BUS41100 Applied Regression Analysis

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

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

Instead

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

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: Fuel Economy in 2005: engine size and MPG

```r
> mpg.2005 <- read.csv("fueleconomy_2005.csv")
> head(mpg.2005)

<table>
<thead>
<tr>
<th>make</th>
<th>trans</th>
<th>cyl</th>
<th>displ</th>
<th>cty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acura</td>
<td>Auto</td>
<td>6</td>
<td>3.5</td>
<td>15</td>
</tr>
<tr>
<td>Acura</td>
<td>Auto</td>
<td>6</td>
<td>3.0</td>
<td>16</td>
</tr>
<tr>
<td>Acura</td>
<td>Manual</td>
<td>6</td>
<td>3.2</td>
<td>16</td>
</tr>
<tr>
<td>Acura</td>
<td>Auto</td>
<td>6</td>
<td>3.5</td>
<td>16</td>
</tr>
</tbody>
</table>

Are these rows independent? If they were ...

```r
> summary(slr <- lm(cty ~ displ, data=mpg.2005))
```

Coefficients:

|          | Estimate | Std. Error | t value | Pr(>|t|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 25.38236 | 0.23945 | 106.00 | <2e-16   |
| displ     | -2.49434 | 0.06729 | -37.07 | <2e-16   |

Residual standard error: 2.657 on 1090 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
> # install.packages("multiwayvcov")
> library(multiwayvcov)
> library(lmtest)
> vcov.slr <- cluster.vcov(slr, mpg.2005$make)
> coeftest(slr, vcov.slr)

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 25.38236   | 1.07708 | 23.5659  | < 2.2e-16 |
| displ     | -2.49434   | 0.26587 | -9.3818  | < 2.2e-16 |

> summary(slr)

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|----------|
| (Intercept) | 25.38236   | 0.23945 | 106.00   | <2e-16    |
| displ     | -2.49434   | 0.06729 | -37.07   | <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) 25.38236   1.07708  23.566  < 2.2e-16
     displ -2.49434   0.26587  -9.382  < 2.2e-16

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

           Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.12278    0.74731  33.618  < 2e-16
     displ -2.37995    0.07693  -30.936  < 2e-16
makeAudi  -1.31496    0.80687  -1.630    0.1035
makeBMW  -1.15510    0.78133  -1.478    0.1396
... ... ...
```
Can we just control for clusters? No!

- Not different slopes (and intercepts?) for each cluster... we want one slope with the right standard error!
Suppose our model has all two-way interactions:

\[
\text{mlr1} \leftarrow \text{lm}(\text{cty} \sim \text{trans}\times\text{cyl} + \text{trans}\times\text{displ} + \text{cyl}\times\text{displ})
\]

Should we add the three-way interaction, \text{trans}\times\text{cyl}\times\text{displ}?

\[
\text{summary(mlr2} \leftarrow \text{lm}(\text{cty} \sim \text{trans}\times\text{cyl}\times\text{displ}, \text{data=mpg.2005}))
\]

Coefficients:

|                  | Estimate | Std. Error | t value | Pr(>|t|)  |
|------------------|----------|------------|---------|-----------|
| (Intercept)      | 33.73003 | 1.03810    | 32.492  | < 2e-16   |
| transManual      | 4.66041  | 1.67316    | 2.785   | 0.00544   |
| cyl              | -1.57417 | 0.21070    | -7.471  | 1.63e-13  |
| displ            | -4.74922 | 0.31639    | -15.011 | < 2e-16   |
| transManual:cyl  | -0.51235 | 0.34995    | -1.464  | 0.14346   |
| transManual:displ| -1.95568 | 0.65168    | -3.001  | 0.00275   |
| cyl:displ        | 0.39639  | 0.04344    | 9.125   | < 2e-16   |
| transManual:cyl:displ | 0.21558 | 0.08122    | 2.654   | 0.00806   |

Yes! Add it to the model.
But using **robust** standard errors ...

```r
> vcov.mlr2 <- cluster.vcov(mlr2, mpg.2005$make)
> coeftest(mlr2, vcov.mlr2)
```

**t test of coefficients:**

|                      | Estimate | Std. Error | t value | Pr(>|t|)   |
|----------------------|----------|------------|---------|------------|
| (Intercept)          | 33.73003 | 3.39520    | 9.9346  | < 2.2e-16  |
| transManual          | 4.66041  | 2.08536    | 2.2348  | 0.0256319  |
| cyl                  | -1.57417 | 0.43681    | -3.6038 | 0.0003278  |
| displ                | -4.74922 | 1.09705    | -4.3291 | 1.636e-05  |
| transManual:cyl      | -0.51235 | 0.31930    | -1.6046 | 0.1088683  |
| transManual:displ    | -1.95568 | 1.30011    | -1.5042 | 0.1328113  |
| cyl:displ            | 0.39639  | 0.12843    | 3.0864  | 0.0020772  |
| transManual:cyl:displ| 0.21558  | 0.13515    | 1.5951  | 0.1109780  |

No!
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 real fuel economy data is a panel with clustering:

- unit is a model: \( i = 1, \ldots, 82 \)
- time is year = 2001, \ldots, 2010
- clustered by make

```r
> fuel.panel
       make model year class trans  cyl displ cty
1      1 Audi  A4  2001  compact    Auto   4   1.8   18
2      2 Audi  A4  2002  compact    Auto   4   1.8   18
3      3 Audi  A4  2003  compact    Auto   4   1.8   20
4      4 Audi  A4  2004  compact    Auto   4   1.8   20
5      5 Audi  A4  2005  compact    Auto   4   1.8   20
6      6 Audi  A4  2006  compact    Auto   4   2.0   21
7      7 Audi  A4  2007  compact    Auto   4   2.0   21
8      8 Audi  A4  2008  compact    Auto   4   2.0   21
9      9 Audi  A4  2009  compact    Auto   4   2.0   23
10    10 Audi  A4  2010  compact    Auto   4   2.0   23
11    11 Audi A4 Avant quattro  2001   wagon    Auto   4   1.8   18
12    12 Audi A4 Avant quattro  2002   wagon    Auto   4   1.8   17
13    13 Audi A4 Avant quattro  2003   wagon    Auto   4   1.8   18
14    14 Audi A4 Avant quattro  2004   wagon    Auto   4   1.8   18
15    15 Audi A4 Avant quattro  2005   wagon    Auto   4   1.8   18
...
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 \textbf{general} trend, same as always. \((\text{Where’s } \beta_0?)\)
- \( Y_{i,t}, X_{i,t}, \varepsilon_{i,t} \) – Outcome, predictor, mean zero idiosyncratic shock (clustered?)

What’s new:

- \( \alpha_i \) – \textbf{unit}-specific effects. Different \textbf{people} are different!
  - Cars: Camry/Tundra/Sienna. S&P500: Hershey/UPS/Wynn
- \( \gamma_t \) – \textbf{time}-specific effects. Different \textbf{years} are different!
- For now, \( \gamma_t = 0 \). Same concepts/methods.

Just the familiar \textbf{same slope}, \textbf{different intercepts} model!

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

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

2. random effects: $\text{cor}(\alpha_i, X_{i,t}) = 0$
   - Still possible to use $\text{lm}$ on $N = nT$ (and cluster on unit) …

   $Y_{i,t} = \beta_1 X_{i,t} + \tilde{\varepsilon}_{i,t}, \quad \tilde{\varepsilon}_{i,t} = \alpha_i + \varepsilon_{i,t}$
   - … 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. fuel data has $n = 82$.
   - No time-invariant $X_{i,t} = X_i$. 

13
Estimation in R: using \texttt{lm} or the \texttt{plm} package.

1. $\alpha_i = 0$
   
   \begin{verbatim}
   > slr <- lm(cty ~ displ, data=fuel.panel)
   > plm.pooled <- plm(cty ~ displ, data=fuel.panel, 
   + index=c("model", "year"), model="pooling")
   \end{verbatim}

2. random effects: $\text{cor}(\alpha_i, X_{i,t}) = 0$
   
   \begin{verbatim}
   > vcov.model <- cluster.vcov(slr, fuel.panel$model)
   > coeftest(slr, vcov.model)
   > plm.random <- plm(cty ~ displ, data=fuel.panel, 
   + index=c("model", "year"), model="random")
   \end{verbatim}

3. fixed effects: $\text{cor}(\alpha_i, X_{i,t}) \neq 0$
   
   \begin{verbatim}
   > dummies <- lm(cty ~ displ + as.factor(model), data=fuel.panel)
   > plm.fixed <- plm(cty ~ displ, data=fuel.panel, 
   + index=c("model", "year"), model="within")
   \end{verbatim}
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:  cty ~ displ
chisq = 71.144, df = 1, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent

Adding year fixed effects ($\gamma_t$).

```r
> lm(cty ~ displ + as.factor(year), data=fuel.panel)
> plm(cty ~ displ, data=fuel.panel,
+ index=c("model", "year"), model="within", effect="time")
```
Prediction

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)