Measuring the Social Return of Higher Education

鄭泊聲、倪楷恩、陳致宇、白崇佑

國立台灣大學

May 27, 2022

External versus Internal

- Government subsides for higher education is high. Is this justified?
- Basic demand and supply model tells us: government subside for external benefits maximize welfare.
- How can we measure the external benefits?
    - Education increases personal wage, but that's **internal**.
    - How about average wage in regions with different amount of higher education? This might include **external** effect.

Much of this work is based on Moretti 2004.

Data and Variables

- Wage data: a data based on workplace (instead of household) location city average wage data calculated by DGBAS. Only 2018-2020.
- Education data: How to measure "the amount of higher education"?
  - No. of college gradutes in city[1]
  - Share of college level worker in city workforce[2]
  - City population education level - above college share [3]
- Other city characteristic data as controls

[1] Depart. of Statistics, Ministry of Education
[2] 縣市重要統計指標查詢系統, DGBAS
[3] 人口統計資料, Dept. of Household Registration


2020 City Data
Hsinchu City
Hsinchu County
Average yearly wage ($10,000)
50
60
70
80
90
40
50
60
70
80
Share of college worker (%)

Model Specification

Dependent variable is city wage at 2020, *wage*2020. The full MLR model is

$$wage2020 = \beta_0 + workforceCollege\_2020 + \beta\mathbf{X} + u \qquad (1)$$

with *workforceCollege_2020*, the share of college educated worker in city workforce, as main explanatory. **X** is a vector containing various city characteristic, including

- *direct*: a dummy for the 6 Special Municipality
- *directEdu*2020: a interaction term between *direct* and *workforceCollege_2020* to allow different slope.

All variables

Table: MLR on all variables

	Dependent variable:
	wage2020
workforceCollege_2020	−0.818
	(1.814)
direct	81.297**
	(23.442)
hired2020	1.486**
	(0.608)
manufecture2020	−2.045**
	(0.850)
service2020	−1.049
	(0.915)
gender2020	2.499**
	(0.994)
eduExpense2020	1.162*
	(0.572)
eduLevel2020	1.866
	(1.746)
married2020	1.504
	(1.141)
expensePerCapita2020	0.003***
	(0.001)
unemployment2020	−31.468
	(21.214)
directEdu2020	−1.531***
	(0.416)
Constant	−268.635
	(171.289)
Observations	20
R^2	0.959
Adjusted R^2	0.888
Residual Std. Error	4.055 (df = 7)
F Statistic	13.604*** (df = 12; 7)
Note:	*p<0.1; **p<0.05; ***p<0.01

Joint significance of education

Table:

	Dependent variable:
	wage2020
workforceCollege_2020	0.946
	(1.231)
eduExpense2020	0.650
	(0.530)
eduLevel2020	−0.333
	(1.259)
Constant	9.424
	(18.926)
Observations	20
R^2	0.525
Adjusted R^2	0.436
Residual Std. Error	9.118 (df = 16)
F Statistic	5.889^{***} (df = 3; 16)
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

The Model

We take only reasonable and strong variables to form a compelling model as

$$\begin{aligned} wage2020 &= \beta_0 + \beta_1 workforceCollege\_2020 + \delta_0 direct \\ &+ \beta_2 wage2018 + \beta_3 manufecture2020 + \beta_4 hired2020 \end{aligned} \quad (2)$$

- $wage2018$: a lagged dependent as proxy to most of the city characteristics
- $manufecture2018$ is the share of manufecturing industry in gross production, $hired2020$ is the share of workforce classified as hired (instead of being employer or self-employed)

MLR

Table:

	Dependent variable:
	wage2020
workforceCollege_2020	0.075**
	(0.029)
direct	−0.194
	(0.413)
wage2018	1.017***
	(0.021)
manufecture2020	0.025
	(0.022)
hired2020	−0.032
	(0.038)
Constant	−1.259
	(2.005)

Observations	20
R^2	0.998
Adjusted R^2	0.997
Residual Std. Error	0.689 (df = 14)
F Statistic	1,175.873*** (df = 5; 14)
Note:	*p<0.1; **p<0.05; ***p<0.01

Heteroskedasticity Robust

Breusch-Pagan test: BP = 10.854, df = 5, p-value = 0.05435

Table:

	Dependent variable:
	wage2020
workforceCollege_2020	0.080***
	(0.013)
direct	0.020
	(0.240)
wage2018	0.997***
	(0.011)
manufecture2020	0.009
	(0.022)
hired2020	−0.034
	(0.042)
Constant	0.290
	(1.613)

Observations	20
R^2	0.998
Adjusted R^2	0.998
Residual Std. Error	0.474 (df = 14)
Note:	*p<0.1; **p<0.05; ***p<0.01

MLR

Table:

	Dependent variable:
	wage2020
direct	−0.322
	(0.421)
manufecture2020	0.022
	(0.020)
eduLevel2020	0.071**
	(0.026)
hired2020	−0.012
	(0.033)
wage2018	1.015***
	(0.021)
Constant	−1.967
	(1.876)

Observations	20
R^2	0.998
Adjusted R^2	0.997
Residual Std. Error	0.677 (df = 14)
F Statistic	1,220.078*** (df = 5; 14)
Note:	*p<0.1; **p<0.05; ***p<0.01

Heteroskedasticity Robust

Breusch-Pagan test: BP = 11.97, df = 5, p-value = 0.0352

Table:

	Dependent variable:
	wage2020
direct	−0.107
	(0.213)
manufecture2020	0.008
	(0.021)
eduLevel2020	0.080***
	(0.015)
hired2020	−0.017
	(0.041)
wage2018	0.992***
	(0.009)
Constant	−0.248
	(1.673)
Observations	20
R^2	0.999
Adjusted R^2	0.998
Residual Std. Error	0.372 (df = 14)
Note:	*p<0.1; **p<0.05; ***p<0.01

IV: Lagged Age Structure

Important criteria for an IV, z

- $cov(x, z) \neq 0$: As proportion of college gradutes in population grows in time, younger workforce may have more college graduates than older one.
- $cov(u, z) = 0$: Wage is unlikely to be correlated with age.

We use the lagged share of worker aged 15-24, *workforceYoung*_2010, as the IV.

First Stage

With a t-Statistic of -2.199, this choice of IV may not be strong enough.

Table:

	Dependent variable:
	workforceCollege_2020
workforceYoung_2010	-5.845^{**}
	(2.658)
Constant	91.140^{***}
	(19.525)
Observations	20
R^2	0.212
Adjusted R^2	0.168
Residual Std. Error	10.578 (df = 18)
F Statistic	4.835^{**} (df = 1; 18)
Note:	*p<0.1; **p<0.05; ***p<0.01

2SLS - College Share

Table:

	Dependent variable:
	wage2020
workforceCollege_2020	0.098
	(0.103)
direct	−0.308
	(0.650)
wage2018	1.007***
	(0.049)
manufecture2020	0.035
	(0.048)
hired2020	−0.050
	(0.087)
Constant	−0.643
	(3.368)

Observations	20
R^2	0.998
Adjusted R^2	0.997
Residual Std. Error	0.704 (df = 14)
Wald test	1125 on 5 and 14 DF, p-value: < 2.2e-16

2SLS - Edu Level

Table:

	Dependent variable:
	wage2020
eduLevel2020	0.097
	(0.101)
direct	−0.505
	(0.820)
wage2018	1.002***
	(0.054)
manufecture2020	0.033
	(0.045)
hired2020	−0.026
	(0.063)
Constant	−1.483
	(2.668)

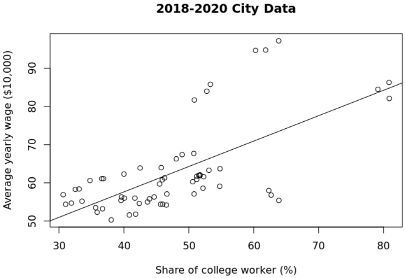
Observations	20
R^2	0.998
Adjusted R^2	0.997
Residual Std. Error	0.700 (df = 14)
Wald test	1138 on 5 and 14 DF, p-value: < 2.2e-16

Main Takeaway

- Choosing either share of college worker or share of college city population as main independent yield similar results.
- Either way they have significant positive effect.
- A lagged dependent variable is very explanatory.
- IV analysis suggest that actual effect may be even larger.

Panel Data

- Same data from 2018-2020.
- Model are specified the same except the the lagged dependent is removed.
- We used an unobserved effect panel data model.



First Differenced HAC

Table:

	Dependent variable:
	wageDiff
workforceCollegeDiff	−0.109
	(0.086)
direct2	0.341
	(0.283)
serviceDiff	0.358***
	(0.103)
manufectDiff	−0.028
	(0.133)
hiredDiff	−0.092
	(0.120)
Constant	0.671***
	(0.141)
Observations	40
R^2	0.279
Adjusted R^2	0.173
Residual Std. Error	0.601 (df = 34)
Note:	*p<0.1; **p<0.05; ***p<0.01

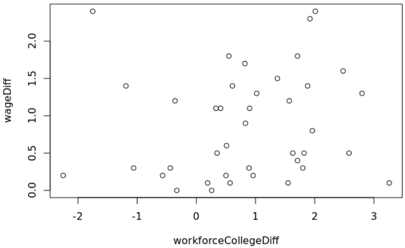
First Differenced with IV

Table:

	Dependent variable:
	wageDiff
workforceCollegeDiff	0.223
	(1.203)
direct2	0.294
	(0.433)
serviceDiff	0.073
	(0.816)
manufectDiff	−0.073
	(0.140)
hiredDiff	−0.067
	(0.143)
Constant	0.558
	(0.552)
Observations	40
R^2	0.024
Adjusted R^2	−0.119
Residual Std. Error	0.748 (df = 34)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

First Differenced

Very large standard error in first differencing estimation may be caused by very small variation in explanatory variable.



So we turned to random effect estimator.

Random Effect

Table:

	Dependent variable:
	wage2018
workforceCollege_2018	0.403***
	(0.114)
manufecture2018	0.122
	(0.127)
hired2018	−0.010
	(0.169)
direct	−8.666
	(10.479)
directEdu2018	0.098
	(0.187)
expensePerCapita2018	0.001***
	(0.0002)
Constant	24.754**
	(10.832)

Observations	60
R^2	0.539
Adjusted R^2	0.487
F Statistic	62.031***
Note:	*p<0.1; **p<0.05; ***p<0.01

RE with IV - First Stage

The IV in even less significant here.

Table:

	Dependent variable:
	workforceCollege_2018
workforceYoung_2013	0.082
	(0.584)
Constant	47.100***
	(4.891)
Observations	60
R^2	0.0003
Adjusted R^2	−0.017
F Statistic	0.020
Note:	*p<0.1; **p<0.05; ***p<0.01

RE with IV - 2SLS

Table:

	Dependent variable:
	wage2018
workforceCollege_2018	1.776
	(7.636)
manufecture2018	0.585
	(2.097)
hired2018	−0.850
	(4.705)
direct	35.758
	(264.171)
directEdu2018	−0.853
	(5.534)
expensePerCapita2018	0.0003
	(0.003)
Constant	21.315
	(22.243)
Observations	60
R^2	0.352
Adjusted R^2	0.279
F Statistic	17.148***
Note:	*p<0.1; **p<0.05; ***p<0.01

Takeaways

- Random effect estimation gives a larger significant coefficient than cross sectional OLS.
- IV using RE also suggest a even larger causal effect.

Other Panel Data Methods

- We find Fixed Effect results vary similar to RE. This can also be seen by the close to one $\hat{\theta}$.
- Pooling independent method finds only the effect of industry structure, *manufecture*2018, has a significant coefficient of 0.4280.

	var	std.dev	share
idiosyncratic	0.5627	0.7501	0.015
individual	37.9300	6.1587	0.985

$\hat{\theta}$= 0.9299

Table: Random effect estimation (without IV)

Remarks

- Our panel data analysis requires strict exogeneity, for which we kind of take it for granted.
- The R^2 in our RE estimation is only 0.487, which definitely leaves room for improvement.

Data

- Between 1982-2020
- National data
- Due to data limitation, after some test we decided to choose:
    - **Average household income** ($income_t$) as the dependent variable.
    - **Government education budget** ($edufund_t$) as the main explanatory variable.

Autoregression

- We find $income_t$ and $income_{t-1}$ to be highly correlated with 0.9975 correlation.
- So we perform first differencing on all variable and define

$$cincome_t = income_t - income_{t-1} \tag{3}$$

Model 1 Specification

$$cincome = cincome\_lag + cedufund + cunem + cindpd + cavgGDP \quad (4)$$

- *cincome_lag* in a 2 periods lag term of the dependent.
- *cedufund* is our main explanatory.
- *cunem* refers to unemployment rate.
- *cindpd* is the gross production of manufacturing industry, in million NTD.
- *cavgGDP* is the average GDP per capita.

Model 1 Results

Table: First Differenced MLR

	Dependent variable:
	cincome[3:38]
cincome_lag	0.555***
	(0.121)
cedufund[3:38]	0.0001**
	(0.00003)
cunem[3:38]	−8,050.278***
	(1,597.252)
cindpd[3:38]	−0.001
	(0.001)
cavgGDP[3:38]	0.149**
	(0.061)
Constant	178.333
	(1,437.168)

Observations	36
R^2	0.744
Adjusted R^2	0.702
Residual Std. Error	3,445.707 (df = 30)
F Statistic	17.451*** (df = 5; 30)
Note:	*p<0.1; **p<0.05; ***p<0.01

Serial Correlation

Define u_t as the residuals of the previous model.

Table:

	Dependent variable:
	ut
ut_1	0.023
	(0.194)
Constant	−52.015
	(553.908)
Observations	35
R^2	0.0004
Adjusted R^2	−0.030
Residual Std. Error	3,265.292 (df = 33)
F Statistic	0.014 (df = 1; 33)
Note:	*p<0.1; **p<0.05; ***p<0.01

Our model doesn't seem to be affected by SC.

Heteroskedasticity

Breusch-Pagan test:
BP = 5.4892, df = 5, p-value = 0.3591
Our model doesn't seem to be affected by heteroskedasticity.

Model 2 Specification

$$cincome = cincome\_lag + cedufund + cunem + cindpd + cindpd\_lag + cavgGDP \quad (5)$$

- *cincome_lag* is a 2 periods lag term of the dependent.
- *cedufund* is our main explanatory.
- *cunem* refers to unemployment rate.
- *cindpd* is the gross production of manufecturing industry, in million NTD.
- *cindpd_lag* is a 1 period lag term of *cindpd*.
- *cavgGDP* is the average GDP per capita.

Model 2 Results

Table: First Differenced MLR model 2

	Dependent variable:
	cincome[3:38]
cincome_lag	0.553***
	(0.112)
cedufund[3:38]	0.0001***
	(0.00003)
cunem[3:38]	−9,850.055***
	(1,643.410)
cindpd[3:38]	−0.001*
	(0.001)
cindpd_lag	−0.001**
	(0.001)
cavgGDP[3:38]	0.077
	(0.063)
Constant	2,159.014
	(1,548.312)
Observations	36
R^2	0.789
Adjusted R^2	0.745
Residual Std. Error	3,182.269 (df = 29)
F Statistic	18.079*** (df = 6; 29)
Note:	*p<0.1; **p<0.05; ***p<0.01

Serial Correlation

Define u_t as the residuals of the previous model.

Table:

	Dependent variable:
	ut
ut_1	0.019
	(0.198)
Constant	−43.651
	(503.820)
Observations	35
R^2	0.0003
Adjusted R^2	−0.030
Residual Std. Error	2,967.877 (df = 33)
F Statistic	0.009 (df = 1; 33)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Our model doesn't seem to be affected by SC.

Heteroskedasticity

Breusch-Pagan test:
BP = 5.3746, df = 6, p-value = 0.4967
Our model doesn't seem to be affected by heteroskedasticity.

Model 3 Specification

$$cincome = cincome\_lag + cedufund + cunem + cservpd + cavgGDP \quad (6)$$

- *cincome_lag* in a 2 periods lag term of the dependent.
- *cedufund* is our main explanatory.
- *cunem* refers to unemployment rate.
- *cservpd* is the gross production of service industry, in million NTD.
- *cavgGDP* is the average GDP per capita.

Model 3 Results

Table: First Differenced MLR model 3

	Dependent variable:
	cincome[3:38]
cincome_lag	0.729***
	(0.126)
cedufund[3:38]	0.0001**
	(0.00003)
cunem[3:38]	−9,225.299***
	(1,609.773)
cservpd[3:38]	−0.009**
	(0.004)
cavgGDP[3:38]	0.201***
	(0.063)
Constant	1,517.859
	(1,488.989)

Observations	36
R^2	0.774
Adjusted R^2	0.737
Residual Std. Error	3,235.215 (df = 30)
F Statistic	20.602*** (df = 5; 30)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Serial Correlation

Define u_t as the residuals of the previous model.

Table:

	Dependent variable:
	ut
ut_1	0.119
	(0.181)
Constant	−19.639
	(517.243)
Observations	35
R^2	0.013
Adjusted R^2	−0.017
Residual Std. Error	3,055.879 (df = 33)
F Statistic	0.435 (df = 1; 33)
Note:	*p<0.1; **p<0.05; ***p<0.01

Our model doesn't seem to be affected by SC.

Heteroskedasticity

Breusch-Pagan test:
BP = 10.115, df = 5, p-value = 0.07205
Our model doesn't seem to be affected by heteroskedasticity.

Takeaways

- All 3 models yields similar results.
- Government education budget has a positive, significant, but small effect.
- The models are free from heteroskedasticity and serial correlation.
- R^2s lies in the 0.7 range, still room for improvement.
- The main explanatory accounts for all education level, not just higher.
- The dependent also doesn't directly indicate wage.

Comparisons

In comparison to our cross sectional analysis,

- TS obtains smaller coefficient,
- TS models have lower R^2.

In comparison to our panel analysis,

- TS obtains smaller coefficient,
- TS models have higher R^2.

From several models using cross section, panel, time series methods, we find (higher)education does have a positive effect on average wage. How much the effect is distributed to the external is beyond the scope of this work.

1. **The Idea**
2. **Cross Sectional MLR**
   - College Worker Share
   - City Higher Education Level
3. **Instrumental Variable**
4. **2018-2020 Panel Data**
   - First Differenced
   - Random Effect
5. **1982-2020 Time Series Data**
   - Model 1
   - Model 2
   - Model 3
   - Detrending

Detrending MLR

Table:

	Dependent variable:
	income
edufund	0.0003***
	(0.00002)
unem	−4,634.824***
	(1,311.579)
indpd	−0.004***
	(0.001)
avgGDP	0.275***
	(0.066)
t	−2,095.528
	(1,398.582)
Constant	32,611.320***
	(5,618.118)

Observations	39
R^2	0.996
Adjusted R^2	0.995
Residual Std. Error	6,269.286 (df = 33)
F Statistic	1,575.655*** (df = 5; 33)
Note:	*p<0.1; **p<0.05; ***p<0.01

Detrended AR

First, we obtain the detrended variables $\ddot{y}_t$ by regressing

$$y_t = \alpha_0 + \alpha_1 t + e_t \tag{7}$$

and specify $\ddot{y}_t = e_t$.
Subsequently, we found detrended $\ddot{income}_t$ highly correlated to $\ddot{income}_{t-1}$ with correlation 0.9570.
So we performed first differencing and define

$$cincome\_dt_t = \ddot{income}_t - \ddot{income}_{t-1} \tag{8}$$

FD Detrending MLR

Then the regression result is **exactly the same** as the non-detrending FD MLR (in table 16).

Table: First differenced detrending MLR

	Dependent variable:
	cincome_dt[4:38]
cincome_dt[2:36]	0.558***
	(0.122)
cedufund_dt[4:38]	0.0001**
	(0.00003)
cunem_dt[4:38]	$-$8,154.249***
	(1,623.290)
cindpd_dt[4:38]	$-$0.001
	(0.001)
cavgGDP_dt[4:38]	0.154**
	(0.062)
Constant	152.261
	(602.216)

Observations	35
R^2	0.741
Adjusted R^2	0.696
Residual Std. Error	3,482.437 (df = 29)
F Statistic	16.602*** (df = 5; 29)
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

SC, Heteroskedasticity and Our Question

- Serial correlation and heteroskedasticity result are also the same.
- So, is detrending redundant after first-differencing?

Reference I

[1] Enrico Moretti. "Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data". In: *Journal of Econometrics* 121.1 (2004). Higher education (Annals issue), pp. 175–212. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2003.10.015. URL: https://www.sciencedirect.com/science/article/pii/S0304407603002653.