労働経済学

Lecture 7 実証研究における因果的効果の識別
固定効果モデル

張　俊超

26th May 2017

最小二乗法推定の問題点

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- $Cov(x_i, \varepsilon_i) \neq 0$
  説明変数と誤差項との間に相関があれば、OLS はバイアスがあり、一致性も持たない。x_i は教育、y_i は賃金の場合、OLS で推定された教育リターン β が常にバイアスがかかり、能力バイアス (ability bias) とも呼ばれる。
- できる限り多く（観察された）変数のコントロールしようとしても、観察できない要因は測定できないため、説明変数と誤差項との間に相関が未だに強い。（能力、やる気、運動意識、労働意欲、選好など）
- 観察できない要因を考慮することが非常に重要。

脱落変数バイアス

脱落変数バイアスを回避することは計量経済学のもっとも重要な課題の一つ。脱落変数問題を考慮しないままで、回帰分析をしても見せかけの相関を得る可能性が極めて高い。

- 例えば、政府の企業に対する雇用調整助成金が企業の雇用維持を調べたいとしよう。助成金と離職人数は負の相関があるであろう。しかし、このことは、助成金を増やせば離職人数が減ることを必ずしも意味しない。大企業であれば、資本も助成される金額も大きいことが予想されるため、助成金と離職人数の相関は、企業の資本額からくる見せかけの相関の可能性がある。(助成金、離職人数以外の、両者の間に共通する第三の要因が動いている)

パネルデータ

複数の観測個体を複数の時点に観測したデータは、パネルデータ (Panel Data) **または縦断面データ** (Longitudinal Data) **と呼ばれる。パネルデータのモデルを考えて、**

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$$

- i：**観測個体（個人、企業、国など）**
- t：**時点（年、月、４半期など）**
- **例：**y_{it} **は労働者** i **の時点** t **での対数賃金**

パネルデータ

パネルデータは二つの形式があり、Wide Form と Long Form である。

労働者	対数賃金 2016	対数賃金 2017
1	5.426	6.022
2	3.449	3.667
3	10.863	11.732

労働者	年	対数賃金
1	2016	5.426
1	2017	6.022
2	2016	3.449
2	2017	3.667
3	2016	10.863
3	2017	11.732

パネルデータ

- Wide Form：一人の労働者は一行のデータ（観測値）になり、異なる時点での変数を表すために、変数の名前に時点をつく。（2016 年の賃金、2017 年の賃金...）
- Long Form: 一人の労働者は複数行のデータになり、それぞれの行に年の情報があり、変数の名前には時点がついてない。（賃金...）
- Wide Form と Long Form は変換可能。Stata のreshape コマンドで変換できる。

パネルデータ

それぞれのメリットとデメリット

- Wide Form：異なる時点での変数の計算しやすい、データをチェックしやすい。一応パネルデータだが、パネル構造を使えないため、分析しにくい。

- Long Form: 分析しやすい。ただし、異なる時点のデータの処理は面倒、パネル演算子、または by/bys コマンドを使わなければならない。（それほど難しいではない）

固定効果モデル

基本的なパネルデータのモデルである、固定効果モデルを考えて、

$$y_{it} = \alpha + \beta x_{it} + \theta_i + \epsilon_{it}$$

- y_{it}, x_{it} はデータにある。観察される。
- θ_i は固定効果であり、観察できない。これを除去したい。
- β は係数。これを推定したい。
- $\varepsilon_{it} = \theta_i + \epsilon_{it}$、個人の観測できない要因 θ_i を考慮しないと、x_{it} と ε_{it} が相関し、脱落変数バイアスがかかる。

固定効果モデル

パネルデータがあれば、時間に通じて変化しない個人の観察できない属性 θ_i を考慮した上で、脱落変数バイアスを回避できる。固定効果 θ_i は時間につれて変化しない、観測できない、説明変数 x_{it} と相関しているかもしれないものを指す。

$$y_{it} = \alpha + \beta x_{it} + \theta_i + \epsilon_{it}$$

- y_{it}, x_{it} はデータにある。観察される。
- θ_i は固定効果であり、観察できない。これを除去したい。
- β は係数。これを推定したい。
- $\varepsilon_{it} = \theta_i + \epsilon_{it}$、個人の観測できない要因 θ_i を考慮しないと、x_{it} と ε_{it} が相関し、脱落変数バイアスがかかる。

固定効果モデル

固定効果モデルを推定するために、基本的な手順は以下の通り。

1. θ_i を除去する。

2. θ_i を除去したデータで、最もらしい回帰直線を引く。

固定効果モデル

固定効果を除去するために、まず、各個人の時間を通じた平均を考える。

$$\bar{y}_i = \alpha + \beta \bar{x}_i + \theta_i + \bar{\epsilon}_i$$

この式を平均していない式から引くと、θ_i を消す。

$$y_{it} - \bar{y}_i = \beta(x_{it} - \bar{x}_i) + \epsilon_{it} - \bar{\epsilon}_i$$

$\tilde{y}_{it} = y_{it} - \bar{y}_i, \tilde{x}_{it} = x_{it} - \bar{x}_i, \tilde{\epsilon}_{it} = \epsilon_{it} - \bar{\epsilon}_i$ とすると、

$$\tilde{y}_{it} = \beta \tilde{x}_{it} + \tilde{\epsilon}_{it}$$

チルダが付いているものを新しい変数として想像してほしい、固定効果が取り除いたため、回帰分析で不偏、一致推定量が得られる。

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y$$

この $\hat{\beta}$ は Fixed Effect Estimator（固定効果推定量）、Within Group Estimator, Least Squares Dummy Variables Estimator とも呼ばれる。

固定効果モデル

個人の間では共通するが、時間を通じて変化する要素を、時間効果で表現できる。時間効果をモデルの中に入れると、

$$y_{it} = \alpha + \beta x_{it} + \theta_i + \lambda_t + \epsilon_{it}$$

個人の固定効果と時間の固定効果両方とも入った。$y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$ のような変換することで、個人の固定効果と時間の固定効果を除去できる。両方の固定効果を考慮するモデルは最も一般的に使われる固定効果モデル。

固定効果モデルを扱う実証研究の紹介

Asai, Kambayashi and Yamaguchi(JJIE 2015) では、日本の都道府県別のパネールデータを用いて、保育園定員率と母親の就業率の関係を調べた。単純な OLS での結果は

$$\hat{L}_{it} = 0.686(0.077)Capacity_{it} + Constant$$

保育園定員率と母親の就業率との間に正の相関を観測した。この相関関係には、脱落変数バイアスがかかる可能性が高い。保育員の定員率を増やすと、必ずしも母親の就業率は増えるとは言えない。

1. 都道府県ごとに、女性の労働供給・育児の慣習、文化などが潜在的に違う。保育園定員率と母親の就業率との間に正の相関は慣習・文化によるかもしれない。
2. 女性の労働供給・育児への考え方が時間を通じて変わってしまうかもしれない。

固定効果モデルを扱う実証研究の紹介

個人固定効果と時間固定効果を考慮すると、係数が負となり、統計的にも有意でなくなる。

$$\hat{L}_{it} = -0.147(0.110)Capacity_{it} + prefecture_i + year_t + other\ variables$$

- 都道府県固定効果と年固定効果を両方考慮した上で、単純な OLS の結果と結構違う。
- 結論は、日本では保育園の定員率の効果は母親の労働供給に効果はない。

双生児固定効果モデル

パネルデータを用いた固定効果モデルを紹介したが、横断面データでも使える双生児固定効果モデルがある。1 年分のデータの中、各家庭 i について、j 番目（j=1,2）の双生児がいる。個人の観察できない属性 θ_i は双子の間に共通する。

$$y_{ij} = \alpha + \beta x_{ij} + \theta_i + \epsilon_{ij}$$

各家庭について、平均をとると、

$$\bar{y}_i = \alpha + \beta \bar{x}_i + \theta_i + \bar{\epsilon}_i$$

$y_{ij} - \bar{y}_i$ により、双子の固定効果を除去できる。普通のパネルデータを用いた固定効果モデルと同様に推定できる。

双生児固定効果モデルの応用

先行研究では、双生児固定効果モデルで能力バイアスを回避して、教育が賃金に与える因果効果を分析した。以下のモデルを考える。

$$log(wage_{ij}) = \alpha + \beta Schooling_{ij} + \theta_i + \epsilon_{ij}$$

θ_i はここで、生まれつきの能力だと考えてよい。時間につれて変化しないとする。能力バイアスを考慮しないならば、θ_i は教育年数、賃金の両方に相関があり、$\hat{\beta}$ は上方バイアスがかかると予測される。家庭内で平均を取って、差分することで、観察できない「能力」を消去できる。

固定効果モデルの他の計算方法 (1) 一階差分

双子兄弟二人の推定式を考える。

$$log(wage_{i1}) = \alpha + \beta Schooling_{i1} + \theta_i + \epsilon_{i1}$$

$$log(wage_{i2}) = \alpha + \beta Schooling_{i2} + \theta_i + \epsilon_{i2}$$

- 1,2 は双子のうち、1 番目のこ、2 番目のこを意味する。一階差分（First-Difference）によって、能力バイアスを消去することもできる。
- 双生児の場合、固定効果推定量と一回差分推定量は同じ。普通のパネルデータの場合、時点が三つ以上であれば、固定効果推定量と一回差分推定量から得られた係数が違うけれど、非常に近い。二つの時点の場合、固定効果と一階差分は同じ。

$$\Delta log(wage_i) = \beta \Delta Schooling_i + \Delta \epsilon_i$$

固定効果モデルの他の計算方法 (2) 最小二乗ダミー変数推定

双子データの代わりに、一般的なパネルデータに戻る。

$$y_{it} = \alpha + \beta x_{it} + \theta_i + \epsilon_{it}$$

固定効果推定量には、観測個体（労働者）を表すダミー変数を用いた計算方法がある。まずは、労働者ごとのダミー変数を作る。

$$D1_{it} = 1 \; if \; i = 1 \; otherwise \; D1_{it} = 0$$

同様に労働者 2、労働者 3... 労働者 n について、n 個のダミー変数を作れる。ダミー変数を用いた固定効果モデルは

$$y_{it} = \sum_{i=1}^{n} \alpha_i DI_{it} + \beta x_{it} + \epsilon_{it}$$

固定効果モデルの他の計算方法 (2) 最小二乗ダミー変数推定

最小二乗ダミー変数推定で個人固定効果を考慮した上で、さらに時間効果を同時に考慮したい場合、

$$y_{it} = \sum_{i=1}^{n} \alpha_i DI_{it} + \sum_{t=1}^{T} \tau_t DT_{it} + \beta x_{it} + \epsilon_{it}$$

時間効果を表す年ダミーをさらにいれたら、推定可能。