労働経済学

Lecture 6 実証研究における因果的効果の識別
識別問題、内生性

張　俊超

18th May 2017

注意してほしいこと

1. 本講義の実証分析の部分では、教科書（労働経済学　日本評論社）と違う表記を扱う。説明しやすいため、本講義では、内生変数以外の説明変数をすべてコントロール変数として考える。
2. 課題と期末試験にどちらの表記を使っても構わない。

実証モデル

- **単回帰**

$$L_i = \alpha + \beta w_i + \varepsilon_i$$

誤差項以外、労働供給時間は賃金だけに依存する。（明らかに間違っている）

- **重回帰**

$$L_i = \alpha + \beta w_i + \gamma P_i + \delta I_i + \kappa X_i + \varepsilon_i$$

誤差項以外、労働供給時間は賃金、消費者物価、非労働所得、他の観察可能の変数に依存する。X_i は複数の変数を含めるベクトル。

条件付き期待値

回帰モデルは条件付き期待値で表現できる

$$L_i = E(L_i|w_i, P_i, I_i, X_i) + \varepsilon_i = \alpha + \beta w_i + \gamma P_i + \delta I_i + \kappa X_i + \varepsilon_i$$

$E(L_i|w_i, P_i, I_i, X_i)$ は賃金、消費者物価、非労働所得、その他の変数が一定の場合、労働供給時間 L_i の平均値。図でも説明できる (板書だけ)。

実証モデル

一般的に、実証分析では、一つの内生変数を着目して、その変数が従属変数に与える効果を推定する。仮に、賃金が労働供給時間に与える効果を見るために、

$$L_i = \alpha + \beta w_i + \tau T_i + \varepsilon_i \qquad (1)$$

- w_i 賃金は内生変数となる。(理論モデルでの内生変数と違う)
- T_i はコントロール変数のベクトル、データ上観察可能な、P_i, I_i, X_i などを含む。説明を簡単化するために、これからは (1) 式のような表記を扱う。(教科書と違う)

識別問題と識別戦略

- **識別問題**

  データから、推定したい未知のパラメータ（β、τ など）を一意的に定めることができるどうかという理論的問題です。

- **識別戦略**

  未知のパラメータ（β、τ など）を一意的に定めるための統計的方法。

データの種類

- 実験データ

  自然科学分野によく使われる。無作為に処置群、対照群を抽出し、その二つのグループの間の平均の差を比較する。実験は完璧な場合、処置変数以外のすべての変数は一定のまま、識別戦略は不要。(教科書では、実験データでの識別戦略も説明したが、実験が完璧でない時の統計的方法だと考えてよい)

- 観察データ

  非実験データ。観察データは労働経済学において、賃金、非労働所得、消費者物価などの観察できるものを記録したデータ。特徴１、記録したデータは、労働者が自分が選択した結果。特徴２、一つの変数が変わる時に、一般的に、その他の変数も同時に変わる。

観察データにおける識別問題

単回帰を考えて、賃金が労働供給時間に与える効果を見る

$$L_i = \alpha + \beta w_i + \varepsilon_i$$

- 最小二乗法（OLS）で推定した $\hat{\beta}$ は BLUE 推定量であるために、五つの仮定を満たさないといけない。
- その中、$Cov(w_i, \varepsilon_i) = 0$ という仮定が極めて重要。$Cov(w_i, \varepsilon_i) = 0$ が満たされない場合、OLS は不偏推定量ではなく、一致推定量でもない。
- $\hat{\beta} = \frac{\sum(w_i - \bar{w})(L_i - \bar{L})}{\sum(w_i - \bar{w})^2} = \beta + \frac{\sum(w_i - \bar{w})\varepsilon_i}{\sum(w_i - \bar{w})^2}$
- $Cov(w_i, \varepsilon_i) = 0$ が満たさない場合、$E(\hat{\beta}) = \beta + E(\frac{\sum(w_i - \bar{w})\varepsilon_i}{\sum(w_i - \bar{w})^2})$ のため、第二項は 0 にならず、不偏ではない。
- 一致性について、$plim\hat{\beta} = \beta + \frac{Cov(w_i, \varepsilon_i)}{Var(w_i)}$ のため、第二項は 0 にならず、一致ではない。

内生性

説明変数と誤差項との間の相関は内生性と呼ぶ。一般的に、以下の三種類の内生性がある。

- 脱落変数バイアス
- サンプルセレクッションバイアス
- 測定誤差バイアス

脱落変数バイアス

$$L_i = \alpha_0 + \beta_0 w_i + \varepsilon_i$$

$$L_i = \alpha_1 + \beta_1 w_i + \tau_1 T_i + u_i$$

上の式は間違って、下の式は正しいとする。下の式から見れば、労働供給時間は、賃金と T_i に含むいろんな変数で正しく解釈できる。

上の式は、入れるべき変数 T_i ベクトルを脱落し、w_i と ε_i の間に相関がある。$Cov(w_i, \varepsilon_i) = 0$ は満たされない。

しかし、現実には、すべての変数を T_i ベクトルに入れるのが難しい。観測できない変数もある。(やる気、野心など)

バイアスの方向

$Cov(w_i, \varepsilon_i) = 0$ が満たされない以下の推定式を考える

$$L_i = \alpha_0 + \beta_0 w_i + \varepsilon_i$$

バイアスの方向は $Cov(L_i, \varepsilon_i)$ と $Cov(w_i, \varepsilon_i)$ から予測できる。
上方バイアス：$\hat{\beta}_0 > \beta_0$
下方バイアス：$\hat{\beta}_0 < \beta_0$
表で説明。（板書だけ）

バイアスの方向: 例 (1) 出席と期末試験の得点

真のモデルが以下の式とする。得点は出席と勉強時間に依存する。

$$Score_i = \beta_0 + \beta_1 Attend_i + \beta_2 Study_i + u_i$$

出席は観測できるが、個人の勉強時間はデータ分析者にとって一般的に観測不可能。勉強時間を真のモデルから脱落して推定すると、

$$\hat{Score}_i = \hat{\beta}_0 + \hat{\beta}_1 Attend_i$$

- $\hat{\beta}_1 > 0$ であれば、出席することが得点に正の効果があると意味する。これが正しい？
- 勉強時間 $Study_i$ と得点との間に、正の相関。（得点は出席と勉強時間だけに依存するので、個人の能力差がない。）
- 勉強時間 $Study_i$ と出席との間に、正の相関。
- $\hat{\beta}_1 > \beta_1$：上方バイアスがかかる。
- $\beta_1 > 0$ は保証できない。なぜ？

バイアスの方向: 例 (2) 教育と犯罪

真のモデルが以下の式とする。犯罪は教育と家庭環境に依存する。

$$Crime_i = \beta_0 + \beta_1 Educ_i + \beta_2 Family_i + u_i$$

教育は観測できるが、家庭環境はデータ分析者にとって一般的に観測不可能。家庭環境を真のモデルから脱落して推定すると、

$$\hat{Crime}_i = \hat{\beta}_0 + \hat{\beta}_1 Educ_i$$

- $\hat{\beta}_1 < 0$ であれば、教育を受けることが犯罪に負の効果があると意味する。これが正しい？
- 家庭環境 $Family_i$ と犯罪との間に、負の相関。（お金持ちの犯罪コストが高いから)
- 家庭環境 $Family_i$ と教育との間に、正の相関。（お金持ちの教育に対する価格弾力性が低いから）
- $\hat{\beta}_1 < \beta_1$：下方バイアスがかかる。
- $\beta_1 < 0$ は保証できない。なぜ？

バイアスの方向: 例 (3) 大卒と賃金（能力差がないケース）

真のモデルが以下の式とする。賃金は大卒ダミーと経験に依存する。

$$log(wage_i) = \beta_0 + \beta_1 Undergrad_i + \beta_2 Exp_i + u_i$$

大卒するかどうかは観測できる。分析者は経験の変数を真のモデルから脱落して推定すると、

$$\hat{log}(wage_i) = \hat{\beta}_0 + \hat{\beta}_1 Undergrad_i$$

- $\hat{\beta}_1 > 0$ であれば、大卒が賃金に正の効果があると意味する。これが正しい？
- 経験 Exp_i と賃金との間に、正の相関。(自明)
- 経験 Exp_i と大卒との間に、負の相関。(経験の構造から)
- $\hat{\beta}_1 < \beta_1$：下方バイアスがかかる。
- $\beta_1 > 0$ は保証できる。なぜ？

バイアスの方向: 例 (4) 大卒と賃金（能力差があるケース）

真のモデルが以下の式とする。賃金は大卒ダミー、経験、観測できない能力に依存する。

$$log(wage_i) = \beta_0 + \beta_1 Undergrad_i + \beta_2 Exp_i + ability_i + u_i$$

大卒するかどうか、経験は観測できるとする。分析者は観測できない能力を脱落し、脱落された能力の変数は新しい誤差項 v_i に入る。

$$log(wage_i) = \beta_0 + \beta_1 Undergrad_i + \beta_2 Exp_i + v_i$$

OLS で推定すると

$$\hat{log}(wage_i) = \hat{\beta}_0 + \hat{\beta}_1 Undergrad_i + \hat{\beta}_2 Exp_i$$

バイアスの方向: 例 (4) 大卒と賃金（能力差があるケース）

- $\hat{\beta}_1 > 0$ であれば、大卒が賃金に正の効果があると意味する。これが正しい？
- 誤差項 v_i と賃金との間に、正の相関。
- 経験 Exp_i と大卒との間に、負の相関。（お金持ちの教育に対する価格弾力性が低いから）
- $\hat{\beta}_1 < \beta_1$：下方バイアスがかかる。
- $\beta_1 < 0$ は保証できない。なぜ？

サンプルセレクッションバイアス

サンプルが母集団を代表するためには、サンプルが母集団から無作為に抽出されたものでないといけない。無作為でなければ、OLS 推定量はサンプルセレクッションバイアスをともない、不偏推定量ではなくなる。観察れれる労働者のサンプル上では、何らかの観測できない要因が共通し、説明変数と誤差項との間に相関が生じる。

1. 賃金が働いている人の間しか観測できない。働いていない人は、特殊な属性を持つ。
2. サーベイデータの調査を行うとき、仕事の忙しい人が調査を拒否する傾向が高い。データに残っている人は比較的に時間の余裕がある人。
3. ... などのケースがある。調査設計上の問題はあんまり分析者にとって、解決できないが、ケース 1 のようば場合が多い。

教育リターン

教育が賃金に与える効果を見るために、以下のモデルを考えて。賃金は教育年数に依存する。u_i は誤差項。

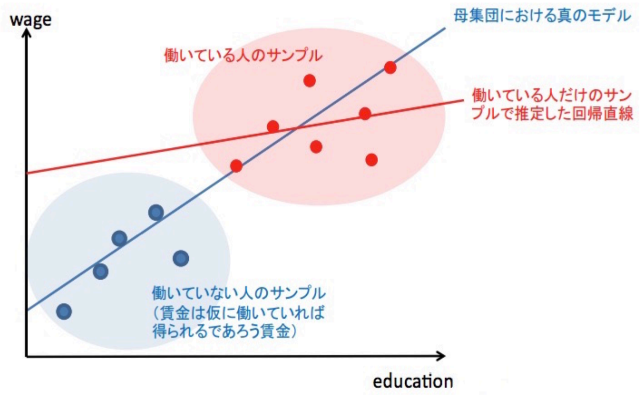
$$wage_i = \beta_0 + \beta_1 Educ_i + u_i$$

仮に、母集団からランダンムサンプルを抽出し、母集団では教育と能力は相関しない。OLS で不偏推定量を得られる。(暗黙的に、働いていない人の留保賃金が観察できるとする)

現実に、働いている人の賃金しか観測できない。つまり、データそのまま扱う場合、偏りのあるサンプルで推定してしまう。(log の問題があるので、ここでは討論しやすいために、wage の log をとっていない。賃金に関する研究では、log をとらなければならない理由は以降の講義で説明する。)

$$wage_i = \beta_0 + \beta_1 Educ_i + v_i \; if \; wage_i > 0$$

教育リターン



働いている人のサンプルで推定した教育リターンが不偏ではない。

測定誤差バイアス

観察データでは説明変数の値が正確に測定されず、誤差を伴って記録されるケースが多い。その時、説明変数と誤差項との間に、相関が出てしまう。以下の式で観測される値と誤差の関係を考えて

$$Educ_i = Educ_i^* + e_i$$

- $Educ_i$ はデータで観察される教育年数。
- $Educ_i^*$ は真の教育年数。
- e_i は測定誤差。

測定誤差バイアス

(真の) 教育年数が賃金に与える効果を見るために、以下のモデルを考えて（前述した脱落変数バイアス（観測できない能力など)、サンプルセレクッションバイアスがないとする）

$$wage_i = \beta_0 + \beta_1 Educ_i^* + u_i$$

$Educ_i^*$ は望ましいが、分析者はその真の値がわからない。データで観測できるものは、労働者が報告した教育年数のみ。推定できるモデルは、

$$wage_i = \beta_0 + \beta_1 Educ_i - \beta_1 e_i + u_i$$

$-\beta_1 e_i + u_i$ を新しい誤差項として読み替えれば、説明変数 $Educ_i$ と誤差項との間に相関が出てしまう。

古典的測定誤差

$Cov(Educ^*, e_i) = 0$ の仮定が満たされれば、古典的測定誤差と呼ばれる。つまり、測定誤差は説明変数の真の値と相関しない。古典的測定誤差であれば、$wage_i = \beta_0 + \beta_1 Educ_i - \beta_1 e_i + u_i$ おける β_1 の絶対値が過少推定され、希釈バイアスとも呼ばれる。

その他の測定誤差

$Cov(Educ^*, e_i) = 0$ の仮定が満たされない場合。

- 現実に、この仮定を満たされないケースが多い。観察データで、労働者は自分の真の教育年数より、もっと大きい数値を報告する傾向が高い。
- また、教育水準が低ければ低いほど、真の値より大きい水準を報告する傾向が高くなるでしょう。
- その時の $\hat{\beta}_1$ のバイアスの方向は？