USC, Fall 2016, Economics 513

Lecture 11: Differences-in-differences

Estimating the effect of a policy change

Workers' Comp is a program run by the states that compensates workers injured on the job for lost income and medical expenses. Most claims are temporary total disability claims that provide income to workers who are expected to make a full recovery.

A key policy question is whether Worker's Comp increases the length of time that injured workers stay out off work.

In 1980 and 1982, respectively, the states of Kentucky and Michigan changed the Workers' Comp benefits, in particular they increased the maximum weekly benefit. The increase affected high income but not low income workers (see figure). We want to use these changes to estimate the effect of Workers' Comp on disability spells. This is done in Meyer, Viscusi, and Durbin, "Workers Compensation and Injury Duration: Evidence from a Natural Experiment", AER (1995).

An obvious comparison that is informative on the effect is the average length of disability spells before and after the change. Because only high income workers are affected we should look at this group. Do you think that this comparison is an unbiased estimate of the effect of the benefit extension?

Another comparison is the difference in average disability spells for high income and low income workers after the benefit extension was introduced. Do you think that this comparison is an unbiased estimate of the effect of the benefit extension?

How can we obtain an estimate of the effect of the benefits extension that is free of these biases?

Counterfactuals and causality

In the example we have two time periods: period 0 before the policy change and period 1 after the policy change. We also have individuals affected by the intervention/treatment, $d = 1$ and individuals not affected by the intervention/treatment, $d = 0$.

Y is the outcome of interest, in the example the disability duration. We omit subscript i. Notation

$$Y_{dt} = \text{individual outcome in period } t \text{ if treatment status is } d$$

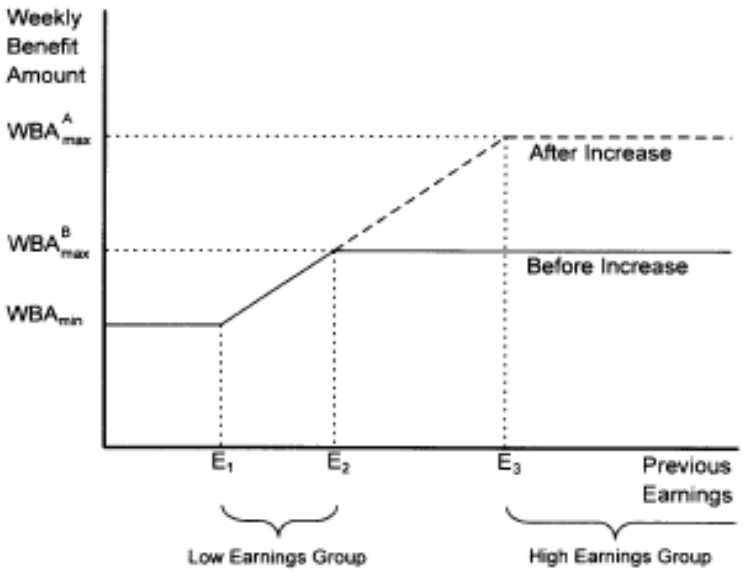


FIGURE 1. TEMPORARY TOTAL BENEFIT SCHEDULE BEFORE AND AFTER AN INCREASE IN THE MAXIMUM WEEKLY BENEFIT

Note that in the example (and that is the case we consider) $d = 0$ in period 0 and $d = 0, 1$ in period 1.

We observe Y_0, Y_1, D with D the assigned/observed treatment status in period 1 (no need for subscript on D)

$$Y_0 = Y_{00}$$

and

$$Y_1 = DY_{11} + (1 - D)Y_{01}$$

The causal effect of the intervention on an individual is the individual treatment effect

$$Y_{11} - Y_{01}$$

This effect cannot be observed, because we observe either Y_{11} or Y_{01} but never both. This is the fundamental problem of causal inference. Note that we do not assume that the treatment effect is constant, i.e. each individual has his/her own treatment effect. We call Y_{11}, Y_{01} the potential outcomes. One of the potential outcomes is a counterfactual and this is the problem.

A solution is to be less ambitious and settle for the average of $Y_{11} - Y_{01}$, the Average Treatment Effect (ATE)

$$\text{ATE} = \text{E}(Y_{11} - Y_{01})$$

We can obtain the ATE if we randomly assign the intervention in period 1, i.e. whether a unit is treated is determined by a coin toss or some similar random experiment. Random assignment implies

$$Y_{00}, Y_{01}, Y_{11} \perp D$$

where $\perp$ denotes stochastic independence. Therefore

$$\text{ATE} = \text{E}(Y_{11} - Y_{01}) = \text{E}(Y_{11}) - \text{E}(Y_{01}) = \text{E}(Y_{11}|D = 1) - \text{E}(Y_{01}|D = 0) =$$

$$\text{E}(Y_1|D = 1) - \text{E}(Y_1|D = 0)$$

The final difference is just the difference of the average outcome of the treated and non-treated, also called the controls. This difference can be estimated using the sample averages.

We obtain the same estimate if we consider the regression model

$$Y_1 = \alpha + \beta D + \varepsilon$$

and estimate β by OLS, i.e. if $n_1 = \sum_{i=1}^{n} D_i$ and $n_0 = \sum_{i=1}^{n}(1 - D_i)$ are the number of the treated and controls, respectively, then

$$\hat{\beta} = \frac{1}{n_1}\sum_{i=1}^{n} Y_i D_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - D_i)$$

Under random assignment $D \perp \varepsilon$ so that the OLS estimator is unbiased for $\beta = \text{E}(Y_{11} - Y_{01})$.

The sampling variance of the estimator depends on the assumption on the variance of ε. If the error is homoskedastic, i.e.

$$\text{Var}(\varepsilon|D = 1) = \text{Var}(\varepsilon|D = 0) = \sigma^2$$

then

$$\text{Var}(\hat{\beta}) = \frac{n_0 + n_1}{n_0 n_1}\sigma^2$$

If the error is heteroskedastic, i.e. the outcomes under treatment and non-treatment have different variances,

$$\text{Var}(\varepsilon|D = 1) = \sigma_1^2 \qquad \text{Var}(\varepsilon|D = 0) = \sigma_0^2$$

then

$$\text{Var}(\hat{\beta}) = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}$$

Note that we get an estimate of the ATE even though the regression model suggests that the treatment effect is constant. We use this type of regression models also in our problem.

Differences-in-differences

In the Workers' Comp example there is no random assignment. This can easily be seen by comparing the average disability spell before the benefit extension, i.e.

$$\mathrm{E}(Y_0|D=1) - \mathrm{E}(Y_0|D=0) = \mathrm{E}(Y_{00}|D=1) - \mathrm{E}(Y_{00}|D=0)$$

Meyer et all find that these averages are 11.18 and 6.27 (weeks).

We assume instead

$$Y_{01} - Y_{00} \perp D$$

i.e. selection for treatment can depend on Y_{00} but is independent of the change in the non-treated outcome. In the example, this implies that the change in the average disability spell would have been the same for low and high earners if the benefits increase had not been implemented.

Under this assumption, because

$$Y_1 - Y_0 = DY_{11} + (1-D)Y_{01} - Y_{00} = Y_{01} - Y_{00} + D(Y_{11} - Y_{01})$$

we have

$$\mathrm{E}(Y_1 - Y_0|D=1) - \mathrm{E}(Y_1 - Y_0|D=0) =$$
$$\mathrm{E}(Y_{01} - Y_{00}|D=1) + \mathrm{E}(Y_{11} - Y_{01}|D=1) - \mathrm{E}(Y_{01} - Y_{00}|D=0) =$$
$$\mathrm{E}(Y_{11} - Y_{01}|D=1)$$

Rewriting the left hand side we have

$$\mathrm{E}(Y_1|D=1) - \mathrm{E}(Y_1|D=0) - (\mathrm{E}(Y_0|D=1) - \mathrm{E}(Y_0|D=0)) = \mathrm{E}(Y_{11} - Y_{01}|D=1)$$

The left hand side is the difference in average outcome between treated and controls in period 1 (a biased estimate of the treatment effect) minus the difference in average outcome between treated and controls in period 0 that corrects for the selective application of the treatment. For obvious reasons the left hand side is called a difference-in-differences (dif-in-dif).

Under the assumption made the dif-in-dif estimator estimates the right-hand side that is not the ATE but the Average Treatment Effect on the Treated individuals (ATET). Often this is a more relevant average effect.

If T is a dummy that indicates the time period, then we can consider the linear model

$$Y = \alpha + \gamma_1 T + \gamma_2 D + \beta T \cdot D + \varepsilon$$

Note that I omit the subscript t on Y, leaving it implicit that we observe Y_1 if $T = 1$ and Y_0 if $T = 0$. The OLS estimate of β is

$$\hat{\beta} = \frac{1}{n_{11}} \sum_{i=1}^{n} Y_i D_i T_i - \frac{1}{n_{01}} \sum_{i=1}^{n} Y_i (1-D_i) T_i - \frac{1}{n_{10}} \sum_{i=1}^{n} Y_i D_i (1-T_i) + \frac{1}{n_{00}} \sum_{i=1}^{n} Y_i (1-D_i)(1-T_i)$$

with

$$n_{11} = \sum_{i=1}^{n} D_i T_i$$

and n_{01}, n_{10}, n_{00} defined similarly.

If we assume a homoskedastic error

$$\text{Var}(\varepsilon | D = d, T = t) = \sigma^2$$

for $d = 0, 1$ and $t = 0, 1$, then

$$\text{Var}(\hat{\beta}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}} \right) \sigma^2$$

If the error is heteroskedastic

$$\text{Var}(\varepsilon | D = d, T = t) = \sigma^2_{dt}$$

then

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2_{11}}{n_{11}} + \frac{\sigma^2_{10}}{n_{10}} + \frac{\sigma^2_{01}}{n_{01}} + \frac{\sigma^2_{00}}{n_{00}}$$

Note that the dif-in-dif estimator can be computed with panel data, but also with two cross-sections from the same population. We call the latter a repeated cross-section. If we use a repeated cross-section we must be able to determine the future treatment status of the individuals observed in period 0. This is not an issue with panel data.

The Workers' Comp data are a repeated cross-section. Note that due to income changes we may not be able to accurately determine the future treatment status in period 0.

	Period 0	Period 1	Dif
Treated	6.27	7.04	.77
Controls	11.18	12.89	1.72
Dif	4.91	5.86	.95

The OLS estimate is

$$\widehat{\text{duration}}_i = \underset{(0.52)}{6.27} + \underset{(0.81)}{4.91 D_i} + \underset{(0.76)}{0.77 T_i} + \underset{(1.17)}{0.95 D_i \cdot T_i}$$

The robust standard error is 1.28.

Dif-in-dif-in-dif

In the Workers' Comp example we compared high and low earners in two states that implemented in increase in benefits for high earners. The key assumption was that the trend in disability durations for the high and low earners would have been the same if the benefits increase for the high earners had not been implemented.

If there is doubt that this assumption holds, i.e. the trend for high earners would have been different anyway, one could look for a second control group. For example a second control group would be high earners in a comparable state.

The assumption now is that the difference of the changes from periods 0 to 1 of the non-treated outcome for the treated and controls (that is not 0), is the same in the treatment and control states.

Define $S = 1$ for the treatment state $S = 0$ for the control state and consider the regression

$$Y = \alpha + \gamma_1 T + \gamma_2 D + \gamma_3 S + \gamma_4 T \cdot D + \gamma_5 D \cdot S + \gamma_6 T \cdot S + \beta T \cdot D \cdot S + \varepsilon$$

The OLS estimator of β is (the subscripts are in the order d, s, t)

$$\hat{\beta} = \frac{1}{n_{111}} \sum_{i=1}^{n} Y_i D_i T_i S_i - \frac{1}{n_{110}} \sum_{i=1}^{n} Y_i D_i S_i (1 - T_i) -$$

$$\left(\frac{1}{n_{011}} \sum_{i=1}^{n} Y_i (1 - D_i) S_i T_i - \frac{1}{n_{010}} \sum_{i=1}^{n} Y_i (1 - D_i) S_i (1 - T_i) \right) -$$

$$\left(\frac{1}{n_{101}} \sum_{i=1}^{n} Y_i D_i (1 - S_i) T_i - \frac{1}{n_{100}} \sum_{i=1}^{n} Y_i D_i (1 - S_i)(1 - T_i) \right)$$

$$+ \left(\frac{1}{n_{011}} \sum_{i=1}^{n} Y_i (1 - D_i)(1 - S_i) T_i - \frac{1}{n_{010}} \sum_{i=1}^{n} Y_i (1 - D_i)(1 - S_i)(1 - T_i) \right)$$

For obvious reasons this is called the dif-in-dif-in-dif estimator.

Conditional dif-in-dif

Key assumption for dif-in-dif

$$Y_{01} - Y_{00} \perp D$$

i.e. selection for treatment independent of the change in the non-treated outcome or the treated would have had the same trend as the controls if they had

not been treated.

This assumption does not hold if the change in average outcomes are as in the figure.

This situation is not uncommon. Often changes are made when outcomes are relatively (compared to some other group) bad. Examples: replacing management, firing coaches. The dip in outcome just before the intervention biases the dif-in-dif estimate of the effect of the intervention upwards. It may seem that the intervention is effective, but this is a temporary effect.

To analyze this (and suggest a solution) we take a specific model for the non-treated outcome

$$Y_{01} = \alpha Y_{00} + \eta$$

with η and Y_{00} independent. This implies

$$Y_{01} - Y_{00} = (\alpha - 1)Y_{00} + \eta$$

The selection for treatment is

$$D = I(Y_{00} \leq c)$$

i.e. selection occurs if the non-treated outcome is below a threshold c.

This implies

$$\begin{aligned} \mathrm{E}(Y_{01} - Y_{00}|D = 1) &= (\alpha - 1)\mathrm{E}(Y_{00}|Y_{00} \leq c) \\ \mathrm{E}(Y_{01} - Y_{00}|D = 0) &= (\alpha - 1)\mathrm{E}(Y_{00}|Y_{00} > c) \end{aligned}$$

so that

$$\mathrm{E}(Y_{01} - Y_{00}|D = 1) \neq \mathrm{E}(Y_{01} - Y_{00}|D = 0)$$

However

$$\mathrm{E}(Y_{01} - Y_{00}|Y_{00}, D = 1) = (\alpha - 1)Y_{00} = \mathrm{E}(Y_{01} - Y_{00}|Y_{00}, D = 0)$$

Note also that $Y_{00} = Y_0$ and is therefore observed. We have in this case

$$Y_{01} - Y_{00} \perp D|Y_0$$

In general we can make the assumption

$$Y_{01} - Y_{00} \perp D|X$$

with X a vector of observable variables.

Because

$$Y_1 - Y_0 = (Y_{01} - Y_{00}) + D(Y_{11} - Y_{01})$$

we have

$$\mathrm{E}(Y_1 - Y_0|D = 1, X) = \mathrm{E}(Y_{01} - Y_{00}|D = 1, X) + \mathrm{E}(Y_{11} - Y_{01}|D = 1, X)$$

$$\mathrm{E}(Y_1 - Y_0|D = 0, X) = \mathrm{E}(Y_{01} - Y_{00}|D = 0, X)$$

Subtract to obtain

$$\mathrm{E}(Y_1 - Y_0|D = 1, X) - \mathrm{E}(Y_1 - Y_0|D = 0, X) = \mathrm{E}(Y_{11} - Y_{01}|D = 1, X)$$

The right hand side is the ATET given X. To find the ATET

$$\mathrm{E}(Y_{11} - Y_{01}|D = 1) = \mathrm{E}_X[\mathrm{E}(Y_{11} - Y_{01}|D = 1, X)|D = 1]$$

To implement this we take the steps

1. Regress Y_0 on X for $D = 0$ and $D = 1$ and regress Y_1 on X for $D = 0$ and $D = 1$, i.e. four separate regressions. Call the estimated regression models $\hat{\mu}_{dt}(x)$.

2. The estimate is

$$\hat{\beta} = \frac{1}{n_1} \sum_{i=1}^{n} D_i(\hat{\mu}_{11}(X_i) - \hat{\mu}_{01}(X_i) - (\hat{\mu}_{10}(X_i) - \hat{\mu}_{00}(X_i)))$$

A simpler but less flexible approach is to estimate

$$Y = \alpha + \gamma_1 T + \gamma_2 D + \gamma_3' X + \beta D \cdot T + \varepsilon$$

Panel data

All estimators (except if X contains lagged outcomes) until now can be estimated with repeated cross-section data. With panel data we can use lagged outcomes in conditional dif-in-dif. There are other differences as well.

Consider the model

$$Y_{it} = \alpha + \gamma_1 T_t + \gamma_2 X_{it} + \beta T_t D_i + \delta D_i + \eta_i + \varepsilon_{it}$$

with η_i an individual effect. We allow D_i to be correlated with η_i (but not with ε_{it}).

With panel data we can consider the FD estimator

$$\Delta Y_{it} = \gamma_1 + \gamma_2 \Delta X_{it} + \beta D_i + \Delta \varepsilon_{it}$$

Without X we get the same dif-in-dif estimator. However the sampling variance of this estimator is now

$$\mathrm{Var}(\hat{\beta}) = \frac{\mathrm{Var}(Y_1 - Y_0|D = 1)}{n_1} + \frac{\mathrm{Var}(Y_1 - Y_0|D = 0)}{n_0}$$

and this is smaller than the repeated cross-section variance if the variation in η_i is large.

Application to minimum wage data

These data are from David Card and Alan B. Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," American Economic Review, Volume 84, no. 4 (September 1994), pp. 774-775. Card and Krueger study the effect of a change in the minimum wage in New Jersey. They collected data on employment at fast food restaurants in New Jersey and Pennsylvania before and after the change in the minimum wage. The data are converted to fulltime employment, counting parttime employees as half. The group sizes are $n_{\mathrm{P}A} = 76$ and $n_{\mathrm{N}J} = 315$.

Table 1 presents the averages in the four groups, and the single and double differences.

Table 1: Card-Krueger Data: Means by Group and Time Period

		Period		Dif
		before	after	
Group	PA	20.01	17.52	-2.49
	NJ	17.05	17.50	0.45
	Dif	-2.97	-0.02	2.94

The dif-in-dif estimate is 2.94 with a robust standard error of 1.32.

The OLS regression leads to

$$\begin{array}{rcccc} \widehat{\Delta \text{employment}}_i & = & -2.49 & + & 2.94 \cdot D_i \\ & & (1.01) & & (1.12) \end{array}$$

again with a slightly smaller standard error than we get under the robust variance calculation.

We can consider a number of conditional dif-in-dif estimators that make different assumptions on the selection for treatment

> Assumption 1: Without the minimum wage increase the employment growth in NJ would have been the same as that in PA in establishments of the same size in year 0.

Assumption 2: Without the minimum wage increase the employment growth in NJ would have been the same as that in PA in establishments in the same chain (Burger King, KFC, Roy Rogers or Wendy's).

Under assumption 1 we regress the change in employment on the state dummy and employment in year 0

$$\begin{array}{rcccc} \widehat{\Delta\text{employment}}_i & = & 7.77 & + & 1.42 \cdot D_i - \quad .51 \cdot \text{employment}_{0i} \\ & & (1.15) & & (.94) \qquad\qquad (.04) \end{array}$$

Under assumption 2 we regress the change in employment on the state dummy and chain dummies

$$\begin{array}{rcccccccc} \widehat{\Delta\text{employment}}_i & = & -2.11 & + & 2.97 \cdot D_i + & .060 \cdot \text{BK}_i + & .55 \cdot \text{KFC}_i - & 2.20 \cdot \text{ROYS}_i \\ & & (1.47) & & (1.12) & (1.39) & (1.55) & (1.50) \end{array}$$

Clustering and the variance of OLS estimators

The dif-in-dif estimator compares groups of units. Recently it has been recognized that the usual standard errors may underestimate the true sampling variation in the OLS estimates if the data is clustered and variables are defined at the cluster level.

We consider the case that the observations $i = 1, \ldots, n$ are from G groups or clusters. Examples are classrooms, cities, states etc. Let $n_g, g = 1, \ldots, G$ be the number of observations in cluster g. We distinguish between two cases: (i) G is large (and conceivably $G \to \infty$) and n_g is small (and fixed or bounded), (ii) G is small (and fixed or bounded) and n_g is large (and conceivably $n_g \to \infty$).

It is also important to distinguish between independent variables that are group specific, i.e. are constant in a cluster, and variables that vary within a cluster. Examples of the former are treatment dummies, state dummies, group averages, population size, and example of the latter are individual age, level of education.

Consider the linear regression model with x_g group specific and w_i individual specific variables

$$y_{ig} = x_g'\beta + w_{ig}'\gamma + \eta_g + \varepsilon_{ig} \qquad (1)$$

Note that we order the data by cluster. Just as in panel data we have an error component error term that is the sum of a group error and an individual specific error, i.e. η_g captures omitted cluster specific variables and ε_{ig} captures omitted individual specific variables. This implies that there is correlation between the errors in a cluster (but not between the errors in different clusters).

We have the same model if we define dummy variables

$$\begin{array}{rcll} d_{ig} & = & 1 & \text{if } i \text{ is in group } g \\ & = & 0 & \text{otherwise} \end{array}$$

so that (note that we can omit the subscript g)

$$y_i = \sum_{g=1}^{G} d_{ig} x_g' \beta + w_i' \gamma + \sum_{g=1}^{G} d_{ig} \eta_g + \varepsilon_i \tag{2}$$

or if we define the $n \times G$ matrix

$$Z = \begin{pmatrix} d_{11} & \cdots & d_{1G} \\ \vdots & & \vdots \\ d_{n1} & \cdots & d_{nG} \end{pmatrix}$$

the $G \times K$ matrix

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_G' \end{pmatrix}$$

and the $G \times 1$ vector

$$\eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_G \end{pmatrix}$$

so that in matrix notation

$$y = ZX\beta + W\gamma + Z\eta + \varepsilon \tag{3}$$

We have

$$Z'Z = \begin{pmatrix} n_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & n_G \end{pmatrix}$$

and ZZ' is the $n \times n$ matrix in which the first row (and column) has component 1 if the corresponding unit is in group 1 etc. This formulation is useful if $n_g \to \infty$.

Large G

We consider the OLS estimator first for G large with $G \to \infty$ and n_g is small and fixed. From (1) we see that this is the same as having (unbalanced) panel data with g the individual and i the time period. We can use pooled OLS to estimate the regression coefficients if

$$\mathrm{E}[\varepsilon_{ig} + \eta_g | x_g, w_{ig}] = 0$$

Note that we can have that ε_{ig} is correlated with w_{ih} for $h \neq g$. If

$$V_g = \begin{pmatrix} x_g & w_{1g} \\ \vdots & \vdots \\ x_g & w_{n_g g} \end{pmatrix} \qquad \theta = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$$

and u_g is the n_g vector of the combined random errors for group g, then we estimate the robust variance matrix

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \left(\sum_{g=1}^{G} V_g' V_g\right)^{-1} \left(\sum_{g=1}^{G} V_g' \hat{u}_g \hat{u}_g' V_g\right) \left(\sum_{g=1}^{G} V_g' V_g\right)^{-1}$$

This estimator allows for arbitrary within cluster correlations between the random errors, i.e. the random error need not be the sum of a cluster and individual error, and it also allows for arbitrary heteroskedasticity.

This estimator is not efficient if the $u_{ig} = \eta_g + \varepsilon_{ig}$ with

$$\mathrm{E}(\eta_g \eta_{g'}) = 0, \quad g \neq g' \qquad \mathrm{E}(\varepsilon_{ig}\varepsilon_{i'g'}) = 0, \quad g \neq g' \text{ or } g = g', i \neq i' \qquad \mathrm{E}(\varepsilon_{ig}\eta_{g'}) = 0$$

In that case we can use a similar transformation as in the RE panel data model

$$H_g = I_g - \frac{\lambda_g}{n_g} \iota_g \iota_g'$$

with I_g the identity matrix of order n_g, ι_g an $n_g \times 1$ vector of 1-s, and

$$\lambda_g = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + n_g \sigma_\eta^2}}$$

and λ_g can be estimated as in RE. Note that with this estimator we need a strict exogeneity assumption

$$\mathrm{E}[\varepsilon_{ig} + \eta_g | x_g, w_{1g}, \ldots, w_{n_g g}] = 0$$

The variance matrix of this RE type estimator can be made robust. See lecture 8.

Fixed G

The case that G is fixed and $n_g \to \infty$ is different. Instead of (2) we consider the more general model

$$y_i = \sum_{g=1}^{G} d_{ig} \delta_g + w_i' \gamma_g + \varepsilon_i$$

with

$$\delta_g = x_g' \beta + \eta_g$$

By choosing η_g appropriately this always holds. If $\gamma_g = \gamma$ we have a model with cluster fixed effects (take y_i and w_i in deviation from their cluster mean and estimate by pooled OLS on the transformed variables), but because n_g is large we can estimate the regression coefficients by cluster and test whether this

equality holds.

This shows that γ_g or the common γ, i.e. the coefficients on the individual specific variables can be estimated consistently and that their sampling variance is not affected by the cluster effects, i.e. can be computed in the usual way.

Now consider a model with only cluster specific variables

$$y_i = \sum_{g=1}^{G} d_{ig} x_g' \beta + \sum_{g=1}^{G} d_{ig} \eta_g + \varepsilon_i$$

and we assume

$$\mathrm{E}(\eta_g + \varepsilon_i | x_g) = 0$$

The pooled OLS estimator is

$$\hat{\beta} = (X'Z'ZX)^{-1} X'Z'y = \left(\sum_{g=1}^{G} n_g x_g x_g' \right)^{-1} \left(\sum_{g=1}^{G} n_g x_g \overline{y}_g \right)$$

with $\overline{y}_g$ the cluster mean of y. Hence if $n_g = L$, i.e. all clusters have equal size this is the OLS estimator if we regress $\overline{y}_g$ on x_g. In this case we have if $n_g \to \infty$ the regression

$$\mu_g = x_g' \beta + \eta_g$$

with μ_g the population mean of y in group g. Because G is fixed we cannot estimate β consistently, but under the assumption OLS is unbiased.

But what about the variance of the OLS estimator? Substitution of the model gives

$$\hat{\beta} - \beta = \left(\sum_{g=1}^{G} n_g x_g x_g' \right)^{-1} \left(\sum_{g=1}^{G} n_g x_g \overline{u}_g \right)$$

with $\overline{u}_g = \eta_g + \overline{\varepsilon}_g$ and $\overline{\varepsilon}_g$ the cluster average. Therefore

$$\mathrm{Var}(\hat{\beta}) = \left(\sum_{g=1}^{G} n_g x_g x_g' \right)^{-1} \left(\sum_{g=1}^{G} n_g^2 \left(\sigma_\eta^2 + \frac{\sigma_\varepsilon^2}{n_g} \right) x_g x_g' \right) \left(\sum_{g=1}^{G} n_g x_g x_g' \right)^{-1}$$

This simplifies if $n_g = L$

$$\mathrm{Var}(\hat{\beta}) = \left(\sigma_\eta^2 + \frac{\sigma_\varepsilon^2}{L} \right) \left(\sum_{g=1}^{G} x_g x_g' \right)^{-1}$$

If we had used the OLS formula in (3) we would have have ignored the correlation between the $u_{ig} = \eta_g + \varepsilon_{ig}$ and taken

$$\mathrm{Var}(\overline{u}_g) = \frac{\sigma_\eta^2 + \sigma_\varepsilon^2}{L}$$

so that

$$\text{Var}(\hat{\beta})_{OLS} = \frac{\sigma_\eta^2 + \sigma_\varepsilon^2}{L}(X'Z'Z'X)^{-1} = \frac{\sigma_\eta^2 + \sigma_\varepsilon^2}{L}\left(\sum_{g=1}^{G} x_g x_g'\right)^{-1}$$

If we define the within cluster correlation of the errors as

$$\rho = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2 + \sigma_\eta^2}$$

then

$$\text{Var}(\hat{\beta}) = (L\rho + (1-\rho))\text{Var}(\hat{\beta})_{OLS}$$

We conclude that the OLS formula that applies if $\rho = 0$ in general underestimates the variance of the OLS estimator. If $\rho = .01$, i.e. hardly any correlation and $L = 5000$ the standard errors are 7 times as large as suggested by the OLS formula.

For the general case in (2) we can use a two step approach. First use OLS to estimate

$$y_i = \sum_{g=1}^{G} d_{ig}\delta_g + w_i'\gamma + \varepsilon_{ig}$$

In the second step estimate β in

$$\hat{\delta}_g = x_g'\beta + \tilde{\eta}_g$$

If n_g is large $\hat{\delta}_g \approx \delta_g$ and $\tilde{\eta}_g \approx \eta_g$. This will give asymptotically correct standard errors.

We can also use (3) directly. Define $V = (ZX\ W)$ so that

$$\text{Var}(\hat{\theta}) = (V'V)^{-1}(\sigma_\varepsilon^2 I_n + \sigma_\eta^2 ZZ')(V'V)^{-1}$$

The correlation coefficient ρ is estimated from the OLS residuals e_i in (3). The sample variance of these residuals $\hat{\sigma}^2$ is an estimate of $\sigma_\varepsilon^2 + \sigma_\eta^2$. Subtract the cluster averages from the residuals, i.e.

$$\tilde{e} = (I_n - Z(Z'Z)^{-1}Z')e$$

The sample variance

$$\tilde{\sigma}^2 = \frac{\tilde{e}'\tilde{e}}{n - G - 1}$$

is an estimator for σ_ε^2 (note that we subtract the number of clusters in the denominator). Now

$$\hat{\rho} = \frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\hat{\sigma}^2}$$

and

$$\hat{\sigma}_\eta^2 = \hat{\rho}\hat{\sigma}^2$$

Empirical illustration

We use again the Angrist-Krueger data with 329,509 observations, 51 clusters (states), and on average 6,461 observations per cluster (but with substantial variation in the cluster size, ranging from 78 to 29,015). We regress log earnings on individual education and average education in the state of residence. The estimated cluster correlation coefficient ρ is very small, approximately 0.0005. The correction factor for OLS is substantial because of the large sample size: $6,461 \times 0.0005 \approx 4$, or a doubling of the standard errors.

The standard errors are computed in three ways. First the conventional OLS standard errors. Second the correct standard errors, and finally the standard errors suggested by the correction factor with equal cluster size (use the average cluster size).

Table 2: Estimates and Standard Errors

	intercept	own education	average state education
estimate	4.2584	0.0656	0.0665
OLS	(0.0542)	(0.0011)	(0.0045)
Correct	(0.1345)	(0.0011)	(0.0110)
Correction factor	(0.1105)	(0.0022)	(0.0091)

We also estimate the coefficient on average education in a regression on the estimated state dummies. This leads to

Table 3: Second Stage Regression

	intercept	average state education
estimate	4.3510	0.0585
s.e.	(0.1606)	(0.0129)

Application to dif-in-dif

We first consider a simpler setting: the effect of a randomly assigned intervention on an outcome y. If d is the treatment dummy we estimate the treatment effect from the regression

$$y_i = \alpha + \beta d_i + \varepsilon_i$$

The data have two clusters: the treated and the control group. So we could consider

$$y_{di} = \alpha + \beta d + \eta_d + \varepsilon_{di}$$

with η_d and ε_{di} uncorrelated/independent and the same for η_0, η_1 and also the ε_{di} uncorrelated/independent. Does this make sense?

In large samples the average outcome for the treated estimates

$$\mathrm{E}(y_{1i}) = \alpha + \beta + \eta_1$$

and the average outcome for the controls estimates

$$\mathrm{E}(y_{0i}) = \alpha + \eta_0$$

The ATE is therefore

$$\beta + \eta_1 - \eta_0$$

The idea seems to be that if we repeat the experiment in a number of populations, each with an η_1, η_0 then the average of these ATE-s over these populations is β if $\mathrm{E}(\eta_1 - \eta_0) = 0$ where the average is over these populations. However, random assignment of treatment in a given population should make the treated and control groups identical, except for the treatment, so that $\eta_1 = \eta_0$. With random assignment of treatment there is no need to consider the group structure of the data.

For dif-in-dif we consider the linear regression (with t the time dummy)

$$y_i = \alpha + \gamma_1 t_i + \gamma_2 d_i + \beta d_i \cdot t_i + \varepsilon_i$$

The data have four clusters: the treated and controls in periods 0 and 1. To reflect this we write the linear regression as

$$y_{dti} = \alpha + \gamma_1 t_i + \gamma_2 d_i + \beta d_i \cdot t_i + \eta_{dt} + \varepsilon_{idt}$$

where the η_{dt} capture the omitted cluster specific variables. In a large sample the group means are equal to the population means so that

$$\mathrm{E}(y_{11i}) - \mathrm{E}(y_{10i}) - \mathrm{E}(y_{01i}) + \mathrm{E}(y_{00i}) = \beta + \eta_{11} - \eta_{10} - \eta_{01} - \eta_{00}$$

Again if we had access to many populations and $\mathrm{E}(\eta_{11} - \eta_{10} - \eta_{01} - \eta_{00}) = 0$ on average over these populations, then the dif-in-dif estimator is unbiased on average. The assumption that justified the dif-in-dif estimator states that $\eta_{11} - \eta_{10} - \eta_{01} - \eta_{00} = 0$ in any population.

If the OLS residual of the dif-in-dif regression is e_i then by the properties of OLS

$$\frac{1}{n}\sum_{i=1}^{n} e_i d_i t_i = 0$$

so that the OLS estimator of η_{dt} is

$$\hat{\eta}_{dt} = \frac{1}{n}\sum_{i=1}^{n} e_i d_i t_i = 0$$

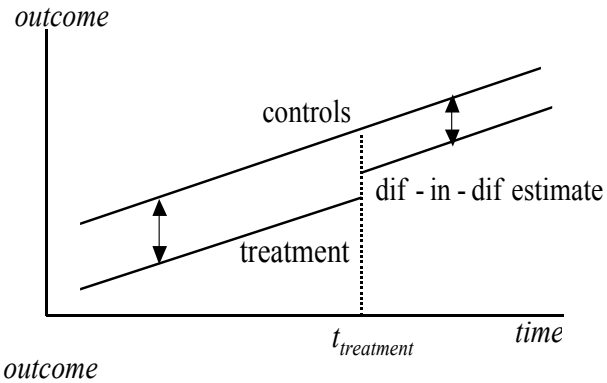
so that in a population we cannot estimate the variance of η_{dt} (over d and t).

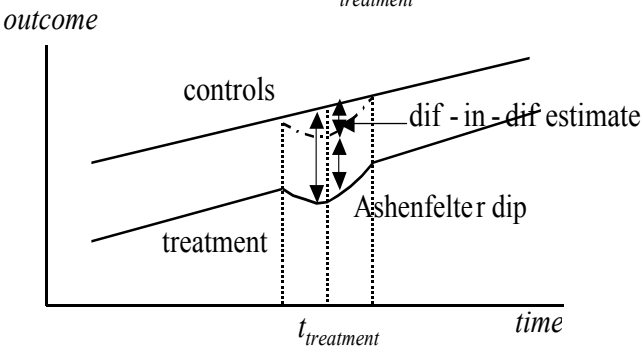
The conclusion is that although the data used for dif-in-dif estimation have a group structure we cannot estimate the cluster variances.

Now consider the conditional dif-in-dif model for the Card-Krueger data

$$\Delta y_i = \beta d_i + \gamma_1 \mathrm{BK}_i + \gamma_2 \mathrm{KFC}_i + \gamma_3 \mathrm{ROYS}_i + \gamma_4 \mathrm{WEND}_i + \eta_{cd} + \varepsilon_i$$

with c is the indicator of the chain (note that we omitted the intercept in the relation). Here a cluster is a chain-treatment/control group combination, i.e. there are 8 clusters. Instead of an intercept we have a complete set of chain dummies. This exactly the model with independent variables that are all cluster specific and we can use the variance formula derived for that case.


outcome
controls
dif - in - dif estimate
treatment
$t_{treatment}$
time


outcome
controls
dif - in - dif estimate
Ashenfelter dip
treatment
$t_{treatment}$
time