PLSC 504 – Fall 2017
Introduction to Panel/TSCS Data

October 17, 2017

Starting Points

- "Longitudinal" $\neq$ "Time Series"
- Terminology:
  - "Unit" / "Units" / "Units of observation" / "Panels" $=$ Things we observe repeatedly
  - "Observations" $=$ Each (one) measurement of a unit
  - "Time points" $=$ When each observation on a unit is made
  - $i \in \{1....N\}$ indexes units
  - $t \in \{1...T\}$ or $\{1...T_i\}$ indexes observations / time points
  - If $T_i = T \; \forall i$ then we have "balanced" panels / units
  - $NT =$ Total number of observations (if balanced)
- Averages:
  - Y_{it} indicates a variable that varies over both units and time,
  - $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it} =$ the over-time mean of Y,
  - $\bar{Y}_t = \frac{1}{N} \sum_{i=1}^{N} Y_{it} =$ the across-unit mean of Y, and
  - $\bar{Y} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it} =$ the grand mean of Y.

More Terminology

- $N >> T \ \rightarrow$ "panel" data
  - NES panel studies ($N = 2000$, $T = 3$)
  - Panel Study of Income Dynamics ($N =$ large, $T \approx 12$)

- $T >> N$ or $T \approx N \ \rightarrow$ "time-series cross-sectional" ("TSCS") data

- $N = 1 \ \rightarrow$ "time series" data

Panel/TSCS Data Structure

id	t	Y	X_1	...
1	1	250	3.4	...
1	2	290	3.3	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...
2	1	160	4.7	...
2	2	150	4.9	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...

Variation: A Tiny (Fake) Example

id	year	gender	pres	pid	approve
1	1998	female	clinton	dem	3
1	2000	female	clinton	dem	3
1	2002	female	bush	dem	5
1	2004	female	bush	dem	3
2	1998	male	clinton	gop	2
2	2000	male	clinton	gop	1
2	2002	male	bush	gop	4
2	2004	male	bush	gop	3
3	1998	male	clinton	gop	2
3	2000	male	clinton	gop	2
3	2002	male	bush	gop	4
3	2004	male	bush	dem	1

Aggregation: Cross-Sectional

id	gender	pres	pid	approve
1	female	?	dem	3.50
2	male	?	gop	2.50
3	male	?	?	2.25

year	female	pres	pid	approve
1998	0.33	clinton	0.66(?)	2.33
2000	0.33	clinton	0.66(?)	2.00
2002	0.33	bush	0.66(?)	4.33
2004	0.33	bush	0.33(?)	2.33

The Point

Aggregation:

- Loses information
- Distorts relationships
- Forces arbitrary decisions

If you have variation in multiple dimensions, use it.

Within- and Between-Unit Variation

Define:

$$\bar{Y}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} Y_{it}$$

Then:

$$Y_{it} = \bar{Y}_i + (Y_{it} - \bar{Y}_i).$$

- The *total* variation in Y_{it} can be decomposed into
- The *between-unit* variation in the $\bar{Y}_i$s, and
- The *within-unit* variation around $\bar{Y}_i$ (that is, $Y_{it} - \bar{Y}_i$).

Variation (SCOTUS Tenure Remix)

"Total" Variation:

```
> with(scotus, describe(service))
   vars    n  mean   sd median trimmed mad min max range skew kurtosis
X1    1 1765 11.74 8.34     10   10.93 8.9   1  37    36 0.73    -0.28
    se
X1 0.2
```

"Between" Variation:

```
> scmeans <- ddply(scotus,.(justice),summarise,
+                  service = mean(service))
> with(scmeans, describe(service))
   vars   n mean   sd median trimmed  mad min max range skew kurtosis
X1    1 107 8.87 4.99    8.5    8.59 5.93 1.5  21  19.5  0.4    -0.92
     se
X1 0.48
```

"Within" Variation:

```
> scotus <- ddply(scotus, .(justice), mutate,
+                 servmean = mean(service))
> scotus$within <- with(scotus, service-servmean)
> with(scotus, describe(within))
   vars    n mean   sd median trimmed  mad min max range skew kurtosis
X1    1 1765    0 6.92      0       0 6.67 -18  18    36    0    -0.36
     se
X1 0.16
```

Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

assumes:

- All the usual OLS assumptions, plus
- $\beta_{0i} = \beta_0 \;\forall\; i$s
- $\beta_{1i} = \beta_1 \;\forall\; i$s

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

(same)

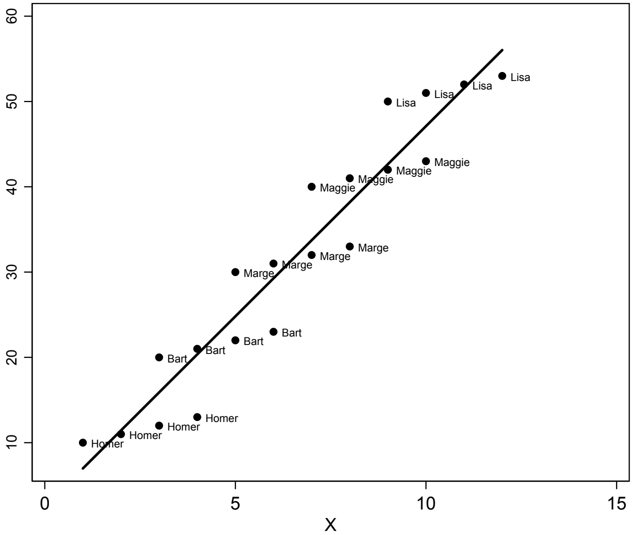
Variable Intercepts

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + u_{it}$$

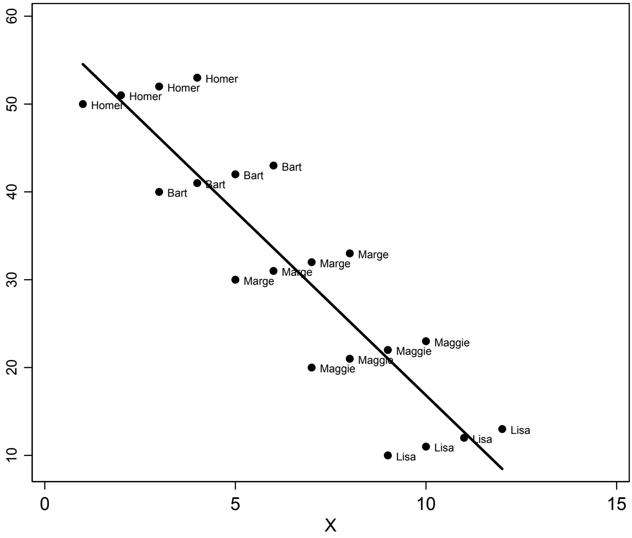
$$Y_{it} = \beta_{0t} + \beta_1 X_{it} + u_{it}$$

$$Y_{it} = \beta_{0it} + \beta_1 X_{it} + u_{it}$$

Varying Intercepts



Varying Intercepts



Varying Slopes (+ Intercepts)

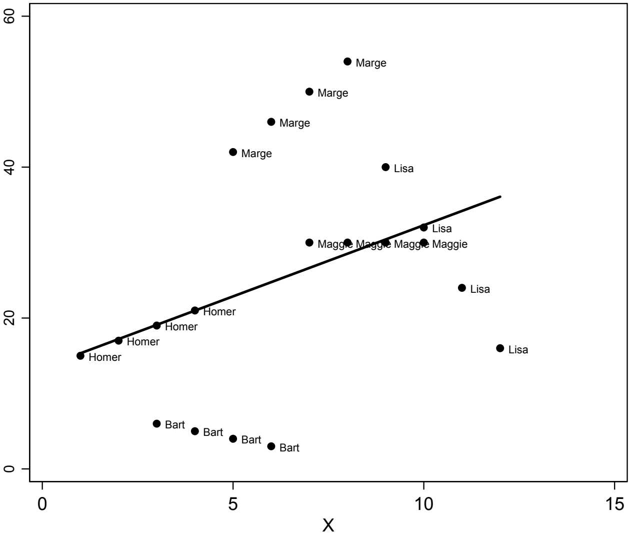
$$Y_{it} = \beta_0 + \beta_{1i} X_{it} + u_{it}$$

$$Y_{it} = \beta_{0i} + \beta_{1i} X_{it} + u_{it}$$

$$Y_{it} = \beta_{0t} + \beta_{1t} X_{it} + u_{it}$$

$$Y_{it} = \beta_{0it} + \beta_{1it} X_{it} + u_{it}$$

Varying Slopes + Intercepts



The Error

$$u_{it} \sim \text{i.i.d.} N(0, \sigma^2) \; \forall \; i, t$$

$$\begin{aligned}
\text{Var}(u_{it}) &= \text{Var}(u_{jt}) \; \forall \; i \neq j \text{ (i.e., no cross-unit heteroscedasticity)} \\
\text{Var}(u_{it}) &= \text{Var}(u_{is}) \; \forall \; t \neq s \text{ (i.e., no temporal heteroscedasticity)} \\
\text{Cov}(u_{it}, u_{js}) &= 0 \; \forall \; i \neq j, \; \forall \; t \neq s \text{ (i.e., no auto- or spatial correlation)}
\end{aligned}$$

- Adds data
- Generalizability

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

Implies

- that the process governing the relationship between X and Y is exactly the same for each i,
- that the process governing the relationship between X and Y is the same for all t,
- that the process governing the us is the same $\forall$ i and t as well.

"Partial" Pooling

Two regimes:

$$Y_A = \beta'_A \mathbf{X}_A + u_A$$

$$Y_B = \beta'_B \mathbf{X}_B + u_B$$

with $\sigma^2_A = \sigma^2_B$, and $\mathrm{Cov}(u_A, u_B) = 0$.

Estimators:

$$\hat{\beta}_{A,B} = (\mathbf{X}'_{A,B}\mathbf{X}_{A,B})^{-1}\mathbf{X}'_{A,B}Y_{A,B}$$

and

$$\widehat{\mathrm{Var}(\beta_{A,B})} = \hat{\sigma}^2_{A,B}(\mathbf{X}'_{A,B}\mathbf{X}_{A,B})^{-1},$$

A Pooled Estimator

$$\begin{aligned}
\hat{\beta}_P &= (\mathbf{X}_A'\mathbf{X}_A + \mathbf{X}_B'\mathbf{X}_B)^{-1}(\mathbf{X}_A' Y_A + \mathbf{X}_B' Y_B) \\
&= (\mathbf{X}_A'\mathbf{X}_A + \mathbf{X}_B'\mathbf{X}_B)^{-1}[\beta_A(\mathbf{X}_A'\mathbf{X}_A) + \beta_B(\mathbf{X}_B'\mathbf{X}_B)],
\end{aligned}$$

$$\begin{aligned}
\mathsf{E}(\hat{\beta}_P) &= \beta_A + (\mathbf{X}_A'\mathbf{X}_A + \mathbf{X}_B'\mathbf{X}_B)^{-1}\mathbf{X}_B'\mathbf{X}_B(\beta_B - \beta_A) \\
&= \beta_B + (\mathbf{X}_A'\mathbf{X}_A + \mathbf{X}_B'\mathbf{X}_B)^{-1}\mathbf{X}_A'\mathbf{X}_A(\beta_A - \beta_B)
\end{aligned}$$

$$F = \frac{\frac{\hat{\mathbf{u}}_P'\hat{\mathbf{u}}_P - (\hat{\mathbf{u}}_A'\hat{\mathbf{u}}_A + \hat{\mathbf{u}}_B'\hat{\mathbf{u}}_B)}{K}}{\frac{(\hat{\mathbf{u}}_A'\hat{\mathbf{u}}_A + \hat{\mathbf{u}}_B'\hat{\mathbf{u}}_B)}{(N_A + N_B - 2K)}} \sim F_{[K,(N_A+N_B-2K)]}$$

Fractional Pooling

$$\hat{\beta}_\lambda = (\lambda^2 \mathbf{X}'_A \mathbf{X}_A + \mathbf{X}'_B \mathbf{X}_B)^{-1} (\lambda^2 \mathbf{X}'_A Y_A + \mathbf{X}'_B Y_B)$$

with $\lambda \in [0, 1]$:

- $\lambda = 0 \rightarrow$ separate estimators for $\hat{\beta}_A$ and $\hat{\beta}_B$,
- $\lambda = 1 \rightarrow$ "fully pooled" estimator $\hat{\beta}_P$,
- $0 < \lambda < 1 \rightarrow$ a regression where data in regime A are given some "partial" weighting in their contribution towards an estimate of β.

Pooling, Summarized

"(R)oughly speaking, it makes sense to pool disparate observations if the underlying parameters governing those observations are sufficiently similar, but not otherwise."