13 Basic Panel Data Methods

Independently pooled cross section data is obtained by sampling randomly from a large population at different points in time. This likely leads to observations which are not identically distributed i.e. the distribution of x or y changes over time.

Panel data is obtained by following the same observation overtime. The unit of observation could be individuals, families, firms, products, cities, states or even countries. The crucial difference between panel and independently pooled cross sections is that the former is not independently distributed across time.

13.1 Pooling Independent Cross Sections across Time

1. The Current Population Survey for the US and the Consumption Rounds for the National Sample Survey, MOSPI are examples of independent cross sections which could be pooled for analysis.

2. Why would we want to pool?

   (a) To increase the sample size assuming that the relation between the y and the x's has not changed overtime.

   (b) In case we suspect instability in relationship, we could put dummy for different time periods and examine the pattern of coefficients. Such dummies are called 'year dummies' (or time dummies) and they capture the conditional variability in y over the repeated cross sections.

3. We will use methods used in Chapter 7, but now instead of group difference, with time differences. Go back to the notes to refresh the use of dummy variable and the testing procedure used therein.

4. Two other methods of using year dummies:

   (a) We interact a time dummy with a key explanatory variable to see if the effect of that variable has changed over the chosen time. (Example 13.2)

   (b) We interact a time dummy with all independent variables. The key difference in this strategy with the earlier one is that in this one we suspect y to be significantly different before and after the time dummy conditional on all x's. In the earlier strategy we think y is significantly different before and after the time dummy conditional on only one x. (Krueger, 1993)

13.1.1 Chow Test for Structural Change across Time

1. Chow test: is for two groups. It could be one group over two different time periods or two groups in the same time period.

   (a) Method 1: One form of the test obtains the sum of squared residuals from the pooled estimation as the restricted SSR. The unrestricted SSR is the sum of the SSRs for the two separately estimated time periods.(Section 7.4)

   (b) Method 2: Interacting each x with a dummy variable for one of the two years, along with having these dummies in the regression and testing joint significance of the year dummies and interaction terms. (Example 13.2)

2. Chow test for multiple time periods: interact all time dummies with one or all the x's and test the joint significance of the interaction terms.

   (a) Alternatively: estimate SSR_r by pooling and allowing for different time intercepts i.e. including time dummies. Then obtain SSR on a regression with $t = 1$ and call it SSR_1. Obtain SSR_2 for $t = 2$ and so on. The SSR_{ur} is just the sum of these individual SSR's.

   (b) If there are k explanatory variables (not including the intercept or the time dummies) with T time periods, then we are testing (T-1)k restrictions, and there are $T(1 + k)$ parameters estimated in the unrestricted model.

   (c) The F-stat is the usual

$$F = \frac{(SSR_r - SSR_{ur})/SSR_{ur}}{(n - (T(1 + k)))/(T - 1)k}$$

   (d) There is no heterskedasticity-robust version of this test.

13.2 Policy Analysis with Pooled Cross Sections

1. Effect of a Garbage Incinerator's Location on Housing Prices: Consider the study by Kiel and McClain (1995) who study the effect of house prices before and after the construction of a garbage dump in the area. The claim was that after construction of the garbage disposal site, the house prices around the site reduced drastically.

   This was in North Andover and the garbage disposal site was proposed in 1978, the construction started in 1981 and operations began in 1985.

   They use house prices for houses sold in 1978 and those that were sold in 1981, when the news for construction could have started affecting house prices.

2. A simple regression of real house prices on an index of nearness - defined as 1 if the house is located within a 3 mile radius and zero otherwise for the year 1981 yield the following result:

```
> summary(lm(rprice ~ nearinc, data = kielmc, subset = (y81 == 1)))

Call:
lm(formula = rprice ~ nearinc, data = kielmc, subset = (y81 ==
    1))

Residuals:
   Min     1Q Median     3Q    Max
-60678 -19832  -2997  21139 136754

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   101308       3093  32.754  < 2e-16 ***
nearinc       -30688       5828  -5.266 5.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31240 on 140 degrees of freedom
Multiple R-squared:  0.1653,    Adjusted R-squared:  0.1594
F-statistic: 27.73 on 1 and 140 DF,  p-value: 5.139e-07
```

3. Interpretation: the intercept is the average selling price for homes not near the incinerator, and the coefficient on *nearinc* is the difference in the average selling price between homes near the incinerator and those that are not.

   (a) The partial effect of nearness is a significant 30,688$ less. However, this does not confirm our hypothesis that the garbage disposal caused this difference.

   (b) If we run the same regression for the year 1978, we get the distance dummy to be significant, but the price difference is $18,824. This implies that the area already had already houses with lower prices.

   The problem deriving causality is this: Did the garbage disposal depress prices further?

```
> summary(lm(rprice ~ nearinc, data = kielmc, subset = (year == 1978)))

Call:
lm(formula = rprice ~ nearinc, data = kielmc, subset = (year ==
    1978))

Residuals:
   Min     1Q Median     3Q    Max
-56517 -16605  -3193   8683 236307

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    82517       2654  31.094  < 2e-16 ***
nearinc       -18824       4745  -3.968 0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29430 on 177 degrees of freedom
Multiple R-squared:  0.08167,   Adjusted R-squared:  0.07648
F-statistic: 15.74 on 1 and 177 DF,  p-value: 0.0001054
```

4. Testing whether difference was increased by garbage site or not by obtaining the 'Difference-in-difference' estimator:

$$\hat{\delta_1} = -30688.27 - (-18824.37)$$

This can also be obtained as the difference over time in the average difference of housing prices in the two locations, near (nr) and far (fr):

$$\widehat{\delta_1} = (\overline{rprice_{81,nr}} - \overline{rprice_{81,fr}}) - (\overline{rprice_{78,nr}} - \overline{rprice_{78,fr}})$$

The standard error of $\widehat{\delta_1}$ can be obtained by the following regression with data pooled over both years

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + u$$

where δ_0 captures average rprice change from 1978 to 1981, β_1 captures the location effect that has nothing to do with the garbage disposal site and δ_1 captures decline in housing value due to the new incinerator; **provided we assume that the house prices in near and far did not change due to any other reasons**.

```
> summary(lm(rprice ~ y81 + nearinc + I(y81 * nearinc), data = kielmc)
+ )

Call:
lm(formula = rprice ~ y81 + nearinc + I(y81 * nearinc), data = kielmc)

Residuals:
   Min     1Q Median     3Q    Max
-60678 -17693  -3031  12483 236307

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         82517       2727  30.260  < 2e-16 ***
y81                 18790       4050   4.640 5.12e-06 ***
nearinc            -18824       4875  -3.861 0.000137 ***
I(y81 * nearinc)   -11864       7457  -1.591 0.112595
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30240 on 317 degrees of freedom
Multiple R-squared:  0.1739,    Adjusted R-squared:  0.1661
F-statistic: 22.25 on 3 and 317 DF,  p-value: 4.224e-13
```

5. Kiel and McClain (1995) included a bunch of controls in order to curb the standard errors and to check whether there are other factors which cause a difference in prices between the two houses.

```
> summary(lm(rprice ~ y81 + nearinc + I(y81 * nearinc) + age + agesq + area +
 land+ intst + rooms + baths, data = kielmc))

Call:
lm(formula = rprice ~ y81 + nearinc + I(y81 * nearinc) + age +
    agesq + area + land + intst + rooms + baths, data = kielmc)

Residuals:
   Min     1Q Median     3Q    Max
-76721  -8885   -252   8433 136649

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.381e+04  1.117e+04   1.237  0.21720
y81               1.393e+04  2.799e+03   4.977 1.07e-06 ***
nearinc           3.780e+03  4.453e+03   0.849  0.39661
I(y81 * nearinc) -1.418e+04  4.987e+03  -2.843  0.00477 **
age              -7.395e+02  1.311e+02  -5.639 3.85e-08 ***
agesq             3.453e+00  8.128e-01   4.248 2.86e-05 ***
area              1.809e+01  2.306e+00   7.843 7.16e-14 ***
land              1.414e-01  3.108e-02   4.551 7.69e-06 ***
intst            -5.386e-01  1.963e-01  -2.743  0.00643 **
rooms             3.304e+03  1.661e+03   1.989  0.04758 *
baths             6.977e+03  2.581e+03   2.703  0.00725 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19620 on 310 degrees of freedom
Multiple R-squared:   0.66,     Adjusted R-squared:  0.6491
F-statistic: 60.19 on 10 and 310 DF,  p-value: < 2.2e-16
```

6. Controlling for age, age2, intst, land, area, rooms, baths to obtain a more precise estimate of δ_1 witha a t-stat of about -2.84 now.

7. Functional form convention: log(price) [or log(rprice)] in the analysis in order to get an approximate percentage effect of the garbage dumpster.

8. Natural or Quasi Experimental data: A natural experiment occurs when some exogenous event—often a change in government policy— changes the environment in which individuals, families, firms, or cities operate.

   (a) A natural experiment always has a control group, which is not affected by the policy change, and a treatment group, which is thought to be affected by the policy change.

   (b) If you were to design an actual experiment, the control and treatment group are chosen apriori, but in a natural experiment we hope that such groups automatically arise from a policy change.

   (c) Natural experiment requires at least 2 years of data - one before and another after the experiment.

   (d) This leads us to have 4 categories: (a) Control (C) before, (b) treatment(T) before, (c) control after and (d) treatment after.

   (e) $dT = 1$ for treatment and $d2 = 1$ for after time period.

9. We run the following regression:

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + \text{other factors}$$

where,

$$\hat{\delta}_1 = (\overline{y_{2,T}} - \overline{y_{1,T}}) - (\overline{y_{2,C}} - \overline{y_{1,C}})$$

The first term, $(\overline{y_{2,T}} - \overline{y_{1,T}})$, is the difference in means over time for the treated group, shows the effect of policy assuming there are no other changes between the two time periods.

- ▷ If there are other changes then subtracting $(\overline{y_{2,C}} - \overline{y_{1,C}})$ from $(\overline{y_{2,T}} - \overline{y_{1,T}})$ would get rid of that to get us only policy induced change.

TABLE 13.3 Illustration of the Difference-in-Differences Estimator

	Before	After	After – Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment—Control	β_1	$\beta_1 + \delta_1$	δ_1

10. The parameter $\hat{\delta}_1$ can be given an interpretation as an average treatment effect (ATE), where the "treatment" is being in group T in the second time period.

13.2.1 Adding an Additional Control Group

1. Difference in difference assumes that the variable y would trend at the same rate in absence of the policy intervention i.e. y_t would follow a stable time path.

2. Myer, Viscusi and Durbin (1995) studied the effect of an increase in paid compensation for injured workers in Kansas.

   - The absolute amount of compensation implied that low wage workers are already getting all their wage back in the case of an injury leave.
   - But for the high wage workers, only a part of their income is given as compensation in the event of an injury.
   - When the cap on injury compensation was increased, the low wage workers were not affected, but the high wage workers were. This gave them more incentive to not return to work and take the compensated leave.
   - This gave rise to a natural control and treatment group and a difference in difference estimate.

3. Adding a control: for middle income groups who, like their low income counterpart are not affected by the policy is just an additional control. They should also be not affected by their counterparts.

   This is known as the *parallel trends assumption*. Violation of this assumption implies that there are other things which affect the difference between control and treated group before and after the experiment.

4. Another control: We could also examine the control and treatment group in two states. We expect no difference to exist in the control and treatment between the two states. Let the other state be Missouri and let dL represent dummy variable for low income families and dB represent state of Kansas, d2 represent time after policy change in Kansas. Now we estimate the following

$$\begin{aligned} y =& \beta_0 + \beta_1 dL + \beta_2 dB + \beta_3 dL \cdot dB + \delta_0 d2 + \delta_1 d2 \cdot dL \\ &+ \delta_2 d2 \cdot dB + \delta_3 d2 \cdot dL \cdot dB + u \end{aligned}$$

   where δ_3 is the policy effect such that d2 = 1, dL = 1 and dB = 1 is the difference in difference estimator.

This can be interpreted as:

$$\begin{aligned}\hat{\delta_3} =&[(\overline{y_{2,L,B}} - \overline{y_{1,L,B}}) - (\overline{y_{2,M,B}} - \overline{y_{1,L,B}})] \\ &- [(\overline{y_{2,L,A}} - \overline{y_{1,L,A}}) - (\overline{y_{2,M,A}} - \overline{y_{1,L,A}})] \\ =&\widehat{\delta_{DD,B}} - \widehat{\delta_{DD,A}} = \widehat{\delta_{DDD}}\end{aligned}$$

Checking parallel trend: If health trends between the L and M groups do not differ in state A, and there were no other intervention that would affect health outcomes, then $\widehat{\delta_{DD,A}} = 0$

The estimator $\hat{\delta_3}$ is called the difference-in-difference estimator. You could estimate this by running OLS with the model specified above with heteroskedasticity robust standard errors.

13.2.2 A General Framework

1. Generalizing the idea of treatment and control: We can create a very general framework for policy analysis by allowing a general pattern of interventions, where some units are never "treated" and others may be treated in different time periods.

2. General representation: i individual, who belongs to a group (g) undergoing policy change at time t. Policy intervention is a dummy x_{gt} for the group at time t. The estimable model is:

$$y_{igt} = \lambda_t + \alpha_g + \beta x_{gt} + \mathbf{z_{igt}}\boldsymbol{\gamma} + u_{igt}$$

λ_t is the **aggregate time effect** - a dummy that captures the parallel trend assumption. This is a factor which controls for anything else that changes across the two time periods, which when not accounted for leads to spurious conclusions on policy.

α_g is the **aggregate group effect** - again a state dummy i.e. it accounts for systematic differences in groups that are constant across time.

The variables $\mathbf{z_{igt}}$ can include measured variables that change only at the (g, t) level but also, as the i subscript indicates, individual-specific covariates.

The **estimation follows pooled OLS** with adequate adjustments for heteroskedasticity robust standard error.

3. Group Specific Linear Time Trend: Notice that variables which change over time are included in (a) x_{gt} - our policy intervention, (b) z_{igt} - our individual level covariates and (c) λ_t - everything else which changes 'systematically' overtime in y_{igt}.

   The λ_t forces our groups (g) to change over time through λ_t. But what if we know the way our groups differ overtime. Further, we want to say that the trend is linear. We could include a group-specific linear trend if we have at least $T \geq 3$ and run the following model:

$$y_{igt} = \lambda_t + \alpha_g + \psi_g t + \beta x_{gt} + \mathbf{z_{igt}}\boldsymbol{\gamma} + u_{igt}$$

   A quadratic group time trend is

$$y_{igt} = \lambda_t + \alpha_g + \psi_g t^2 + \beta x_{gt} + \mathbf{z_{igt}}\boldsymbol{\gamma} + u_{igt}$$

4. Problem of identifying β i.e. the effect of policy is that we need more variation in x_{gt} over both g and t to include complicated functional forms of time and group trends. A more general specification which looses the β would then be

$$y_{igt} = \theta_{gt} + \beta x_{gt} + \mathbf{z_{igt}}\boldsymbol{\gamma} + u_{igt}$$

where θ_{gt} is a different intercept for each (g, t) pair.

5. What if we are interested in γ and not β?

Then we could estimate the following models:

$$\begin{aligned} y_{igt} =& \theta_{gt} + \beta x_{gt} + \mathbf{z_{igt}}\boldsymbol{\gamma} + u_{igt} \\ y_{igt} =& \beta_1 dG + \beta_2 dT + \beta_3 dG \cdot dT + \beta x_{gt} + \mathbf{z_{igt}}\boldsymbol{\gamma} + u_{igt} \end{aligned}$$

6. We could extend this set-up to multiple policies which were implemented together to examine their individual and joint impact

$$y_{igt} = \theta_{gt} + \mathbf{x_{gt}}\boldsymbol{\beta} + \mathbf{z_{igt}}\boldsymbol{\gamma} + u_{igt}$$

where $\mathbf{x_{gt}}$ is a vector of two policies which were implemented together.

The policy variable need not be binary but could have a continuous representation.

The policy could also have lagged effect:

$$y_{igt} = \theta_{gt} + \mathbf{x_{gt}}\boldsymbol{\beta} + \mathbf{x_{gt-1}}\boldsymbol{\beta_1} + \mathbf{x_{gt-2}}\boldsymbol{\beta_2} + \mathbf{z_{igt}}\boldsymbol{\gamma} + u_{igt}$$

where the policy was implemented at $t-2$ and β measures its effect only 2 period after the implementation took place.