Dummy Variables and Difference-in-Differences

EC 320: Introduction to Econometrics

David Hall
University of Oregon

Review and Today's Goal

Last time, we looked at:

- log-log models
- semi-log models
- quadratic models
- interaction variables

These helped us deal with non-linearity and curvature in regression.

Today: We introduce **dummy variables** and **dummy interactions** — tools for modeling qualitative or categorical differences.

Motivating Examples

Consider:

- A tax on feminine hygiene products might affect men and women differently.
- A new alcohol tax in Eugene might affect UO freshmen, sophomores, juniors, and seniors differently.

Question: How do we model these qualitative group differences?

Two Options

Option 1: Run separate regressions for each group.

- This reduces efficiency — each regression discards data.
- We lose precision and cannot test group differences directly.

Option 2: Use **dummy variables** to estimate group differences in one model.

This keeps all data and allows direct hypothesis testing.

Example from HW2

You already estimated the following regression:

$$\text{LifeExpectancy}_i = \beta_0 + \beta_1 D_i + \beta X_i + \varepsilon_i,$$

where:

- $D_i = 1$ if parents are cousins
- $D_i = 0$ otherwise
- X_i represents control variables

This is the simplest dummy-variable model.

Regression Output

TABLE 2—THE EFFECT OF COUSIN MARRIAGE ON OFFSPRING LONGEVITY

	Baseline specifications			Robustness checks				
	Raw	Controls	Parental fixed effects	Flexible controls	Parent longevity	Age heaping	County-decade fixed effects	Same surname
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Life expectancy at age 5								
Parents are first cousins	−2.77	−2.19	−2.21	−2.20	−2.10	−2.09	−2.03	−1.86
	(0.07)	(0.07)	(0.29)	(0.29)	(0.30)	(0.32)	(0.36)	(0.19)
Control mean	63.73	63.73	63.73	63.73	63.73	64.00	64.13	63.86
Observations (thousands)	5,914	5,914	5,914	5,914	5,845	5,186	4,594	12,750
Panel B. Life expectancy at age 20								
Parents are first cousins	−2.40	−1.87	−1.82	−1.83	−1.64	−1.53	−1.89	−1.52
	(0.06)	(0.06)	(0.26)	(0.26)	(0.28)	(0.29)	(0.33)	(0.17)
Control mean	66.75	66.75	66.75	66.75	66.75	66.87	67.06	66.77
Observations (thousands)	5,569	5,569	5,569	5,569	5,504	4,894	4,331	12,034
Individual controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Paternal fixed effects	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Maternal fixed effects	No	No	Yes	Yes	Yes	Yes	Yes	Yes

Interpreting the Dummy

A dummy variable creates two estimated equations:

$$D = 0: \quad \widehat{Y}_0 = \hat{\beta}_0 + \hat{\beta} X$$

$$D = 1: \quad \widehat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta} X$$

- $\hat{\beta}_1$ is the estimated **difference** between the two groups.
- Everything else in the model stays the same.

Interaction Example

Suppose returns to education differ for men and women.

We estimate:

$$Wage_i = \beta_0 + \beta_1 Educ_i + \beta_2 Female_i + \beta_3 (Educ_i \times Female_i) + \varepsilon_i$$

This allows the slope on education to differ by gender.

Interpreting the Interaction

The return to education becomes:

$$\text{Men: } \beta_1$$

$$\text{Women: } \beta_1 + \beta_3$$

Group Question
Based on the equation, what are the average wages for men and women without considering education at all?

Example: Alcohol Tax and Dating Outcomes

Suppose an alcohol tax goes into effect in Eugene.

We think:

- Freshmen/sophomores are mostly unaffected (too young to drink).
- Juniors/seniors previously used alcohol as a "social lubricant," so the tax might reduce successful first dates and thus probability of marriage.

We model this using class-year dummies.

Class-Year Dummy Specification

We include:

$$\textit{Freshman}, \ \textit{Sophomore}, \ \textit{Junior}, \ \textit{Senior}$$

but we **omit** one group (say Freshmen), making it the baseline.

Key Question
Why must we omit one group? What happens if we include all four dummies *and* an intercept?

Comparing Two Regressions

	Regression A (All 4 dummies included)	**Regression B** (Freshman omitted)
Intercept	72.4	72.4
Freshman	**NA (dropped)**	–
Sophomore	3.1	3.1
Junior	4.7	4.7
Senior	6.5	6.5
Observations	400	400

Omission of Dummy

Why is Freshman NA in Regression A?
Including all four class dummies **and** an intercept causes:

$$\text{Freshman} + \text{Sophomore} + \text{Junior} + \text{Senior} = 1,$$

which creates **perfect multicollinearity**.

The Dummy Variable Trap

Dummy Variable Trap: Perfect multicollinearity occurs when:

$$\text{All group dummies} + \text{Intercept} = 1$$

To avoid this:

- Omit one category ("base group"), or
    - Interpret the intercept as the baseline.
- Drop the intercept.

Additional Notes (Dougherty Chapter 5)

- You use t-tests on dummy coefficients just like any other regressor.
- F-tests can jointly test whether categories differ.
- In time series: dummy breaks can test slope changes (Chow test).

Transition: Enter DiD

We now know how to:

- use dummy variables,
- interpret group differences,
- interact dummies to allow effects to vary.

Difference-in-Differences (DiD) uses exactly these tools:

- A dummy for "treated"
- A dummy for "post"
- Their interaction: Treated $\times$ Post

Why Do We Need a Control Group?

Difference-in-Differences is widely used in **medicine** and public health.

- We introduce a treatment (a new drug, a therapy, a clinical protocol).
- We compare changes in outcomes for the treated group to a **control group**.
- The goal is to understand what would have happened *without* treatment.

Question for the class: Why isn't it enough to just look at the treated group before and after?

- Why is having a control group so important?

Motivation Example: Eugene Alcohol Tax

Suppose Eugene implements a new alcohol tax and we want to study its effect on social interaction, dating success, or long-run marriage probabilities.

Potential problem: Many other things could change in Eugene at the same time:

- UO changes its admissions criteria $\rightarrow$ different types of students.
- A shift in UO staffing or campus culture.
- New housing constraints, nightlife changes, or campus policies.
- Social trends affecting communication (phones, apps, etc.).

Any of these could affect dating/marriage outcomes, even without the tax.

Motivation Example Cont'd.

Solution: Choose a comparison group that shows how Oregon *would have changed* without the policy.

- If Oregon looked like another state (or similar universities) *before* the tax,
- then we can use that group to approximate Oregon's counterfactual trend.
- e.g.: Use OSU since it is another similarly large school in OR.

Why DiD? Connecting to What We Know

DiD is simply:

a treatment dummy $\quad + \quad$ a pre-post dummy $\quad + \quad$ their interaction.

The interaction isolates the causal effect of a policy that occurs at a point in time.

Recall: Omitted Variable Bias

OLS implies:

$$\hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(X, u)}{\text{Var}(X)}.$$

DiD works by **differencing out** the time-invariant portion of u.

The Standard DiD Regression

$$Y_{st} = \beta_0 + \beta_1 Treated_s + \beta_2 Post_t + \beta_3 (Treated_s \times Post_t) + \varepsilon_{st}$$

- β_1: baseline difference
- β_2: overall time trend
- β_3: **treatment effect**

The 2×2 Structure

	Pre (0)	Post (1)
Control (0)	$\beta_0 + \text{OVB}_c$	$\beta_0 + \beta_2 + \text{OVB}_c$
Treated (1)	$\beta_0 + \beta_1 + \text{OVB}_t$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \text{OVB}_t$

Step 1: Within-Group Differences

Treated change:

$$(\beta_2 + \beta_3) + (\mathrm{OVB}_t - \mathrm{OVB}_t) = \beta_2 + \beta_3$$

Control change:

$$\beta_2 + (\mathrm{OVB}_c - \mathrm{OVB}_c) = \beta_2$$

Step 2: Difference the Differences

$$(\beta_2 + \beta_3) - \beta_2 = \beta_3$$

β_3 is the treatment effect.

What Does DiD Remove?

DiD automatically removes:

- group-specific,
- time-invariant factors.

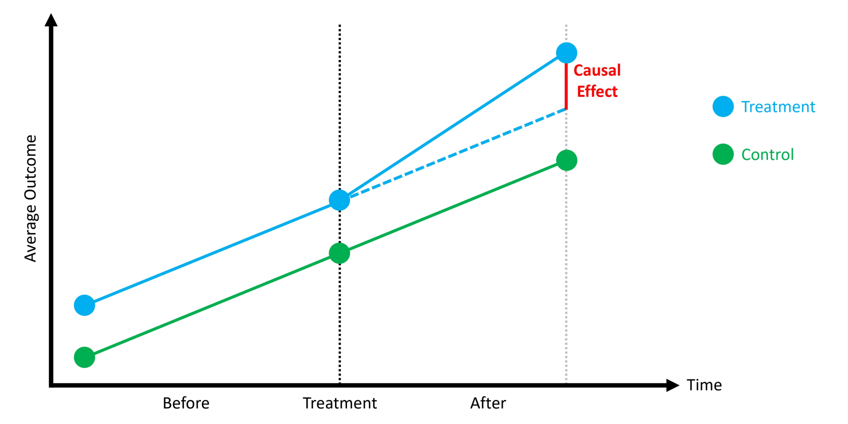
Culture, geography, stable health differences, etc.

Key Assumption: Parallel Trends

The control group must represent what *would have happened* to the treated group without treatment.

If pre-treatment trends move together, the control group's change is the counterfactual.

Visualizing Parallel Trends



Interpreting β_3

$$Y = \beta_0 + \beta_1 Treated + \beta_2 Post + \beta_3(Treated \times Post)$$

- β_1: baseline difference
- β_2: time trend
- β_3: **treatment effect**

Why We Use DiD

- Controls for unobserved, time-invariant differences
- Simple to implement
- Widely used in applied microeconomics

Summary

- Dummy variables capture treatment differences
- Interactions allow effects to differ before/after some time (treatment)
- The exogeneity assumption is hard – this (sometimes) makes it easier.
- Parallel trends is key (note: there are other assumptions, but this is the one you have to defend the most).

If Time

If anyone has read, let's discuss how Difference in Differences might be used in our final project. What is the comparison being made?

Next Class

- Discuss Chapter 6 in part.
- HW 4 will be out by end of the week and will cover Chapters 4-6 (and sort of Chapter 3 since they will all just be multiple regressions)