Econometrics of Policy Evaluation: Matching

Cristian Huse

- What if the assignment to the treatment is done not randomly, but **on the basis of observables**?
- Then, one can use **matching**
- **Matching** is a set of statistical techniques to construct the "best possible" comparison group to a given treatment group
- **Matching** relies on the **Unconfoundedness** (or **Selection on Observables**) assumption. This assumes that after we control for all relevant **observed characteristics** (like age, education, etc.), there are **no unobserved differences** between the two groups that are correlated with both program participation and the outcome.

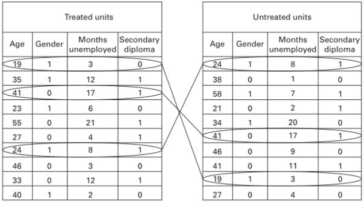
- **Stepping back:**
  - The main difficulty in getting a proper identification comes from the fact that we do not observe the **counterfactual**
  - The true counterfactual could only be observed if we could randomly assign treatment and "rewind the clock"
  - The goal of matching techniques is to estimate the missing counterfactual by using the information of subjects from a control group that are "close" to the treated units in some sense

- **Example:** What is the effect of a new diet on weight loss?
  - **Perfect experiment**
    - Randomly assign the new diet to some people and observe their weight loss
    - "Rewind the clock" and observe the weight loss without the new diet
  - **Matching**
    - For each person who followed the diet, find a "similar" person (on height, weight, occupation, health, etc) who did not (control/benchmark sample)
    - Compute the difference between the average weight loss for the dieters and the non-dieters is the weight loss (gain?) effect of the diet

- **Question:** What is the effect of treatment on the treated when the assignment to the treatment is based on observable variables?
- Key points to note:
  - The comparison group needs to be as similar as possible to the treatment group, in terms of the observables **before the start of the treatment**
    - Using **ex post** data (observed/collected after start of treatment) is dangerous as it might have been influenced by the treatment itself...
  - The method's validity rests on the strong assumption that, **conditional on the set of observable characteristics**, there are no remaining unobservable differences between the treatment and control groups.

- **Intuition:**
  - For every possible unit under treatment, matching attempts to find a non-treatment unit (or set thereof) that has the most similar characteristics possible, based on characteristics available on one's data set
  - These matched non-treated individuals then become the comparison group that you use to estimate the counterfactual

Figure 8.1 Exact Matching on Four Characteristics



Treated units			
Age	Gender	Months unemployed	Secondary diploma
19	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Untreated units			
Age	Gender	Months unemployed	Secondary diploma
24	1	8	1
38	0	1	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

- **Note:**
  - The above assumes (and finds) **exact** matching of individuals
  - What is the potential problem with this?

- **Problem:**
  - If the list of relevant observed characteristics is very large, or if each characteristic takes on many values, it may be hard to identify **an exact match** for each of the units in the treatment group
    - As you increase the number of characteristics or dimensions against which you want to match units that enrolled in the program, you may run into what is called the **curse of dimensionality**
- **Solution:**
  - Use a probabilistic approach that – instead of exact matching – assigns to each individual the probability to be part of the treatment group based on the observed values of its characteristics (the explanatory variables)
  - One such method is called **Propensity Score Matching (PSM)**

- **Intuition:**
  - For each unit in the treatment group, take the group of non-treated and compute the probability that this unit will enroll in the program (the **propensity score**) based on the observed values of its characteristics (the explanatory variables)
  - The PS is a real number between 0 and 1 that summarizes the influence of all of the observed characteristics on the likelihood of enrolling in the program
  - Non-treated units with a PS "close" to that of a treated observation will be matched to it
- **Concretely:**
  - Run a regression (e.g., logit) for all observations in the sample where the dependent variable is an indicator of treatment and the explanatory variables are the observed characteristics
  - The fitted values are the estimated probabilities that each unit in the sample enrolls in the program based on its observed characteristics, i.e., the **propensity score (PS)**
  - Select the closest non-treated unit to each treated unit (also >1)
  - Note that not all non-treated units will necessarily be matched to the treated units, so some observations may not be used

- Rosenbaum and Rubin (1983): assuming **unconfoundedness** (discussed below), then conditioning on the entire k-dimensional vector X is unnecessary. One can instead condition on the one-dimensional PS, i.e., the probability of receiving the treatment conditional on the covariates:
  - *"The propensity score allows to convert the multidimensional setup of matching into a one-dimensional setup. In that way, it allows to reduce the dimensionality problem."*
- **More intuition:**
  - Once the PS has been computed for all units, then units in the treatment group can be matched with units in the pool of non-treated that have the closest PS
  - These closest units become the control group and are used to produce an estimate of the counterfactual
  - The average difference in outcomes between the treatment units and their matched comparison units produces the estimated impact of the program
    - i.e., the program's impact is estimated by comparing the average outcomes of a treatment group and the average outcomes among a statistically matched subgroup of units, the match being based on observed characteristics available in the data at hand

- **Comments:**
  - Use only baseline (pre-intervention) observed characteristics to calculate the propensity score. Why?
    - Because post-treatment characteristics might have been affected by the program itself – using them would bias the results
    - i.e., when treatment affects individual characteristics and we use those to match, we choose a control group that looks similar to the treated group because of the treatment itself. Without the treatment, those characteristics would look more different
    - i.e., this violates the basic requirement for a good estimate of the counterfactual: the control group must be similar in all aspects, except for the fact that the treatment group receives the treatment and the control group does not
  - There are different methods to match individuals – e.g., nearest neighbours, distance-based or kernel-based techniques – and it is good practice to check robustness by comparing different methods

- **Matching in practice**
  - Assume you have a sample of observations on outcomes, covariates and the assignment indicator, (y_i, X_i, d_i)
    - Can assume $d_i = 1$ (say, treated) to fix ideas ($d_i = 0$ would have meant non-treated)
  - Next step is to take observations l from the **other group** ($d_l = 0$ in this case) and rank them according to closeness to observation i
  - You will rank those observations according to their distance from observation i as measured by the distance in terms of **characteristics**: $||X_l - X_i||$, where $||.||$ measures the distance (norm) between two matrices
  - You will then choose the M **closest observations** to observation i, which we can write as $L_M(i) = \{l_1(i), ..., l_M(i)\}$ – these are the matched observations for observation i
  - You will do this for every observation among the treated ones. Upon finishing, you will have the matched sample to the treated observations, i.e., the control group constructed via matching
- **Note:**
  - Observations belong to the **other group**
  - Observations are the **closest** to the observation you are focusing on w.r.t **observable characteristics** (multivariate)

- **Matching in practice (cont'd)**
  - Recall that we can **match on the propensity score** rather than matching directly on all the covariates X. Rewrite previous slide:
  - Assume you have a sample of observations on outcomes, covariates and the assignment indicator, $(y_i, X_i, d_i, \hat{p}(X_i))$
    - Can assume $d_i = 1$ (say, treated) to fix ideas
  - Next step is to take observations l from the **other group** ($d_l = 0$ in this case) and rank them according to closeness to observation i
  - You will rank those observations according to their distance from observation i as measured by the difference in **propensity scores**: $|\hat{p}(X_l) - \hat{p}(X_i)|$
  - You will then choose the M **closest observations** to observation i, which we can write as $L_M(i) = \{l_1(i), ..., l_M(i)\}$ – these are the matched observations for observation i
  - You will do this for every observation among the treated ones. Upon finishing, you will have the matched sample to the treated observations, i.e., the control group constructed via PS matching
- **Note:**
  - Observations belong to the **other group**
  - Observations are the **closest** to the observation you are focusing on w.r.t **propensity score** (univariate)

- **Matching in practice (cont'd)**
  - Define $L_M(i) = \{l_1(i), ..., l_M(i)\}$ to be the set of indices for the first M matches to observation i
    - M often depends on data availability, but let $M = 2$ to fix ideas
  - The estimates for **potential outcome** for observation i is

$$\hat{y}_i(0) = \begin{cases} y_i, d_i = 0 \\ \frac{1}{M} \sum_{j \in L_M(i)} y_j, d = 1 \end{cases}$$

$$\hat{y}_i(1) = \begin{cases} \frac{1}{M} \sum_{j \in L_M(i)} y_j, d_i = 0 \\ y_i, d = 1 \end{cases}$$

  - In words:
    - When observation i is in the treatment group, there is no need to estimate the potential outcome (we observe it)
    - When we do not observe $y_i(0)$, we estimate it as the average outcome of the M closest matches to observation i
      in the control group

- **Matching in practice: ATE, ATT, ATU**
  - ATE (avg. treatment effect):

$$\frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i(1)-\hat{y}_i(0)]$$

  - ATT (ATE for the treated):

$$\frac{1}{N_1}\sum_{i=1}^{N}[y_i-\hat{y}_i(0)]$$

    where N_1 is the *number of treated observations.*
  - ATU (ATE for the untreated):

$$\frac{1}{N_0}\sum_{i=1}^{N}[\hat{y}_i(1)-y_i]$$

    where N_0 is the *number of untreated observations.*

- **How to estimate the PS?**
  - **Aim:** NOT sign, magnitude, and significance of the covariates
  - **Rather:** Estimate the PS as precisely as possible to mitigate any selection bias
- **Estimation methods**
  - OLS (not great to fit probability distributions – Why?)
  - Maximum likelihood, e.g., probit, logit
  - Nonparametric approaches, e.g., kernel estimator, series estimator
- **Key considerations:** Accuracy and robustness
  - (which need to be assessed in your work!)

- **How many matches (M)?**
  - No objective rule for the optimal number of matches
    - A single (e.g. the best) match leads the less biased and more credible estimates, but the least precise (Usual **bias-variance trade-off**)
    - Goal: choose as many matches as possible, without sacrificing too much in terms of accuracy of the matches
  - Several possibilities
    - Closest neighbor matching chooses the m closest matches
    - Caliper-matching where all the comparison observations falling within a radius are chose as match (e.g., all matches with a PS within 1%)

- In practice, examine results with different choices (If bias is a big concern, there will be variation in the estimated effect)

- Match with or without replacement? (replacement is preferred since the primary objective is to proper identification)
  - i.e., return or not the observation of the treatment group to the "pile of observations" before selecting another one

- **Choice of covariates**
  - As in regressions, this choice is dictated by the particular question
  - General rules
    - Variables affected by the treatment **should not** be included in the set of covariates (match on observables **before** treatment). Careful with PS estimation (only before...)
    - To have low bias, a rich set of variables related to treatment assignment is needed
    - Identification of treatment effects rests on the ability to absorb observable heterogeneity
    - PS matching may be preferred in many applications

- **Matching vs. OLS regression**
  - Matching is not very different from attempting to control for observed covariates in a regression framework
    - Rather than including the imbalanced characteristics in the regression, units with similar characteristics are matched based on these characteristics
  - Both rely on the same assumption: Conditional on the control variables (covariates), the variable of interest (treatment) is **independent** of the error term
  - Distribution versus Variance
    - Matching uses the distribution of covariates to weight covariate-specific estimates into an estimate of the treatment effect
    - Regression produce a variance-weighted average of the treatment effect
  - Matching makes no assumption about the functional form of the tested relationships
  - In practice, it might be wise to perform both
- **Reminder: Matching does not *per se* mitigate endogeneity issues!**
  - It has to be used with something else, e.g., exogenous variation

- Matching methods do not rely on a **clear source of exogenous variation** for identification
    - i.e., no change in policy/ regulation triggering reaction of consumers over time or in the cross-section
- Matching is unlikely to solve an endogeneity problem since it relies only on observables
    - Hypothesis 1: Unconfoundedness
    - Hypothesis 2: Overlap
- Practical uses
    - In conjunction with an exogenous shock
        - (to obtain exogenous variation!)
    - Useful robustness test

- The first key assumption for the identification of treatment effect is referred to as **unconfoundedness**:

$$(Y_1, Y_0) \perp D|X$$

where $y(1), y(0)$ are the potential outcomes under treatment and control, respectively

- **Unconfoundedness** states that *the potential outcomes are independent of the treatment assignment, conditional on the observable covariates*
- Unconfoundedness is equivalent to saying that:
  1. Within each cell defined by X: treatment is random
  2. The selection into treatment depends only on the observables X
- **Intuition:** Consider a regression model of the form

$$y = \beta_0 + \beta_1 d + \beta_2 x_2 + ... + \beta_k x_k + u$$

unconfoundedness is akin to a stronger version of the orthogonality assumption in OLS ($E(d.u|X) = 0$)

- The second identifying assumption is referred to as **overlap**:

$$0 < Pr(d = 1|X) < 1$$

- **In words:** *For each value of the covariates, there is a positive probability of being in the treatment group and in the control group ("common support")*
- **What if this does not hold?**
  - It could be that there is no control unit with some covariate, i.e., there is no unit in the control group similar to a treated unit in terms of covariates
  - It could be that there is no treatment unit matching with the controls for some covariate
  - So, no good data to estimate the counterfactual!

- Under both assumptions, we can treat the outcome of the non-participant that has similar covariate as the participants as if it was the counterfactual outcome for the participants

- **Intuition:**
  - Estimate the treatment effect within each cell defined by X
  - Take the average over the different cells

- The average treatment effect (**ATE**) for a subset of observations with certain covariates ($X = X'$)

$$
\begin{aligned}
ATE(X') &:= E[y(1) - y(0)|X = X'] \\
&= E[y(1) - y(0)|d = d', X = X'] \\
&= E[y|d = 1, X = X'] - E[y|d = 0, X = X']
\end{aligned}
$$

- The first equality follows from unconfoundedness and the second from overlap (requires common values for the X's)

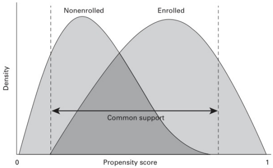
- **Comments:**
  - Agents' optimizing behavior precludes choices being independent of potential outcomes, so unconfoundedness is almost always violated (while being **untestable**)
  - This is the **"Achilles' heel"** of Matching. Because the unconfoundedness assumption is untestable, many econometricians view Matching (by itself) as a "last resort" method. It is often seen as a sophisticated tool for data description or for improving balance, not for identifying causal effects on its own.
  - Still several reasons to investigate ATE
    - Data description (non-causality)
    - Economic theory could help identify the relevant covariates (non-exhaustive though)
    - Not a problem if the choices are driven by unobservables that are unrelated to the outcomes
    - Can be used in conjunction with some exogenous shocks

- **Unconfoundedness**
  - Is untestable because the counterfactual is not observed (similar to the orthogonality condition in regression)
  - If selection is based on unobservables, matching does not mitigate endogeneity
  - Use of placebo tests (estimate treatment effect when there is *a priori* no treatment)
- **Overlap**
  - Plot the distribution of covariates (or $p(x)$) by treatment group. Look at pairs of marginal distributions. But overlap is really about the joint, and not marginal, distribution of the covariates
  - Inspect the quality of the worst matches. If the difference is large relative to the standard deviation of the covariate, one might be concerned about the quality of the matches

- **Important: Lack of common support**
  - For propensity score matching to produce estimates of a program's impact for all treated observations, each treatment or enrolled unit needs to be successfully matched to a non-treated unit (or several of them)
  - In practice, some treated units might find no match
  - In technical terms, there may be a **lack of common support**, or lack of overlap, between the propensity scores of the treatment group and those of the pool of non-treated
- **Intuition:**
  - Sub-samples of treated and control units have to have a similar distribution of propensity scores, so that one can successfully argue that their observed characteristics are similar

- **Illustration:** The figure shows the distribution of propensity scores separately for enrolled and non-enrolled

Figure 8.2 Propensity Score Matching and Common Support



- **Problem:** these distributions do not overlap perfectly
  - Middle: matches are relatively easy to find because there are both the enrolled and non-enrolled with these levels of propensity scores
  - Extremes: enrollees with propensity scores close to 0/1 cannot be matched to any non-enrolled because there are no non-enrolled with such low/high propensity scores, respectively
- **Consequence:**
  - The lack of common support at the tails of the distribution of propensity scores implies that the matching procedure estimates the **Average Treatment Effect on the Treated (ATT)**, and only for the sub-population of treated units on the common support.

- **Summary of procedure:**
  1. Obtain data in which it is possible to identify the units that enrolled in the program and those that did not;
  2. **[Get propensity score]** Estimate the probability that each individual enrolls in the program, based on individual characteristics observed in the survey, e.g., using a logit model;
  3. Restrict the sample to units for which common support appears in the propensity score distribution;
  4. For each enrolled unit, locate a sub-group of non-enrolled units that have **similar** propensity scores;
  5. Compare the outcomes for the treatment/enrolled units and their matched comparison/non-enrolled units. The difference in average outcomes for these two sub-groups is the measure of the impact that can be attributed to the program for that particular treated observation;
  6. The mean of these individual impacts yields an estimate of the local average treatment effect.

- Most softwares have commands that run steps 2-6 automatically

1. Matching relies only on observed chars to construct a comparison group – unobserved chars cannot be taken into account
   - If there are any unobserved chars that affect whether a unit enrolls in the program and the outcome, then the impact estimates obtained with the matched comparison group would be biased
   - For unbiasedness: need strong assumption that there are no unobserved differences in the treatment and comparison groups that are also associated with the outcomes of interest
2. Matching must be done using only chars that are not affected by the program
   - Most chars that are measured after the start of the program would not fall into that category
   - If baseline (pre-intervention) data are not available, the only chars we will be able to use to construct a matched sample are those (usually few) chars that are unaffected by a program, e.g., age and gender. We cannot match using other chars because those are potentially affected by the intervention
   - If baseline data on outcomes are available, combine matching with DD to reduce the risk of bias (see below)
3. Results are only as good as the characteristics that are used
   - This is why the **most robust** applications of Matching **combine** it

- **Idea:** Combine both methods to reduce the risk of bias in the estimation
- **Why:** Because DD takes care of any unobserved characteristics that are constant across time between the two groups
  - i.e., **time-invariant unobservables** common across treatment and control groups
- How?

- **Summary of procedure:**
  1. Perform matching based on observed baseline characteristics;
  2. For each enrolled unit, compute the change in outcomes between the before and after periods (first difference);
  3. For each enrolled unit, compute the change in outcomes between the before and after periods for this unit's matched comparison (second difference);
  4. Subtract the second difference from the first difference; i.e, apply the DD method;
  5. Finally, average out those double differences.

- The **synthetic control method** allows for impact estimation in settings where a single unit (such as a country, a firm, or a hospital) receives an intervention or is exposed to an event
    - Instead of comparing this treated unit to a group of untreated units, the method uses information about the characteristics of the treated unit and the untreated units to construct a "synthetic," or artificial, comparison unit by weighting each untreated unit in such a way that the synthetic comparison unit most closely resembles the treated unit
    - This requires a long series of observations over time of the characteristics of both the treated unit and the untreated units
    - This combination of comparison units into a synthetic unit provides a better comparison for the treated unit than any untreated unit individually

- To be covered later, if time allows

- Matching typically requires extensive data sets on large samples of units
    - Even when those are available, there may be a lack of common support between the treatment and the pool of non-participants
- Matching can only be performed based on observed characteristics
    - By definition, we cannot incorporate unobserved characteristics in the calculation of the propensity score
    - So for the matching procedure to identify a valid comparison group, we must be sure that no systematic differences in unobserved characteristics between the treatment units and the matched comparison units exist that could influence the outcome (Y)
    - Since we cannot prove that there are no such unobserved characteristics that affect both participation and outcomes, we must assume that none exist. This is usually a very strong assumption

- Although matching helps control for observed background characteristics, we can never rule out bias that stems from unobserved characteristics
- Matching alone is generally less robust than the other evaluation methods, since it requires the strong assumption that there are no unobserved characteristics that simultaneously affect program participation and outcomes
    - Randomized assignment, instrumental variable, and regression discontinuity design, on the other hand, do not require the untestable assumption that there are no such unobserved variables
    - They also do not require such large samples or as extensive background characteristics as propensity score matching
    - Thus, matching is used when the use of above methods is not possible
- In summary, the assumption that no selection bias has occurred stemming from unobserved characteristics is overly strong, and most problematically, it cannot be tested

- **Ex post matching** is risky when no baseline data are available neither on the outcome variable nor on background characteristics
  - If an evaluation uses ex post survey data (collected only after the start of program) to infer what people's background characteristics were at baseline, and then matches the treated group to a comparison group using those inferred characteristics, it may inadvertently match based on characteristics that were also affected by the program ⇒ invalid or biased estimation result
- Matching is more reliable when both program assignment rule and underlying variables are known, so matching can be performed on those variables
- **Conclusion:** impact evaluations are best designed **before** a program begins to be implemented
  - Once the program has started, if one has no way to influence how it is allocated and no baseline data have been collected, few, if any, rigorous options for the impact evaluation will be available

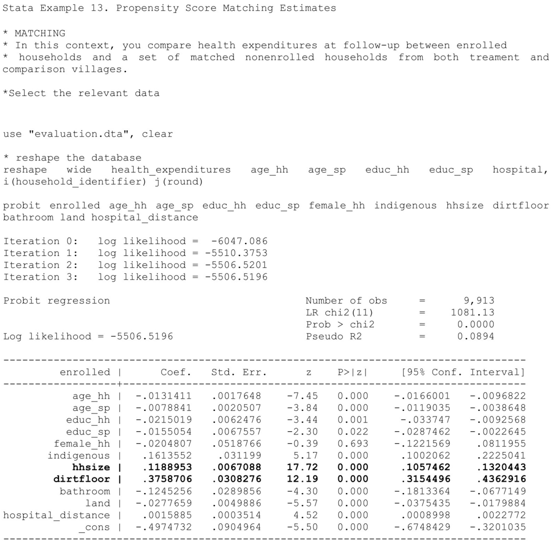
- Matching relies on the assumption that enrolled and non-enrolled units are similar in terms of any unobserved variables that could affect both the probability of participating in the program and the outcome
- Is program participation determined by variables that cannot be observed?
  - This cannot be directly tested, so you will need to rely on theory, common sense, and good knowledge of the setting of the impact evaluation for guidance
- Are the observed characteristics well balanced between matched sub-groups?
  - Compare the observed characteristics of each treatment and its matched comparison group of units at baseline (pre-intervention)
- Can a matched comparison unit be found for each treatment unit?
  - Check whether sufficient **common support** exists in the distribution of the propensity scores
  - Small areas of common support indicate that enrolled and non-enrolled persons are very different, and that casts doubt as to whether matching is a credible method

Classic Application: Evaluating Job Training Programs

- **The Problem:** A famous paper by Lalonde (1986) showed that standard econometric methods (like OLS) failed to replicate the "true" causal effect of a job training program, which was known from an RCT.
- **The Challenge:** The treated group (from the National Supported Work, NSW, demonstration) was extremely disadvantaged. Standard "control" groups from public surveys (like the PSID) were very different, leading to massive selection bias.

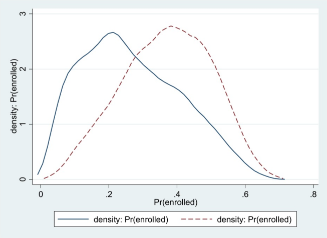
The PSM Solution: Dehejia & Wahba (1999)

- Dehejia and Wahba (1999) re-analyzed this data using Propensity Score Matching.
  - **Step 1 (Estimate PS):** They ran a logit model to predict who was in the treated (NSW) group based on pre-treatment observables (age, education, race, 1975 earnings, etc.).
  - **Step 2 (Common Support):** They enforced the common support condition, dropping many individuals from the PSID group who were "too different" (e.g., had propensity scores outside the range of the treated group).
  - **Step 3 (Match & Estimate):** They matched each treated participant to the "closest" control individuals based on this score.
- **The Result:** After matching, their simple PSM estimate of the treatment effect was remarkably close to the "true" experimental effect from the RCT. This paper is the canonical example that demonstrated the power (and limitations) of PSM.



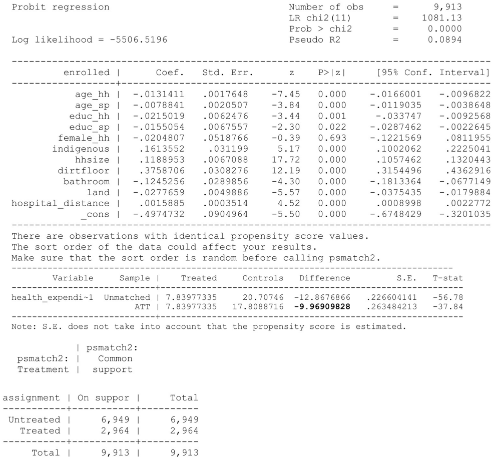
```
Stata Example 13. Propensity Score Matching Estimates

* MATCHING
* In this context, you compare health expenditures at follow-up between enrolled
*  households and a set of matched nonenrolled households from both treament and
comparison villages.

*Select the relevant data


use "evaluation.dta", clear

* reshape the database
reshape  wide  health_expenditures  age_hh  age_sp  educ_hh  educ_sp  hospital,
i(household_identifier) j(round)

probit enrolled age_hh age_sp educ_hh educ_sp female_hh indigenous hhsize dirtfloor
bathroom land hospital_distance

Iteration 0:   log likelihood =  -6047.086
Iteration 1:   log likelihood = -5510.3753
Iteration 2:   log likelihood = -5506.5201
Iteration 3:   log likelihood = -5506.5196

Probit regression                               Number of obs     =      9,913
                                                LR chi2(11)       =    1081.13
                                                Prob > chi2       =     0.0000
Log likelihood = -5506.5196                     Pseudo R2         =     0.0894

------------------------------------------------------------------------------------
         enrolled |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
           age_hh |  -.0131411   .0017648    -7.45   0.000    -.0166001   -.0096822
           age_sp |  -.0078841   .0020507    -3.84   0.000    -.0119035   -.0038648
          educ_hh |  -.0215019   .0062476    -3.44   0.001     -.033747   -.0092568
          educ_sp |  -.0155054   .0067557    -2.30   0.022    -.0287462   -.0022645
        female_hh |  -.0204807   .0518766    -0.39   0.693    -.1221569    .0811955
       indigenous |   .1613552    .031199     5.17   0.000     .1002062    .2225041
           hhsize |   .1188953   .0067088    17.72   0.000     .1057462    .1320443
        dirtfloor |   .3758706   .0308276    12.19   0.000     .3154496    .4362916
         bathroom |  -.1245256   .0289856    -4.30   0.000    -.1813364   -.0677149
             land |  -.0277659   .0049886    -5.57   0.000    -.0375435   -.0179884
hospital_distance |   .0015885   .0003514     4.52   0.000     .0008998    .0022772
            _cons |  -.4974732   .0904964    -5.50   0.000    -.6748429   -.3201035
------------------------------------------------------------------------------------
```

```
predict pscore

kdensity pscore if enrolled ==1, gen(take1 den1)

kdensity pscore if enrolled ==0, gen(take0 den0)

graph twoway (line den0 take0, lpattern(solid)) (line den1 take1, lpattern(dash))
```



-

```
set seed 100
generate u=runiform()
sort u

psmatch2 enrolled age_hh age_sp educ_hh educ_sp female_hh indigenous hhsize dirtfloor
bathroom land hospital_distance, out(health_expenditures1)
```



```
Probit regression                               Number of obs     =      9,913
                                                LR chi2(11)       =    1081.13
                                                Prob > chi2       =     0.0000
Log likelihood = -5506.5196                     Pseudo R2         =     0.0894

-----------------------------------------------------------------------------------
         enrolled |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------+----------------------------------------------------------------
           age_hh |  -.0131411   .0017648    -7.45   0.000    -.0166001   -.0096822
           age_sp |  -.0078841   .0020507    -3.84   0.000    -.0119035   -.0038648
          educ_hh |  -.0215019   .0062476    -3.44   0.001     -.033747   -.0092568
          educ_sp |  -.0155054   .0067557    -2.30   0.022    -.0287462   -.0022645
        female_hh |  -.0204807   .0518766    -0.39   0.693    -.1221569    .0811955
       indigenous |   .1613552    .031199     5.17   0.000     .1002062    .2225041
           hhsize |   .1188953   .0067088    17.72   0.000     .1057462    .1320443
        dirtfloor |   .3758706   .0308276    12.19   0.000     .3154496    .4362916
         bathroom |  -.1245256   .0289856    -4.30   0.000    -.1813364   -.0677149
             land |  -.0277659   .0049886    -5.57   0.000    -.0375435   -.0179884
hospital_distance |   .0015885   .0003514     4.52   0.000     .0008998    .0022772
            _cons |  -.4974732   .0904964    -5.50   0.000    -.6748429   -.3201035
-----------------------------------------------------------------------------------
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
----------------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
----------------------------+-----------------------------------------------------------
health_expendi~1  Unmatched | 7.83977335     20.70746  -12.8676866   .226604141   -56.78
                        ATT | 7.83977335   17.8088716  -9.96909828   .263484213   -37.84
----------------------------+-----------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support

assignment | On suppor |     Total
-----------+-----------+----------
 Untreated |     6,949 |     6,949
   Treated |     2,964 |     2,964
-----------+-----------+----------
     Total |     9,913 |     9,913
```

- STATA program `pscore.ado` is available from http://www.iue.it/Personal/Ichino/Welcome.html
- The table shows the simple PSM estimate for the ATT. This compares only post-treatment outcomes. This is risky, as it relies 100% on the unconfoundedness assumption.

- A more robust method is the Matched DD (slides 28-29).
  - This uses the pre-treatment data to difference out any time-invariant unobserved differences between the groups, relaxing our assumptions (see Appendix for conceptual R Code and the Lab).
- This is the estimator used in the Fowlie et al (2012) paper discussed below.
  - see also Huse, C. and N. Koptyug (2017). "Bailing On The Car That Wasn't Bailed Out: Bounding Consumer Reactions to Financial Distress". Journal of Economics and Management Strategy 26 (2), 337-374. https://doi.org/10.1111/jems.12184

Overview

- **Paper:** Fowlie, Holland, and Mansur (2012). What Do Emissions Markets Deliver and to Whom? Evidence from Southern California's NOx Trading Program. American Economic Review.
- **Research questions:** (CAC=command-and-control)
  - Did emissions reductions at facilities subject to Southern California's RECLAIM program exceed emissions reductions achieved at very similar facilities subject to CAC regulation over the same time period?
  - Has the compliance flexibility afforded by market-based environmental regulation resulted in more (or less) pollution in traditionally disadvantaged communities?
- **Objective:** identify the causal effects of this emissions trading program on facility-level emissions *vis-à-vis* the CAC regulations it replaced
- **Empirical strategy:** Match facilities in the RECLAIM program with similar California facilities also in non-attainment areas

Overview cont'd

- **Motivation:** An advantage of cap-and-trade programs over more prescriptive environmental regulation (e.g., CAC) is that **compliance flexibility** and **cost effectiveness** can make more stringent emissions reductions politically feasible. However, when markets (versus regulators) determine where emissions occur, it becomes more difficult to assure that mandated emissions reductions are **equitably** achieved
- **Results:**
  - Average emissions fell 20 percent at RECLAIM facilities relative to the counterfactual
  - Observed changes in emissions do not vary significantly with neighborhood demographic characteristics
- **Big picture:** After a long period under which CACs were predominant, the 1990 CAA Amendments introduced market-based policy instruments, thus the interest to evaluate/compare them

Institutional Background

- **RECLAIM**: REgional CLean Air Incentives Market (1994)
  - In short, an emissions-trading programme
- Recall aim: want to identify the causal effects of RECLAIM *vis-à-vis* the CAC regulations it replaced
- Argument for matching: only a subset of industrial facilities located in non-attainment counties in California were removed from a CAC regime and required to participate in RECLAIM
- Control group: similar California facilities that remained subject to CAC regulation over the duration of the study period
- Importance of RECLAIM:
  - First mandatory trading program to supplant a preexisting CAC regime that was, in theory, capable of achieving the same environmental objectives
  - First program to include a broad and diverse population of sources
  - First emissions trading program to be challenged on the grounds of environmental injustice and noncompliance

Institutional Background

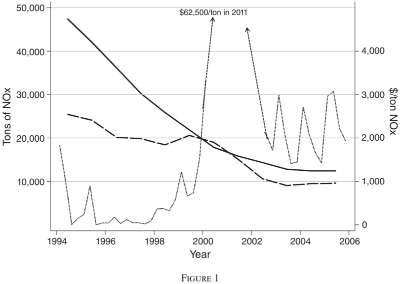


FIGURE 1

Note: Trends in nitrogen oxides emissions (dashed), allocations (solid), and permit price (dotted).

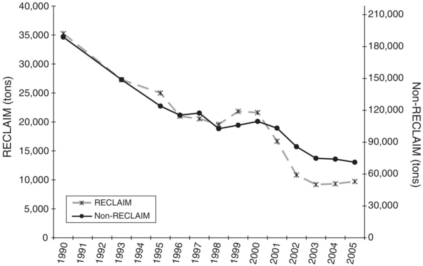


FIGURE 2. TOTAL NO_x EMISSIONS IN RECLAIM AND IN THE REST OF CALIFORNIA

- Fig. 1: Note how prices shoot up as demand increases towards supply (late 1990s)
- Fig. 2: Note evidence of pre-1994 common trends, downward trends in both groups even pre-RECLAIM

Data

- CARB (California Air Resources Board) maintains a database of emissions reports from the Air Quality Management Districts
    - Data also includes information on industry classification
- Use addresses, geocodes, and industry classifications to ensure a consistent coding of facilities over time
- Also use separate emissions data from RECLAIM to verify the emissions reported to the ARB database (SCAQMD 2007)
- Demographics from Census (1990, 2000) at the block group level
    - Median HH income (1989, 1999), population by ethnicity and race
    - Construct measure of percent minority as the percent of the population that is either non-Hispanic black or Hispanic
    - To account for the possibility that HH can sort based on pollution exposure, emphasize the 1990 data, versus the more recent 2000 data that may be endogenous to emissions due to sorting
- Demographic data is merged with emissions data using circles surrounding each facility (0.5, 1, 2 miles) – see paper for details

Empirical Strategy

- Paper combines matching and difference-in-differences (DD) approaches
  - In particular, use nearest-neighbor matching estimator which weights control facilities according to their similarity to treated facilities, where similarity is based on X, a vector of observable facility and neighborhood characteristics (such as historic emissions, industry classification, county attainment status)
- Crucial assumptions (check R code in Appendix and Lab):
  1. **Parallel trends:** The "DD" part (testable with pre-treatment data, see Fig 2).
  2. **Unconfoundedness:** The "Matching part" (assumed, relies on good X's). Any differences between TG and CG can be removed by adjusting for differences in observable covariates
  3. **SUTVA:** No spillover between plants, no general equilibrium effects.

Summary Statistics

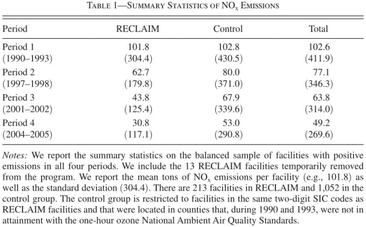


Table 1—Summary Statistics of NO_x Emissions

Period	RECLAIM	Control	Total
Period 1	101.8	102.8	102.6
(1990–1993)	(304.4)	(430.5)	(411.9)
Period 2	62.7	80.0	77.1
(1997–1998)	(179.8)	(371.0)	(346.3)
Period 3	43.8	67.9	63.8
(2001–2002)	(125.4)	(339.6)	(314.0)
Period 4	30.8	53.0	49.2
(2004–2005)	(117.1)	(290.8)	(269.6)

Notes: We report the summary statistics on the balanced sample of facilities with positive emissions in all four periods. We include the 13 RECLAIM facilities temporarily removed from the program. We report the mean tons of NO_x emissions per facility (e.g., 101.8) as well as the standard deviation (304.4). There are 213 facilities in RECLAIM and 1,052 in the control group. The control group is restricted to facilities in the same two-digit SIC codes as RECLAIM facilities and that were located in counties that, during 1990 and 1993, were not in attainment with the one-hour ozone National Ambient Air Quality Standards.

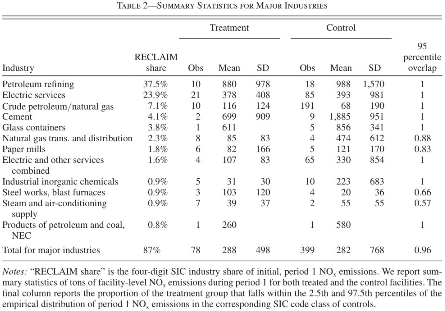


Table 2—Summary Statistics for Major Industries

		Treatment			Control			
Industry	RECLAIM share	Obs	Mean	SD	Obs	Mean	SD	95 percentile overlap
Petroleum refining	37.5%	10	880	978	18	988	1,570	1
Electric services	23.9%	21	378	408	85	393	981	1
Crude petroleum/natural gas	7.1%	10	116	124	191	68	190	1
Cement	4.1%	2	699	909	9	1,885	951	1
Glass containers	3.8%	1	611		5	856	341	1
Natural gas trans. and distribution	2.3%	8	85	83	4	474	612	0.88
Paper mills	1.8%	6	82	166	5	121	170	0.83
Electric and other services combined	1.6%	4	107	83	65	330	854	1
Industrial inorganic chemicals	0.9%	5	31	30	10	223	683	1
Steel works, blast furnaces	0.9%	3	103	120	4	20	36	0.66
Steam and air-conditioning supply	0.9%	7	39	37	2	55	55	0.57
Products of petroleum and coal, NEC	0.8%	1	260		1	580		1
Total for major industries	87%	78	288	498	399	282	768	0.96

Notes: "RECLAIM share" is the four-digit SIC industry share of initial, period 1 NO_x emissions. We report summary statistics of tons of facility-level NO_x emissions during period 1 for both treated and the control facilities. The final column reports the proportion of the treatment group that falls within the 2.5th and 97.5th percentiles of the empirical distribution of period 1 NO_x emissions in the corresponding SIC code class of controls.

Main Results

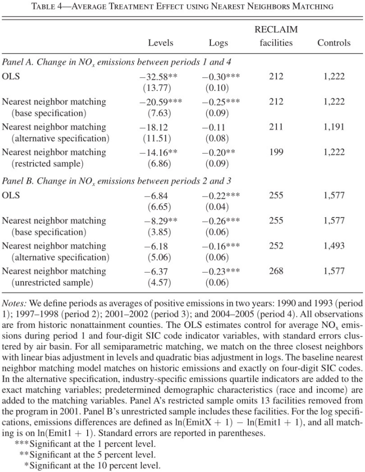


Table 4—Average Treatment Effect using Nearest Neighbors Matching

	Levels	Logs	RECLAIM facilities	Controls
Panel A. Change in NO$_x$ emissions between periods 1 and 4				
OLS	−32.58** (13.77)	−0.30*** (0.10)	212	1,222
Nearest neighbor matching (base specification)	−20.59*** (7.63)	−0.25*** (0.09)	212	1,222
Nearest neighbor matching (alternative specification)	−18.12 (11.51)	−0.11 (0.08)	211	1,191
Nearest neighbor matching (restricted sample)	−14.16** (6.86)	−0.20** (0.09)	199	1,222
Panel B. Change in NO$_x$ emissions between periods 2 and 3				
OLS	−6.84 (6.65)	−0.22*** (0.04)	255	1,577
Nearest neighbor matching (base specification)	−8.29** (3.85)	−0.26*** (0.06)	255	1,577
Nearest neighbor matching (alternative specification)	−6.18 (5.06)	−0.16*** (0.06)	252	1,493
Nearest neighbor matching (unrestricted sample)	−6.37 (4.57)	−0.23*** (0.06)	268	1,577

Notes: We define periods as averages of positive emissions in two years: 1990 and 1993 (period 1); 1997–1998 (period 2); 2001–2002 (period 3); and 2004–2005 (period 4). All observations are from historic nonattainment counties. The OLS estimates control for average NO$_x$ emissions during period 1 and four-digit SIC code indicator variables, with standard errors clustered by air basin. For all semiparametric matching, we match on the three closest neighbors with linear bias adjustment in levels and quadratic bias adjustment in logs. The baseline nearest neighbor matching model matches on historic emissions and exactly on four-digit SIC codes. In the alternative specification, industry-specific emissions quartile indicators are added to the exact matching variables; predetermined demographic characteristics (race and income) are added to the matching variables. Panel A's restricted sample omits 13 facilities removed from the program in 2001. Panel B's unrestricted sample includes these facilities. For the log specifications, emissions differences are defined as ln(EmitX + 1) − ln(Emit1 + 1), and all matching is on ln(Emit1 + 1). Standard errors are reported in parentheses.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

- Stronger evidence of long-run effect (periods 1-4) in levels
- Effect in logs are typically significant

Robustness

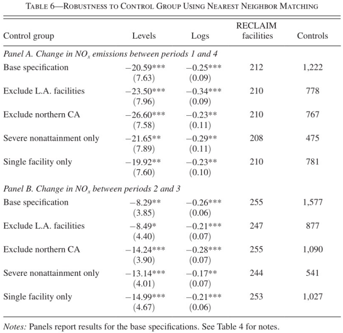


Table 6—Robustness to Control Group Using Nearest Neighbor Matching

Control group	Levels	Logs	RECLAIM facilities	Controls
Panel A. Change in NO_x emissions between periods 1 and 4				
Base specification	−20.59*** (7.63)	−0.25*** (0.09)	212	1,222
Exclude L.A. facilities	−23.50*** (7.96)	−0.34*** (0.09)	210	778
Exclude northern CA	−26.60*** (7.58)	−0.23** (0.11)	210	767
Severe nonattainment only	−21.65** (7.89)	−0.29** (0.11)	208	475
Single facility only	−19.92** (7.60)	−0.23** (0.10)	210	781
Panel B. Change in NO_x between periods 2 and 3				
Base specification	−8.29** (3.85)	−0.26*** (0.06)	255	1,577
Exclude L.A. facilities	−8.49* (4.40)	−0.21*** (0.07)	247	877
Exclude northern CA	−14.24*** (3.90)	−0.28*** (0.07)	255	1,090
Severe nonattainment only	−13.14*** (4.01)	−0.17** (0.07)	244	541
Single facility only	−14.99*** (4.67)	−0.21*** (0.06)	253	1,027

Notes: Panels report results for the base specifications. See Table 4 for notes.

- Robustness checks:
  - Row 1: dropping the closest facilities in the control group
  - Row 2: dropping the facilities farthest away
  - Row 3: using only data from facilities in severe (versus moderate) non-attainment areas as controls
  - Row 4: control group is restricted to single plant firms (no way to shift production)

Heterogeneous Treatment Effects

- Are the reduction in emissions that occurred under RECLAIM, in comparison to those in the control group, correlated with demographics? Focus on the θ parameters of equation (4):

$$Y_{it'} - Y_{it^0} = \delta_j + \boldsymbol{\beta}'\boldsymbol{X}_i + \boldsymbol{\theta}'\boldsymbol{X}_i D_i + \alpha D_i + \varepsilon_i$$

- Focus: historic emissions, income, and percent minority

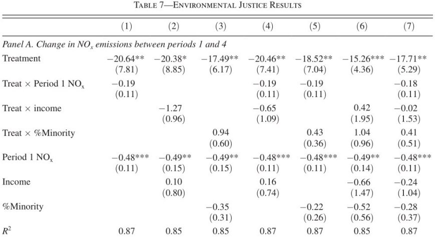
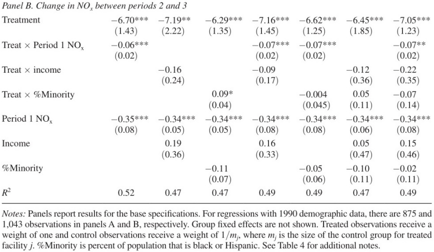


TABLE 7—ENVIRONMENTAL JUSTICE RESULTS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Change in NO_x emissions between periods 1 and 4							
Treatment	−20.64** (7.81)	−20.38* (8.85)	−17.49** (6.17)	−20.46** (7.41)	−18.52** (7.04)	−15.26*** (4.36)	−17.71** (5.29)
Treat × Period 1 NO_x	−0.19 (0.11)			−0.19 (0.11)	−0.19 (0.11)		−0.18 (0.11)
Treat × income		−1.27 (0.96)		−0.65 (1.09)		0.42 (1.95)	−0.02 (1.53)
Treat × %Minority			0.94 (0.60)		0.43 (0.36)	1.04 (0.96)	0.41 (0.51)
Period 1 NO_x	−0.48*** (0.11)	−0.49** (0.15)	−0.49** (0.15)	−0.48*** (0.11)	−0.48*** (0.11)	−0.49** (0.14)	−0.48*** (0.11)
Income		0.10 (0.80)		0.16 (0.74)		−0.66 (1.47)	−0.24 (1.04)
%Minority			−0.35 (0.31)		−0.22 (0.26)	−0.52 (0.56)	−0.28 (0.37)
R^2	0.87	0.85	0.85	0.87	0.87	0.85	0.87



	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel B. Change in NO_x between periods 2 and 3							
Treatment	−6.70*** (1.43)	−7.19** (2.22)	−6.29*** (1.35)	−7.16*** (1.45)	−6.62*** (1.25)	−6.45*** (1.85)	−7.05*** (1.23)
Treat × Period 1 NO_x	−0.06*** (0.02)			−0.07*** (0.02)	−0.07*** (0.02)		−0.07** (0.02)
Treat × income		−0.16 (0.24)		−0.09 (0.17)		−0.12 (0.36)	−0.22 (0.35)
Treat × %Minority			0.09* (0.04)		−0.004 (0.045)	0.05 (0.11)	−0.07 (0.14)
Period 1 NO_x	−0.35*** (0.08)	−0.34*** (0.05)	−0.34*** (0.05)	−0.34*** (0.08)	−0.34*** (0.08)	−0.34*** (0.06)	−0.34*** (0.08)
Income		0.19 (0.36)		0.16 (0.33)		0.05 (0.47)	0.15 (0.46)
%Minority			−0.11 (0.07)		−0.05 (0.06)	−0.10 (0.11)	−0.02 (0.11)
R^2	0.52	0.47	0.47	0.49	0.49	0.47	0.49

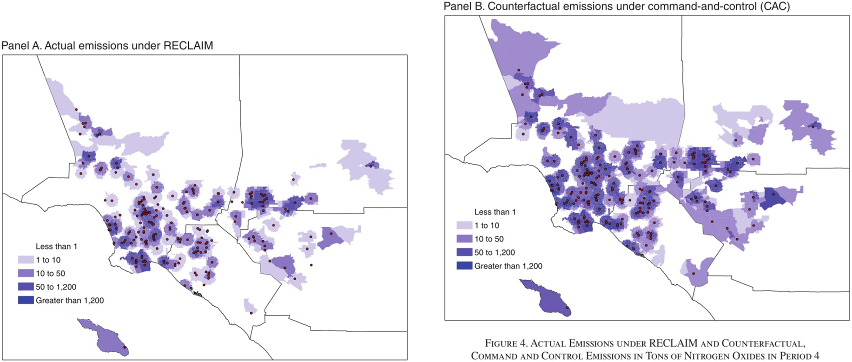
Notes: Panels report results for the base specifications. For regressions with 1990 demographic data, there are 875 and 1,043 observations in panels A and B, respectively. Group fixed effects are not shown. Treated observations receive a weight of one and control observations receive a weight of $1/m_j$, where m_j is the size of the control group for treated facility *j*. %Minority is percent of population that is black or Hispanic. See Table 4 for additional notes.

Heterogeneous Treatment Effects cont'd

- Panel B: variable $\mathsf{Treat} \times \mathsf{period1NOx}$ is statistically significant, indicating larger emissions reductions at larger facilities
- Overall, finding that RECLAIM facilities polluting more in period 1 reduced emissions more during this time period
- However, no evidence of 1990 demographics being a significant determinant of which facilities reduced emissions

Counterfactuals

- Geographic distribution of emissions under RECLAIM and the CAC counterfactual
  - Evidence of spatial clustering either way – so no “perverse effects” or creation of hot-spots due to use of the market mechanism
  - Moreover, recall that avg. emissions are lower under RECLAIM as compared to the CAC counterfactual


Panel A. Actual emissions under RECLAIM

Panel B. Counterfactual emissions under command-and-control (CAC)

Figure 4. Actual Emissions under RECLAIM and Counterfactual, Command and Control Emissions in Tons of Nitrogen Oxides in Period 4

- Matching is an appropriate method whenever observable characteristics are rich enough to control for treated vs. control group heterogeneity
- Matching is also intuitive, and advanced matching techniques have received substantial attention
  - e.g., synthetic matching
- The main concern when using the method is the role played by unobserved heterogeneity – how large, how important is it in your particular case?

- Gertler et al (2016). Impact Evaluation in Practice, 2nd. Edition. Washington, DC: Inter-American Development Bank and World Bank
  - Chapter 8
- Gertler et al (2016). Impact Evaluation in Practice, 2nd. Edition, Technical Companion (Version 1.0). Washington, DC: Inter-American Development Bank and World Bank.
  - p. 28-32
- Jalan, Jyotsna, and Martin Ravallion. 2003. “Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching.” *Journal of Business & Economic Statistics* 21 (1): 19–30.
- Rosenbaum, P.R. and D.B. Rubin (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects”, *Biometrika* 70, 1, 41–55.

- **Matching on the propensity score: Intuition**
    - Consider a regression model of the form

$$y = \beta_0 + \beta_1 d + \beta_2 x_2 + ... + \beta_k x_k + u$$

    - Omitting the controls X leads to bias in the estimated treatment effect (β_1)
    - By conditioning on PS, we remove the correlation between X and d because $X \perp d | PS$ (see slide on unconfoundedness)
    - Omitting X after conditioning on PS no longer leads to bias – but can be inefficient though
- **Note:**
    - The dimensionality of X is typically large so an exact match is typically impossible. In fact, it becomes difficult even to find "close matches"
    - PS avoids the choice of a weighting scheme or norm to account for differences in covariate

- **Choice of the distance metric**
  - There are several alternative for measuring the distance between units, each one with its pros and cons (see Imbens 2004)
  - Euclidean metric (no scaling): $||X_i - X_j|| = \sqrt{(X_i - X_j)'(X_i - X_j)}$
  - Euclidean metric (scaled by variance): $||X_i - X_j|| = \sqrt{(X_i - X_j)' diag(\Sigma_X^{-1})(X_i - X_j)}$
  - Mahalanobis metric (quite popular): $||X_i - X_j|| = \sqrt{(X_i - X_j)'\Sigma_X^{-1}(X_i - X_j)}$

- A more robust method is the Matched DD (slides 28-29). This uses the pre-treatment data to difference out any time-invariant unobserved differences between the groups, relaxing our assumptions.
- Conceptual R Code:

```
# Assumes your full panel data is in a dataframe ’df’ with columns: id,
time (0=pre, 1=post), outcome, enrolled, x_vars...
# 1. Load necessary libraries
library(dplyr)
library(MatchIt)
library(fixest) # For regression with clustered SEs
# 2. Create a baseline (pre-treatment) dataframe
df_baseline <- filter(df, time == 0)
# 3. Run Propensity Score Matching on baseline data
# We match on pre-treatment observables (x_vars from slides 37-39)
m.out <- matchit(enrolled ~ age_hh + age_sp + educ_hh + educ_sp
+ female_hh + indigenous + hhsize + dirtfloor + bathroom + land +
hospital_distance,
data = df_baseline,
method = "nearest", # 1-to-1 nearest neighbor
distance = "logit") # Use logit for propensity score
# 4. Get the IDs of the matched units (treated + their controls)
matched_ids <- match.data(m.out)$id # Assumes a unit ID column
’id’
# 5. Create the new "Matched Panel" by filtering the *original*
# full panel dataframe to keep only the matched units.
matched_panel_data <- filter(df, id %in% matched_ids)
```

```
# 6. Run the final DD regression on this matched panel
# We cluster standard errors by unit ’id’
model_matched_dd <- feols(outcome ~ enrolled + post +
enrolled:post,
data = matched_panel_data,
cluster = ~id)
summary(model_matched_dd)
# The coefficient on ’enrolled:post’ is the Matched DD estimate (ATT).
# This is the estimator used in the Fowlie (2012) paper.
```