

操作変数法

Instrumental Variables Method

誤差項と説明変数の相関

- 誤差項と説明変数の間に相関がある場合には、最小二乗法による推定値はバイアスを持つ
→ 操作変数法(Instrumental Variables Method)という手法で回帰分析を行う
- 誤差項と説明変数の相関の原因
 1. 説明変数の誤差
 2. 説明変数が内生変数（連立方程式モデル）
 3. 欠落変数の影響
- 以下では、まず、上記の誤差項と説明変数の相関の原因について説明する

説明変数の誤差

真のモデル

$$y_i = \alpha + \beta x_i^* + u_i \quad (1)$$

ただし，説明変数 x_i^* は観察できない。そのかわり x_i が観察できる。 x_i^* ， x_i は次のような関係にあるとする(つまり説明変数が誤差をもって観測される)

$$x_i = x_i^* + v_i \quad (2)$$

$$E(v_i) = 0, \text{cov}(u_i, v_j) = 0 \quad \text{for all } i, j$$

(2)を(1)に代入すると

$$\begin{aligned} y_i &= \alpha + \beta(x_i - v_i) + u_i = \alpha + \beta x_i + (u_i - \beta v_i) \\ &\equiv \alpha + \beta x_i + w_i \end{aligned}$$

誤差項 w_i の期待値は0，分散は一定。しかし， w_i と x_i には相関がある(次ページ参照) → OLSの前提が満たされない

説明変数の誤差(2)

説明変数の誤差→誤差項と説明変数の相関→OLS推定量（最小二乗推定量）にはバイアス

$$\text{OLS推定量} \quad b = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'w$$

特に単回帰の場合は次の通りになる(reg.pdfを参照せよ)

$$\begin{aligned} \text{plim } b &= \beta + \frac{\text{cov}(x, w)}{\text{var}(x)} = \beta + \frac{\text{cov}(x^* + v, u - \beta v)}{\text{var}(x^* + v)} \\ &= \beta - \frac{\beta \sigma_v^2}{\sigma_{x^*}^2 + \sigma_v^2} = \beta \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2} \right) < \beta \end{aligned}$$

plimはprobability limitの意味。plim bはサンプル数が限りなく大きくなる時のbの確率分布の収束値。上の式はOLS推定量bはバイアスを持つことを示している。

OLSの前提が満たされている場合には,

$$E(b) = \beta, \quad \text{var}(b) = \sigma^2/S_{xx} \rightarrow 0 \text{ (as } n \rightarrow \infty)$$

が成り立つので, $\text{plim } b = \beta$ が成り立つ（一貫性 consistencyがあるという）

説明変数の誤差(3)

例) 恒常所得仮説

$$C_i = kY_i^P + u_i$$

$$Y_i = Y_i^P + Y_i^T$$

$$E[Y_i^T] = 0, \text{cov}(Y_i^P, Y_i^T) = \text{cov}(Y_i^T, u_i) = 0$$

Y : 観察される所得

Y^P : 恒常所得 (観察不可能)

Y^T : 変動所得

- 恒常所得仮説によれば, 消費は観察不可能な変数である恒常所得によって決まる (k はほぼ1に近い)。
- しかし, 実際に観測できるのは Y で, Y は誤差 (変動所得) を含む
- Y を説明変数にして消費関数を推計すると, 消費性向はケインズ型消費関数の消費性向 (0.6~0.7)に近い値が推定されてしまう。しかし, それはケインズ型消費関数が妥当だということを意味するものではない。

連立方程式モデル（説明変数の内生性）

例）Keynes型マクロモデルを考える

$$C = \alpha + \beta Y + u \quad (1)$$

$$Y = C + I + G \quad (2)$$

上のモデルから Y の均衡値を求めると

$$Y = \frac{1}{1-\beta} (\alpha + I + G) + \frac{1}{1-\beta} u \quad (3)$$

となり， Y は(1),(2)の連立方程式の解として求まる

→(1)のケインズ型消費関数の Y は内生変数

→ Y は(3)式のように u の関数なので，ケインズ型消費関数の誤差項 u と説明変数 Y は相関を持つ

→回帰分析の前提が満たされない

→(1)式をOLSで推定すると， Y の係数の推定値はバイアスを持つ

連立方程式 (2)

社会資本の生産性

社会資本の生産力効果の推計→次のような方程式を推計

$$\ln Y_i = \alpha + \beta_1 \ln L_i + \beta_2 \ln K_i^P + \beta_3 \ln K_i^G + \gamma Z_i + u_i$$

Y :県民所得, L :労働力, K^P :民間資本, K^G :社会資本

社会資本の生産性に関する多くの研究では, 低い (場合によってはマイナスの) β_3 の値が報告されている

しかし, K^G は政治的に決定されているかもしれない (過疎地や低所得地域に手厚い再分配が日本の公共投資の特徴) → K^G は内生変数で, 次のように決まるかもしれない

$$\ln K_i^G = \delta_0 + \delta_1 \ln Y_i + \delta_2 \ln POP_i + \cdots + v_i$$

Y :所得, POP :人口

→ OLSによる β_3 の値はバイアスを持つ

欠落変数 omitted variables

例) 賃金方程式

真のモデルが次の方程式で表されとする

$$\ln(wage) = a + b \cdot educ + c \cdot ability + u$$

educ: 教育年数, *ability*: 能力

しかし, *ability*が観測できず,

$$\ln(wage) = a + b \cdot educ + v$$

を推定したとすると, 誤差項*v*には欠落した変数の*ability*の影響が含まれる

この場合, 高い能力→高学歴が成立→*ability*と*educ*には正の相関がある→*v*と*educ*には相関→OLSの前提が満たされず, 賃金方程式の係数*b*はバイアスを持って推計されてしまう

(教育の影響を過大に評価してしまうので, 今のケースではプラスのバイアスがある)

操作変数法

Instrumental Variable Method

- 説明変数と誤差項に相関がある場合を考える

$$y = \alpha + \beta x + u, \quad \text{cov}(x, u) \neq 0$$

- 操作変数 z は次の性質を満たすような変数である

$$\text{cov}(z, u) = 0$$

$$\text{cov}(z, x) \neq 0$$

- 操作変数 z は誤差項と相関が無い
 - 操作変数 z は説明変数 x と相関がある
- IV法の推定方法

$$\hat{\beta} = \frac{\text{cov}(z, y)}{\text{cov}(z, x)} = \frac{\text{cov}(z, \alpha + \beta x + u)}{\text{cov}(z, x)} = \beta + \frac{\text{cov}(z, u)}{\text{cov}(z, x)} \rightarrow \beta$$

- 一方OLSは

$$\hat{\beta}_{OLS} = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{\text{cov}(x, \alpha + \beta x + u)}{\text{var}(x)} = \beta + \frac{\text{cov}(x, u)}{\text{var}(x)}$$

操作変数法(2)

- 賃金方程式の場合

$$\ln(wage) = a + b \cdot educ + u$$

誤差項 $u \rightarrow$ 能力を反映しているとする

- 操作変数として望ましい性質

(a) u (能力等) と無相関

(b) $educ$ と相関がある

- どの変数が望ましいかはわからない。(a)の $cov(u, z)=0$ をテストすることはできない。しかし、(b)は確かめることができる。

- 操作変数の候補

- 誕生日 (b)が満たされない
- 父親・母親の学歴 (a)が満たされない (遺伝的要因)
- 兄弟の数 (a) も (b)も満たされる?
 - 兄弟の数 $\rightarrow educ$ と相関あり (兄弟数が多い $\rightarrow$ 学費がかさむ $\rightarrow$ マイナスの相関), 一方で能力と無相関と想定される

操作変数法(3)

重回帰の場合

$$y = X\beta + u$$

操作変数の満たすべき条件

$$\text{plim} \left(\frac{1}{n} Z'X \right) \neq 0, \text{plim} \left(\frac{1}{n} Z'u \right) = 0$$

操作変数法とOLSによる推定量の比較

$$b_{IV} = (Z'X)^{-1}Z'y = \beta + (Z'X)^{-1}Z'u$$

$$b_{OLS} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u$$

誤差項と説明変数に相関がある場合、操作変数法による推定量はバイアスを持たない（標本数が大きいとき；もちろん、誤差項と相関を持たない操作変数が選べればの話）。一方、OLSの推定量はバイアスを持つ。

2段階最小二乗法 Two Stage Least Square Method

2段階最小二乗法

$$y_1 = \alpha + \beta_1 y_2 + \beta_2 x + u$$

上のモデルで y_2 が内生変数である場合、 y_2 をそのまま使うのではなく、 y_2 を外生変数（操作変数）に回帰させ、その予測値 $\hat{y}_2$ を説明変数として用い、回帰分析を行う方法。

すなわち、次の方程式を推計する。

$$y_1 = \alpha + \beta_1 \hat{y}_2 + \beta_2 x + u$$

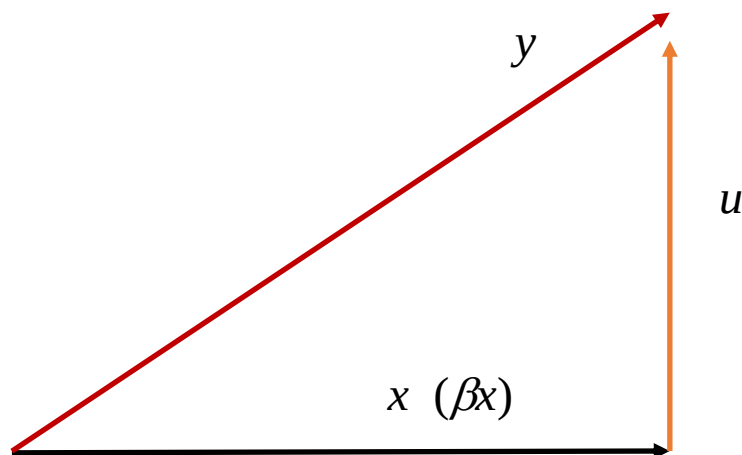
$\hat{y}_2$: y_2 を外生変数（操作変数）に回帰させた場合の予測値

- 社会資本の生産性の計測の例
 - 社会資本ストックは政治的に決定される内生変数
 - 社会資本ストックを決める政治的ルールを計測し（交付税，補助金，人口，面積，所得等），その予測値を説明変数として用いる
- 2段階最小二乗法は操作変数法の一つ
 - 多くの統計ソフトでは，操作変数を指定すれば， $\hat{y}_2$ を自動的に計算してTSLSの結果を報告してくれる

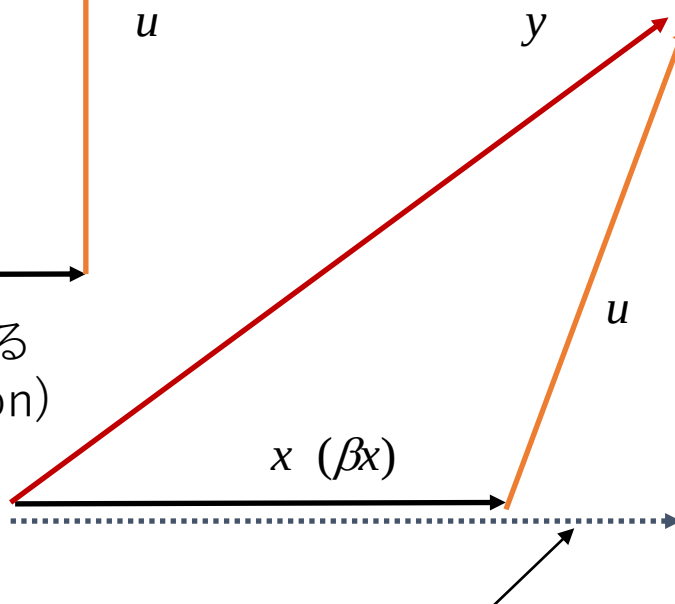
操作変数法 (4)

OLSの前提

説明変数 x と誤差項 u は独立（直交する）



説明変数と誤差項に相関がある場合

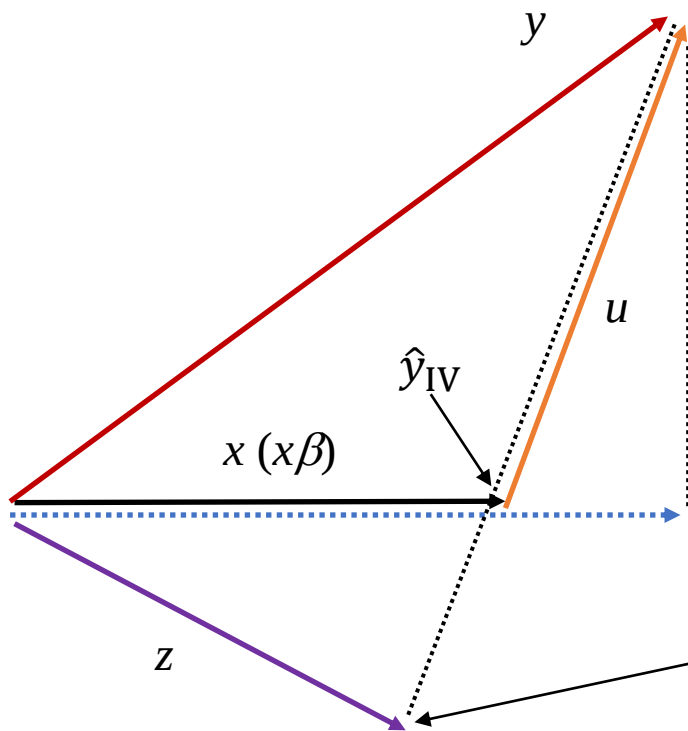


数学的には、 y を x に回帰させる
→ y の x 方向への射影(projection)

誤差項と説明変数に相関がある場合のOLS推定量には バイアスが生じる

操作変数法 (5)

説明変数 x と誤差項 u に相関がある場合



誤差項 u と無相関（直交する）な変数 z を用い，変数 z （操作変数）の方向に射影するとバイアスの無い推定量が得られる

ただし，理想的な操作変数を見つけるのは容易ではない → z と u が無相関だという仮説検定はできない

y を z に回帰した場合のOLS推定値

操作変数は，誤差項と無相関， x と相関があることが必要
操作変数法と2段階最小二乗法の同等性

Rでの操作変数法

- `ivreg()`を用いる（パッケージAERが必要）

パッケージAERがインストールされていない場合はインストールする

`library("AER")`でロード

`ivreg(y~x1+x2+x3 | z1 + x2 +x3)`

x1に内生性，操作変数としてz1,x2,x3を選んだ場合

- `tsls()`を用いる（パッケージsemが必要）

パッケージsemをインストール

`library("sem")`でパッケージsemをロード

`tsls(y~x1+x2+x3, ~z1+ x2 +x3)`

`tsls( モデル式 , 操作変数のリスト )`

操作変数のリストは `~ z1 + x2 + x3` のように書く

注意

- 操作変数 z の選択基準

- 説明変数 x と相関

- これはデータからチェックできる
 - x を z に回帰させて有意がどうかみればよい

- 誤差項と無相関

- データからチェックできない
 - 相関が無いという想定がもっともらしい理由が必要

- 操作変数の数

- 推定する方程式の説明変数と（少なくとも）同じ数を指定
 - 説明変数が $x_1, x_2, \dots, x_k$ で、 x_1 に内生性があり、他の説明変数は外生変数だとした場合、操作変数として z 以外に $x_2, \dots, x_k$ を指定する（前のページを参照）

mroz.xls

- 既婚女性の教育の収益率の分析
- 基本モデル：賃金の対数値($\ln wage$)を教育年数($educ$)で回帰
- 問題点： $educ$ は本人の能力（これは観察不可能な変数）と相関があるかもしれない
 - 誤差項と $educ$ の間の相関
 - 操作変数法による推計が必要

注意：Rでの欠損値の扱い

データセットの中に欠損値が含まれている場合 (mroz.xlsに欠損値あり)

データのimportの画面で、
n.a.stringsの欄に欠損値の数値 (文字列) を指定する

左図は欠損値が"."の場合

欠損値としてよく使われるのは
-999

のようなありえない数値

Import Dataset

Name: card

Input File: id,nearc2,nearc4,educ,age,fatheduc,motheduc,weight,momdad1,2,0,0,7,29,...,158413,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,54,3,0,0,12,27,8,8,380166,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,1,4,4,0,0,12,34,14,12,367470,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,1,5,1,1,11,27,11,12,380166,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,6,1,1,12,34,8,7,367470,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,7,7,1,1,12,26,9,12,380166,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,8,1,1,18,33,14,14,367470,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,9,1,1,14,29,14,14,496635,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,10,1,1,12,28,12,12,367772,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,11,1,1,12,29,12,12,480445,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,12,1,1,9,28,11,12,380166,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,13,1,1,12,26,11,6,380166,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,14,1,1,11,24,11,6,369567,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,15,1,1,11,30,11,6,496635,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,16,1,1,16,31,...,8,648053,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,0,1,17,1,1,14,24,15,12,357519,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,

Encoding: Automatic

Heading: ☒ Yes ☐ No

Row names: Automatic

Separator: Comma

Decimal: Period

Quote: Double quote (")

Comment: None

na.strings: .

☒ Strings as factors

Data Frame

id	nearc2	nearc4	educ	age	fatheduc	motheduc
2	0	0	7	29	.	.
3	0	0	12	27	8	8
4	0	0	12	34	14	12
5	1	1	11	27	11	12
6	1	1	12	34	8	7
7	1	1	12	26	9	12
8	1	1	18	33	14	14
9	1	1	14	29	14	14
10	1	1	12	28	12	12
11	1	1	12	29	12	12
12	1	1	9	28	11	12
13	1	1	12	26	11	6
14	1	1	11	24	11	6
15	1	1	11	30	11	6
16	1	1	16	31	.	8
17	1	1	14	24	15	12

Import Cancel

Rでの欠損値 後から欠損値を指定する場合

データフレームmroz中の変数x

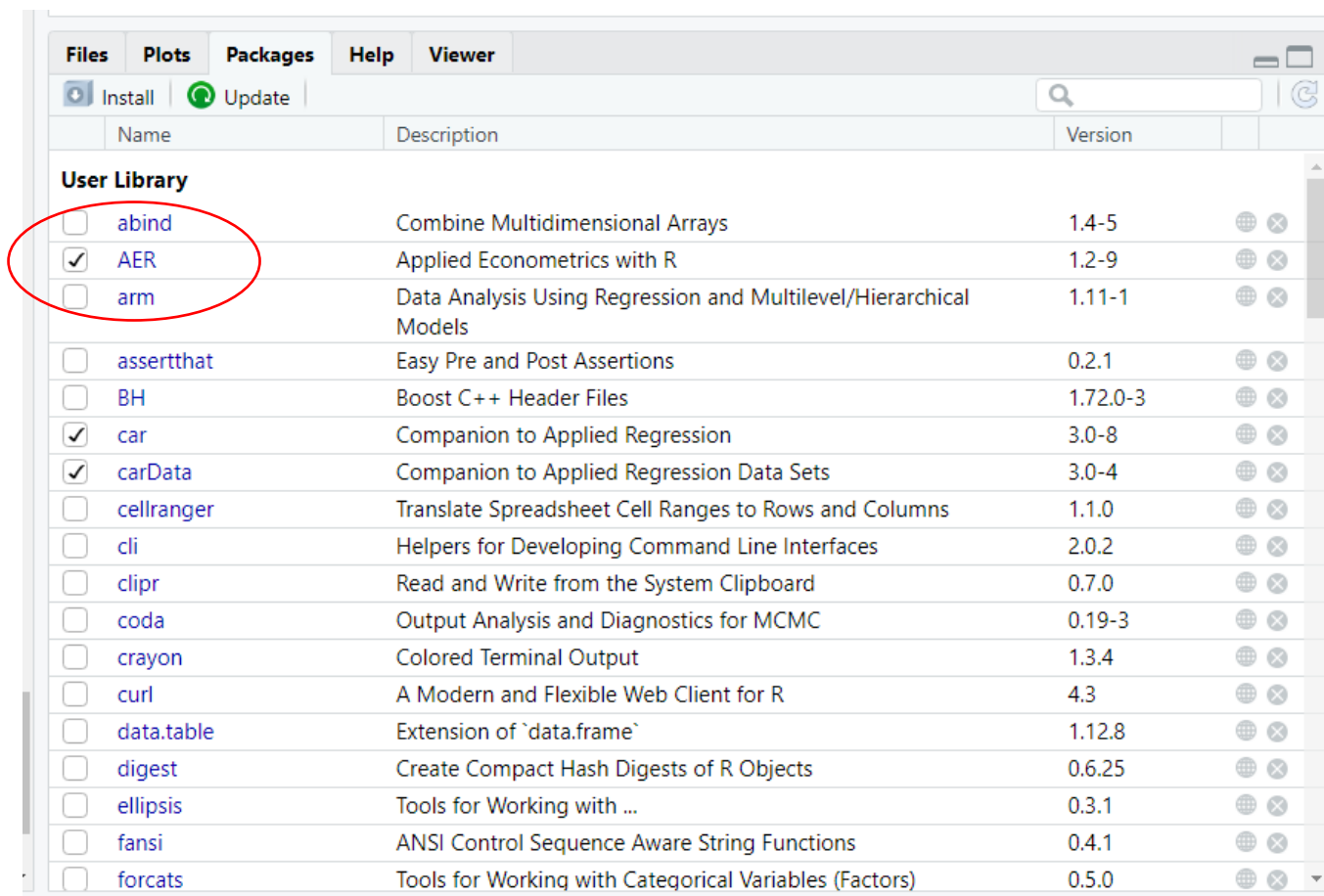
- 欠損値が-999の場合
 - `mroz$x[mroz$x == -999] <- NA`
 - `mroz$x`はベクトル:その要素が-999に等しいものにNA（欠損値：Not Available）を代入するというコマンド
 - いきなり置き換えるのが危険な場合は別の変数にxを代入してから行う
 - `mroz$y <- mroz$x` としてから`mroz$y`について上記の代入
- 欠損値が.のような文字列の場合
 - 変数xは文字列のベクトルとして読み込まれる
 - `mroz$x[mroz$x == "."] <- NA` として（文字列は” “ で囲む）
 - `mroz$x <- as.numeric(mroz$x)` で数値データに変換
- 新たに作成した変数を含めたデータセットを保存 → `write.csv()` 等で

R: パッケージのロード

ivreg()を使用するためにはパッケージAERが必要

Rstudioの右下のwindowでAERにチェックを入れればパッケージがロードされます

tsls()にはsemというパッケージが必要
右の画面になればinstallすること



OLSの結果

```
lm(formula = lwage ~ educ)
```

educと能力に相関がある場合、推定された係数は教育の効果だけでなく、能力の効果も反映されている

```
Residuals:
```

	Min	1Q	Median	3Q	Max
	-3.10256	-0.31473	0.06434	0.40081	2.10029

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.1852	0.1852	-1.000	0.318
educ	0.1086	0.0144	7.545	2.76e-13 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.68 on 426 degrees of freedom  
(325 observations deleted due to missingness)
```

```
Multiple R-squared:  0.1179, Adjusted R-squared:  
0.1158
```

```
F-statistic: 56.93 on 1 and 426 DF, p-value: 2.761e-13
```

操作変数と説明変数の相関を確認

Call:

```
lm(formula = educ ~ fatheduc)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.1881	-1.1881	0.2240	0.8119	6.3537

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.79901	0.19854	49.36	<2e-16 ***
fatheduc	0.28243	0.02089	13.52	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.046 on 751 degrees of freedom

Multiple R-squared: 0.1958, Adjusted R-squared: 0.1947

F-statistic: 182.8 on 1 and 751 DF, p-value: < 2.2e-16

ここでは操作変数としてfatheduc
(父親の教育年数)をとる；父親
の遺伝的性質
賃金方程式の説明変数である本人
の教育年数(educ)とfatheducには
相関があることが確認された

操作変数法による推計結果 ivreg()

```
ivreg(formula = lwage ~ educ | fatheduc)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.0870	-0.3393	0.0525	0.4042	2.0677

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.44110	0.44610	0.989	0.3233
educ	0.05917	0.03514	1.684	0.0929 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6894 on 426 degrees of freedom

Multiple R-Squared: 0.09344, Adjusted R-squared: 0.09131

Wald test: 2.835 on 1 and 426 DF, p-value: 0.09294

OLSよりeducの係数は低めに推定された。ただし、s.e.は大きい。

パッケージAERをロードする必要あり

操作変数法による推計結果 `tsls()`

結果は`ivreg()`と同じ。
`tsls()`の実行にはパッケージ`sem`
をロードしておく必要あり

2SLS Estimates

Model Formula: `lwage ~ educ`

Instruments: `~fatheduc`

Residuals:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.0870	-0.3393	0.0525	0.0000	0.4042	2.0677

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.44110350	0.44610178	0.98880	0.323324
educ	0.05917347	0.03514177	1.68385	0.092943

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6893899 on 426 degrees of freedom

問題1

1. mroz.xlsのデータを用い、次の方程式を通常の新最小二乗法で推計せよ。
 - 被説明変数：lwage（賃金の対数値）
 - 説明変数：educ（本人の教育年数）
2. educとfatheduc（父親の教育年数）の相関をみるため次の回帰を行え
 - 被説明変数：educ
 - 説明変数: fatheduc
3. 操作変数法を用いて、教育の収益率を推計せよ。
 - 被説明変数：lwage
 - 説明変数：educ
 - 操作変数: fatheduc

card.xls：教育の収益率の測定

- Cardは次のような賃金方程式の推計を行った
 - 被説明変数：賃金の対数値
 - 説明変数：教育年数，仕事の経験年数，人種，地域
 - ところが，教育年数は賃金方程式の誤差項と相関があるかもしれない？
 - 欠落変数（能力）の問題： 能力が高い→高学歴
 - 教育年数は親の所得，家庭環境によって決まる内生変数かもしれない
- そこで，Cardは教育年数(educ)の操作変数として，「17歳時に大学の近くに住んでいた」というダミー変数(nearc4)を操作変数として，操作変数法によって賃金方程式を推計した
 - nearc4 は教育年数と相関があった（大学への進学が容易）
 - nearc4 は本人の能力とは無関係だと考えられる

問題2

- 使用データ : card.xls

1. 次の賃金方程式をOLSで推計

- 被説明変数: lwage
- 説明変数: educ, exper, expersq, black, south, smsa, smsa66, reg662-reg669
- smsaは大都市圏ダミー, reg662-reg669は地域ダミーですが, 詳細な情報は無いので説明変数に加えるだけ; reg661は含めない

2. educ を nearc4(17歳時点で大学の近くに住んでいたというダミー変数) で回帰し, 相関があることを確かめる

3. 操作変数法で賃金方程式を推計し, OLSの結果と比較する

- 操作変数 near4c, exper, expersq, black, south, smsa, smsa66, reg662-reg669