DS-UA 201: Causal Inference: Regression and grouping

Parijat Dube

New York University Center for Data Science

August 8, 2022

Acknowledgement: Slides including material from DS-US 201 Fall 2021 offered by Marco Morucci.

Recall

So far we have studied **selection-on-observables** designs. These are settings in which

- $Y_i(d) \perp\!\!\!\perp D_i | X_i$ ...
- ...and X_i is observed **in full**.

Starting with today we will be looking at settings in which X_i is not **fully observed**

- We have seen that we often incur in omitted variable bias in such settings
- However we will see that if the **unobserved confounders** follow certain conditions then causal inference is still possible!

What types of unobservables?

Today and in the next few lectures we will see that causal inference is still possible if there is **unobserved confounding** but:

- Unobservables are constant **within groups**. . .
- Unobservables are constant **over time**. . .
- The relationship between unobservables and the treatment assignment can be accounted for with some other variable . . .
- Treatment is assigned across an arbitrary threshold.

Today: Observations are grouped in such a way that unobserved confounders are constant within groups.

- and how to use regression to estimate the ATE in this setting.

Recall: Regression

Last lecture: we talked about linear regression:

$$Y_i = \alpha + X_i\beta + D_i\tau + \epsilon_i$$

- We saw how linear regression can be used as an imputation estimator for the ATE
- We also saw that the coefficient on D_i can be interpreted as the ATE under certain conditions ...
- ...and what to do when we don't think those conditions are met
- In all cases: OLS requires some **strong assumptions**!

Grouped observations

Suppose we are in this setting:

- We have our usual n units
- But they are grouped into G groups: $S_1, \ldots, S_G$.
- There is an **unobserved** confounder, U

$$\begin{tikzcd} & U \arrow[dl, dashed] \arrow[dr, dashed] & \\ D \arrow[rr] & & Y \end{tikzcd}$$

Key Assumption

Assume that U is **constant within groups**, that is, if $i, j \in S_g$, then $U_i = U_j = U_g$.

Grouped observations

We do not observe U

$$\begin{tikzcd} & U \arrow[dl, dashed] \arrow[dr, dashed] & \\ D \arrow[rr] & & Y \end{tikzcd}$$

- We cannot **condition** on U explicitly
- But if we do not account for U then we will have OVB

Can we exploit the fact that U is constant within groups to solve this problem?

- Intuitively . . .

What kinds of groups could we have?

There are many examples of **grouped** data that exist in real-world research:

- The same unit is observed multiple times over time
- Households, cities, villages, neighborhoods . . . (geographic groups)
- Days, weeks, years . . . (temporal groups)
- Social networks, clusters . . .

Stratified vs. Grouped data

In this class we will use the word "grouped" to denote data such that:

- The units can be divided into **mutually exclusive** groups
- Unobserved confounders are **constant** within groups

What are the differences between **grouped** and **stratified** data?

- Differences are almost exclusively **conceptual**
- In practice, methods for stratified data will work on grouped data
- The distinction is still useful to design and understand a study

Stratified vs. Grouped data

Stratified vs. grouped

In stratified data we **observe** the covariates that we stratify on, whereas in grouped data we do not.

- In the first case, we create strata so that observed confounders are constant within each strata
- In the second case, we assume that confounders are constant within a group
- In stratified data we know the value of the observed confounder that is associated with a stratum
- In grouped data we do not know the value of the confounder

Grouped observations and linear models

Today we make an additional assumption, that our outcome model is **linear**:

$$Y_i = \alpha + D_i\beta + U_i + \epsilon_i,$$

with $E[\epsilon_i] = 0$.

- This enables us to estimate the ATE with OLS
- Grouped observations can also be used without this assumption, but other estimators are needed.

If we use our assumption that $U_i = U_g$ for all $i \in S_g$:

$$Y_i = \alpha + D_i\beta + \sum_{g=1}^{G} \mathbb{1}(i \in S_g) U_g + \epsilon_i$$

Can we use this to get rid of U somehow?

"Within" estimator

What happens if we average all the outcomes within one group?

$$\bar{Y}_g = \alpha + \beta \bar{D}_g + U_g + \bar{\epsilon}_g,$$

where $\bar{D}_g = \frac{1}{N_g} \sum_{i \in S_g} D_i$, and $\bar{\epsilon}_g = \frac{1}{N_g} \sum_{i \in S_g} \epsilon_i$.

If we de-mean the units in group g with $\bar{Y}_g$ we get ...

$$\begin{aligned}
\tilde{Y}_i &= Y_i - \bar{Y}_g \\
&= \alpha + \beta D_i + U_g - \epsilon_i - \alpha - \beta \bar{D}_g - U_g - \bar{\epsilon}_g \\
&= \beta (D_i - \bar{D}_g) + (\epsilon_i - \bar{\epsilon}_g) \\
&= \beta \tilde{D}_i + \tilde{\epsilon}_i.
\end{aligned}$$

- U_g is gone!
- $E[\epsilon_i - \bar{\epsilon}] = 0$ by assumption
- This is still a linear regression!

“Within” estimator: summary

To fit a **within** estimator to grouped data you must:

1. Compute the mean outcome $\bar{Y}_g$ and mean treatment $\bar{D}_g$ in each group g
2. For each unit, subtract the respective group means from outcome and treatment:
   - $\tilde{Y}_i = Y_i - \bar{Y}_g$
   - $\tilde{D}_i = D_i - \bar{D}_g$
3. estimate the regression $\tilde{Y}_i = \alpha + \beta \tilde{D}_i + \tilde{\epsilon}_i$

The resulting coefficient on $\tilde{D}$, $\hat{\beta}$ will be a consistent estimate of the ATE.

Fixed-effects regression

There is an **alternative** way to formulate and implement regression for grouped data

- Results are going to be **eqivalent** to the “within” estimator
- Important to know because the two are used interchangeably both in science and industry

For each unit, create a new set of G **binary** variables: $W_{i1}, \dots, W_{iG}$ such that: $W_{ig} = \mathbb{1}(i \in S_g)$.

- For each unit, only one of these variables is 1 and all others are 0

Then we estimate the model:

$$Y_i = \alpha + \beta D_i + \sum_{g=1}^{G} \lambda_g W_{ig} + \epsilon_i$$

Fixed-effects estimator

The **fixed-effects** model looks like this:

$$Y_i = \alpha + \beta D_i + \sum_{g=1}^{G} \lambda_g W_{ig} + \epsilon_i$$

- Essentially fitting a different intercept for each unit i.

$$Y_i = \beta D_i + \lambda_{g_i} + \epsilon_i$$

- Again, this is equivalent to the "within" estimator.
- Coefficients on each dummy variable (λ_g) are often referred to as "fixed-effects"

“Within” estimator: summary

To fit a **fixed effect** estimator to grouped data you must:

1. Create G binary variables for each unit, such that: $W_{ig} = \mathbb{1}(i \in S_g)$.

2. Estimate the regression: $Y_i = \alpha + \beta D_i + \sum_{g=1}^{G} \lambda_g W_{ig} + \epsilon_i$

The resulting coefficient on D, $\hat{\beta}$ will be a consistent estimate of the ATE.

Treatment effect heterogeneity

We've assumed so far that the TE is constant: β. What if we have heterogeneous effects β_i?

Recall: last lecture we have seen that if $Y_i = \alpha + \beta_i D_i + \epsilon_i$, then

$$\hat{\beta} \xrightarrow{p} \frac{E[w_i \beta_i]}{E[w_i]},$$

and this is **not the ATE**!

This problem still persist in **fixed-effects** regression, i.e., if $Y_i = \alpha + \beta_i D_i + \lambda_g + \epsilon_i$, then $\hat{\beta}$ will still not be consistent for the ATE.

- Intuitively, this happens because $\Pr(D_i = 1)$ is different for units in different groups!

Treatment effect heterogeneity

Solution: Adapt the Lin estimator to this setting (Gibbons, Serrato, and Urbancic, 2019).

- We create the G variables denoting group membership
- We compute the mean of each W_{ig} **across all data**:

$$\bar{W}_g = \frac{1}{n}\sum_{i=1}^{n} W_{ig}.$$

- We de-mean each unit's binary variable like we did with covariates:

$$\tilde{W}_{ig} = W_{ig} - \tilde{W}_g.$$

Then we fit the regression:

$$Y_i = \alpha + \beta D_i + \sum_{g=1}^{G} \lambda_g \tilde{W}_{ig} + \sum_{g=1}^{G} \omega_g \left[\tilde{W}_{ig} D_i\right] + \epsilon_i$$

Treatment effect heterogeneity

The regression:

$$Y_i = \alpha + \beta D_i + \sum_{g=1}^{G} \lambda_g \tilde{W}_{ig} + \sum_{g=1}^{G} \omega_g \left[\tilde{W}_{ig} D_i \right] + \epsilon_i$$

- It's the "Lin" estimator with the "dummy" variables as the covariates!
- Easily implemented in `lm_lin()` in `estimatr`.

This model will lead to consistent estimates of the ATE even when individual TEs vary across units.

Relationship with stratification

The regression:

$$Y_i = \alpha + \beta D_i + \sum_{g=1}^{G} \lambda_g \tilde{W}_{ig} + \sum_{g=1}^{G} \omega_g \left[\tilde{W}_{ig} D_i \right] + \epsilon_i$$

Is in effect the same as applying the **stratification** estimator to the strata defined by the groups.

- Estimate $\hat{\tau}(g)$ by taking the difference-in-means in each group
- Aggregate with a weighted average: $\hat{\tau}_{block} = \sum_{g=1}^{G} \hat{\tau}(g) \frac{N_g}{n}$
- $\hat{\beta}$ and $\hat{\tau}_{block}$ will have the same value

Relationship with stratification

If de-meaned FE regression is the same as the stratification estimator then why use FE regression at all?

- FE regression requires stronger assumptions on Y than the stratification estimator!

But:

- FE regression can give use precise variance estimates even when **outcomes are correlated** within groups
  - In this case the Neyman variance estimator will not be consistent for the ATE variance

- FE regression can handle **multiple groups** at once

Fixed-effects regression variance

For simplicity define: $\mathbf{X}_i = [1, D_i, W_{i1}, \ldots, W_{iG}]$. Then:

$$Y_i = \mathbf{X}_i \gamma + \epsilon_i.$$

Recall: Estimating the variance of $\hat{\gamma}$ in OLS requires at least one assumption, that errors in the outcome are **uncorrelated** across units:

$$Cov(\epsilon_i, \epsilon_j | \mathbf{X}_i, \mathbf{X}_j) = 0.$$

However sometimes it makes sense to believe that errors for units **within the same group** are correlated:

$$Cov(\epsilon_i, \epsilon_j | i, j \in S_g) \neq 0$$

- For example, when groups are social networks and units are individuals
- In general reasonable to assume this when there is significant interaction between units

Fixed-effects regression variance

Recall: Estimating the variance of $\hat{\gamma}$ in OLS requires at least one assumption, that errors in the outcome are **uncorrelated** across units:

$$Cov(\epsilon_i, \epsilon_j | X_i, X_j) = 0,$$

Under this assumption, we saw that a **consistent** estimator for the variance matrix of $\hat{\gamma}$ is:

$$\widehat{Var}_{HC0}[\hat{\gamma}] = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{S}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1},$$

where:

$$\mathbf{S} = \begin{bmatrix} \hat{\epsilon}_1, & \dots, & 0 \\ 0, & \hat{\epsilon}_2, \dots, & 0 \\ \vdots, & & , \vdots \\ 0, & \dots, & \hat{\epsilon}_n, \end{bmatrix}$$

and $\hat{\epsilon}_i = Y_i - \mathbf{X}_i\hat{\gamma}$.

Fixed-effects regression variance

To account for correlation of errors within the same group, we change our assumption to

$$Cov(\epsilon_i, \epsilon_j | G_i \neq G_j) = 0,$$

That is errors are uncorrelated only if units are in **different** groups.

Under this assumption, we can extend the previous variance estimator by defining:

$$\hat{\epsilon}_{ij} = \begin{cases} (Y_i - \mathbf{X}_i\hat{\gamma})(Y_j - \mathbf{X}_j\hat{\gamma}) & \text{if } G_i = G_j \\ 0 & \text{otherwise .} \end{cases}$$

Note that $\hat{\epsilon}_{ii} = \hat{\epsilon}_i^2$.

Fixed-effects regression variance

Then we construct the matrix:

$$\mathbf{S}_{grp} = \begin{bmatrix} \hat{\epsilon}_1^2, & \hat{\epsilon}_{1,2}, \dots & \hat{\epsilon}_{1,n} \\ \vdots, & & , \vdots \\ \hat{\epsilon}_{1n}, & \dots, & \hat{\epsilon}_n^2, \end{bmatrix}$$

and finally, we can just **plug-in** $\mathbf{S}_{grp}$ in the previous variance estimator to obtain a cluster-robust estimator:

$$\widehat{Var}_{grp}[\hat{\gamma}] = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{S}_{grp}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$$

- This will account for error correlation across units in the same group

Multiple groups

Sometimes **multiple groups** are present within the same dataset

- For example, if the same units are observed at multiple times, one group is the unit and another group is the day of the observation

- The same unit is member of **multiple groups** at once
- Mutually-exclusive strata would have only one observation in them!
  - Simple stratification won't work!

Many fixed effects

Fixed-effects regression can handle this setting easily.

- We simply include as many fixed-effects as we have grouping variables:

$$Y_i = \beta D_i + \sum_{g=1}^{G} \lambda_g W_{ig} + \sum_{h=1}^{H} \delta_h V_{ih} + \epsilon_i$$

Where:

- there are two grouping variables g and h

- G is the total number of groups defined by g and H is the same for groups defined by H
- W_{ig} is 1 if $i \in S_g$ and V_{ig} is 1 if $i \in L_h$.

Many fixed effects

Fixed-effects regression can handle multiple groups easily . . .

. . . but we have to be **careful**:

- Intuitively, FE regression solves the problem of unique groups having only one member by **extrapolating** from the observations
- Units in different groups contribute to estimates for units in other groups
- If we believe there to be error correlation within multiple groups variance estimation is a problem
- Many strong modeling assumptions!

The BIMAS program

Dataset of 1027 observations of **rice farms** in indonesia

- Question: does an intensive rice production program increase output?
- Treatment: is the farm taking part in the program?
- Outcome: Gross rice output in Kilograms

We will group on the The **geographic region** of the farm

- Some regions inherently are more favorable to rice production
- Because of weather, economic conditions...

Data loading and Naive estimator

```r
library(estimatr)
library(plm)

data(RiceFarms)

# Define treatment, outcome and grouping variable
D = ifelse(RiceFarms$bimas=="no", 0, 1.0)
Y = RiceFarms$goutput
S = RiceFarms$region

# Naive estimate
lm_robust(Y ~ D)
```


- Estimate: 770.5
- SE: 164.2
- 95% CI: [448.2, 1092.8]

“within” estimator

```
Ytilde = rep(NA, nrow(ChickWeight))
Dtilde = rep(NA, nrow(ChickWeight))
for (chk in unique(S)){
  ss = which(S == chk)
  Ytilde[ss] = Y[ss] - mean(Y[ss])
  Dtilde[ss] = D[ss] - mean(D[ss])
}

lm_robust(Ytilde ~ Dtilde)
```


- Estimate: 632.2
- SE: 167.5
- 95% CI: [303.5, 961.0]

FE estimator


```
# FE estimator
lm_robust(Y ~ D + factor(S))
```


- Estimate: 632.2
- SE: 167.5
- 95% CI: [303.5, 961.0]

Lin FE estimator


```
# Lin FE estimator
lm_lin(Y ~ D, covariates=~ S)
```


- Estimate: 539.6
- SE: 150.9
- 95% CI: [243.4, 835.7]

Allowing for in-group correlations


```r
# Robust variance
lm_robust(Y ~ D + factor(S), clusters = S)
```


- Estimate: 632.2
- SE: 258.9
- 95% CI: [-81.7, 1346.3]

Multiple groupings

Now we also group by the **variety** of rice produced by the farm

- traditional, high yield, and mixed rice varieties

```
# Multiple FEs
L = RiceFarms$varieties
lm_robust(Y ~ D + factor(S) + factor(L))
```

- Estimate: 665.7
- SE: 163.3
- 95% CI: [345.2, 986.3]

Summary

Today we have introduced **linear regression** for **grouped data**

- When there are unobserved confounders that are **constant** within groups, then we can still estimate the ATE
- Fixed effects or "within" estimators extend regression to these settings
- The Lin estimator still allows for valid OLS with TE heterogeneity
- Variance of FE regression can account for correlation inside groups
- FE regression can handle multiple groups