Module 8: Difference-in-Differences Designs

Contents

1. INTRODUCTION

In previous modules, we have argued that Randomized Control Trials (RCT) are a gold standard because they make a minimal set of assumptions to infer causality: namely, under the randomization assumption, there is no selection bias (which arises from pre-existing differences between the treatment and control groups). However, randomization does not always result in balanced groups, and without balance in observed covariates it is also less likely that unobserved covariates are balanced. Earlier, we explored Regression Discontinuity Designs (RDD) as a quasi-experimental approach when randomization is not feasible, allowing us to use a forcing variable to estimate the (local) causal effects around the discontinuity in eligibility for study participation. In RDD, we use our knowledge of the assignment rule to estimate causal effects.

In this module, we cover the popular quasi- or non-experimental method of Difference-in-Differences (DID) regression, which is used to estimate causal effect – under certain assumptions – through the analysis of panel data. DID is typically used when randomization is not feasible. However, DID can also be used in analyzing RCT data, especially when we believe that randomization fails to balance the treatment and control groups at the baseline (particularly in observed or unobserved effect modifiers and confounders). DID approaches can be used with multi-period panel data and data with multiple treatment groups, but we will demonstrate a typical two-period and two-group DID design in this module.

We present analytical methods to estimate causal effects using DID designs and introduce you to extensions to improve the precision and reduce the bias of such designs. We conclude the module with a discussion of Triple-Differences Designs (DDD) to introduce analysis allowing more than two groups or periods to be analyzed in DID designs.

The learning objectives of this module are:

- ✓ Understanding the basics of DID designs
- ✓ Estimating causal effects using regression analysis
- ✓ Incorporating “matching” techniques to improve precision and reduce bias in DID designs
- ✓ Introducing Triple-Differences Designs.

2. BASICS OF DID DESIGNS

Imagine that we have data from a treatment groups and a control group at the baseline and endline. If we conduct a simple before-and-after comparison using the treatment group alone, then we likely cannot “attribute” the outcomes or impacts to the intervention. For example, if income from agricultural activities increases at the endline, then is this change attributable to the agriculture-based intervention or to a better market (higher demand and price), season, or something else that the intervention did not impact? If children’s health improved over time, is it simply because they are getting older and having improved immune system or because of the intervention? In many cases, such baseline-endline comparison can be highly biased when evaluating causal effects on outcomes affected over time by factors other than the intervention.

A comparison at the endline between the treatment and control groups, on the other hand, may also be biased if these groups are unbalanced at the baseline. DID designs compare ***changes over time*** in treatment and control outcomes. Even under these circumstances, there often exist plausible assumptions under which we can control for time-invariant differences in the treatment and control groups and estimate the causal effects of the intervention. Consider the following math to better understand the DID design concept.

- ✓ The outcome Y_{igt} for an individual i at time t in group g(treatment or control) can be written as a function of:

$$Y_{igt} = \alpha_g + \theta_t + \beta_1 G + \beta_2 t + \beta_3 G.t + U_{igt} + \varepsilon_{igt}$$

where α_g captures group-level time-invariant (not changing over time) "fixed effects" (think of these as distinct Y-intercepts of the baseline outcome for each group); θ_t captures period time-invariant fixed effects (e.g., election effects if the baseline was an election year); G is an indicator variable for treatment (=1) or control (=0) groups; t is an indicator variable for baseline (=0) or endline/ (=1) measurements, the βs are the regression coefficients to be estimated; U_{igt} captures individual-level factors that vary across groups and over time; and ε_{igt} captures random error. Let's denote the outcomes for the following four conditions as,

- ✓ At baseline in treatment group:

$$Y_{i10} = \alpha_1 + \theta_0 + \beta_1.1 + \beta_2.0 + \beta_3.1.0 + U_{i10} + \varepsilon_{i10}$$

- ✓ Individual at baseline in control group:

$$Y_{i00} = \alpha_0 + \theta_0 + \beta_1.0 + \beta_2.0 + \beta_3.0.0 + U_{i00} + \varepsilon_{i00}$$

- ✓ Individual at follow-up in treatment group:

$$Y_{i11} = \alpha_1 + \theta_1 + \beta_1.1 + \beta_2.1 + \beta_3.1.1 + U_{i11} + \varepsilon_{i11}$$

- ✓ Individual at follow-up in control group:

$$Y_{i01} = \alpha_0 + \theta_1 + \beta_1.0 + \beta_2.1 + \beta_3.0.1 + U_{i01} + \varepsilon_{i01}$$

- ✓ Change over time in outcome in treatment group = (4) – (2):

$$\begin{aligned} Y_{i11} - Y_{i10} &= (\alpha_1 + \theta_1 + \beta_1.1 + \beta_2.1 + \beta_3.1.1 + U_{i11} + \varepsilon_{i11}) \\ &\quad - (\alpha_1 + \theta_0 + \beta_1.1 + \beta_2.0 + \beta_3.1.0 + U_{i10} + \varepsilon_{i10}) \\ &= (\theta_1 - \theta_0) + \beta_2 + \beta_3 + (U_{i11} - U_{i10}) + (\varepsilon_{i11} - \varepsilon_{i10}) \end{aligned}$$

- ✓ Change over time in outcome in control group = (5) – (3):

$$Y_{i01} - Y_{i00} = (\theta_1 - \theta_0) + \beta_2 + (U_{i01} - U_{i00}) + (\varepsilon_{i01} - \varepsilon_{i00})$$

- ✓ The average treatment effect (or the DID impact) = (6) – (7)

$$(Y_{i11} - Y_{i10}) - (Y_{i01} - Y_{i00}) = \beta_3 + (U_{i11} - U_{i10} - U_{i01} + U_{i00}) + (\varepsilon_{i11} - \varepsilon_{i10} - \varepsilon_{i01} + \varepsilon_{i00})$$

- ✓ $\boldsymbol{DID\ Impact\ or\ ATE = \beta_3 + (U_*) + (\varepsilon_*)}$

The final equation specified clarifies the assumptions needed in order to infer causality from DID designs. First, we expect that the regression error term has a distribution with mean 0, so that $\boldsymbol{\varepsilon}_*$ is also distributed with mean 0. Second, we assume that the time-variant differences over time in the treatment and control groups are equal, thus cancelling each other out (U_* = 0). This is a critical assumption made in DID analysis, allowing for causal analysis despite the absence of randomization, and in some cases we may not believe it to be true.

The concept of DID is displayed in Figure 1. The solid red line shows how the outcome (some outcome of interest, measured in percentages) would change over time without the treatment (as measured in the control group), while the solid blue line displays the change over time in the treatment group. By shifting the red dotted line upwards from the solid red line, we remove the change over time attributable to other-than-treatment factors. Therefore, DID design estimates the outcome attributable to the intervention. However, if the assumption that the **changes in time-variant** factors in treatment and control groups are equal does not hold (known as the Parallel Trend Assumption), then the true control outcome could track the red dashed line. As the figure demonstrates, we could overestimate (or underestimate) the causal effect using DID if the above assumption is violated.

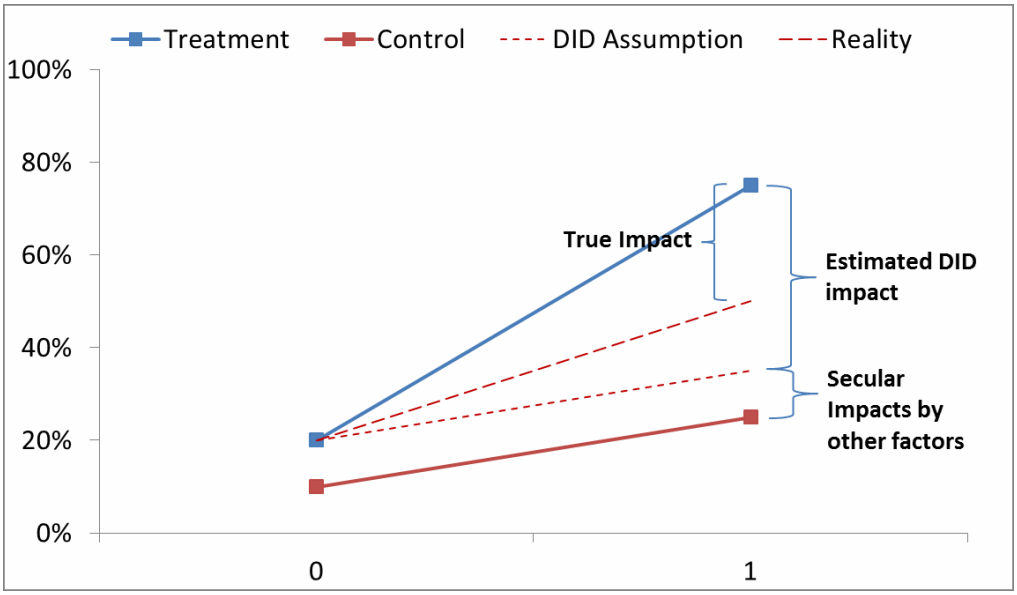


Figure 1. Graphical demonstration of difference-in-difference

It is possible to "control" for factors that may vary or change over time differently between the treatment and control groups in regression analysis but one can always be concerned about immeasurable or unmeasured factors causing time variant changes. Also, mathematically, DID can also be shown as subtracting from the mean difference at the endline between treatment and control groups the pre-existing differences in these groups at the baseline.

3. DEMONSTRATION: DID IN OPORTUNIDADES

We will demonstrate application of DID with dataset for OPORTUNIDADES (`DID_OPORTUNIDADES.csv`). This is a panel dataset of household and individuals tracked in years 2000, 2003 and 2007. Year 2000 was actually the final year of a previous version of OPORTUNIDADES called PROGRESA, which we studied in Modules 2.2, 2.3, and 2.4. The PROGRESA treatment was randomized to 320 villages and 186 control villages. By the fall of 2000 all 506 treatment and control villages were included in OPORTUNIDADES. However, it wasn't decided to track the long term impacts of OPORTUNIDADES until 2003, but by that time the original controls had become the "treatment," leaving only one option: to find a new control group. The study team used matching methods to find 150 control villages for the 506 treatment villages in 2003.

For this demonstration, we will apply DID method to compare outcomes between treatment and control villages in 2000 with those in 2003. Interestingly, the "baseline" year of 2000 already has 320 villages, which were exposed to the treatment. This is different from the typical setting in which baseline measurements are collected prior to program activities in the treatment villages. This is a challenging case for DID because of such contamination in the baseline, as well as because control villages in the baseline receiving subsequent treatment (under OPORTUNIDADES) and each control's being matched to multiple treatment villages.

Please implement the following steps in R.

- ✓ Open the dataset and create flags that identify unique villages and households in our sample. The code below cross-tabulates the treatment and control villages by year. The table displayed shows the distribution of villages by year and comparison groups.

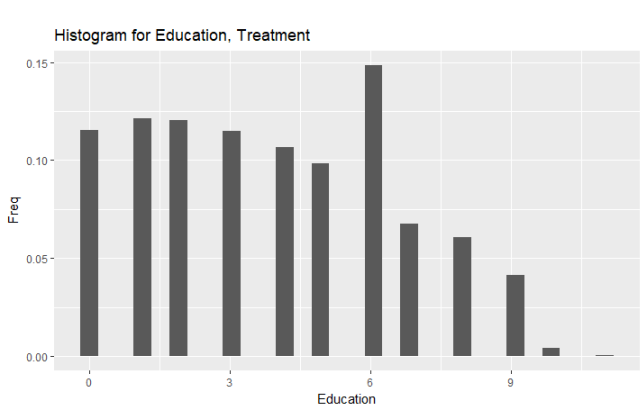
```
Panel_Data_uniqvil <- subset(Panel_Data, uniqvill == 1)

table(Panel_Data_uniqvil$D, Panel_Data_uniqvil$year, useNA = "ifany")
```

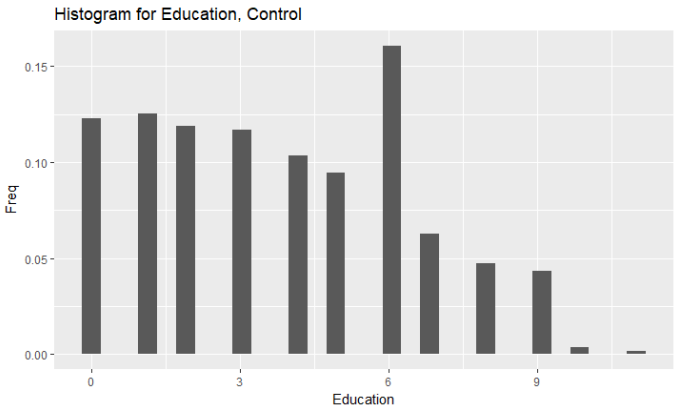
```
> table(Panel_Data_uniqvil$D, Panel_Data_uniqvil$year, useNA = "ifany")

          2000 2003 2007
  Control  186  150   41
  Treated  319  506  767
```

- ✓ In the next figures, we compare the distribution of children's education in the comparison groups in year 2000. The distributions are very similar for lower education levels. This is expected because changes in the number of years of education can be expected only in the long term. However, we notice that treatment villages may have had somewhat better outcomes in higher grades. This is expected, because we know that 2000 is not a "true" baseline and that several of treatment villages had benefitted from PROGRESA in the past. The R code is as follows,

```
# Compare distribution at baseline year 2000
Panel_Data_2000_0 <- subset(Panel_Data, D=="Control" & year == 2000)
Panel_Data_2000_1 <- subset(Panel_Data, D=="Treated" & year == 2000)
```

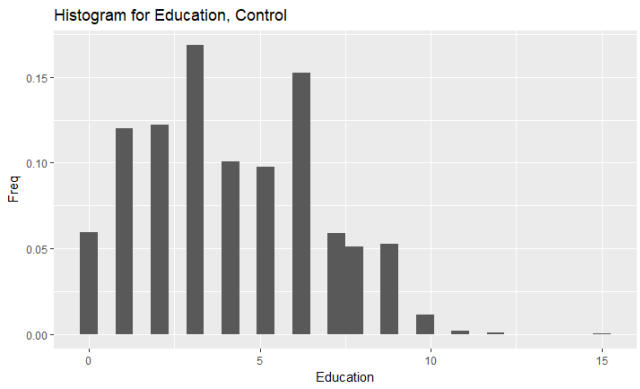
```
# Calculating proportions between control and treatment for different Education
ggplot(data=Panel_Data_2000_0, aes(Panel_Data_2000_0$edu_child)) +
  geom_histogram(aes(y=..count../sum(..count..))) +
  labs(title="Histogram for Education, Control") +
  labs(x="Education", y="Freq")

ggplot(data=Panel_Data_2000_1, aes(Panel_Data_2000_1$edu_child)) +
  geom_histogram(aes(y=..count../sum(..count..))) +
  labs(title="Histogram for Education, Treatment") +
  labs(x="Education", y="Freq")
```

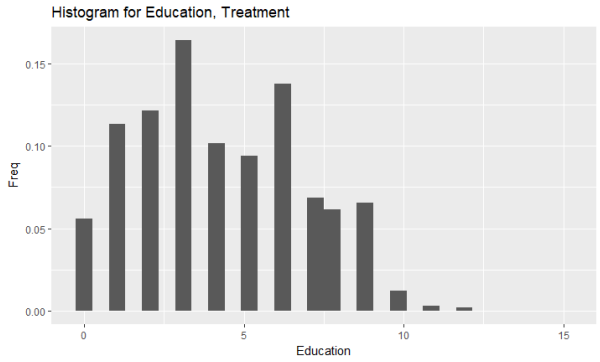




✓ In the following, we compare the same distributions in the follow-up year 2003. Here, we see that the treatment villages are faring better than the controls. The R code is as follows,

```
# Compare distribution at followup year 2003
Panel_Data_2003_0 <- subset(Panel_Data, D=="Control" & year == 2003)
Panel_Data_2003_1 <- subset(Panel_Data, D=="Treated" &  year == 2003)

# Calculating proportions between control and treatment for different Education
ggplot(data=Panel_Data_2003_0, aes(Panel_Data_2003_0$edu_child)) +
  geom_histogram(aes(y=..count../sum(..count..))) +
  labs(title="Histogram for Education, Control") +
  labs(x="Education", y="Freq")

ggplot(data=Panel_Data_2003_1, aes(Panel_Data_2003_1$edu_child)) +
  geom_histogram(aes(y=..count../sum(..count..))) +
  labs(title="Histogram for Education, Treatment") +
  labs(x="Education", y="Freq")
```





- ✓ Next, we check the baseline balance in covariates. We use the lapply command to apply the t-test function to many covariates at once

```
# T-tests for different covariates at Baseline
lapply(Panel_Data_2000[,c("edu_child","age", "sex", "agehead",
"sexhead")], function(x) t.test(x ~ Panel_Data_2000$D, var.equal = TRUE))
```

The figure below shows that the key covariates were reasonably balanced at the baseline in an economic sense, even though the differences are significantly different from 0 for the years of child education, age of head of the household, and age and sex of the child. However, we also know that 2000 was the final year of PROGRESA for this sample, for which reason there are certainly differences between treatment and control groups. Further, we tested only a few covariates, and an educational baseline test should be conducted simultaneously with several covariates and the outcomes of interest

```
$edu_child

        Two Sample t-test

data:  x by Panel_Data_2000$D
t = -2.765, df = 34904, p-value = 0.005696
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.13702955 -0.02334381
sample estimates:
mean in group Control mean in group Treated
             3.815461              3.895648


$age

        Two Sample t-test

data:  x by Panel_Data_2000$D
t = 1.8414, df = 111930, p-value = 0.06557
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01500806  0.48106650
sample estimates:
mean in group Control mean in group Treated
             25.45909              25.22606


$sex

        Two Sample t-test

data:  x by Panel_Data_2000$D
t = 2.627, df = 100830, p-value = 0.008615
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.002142699 0.014735157
sample estimates:
mean in group Control mean in group Treated
            0.5055985             0.4971596
```

```
$agehead

        Two Sample t-test

data:  x by Panel_Data_2000$D
t = 5.358, df = 106590, p-value = 8.434e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3037183 0.6540936
sample estimates:
mean in group Control mean in group Treated
             47.33369              46.85479


$sexhead

        Two Sample t-test

data:  x by Panel_Data_2000$D
t = 0.63324, df = 99182, p-value = 0.5266
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.002361874  0.004616444
sample estimates:
mean in group Control mean in group Treated
           0.08289686            0.08176957
```

- ✓ To estimate the average treatment effect using DID method, we specify the following regression model:

```
Panel_Data$D <- ifelse(Panel_Data$D == "Control", 0,
   ifelse(Panel_Data$D == "Treated", 1, NA))

Panel_Data$D_period <- Panel_Data$D * Panel_Data$period

Panel_Data_model <- lm(edu_child ~ D_period+D+period, data =
   Panel_Data)

summary(Panel_Data_model)
```

- ✓ where `D_period` is an interaction variable created by multiplying the `D` and `period` variables (See Equation 1 in Section 2). Below is the output of the DID analysis. As discussed in Section 2, the interaction coefficient (β_3 in models presented earlier) provides the average

treatment effect. We find an increase of 0.075 years of education for children in treatment villages compared to those in control villages.

```
Call:
lm(formula = edu_child ~ D_period + D + period, data = Panel_Data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2882 -2.2882 -0.2882  2.1044 10.8670

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.81546    0.02261 168.719  < 2e-16 ***
D_period     0.07501    0.04196   1.788  0.07385 .
D            0.08019    0.02901   2.764  0.00571 **
period       0.31759    0.03531   8.993  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.647 on 82615 degrees of freedom
  (384219 observations deleted due to missingness)
Multiple R-squared:  0.005758,  Adjusted R-squared:  0.005722
F-statistic: 159.5 on 3 and 82615 DF,  p-value: < 2.2e-16
```

- ✓ Let's discuss a way of mitigate bias resulting from baseline imbalance in DID analysis. We can do so by including covariates which (we believe) to have been imbalanced, or those which (we believe) could explain the imbalance between the groups well in the regression model specification. For demonstration sake, let's assume that age, sex of child, and the household head age and sex were imbalanced at the baseline. We re-estimate the DID model. Now the coefficient for the interaction term (D_period) is not statistically significant. The estimated magnitude of effect is also very small.

```
Call:
lm(formula = edu_child ~ D_period + D + period + age + sex +
    agehead + sexhead, data = Panel_Data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9457 -0.6999  0.0630  0.9604  8.6083

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -3.908810   0.029793 -131.197  < 2e-16 ***
D_period     0.015746   0.024487    0.643  0.52019
D            0.103376   0.017377    5.949 2.71e-09 ***
period       0.295866   0.020491   14.439  < 2e-16 ***
age          0.710285   0.001788  397.316  < 2e-16 ***
sex          0.096671   0.010725    9.014  < 2e-16 ***
agehead     -0.001359   0.000480   -2.831  0.00464 **
sexhead     -0.094090   0.019249   -4.888 1.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.497 on 77919 degrees of freedom
  (388911 observations deleted due to missingness)
Multiple R-squared:  0.6802,    Adjusted R-squared:  0.6802
F-statistic: 2.368e+04 on 7 and 77919 DF,  p-value: < 2.2e-16
```

Exercise 8.1

Here, we will repeat the analysis above, but for a new outcome variable, Income_HH_per (income per household member).

(a) Create histograms to see the distribution of income per household member by treatment and control in 2000 and then in 2003. You will need to add

```
+ scale_x_continuous(limits = c(0, 5000))
```

to your ggplot code in order to restrict the range of the x-axis so you can better see the data. Compare and interpret the distributions in the pre and post period.

(b) Use a difference in differences regression model to estimate the average treatment effect on income per household member (use controls as above). Interpret the results of the regression.

4. TRIPLE DIFFERENCE IN DIFFERENCES

Triple-Differences Design (DDD) is an extension to the basic DID analysis covered above to multiple groups and for multiple time periods. For example, let's compare three groups. We know that in OPORTUNIDADES only a fraction of the households in the treatment villages are eligible for treatment, so we can compare these three groups: (1) control villages and households; (2) participating households from treatment villages; and (3) non-participating households from treatment villages. While we agree that there are concerns about selection bias across these groups, here we only wish to demonstrate an application of DDD.

Consider the following conceptual model.

(10) The outcome Y_{igst} for an individual i at time t in group g (treatment or control) and subgroup (s) denoting participating in program itself:

$$E[Y_{igt}|G,S,t] = \alpha_g + \pi_s + \theta_t + \beta_1 G + \beta_2 S + \beta_3 t + \beta_4 GS + \beta_5 Gt + \beta_6 St + \beta_7 GSt$$

Where α_g represents group-level time-invariant "fixed" effects; π_s represents time-invariant fixed effects between participating and non-participating households; θ_t represents period fixed effects; G is indicator variable for the treatment (=1) or control (=0) groups; S is indicator variable for the participation (=1) group; t is indicator variable for baseline (=0) or endline/follow-up (=1) measurements, and the βs are the regression coefficients to be estimated. We ignore time-variant effects and assume that the expected mean error in above model is 0. We'll spare you the math, but it can be shown that the average treatment effect of the intervention on the participating households from the treatment villages compared to the control households from control villages is given by coefficient β_7.

5. BIBLIOGRAPHY/FURTHER READING

1. Bertrand, Marianne; Esther Duflo and Sendhil Mullainathan (2004). "How Much Should We Trust Differences-in-Differences Estimates?" Quarterly Journal of Economics, 119, 249-275.
2. Donald, S. G. and K. Lang (2007). "Inference with Difference-in-Differences and Other Panel Data", Review of Economics and Statistics, 89, 221-233.

3. Gerber, Alan S., and Donald P. Green. Field experiments: Design, analysis, and interpretation. WW Norton, 2012.
4. McKinnish, T. (2000). “Model Sensitivity in Panel Data Analysis: Some Caveats About the Interpretation of Fixed Effects and Differences Estimators”, Mimeo, University of Colorado.