UCLA Anderson
SCHOOL of MANAGEMENT

PMDA

Lecture 5: Interaction Terms & Panel Data

Agenda today

- Interaction terms & treatment effect heterogeneity
- Panel data
  - Application: Price responsiveness in retail data
  - Two-way fixed effects as a special case of regression with controls

Quick Recap from last week: Sequential estimation of multivariate regression

- Let's drop the "i" subscript

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + e$$

- Frisch-Waugh Theorem: In principle, we could estimate each β sequentially. E.g.,
  1. First regress X_1 on all other X-variables

$$X_1 = a_0 + a_1 X_2 + a_2 X_3 + \cdots + \widetilde{X}_1$$

  2. Then regress Y on the residual $\widetilde{X}_1$ from the regression in step 1

$$Y = \beta_0 + \beta_1 \widetilde{X}_1 + v$$

The coefficient on the residual in the regression in step 2 equals β_1

- In practice, we estimate all the β's jointly not sequentially, but the theorem helps understand what a multivariate regression does in the background

Quick Recap from last week: Usefulness of FW theorem

- Frisch-Waugh theorem gives an expression for the precision of the coefficient estimators in multivariate regressions that is based on sequential estimation

- Multivariate regression thus gives you the effect of X_1 on Y after removing the variation in X_1 that can be explained by the other X -variables

- Explains why you can control for omitted variables and thus mitigate OVB when you know that the treatment was not random because it was assigned based on some observable characteristics

Where we left off last time ...

- We had a partially randomized treatment (= showing ad) due to age-based targeting
- We isolated variation in treatment that was uncorrelated with age bins
- Causal interpretation:
  - Conditional on age bins, treatment is random → the treatment coefficient has a causal interpretation
  - The treatment coefficient is also statistically significant (NOTE: significance has nothing to do with causality)

	coef	std err	t	P>|t|
Intercept	9.7054	0.120	80.955	0.000
treatment	1.2597	0.118	10.654	0.000
age_25to40	-1.1629	0.129	-9.033	0.000
age_above40	-0.8772	0.136	-6.469	0.000

  - We conclude that the ad campaign was effective since showing the ad increases revenue by $1.2597

Interaction terms

- New question: do consumers in different age bins respond differently to the ad?
- If they do, this justifies targeting based on age (which the firm did in the past)
- In the regression, we want to test whether the response to the ad (the coefficient of the treatment variable) is different for different age groups
- We can test this by adding interaction terms to the regression
    - Interaction terms are new variables that we generate by multiplying two existing variables
    - Here we multiply treatment with dummies for two age bins
    - NOTE: age and treatment also enter “on their own”

$$Revenue = \beta_0 + \beta_1 treatment + \beta_2 age25to40 + \beta_3 ageAbove40 + \beta_4 treatment \times age25to40 + \beta_5 treatment \times ageAbove40 + e$$

Interpreting coefficients if you have interaction terms/1

$$Revenue = \beta_0 + \beta_1 treatment + \beta_2 age25to40 + \beta_3 ageAbove40 + \beta_4 treatment \times age25to40 + \beta_5 treatment \times ageAbove40 + e$$

Let’s interpret the coefficients by using the regression to predict expected revenue for different types of users:

- E(revenue | treatment=0 & age<25) = β_0
- E(revenue | treatment=1 & age<25) = $\beta_0 + \beta_1$
- → treatment effect for age<25 group is β_1

- E(revenue | treatment=0 & 25≤age ≤ 40) = $\beta_0 + \beta_2$
- E(revenue | treatment=1 & 25≤age ≤ 40) = $\beta_0 + \beta_1 + \beta_2 + \beta_4$
- → treatment effect for 25≤age ≤ 40 group = $\beta_1 + \beta_4$
- → difference in treatment effect for 25≤age ≤ 40 group relative to age<25 = β_4
- → difference in treatment effect for age>40 group relative to age<25 = β_5

Interpreting coefficients if you have interaction terms/2

$$Revenue = \beta_0 + \beta_1 treatment + \beta_2 age25to40 + \beta_3 ageAbove40 + \beta_4 treatment \times age25to40 + \beta_5 treatment \times ageAbove40 + e$$

- Another way to get to the interpretation of the coefficients is to rewrite the regression by collecting the variable treatment:

$$Revenue = \beta_0 + \beta_2 age25to40 + \beta_3 ageAbove40 + (\beta_1 + \beta_4 age25to40 + \beta_5 ageAbove40) \times treatment + e$$

- $\beta_1 + \beta_4 age25to40 + \beta_5 ageAbove40$ is the treatment effect, which is allowed to depend on age
- $\rightarrow$ treatment effect for age<25 group is = β_1
- $\rightarrow$ treatment effect for 25≤age ≤ 40 group = $\beta_1 + \beta_4$
- $\rightarrow$ treatment effect for age > 40 group = $\beta_1 + \beta_5$
- $\rightarrow$ difference in treatment effect for 25≤age ≤ 40 group relative to age<25 = β_4
- $\rightarrow$ difference in treatment effect for age>40 group relative to age<25 = β_5

Results

- Let’s run this regression in R

	coef	std err	t	P>|t|
Intercept	9.8791	0.161	61.502	0.000
treatment	1.0283	0.185	5.547	0.000
age_25to40	-1.2998	0.181	-7.162	0.000
age_above40	-1.1305	0.180	-6.272	0.000
age_25to40:treatment	0.0439	0.266	0.165	0.869
age_above40:treatment	0.9952	0.314	3.170	0.002

slido

What is the effect for age above 40?

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Results

- Let's run this regression in R

	coef	std err	t	P>|t|
Intercept	9.8791	0.161	61.502	0.000
treatment	1.0283	0.185	5.547	0.000
age_25to40	-1.2998	0.181	-7.162	0.000
age_above40	-1.1305	0.180	-6.272	0.000
age_25to40:treatment	0.0439	0.266	0.165	0.869
age_above40:treatment	0.9952	0.314	3.170	0.002

- **Q:** Interpret the findings: is the effect of the ad different across age groups? How?
- **A:**
  - 25to40 group reacts similarly to treatment as <25 group (interaction term is insignificant)
  - Above 40 group has reaction that is twice as large relative to <25 group

Treatment effect heterogeneity

	coef	std err	t	P>|t|
Intercept	9.8791	0.161	61.502	0.000
treatment	1.0283	0.185	5.547	0.000
age_25to40	-1.2998	0.181	-7.162	0.000
age_above40	-1.1305	0.180	-6.272	0.000
age_25to40:treatment	0.0439	0.266	0.165	0.869
age_above40:treatment	0.9952	0.314	3.170	0.002

- The interaction terms allow us to assess whether different groups react differently to treatment
  - This type of analysis is at the heart of targeting / segmentation analysis
- Subtle, but very important point: the following statements are not the same
  1. The treatment is correlated with age
  2. The treatment effect depends on age
  - Statement 1. causes OVB and requires us to control for age
  - Statement 2. implies that the treatment affects different customers differently and can be captured via interaction terms

Managerial insight: who should you target?

	coef	std err	t	P>|t|
Intercept	9.8791	0.161	61.502	0.000
treatment	1.0283	0.185	5.547	0.000
age_25to40	-1.2998	0.181	-7.162	0.000
age_above40	-1.1305	0.180	-6.272	0.000
age_25to40:treatment	0.0439	0.266	0.165	0.869
age_above40:treatment	0.9952	0.314	3.170	0.002

slido

Which group should you target with advertising?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Managerial insight: who should you target?

	coef	std err	t	P>|t|
Intercept	9.8791	0.161	61.502	0.000
treatment	1.0283	0.185	5.547	0.000
age_25to40	-1.2998	0.181	-7.162	0.000
age_above40	-1.1305	0.180	-6.272	0.000
age_25to40:treatment	0.0439	0.266	0.165	0.869
age_above40:treatment	0.9952	0.314	3.170	0.002

Q: Which group should you target with advertising?

A:

- The results show that older customers spend less, but they respond more strongly to advertising
- Should focus on incremental effect, i.e. additional revenue generated by ad
- Incremental effect and revenue levels need not coincide (here they run in opposite directions, so the company should have targeted older users)
- Targeting based on levels is a common mistake (think of casinos targeting “high-rollers” with discounts)

Agenda today

- Interaction terms & treatment effect heterogeneity
- Panel data
  - Application: Price responsiveness in retail data
  - Two-way fixed effects as a special case of regression with controls

What are panel data?

- Data with a specific structure:
    - We observe multiple units (customers, stores, …)
    - We observe each unit for multiple time periods
    - We refer to these respectively as cross-sectional- and time-dimension
- Notation:
    - Use two subscripts to denote level at which variables vary
- **Example:** retail scanner data on sales and prices in i =1,…,100 stores measured each week t = 1,…,52 in a year, for a total of 5200 observations

$$Sales_{it} = \beta_0 + \beta_1 Price_{it} + e_{it}$$

    - Sales (and price) at store i in week t
- What panel data look like in Python

Balanced and unbalanced panels

- Balanced panel: we have data for all units i =1,…,N for all time periods t = 1,…,T

- Unbalanced panel: some units i are not observed in some time period t (missing data)
  - This could cause problems unless we can assume that data are missing at random

Why panel data are useful?

Panel data allows us to control for factors that:

- Vary across i (stores) but not over time, or that vary over time but not across i. Denote these with only subscript i or t

- Cause omitted variable bias if they are omitted from the regression

- May be difficult to measure → cannot be included in the regression as controls

Omitted variables in panel data

Q: For the sales on price regression, is this a potential omitted variable? Does it vary in both cross-sectional and time dimensions? Is it measurable?

- Store size:
  - affects both sales and price so potential OVB
  - does not vary over time
  - measurable
- Season (e.g. summer):
  - affects both sales and price so potential OVB
  - does not vary across stores
  - measurable

→ Since these are measurable, they could be included as control variables in a multivariate regression

Omitted variables in panel data

Q: For the sales on price regression, is this a potential omitted variable? Does it vary in both cross-sectional and time dimensions? Is it measurable?

- Store quality
    - affects both sales and price so potential OVB
    - does not vary over time
    - hard to measure
- Macroeconomic shock
    - affects both sales and price so potential OVB
    - does not vary across stores (affects all stores)
    - hard to measure

→ Since these are hard to measure, they cannot be included as control variables in a multivariate regression

→ However, panel data estimation can control for them because they only vary in one dimension

Omitted variables in panel data

Q: For the sales on price regression, is this a potential omitted variable? Does it vary in both cross-sectional and time dimensions? Is it measurable?

- Promotional campaigns/discounts
  - affects both sales and price so potential OVB
  - varies across both stores and time
  - maybe measurable, maybe not

→ If measurable, include it in the regression

→ If not measurable, panel data estimation cannot do anything about it so you will still have OVB

Omitted variables in panel data

- In summary, if Z is a possible omitted variable that would cause bias:
    - Panel data methods control for measurable or unmeasurable Z, as long as it varies in only one dimension: Z_i or Z_t
    - You can control for measurable Z that varies in both dimensions: measurable Z_{it} can be included as control in the panel regression
    - Panel data methods cannot control for unmeasurable Z that varies in both dimensions: unmeasurable Z_{it} cause OVB

Omitted variable bias in the scanner dataset

- In the scanner dataset, we can measure store size and the summer dummy so we could in principle add them to the regression
- Run the following three regressions:
  - (1) Sales on price
  - (2) Sales on price and store size
  - (3) Sales on price and stores size and summer dummy

Q: How do you expect the coefficient on price to change across the three regressions? Explain the direction of change based on the OVB formula

A:

- (1) → (2): coefficient goes up
  - Store size has positive impact on sales
  - Prices are lower in larger stores
- (2) → (3): coefficient goes down
  - Price and sales are higher in summer

OVB: changes in the price coefficient

	coef	std err	t	P>|t|
Intercept	6955.3832	17.189	404.631	0.000
price	$\boxed{-99.1455}$	7.548	-13.135	0.000

Goes up

	coef	std err	t	P>|t|
Intercept	658.9133	270.862	2.433	0.015
price	$\boxed{-45.8590}$	7.539	-6.083	0.000
store_size	137.5917	5.908	23.289	0.000

Goes down

	coef	std err	t	P>|t|
Intercept	1864.1927	236.606	7.879	0.000
price	$\boxed{-147.6197}$	6.981	-21.145	0.000
store_size	113.3881	5.155	21.996	0.000
summer_dummy	415.6049	10.023	41.463	0.000

The fixed effects regression models

- Regular regression model: $Y_{it} = \beta_0 + \beta_1 X_{it} + e_{it}$
- With panel data we can estimate more general regression models that allow for unobservable omitted variables

1. Fixed effects model

$$Y_{it} = \alpha_i + \beta_1 X_{it} + e_{it}$$

   - α_i is called the unit fixed effect (or simply fixed effect)
   - It includes omitted variables that only vary across i: $\alpha_i = \beta_0 + \beta_2 Z_i$

2. Time fixed effects model

$$Y_{it} = \delta_t + \beta_1 X_{it} + e_{it}$$

   - δ_t is called the time fixed effect
   - It includes omitted variables that only vary across t: $\delta_t = \beta_0 + \beta_2 Z_t$

3. Two-way fixed effects model

$$Y_{it} = \alpha_i + \delta_t + \beta_1 X_{it} + e_{it}$$

Estimation of fixed effects regression models

- The fixed effect models can be rewritten using dummy variables
- For example, the fixed effects model

$$Y_{it} = \alpha_i + \beta_1 X_{it} + e_{it}$$

  has a different intercept for each unit (e.g., store) i=1,…, N

- This means that it can be rewritten using N-1 different dummies for each unit

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 D2_i + \cdots + \beta_2 DN_i + e_{it}$$

  - α_1= β_0 is the intercept for the unit left out
  - α_i = β_0 + β_i are the intercepts for the other units

- You can see how this dummy representation generalizes to the time fixed effect and the two-way fixed effect models

Back to the scanner data

- Estimate a two-way fixed effect regression for the scanner data, which includes
  - One dummy for every store (minus 1)
  - One dummy for every week (minus 1)
  - NOTE: we do not add summer-dummy and store size as controls for now
- How do we implement this regression in Python?(can also use PanelOLS)
- How do we implement this regression in R?
  - Use special command `plm` (panel version of `lm` command)

```
result_price_two_way = smf.ols(formula = 'sales ~ price + C(store_id) + C(week)', data = retail).fit()
print(result_price_two_way.summary())
```

```
library(plm)
two_way_FE_reg = plm(sales ~ price, data=retail_data, index=c("store_id","week"),
model="within", effect="twoways")
summary(two_way_FE_reg)
```

- NOTE: It is common for fixed effect regressions to only report the estimate of the slope coefficient of interest β_1, not the estimates of the fixed effects

Comparing regressions

- Let's compare results from the two-way fixed effect regression with the earlier regular regression with store-size and summer controls
- Interestingly, price coefficient is very similar…

- Regular regression with controls

	coef	std err	t	P>|t|
Intercept	1864.1927	236.606	7.879	0.000
price	-147.6197	6.981	-21.145	0.000
store_size	113.3881	5.155	21.996	0.000
summer_dummy	415.6049	10.023	41.463	0.000

- Two-way FEs

price	-152.6799	7.355	-20.759	0.000

Fixed effects and other controls

Q: What happens if we control for store size in the two-way FE regression?

A: Store size effect will not be estimated (Python will report a multicollinearity issue) because the store dummies perfectly explain every omitted variable that does not change over time, such as store size

- Two-way fixed effects regression
  - Implicitly controls for store size and summer dummy by including unit and time fixed effects (i.e., store and week dummies)
  - In fact, it controls for any variable that varies either only across stores or only over time (e.g. store quality, other seasonal effects such as holidays)
  - → Powerful tool that eliminates bias from a variety of possible (and unobservable) omitted variables that do not vary in both dimensions

Do panel data give you causal estimates?

- **Q:** Can we say that the price coefficient has a causal interpretation, given that we have estimated it using a panel regression?
- **A:** Not necessarily! We have only eliminated bias caused by variables that vary only in one dimension
- Causal interpretation requires that price is also uncorrelated with all other determinants of sales that vary across both stores and times

- **Q:** In general, when should you include controls in a two-way fixed effects regression?
- **A:** If you can measure omitted variables that vary across time and units, you should include them in the regression as controls
- Any control that only varies in one dimension should not be included because it is fully explained by the fixed effects (the unit and time dummies)

Summary

- **Interaction terms**
  - Implementation: include product of two existing variables as a separate variable in the regression
  - Interpretation: can be used to analyze whether treatment affects different consumers differently (→ targeting & segmentation)

- **Panel regression with two-way fixed effects**
  - Way to eliminate bias due to omitted variables (observable and unobservable) that do not vary over time / do not vary across units
  - Avoids having to look for controls (unless they vary in both dimensions)
  - Careful: causality not guaranteed!

Workshop

- Go to BruinLearn Module 5/Workshop
- Download dataset and Jupyter Notebook