I. Data Types

Data can come to us in various formats. Let's start by visualizing what these kinds of data look like. We will refer to "unit" as the individual people, cities, firms, etc. in a given dataset.

Cross-sectional data is a snapshot of a bunch of individuals/units at *one point in time*. When a sample of units is randomly selected, we can think of it as a representative "cross section" of the population. Since we use *i* to index units (people, firms, cities, etc.), the notation for cross sectional data is what you've seen before. The following example is for a cross-section of individuals.

$$wage_i = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 female_i + u_i$$

indiv	wage	edu	exper	female
1	3.10	11	2	1
2	3.24	12	22	1
.	.	.	.	.
100	5.30	12	7	0

Pooled cross-sectional data is multiple snapshots of multiple bunches of (randomly selected) units at *many points in time*. Suppose we have two cross-sectional datasets from two different years; *pooling* the data means to treat them as one larger sample and control for the fact that some observations are from a different year. We can use the same notation here as in cross-section, indexing each person, firm, city, etc. by *i*. In the example below, the variable $y2010_i$ captures whether a house from the pooled sample was observed in 2010 (as opposed to 2000).

$$hprice_i = \beta_0 + \beta_1 bdrms_i + \beta_2 bthrms_i + \beta_3 sqrft_i + \delta y2010_i + u_i$$

house	year	hprice	bdrms	bthrms	sqrft
1	2000	85,500	3	2.0	1600
2	2000	67,300	3	2.5	1400
.	.	.	.	.	.
100	2000	134,000	4	2.5	2000
101	2010	243,000	4	3.0	2600
102	2010	65,000	2	1.0	1250
.	.	.	.	.	.
200	2010	144,000	3	2	2000

Finally, **panel data** is more like a movie than a snapshot because it tracks particular people, firms, cities, etc. over time. We observe the *same cross-section* in multiple time periods. With panel data we start indexing observations by *t* as well as *i* to distinguish between our observations of unit *i* at various points in time. The following example is for panel data from cities.

$$murders_{it} = \beta_0 + \beta_1 pop_{it} + \beta_2 police_{it} + a_i + d_t + u_{it}$$

a_i is an vector of dummy variables for each unit *i* (except one reference unit) and d_t is a vector of dummy variables for each time period *t* (except one reference period). We call these "fixed effects."

i	t	murder rate	pop density	police
1	2000	9.3	2.24	440
1	2001	11.6	2.38	471
2	2000	7.6	1.61	75
2	2001	10.3	1.73	75
.	.	.	.	.
100	2000	11.1	3.12	520
100	2001	17.2	3.34	493

II. Using Panel Data and Fixed Effects

A. Example: Two-Period Panel Data Analysis

Let's consider an example of panel data where you have data on crime and unemployment rates for 46 cities for 1982 and 1987. Therefore we have two time periods, and we can label them as $t = 1$ for 1982, and $t = 2$ for 1987.

Cross-sectional analysis: What happens if we use the 1987 cross section and run a simple regression of crime on unemployment?

$$\widehat{crmrte} = \underset{(20.76)}{128.38} - \underset{(3.42)}{4.16} unemp$$

If we were to interpret the coefficient on unemployment, we would infer that higher unemployment is associated with less crime. This seems backwards. The culprit? Well we might first think about omitted variable bias. The first solution that comes to mind is to control for more factors, such as land area, part of the country (West or East), police officers per square mile, law enforcement expenditure, and per capita income. We get the following result:

$$\widehat{crmrte} = \underset{(2.74)}{140.06} - \underset{(1.80)}{6.7} unem + \underset{(1.23)}{0.059} area - \underset{(1.79)}{21.963} west - \underset{(0.17)}{0.114} offarea + \underset{(1.15)}{0.021} lawexp - \underset{(0.53)}{0.002} pcinc$$

We still get this puzzling negative relationship between unemployment and crime. Is this the true relationship, or are there still omitted variables we are missing?

Pooled cross-sectional analysis: Since we have two time periods, we can take advantage of this by using both years of data and controlling for what time period an observation is in. This would account for factors that changed over time and are associated with both unemployment and crime. Doing this, we obtain

$$\widehat{crmrte} = 93.42 + 7.94 d87 + 0.427 unem$$

Here we recover the positive relationship we expected! But we are effectively treating the data as a pooled cross-section and not taking advantage of the fact that we observe a panel of the same cities multiple times. Doing so could help us address some remaining concerns about OVB even after controlling for time period.

Panel data analysis - first differences: One potential solution we can use with panel data is to take first differences. Because the set of cities in our dataset is constant over time, we can difference the data across the two years. Taking first differences (within cities) tells us how variables are changing within cities over time, and controls for all city characteristics that don't change over time. This is useful, since what we want to estimate is how a change in unemployment affects crime.

For an observation i measured in two time periods (where t is period 1 or period 2), we can think of separate regressions for each time period as follows:

$$\begin{aligned} y_{i2} &= (\beta_0 + \delta_0) + \beta_1 x_{i2} + \alpha_i + u_{i2} \\ y_{i1} &= \beta_0 + \beta_1 x_{i1} + \alpha_i + u_{i1} \end{aligned}$$

In this specification, α_i is the subset of variables in $u_{i,t}$ that includes all time-constant characteristics of unit i that affect the outcome y. We don't know what these are since we don't observe u_i, but we can assume that u_i includes some time-invariant variables, and for notational purposes we group them together in α_i. Note that we assume that the effect of x on y is constant over time (β_1), but we allow there to be a different baseline level (intercept) of y in the two time periods, where that difference is captured by δ_0.

Subtracting the second equation from the first gives:

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

The most important thing to note about this formula is that α_i has been "differenced away". We don't even have to know what those time-constant omitted variables are - we have accounted for them! We can analyze this expressing using the same methods as before in the class, just defining the changes in variables as random variables (i.e., let $\tilde{y} = \Delta y$, etc.). You will recover causal estimates as long as the usual assumptions hold, with some changes to account for the data structure. Most importantly, MLR4 now requires that Δu_i is uncorrelated with Δx_i.

After taking first differences in our data, we obtain:

$$\Delta \widehat{crmrte} = 15.40 + 2.22 \Delta unem$$

Here a one unit change in the unemployment rate over time within cities is associated with a 2.22 unit increase in the change in crime rate over time. By taking first differences, we have controlled for all city characteristics that do not change over time.

Panel data analysis - fixed effects: First differences allowed us to eliminate all unobserved, time-constant factors that affect crime rates in a given city i from our estimation. An alternative method to accomplish this is to directly control for those city-specific time-constant factors. We can do this by introducing *unit fixed effects* - individual dummies that control for the unit of interest - directly into our regression. In this case, with cities, we can create a dummy for each city. We get the following result:

$$\widehat{crmrte} = \underset{(1.95)}{91.618} + \underset{(1.80)}{2.932 unem} + \underset{(1.03)}{1.838 offarea} - \underset{(0.51)}{0.007 lawexp} - \underset{(1.00)}{0.006 pcinc} + \delta_2 city2 + \cdots + \delta_{46} city46 + d87$$

As with the first differences approach, we now have something that makes much more sense: an increase in the unemployment rate is associated with an increase in the crime rate. In fact, with two time periods, fixed effects and first differences give the same results for the independent variables of interest: the difference with the previous regression using first differences is due to added time-varying controls in this fixed effects specification.

Comparing this regression to those above, you can see that we include a control for year ($d87$) to capture any factors broadly affecting crime rate over time, as we could with repeated cross-sections. The panel nature of the data allows us to include unit fixed effects, which accomplish the same thing as taking first differences. Because unit fixed effects capture time-constant variables, we can no longer include *area* or *west* (included in our initial cross-sectional attempt to deal with OVB) because these do not change over time so are absorbed in the fixed effect.

B. Fixed Effects Definition

In panel data where we observe the same units in multiple periods, we can include controls for specific units. We do this with what we call a *unit fixed effect*, which we denote going forward as a_i or α_i (for notational simplicity). A unit fixed effect is a vector of $n - 1$ dummy variables, where n is the number of captures all unobserved. Thus a_i represents not just a single dummy variable, but multiple dummy variables for each i except for one reference unit (notice how in the example above, we omit $city1$ - this is the reference unit). A given dummy variable a_j is coded as 1 for unit j and 0 for all other units. The unit fixed effect captures all time-constant factors within the unit that affect y_{it} (the fact that this term is not indexed by a time subscript t reminds us that it does not change over time). In the example above, this would be all city characteristics that don't change over time. Note that including unit fixed effects means that we will not include any additional x variables that don't change over time within units because they will already be subsumed by the fixed effect for our unit of interest.

Unit fixed effects are extremely useful because they reduce the number of potential omitted variables we could be concerned about. Many time-constant variables could affect y_{it} and cause bias in estimates of a particular coefficient of interest if we are using cross-sectional data. With panel data, we can include the unit fixed effects and control for all of those variables at once. We still will be concerned about time-varying omitted variables, but we have at least reduced the set of possible omitted variables.

Note that we can only include unit fixed effects with panel data where the outcome variable changes over time. Otherwise, including the fixed effect perfectly predicts the outcome. Similarly, if you don't have panel data (you only observe each unit once) and control for what unit you are looking at in your regression, this directly tells you the value of the outcome variable.

In addition to unit fixed effects, with panel or repeated cross-sectional data we can also include *time fixed effects*. These follow the same concept as unit fixed effects, but are dummy variables for each time period observed. We often represent the vector of time period dummies (fixed effects) by δ_t or d_t. These time fixed effects capture all variables that change over time in the same way across units. If a variable changes over time in a different way across units, it will not be included in the time fixed effects.

C. Example: General Period Panel Data Analysis

Let's consider an example of panel data, where the unit of observation is a city-year, and suppose we have data for 3 cities for 3 years—so 9 total observations in our dataset. So in contrast to the previous example, we now have multiple years of data. The data look as follows (note the unit and time period dummy variables):

i	t	murder rate	pop density	City1	City2	City3	Yr00	Yr01	Yr02
1	2000	9.3	2.24	1	0	0	1	0	0
1	2001	11.6	2.38	1	0	0	0	1	0
1	2002	11.8	2.42	1	0	0	0	0	1
2	2000	7.6	1.61	0	1	0	1	0	0
2	2001	10.3	1.73	0	1	0	0	1	0
2	2002	11.9	1.81	0	1	0	0	0	1
3	2000	11.1	6.00	0	0	1	1	0	0
3	2001	17.2	6.33	0	0	1	0	1	0
3	2002	20.3	6.42	0	0	1	0	0	1

Since we have multiple observations for each city, we can run the following regression:

$$murders_{it} = \beta_0 + \beta_1 popden_{it} + \alpha_2 City2 \ + \alpha_3 City3 \ + \delta_2 Yr01 \ + \delta_3 Yr02 \ + u_{it}$$

1. How do we interpret β_1, α_3 or δ_3 here?

2. How would we get the predicted murder rate for city 3 in the year 2002?

For fixed effect regressions, we usually save time by writing an α_i instead of writing out each dummy variable for the unit fixed effects. You can imagine that if we had 40 cities instead of 3, writing out each dummy variable would get super tedious. Similarly, we usually write δ_t instead of writing out each dummy variable for the time fixed effects.

$$murder_{it} = \beta_0 + \beta_1 popden_{it} + a_i + d_t + u_{it}$$

Note the subscripts on these variables: for a given city, its city dummy variable isn't going to vary by year, and for a given year, its year dummy variable isn't going to vary by city. Now that we have both time and city fixed effects, we cannot include any additional x variables that do not vary across both time t and across units i. x variables that vary across time in the same way for all cities are captured by the time fixed effects. x variables that do not vary across time within cities are captured by the city fixed effects.

III. Panel Regressions and Fixed Effects in R

There are a few ways to implement a regression that totally controls for city and time effects. In these examples, I'll use a dataset about murder rates and unemployment rates across US states in the years 1987, 1990, and 1993.

1. Directly including dummies for units and time periods

$$\widehat{mrdrte}_{it} = \hat{\beta}_0 + \hat{\beta}_1 unem_{it} + \underbrace{\alpha_2 State2 + ...\alpha_{50} State50}_{\text{Dummies for all but one state}} + \underbrace{\delta_1 Yr1990 + \delta_2 Yr1993}_{\text{Dummies for all but one year}} + u_{it}$$

In R:

To include dummies for states and years in a regression (or to include them as fixed effects), you need to make sure those variables are treated as "factor" (categorical variables) in R. Sometimes a variable you would like to treat as a categorical variable in R will be coded automatically as another data type. We can check an object's data type using the class() command. For example, R might see a year as a number, whereas we would like to treat it as a factor (categorical variable). To fix this, we can run the following code:

```
class(mrdr$year)
mrdr$year<-as.factor(mrdr$year)
mrdr$state<-as.factor(mrdr$state)
```

After checking that your categorical variables are factors, run the regression (you don't need to do anything special once R knows the state and year variables are factors).

```
Call:
lm(formula = mrdrte ~ unem + state + year, data = mrdr)

Residuals:
Min       1Q   Median       3Q      Max
-26.7121  -0.6385  -0.0904   0.5954  13.4617

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.07729    3.30035   1.841   0.0686 .
unem         0.20194    0.29476   0.685   0.4949
stateAL      2.18207    2.88674   0.756   0.4515
stateAR      0.77599    2.89771   0.268   0.7894
stateAZ     -0.42174    2.96514  -0.142   0.8872
stateCA      3.31638    2.90649   1.141   0.2566
stateCO     -3.09514    2.96268  -1.045   0.2987
```

```
stateCT     -2.71307    3.05609  -0.888   0.3768
stateDC     55.56253    2.89504  19.192   <2e-16 ***
stateDE     -3.04576    3.09126  -0.985   0.3269
stateFL      1.94460    2.95315   0.658   0.5118
stateGA      3.37890    2.99139   1.130   0.2614
stateHI     -3.69061    3.20376  -1.152   0.2521
stateIA     -5.98582    3.08762  -1.939   0.0554 .
stateID     -5.60317    2.91821  -1.920   0.0577 .
... (Truncated for length)
stateWA     -3.27657    2.91644  -1.123   0.2639
stateWI     -4.02012    3.03631  -1.324   0.1885
stateWV     -3.37631    2.90495  -1.162   0.2479
stateWY     -5.03618    2.92754  -1.720   0.0885 .
year90       1.57702    0.74339   2.121   0.0364 *
year93       1.68194    0.69598   2.417   0.0175 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.514 on 99 degrees of freedom
Multiple R-squared:  0.9048,Adjusted R-squared:  0.8538
F-statistic: 17.75 on 53 and 99 DF,  p-value: < 2.2e-16
```

Note that R will automatically choose one category for your units and time periods to exclude from the regression and serve as the reference category. In this example, R excluded the dummy for the state AK and the year 1987.

2. Including unit (state) and time (year) fixed effects

$$\widehat{mrdrte}_{it} = \hat{\beta}_0 + \hat{\beta}_1 unem_{it} + \underbrace{\delta_t}_{\text{Year “fixed effect”}} + \underbrace{\alpha_i}_{\text{State “fixed effect”}} + u_{it}$$

In R, we now use the $felm()$ function (after loading the lfe package) instead of $lm()$. This is the fixed effects linear model. In this function, you write your regression equation as usual excluding the fixed effect, then add the fixed effect variable(s) after a bar, as below. Note that most of the commands like $summary()$ that you can run on an lm object can also be run on a felm object. Again, make sure your fixed effect variable is a factor variable in R (see method 1 above).

```
> library(lfe)
> summary(felm(mrdrte~unem|year+state,data=mrdr))

Call:
felm(formula = mrdrte ~ unem | year + state, data = mrdr)

Residuals:
Min       1Q   Median       3Q      Max
-26.7121  -0.6385  -0.0904   0.5954  13.4617
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
unem   0.2019     0.2948   0.685    0.495

Residual standard error: 3.514 on 99 degrees of freedom
Multiple R-squared(full model): 0.9048   Adjusted R-squared: 0.8538
Multiple R-squared(proj model): 0.004719   Adjusted R-squared: -0.5281
F-statistic(full model):17.75 on 53 and 99 DF, p-value: < 2.2e-16
F-statistic(proj model): 0.4694 on 1 and 99 DF, p-value: 0.4949
```

Note that we no longer see estimated coefficients for each state and year. They are included in the model but not shown, which is fine because we usually aren't interested in those coefficients, and only include fixed effects to help isolate the causal effect of a particular independent variable. You can see that the estimated coefficient for unemployment is the same as in the regression with all the included dummy variables! This is because we control for the same state and time effects in both regressions, just in different ways.

IV. Assumptions for FE model

Consider the following model:

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_k x_{itk} + a_i + \delta_t + u_{it}$$

1. Assumption 1: Model is linear in parameters

2. Assumption 2: Random sample

3. Assumption 3: Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables. This is important: we can't use first differences or fixed effects to analyze impacts of independent variables that don't change over time.

4. Assumption 4: $E(u_{it}|x_{it}, a_i, \delta_t) = 0$, or equivalently $E(\Delta u_i|\Delta x_i) = 0$ (since the former case rules out the time-invariant parts of u and x and the common trends over time in these variables). This assumption says that we don't want changes in the u's to be correlated with changes in the x's.

5. Assumption 5: $Var(u_{it}|x_{it}, a_i, \delta_t) = \sigma_u^2$

As before, from Assumption $A1 \rightarrow A4$ we get that β is unbiased. From Assumption A5 we get an expression we can estimate for $var(\hat{\beta})$.

Exercise

Consider the two panel data regressions below, where *i* indexes individuals and *t* indexes time in months:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + u_{it} \tag{1}$$
$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + a_i + u_{it} \tag{2}$$

1. What are the MLR4 assumptions for each model?

2. We've talked about how OVB is a violation of the MLR4 assumption. What kind of omitted variable bias is mitigated by using model (2) instead of model (1)? [Why is model (2) *better* than model (1)?]

V. Example

The scrap rate for a manufacturing firm is the number of defective items, products that must be discarded, out of every 100 produced. Thus, for a given number of items produced, a decrease in the scrap rate reflects higher worker productivity. A labor economist would like to examine the effects of job training on worker productivity. We can use the scrap rate to measure the effect of worker training on productivity. We can also take advantage of a program that offered firms a job training grant in certain years.

Let's use the JTRAIN.DTA dataset. We use the data for three years, 1987, 1988, and 1989, on the 54 firms that reported scrap rates in each year. No firms received job training grants prior to 1988; in 1988, 19 firms received grants; in 1989, 10 different firms received grants. Consider the following model:

$$\ln(scrap)_{it} = \beta_0 + \beta_1 grant_t + d88 + d89 + a_i + \varepsilon_{it}$$

Below is the R code used to run this regression. Notice that you need to create and include a lag of grant, include year dummies, and include firm fixed effects.

```
> summary(felm(lscrap~ grant+d88+d89|fcode, data=jtrain))

Call:
felm(formula = lscrap ~ grant + d88 + d89 | fcode,      data = jtrain)
```

```
Residuals:
Min       1Q   Median       3Q      Max
-2.28694 -0.11239 -0.01784  0.14427  1.42667

Coefficients:
Estimate Std. Error t value Pr(>|t|)
grant   -0.25231    0.15063  -1.675   0.0969 .
d88     -0.08022    0.10948  -0.733   0.4654
d89     -0.24720    0.13322  -1.856   0.0663 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4977 on 104 degrees of freedom
(309 observations deleted due to missingness)
Multiple R-squared(full model): 0.9276   Adjusted R-squared: 0.8879
Multiple R-squared(proj model): 0.201   Adjusted R-squared: -0.2368
F-statistic(full model):23.37 on 57 and 104 DF, p-value: < 2.2e-16
F-statistic(proj model): 6.543 on 4 and 104 DF, p-value: 9.774e-05
```

Note that we could equivalently run:

```
summary(felm(lscrap˜ grant|year+fcode, data=jtrain))
```

and obtain the same results: including year dummies or year fixed effects does the same thing. Whether we include the dummies or the fixed effects depends on whether we are interested in seeing the coefficients on years (in this case, do we care about how scrap rates are changing over time on average, holding firms and grants constant?).

Use logical reasoning and the regression results above to answer the following questions:

1. Interpret the coefficient on *grant*. Think about what we are holding constant.

2. Does it seem important to add indicators for time? What does this control for?

3. Why might we control for a particular firm? What types of potential omitted variables do firm fixed effects control for?

4. Is the coefficient on *grant* likely to represent the causal impact of job training on scrap rates? Give a reason why or why not.

Solutions

Section II.C

1. How do we interpret β_1, α_3 or δ_3 here?

- β_1 is the partial/marginal effect of population density on expected murders controlling for the year and the city.
- We can interpret α_3 as the "effect" (difference in mean murder rate) of City3 relative to the omitted group (City1)
- Similar to above, we can interpret δ_3 as the "effect" of Year02 relative to the omitted group (Year00)

2. How would we get the predicted murder rate for city 3 in the year 2002?

- Set $City1 = 0, City2 = 0, City3 = 1$
- Set $Year01 = 0, Year02 = 0, Year3 = 1$
- Plug in $popden_{3,2002}$
- $\widehat{murders}_{3,2002} = \beta_0 + \beta_1 popden_{3,2002} + \alpha_3 \ + \delta_{2002}$

Section IV

Consider the two panel data regressions below, where *i* indexes individuals and *t* indexes time in months:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + u_{it} \tag{3}$$
$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + a_i + u_{it} \tag{4}$$

1. What are the MLR4 assumptions for each model?

   For (1): $\mathbb{E}[u_{it}|x_{it1}, ..., x_{itk}] = 0$.

   For (2): $\mathbb{E}[u_{it}|x_{it1}, ..., x_{itk}, a_i] = 0$

2. We've talked about how OVB is a violation of the MLR4 assumption. What kind of omitted variable bias is mitigated by using model (2) instead of model (1)? [Why is model (2) *better* than model (1)?]

   Any omitted variable that is constant (or relatively constant) over time for a unit i will bias (1), but will not bias (2) because the fixed effect will capture any effect they have. This is the average over all the time periods for this unit.

Section V

1. Interpret the coefficient on *grant*.

   Having received job training grant decreases scrap rates by about 25% in the year the grant was received, holding constant the year the firm is observed, and accounting for differences in average scrap rates across firms. Effectively, we hold constant all attributes about the firms that don't change over time and all attributes of a year that affect all firms equally. This estimate is statistically significant at a 10% level (p=0.0969).

2. Does it seem important to add indicators for time? What does this control for?

   Time dummies capture factors that are changing generally over time across firms that affect worker productivity. For example, this could include technological advances in this manufacturing industry. The coefficient on d89 indicates that the scrap rate was substantially lower in 1989 than in the base year, 1987, even in the absence of job training grants. Thus, it is important to allow for these aggregate effects. If we omitted the year dummies, the secular increase in worker productivity would be attributed to the job training grants.

3. Why might we control for a particular firm? What types of potential omitted variables do firm fixed effects control for?

   We may be worried that certain types of firms are more likely to get the grants. For example, perhaps more productive firms on average are more likely to apply for a job training grant. We can control for this by controlling for time-invariant firm characteristics. This helps us isolate the effect of the training itself.

4. Is the coefficient on *grant* likely to represent the causal impact of job training on scrap rates? Give a reason why or why not.

   This is unlikely to represent the true causal impact. We have eliminated two sources of possible omitted variables bias: time-invariant differences across firms associated with scrap rates, and "secular trends" on average over time across the sample affecting changes in scrap rates. But there may be other omitted variables that remain. For example, firms that seek out job training grants may also pursue other methods to increase worker productivity over time, such as developing new methods to reduce scrap rates. If these change over time within a subset of firms that received grants, not measuring such efforts could lead to omitted variables bias (in this case, the estimated impacts would be biased downward - larger in absolute magnitude).