Panel Data Models with Fixed Effects

Yuxin Wang, Zelin Ren

Febrary 2021

Content

- Pooled Estimator
- Fixed Effects Models
- Monte Carlo Simulations

Panel Data

- Panel data consist of observations on many individual economic units over two or more periods of time.
- Common panel data include consumption y_{it} of a household i in a period t, or income x_{it} of household i in a period t.
- We are interested in the linear relationship between x_{it} and y_{it}:

$$y_{it} = \beta x_{it} + u_{it},$$

where u_{it} is an error term.

Pooled Estimator

$$y_{it} = \beta x_{it} + u_{it}$$

The pooled estimator $\hat{\beta}$ based on the observations $\{x_{it}, y_{it}\}$ is the OLS estimator defined by

$$\hat{\beta} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x})^2\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x})(y_{it} - \bar{y})\right],$$

where

$$\bar{x} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}$$

is the overall average of the observation, and $\bar{y}$ is defined analogously.

Consistency Pooled Estimator

$$y_{it} = \beta x_{it} + u_{it}$$
$$\hat{\beta} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x})^2\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x})(y_{it} - \bar{y})\right]$$

- The pooled estimator is an OLS estimator.
- By assuming the full rank condition and that the random variables $\{x_{it}\}$ and $\{u_{it}\}$ are i.i.d. across i and that $E[u_{it}\boldsymbol{x}_i] = 0$, we can deduce that $\hat{\beta}$ converges to β in probability as $N \to \infty$ for any fixed T.

Motivation of Fixed Effects

Suppose that we have the data set y_{it} and x_{it}, where

- y_{it}: Consumption of household i in one country in period t;
- x_{it}: Income of household i in period t.

The true relationship between x_{it} and y_{it} is

$$y_{it} = \beta x_{it} + u_{it},$$

where $u_{it} = \lambda_i f_t + \epsilon_{it}$ and

- λ_i: (Demeaned) Education of employed members of household i;
- f_t: Economic situation of the country in period t;
- ϵ_{it}: Unobserved error.

Should we use pooled estimator to estimate β?

Motivation of Fixed Effects

$$y_{it} = \beta x_{it} + u_{it}, \qquad u_{it} = \lambda_i f_t + \epsilon_{it}$$

y_{it}: consumption, x_{it}: income, λ_i: education, f_t: economic situation

- The education λ_i and economic situation f_t are correlated with the income x_{it} of household.
- In particular, the regressor x_{it} is not exogenous.
- And the pooled estimator is not consistent in general.

Fixed Effects Models

The general interactive fixed effects models with r factors take the following form

$$\begin{aligned} y_{it} &= \beta x_{it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \epsilon_{it} \\ &= \beta x_{it} + \sum_{s=1}^{r} \lambda_{is} f_{ts} + \epsilon_{it}. \end{aligned}$$

- y_{it}, x_{it} are observable.
- The fixed effects λ_{is} and f_{ts} are unknown.
- We are interested in the true value of β.

An Accompanying Example

To make the notations as simple as possible, we look at the case where $r = 2$:

$$y_{it} = \beta x_{it} + \begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix}' \begin{pmatrix} f_{t1} \\ f_{t2} \end{pmatrix} + \epsilon_{it}$$

In the following, we are going to see different estimating strategy based on different forms of fixed effects.

Time Invariant Fixed Effects Model

$$y_{it} = \beta x_{it} + \begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix}' \begin{pmatrix} f_{t1} \\ f_{t2} \end{pmatrix} + \epsilon_{it}$$

By setting

$$\begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix} = \begin{pmatrix} \alpha_i \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} f_{t1} \\ f_{t2} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

the model reads

$$y_{it} = x_{it}\beta + \alpha_i + \epsilon_{it},$$

which is the time invariant fixed effects model.

Additive Fixed Effects Model

$$y_{it} = \beta x_{it} + \begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix}' \begin{pmatrix} f_{t1} \\ f_{t2} \end{pmatrix} + \epsilon_{it}$$

By setting

$$\begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix} = \begin{pmatrix} \alpha_i \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} f_{t1} \\ f_{t2} \end{pmatrix} = \begin{pmatrix} 1 \\ d_t \end{pmatrix},$$

the model reads

$$y_{it} = x_{it}\beta + \alpha_i + d_t + \epsilon_{it},$$

which is the additive fixed effects model.

Least Square Estimator with Known Numbers of Factors

$$y_{it} = \beta x_{it} + \begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix}' \begin{pmatrix} f_{t1} \\ f_{t2} \end{pmatrix} + \epsilon_{it}$$

The least square estimator $(\hat{\beta}, \hat{f}_{t1}, \hat{f}_{t2}, \hat{\lambda}_{i1}, \hat{\lambda}_{i2})$ is a minimiser of the objective function

$$\sum_{i=1}^{N} \sum_{t=1}^{T} |y_{it} - \beta x_{it} - f_{t1}\lambda_{i1} - f_{t2}\lambda_{i2}|^2 ,$$

with the condition $\hat{\boldsymbol{f}}_{\boldsymbol{t}}' \hat{\boldsymbol{f}}_{\boldsymbol{t}}/T = I$ and $\hat{\boldsymbol{\lambda}}_{\boldsymbol{i}}{}' \hat{\boldsymbol{\lambda}}_{\boldsymbol{i}}$ is diagonal.

Asymptotic Theory of the Least Squares Estimator

1. Full rank assumption
2. (i.i.d. errors) ϵ_{it} is i.i.d. with mean zero and the eighth moment;
3. (Exogeneity) ϵ_{it} is independent of $x_{it}, \boldsymbol{f}_t, \boldsymbol{\lambda}_i$;
4. The fixed effects $\boldsymbol{f}_t$ and $\boldsymbol{\lambda}_i$ have uniformly bounded eighth moment. And they obey the law of large numbers.

- The least square estimator $\hat{\beta}$ converges to the true value in probability as $N \to \infty$ and $T \to \infty$.
- There exists a symmetric matrix Ω such that

$$\sqrt{NT}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Omega).$$

Estimation Strategy

$$\mathrm{argmin}_{\beta, f_t, \lambda_i} \sum_{i=1}^{N} \sum_{t=1}^{T} |y_{it} - \beta x_{it} - f_{t1}\lambda_{i1} - f_{t2}\lambda_{i2}|^2 ,$$

- We start with some $\beta^{(0)}$. Then we can choose $\hat{\boldsymbol{\lambda}}_{\boldsymbol{i}}^{\mathbf{(1)}}$ and $\hat{\boldsymbol{f}}_{\boldsymbol{t}}^{\mathbf{(1)}}$ that minimise the residual

$$\sum_{i,t} \left(y_{it} - \beta^{(0)} x_{it} - \lambda_{i1} f_{t1} - \lambda_{i2} f_{t2} \right)^2 .$$

- Keep both $\hat{\boldsymbol{\lambda}}_{\boldsymbol{i}}^{\mathbf{(1)}}$ and $\hat{\boldsymbol{f}}_{\boldsymbol{t}}^{\mathbf{(1)}}$. Determine the minimiser $\hat{\beta}^{(1)}$ of

$$\sum_{i,t} \left(y_{it} - \beta x_{it} - \hat{\lambda}_{i1}^{(1)} \hat{f}_{t1}^{(1)} - \hat{\lambda}_{i2}^{(1)} \hat{f}_{t2}^{(1)} \right)^2 .$$

- Do the above step inductively.

Remarks: Starting Values

- The algorithm generates a sequence of random variables.
- Whether it converges to the true value depends on the starting value.
- A popular choice of starting value is the pooled estimator. It is not always optimal.

Remarks: Number of Factors

- So far we have seen the interactive effects estimator $\hat{\beta}(r)$ with known number of factors r.
- In practice, r is usually unknown.
- However, $\hat{\beta}(s)$ are still consistent if $s \geq r$, where r is the true number of factors. The estimator $\hat{\beta}(s)$ is less efficient than $\hat{\beta}(r)$.

Estimation of the Number of Factors

$$y_{it} = \beta x_{it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \epsilon_{it}$$

- The number r can be estimated consistently, if β is known.
- For every natural number s, estimate $\boldsymbol{\lambda_i^{(s)}}$ and $\boldsymbol{f_t^{(s)}}$, as s dimensional vectors.
- Calculate the residual $V(s) = y_{it} - \beta x_{it} - \boldsymbol{\hat{\lambda}_i'^{(s)}} \boldsymbol{\hat{f}_t^{(s)}}$
- Choose a function $g(N, T)$ that converges to 0 sufficiently fast as N and $T \to \infty$.
- Define the criterion function $C(s) = V(s) + sg(N, T)$.
- The quantity $\hat{r}$ that minimises C is a consistent estimator of r.

Estimation of Interactive Fixed Effects Models with Unknown Numbers of Factors

$$y_{it} = \beta x_{it} + \boldsymbol{\lambda}_i' \boldsymbol{f}_t + \epsilon_{it}$$

We can start with a large s, compute $\hat{\beta}(s)$, estimate $\hat{r}$ based on $\hat{\beta}(s)$, and compute $\hat{\beta}(\hat{r})$.

Monte Carlo Simulations

- Compare the performance of the estimators in different models.
- All the plots and tables are replicated under 100 times of simulations.

Additive Fixed Effects Model

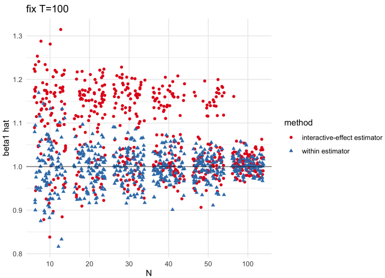
$$y_{it} = \beta x_{it} + u_{it}$$

First let us look at an additive fixed effects model

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \alpha_i + \xi_t + \epsilon_{it}.$$

- $\beta_1 = 1, \quad \beta_2 = 3.$
- Fixed effects: $\alpha_i, \xi_t \overset{\text{i.i.d}}{\sim} N(0, 1)$.
- Regressor:
  $x_{it,1} = 3 + 2\alpha_i + 2\xi_t + \eta_{it,1}$, $\eta_{it,1} \overset{\text{i.i.d}}{\sim} N(0, 1)$.
  $x_{it,2} = 3 + 2\alpha_i + 2\xi_t + \eta_{it,2}$, $\eta_{it,2} \overset{\text{i.i.d}}{\sim} N(0, 1)$.
- Error term: $\epsilon_{it} \overset{\text{i.i.d}}{\sim} N(0, 4)$.

Additive Fixed Effects Model



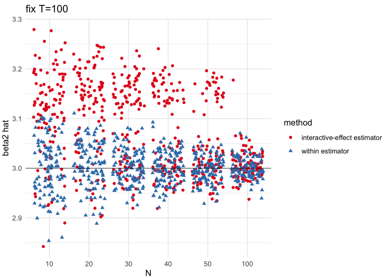


Figure: Estimation of $\beta_1 = 1$ and $\beta_2 = 3$

- Both estimators are consistent, but within estimator is more efficient than interactive-effects estimator.
- Interactive-effects estimator does not work well in small N, but it shows consistency under large sample size.

Additive Fixed Effects Model

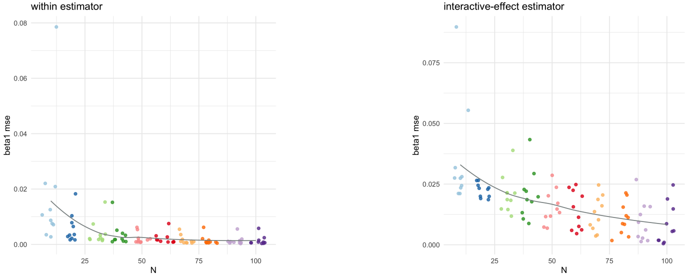


Figure: Estimation of $\beta_1 = 1$ and $\beta_2 = 3$

- Both estimators are consistent, but within estimator is more efficient than interactive-effects estimator.
- Interactive-effects estimator does not work well in small N, but it shows consistency under large sample size.

Starting Values of Interactive-effects Estimator

- A popular choice of starting value is the pooled estimator. It is not always optimal.

		pooled			two-way	
N	T	$\beta_1 = 1$	$\beta_2 = 3$		$\beta_1 = 1$	$\beta_2 = 3$
50	100	1.155	3.156		1.057	3.060
100	50	1.150	3.149		1.046	3.043
100	100	1.120	3.119		1.006	3.007

- We found that two-way estimator works better than pooled estimator in additive fixed effects model.

Interactive Fixed Effects Model

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \alpha_i + \xi_t + \epsilon_{it}$$

Now let us move to a more complex model

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + x_i \gamma + w_t \delta + \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + \epsilon_{it}.$$

- $\beta_0 = 5, \beta_1 = 1, \beta_2 = 3, \gamma = 2, \delta = 4.$
- Fixed effects: $\lambda_{i,1}, \lambda_{i,2}, f_{t,1}, f_{t,2} \overset{\text{i.i.d}}{\sim} N(0,1).$
- Regressor:
  $x_{it,1} = 1 + \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + \lambda_{i1} + \lambda_{i2} + f_{t1} + f_{t2} + \eta_{it,1}.$
  $x_{it,2} = 1 + \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + \lambda_{i1} + \lambda_{i2} + f_{t1} + f_{t2} + \eta_{it,2}.$
  $x_i = \lambda_{i1} + \lambda_{i2} + e_i,\ e_i \overset{\text{i.i.d}}{\sim} N(0,1).$
  $w_t = f_{t1} + f_{t2} + \eta_t,\ \eta_t \overset{\text{i.i.d}}{\sim} N(0,1).$
- Error term: $\epsilon_{it} \overset{\text{i.i.d}}{\sim} N(0,4).$

Interactive Fixed Effects Model

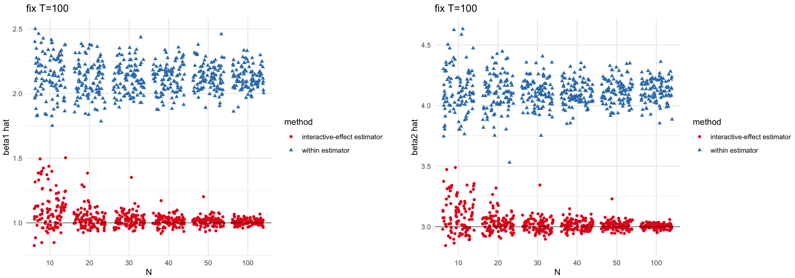


Figure: Estimation of $\beta_1 = 1$ and $\beta_2 = 3$

- Within estimator can only estimate β_1 and β_2, and it is inconsistent.
- Interactive-effects estimator can estimate all the coefficients (β_0, β_1, β_2, γ, δ) and give consistent estimations.

Interactive Fixed Effects Model

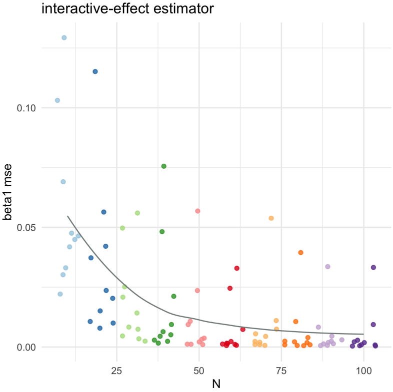


Figure: MSE of β_1

- MSE decreases as N increases.

Least Square Estimator with Unknown Numbers of Factors

- Previously in the models, we know that real factor number is equal to 2. But what would happen if we do not know the real value of r?
- Let us look at the cases where number of factors is not correctly estimated.

Least Square Estimator with Unknown Numbers of Factors

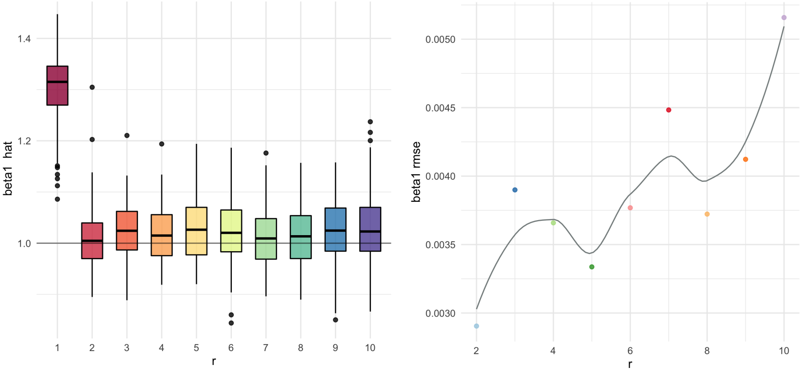


Figure: Estimation of $\beta_1 = 1$, true $r = 2$

- With fewer factor number, it will be biased and inconsistent.
- With more factor number, we have similar bias as the real one but the mean square error is higher.

Determine the Number of Factors

- We use the method introduced before to estimate the number r of factors in the interactive fixed effect model

$$y_{it} = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2} + x_i\gamma + w_t\delta + \lambda_{i1} f_{t1} + \lambda_{i2} f_{t2} + \epsilon_{it}$$

- Choose a function $g(N,T)$ that converges to 0 sufficiently fast as N and $T \to \infty$.
- Several choice of g are chosen to estimate the number of factors.
  Example I:

$$g(N,T) = \frac{N+T}{NT} \log \frac{NT}{N+T}$$

Determine the Number of Factors

- True $r = 2$.

N	T	I	II	III	IV	V	VI
100	10	8	8	8	8	8	8
100	20	5.1	4.22	6.58	1.88	1.78	1.96
100	50	**2**	**2**	2.94	**2**	**2**	**2**
100	100	**2**	**2**	3.5	**2**	**2**	**2**
10	100	8	8	8	8	8	8
20	100	5.26	4.52	6.72	1.82	1.74	1.98
50	100	**2**	**2**	2.96	**2**	**2**	**2**

- The tables shows that the estimator $\hat{r}$ is consistent.
- The biased ones are not that bad as well since they overestimate the result.

Determine the Number of Factors

- True $r = 2$.

N	T	I	II	III	IV	V	VI
100	40	**2**	**2**	3.08	1.98	1.94	**2**
100	60	**2**	**2**	2.88	**2**	**2**	**2**
200	60	**2**	**2**	**2**	**2**	**2**	**2**
500	60	**2**	**2**	**2**	**2**	**2**	**2**

- If we increase the sample size further, we see that all estimators yield the correct number of factors.