Chapter 13
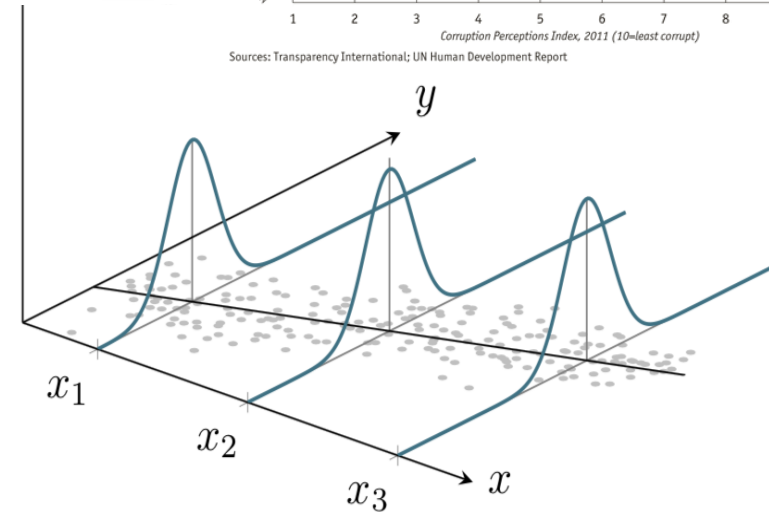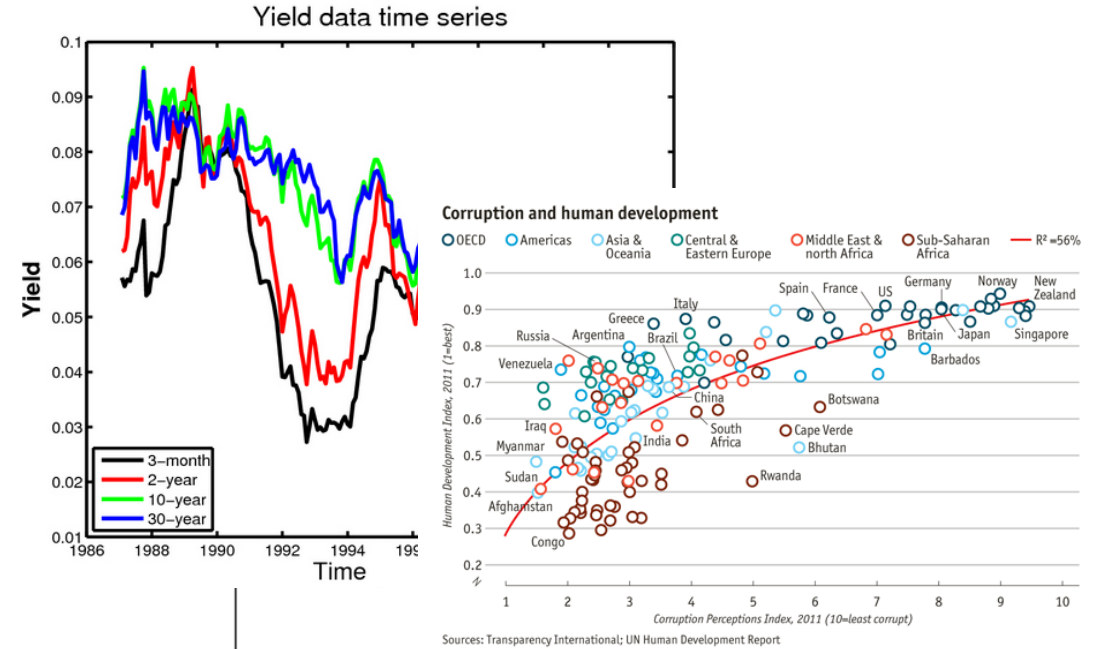
Pooling Cross Sections across Time: Simple Panel Data Methods

- **Policy analysis with pooled cross sections**
  - Two or more independently sampled cross sections can be used to evaluate the impact of a certain event or policy change.

  - Example: Effect of new garbage incinerator's location on housing prices.
  - Examine the effect of the location of a house on its price before and after the garbage incinerator was built:

$$\widehat{rprice} = \begin{array}{cc} 101,307.5 & - & 30,688.27 \ nearinc \\ (3,093.0) & & (5,827.71) \end{array}$$

After incinerator was built

$$n = 142, R^2 = .165$$

$$\widehat{rprice} = \begin{array}{cc} 82,517.23 & - & 18,824.37 \ nearinc \\ (2,653.79) & & (4,744.59) \end{array}$$

Before incinerator was built

$$n = 179, R^2 = .082$$

- **Example: Garbage incinerator and housing prices (cont.)**
  - It would be wrong to conclude from the regression after the incinerator is there that being near the incinerator depresses prices so strongly.
  - One has to compare with the situation before the incinerator was built:

$$\hat{\delta}_1 = -30,688.27 - (-18,824.37) = -11,863.9$$

Incinerator depresses prices but location
was one with lower prices anyway

- In the given case, this is equivalent to

$$\hat{\delta}_1 = (\overline{rprice}_{1,nr} - \overline{rprice}_{1,fr}) - (\overline{rprice}_{0,nr} - \overline{rprice}_{0,fr})$$

- This is called the difference-in-differences estimator (DiD)

- **Difference-in-differences in a regression framework**

$$rprice = \beta_0 + \delta_0\, after + \beta_1\, nearinc + \delta_1\, after \cdot nearinc + u$$

Differential effect of being in the location <u>and</u> after the incinerator was built

- In this way standard errors for the DiD-effect can be obtained.
- If houses sold before and after the incinerator was built were systematically different, further explanatory variables should be included.
- This will also reduce the error variance and thus standard errors.

- Before/After comparisons in "natural experiments"
  - DiD can be used to evaluate policy changes or other exogenous events.

- **Policy evaluation using difference-in-differences**

$$y = \beta_0 + \delta_0 after + \beta_1 treated + \delta_1 after \cdot treated + other factors$$

$$\hat{\delta}_1 = (\overline{y}_{1,T} - \overline{y}_{1,C}) - (\overline{y}_{0,T} - \overline{y}_{0,C}) \leftarrow$$ Compare outcomes of the two groups before and after the policy change

- Compare the difference in outcomes of the units that are affected by the policy change (= treatment group) and those who are not affected (= control group) before and after the policy was enacted.

- For example, the level of unemployment benefits is cut but only for group A (= treatment group). Group A normally has longer unemployment durations than group B (= control group). If the diffe-rence in unemployment durations between group A and group B becomes smaller after the reform, reducing unemployment benefits reduces unemployment duration for those affected.

- Caution: Difference-in-differences only works if the difference in outcomes between the two groups is not changed by other factors than the policy change (e.g. there must be no differential trends).

- **Another interpretation of the difference-in-differences estimator**
  - We can re-write the DiD estimator as:

$$\hat{\delta}_1 = \left(\bar{y}_{1,T} - \bar{y}_{0,T} \right) - \left(\bar{y}_{1,C} - \bar{y}_{0,C} \right)$$

- The first term is the difference in means over time for the treated group
  - This would be a good estimator of the policy effect only if no external factors changed across the two time periods.
- The second term is the difference in means over time for the control group.
  - Subtracting off this term hopefully controls for any changes in external factors that are common to both the treated and control groups, which will be the case when we have random assignment.
  - In this case, the DiD estimator can be interpreted as the average treatment effect.

- **Adding an additional control group**
- The standard two-group, two period difference-in-differences setup relies on the assumption of parallel trends.
  - Parallel trends assumes that any trends in the outcome y would trend at the same rate in the absence of the intervention.
  - Prior to the intervention, y should move in the same direction for both groups.
- The standard DiD estimator measures the difference in estimated trends between the two groups.
  - If the parallel trends assumption is violated, we cannot be sure that the DiD estimator is identifying the effects of the policy or simply some other unaccounted factor causing different trends between these groups.
- We can add flexibility by adding an additional control group.

- **Adding an additional control group (contd)**
- Example: The effects of expanding health care for low income families in a particular state.
  - Let L denote low-income families (eligible for the policy) and M be middle-income families (not eligible).
  - Let B denote states that implemented the policy and A be states that did not implement the policy.
  - The policy is implemented in period 1 , but no policy exists in period 0.

- The additional control group (income level) allows for more flexibility if we assume that any difference in trends in health outcomes between low and middle income families is similar across states.

- **Adding an additional control group (contd)**

$$y = \beta_0 + \beta_1 dL + \beta_2 dB + \beta_3 dL * dB + \delta_0 d1 + \delta_1 d1 * dL + \delta_2 d1 * dB + \delta_3 d1 * dL * dB + u$$

$$\hat{\delta}_3 = \left[(\bar{y}_{1,L,B} - \bar{y}_{0,L,B}) - (\bar{y}_{1,M,B} - \bar{y}_{0,M,B})\right] - \left[(\bar{y}_{1,L,A} - \bar{y}_{0,L,A}) - (\bar{y}_{1,M,A} - \bar{y}_{0,M,A})\right]$$

$$\hat{\delta}_3 = \hat{\delta}_{DD,B} - \hat{\delta}_{DD,A} = \hat{\delta}_{DDD}$$

  - The difference-in-difference-in differences estimator has two components
    - A DD estimator looking only at states that implemented the policy.
    - A DD estimator looking only at states that did not implement the policy.
  - If health trends between the L and M groups do not differ in non-implementation states, then the second component vanishes and we are back to the standard DiD setup.
    - However, we include this second term to account for possibly different trends in the L and M groups that are common across both states A and B.

- **A General Framework for Policy Analysis**
- A more general approach to policy analysis is to include multiple control and treatment groups as well as more than two time periods.
  - Some units may never be treated and others may be treated in different time periods.
  - With this general framework, we should avoid trying to fit the problem into the basic DD setup.
- Let each observation i belong to a pair (g, t) , where g is a group and t is a time period.
  - We are interested in policy interventions that occur at the group level and in order to be convincing, there should be before and after periods for at least some of the groups in our study.

- **A General Framework for Policy Analysis (contd)**
  - Indicate the policy with a dummy variable $x_{g,t}$ for group g and time t

$$y_{i,g,t} = \lambda_t + \alpha_g + \beta x_{g,t} + \boldsymbol{z_{i,g,t}\gamma} + u_{i,g,t}$$

Aggregate time effect common to all groups

Fixed effects specific to each group

Policy effect

Explanatory variables measured at the individual and group level

  - This model can be expanded upon with the addition of lagged policy effects or allowing the policy to have different effects between groups.

- The parallel trends assumption can also be relaxed (for T>2):

$$y_{i,g,t} = \lambda_t + \alpha_g + \psi_g t + \beta x_{g,t} + \boldsymbol{z_{i,g,t}\gamma} + u_{i,g,t}$$ ⟵ ψ_g captures the linear time trend by group

- **Two-period panel data analysis**

- Example: Effect of unemployment on city crime rate
  - Assume that no other explanatory variables are available. Will it be possible to estimate the causal effect of unemployment on crime?
  - Yes, if cities are observed for at least two periods and other factors affecting crime stay approximately constant over those periods:

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}, \quad t = 1982, 1987$$

Time dummy for the second period

Unobserved time-constant factors (= <u>fixed effect</u>)

Other unobserved factors (= idiosyncratic error)

- **Example: Effect of unemployment on city crime rate (cont.)**

$$crmrte_{i1987} = \beta_0 + \delta_0 \cdot 1 + \beta_1\, unem_{i1987} + a_i + u_{i1987}$$

$$crmrte_{i1982} = \beta_0 + \delta_0 \cdot 0 + \beta_1\, unem_{i1982} + a_i + u_{i1982}$$

Subtract: $\quad \Rightarrow \quad \Delta crmrte_i = \delta_0 + \beta_1\, \Delta unem_i + \Delta u_i$ ← Fixed effect drops out

- Estimate differenced equation by OLS:

$$\widehat{\Delta crmrte} = \begin{array}{c} 15.40 \\ (4.70) \end{array} + \begin{array}{c} 2.22\ \Delta unem \\ (.88) \end{array}$$

← + 1 percentage point unemployment rate leads to 2.22 more crimes per 1,000 people

$$n = 46, R^2 = .127$$

Secular increase in crime

Simple Panel Data Methods

- **Discussion of first-differenced panel estimator**
- Further explanatory variables may be included in original equation.

- Note that there may be arbitrary correlation between the unobserved time-invariant characteristics and the included explanatory variables.
  - OLS in the original equation would therefore be inconsistent.
  - The first-differenced panel estimator is thus a way to consistently estimate causal effects in the presence of time-invariant endogeneity.

- For consistency, strict exogeneity has to hold in the original equation.
- First-differenced estimates will be imprecise if explanatory variables vary only little over time (no estimate possible if time-invariant).