Replication of "Educational Expansion and Its Heterogeneous Returns for Wage Workers" by Michael Gebel and Friedhelm Pfeiffer

Luisa Hammer and Marcelo Avila

22 Nov 2018

TODO: include outline of present.

Theoretical Part

Introduction

Summary of Gebel & Pfeiffer (2010)

- basic idea: examine evolution of returns to education in West German labour market.
- Focus on change in returns to education over time as a consequence to education expansion in Germany.
- methodology:
  - Wooldrigdge's (2004) **conditional mean independence**
  - Garen's (1984) **control function** approach, that requires an *exclusion restriction*
  - as well as OLS
- data: SOEP 1984-2006

Data and Variables

- Log of hourly wage
- Years of education (constructed from categorical variable)
- Age and age squared
- Gender
- Father's education
- Mother's ecucation
- Father's occupation
- Rural or urban household
- Number of Siblings (as instrument)

TODO: more detailed table? (Comment) not necessary

Background Information

- **increase in educational attainment** in the 1960s. From 1984 to 2006, average years of schooling increased:
  - woman: 11.3 -> 12.8
  - men: 11.9 -> 12.9
- **How can educational expansion affect the returns to education?**
- Standard theory: an increase of labor supply of high-skilled workers should decrease the returns to education
- High-educated workers with higher unobserved motivation / ability which positively affects wages
- More “less talented” accepted to higher education & thereby decreasing the average productivity levels of higher educated workers –> overall effect not clear
- unobserved characteristics leading to **selection bias**:
  - higher ability and motivation to stay longer in education

A few a priori hypotheses

(Comment) imo not super important, could be neglected because we have enough stuff tot alk about - but nice table!

Factors affecting RtE	$\Uparrow$ RTE	$\Downarrow$ RTE
Increase in female labour participation		✓
Birth cohort sizes (Baby boom)		✓
Wage determination processes (entrants)		✓
Skill-biased technological change	✓	

Econometric Approach

Empirical Framework (Derivation) I

The study is based on the **correlated random coefficient model** (Blundell / Dearden / Sianesi, 2005; Heckman / Vytlacil, 1998; Wooldridge, 2004).

$$\ln Y_i = a_i + b_i S_i$$

with $a_i = a' X_i + \varepsilon_{ai}$, and $b_i = b' X_i + \varepsilon_{bi}$

where $\ln Y_i$: log of wages and S_i years of schooling of individual i

- The model has, therefore, an **individual-specific intercept** a_i and **slope** b_i dependent on **observables** X_i and **unobservables** ε_{ai} and ε_{bi}.
- Does not assume that b_i and S_i are independent –> Individuals with higher expected benefits from education are more likely to remain longer in education –> b_i may be correlated with S_i meaning positive self-selection.

Empirical Framework (Derivation) II

- focus: estimate average partial effect (APE), which is the return per aditional year of education for a randomly chosen individual (or averaged across the population)

$$E(\partial \ln Y / \partial S) = E(b_i) = \beta$$

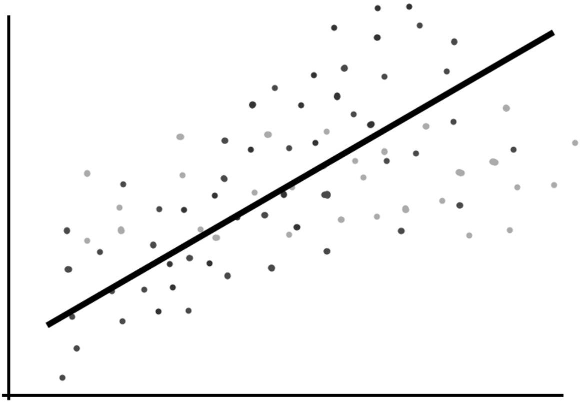
In case of homogenous returns to education the wage equation reduces to:

$$\ln Y_i = a' X_i + \bar{b} S_i + \varepsilon_{ai}$$

- Unobserved heterogeneity may only affect the **intercept** of the wage equation.

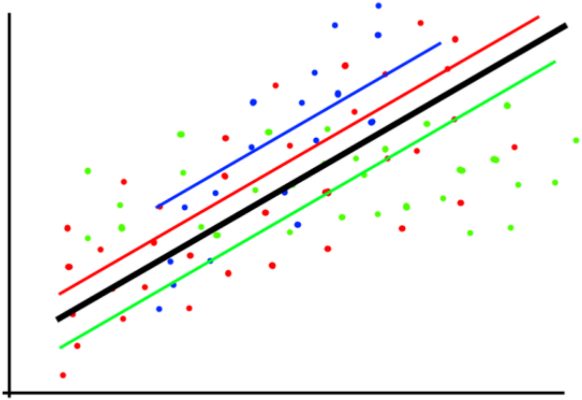
- still potential endogeneity if ε_{ai} correlates with S_i

Empirical Framework (Intuition) I



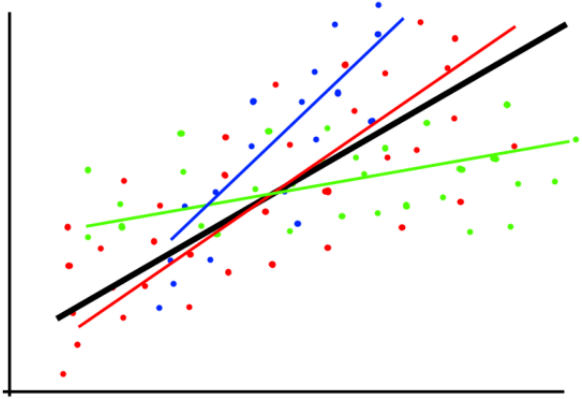
- Simple OLS

Empirical Framework (Intuition) II



- Multiple OLS with homogenous return to Educ

Empirical Framework (Intuition) III



- Correlated Random Coefficient Model

Distinction to conventional methods

- OLS
  - ability and “background” bias
- IV Methods
  - if education is correlated with **unobserved individual heterogeneity**, IV methods may fail to identiy APE.
    - alternative: **L**ocal **A**verage **T**reatment **E**ffect.

Conditional Mean Independence

According to Wooldridge (2004, pg. 7), **APE** is identified by:

$$E(\ln Y_i \mid a_i, b_i, S_i, X_i,) = E(\ln Y_i \mid a_i, b_i, S_i) = a_i + b_i S_i \qquad (A.1)$$

$$E(S_i \mid a_i, b_i, X_i) = E(S_i \mid X_i) \text{ and } \mathrm{Var}(S_i \mid a_i, b_i, X_i) = \mathrm{Var}(S_i \mid X_i) \qquad (A.2)$$

TODO: add interpretation of assumptions

Estimator for β and GLM

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \left(\left(S_i - \hat{E}(S_i \mid X_i) \ln Y_i \right) \Big/ \hat{Var}(S_i \mid X_i) \right)$$

$$E(S_i \mid X_i) = e^{\gamma X_i} \quad \text{and} \quad Var(S_i \mid) = \sigma^2 e^{\gamma X_i}$$

Where σ^2 can be consistently estimated by the mean of squared Pearson residuals and standard errors are bootstrapped.

Control Function Approach I

- Based on proposition by Garen (1984).
- Similar to Heckman two-step estimator.
- Models schooling choice explicitly in first step
- CF approach can identify APE in heterogeneus returns while standard IV approach may not.

First stage: modellation of schooling choice

$$S_i = c' X_i + dZ_i + v_i \quad \text{with} \quad E(v_i \mid Z_i, X_i) = 0$$

where:

- X_i and Z_i influence the educational decision.
- v_i: Error term incorporating unobserved determinants of education choice.
- Z_i: Exclusion restriction (instrument).

Control Function Approach II

- V_i, ε_{ai} and ε_{bi} are normally distributed with zero means and positive variances.
- possible correlation between error terms
- v_i is positive if an individual acquires higher education than expected conditional on observed characteristics

Second step: augmented wage equation

$$\ln Y_i = a_i + \beta S_i + \gamma_1 v_i + \gamma_2 V_i S_i + w_i$$

where:

- $\gamma_1 v_i$ and γ_2 are the **control functions**
  - $\gamma_1 = cov(\varepsilon_{ai}, v_i)/var(v_i)$
  - $\gamma_2 = cov(\varepsilon_{bi}, v_i)/var(v_i)$
- $E(w_i \mid X_i, S_i, v_i) = 0$ (as shown in Heckman / Robb, 1985)

Control Function Approach III

Interpretation of the coefficients of the control functions - γ_1 measures the effect of those unobserved factors that led to over- or under-achievement in education on the wage - Thus, if γ_1 is positive, the unobserved factors affect schooling *and* wages positively - γ_2 describes how this effect changes with increasing levels of education - Positive coefficient would indicate that those with unexpected educational "over-achievement" tend to earn higher wages

TODO: intuition for CF approach

Replication results

Set-up

- We use the same sample: West Germans (not foreign-born or self-employed) between 25 & 60 years who work full-time
- We have less observations than Gebel & Pfeiffer (2010) per survey year after we delete all observations with missing values
- Yet, we extend the observation period until 2016
- Three estimation methods: OLS, CMI & CF
- We are not able to replicate the estimation results of Gebel & Pfeiffer (2010) exactly, yet the trend / shape is similar

Results

- i'm not so sure how to add images / tables here but in the new do-file link https://1drv.ms/u/s!Ap1Tm8513oIthBjgyIALS8Zp3A7G you can just save the graph with all 3 approaches
- & then display on the other side the same graph from GP(2010, p.35)
- also: here is a table with the our & GP estimates for comparisons - your bootstrapped standard errors are already included

https://1drv.ms/x/s!Ap1Tm8513oIthBp5BPId0qO8h3Yj

Estimated returns on education

- Estimates from OLS & CMI are similar, yet, CMI produces lower estimates which points to a positive self-selection bias
- Generally, CF estimates are much more volatile and less precise

Differences between replicated & original estimations - Our OLS estimates are on average larger than those of Gebel & Pfeiffer (2010) by 0.004 percentage points - Our CMI estimates are on average larger than those of Gebel & Pfeiffer (2010) by 0.002 percentage points (first years lower, than larger) - Our CF estimates are on average significantly larger by 0.032 percentage points, though the divergence gets smaller from 2000 onwards

Control function estimates I

Instrumental variable in first stage - *number of siblings* is significant at the 0.1% level for all years - as expected, the number of siblings has a negative impact on the years of schooling (the estimates range between -0.13 & -0.23) - We would assume that the instrument does not directly affect the error term in the wage equation

Coefficients of the control functions - γ_1 is negative for majority of years, yet very small and insignificant in all years - Gebel & Pfeiffer (2010) estimate a positive coefficient in the 1980s and 1990s - but also insignificant - γ_2 is negative and close to zero for most years - indicates that those with unexpectedly high education have lower returns to education - Similarly, they are only slightly significant in the 1980s, and stronger significant in the early 2000s - The estimates are very similar to those of Gebel & Pfeiffer (2010)

- that both coefficients are (mostly) negative hints that educational expansion caused more "less abled" to achieve higher education

Explanations for divergences between replication and Gebel & Pfeiffer (2010)

- sample not the same
- ...

Pro's & Con's of estimation methods

Results and Comparison I

	First Stage		Second Stage					
	IV: Nr. of Siblings		v_i			$v_i S_i$		
year	coeff	s.e.	coeff.1	s.e..1	p	coeff.2	s.e..2	p.1
1984	-0.163	0.035	-0.019	0.036	0.601	-0.003	0.001	0.027
1985	-0.191	0.036	0.005	0.030	0.864	-0.003	0.001	0.024
1986	-0.129	0.034	-0.039	0.041	0.344	-0.001	0.001	0.681
1987	-0.133	0.033	-0.064	0.039	0.105	-0.002	0.001	0.141
1988	-0.150	0.034	-0.031	0.034	0.365	-0.003	0.001	0.038
1989	-0.153	0.033	0.018	0.033	0.590	-0.002	0.001	0.056
1990	-0.164	0.032	-0.027	0.032	0.404	-0.001	0.001	0.341
1991	-0.167	0.033	0.014	0.034	0.685	-0.002	0.001	0.152
1992	-0.178	0.032	-0.007	0.030	0.808	-0.001	0.001	0.298
1993	-0.162	0.033	-0.033	0.033	0.311	-0.001	0.001	0.264
1994	-0.176	0.034	-0.035	0.029	0.233	-0.001	0.001	0.225
1995	-0.172	0.036	-0.026	0.032	0.422	-0.002	0.001	0.077
1996	-0.195	0.037	-0.015	0.031	0.624	-0.003	0.001	0.058
1997	-0.214	0.038	-0.030	0.027	0.268	-0.002	0.001	0.225
1998	-0.200	0.036	-0.049	0.028	0.076	0.000	0.001	0.866

Results and Comparison II

1999	-0.219	0.036	-0.029	0.027	0.282	-0.001	0.001	0.559
2000	-0.211	0.028	-0.010	0.024	0.669	-0.002	0.001	0.057
2001	-0.191	0.027	-0.007	0.023	0.769	-0.002	0.001	0.021
2002	-0.222	0.029	0.005	0.023	0.819	-0.004	0.001	0.001
2003	-0.209	0.029	0.016	0.024	0.499	-0.005	0.001	0.000
2004	-0.228	0.029	-0.025	0.024	0.298	-0.003	0.001	0.006
2005	-0.205	0.030	-0.027	0.026	0.312	-0.002	0.001	0.170
2006	-0.213	0.030	-0.032	0.026	0.221	-0.002	0.001	0.107
2007	-0.222	0.030	-0.015	0.025	0.536	-0.003	0.001	0.008
2008	-0.226	0.031	-0.048	0.027	0.082	-0.002	0.001	0.149
2009	-0.206	0.031	-0.026	0.028	0.343	-0.004	0.001	0.004
2010	-0.222	0.034	-0.020	0.028	0.473	-0.003	0.001	0.023
2011	-0.215	0.032	-0.029	0.028	0.297	-0.004	0.001	0.001
2012	-0.228	0.034	0.020	0.027	0.471	-0.005	0.001	0.000
2013	-0.224	0.033	-0.005	0.028	0.857	-0.003	0.001	0.017
2014	-0.238	0.035	0.015	0.028	0.579	-0.003	0.001	0.038
2015	-0.216	0.037	-0.020	0.031	0.519	-0.003	0.001	0.037
2016	-0.212	0.038	0.027	0.034	0.423	-0.003	0.001	0.037

Results and Comparison III

year	OLS	OLS.se	CMI	CMI_se	CF	CF_se	obs
1984	0.074	0.004	0.066	0.004	0.075	0.079	1.545
1985	0.065	0.004	0.059	0.004	0.032	0.131	1.600
1986	0.069	0.004	0.061	0.004	0.077	0.091	1.682
1987	0.070	0.004	0.063	0.004	0.090	0.048	1.775
1988	0.069	0.004	0.065	0.004	0.081	0.041	1.798
1989	0.067	0.003	0.063	0.004	0.067	0.038	1.922
1990	0.063	0.003	0.059	0.004	0.048	0.031	2.007
1991	0.062	0.003	0.060	0.004	0.037	0.030	2.122
1992	0.064	0.003	0.057	0.004	0.049	0.027	2.107
1993	0.060	0.003	0.057	0.004	0.062	0.026	2.124
1994	0.057	0.003	0.051	0.004	0.055	0.022	2.082
1995	0.058	0.003	0.053	0.004	0.064	0.024	2.075
1996	0.055	0.003	0.049	0.004	0.054	0.025	2.057
1997	0.057	0.003	0.054	0.003	0.074	0.025	2.011
1998	0.054	0.003	0.049	0.003	0.053	0.021	2.145
1999	0.058	0.003	0.054	0.003	0.072	0.023	2.163
2000	0.065	0.002	0.059	0.003	0.108	0.024	3.965
2001	0.065	0.002	0.062	0.003	0.097	0.022	3.961
2002	0.068	0.003	0.066	0.003	0.106	0.030	3.668

Results and Comparison IV

2003	0.064	0.003	0.062	0.003	0.123	0.028	3.476
2004	0.064	0.003	0.062	0.003	0.115	0.030	3.366
2005	0.065	0.003	0.064	0.003	0.113	0.032	3.220
2006	0.064	0.003	0.063	0.003	0.133	0.033	3.477

Results and Comparison V

year	ret_OLS	se_OLS	se_cf	ret_cf	ret_cmi	se_cmi	obs
1984	0.060	0.004	0.030	0.118	0.053	0.006	1.448
1985	0.067	0.003	0.024	0.100	0.060	0.005	1.412
1986	0.067	0.004	0.036	0.104	0.059	0.006	1.463
1987	0.069	0.004	0.034	0.154	0.062	0.005	1.489
1988	0.064	0.003	0.029	0.127	0.058	0.005	1.476
1989	0.064	0.003	0.028	0.073	0.056	0.005	1.553
1990	0.065	0.003	0.026	0.100	0.057	0.005	1.571
1991	0.060	0.004	0.028	0.067	0.053	0.005	1.602
1992	0.067	0.003	0.024	0.082	0.057	0.005	1.555
1993	0.070	0.004	0.027	0.115	0.063	0.005	1.527
1994	0.068	0.003	0.024	0.117	0.062	0.005	1.491
1995	0.062	0.003	0.026	0.113	0.057	0.005	1.444
1996	0.069	0.003	0.025	0.112	0.062	0.005	1.383
1997	0.067	0.003	0.021	0.112	0.061	0.005	1.285
1998	0.062	0.003	0.022	0.110	0.060	0.005	1.452
1999	0.065	0.003	0.021	0.098	0.061	0.005	1.452
2000	0.067	0.003	0.019	0.102	0.063	0.004	2.701
2001	0.069	0.003	0.019	0.106	0.066	0.004	2.659
2002	0.073	0.003	0.019	0.115	0.069	0.004	2.818

Results and Comparison VI

2003	0.074	0.003	0.020	0.122	0.071	0.004	2.741
2004	0.069	0.003	0.018	0.133	0.065	0.004	2.558
2005	0.069	0.003	0.021	0.116	0.066	0.004	2.457
2006	0.071	0.003	0.020	0.126	0.068	0.004	2.525
2007	0.072	0.003	0.020	0.132	0.071	0.004	2.462
2008	0.076	0.003	0.021	0.147	0.072	0.005	2.316
2009	0.075	0.003	0.023	0.153	0.075	0.004	2.367
2010	0.077	0.003	0.023	0.139	0.074	0.005	2.183
2011	0.078	0.003	0.023	0.162	0.075	0.004	2.523
2012	0.076	0.003	0.022	0.127	0.076	0.004	2.493
2013	0.072	0.003	0.022	0.117	0.071	0.004	2.477
2014	0.074	0.003	0.022	0.091	0.070	0.004	2.353
2015	0.074	0.003	0.025	0.129	0.071	0.005	2.147
2016	0.076	0.003	0.028	0.085	0.071	0.005	1.971

Graphs I

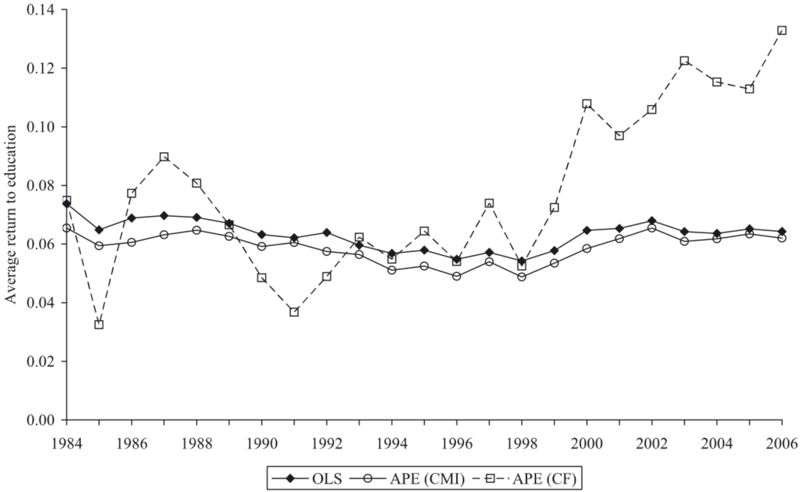
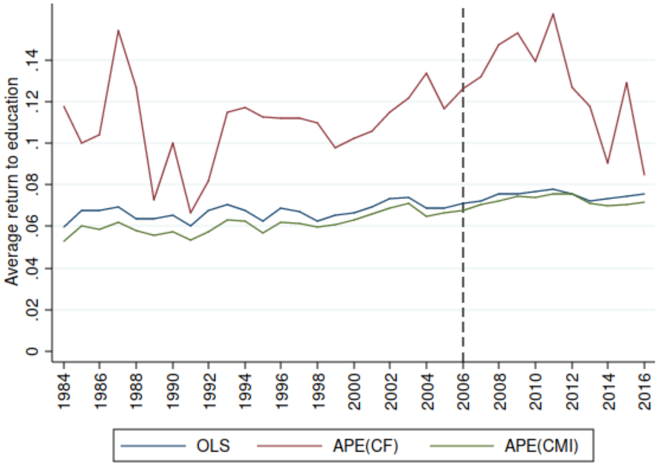


Figure 1: Original Results (GB 2010, pg.30)

Graphs II


Average return to education
0
.02
.04
.06
.08
.1
.12
.14
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
OLS
APE(CF)
APE(CMI)

Graphs III

Figure 2: Replication results: Comparison between OLS, CMI and CF approaches

The end I