Week 6: Clustered Data and Panels

Robust Standard Errors, Fixed and Random Effects

Max H. Farrell
The University of Chicago Booth School of Business
Clustering

No more time series. Back to SLR. Our assumptions were:

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \]

which in particular means

\[ \text{COV}(\varepsilon_i, \varepsilon_j) = 0 \quad \text{for all } i \neq j. \]

Clustering allows each observation to have

- unknown correlation with a small number others
- ... in a known pattern.
- Examples
  - Children in classrooms in schools
  - Firms in industries
  - Products made by companies
- How much independent information?
The SLR model with clustering

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \]

Instead

\[
\text{COV}(\varepsilon_i, \varepsilon_j) = \begin{cases} 
\sigma_i^2 & \text{if } i = j, \quad \text{just } \text{V}[\varepsilon_i] \\
\sigma_{ij} & \text{if } i \neq j, \text{ but in the same cluster} \\
0 & \text{otherwise.} 
\end{cases}
\]

So **only** standard errors change!

- Same slope \( \beta_1 \) for everyone

Cluster methods aim for **robustness**:

- No assumptions about \( \sigma_i^2 \) and \( \sigma_{ij} \)

- Assume we have **many** clusters \( G \), each with a **small** number of observations \( n_g \):

\[ n = \sum_{g=1}^{G} n_g \]
Example: Patents and R&D in 1991, by firm.id

> head(D91)

<table>
<thead>
<tr>
<th>year</th>
<th>sector</th>
<th>rdexp</th>
<th>firm.id</th>
<th>patents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1449</td>
<td>1991</td>
<td>4</td>
<td>6.287435</td>
<td>1</td>
</tr>
<tr>
<td>1450</td>
<td>1991</td>
<td>5</td>
<td>5.150736</td>
<td>2</td>
</tr>
<tr>
<td>1451</td>
<td>1991</td>
<td>2</td>
<td>4.172710</td>
<td>3</td>
</tr>
<tr>
<td>1452</td>
<td>1991</td>
<td>2</td>
<td>6.127538</td>
<td>4</td>
</tr>
<tr>
<td>1453</td>
<td>1991</td>
<td>11</td>
<td>4.866621</td>
<td>5</td>
</tr>
<tr>
<td>1454</td>
<td>1991</td>
<td>5</td>
<td>7.696947</td>
<td>6</td>
</tr>
</tbody>
</table>

Are these rows independent? If they were ...

> D91$newY <- log(D91$patents + 1)
> summary(slr <- lm(newY ~ log(rdexp), data=D91))

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -3.9226 | 0.7551 | -5.195 | 5.54e-07 |
| log(rdexp) | 4.1723 | 0.4531 | 9.208 | < 2e-16 |

Residual standard error: 1.451 on 179 degrees of freedom
What happens when errors are correlated?

▶ If $\varepsilon_i > 0$ we expect $\varepsilon_j > 0$. (if $\sigma_{ij} > 0$)

$\Rightarrow$ Both observation $i$ and $j$ are above the line.
We want our inference to be **robust** to this problem.

```r
> library(multiwayvcov); library(lmtest)
> vcov.slr <- cluster.vcov(slr, D91$sector)
> coeftest(slr, vcov.slr)

**t test of coefficients:**

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -3.923   | 0.909      | -4.31   | 2.649e-05|
| log(rdexp)     | 4.172    | 0.560      | 7.44     | 3.920e-12|
```

```r
> summary(slr)

**Coefficients:**

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -3.923   | 0.755      | -5.19   | 5.54e-07 |
| log(rdexp)     | 4.172    | 0.453      | 9.20    | < 2e-16  |
```
Can we just control for clusters? No!

- Not different slopes (and intercepts?) for each cluster ... we want one slope with the right standard error!

```r
> coeftest(slr, vcov.slr)

                  Estimate Std. Error    t value  Pr(>|t|)
(Intercept)   -3.9226303  0.9093307 -4.3137729 2.648864e-05
log(rdexp)       4.1722578  0.5603655  7.4456787 3.919954e-12

> slr.dummies <- lm(cty ~ displ + make, data=mpg.2005)
> summary(slr.dummies)

                  Estimate Std. Error    t value  Pr(>|t|)
log(rdexp)       4.500747  0.5145317  8.7471664 2.432794e-15
as.factor(sector)1 -5.880060  0.9234920 -6.3670073 1.829341e-09
as.factor(sector)2 -3.471446  0.8794350 -3.9469461 4.521279e-04
...                ...         ...            ...         ...
```
Can we just control for clusters? No!

- Not different slopes (and intercepts?) for each cluster . . . we want one slope with the right standard error!
Panel Data

So far we have seen i.i.d. data and time series data. **Panel** data combines these:

- units $i = 1, \ldots, n$
- followed over time periods $t = 1, \ldots, T$

$\Rightarrow$ dependent over time, possibly clustered

More and more datasets are **panels**, also called **longitudinal**

- Tracking consumer decisions
- Firm financials over time
- Macro data across countries
- Students in classrooms over several grades

Distinct from a *repeated cross-section*:

- New units sampled each time $\Rightarrow$ independent over time
The linear regression model for panel data:

\[ Y_{i,t} = \beta_1 X_{i,t} + \alpha_i + \gamma_t + \varepsilon_{i,t} \]

Familiar pieces, just like SLR:

▶ \( \beta_1 \) – the **general** trend, same as always. (*Where’s \( \beta_0 \)‽*)

▶ \( Y_{i,t}, X_{i,t}, \varepsilon_{i,t} \) – Outcome, predictor, mean zero idiosyncratic shock (clustered?)

What’s new:

▶ \( \alpha_i \) – **unit**-specific effects. Different **people** are different!
  
  ▶ Cars: Camry/Tundra/Sienna. S&P500: Hershey/UPS/Wynn

▶ \( \gamma_t \) – **time**-specific effects. Different **years** are different!

▶ For now, \( \gamma_t = 0 \). Same concepts/methods.

Just the familiar **same slope**, **different intercepts** model!

Well, almost . . .
Estimation strategy depends on how we think about $\alpha_i$

1. $\alpha_i = 0 \implies Y_{i,t} = \beta_1 X_{i,t} + \varepsilon_{i,t}$
   - $\textbf{lm}$ on $N = nT$ observations. Cluster if needed.

2. random effects: $\text{cor}(\alpha_i, X_{i,t}) = 0$
   - Still possible to use $\textbf{lm}$ on $N = nT$ (and cluster on unit) . . .
     
     $Y_{i,t} = \beta_1 X_{i,t} + \tilde{\varepsilon}_{i,t}$, $\tilde{\varepsilon}_{i,t} = \alpha_i + \varepsilon_{i,t}$
   - . . . but lots of variance!

3. fixed effects: $\text{cor}(\alpha_i, X_{i,t}) \neq 0$
   - same slope, but $n$ different intercepts!
     
     $Y_{i,t} = \beta_1 X_{i,t} + \alpha_i + \varepsilon_{i,t}$
   - Too many parameters to estimate. patent data has $n = 181$.
   - No time-invariant $X_{i,t} = X_i$. 
The real **patent** data is a **panel** with clustering:

- unit is a **firm**: \( i = 1, \ldots, 181 \)
- time is **year** = 1983, \ldots, 1991
- clustered by **sector**?

\[
\begin{array}{cccccccccccc}
\end{array}
\]

\[
\begin{array}{cccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
4 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
5 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\cdots \\
\end{array}
\]
Estimation in R: using `lm` or the `plm` package.

1. $\alpha_i = 0$

```r
> slr <- lm(newY ~ log(rdexp), data=D)
> plm.pooled <- plm(newY ~ log(rdexp), data=D,
+     index=c("firm.id", "year"), model="pooling")
```

2. **random effects**: $\text{cor}(\alpha_i, X_{i,t}) = 0$

```r
> vcov.model <- cluster.vcov(slr, D$firm.id)
> coeftest(slr, vcov.model)
> plm.random <- plm(newY ~ log(rdexp), data=D,
+     index=c("firm.id", "year"), model="random")
```

3. **fixed effects**: $\text{cor}(\alpha_i, X_{i,t}) \neq 0$

```r
> many.dummies <- lm(newY ~ log(rdexp) + as.factor(firm.id) - 1,
> plm.fixed <- plm(newY ~ log(rdexp), data=D,
+     index=c("firm.id", "year"), model="within")
```
Choosing between fixed or random effects.

- Fixed effects are more general, more realistic: isolate changes due to $X$ vs due to specific person.
- If $\alpha_i$ don't matter, then $b_{RE} \approx b_{FE}$

```r
> phtest(plm.random, plm.fixed)
```

Hausman Test

data:  newY ~ log(rdexp)
chisq = 22.162, df = 1, p-value = 2.506e-06
alternative hypothesis: one model is inconsistent

Using year fixed effects ($\gamma_t$).

```r
> lm(newY ~ log(rdexp) + as.factor(year) - 1, data=D)
> plm(newY ~ log(rdexp), data=D, 
+   index=c("firm.id", "year"), model="within", effect="time")
```

Both firm and year fixed effects $\rightarrow$ effect="twoways"
Clustered Panels

A panel is not exempt from the concern of clustered data.

\[ Y_{i,t} = \beta_1 X_{i,t} + \alpha_i + \gamma_t + \varepsilon_{i,t} \quad \text{cor}(\varepsilon_{i_1,t_1}, \varepsilon_{i_2,t_2}) \neq 0 \]

> summary(plm.fixed)

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| log(rdexp) | 2.22611 | 0.22642 | 9.832 | < 2.2e-16 |

> vcov <- cluster.vcov(many.dummies, D$sector)
> coeftest(plm.fixed, vcov)

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| log(rdexp) | 2.22611 | 0.80872 | 2.7527 | 0.005985 |

\[ \rightarrow \text{Four times less information!} \]
Prediction in Panels

Just use the usual prediction?

\[ \hat{Y}_{f,i,t} = b_1 X_{f,i,t} + \hat{\alpha}_i + \hat{\gamma}_t \]

Predicting for who? when?

Only works if \( \hat{\alpha}_i \approx \alpha_i \) and \( \hat{\gamma}_t \approx \gamma_t \)

- Long panels (large \( T \)) and no \( \gamma_t \)
- Many units (large \( n \)) and no \( \alpha_i \)
- How big is big enough?

Uncertainty, same idea as before.

- Prediction intervals: same logic, similar formula, but **more** uncertainty.
- Intervals can be **wide**!
Further Issues in Panel Data

More general models
- Dynamic models – adding $X_{i,t} = Y_{i,t-1}$?
- Nonlinear model – binary $Y$?
- ... lots more.

Specification Tests
- Breusch-Pagan – time effects
- Wooldridge – serial correlation
- Dickey-Fuller – non-stationarity over time
- ... lots more.
Coming Up

First, take a well-earned break!

OK, that’s long enough. Back to the grind . . .

- Project proposals in two weeks
- Keep questions coming over email/Piazza
- Midterms coming back in your mailfolders (eventually)