Unobserved Heterogeneity

An Introduction to Panel Data

The Population Regression Equation

- In an OLS assume *COMPLETE PARAMETER HOMOGENEITY*
  - Each and every cross-section unit is assumed to share the behavioral parameters β
  - Strong Assumption: Everyone reacts to a change in gasoline prices in exactly the same way (in a strictly linear specification)

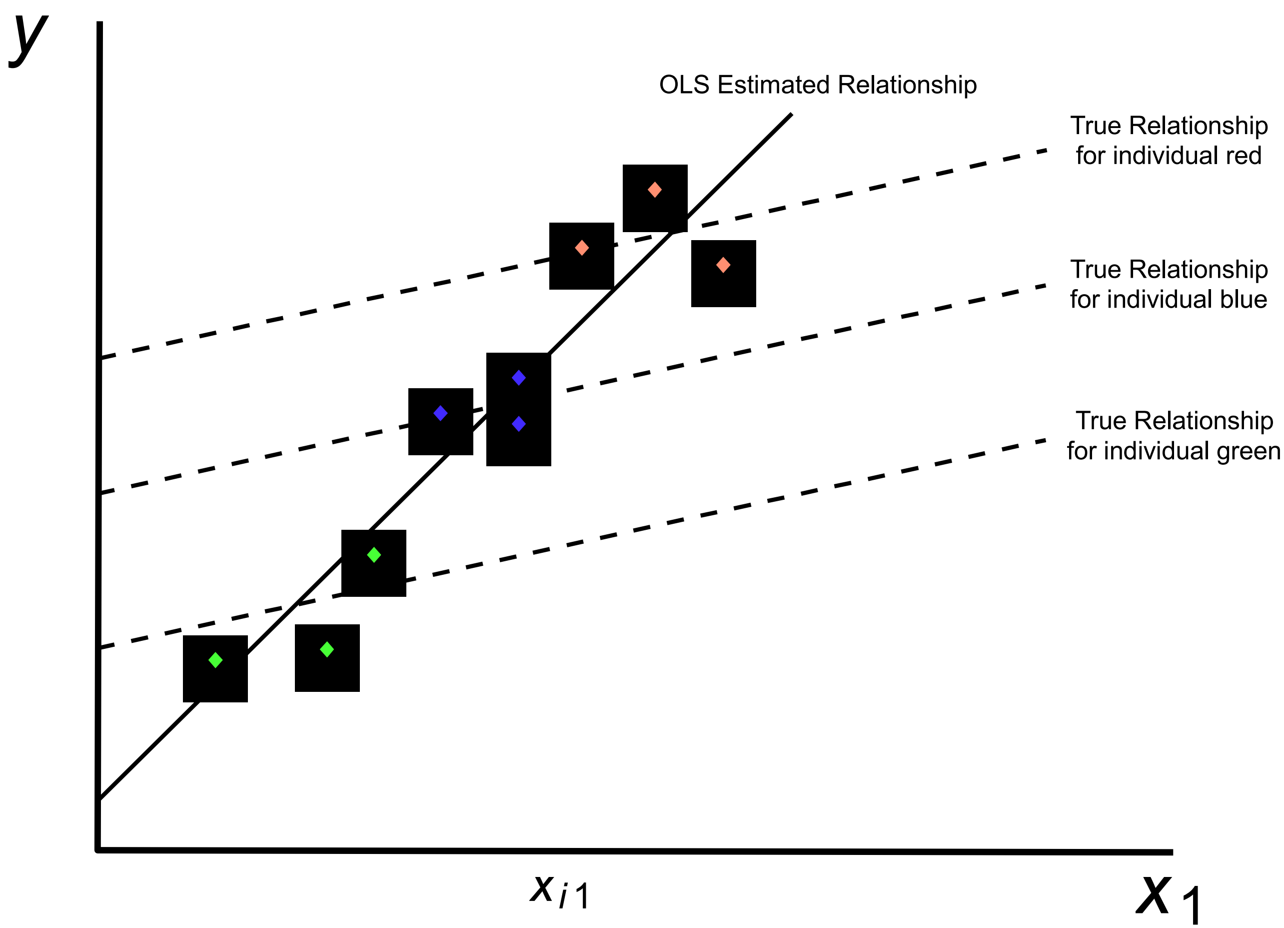
Limitations of Cross Section Data

- In a linear regression setting, with such unobserved heterogeneity, perhaps difficult to estimate the correct slope coefficient, because we have a sample of N individuals but really need to estimate N constants plus coefficients on the other independent variables
- Unobserved heterogeneity not identified
- Note:
  - There are some advanced models that can introduce heterogeneity in a purely cross section setting (random parameter models)

Partial Heterogeneity

- The classic case is that each observation i's population regression function looks like

$$y_{it} = \mathbf{x}_{it}\beta + \underbrace{c_i + \epsilon_{it}}_{u_{it}}$$


y
OLS Estimated Relationship
True Relationship
for individual red
True Relationship
for individual blue
True Relationship
for individual green
x_{i1}
x_1

What is c_i?

- This unobserved, individual-specific effect does not vary across time
  - Ex: Labor market study: cognitive ability, drive, and determination
  - Ex: Firm productivity study: each firm's organization or managerial structure

Marginal Effects

- We continue to maintain our focus on marginal effects

$$\beta = \frac{\partial E[y_{it}|x_{it}]}{\partial x_{it}}$$

- 3 Approaches:
  - Pooled OLS
  - Random Effects
  - Fixed Effects

Method 1: Pooled OLS

- OLS (termed Pooled OLS) performs ok when
  - Independent Variables Exogenous

when $$E[\mathbf{x}'_{it} u_{it}] = 0$$

$$E[\mathbf{x}'_{it} \epsilon_{it}] = 0 \quad \& \quad E[\mathbf{x}'_{it} c_i] = 0$$

Run this Regression Using OLS

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{u}$$

$$y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1T} \\ \vdots \\ y_{i1} \\ \vdots \\ y_{iT} \\ \vdots \\ y_{N1} \\ \vdots \\ y_{NT} \end{bmatrix} \quad x = \begin{bmatrix} 1 \ldots x_{11k} \ldots x_{11K} \\ \ddots \\ 1 \ldots x_{1Tk} \ldots x_{1TK} \\ \vdots \\ 1 \ldots x_{i1k} \ldots x_{i1K} \\ \ddots \\ 1 \ldots x_{iTk} \ldots x_{iTK} \\ \vdots \\ 1 \ldots x_{N1k} \ldots x_{N1K} \\ \ddots \\ 1 \ldots x_{NTk} \ldots x_{NTK} \end{bmatrix}$$

Pooled OLS

- The approach is unbiased but consistent
- Standard errors are incorrect *and* the model is inefficient
- Instead, use robust standard errors – since we might expect there to be different variances of the errors amongst cross section units.
  - Even if the method ignores possible correlations or errors within cross section units.
- This is a useful benchmark model, but usually we can do better

Method 2: Random Effects

- Exploiting what we know about the error structure
  - There must be correlation within cross section units in the error structure:

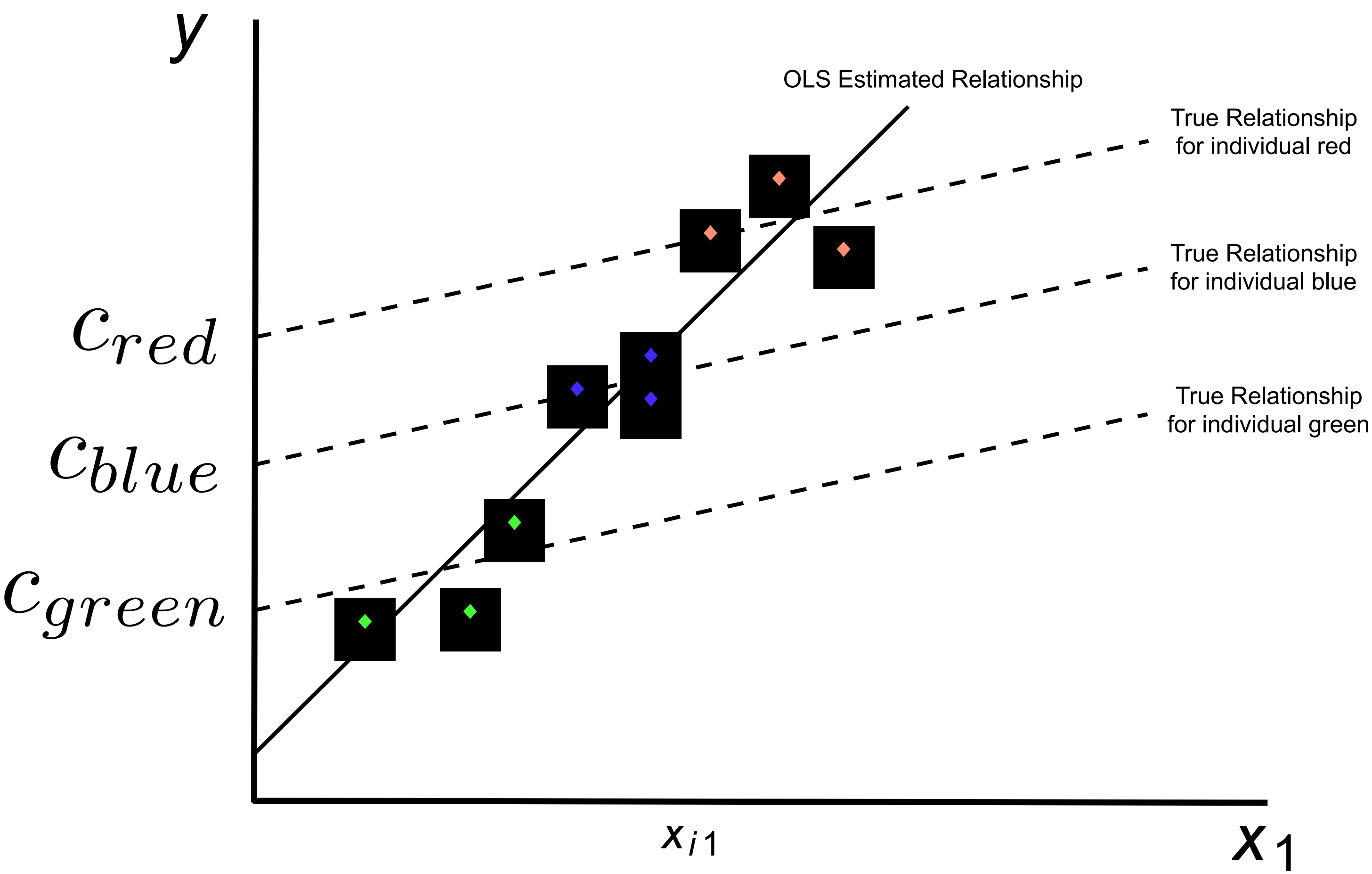
$$
\begin{aligned}
E[u_{it}u_{it+1}] &= E[(\epsilon_{it} + c_i)(\epsilon_{it+1} + c_i)] \\
&= E[\epsilon_{it}\epsilon_{it+1} + c_i\epsilon_{it} + c_i\epsilon_{it+1} + c_i^2] \\
&= E[c_i^2] = E[(c_i - 0)(c_i - 0)] = \sigma_c^2
\end{aligned}
$$

The Random Effects Model

- Fixes the standard errors
- Puts all the unobserved heterogeneity in the error term
- For unbiasedness, relies on the condition

$$E[\mathbf{x}'_{it} c_i] = 0$$

OLS Bias and Partial Heterogeneity



Method 3: Fixed Effects

Directly dealing with the unobservable c_i

A Two Period Example

- Now consider a different approach- we have information for each individual for two periods {t=1,2}

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{i1} \\ y_{i2} \\ \vdots \\ y_{N1} \\ y_{N2} \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} m_{11} & z_{11} \\ m_{12} & z_{12} \\ \vdots & \vdots \\ m_{i1} & z_{i1} \\ m_{i2} & z_{i2} \\ \vdots & \vdots \\ m_{N1} & z_{N1} \\ m_{N2} & z_{N2} \end{bmatrix}$$

First Difference

- The difference is

$$
\begin{aligned}
\Delta y_i = & \quad \beta_m \left(m_{i2} - m_{i1}\right) + \beta_z \left(z_{i2} - z_{i1}\right) + \left(c_i - c_i\right) + \left(\epsilon_{i2} - \epsilon_{i1}\right) \\
= & \quad \Delta x_i \beta + \Delta \epsilon_i
\end{aligned}
$$

- Notice, that the difference is only a function of the $\mathbf{x}_i$ (m and z) and c_i is not in the model anymore

For 2 period case, we now have a cross section

- One observation per cross section unit:

$$\Delta \mathbf{y} = \begin{bmatrix} y_{12} - y_{11} \\ \vdots \\ y_{i2} - y_{i1} \\ \vdots \\ y_{N2} - y_{N1} \end{bmatrix} \quad \Delta \mathbf{x} = \begin{bmatrix} m_{12} - m_{11} & z_{12} - z_{11} \\ \vdots & \vdots \\ m_{i2} - m_{i1} & z_{i2} - z_{i1} \\ \vdots & \vdots \\ m_{N2} - m_{N1} & z_{N2} - z_{N1} \end{bmatrix}$$

We have the OLS Regression

$$\mathbf{\Delta y} = \mathbf{\Delta x}\beta + \mathbf{\Delta u}$$

- Under what conditions is the OLS estimate unbiased?

$$\mathbf{b}^{OLS} = (\mathbf{\Delta x}'\mathbf{\Delta x})^{-\mathbf{1}}\mathbf{\Delta x}'\mathbf{\Delta y}$$

And we are back in the OLS World

- Or are we?
  - Consider endogeneity and the proof of unbiasedness:

$$E(\Delta x' \Delta \epsilon) = 0$$

$$E\left[(\mathbf{x}_{i2} - \mathbf{x}_{i1})'(\epsilon_{i2} - \epsilon_{i1})\right] = 0$$

$$E\left[(\mathbf{x}_2'\epsilon_2 + \mathbf{x}_1'\epsilon_1 - \mathbf{x}_1'\epsilon_2 - \mathbf{x}_2'\epsilon_1)\right] = 0$$

Need Strict Exogeneity

- Unobservable factors in one period can't be related to the observable factors in any other period:

$$\begin{aligned} E(\mathbf{x}_1' \epsilon_2) &= 0 \\ E(\mathbf{x}_2' \epsilon_1) &= 0 \end{aligned}$$

  - For example, unexplained factors in period 1 can't influence educational attainment in period 2.

Fixed Effects

- Deals with cases where unobserved heterogeneity is correlated with the independent variables
- Is unbiased, consistent, and efficient
- Does **NOT** allow the inclusion of time invariant independent variables in the analysis