period. This difference is assumed to be time invariant. $\hat{\beta}_1$ or ($\bar{y}_{01}$ - $\bar{y}_{00}$), is assumed to be that part of the change with time that is common to both the control and treatment groups. And, the estimated effect of the treatment is:

$[(\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3) – (\hat{\alpha} + \hat{\beta}_2)] – [(\hat{\alpha} + \hat{\beta}_1) - \hat{\alpha}\ ] = \hat{\beta}_3$ or $(\bar{y}_{11} - \bar{y}_{10}) - (\bar{y}_{01} - \bar{y}_{00})$

Figure 2



Given the goal of explaining the level of a variable (a pursuit that lends itself to adopting a regression format) rather than measuring the effect of a treatment, there has been a tendency in the analysis of natural experiments by economists to measure the *difference in the averages* between treatment and control group outcomes, measured by the coefficients in the estimated regressions shown above, rather than the *average of the individual differences*. Thus, in analyzing the outcome of interest in a before and after natural experiment, where y_{1i} is the after-treatment value of the ith observation and y_{0i} is the before-treatment value of the ith observation, the measure of the effect of the treatment has typically been the difference in the averages, $1/N\Sigma(y_{1i}) - 1/N\Sigma(y_{0i})$, rather than the average of the differences, $1/N \Sigma (y_{1i} - y_{0i})$.

Of course, the difference in the averages, $1/N\Sigma(y_{1i}) - 1/N\Sigma(y_{0i})$, equals the average of the individual differences, $1/N\Sigma(y_{1i} - y_{0i})$. However, the variance of the average of the individual differences is smaller than the variance of the difference in the averages. Hence, the precision with which we measure the effect of the treatment will be greater if we measure the average of the individual differences. Intuitively, the variance of the average of the differences is less than the difference in the averages because separately averaging the before-treatment values and the after-treatment values before taking the difference throws away the within-group variance. A more formal proof that the variance of the average of the individual differences is less than the variance of the difference in the averages follows.

The variance of the difference in the averages equals

var ($\bar{y}_1$ - $\bar{y}_0$) = var $\bar{y}_1$ + var $\bar{y}_0$ - 2cov $\bar{y}_1$ $\bar{y}_0$

The variance of the average of the individual differences equals

var $1/N\Sigma(y_{1i} - y_{0i})$ or var($\overline{y_{1\iota} - y_{0\iota}}$) = var($y_1$ - y_0)/N (since var $\bar{y}$ = (var y)/N).

var(y_1 - y_0)/N =(vary_1 + vary_0 - 2covy_1y_0)/N = (var y_1)/N + (var y_0)/N - (2cov y_1y_0)/N

= var $\bar{y}_1$ + var $\bar{y}_0$ - (2covy_1y_0)/N

Since (2covy_1y_0)/N , the term subtracted off var $\bar{y}_1$ + var $\bar{y}_0$ in the formula for the var $1/N\Sigma(y_{1i} - y_{0i})$, is greater than 2cov $\bar{y}_1$ $\bar{y}_0$, the term subtracted off var ($\bar{y}_1$ - $\bar{y}_0$), it follows that var $1/N\Sigma(y_1 - y_0)$ < var ($\bar{y}_1$ - $\bar{y}_0$).

Proof that (2covy_1y_0)/N > 2cov $\bar{y}_1$ $\bar{y}_0$ or (covy_1y_0)/N > cov $\bar{y}_1$ $\bar{y}_0$:

Covy_1y_0 = E[y_1 - Ey_1][y_0 - Ey_0] = E[y_1y_0 - y_0Ey_1 - y_1Ey_0 + Ey_1Ey_0]

= E(y_1y_0) - 2Ey_1Ey_0 + Ey_1Ey_0 = E(y_1y_0) - Ey_1Ey_0 Thus, (covy_1y_0)/N = [E(y_1y_0) - Ey_1Ey_0]/N.

Cov $\bar{y}_1\bar{y}_0$ = E[$\bar{y}_1$ - E$\bar{y}_1$][$\bar{y}_0$ - E$\bar{y}_0$] = E($\bar{y}_1\bar{y}_0$) - 2E$\bar{y}_1$E$\bar{y}_0$ + E$\bar{y}_1$E$\bar{y}_0$

= E($\bar{y}_1\bar{y}_0$) - E$\bar{y}_1$E$\bar{y}_0$ = E($\bar{y}_1\bar{y}_0$) - Ey_1Ey_0 (since E$\bar{y}$ = Ey)

But E($\bar{y}_1\bar{y}_0$) = E($1/N\Sigma y_{1i} \cdot 1/N\Sigma y_{0i}$) = $1/N^2$ E$\Sigma y_{1i}\Sigma y_{0i}$ = $1/N^2$ E$\Sigma\Sigma y_{1i}y_{0i}$

$= 1/\text{N}^2\Sigma\Sigma E(y_{1i}y_{0i}) = 1/\text{N}^2(\text{N}\cdot E(y_1y_0)) = 1/\text{N}\ E(y_1y_0)$.

So, cov $\bar{y}_1\bar{y}_0 = E(\bar{y}_1\bar{y}_0) - Ey_1Ey_0 = 1/\text{N}\ E(y_1y_0) - Ey_1Ey_0$

$= [E(y_1y_0) - \text{N}\cdot Ey_1Ey_0]/\text{N}$ versus $(\text{cov}y_1y_0)/\text{N} = [E(y_1y_0) - Ey_1Ey_0]/\text{N}$.

Thus, $(\text{cov}y_1y_0)/\text{N} >$ cov $\bar{y}_1\bar{y}_0$.

The preceding discussion suggests that in a before-treatment/after-treatment design with no control group, rather than adopting the regression format, $y_{it} = \alpha + \beta d_t + \varepsilon_{it}$, and measuring the difference between the average of the after-treatment values and the average of the before-treatment values, a superior approach that will yield more precise estimates of the treatment effect is the more straightforward one of measuring $1/\text{N}\Sigma(y_{1i} - y_{0i})$.

For exactly the same reason, in a before treatment/after treatment analysis with a control group, rather than estimating $y_{it}^j = \alpha + \beta_1 d_t + \beta_2 d_j + \beta_3 d_t^j + \varepsilon_{it}^j$ the preferred approach is to measure $1/\text{N}\Sigma[(y_{11i} - y_{10i}) - (y_{01i} - y_{00i})]$ where the first subscript refers to the group (treatment versus control), the second subscript refers to the time period, and the third subscript denotes each individual treatment-control observation pair. This presumes that the treatment and control observations are matched, the topic of the next section. However, even when the control and treatment observations are not matched, precision of the estimated treatment effect can be improved by estimating

$1/\text{N}\Sigma(y_{11i} - y_{10i}) - 1/\text{N}\Sigma(y_{01j} - y_{00j})$, rather than $(\bar{y}_{11} - \bar{y}_{10}) - (\bar{y}_{01} - \bar{y}_{00})$.

Note, that if the regression were set up as $y_1\text{-}y_0 = \alpha + \beta T$ where $y_1\text{-}y_0$ is an observation pair from either the control or treatment group and T = 1 if the observation pair belongs to the treatment group and 0 if it belongs to the control group, then $\hat{\beta}$ will be measuring the average of the individual differences of the treatment and, separately, the average of the individual differences of the control observation pairs, or $1/N\Sigma(y_{11i} - y_{10i}) - 1/N\Sigma(y_{01j} - y_{00j})$.

The advantage of using the regression format and estimating $y_{it}^j = \alpha + \beta_1 d_t + \beta_2 d_j + \beta_3 d_t^j + \varepsilon_{it}^j$ is that it conveniently decomposes the effects on the outcome variable of the initial difference between the control and treatment groups (β_2), their assumed common time trend effect on the outcome (β_1), and the effect of the treatment on the outcome (β_3). Yet, if our goal is to measure as accurately as possible the effect of a treatment, rather than to explain the level of a variable, then our preferred approach should always be to measure the average of the individual differences, rather than the difference in the averages.

II. Increasing Precision by Using Matched Data

Underlying several analyses of natural experiments is the idea that if the control and treatment groups are similar in respects other than the imposition of the treatment, then we will be more likely to detect the effect of the treatment; its effect will be more likely to surface above the other noise.

In this vein, the effect of a treatment is generally measured by comparing the before-treatment/after-treatment experience of the treatment group with the experience of other, non-treatment, units over the same period of time. To pick the control group from another time period would inject into the comparison an additional source of variation in the dependent variable and make it more difficult to detect the effect of the treatment. For instance, in a natural experiment analysis of the effect of state-imposed price changes on liquor consumption, Simon (1966) compares the change in liquor consumption in state i, which experienced a liquor tax change, with the changes in liquor consumption occurring over the same time period in states that did not experience a liquor tax change.

Analysts have also sought to have a comparison group that was similar to the treatment group in characteristics other than a shared time period. For instance, Card and Krueger (1994) sought to shed light on the effect of a minimum wage change on low-wage employment by comparing the over time employment of fast-food restaurant workers in a state in which a minimum wage increase was legislated, during this time period, to a state with no minimum wage change: the employment in fast-food restaurants before and after a minimum wage increase in New Jersey is compared with changes in the employment of fast-food restaurants over the same time period of neighboring Pennsylvania, which had not instituted a minimum wage increase. Presumably, the economies and populations of neighboring states share more in common than more geographically dispersed states. Similarly, in an analysis of the effect of the sudden influx (with the Mariel boatlift) of Cuban immigrants into Miami on the unemployment and wages of the low-skilled in Miami, Card (1990) compared Miami's employment and wage experience preceding and following the Mariel boatlift with that of another Florida city that was similar to Miami in a number of respects.

As these examples show, the idea of trying to eliminate sources of variation other than the treatment by matching has been incorporated to a varying extent in analyses of natural experiments. Yet, when the analyst has observations on individual units, this fundamental idea can be taken much further.

The ideal experimental design to estimate the effect of a treatment would entail the following specifications. From the relevant population, a sample of pairs of individuals would be selected who were matched in terms of an assortment of characteristics. From this sample of matched individuals, one person (or other unit) in each pair would be randomly assigned to the treatment, the other would act as a control. Matching individuals increases the precision of the estimate of the effect of the treatment. Randomly assigning members of matched pairs to treatment and control groups insures that the treatment is uncorrelated with other variables that affect the outcome.

The advantage of matching before-treatment observations with after-treatment observations (and control observations with treatment observations) can be easily seen from the formula for the variance of a difference. Let us first consider simply a before and after analysis where y_0 is the before treatment outcome and y_1 is the after treatment outcome.

$\text{Var } (y_1 - y_0) = \text{Var } y_1 + \text{Var } y_0 - 2\text{cov}y_1y_0$

Evidently, the greater the covariance between y_1 and y_0, the smaller the variance of the difference. Independently drawn samples will have the largest variance. Thus, matching should be incorporated wherever possible in the collection of information for the analysis of a natural experiment. If possible the same individual units should be followed before and after the treatment, and control observations should be matched with treatment observations. The matched data should then be analyzed by taking the average of the individual differences as discussed in the preceding section; not to do so would fail to take advantage of the matching.

However, even if the analyst of a natural, as opposed to real, experiment has two random samples at their disposal, one before the treatment and one after, the precision of the estimated effect of the treatment can be increased by matching the before/after observations on the basis of their characteristics. Thus, in Peters' before-after study of the effect of a legislative change on divorce, the preciseness of her estimated effect of the treatment—the change in legislation—

could have been improved if she divided her before and after samples according to characteristics such as age, education, and year of marriage.[1] The difference in the probability of divorce for these pairs of before-after observation groups could be measured. Her estimated effect of the divorce law would then be the average of the differences for these pairs of before-after observations.

Similarly, in analyzing control-treatment groups, observations can be paired across the groups according to similar characteristics. Ideally, matching should be incorporated in the initial data collection for analyzing a natural experiment. Thus, in the Card/Krueger analysis of the effect on low-wage employment of the increase in the minimum wage in New Jersey versus Pennsylvania, pairs of New Jersey-Pennsylvania fast-food restaurants could have been chosen that were similar in several characteristics, such as the income level of the neighborhoods they served, how long they had been in business, the number of employees in the initial period, etc. Such pairs of restaurants could have been drawn from the New Jersey-Pennsylvania sample of restaurants, rather than drawing the New Jersey and Pennsylvania samples separately. The analysts could then have computed the difference in the before-after employment for each restaurant pair and taken the average of the differences.

There is, however, no reason why matching treatment and control observations cannot be incorporated into the analysis of a natural experiment even if it was not part of the original data collection effort. In his study of the effect of military service on earnings (in which the random assignment of service numbers was used to distinguish the treatment group from the control group), Angrist (1990) notes a lack of statistical significance in his results. However, the

[1] The values for the age variable for the post-legislative time period would reflect the passage of time.

preciseness of his estimated treatment effect might be increased if the members of the control group (those with non-draft selective service numbers) were matched according to several characteristics (such as initial earnings) to members of the treatment group (those with draft selective service numbers). The differences in the before-after outcomes between the matched individuals would then be measured and averaged across all treatment-control observation pairs. Doing so would likely increase the statistical significance of Angrist's results.

Matching would not be done if we were trying to explain the level of a variable, or changes in a variable. To match in this case would be to throw out relevant information. Thus, matching individuals on the basis of their education in the Peters study eliminates the possibility of estimating the effect of education on the probability of divorce. However, if our goal is to estimate the effect of a treatment—the effect of a certain legislative provision on the probability of divorce—rather than explaining the level of divorce, then matching before and after observations, and treatment and control observations, will increase the precision with which we can estimate that effect.

III. Extraneous Group-Specific Effects

> *If you are studying tree diseases by observing the condition of tree leaves, and you have two trees, one healthy, the other not, each with a 1000 leaves, do you have 2000 observations, or 2?*

Anyone who has done actual experiments realizes that the control and treatment groups may be affected in a variety of unknown and sometimes surprising ways. This will always be true as long as members of the treatment group and the control group are separated in some way in addition to the treatment — as long as there are group-specific extraneous factors. An important protection against the effect of group-specific extraneous factors is to compare the differences in the post-treatment/pre-treatment values of the treatment group with the over time experience of the controls. However, the treatment-control comparison of over-time changes only provides partial protection against group-specific extraneous factors. It provides protection against group-specific extraneous factors associated with the experimental and control group only if these extraneous factors have no effect on the variable of interest or if the differential level of group-specific extraneous factors between the control and treatment groups remains the same over time *and* the initial difference does not interact with other relevant factors that change with the passage of time.

With no change in the levels of the group-specific factors and no interaction between the initial difference and time, there would be no problem in isolating the effect of the treatment. The treatment effect is isolated by comparing the over-time change in the treatment and control groups. Even if the level of the group-specific factors change differentially for the treatment and control groups, or if there is an interaction between the initial difference and the passage of time, as long as we have information on how the extraneous factors affect the outcome of interest or how the difference interacts with time, then we can model it and incorporate it into the analysis. Without such information there would be no way to disentangle the effect of extraneous group-specific factors from the effect of the treatment. This would be true regardless of how many observations are in the control and treatment groups.

To illustrate these thoughts, consider the following experiment based on an actual experiment I did prior to my life as an economist. Given a running wheel, rats will run on it with no inducement whatsoever. In our experiment, the question is, how does rewarding rats affect the number of wheel cycles rats will run?

To address this question, we perform the following experiment. One group of rats, the treatment group, is rewarded for running on the wheel and the control group of rats is not. Information on the daily wheel cycles of the rats is collected for both groups before and after the reward schedule for the treatment rats is commenced. Other than the reward schedule, the treatment of the two groups of rats is the same, except for the position of their cages in the room, the control group being closer to the radiator, something the analyst has given no thought to.

Baseline information on the daily wheel cycles of the treatment and control groups is collected through the summer, when the radiator is off. With onset of winter, the reward schedule for the treatment group begins, and the radiator comes on. Thus there is an interaction between the difference in an initial condition and the effect of the passage of time. Unknown to the experimenter, all of the rats in the control group, whose cages are closer to the radiator, are exposed to a warmer temperature than the rats in the experimental group. The control group rats become more sluggish with the onset of winter, and they reduce their number of daily wheel revolutions. Their compatriots in the cage further away from the radiator increase or maintain their wheel routines, spurred on by the relative coolness in that part of the room. On the basis of this evidence, the experimenter assumes that the change (or lack of change) in the control group’s behavior is due to the onset of winter; they serve their purpose as a control group by controlling for the effect of the change in season. He assumes that in the absence of the reward schedule, the rats in the experimental group would be behaving in much the same way as the control-group rats, and he wrongly concludes that the propensity of rats to do wheel running is affected by rewarding wheel revolutions. In this example there was a differential change for the control and treatment group in the level of an extraneous group-specific relevant factor that the researcher was unaware of (the temperature of the cages that housed the control and treatment rats).

It is also possible for there to be no differential change in the level of relevant extraneous factors that the control and treatment groups are exposed to, and yet, experimental results may be contaminated by extraneous group-specific factors. This can occur if there is an interaction between the difference in the initial conditions of the control and treatment groups and the effect of time.

For instance, in the experiment above, let us pretend that the control rats were placed in a blue room, whereas the treatment rats were placed in a green room. Further imagine that (unknown to the analyst) with the change in season, being in a green room has a different psychological effect on rats (and their propensity to run wheels) than being in a blue room;   the green room provides a more uplifting environment with the onset of winter. In this example, whatever differences there were between the level of the extraneous factors for the control and treatment groups before the treatment remained the same over time. And yet, the interaction between the passage of time and the initial difference in extraneous factors prevents the analyst from separating the experimental effect from the effect a group-specific factor.

Sources of experimental bias due to extraneous group-specific effects can occur if:

1. The level of an extraneous factor changes differentially for the control and treatment group.

2. The difference in the level of extraneous factors between the control and treatment group is constant over the measurement period. However, the initial difference in these levels interacts with changes that occur (for both groups) over time.

3. The difference in the level of the extraneous factor is constant over the measurement period. However, the extraneous factor interacts with the treatment. (A famous example of this is the Hawthorne effect. A more recent potential example is the first “successful” cold fusion experiment in Utah.)

Even though the experimenter uses randomization to insure that the selection of the experimental units versus the control units is uncorrelated with the treatment, it is still possible that external factors that affect the outcome of interest may affect the members of the treatment group or members of the control group. If the group-specific effect interacts with either time or the treatment then part of the measured effect of the treatment will be due to the extraneous group effect. A large sample size of individual units in the control group and experimental group provides no protection whatsoever for this type of problem. More generally, *if the designation of the control and treatment group involves a group separation by something other than the treatment then there is always the possibility of extraneous group effects.*[2]

The potential contaminating role of extraneous group-specific factors is much greater for natural than for real experiments since the conditions in an experiment are amenable to the experimenter’s control. Furthermore, the experimental and control groups in natural experiments are almost always separated by either place—one state increases the minimum wage or institutes a new divorce law, and other states do not—or another group characteristic that makes one group of persons potentially affected by a change in the law, while another group is not.

One natural experiment that completely circumvents the problem of extraneous group-

[2] Note that I am not considering cases where subjects can choose whether or not to participate as discussed in Achen (1986). The natural experiments that are my focus are cases where something is imposed on a state or group versus a decision whether to participate or not. Indeed, I take as the definition of a natural experiment in this paper an exogenously imposed circumstance.

specific effects is the study of the effect of military service by Angrist (1990). In this study, persons in the control group and treatment group are *only* defined by the random selection. Angrist did not use actual military service to define the experimental and control groups. The treatment—receiving a draft-eligible selective service number—involved no differential interaction with the control and treatment group. The more common case, however, is that the control and experimental groups in natural experiments are defined by a characteristic other than the selection for treatment.

To either implicitly or explicitly control for extraneous factors correlated with the control and treatment groups, one approach has been to try to pick a control group that is as similar as possible to the treatment group or that differs in such a way that, a priori, one would suspect the effect of the recognized extraneous factor would work against measuring a treatment effect.

Both precautions against extraneous group-specific effects—choosing control and treatment groups that are similar in important respects other than the treatment, and pursuing a devil's advocate approach of choosing a control group with levels of potentially relevant variables that would be expected to work against finding an effect—are present in the approach used by Card and Krueger (1994) to measure the effect on low-wage employment of increasing the minimum wage in New Jersey. A neighboring state, Pennsylvania, was chosen as the state from which the sample of control fast-food restaurants was chosen. Although the relative state unemployment rates changed over time for both states, New Jersey experienced more of an economic down turn than did Pennsylvania. Thus, it would seem reasonable to suspect that the differential change in the states' overall unemployment rates (the recognized extraneous factor) would work against the reported finding that the minimum wage in New Jersey had no detrimental effect on the employment of low-wage workers in New Jersey. Indeed, the minimum wage increase was found to have a *positive* effect on employment in fast-food restaurants.

A problem with these two approaches for controlling for extraneous group-specific factors (picking a control group that is similar to the treatment group or differs in such a way to work against finding a treatment effect) is that such approaches rely too heavily on the analyst's own knowledge. The analyst logically thinks that the greater decrease in economic activity in New Jersey, relative to Pennsylvania, strengthens his case; the absence of an adverse employment effect on fast food restaurant workers in New Jersey with the minimum wage increase cannot be attributed to New Jersey's economy improving more than Pennsylvania's economy. The fact that New Jersey's economy actually worsened relative to Pennsylvania and the minimum wage increase was associated with an increase in fast-food restaurant employment in New Jersey provides convincing evidence that the imposition of a small change in the minimum wage has no effect on low-wage employment. Yet, other factors may affect the control or treatment group in ways that escape even the most vigilant of researchers.

To proceed with the Card-Krueger analysis, a potential extraneous group-specific factor not considered by the analysts could be an increase in demand for food prepared by low-wage workers *as a result of an economic downturn*. People likely switch from expensive restaurants to fast-food restaurants as the economy sours. This potential extraneous factor could have occurred more in New Jersey than Pennsylvania because the economic downturn in New Jersey was greater than the economic downturn in Pennsylvania. The fact that employment in New Jersey's fast food restaurants increased after the minimum wage increase could have reflected an increase in the demand for fast food as people in New Jersey switched from expensive to inexpensive restaurants. Another potential explanation for the Card-Krueger results is the added worker effect: with an economic downturn, more women work in response to their husbands' unemployment. This decreases their time to prepare meals, and may lead to an increase in the demand for fast food. Thus, the very condition that the analysts believed protected them from an extraneous group-specific effect (a decrease in New Jersey's economy relative to Pennsylvania) could have triggered the increased employment in fast-food restaurants in New Jersey relative to Pennsylvania.

If the variables that affect the change in the outcome were identified, then one could get information on their effects on the outcome in the absence of the treatment. Using previous data, one could model the effect of downturns on the demand for food at low-wage versus high-wage restaurants. Equipped with knowledge of this relationship, the analyst could simulate what the effect of the current change would be on the change in outcome. As long as the extraneous factor is known, and there is variation on it within the group, or from other sources, then its effect can be accounted for in some way in the analysis. If the group-specific factor is known and is present to varying degrees across the units within the control or treatment group, then a solution is readily available. The analyst may simply divide up the units within the treatment and control groups so that they are similar in the level of this variable.

Such approaches, however, depend on the insights of researchers. There are many other state-specific factors other than the one discussed here that could affect over time trends in employment in low wage restaurants, and that may interact with the passage of time. The fact that two states are similar at the outset in terms of characteristics that the researcher thinks are relevant is no guarantee that there might not be differential changes between the time before the treatment and after the treatment is imposed in the level of other relevant group-specific factors or that there might be an interaction with the initial difference in group-specific factors and the passage of time. This discussion suggests that results from natural experiments may be more fragile than otherwise thought, despite the large sample sizes of the control and treatment groups that characterize many recent studies.

How then do we protect ourselves against the effect of extraneous and unknown group-specific factors?

Averaging Over Many Sites

It is easy to think of solutions to *known* group-specific extraneous factors. In the above example, the analyst could have partitioned the New Jersey and Pennsylvania data into areas of similar overall unemployment changes. The problem though is that group-specific extraneous factors that are unknown to the analyst are likely ever present. In this case, the only protection is to bring in other comparisons so that the effects of group-specific factors might be rendered harmless through averaging.

The approach used by Simon (1966) and Lyon and Simon (1968) offers a two-part protection plan against the effects of unknown group-specific extraneous factors. To illustrate this approach, consider Simon’s study of the effect of state price changes on liquor consumption.[3]

First, for any given year, the percentage change in before-after liquor consumption of the

[3] State changes in liquor (and cigarette) prices offer the possibility of measuring the effect of price changes on consumption to the extent that the state changes are not the result of changes in demand.

particular state which changed the price of liquor is compared with the average of the percentage changes in liquor consumption, over the same time period, of all the states that did not change their liquor prices.[4] In the Card and Krueger example, this would be analogous to comparing the change in New Jersey's fast food restaurant employment with the average of the change in fast food employment, over the same time period, for all states that did not have a minimum wage change. Averaging the change across several control groups, rather than one, protects against the measured treatment effect being the result of a unique difference in the effect of group-specific effects between the treatment group and the control group.

Averaging the change over several control groups does not, however, protect against the possibility that the measured treatment effect is the result of a change in group-specific factors unique to the treatment state (e.g. New Jersey, in the Card and Krueger study). The second protection against the effects of group-specific extraneous factors in the Simon (1966) and Lyon and Simon (1968) papers is to average the measured treatment effect across several years in which one state imposed a liquor price change. In each of these relevant years, the liquor consumption change in the state with a liquor tax change is compared with the average of all the changes for states in that same time period that did not have a liquor tax change. Averaging over many treatment-control group experiences reduces the possibility that we will assign the measured effect to the effect of the treatment when in fact it is due to a group-specific extraneous factor or the combined effect of the treatment and group-specific extraneous factors.

Analytically, the advantages of averaging the before-after experiences of numerous

[4] Note that we ignore throughout this paper the possible problem of a situation giving rise to a treatment. (Although averaging of treatment-control experiences may provide some protection against this.) We assume exogeneity of the treatment. See the careful discussion of this issue in Card and Krueger (1994) as applied to the change in the minimum wage law in New Jersey.

control groups and then averaging over several treatment-control group experiences can be seen in the following way.

In an analysis of a single treatment group and a single control group, let T_1 be the post-treatment outcome of the state that receives the treatment, let T_0 be the pre-treatment outcome of the same state, and let C_1 and C_0 be the first and second period outcomes of the control state. Let X be a vector of extraneous group-specific factors affecting the treatment state, and $X_1 - X_0$ the change in the level of extraneous group-specific factors affecting the variable of interest in the treatment state. Let Z be a vector of extraneous group-specific factors affecting the control state, and $Z_1 - Z_0$ the change in the level of extraneous group-specific factors affecting the variable of interest in the control state. Then,

$$(T_1 - T_0) - (C_1 - C_0) = \Delta + [\beta(X_1 - X_0) - \gamma(Z_1 - Z_0)] \qquad (1)$$

or, the measured effect of the treatment is the true effect, Δ , plus the difference in the effects on the outcome variable of changes in the group-specific factors of the treatment and control groups.

Comparing the change in the treatment group to the average change across several comparison groups, our natural experiment becomes

$$(T_1 - T_0) - 1/M\Sigma(C_{1i} - C_{0i}) = \Delta + [\beta(X_1 - X_0) - 1/M\Sigma\gamma_i(Z_{1i} - Z_{0i})] \qquad (2)$$

The expected value of this equals $\Delta + \beta(X_1 - X_0)$, or, the actual treatment effect plus the change in the group-specific factors associated with the treatment group. If we then do several such treatment-control group comparisons and average over their expected values, we get

$$1/N\Sigma\Delta_j + 1/N\Sigma[\beta_j (X_1 - X_0)_j \qquad (3)$$

the expected value of which is Δ.

If it is the case that the treatment is always accompanied by another effect so that the expected value of the second term of (3) is not zero, then it is the case that X, the vector of extraneous factors, always occurs with the treatment. Given this situation, it is impossible to separate the treatment effect from the effect of X. But then if this is the case, it doesn't matter. If X always accompanies the treatment, then for policy purposes, it is the combined effect that we are interested in measuring. Indeed, as suggested by Orcutt and Orcutt (1968), this may be an advantage of natural experiments over real experiments.

In general, recent economic studies of natural experiments have not combined experimental evidence over time and across areas to analyze the effect of a treatment. Rather, many recent analyses of natural experiments consist of one over time comparison of two groups defined by the state in which they live or another group-defining characteristic. Although studies with micro data have many observations in their control and treatment group, the possibility of group-specific extraneous factors makes the findings of these individual studies fragile even if the sample sizes of the individual units within the control and treatment groups are very large.

In the studies by Simon (1966) and Lyon and Simon(1968), the possibility of increasing the precision of their results by measuring the average of the differences (versus the difference in the averages) or by matching before-after observations and treatment-control observations did not exist because only average data for the states were available. With micro samples for the control and treatment groups, as has characterized several recent analyses of natural experiments, there is the potential to use the methods discussed in Part I and II of this paper to obtain more precise estimates of the effect of a treatment. Yet, with the advent of the analysis of large micro samples for the control and treatment groups, analysts may have lost sight of the potential sensitivity of their results to group-specific extraneous effects of any one comparison—regardless of the number of individual observations involved in that comparison.

There is no discussion in Simon (1966) or Lyon and Simon (1968) of how their approach protects against the possible contamination of the estimated treatment effect by extraneous group-specific effects. This may be because in the absence of micro data, the comparison of only two groups would have amounted to two observations, and the fragility of the comparison to group-specific extraneous factors would have been self-evident.

There is also no discussion of the potential contaminating effect of unknown extraneous group-specific factors in more recent studies that have been done with numerous observations. And yet these studies, despite numerous observations on individual units, are vulnerable to this effect. The advent of micro data may have obscured this fragility and instilled a false sense of confidence in the powers of natural experiments. Yet, from the sole perspective of the potential effect of group-specific factors, a comparison of two groups, each with a 1000 observations, is the same as a comparison of two observations.

IV. Summary

The analysis of natural experiments offers considerable promise for shedding light on a number of policy issues. Yet, in recent economic studies, natural experiments could be better exploited in terms of the precision and reliability of the estimated effects from their analysis.

Precision can be increased by analyzing the average of the individual differences between control and treatment observations, as opposed to the difference in the means of control and treatment groups. The advantage of measuring the average of the differences, rather than the difference in averages, may have gone unnoticed by many economists who have analyzed natural experiments because of the generally embraced goal by economists of trying to explain levels of variables or changes in levels of variables—the appropriate domain of the regression format—rather than accurately measuring the effect of a treatment on an outcome.

Matching before and after observations also increases the precision of the estimate. Ideally the same units would be followed over time. But in absence of this, before and after observations could be matched on characteristics. Forming matched pairs between the control and treatment groups also increases the potential precision with which an effect can be measured.

Regardless of how treatment-control outcomes are analyzed, the analysis of natural experiments may be affected by unknown group-specific factors. The best protection against the potential effect of extraneous group-specific effects is the one that relies least on the analyst's knowledge—averaging the experimental evidence over time and over sites as exemplified in Simon (1966) and Lyon and Simon (1968).

In any given comparison, micro studies allow more precise estimates of the effect of the treatment than do studies such as Simon (1966) and Lyon and Simon (1968) that relied on aggregate statistics. With the availability of micro data, precise estimates of the treatment may be obtained by measuring the average of the differences versus the difference of the averages, and by matching before/after observations and control/treatment observations. Yet, regardless of whether the analyst has micro data or aggregate data at his disposal, there is always a danger that the estimated treatment effect will be contaminated by extraneous and unknown group-specific factors when only two groups are compared. Numerous micro observations offer no protection for the problem of contaminating group-specific effects. The ideal situation would be for all three of the approaches discussed in this paper to be used in the analysis of natural experiments.

References

Achen, Christopher H., *The Statistical Analysis of Quasi-Experiments*, Berkeley and Los Angeles, California: University of California Press, 1986.

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence form Social Security Administrative Records, American Economic Review, vol. 80, 1990, pp. 313-336.

Card, David, "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review*, vol. 43, January 1990, pp. 245-57.

Card, David and Alan Krueger, " Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania, American Economic Review, vol. 84, pp. 772-793.

Hunt, Jennifer. 1992. The impact of the 1962 repatriates from algeria on the French labor market. *Industrial and Labor Relations Review* 45(3): 556-572.

Lyon, Herbert L. and Julian L. Simon, "Price Elasticity of the Demand for Cigarettes in the United States," *American Journal of Agricultural Economics*, vol. 50, November 1968, pp. 888-895.

Meyer, Bruce D. "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics*, April 1995, pp. 151-161.

Nakamura, Alice, Masao Nakamura, and Harriet Orcutt Duleep, "Alterative Approaches to Model Choice,"   *Journal of Economic Behavior and Organization*, July, 1990, pp 97-125.

Orcutt, Guy H., "Data, Research, and Government," *American Economic Review, Proceedings*, May 1970.

Orcutt, Guy H. and Alice G. Orcutt, "Incentive and Discincentive Experimentaiton for Income Maintenance Policy Purposes," *American Economic Review*, vol. 58, no. 4, Sept. 1968, pp.754-772.

Simon, Julian L., "The Price Elasticity of Liquor in the U.S. and a Simple Method of Determination," *Econometrica*, vol. 34, January 1966, pp. 193-205.