

## Rによるデータ分析入門：練習問題・回答例 第6章

1. 教育が収入に及ぼす影響は教育の収益率として教育経済学の分野では古くからさかんに研究されてきました。ここでは、第4章で使用した `mroz` データを用いて、教育年数が1年延びると収入(賃金)がどの程度変化するかを計算してみましょう。教育年数と賃金は、IQといった観察できない個人の能力の影響を受けている可能性があります(例: IQが高いと教育年数も長くなり同時に賃金も高くなる)。よって最小二乗法(OLS)では教育の効果をうまく計測できない可能性があります。そこで操作変数を用いて、教育年数が賃金に及ぼす影響はについて分析します。

(1) 賃金の対数値(`wage`)を被説明変数に、教育年数(`educ`)と経験年数(`exper`)を説明変数とする推計式を考える。操作変数は「父親の教育年数(`fatheduc`)」と「母親の教育年数(`matheduc`)」の2つを用意している。なぜ、この2つの変数が操作変数として機能すると考えられるのか説明せよ。

教育年数が長い父親・母親は、教育、特に大学進学に理解があるので、教育年数が長くなる(正の相関を持つ)。一方で、父親・母親が高学歴だからと言って子供の賃金(所得)が高くなるとは限らないので、これらの変数は操作変数としての条件を満たすと考えられる。

(2) (1)の推計式を①OLSによる推定、②「父親の教育年数」を用いた操作変数法による推定、③「父親の教育年数」と「母親の教育年数」を操作変数とする推定を実施せよ。

推定結果は以下の通り。

```
> # ①最小二乗法(OLS)の推定結果
> result_ols <- fixest::feols(log(wage)~educ+exper, data=ataf)
> summary(result_ols)
OLS estimation, Dep. Var.: log(wage)
Observations: 428
Standard-errors: IID

      Estimate Std. Error  t value  Pr(>|t|)
(Intercept) -0.400174    0.190368  -2.10211 3.6132e-02 *
educ          0.109489    0.014167   7.72833 7.9370e-14 ***
exper         0.015674    0.004019   3.89980 1.1186e-04 ***
---
```

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

RMSE: 0.666619 Adj. R2: 0.14435

>

> # ②操作変数法（操作変数=父親の教育年数）の推定結果

> result\_iv1 <- fixest::feols(log(wage)~exper|educ~fatheduc, data=dataf)

> summary(result\_iv1)

TSLS estimation, Dep. Var.: log(wage), Endo.: educ, Instr.: fatheduc

Second stage: Dep. Var.: log(wage)

Observations: 428

Standard-errors: IID

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.035611	0.439654	0.080999	0.93548110
fit_educ	0.075216	0.034232	2.197246	0.02854187 *
exper	0.015526	0.004049	3.834576	0.00014482 ***

---

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

RMSE: 0.671193 Adj. R2: 0.132568

F-test (1st stage), educ: stat = 89.3 , p < 2.2e-16 , on 1 and 425 DoF.

Wu-Hausman: stat = 1.23041, p = 0.267956, on 1 and 424 DoF.

>

> # ③操作変数法（操作変数=父親の教育年数+母親の教育年数）の推定結果

> result\_iv2 <- fixest::feols(log(wage)~exper|educ~fatheduc+motheduc, data=dataf)

> summary(result\_iv2)

TSLS estimation, Dep. Var.: log(wage), Endo.: educ, Instr.: fatheduc, motheduc

Second stage: Dep. Var.: log(wage)

Observations: 428

Standard-errors: IID

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.147841	0.402215	0.367568	0.7133784
fit_educ	0.066389	0.031252	2.124330	0.0342193 *
exper	0.015488	0.004064	3.810595	0.0001591 ***

---

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

RMSE: 0.673838 Adj. R2: 0.125717

F-test (1st stage), educ: stat = 56.3 , p < 2.2e-16 , on 2 and 424 DoF.

Wu-Hausman: stat = 2.46836 , p = 0.116905, on 1 and 424 DoF.

Sargan: stat = 0.383852, p = 0.535549, on 1 DoF.

(3) (2) の推計結果の教育年数 (educ) の係数を比較せよ。OLS による係数の推定値は操作変数による推定値よりも大きくなったか、小さくなったか、また、それはなぜかを説明せよ。

educ の係数は OLS の場合、0.1095 だが、操作変数法による②と③の推定では、0.752、0.066 と小さくなっている。これは最小二乗法の結果は、教育年数と賃金の両方に影響する IQ や集中力といった観察されない要因の存在が、教育年数と賃金の間に相関をもたらしていたが、操作変数ではこうした観察されない要因の影響を排除することができたので係数が小さくなったと解釈できる。

2. 労働経済学では子どもの数が増えることで女性の就業がどの程度抑制されるかについてさかんに分析が試みられてきました。しかし、労働時間が長くなると出生率が抑えられるという逆のメカニズムも存在し、出生率と労働供給は同時決定と考えるのが自然です。ここでは Angrist and Evans (1998) を参考に子どもの数が労働供給におよぼす因果効果を操作変数法で分析します。具体的には Angrist and Evans (1998) では、第一子と第二子が同性である親は第三子を望みやすいという事象を利用した操作変数を用いて推計を行っています。使用するデータは data (Feterlity, package='AER') あるいは Feterlity.csv です。

(1) 「第三子をもつダミー」が「女性の労働時間」に及ぼす影響を分析するにあたり「第一子と第二子の性別が同じダミー」が妥当な操作変数であると考えられる理由について説明せよ。

第一子と第二子がともに男の子、あるいは女の子場合、今度は女の子、あるいは男の子が欲しいという心理が働くため、第三子目を希望する家庭が多いと言われている。こうした選好をふまえると、「第一子と第二子の性別が同じ」ダミーは第三子を持つ確率にプラスの影響を及ぼすと考えられる。一方で、「第一子と第二子の性別が同じ」かどうかは被説明変数である母親の労働時間には影響しないと考えられるので、この変数は操作変数として機能すると考えられる。

(2) 最小二乗法で、被説明変数に work (女性の労働時間)、説明変数に mkids (第三子をもつダミー)、age (年齢) と age の 2 乗値、afam (アフリカ系アメリカ人)、hispanic (ヒスパニック系)、other (その他) を用いた回帰式を推定せよ。なお、afam, hispanic, other

はカテゴリー変数なので説明変数として 導入する際は, factor () 関数でダミー変数とすること.

#### (4)の回答例を参照のこと

(3) gender1 (第一子の性別) と gender2 (第二子の性別) が同じであれば 1 をと る「第一子と第二子の性別が同じダミー (samegender)」を作成せよ.

#### 省略

(4) samegender (第一子と第二子の性別が同じダミー) を操作変数として (2) の推 定式を操作変数法で推定せよ. (2) と (3) の morekids の係数の大きさを比較し, なぜ係数の大きさが変わったかを説明せよ.

```
> # fm1 は OLS, fm2 は操作変数法、両者を比較
> fm1 <- feols(work~morekids+age+I(age^2)+afam+hispanic+other, data = Fertility)
> fm2 <- feols(work~+age+I(age^2)+afam+hispanic+other|morekids ~ samegender, data = Fertility)
>
> fixest::etable(fm1, fm2, stage=1:2, se="HC1", fitstat=~ivf+ivf.p, se.below=TRUE)
```

	fm1	fm2. 1	fm2. 2
IV stages		First	Second
Dependent Var. :	work morekidsyes		work
Constant	-5.455* (2.720)	0.1830** (0.0630)	-5.550* (2.735)
morekidsyes	-6.230*** (0.0862)		-5.819*** (1.247)
age	0.8807*** (0.1880)	-0.0069 (0.0043)	0.8839*** (0.1883)
age square	-0.0007 (0.0032)	0.0004*** (7.38e-5)	-0.0009 (0.0033)
afamy	11.66*** (0.1955)	0.1004*** (0.0044)	11.62*** (0.2318)
hispanicyes	0.4669** (0.1807)	0.1508*** (0.0041)	0.4049 (0.2605)

otheryes	2.142*** (0.2083)	0.0274*** (0.0046)	2.131*** (0.2110)
samegender		0.0680*** (0.0019)	
<hr/>			
S. E. type	Hete. -rob.	Heter. -rob.	Hete. -rob.
F-test (1st stage)	--	1,278.1	--
F-test (1st stage), morekidsyes	--	--	1,278.1
F-test (1st stage), p-value	--	3.3e-279	--
F-test (1st stage), p-value, morekidsyes	--	--	3.3e-279
---			
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

morekidsyes の係数は OLS では-6.23、操作変数法では-5.82 で係数は絶対値で小さくなっている。労働時間と子供の数の関係には、子どもの数が増えると子育てに時間をとられるので労働時間を抑制せざるをえなくなるという子どもの数→労働時間という関係と、長時間労働を厭わず働いている女性の中には子どもを複数持つことを躊躇う人もいるという逆のメカニズムも存在する。労働時間と子供の数の関係を OLS で推計すると両者の関係が含まれてしまうため、本来知りたい子どもの数→労働時間の関係を（絶対値でみて）過大評価してしまう可能性がある。

一方、操作変数による推定の場合、子どもの数の変化が労働時間にどう影響したかという因果効果をみることができる。実際、操作変数法で得られた係数は、過大評価される可能性がある OLS の係数よりも絶対値でみて小さくなっている。

(5) 第一段階の F 検定統計量からみて操作変数法は機能しているといえるか？

第一段階の F 値は 1278.1 であり、操作変数法は機能していると考えられる。

3. 第5章の練習問題2では、Cornwell and Trumbull (1994) のデータを用い、アメリカ・ノースカロライナ州の郡単位のデータで犯罪率の決定要因を分析しました。しかし、そこで用いられている説明変数のうち「人口当たりの警察官の数」と「逮捕される確率」は被説明変数と同時決定になっている可能性があります。これに配慮して固定効果操作変数法で分析することで結果がどのように変わるかを分析してみましょう。

(1) なぜ「人口当たりの警察官の数」と「逮捕される確率」は被説明変数と同時決定になっていると考えられるのか。

警察官が増えると犯罪率が抑制される可能性があるが、一方で、犯罪率の高い地域に警察官を多く配備している可能性もあり両者の間には双方向の関係がある。逮捕される確率も同様に、この確率が高くなると犯罪のコストが高まるので犯罪抑止効果があるが、犯罪が増えると逮捕者も増えるので両者には双方向の関係があると言える。

(2) 操作変数として「郡の一人当たり税収 (ltaxpc)」と「直接顔を合わせる犯罪 (強盗, 暴行, 強姦) の比率 (lmix)」を用いる。これらが操作変数として機能する理由を説明せよ。

郡の一人当たり税収 (ltaxpc) が増えると警察の予算が増える。警察の予算が増えると人口当たりの警察官を増やし、逮捕者を摘発しやすくなる。直接顔を合わせる犯罪 (強盗, 暴行, 強姦) の比率 (lmix=the ratio of crimes involving face-to-face contact) が増えると、これを抑止するために警察予算を警察官の増員により多く振り向けるので、人口当たりの警察官の数が増え、および逮捕される確率が上昇する。

(3) 被説明変数に犯罪発生率 (lcrmrte), 説明変数には逮捕される確率 (lprobarr), 逮捕されたのち有罪になる確率 (lprobconv), 有罪になったのち収監される確率 (lprobpris), 刑期の平均 (lavgsen), 人口当たり警察官の数 (lpolpc), 人口密度 (ldensity) を説明変数とし, ①郡固定効果と年次の両方を含む固定効果モデル, ②固定効果を考慮した操作変数法 (固定効果操作変数法) で推定せよ。なお, 郡と年を示す変数はそれぞれ county と year である。

推計結果は次のページ参照。result1 は最小二乗法 (OLS)、result2 は固定効果モデル、result3.1 と result3.2 は操作変数法の第一段階の推定結果、result3 は第二段階の推定結果

(4) 第一段階の F 検定統計量からみて操作変数法は機能しているといえるか?

操作変数法の推計における F 値は result3 の列の F 値を参照します。lprbarr を被説明変数とする第一段階の推定式の F 値が 28.8、lpolpc の F 値が 18 と、いずれも 10 を上回っているので弱操作変数の問題がないと考える。

(5) 結果がどのように変わったか、警察官の数や逮捕される確率は犯罪発生率に対して因果効果を持つといえるか?

逮捕される確率と警察官の数の係数は OLS では有意だったが、操作変数法により逆の因果性を排除した推定では統計的に非有意になった。よって、これらの変数は因果効果を持たないと考えられる。

```
> fixest::etable(result1,result2,result3,stage=1:2,se="HC1",fitstat=~ivf+ivf.p, se.below=TRUE)
```

	result1	result2	result3.1	result3.2	result3
IV stages			First	First	Second
Dependent Var. :	lcrmrte	lcrmrte	lprbarr	lpolpc	lcrmrte
Constant	-2.035*** (0.3605)				
lprbarr	-0.5227*** (0.0493)	-0.3560*** (0.0486)			-0.5664 (0.6610)
lprbconv	-0.4023*** (0.0329)	-0.2825*** (0.0363)	-0.3063*** (0.0372)	0.3087*** (0.0672)	-0.4176 (0.4121)
lprbpris	-0.0223 (0.0610)	-0.1802*** (0.0437)	-0.2530*** (0.0566)	0.0943 (0.0714)	-0.2542 (0.2324)
lavgsen	-0.1042* (0.0489)	-0.0044 (0.0329)	0.0002 (0.0449)	-0.0484 (0.0620)	0.0070 (0.0473)
lpolpc	0.3221*** (0.0481)	0.4214*** (0.0517)			0.6504 (0.6722)
ldensity	0.2521*** (0.0248)	0.4073 (0.3097)	-0.4579 (0.3889)	0.5511 (0.5911)	0.1746 (0.7145)
factor(region)west	-0.5670*** (0.0328)				
factor(region)central	-0.2317*** (0.0256)				



ltaxpc			0.0418	0.1098	
			(0.0590)	(0.0843)	
lmix			0.1348***	0.1285*	
			(0.0401)	(0.0551)	
Fixed-Effects:	-----	-----	-----	-----	-----
county	No	Yes	Yes	Yes	Yes
year	No	Yes	Yes	Yes	Yes
	-----	-----	-----	-----	-----
S.E. type	Hete.-rob.	Hete.-rob.	Hete.-rob.	Hete.-rob.	Hete.-rob.
F-test (1st stage)	--	--	24.634	15.389	--
F-test (1st stage), lprbarr	--	--	--	--	28.786
F-test (1st stage), lpolpc	--	--	--	--	17.983
F-test (1st stage), p-value	--	--	5.91e-11	3.19e-7	--
F-test (1st stage), p-value, lprbarr	--	--	--	--	1.12e-12
F-test (1st stage), p-value, lpolpc	--	--	--	--	2.56e-8
---					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					