操作変数法１：枠組み

講師：遠山祐太

最終更新：2024-12-08

はじめに

内生性問題と解決策

- 線形回帰モデル

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- 内生性問題： $Cov(x_i, \epsilon_i) \neq 0$
- 内生性問題に対する解決策の一つとして、**操作変数法（instrumental variable）**を学ぶ

授業の流れ

- パート１：基本＋実践
  - フレームワーク: 内生性、操作変数、二段階最小二乗法
  - Rでの実践例
- (補足) パート2: 同時性問題と需要・供給分析への応用
- (補足) パート3: 実証研究例-排出権取引制度におけるコースの定理の検証

内生性

内生性の例

- $Cov(x_k, \epsilon) = 0$ が満たされないケースを **endogeneity problem (内生性問題)**と呼ぶ。
  - また、そのような変数 x_k を **endogenous variable (内生変数)** と呼ぶ。
- 例 1 **[今日のメイン]**：欠落変数バイアス (Omitted variable bias)
- 例 2 ：同時性 (simultaneity)
  - 被説明変数と説明変数が同時に決定される。例：需要・供給曲線
  - パート 2 を参照
- 例 3 ：観測誤差 (measurement error)
  - 説明変数の観測誤差
- 例 4 ：サンプルセレクション (sample selection)
  - 例：賃金は働いている人についてのみ観察される。

例１：欠落変数バイアス

- 賃金方程式を思い出そう。

$$\begin{aligned} \log W_i &= \beta_0 + \beta_1 S_i + \beta_2 A_i + u_i \\ E[u_i | S_i, A_i] &= 0 \end{aligned}$$

  ただし W_i は賃金、 S_i は教育年数、 A_i は能力である。

- A_i を取り除き、次の回帰モデルを推定するとしよう。

$$\log W_i = \alpha_0 + \alpha_1 S_i + v_i$$

  ここで $v_i = \beta_2 A_i + \epsilon_i$ であるから、 S_i と v_i は相関していると考えられる。

説明変数の外生性の議論の難しさ。

- 説明変数と誤差項の無相関は、係数をバイアスなく推定する上で非常に重要。
- しかしながら、観察されない要因(誤差項)に関する仮定のため議論が難しい。
  - フォーマルに検定する方法はない。
  - 問：OLSで推定後、残差(residual)を計算し、それと説明変数の相関を計算する。この方法はうまくいくか？
- 欠落変数バイアスの公式の重要性。バイアスの方向を議論できる。
- どのようにして、この仮定の妥当性を確保するか？
  - 1: コントロール変数する。-> 線形回帰＆パネルデータ回帰
  - 2: 自然実験
  - 3: コントロール変数でも対処できないケース？(次スライド)

コントロール変数による内生性への対処

- 能力を代理する変数：
  - 試験のスコア、GPA、SAT、などなど
- どれだけコントロールすれば、本当に S_i が外生と言えるか？
- 賃金と教育の両方に影響するような「観察されない要因」はコントロールできない。
  - 人間行動の分析においては、「選好」や「能力」などが観察できない。

(参考) パネルデータを用いた固定効果モデル

- パネルデータがあると個人固定効果を捉えられる。

$$y_{it} = \alpha D_{it} + \beta X_{it} + \epsilon_i + \epsilon_t + \epsilon_{it}$$

- もし「時間を通じて変化する要因」 ϵ_{it} がトリートメント変数と相関しているとまずい。
- これは先週のDIDで言うところの「並行トレンド」が満たされないケース。

例２：測定誤差

- 変数が誤差を持って計測される状況
  - 例：報告の誤差・問いの誤解など
- 次の回帰を考える。

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i$$

- ここで、データで観察できる x_i は誤差を伴うとしよう。

$$x_i = x_i^* + e_i$$

  - e_i は ϵ_i と x_i^* から独立である。 (古典的な測定誤差)
  - e_i はデータに加わったノイズと捉えられる。

測定誤差がもたらす問題

- 回帰式は

$$\begin{aligned} y_i &= \beta_0 + \beta_1(x_i - e_i) + \epsilon_i \\ &= \beta_0 + \beta_1 x_i + (\epsilon_i - \beta_1 e_i) \end{aligned}$$

- x_i と誤差 $\epsilon_i - \beta_1 e_i$ に相関があり、平均独立の仮定(説明変数と誤差項の無相関)を満たさない。

- 説明変数の測定誤差は係数の過少推定をもたらす
  - **減衰バイアス（attenuation bias）**

操作変数のアイデア

操作変数法

- 簡単な例からはじめよう。

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ Cov(x_i, \epsilon_i) \neq 0$$

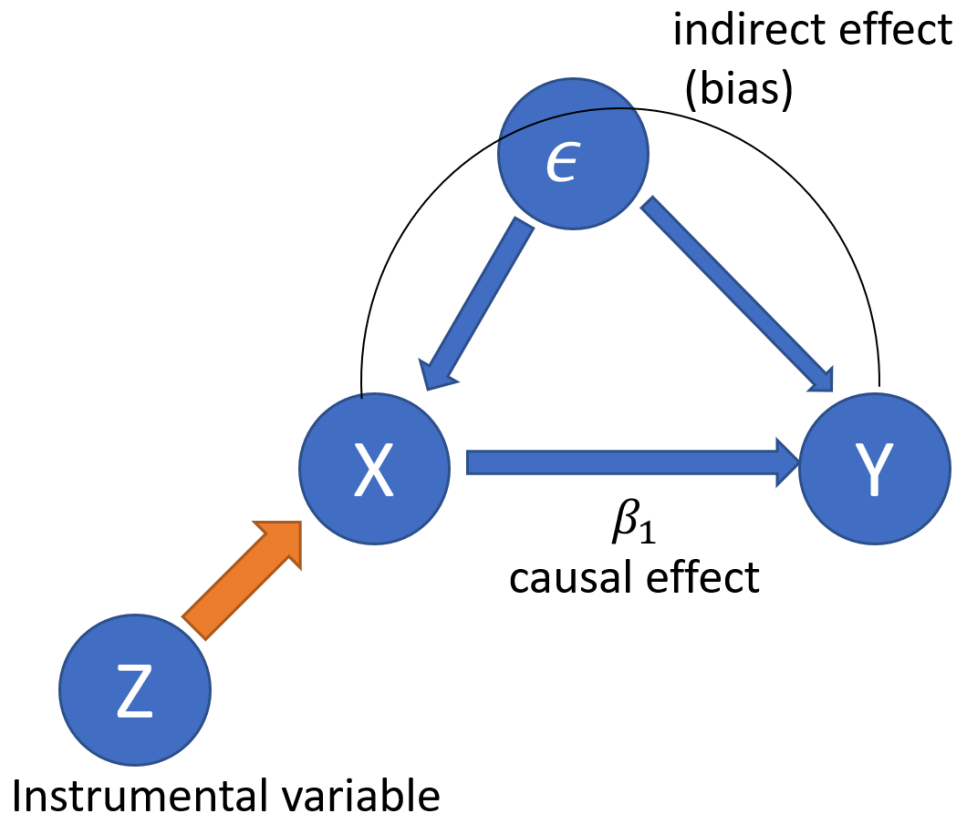
- 変数 z_i は **操作変数（instrumental variable; IV）**とよばれ、次の条件を満たす。

1. **独立性（independence）**： $Cov(z_i, \epsilon_i) = 0$ （操作変数と誤差に相関がない）
2. **関連性（relevance）**： $Cov(z_i, x_i) \neq 0$ （操作変数と内生変数に相関がある）

- アイデア：**操作変数 z_i に引き起こされる**内生変数 x_i の変動を用いて、 x_i の y_i に対する直接的な因果効果（ β_1 ）を推定する！

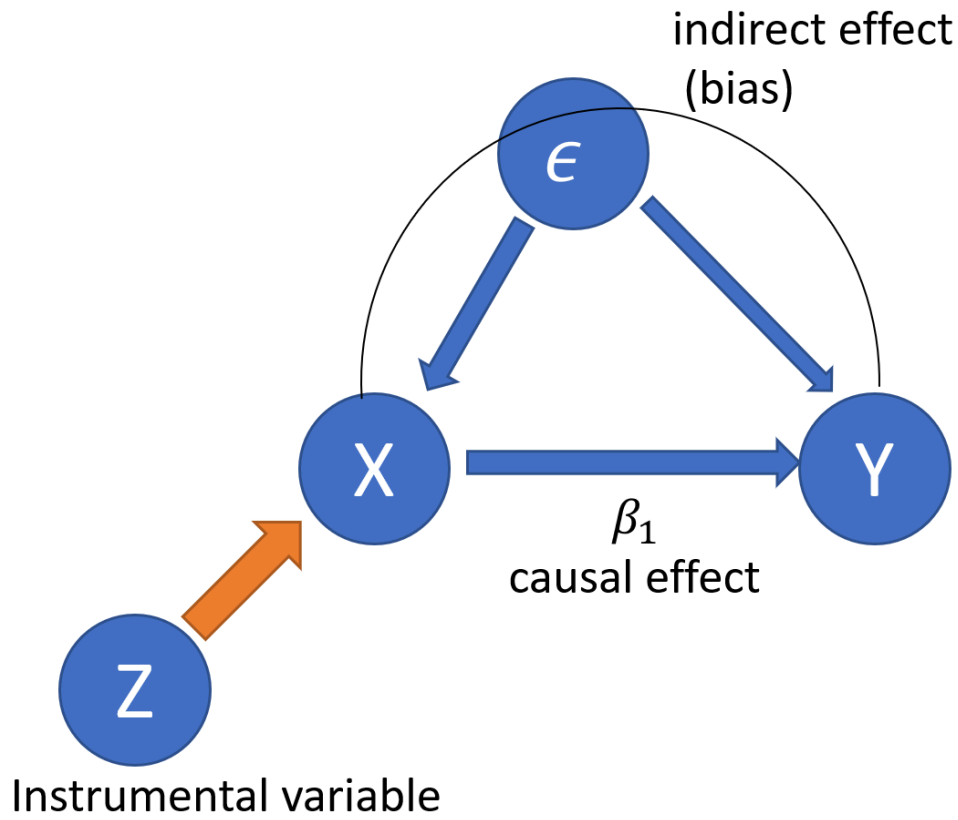
アイデア: OLSの復習

- OLS は x と y の相関を捉える。
- もし x と ϵ に相関がないならば、OLSは因果効果 β_1 を捉える。
- 相関がある場合、OLS は因果効果 β_1 と間接効果(バイアス)を捉える。



アイデア：操作変数があると・・・

- 操作変数 z を考える。
- 今、操作変数 z が動くと、 x が動くとしよう。(**関連性**の仮定)
- 一方、操作変数 z が動いても、誤差項 ϵ は変化しない (**独立性**の仮定)
- 操作変数 z によって引き起こされた x の変化が y とどのように相関するかを見ることによって、因果効果 β_1 を取り出すことができる！



(おまけ) 前スライドの補足

- 前のスライドでは、誤差項からXへの影響するという方向を考えている。
  - 例：誤差項が「生まれつきもった能力」(innate ability)、Xが教育年数
- もし誤差項を「教育によって影響を受ける能力」と解釈するとモデルの解釈が変化する。
  - モデルにおける教育の係数 β_1 は「他の要因を固定したときに(セテリス・パリバス)、教育を１年増やすことで、賃金がどれだけ変化するか？」というもの。
- なお、処置変数がどのような経路によってアウトカムに影響するかについて扱う「媒介分析」(mediator analysis) という手法が、政治学・疫学では存在する。
- また、前ページのような因果関係を捉えるグラフを有向非巡回グラフ(Directed　acyclic graph, DAG)と呼び、DAGを用いた因果推論のフレームワークもある (Judea Pearlらによる)

操作変数法とパラメータの識別

- モデル： $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- y_i と z_i の共分散をとると次のようになる。

$$Cov(y_i, z_i) = \beta_1 Cov(x_i, z_i) + Cov(\epsilon_i, z_i)$$

- 操作変数の条件により、

$$\beta_1 = \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)}$$

- 問い：操作変数の条件は、どのような役割を果たしているだろうか？

一般的な操作変数法の枠組み

内生変数と操作変数が複数ある場合

- 次のモデルを考える。

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \beta_{K+1} W_{1i} + \cdots + \beta_{K+R} W_{Ri} + \epsilon_i,$$

  - Y_i ：被説明変数
  - $X_{1i}, \ldots, X_{Ki}$：K 個の内生的（endogenous）な説明変数で、任意のk について $Cov(X_{ki}, \epsilon_i) \neq 0$
  - $W_{1i}, \ldots, W_{Ri}$ ： R 個の外生的（exogenous）な説明変数で、任意の r について $Cov(W_{ri}, \epsilon_i) = 0$
  - ϵ_i ：誤差項
  - $Z_{1i}, \ldots, Z_{Mi}$ ： M 個の操作変数
  - $\beta_0, \ldots, \beta_{K+R}$ ： $1 + K + R$ 個の未知の回帰係数

二段階最小二乗法

- 一般ケースにおける推定手法：**二段階最小二乗法（two-stage least squares; 2SLS）**
- 上述の操作変数法を特殊ケースとして包含する。
- (上級) 2SLSは一般化モーメント法 (GMM) の特殊ケースでもある。

Stage 1：第一段階（first-stage）の回帰

- それぞれの内生変数 $(X_{1i}, \ldots, X_{Ki})$ を、最小二乗法を用い、すべての操作変数 $(Z_{1i}, \ldots, Z_{Mi})$ および外生変数 $(W_{1i}, \ldots, W_{Ri})$ に回帰する。

$$X_{ki} = \pi_0 + \pi_1 W_{1i} + \cdots + \pi_R W_{R,i} + \pi_{R+1} Z_{1i} + \cdots + \pi_{R+M} Z_{Mi} + \nu_{ki}$$

- 予測値 $(\widehat{X}_{1i}, \ldots, \widehat{X}_{ki})$ を計算する。

Stage 2：第二段階（second-stage）の回帰

- 被説明変数 Y_i を、最小二乗法を用い、すべての**予測された**内生変数 $\left(\widehat{X}_{1i}, \ldots, \widehat{X}_{ki}\right)$ および外生変数 $\left(W_{1i}, \ldots, W_{ri}\right)$ に回帰する。

$$Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \cdots + \beta_K \hat{X}_{Ki} + \beta_{K+1} W_{1i} + \cdots + \beta_{K+R} W_{Ri} + \epsilon_i$$

- 上の手続きにより、係数の二段階最小二乗推定値 $\widehat{\beta}_0^{TSLS}, \ldots, \widehat{\beta}_{k+r}^{TSLS}$ を得る。
- 2SLS推定量は一定の条件下で一致性＆漸近正規性を持つ。

二段階最小二乗法の仕組み

- 単純な例として、内生変数と操作変数がともに一つの場合を考えよう。
- 第一段階では、モデル

$$x_i = \pi_0 + \pi_1 z_i + v_i$$

を最小二乗法で推定し、推定値 $\widehat{x}_i = \widehat{\pi}_0 + \widehat{\pi}_1 z_i$ を得る。

- 第二段階では、モデル

$$y_i = \beta_0 + \beta_1 \widehat{x}_i + \epsilon_i$$

を推定する。

- 内生変数の予測値 $\widehat{x}_i$ は操作変数 z_i にのみ依存し、（第二段階の）誤差項 ϵ_i と相関しないので、第二段階において β_1 をバイアスなく推定することができる。

(補足)なぜ外生変数を第一段階にも入れるのか？

- 元のモデル

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \epsilon_i$$

- 1st stageのモデル

$$x_i = \pi_0 + \pi_1 z_i + \pi_2 w_i + v_i$$

線形射影として $E[v_i] = 0, E[z_i v_i] = 0, E[w_i v_i] = 0$ を満たす係数 π_0, π_1, π_2が得られる。

- yを1st stage のFittedに回帰するモデルは

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \beta_2 w_i + (\epsilon_i + \beta_1 v_i)$$

このモデルで係数をConsistentに推定するには、 ϵ_i のみならず v_i も変数と無相関。

- もし第一段階で w を外すと w が v に含まれてしまうので、二段階目が正しく機能しない。

操作変数の条件

必要条件

- 操作変数の数を M ,内生変数の数を K とする。
- 操作変数の数に応じて、三パターンに分けることができる。
  1. 過剰識別（over-identification）： $M > K$
  2. 丁度識別（just-identification）： $M = K$
  3. 過小識別（under-identification）： $M < K$
- 必要条件は $M \geq K$ である。
  - 内生変数以上の操作変数が必要だということ。

十分条件

- 一般的な枠組みでは、操作変数が妥当であるための十分条件は次のように与えられる。

  1. **独立性**：任意の m について $Cov(Z_{mi}, \epsilon_i) = 0$
  2. **関連性**：第二段階の回帰において、変数

$$\left(\widehat{X}_{1i}, \ldots, \widehat{X}_{Ki}, W_{1i}, \ldots, W_{Ri}, 1\right)$$

  が**完全な多重共線性をもたない。**

- 関連性の条件は何を意味するだろうか？

一変数回帰における関連性

- 既出の単純な例に戻ろう。第一段階は

$$x_i = \pi_0 + \pi_1 z_i + v_i$$

であり、第二段階は

$$y_i = \beta_0 + \beta_1 \widehat{x}_i + \epsilon_i$$

- もし $\pi_1 = 0$ すなわち $\widehat{x}_i = \pi_0$ ならば、第二段階の説明変数は完全な多重共線性をもつ。

内生変数一個のケースにおける関連性

- 内生変数 X_{1i} に対する第一段階回帰モデルは次のように書ける。

$$X_{1i} = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_M Z_{Mi} + \pi_{M+1} W_{1i} + \cdots + \pi_{M+R} W_{Ri}$$

  ここで、係数 $\pi_1, \cdots, \pi_M$ のうち少なくとも一つは 0 でない。

- 直感的には、**外生変数を統制した後でも、操作変数が内生変数と相関していなければならない。**

操作変数の妥当性チェック：関連性

- 操作変数が内生変数の変動をあまり説明しないとき、**弱操作変数（weak instrument）**の問題が生じているという。
- 弱操作変数があると、
  1. 正確でない推定をもたらす。（標準誤差が大きくなる）
  2. サンプルサイズが大きくても、漸近分布が正規分布から乖離する。

関連性の条件のめやす

- 内生変数 X_{1i} が一つの場合を考える。
- 第一段階の回帰は

$$X_k = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_M Z_{Mi} + \pi_{M+1} W_{1i} + \cdots + \pi_{M+R} W_{Ri}$$

- F 検定を実施する。

$$H_0 : \pi_1 = \cdots = \pi_M = 0$$
$$H_1 : otherwise$$

- もし帰無仮説が棄却されるならば、弱操作変数の懸念はおそらくない。

独立性（操作変数の外生性）

- これもまた、データから直接検定できない仮定である。
  - OLSにおけるエラーと説明変数の無相関や、DIDの並行トレンドと同様。
- 仮定の正当化は状況・コンテキスト次第。ここに自然実験的な議論が入ってくる。

例１：賃金方程式

はじめに：パッケージの読み込み

```
library(summarytools)
library(npsf)
library(tidyverse)
library(fixest)
```

例１：賃金方程式

- Wooldridge "*Introductory Econometrics: A Modern Approach*" に付随する、クロスセクションの労働力参加データ `MROZ` を用いる。
- データの出典：Mroz, Thomas A. (1987). "The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions." *Econometrica*
- 詳しくはこちら

```
# npsf パッケージの mrozデータを読み込む
data(mroz)
data <- mroz
```

記述統計

```
descr(data,
      stats = c("mean", "sd", "min", "q1", "med", "q3", "max"),
      transpose = T)
```

```
## Descriptive Statistics
## data
## N: 753
##
##                    Mean     Std.Dev        Min          Q1      Median          Q3         Max
## -------------- ---------- ---------- --------- ---------- ---------- ---------- ----------
##             age      42.54       8.07     30.00      36.00      43.00      49.00      60.00
##            city       0.64       0.48      0.00       0.00       1.00       1.00       1.00
##            educ      12.29       2.28      5.00      12.00      12.00      13.00      17.00
##           exper      10.63       8.07      0.00       4.00       9.00      15.00      45.00
##          faminc   23080.59   12190.20   1500.00   15428.00   20880.00   28200.00   96000.00
##       fatheduc       8.81       3.57      0.00       7.00       7.00      12.00      17.00
##          hours     740.58     871.31      0.00       0.00     288.00    1516.00    4950.00
##         husage      45.12       8.06     30.00      38.00      46.00      52.00      60.00
##        huseduc      12.49       3.02      3.00      11.00      12.00      15.00      17.00
##         hushrs    2267.27     595.57    175.00    1928.00    2164.00    2553.00    5010.00
##        huswage       7.48       4.23      0.41       4.79       6.98       9.17      40.51
##           inlf       0.57       0.50      0.00       0.00       1.00       1.00       1.00
```

賃金方程式 (wage regression)

- モデル：

$$\log(w_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + \epsilon_i$$

  - w_i: 賃金
  - $educ_i$ ：各個人の教育年数
  - $exper_i$ ：各個人の労働市場経験年数

- $exper_i$ は外生変数、 $educ_i$ を内生変数だと扱う。

- $educ_i$ に対する操作変数として、父母の教育年数 $fathereduc_i$ と $mothereduc_i$ を用いる。
- 操作変数の妥当性については後ほど

OLSとIVでの分析

```
# データの抽出・整形
data <- data %>%
    filter( wage > 0 ) %>%
    mutate( lwage = log(wage),
            expersq = exper^2) %>%
    select(lwage, educ, exper, expersq, motheduc, fatheduc)

# 最小二乗法
result_OLS <- feols(lwage ~ educ + exper + expersq | 0,
                    se = "hetero", # heteroskedasticity-robust
                    data = data)

# 操作変数法
result_IV <- feols(lwage ~ exper + expersq | 0 | educ ~ fatheduc + motheduc,
                   se = "hetero",  # heteroskedasticity-robust
                   data = data)
```

結果の表示

```
# 結果の表示
results <- list()
results[["OLS"]] <- result_OLS
results[["IV"]] <- result_IV

table_result <- etable(results, signifCode = NA,
                       dict = c(educ = "教育年数",
                                exper = "経験年数",
                                expersq = "経験年数の二乗",
                                `(Intercept)` = "定数項"),
                       fitstat = c("r2", "n", "ivwald" ),
                       digits = 2,
                       digits.stats = 2
)
```

```
## Warning in etable(results, signifCode = NA, dict = c(educ = "教育年数", : The
## argument 'signifCode' is deprecated. Please use 'signif.code' instead.
```

推定結果

```
print(table_result)
```

```
##                                      OLS                IV
## Dependent Var.:                    lwage             lwage
##
## 定数項                    -0.52** (0.20)       0.05 (0.43)
## 教育年数                 0.11*** (0.01)      0.06. (0.03)
## 経験年数                  0.04** (0.01)     0.04** (0.02)
## 経験年数の二乗  -0.0008. (0.0004) -0.0009* (0.0004)
## ____________________ _________________ _________________
## S.E. type                Heteroskeda.-rob. Heteroskeda.-rob.
## R2                                  0.16              0.14
## Observations                         428               428
## Wald (1st stage), 教育年数                --              49.5
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

第一段階の回帰

```
# 第一段階における操作変数のF検定
fitstat(result_IV, "ivwald")
```

```
## Wald (1st stage), educ: stat = 49.5, p < 2.2e-16, on 2 and 423 DoF, VCOV: Heteroskedasticity-robust.
```

手作業で第一段階のF検定

```
result_1st <- feols(educ ~ exper + expersq + fatheduc + motheduc | 0,
                    se = "hetero", # heteroskedasticity-robust
                    data = data)

wald(result_1st, keep = c("fatheduc", "motheduc"))
```

```
## Wald test, H0: joint nullity of fatheduc and motheduc
##  stat = 49.5, p-value < 2.2e-16, on 2 and 423 DoF, VCOV: Heteroskedasticity-robust.
```

操作変数の議論

- 教育の操作変数として家族の背景情報は妥当なものか？
- **関連性**：第一段階の検定をすればよい。
- **独立性**：疑念が残る。親の学歴は、幼少期の育て方の質を通じて子の能力と相関しうる。
- それでも、この操作変数は欠落変数バイアスを（完全に除去しないまでも）軽減できる。