

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

# INTRODUCTION

# THEORETICAL FRAMEWORK

INTRODUCTION

REPLICATION AND COMPARISON

RESULTS COMPARISON AND CONCLUSION

THEORETICAL FRAMEWORK

ECONOMETRIC APPROACH

# OUTLINE

# SUMMARY OF GEBEL AND 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:
  - Wooldridge's (2004) **conditional mean independence**
  - Garen's (1984) **control function** approach, that requires an *exclusion restriction*
  - as well as OLS
- data: SOEP 1984-2006

# BACKGROUND INFORMATION I

## Increase in educational attainment

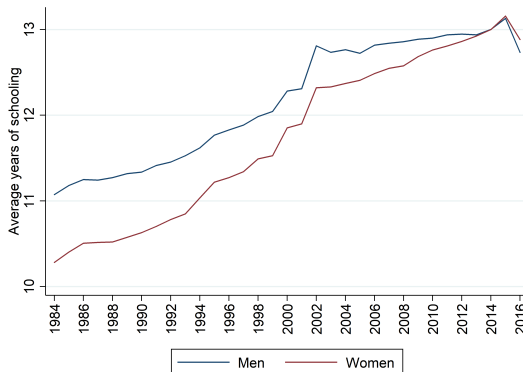


FIGURE 1: Source: SOEP 1984-2016, own estimations.

# BACKGROUND INFORMATION II

## 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
- If more less talented are accepted to higher education, this should decrease the average productivity levels of higher educated workers  
→ overall effect not clear

## Problems in the estimation of returns to education

- unobserved characteristics leading to **selection bias**:
  - higher ability and motivation to stay longer in education.
  - select jobs with higher expected returns.

# ECONOMETRIC APPROACH



# EMPIRICAL FRAMEWORK (DERIVATION) I

The study is based on the **correlated random coefficient model** (Wooldridge, 2004) specified as:

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

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**  $\varepsilon_{ai}$  and  $\varepsilon_{bi}$ .
- Do not assume that  $b_i$  and  $S_i$  are independent  $\rightarrow$  Individuals with higher expected benefits from education are more likely to remain longer in education  $\rightarrow b_i$  may be correlated with  $S_i$  indicating positive self-selection.

# EMPIRICAL FRAMEWORK (DERIVATION) II

- focus: estimate average partial effect (APE), which is the return per additional 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 homogeneous 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  $\varepsilon_{ai}$  correlates with  $S_i$

# EMPIRICAL FRAMEWORK (INTUITION) I

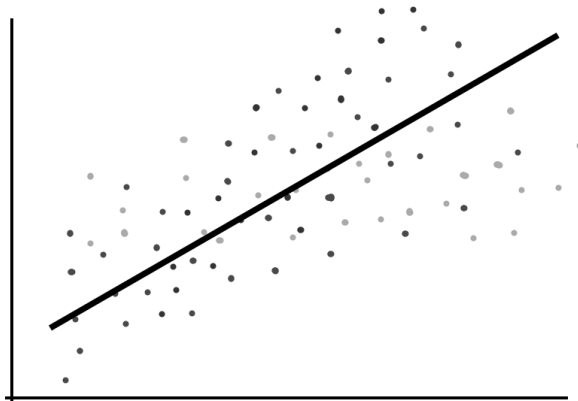


FIGURE 2: Simple OLS

# EMPIRICAL FRAMEWORK (INTUITION) II

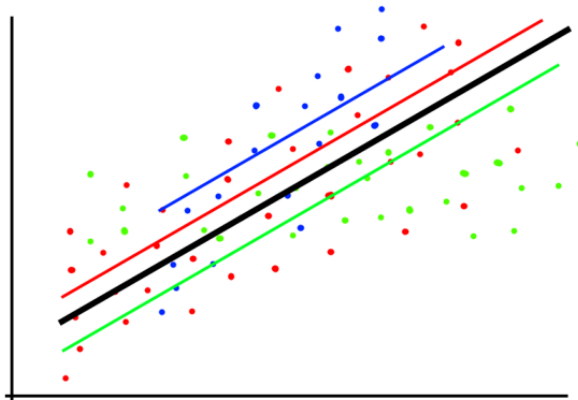


FIGURE 3: Multiple OLS with homogeneous return to Educ

# EMPIRICAL FRAMEWORK (INTUITION) III

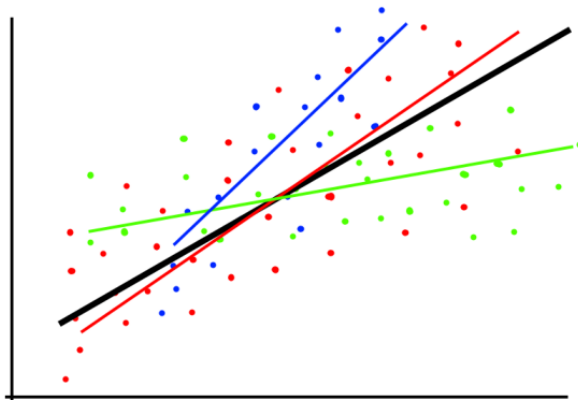


FIGURE 4: Correlated Random Coefficient Model

# DISTINCTION TO CONVENTIONAL METHODS

- OLS
  - ability and “background” bias
- IV Methods:
  - suitable if assume homogeneous returns to education.
  - if education is correlated with **unobserved individual heterogeneity**, IV methods may fail to identify APE.
    - alternative: **Local Average Treatment Effect** if interested in effect of educational policy reforms.

# CONDITIONAL MEAN INDEPENDENCE

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

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

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

- $X_i$  should be “good predictors” of treatment  $S_i$  (Wooldridge 2004, pg.7).
- (A.1): Redundancy of  $X_i$  given  $a_i$  and  $b_i$  and  $S_i$ .
- (A.2): In the first two conditional moments of  $S_i$ ,  $a_i$  and  $b_i$  are redundant  $\rightarrow$  “Staying longer in Education is determined by  $X$  covariates”.

# ESTIMATOR FOR $\beta$ AND GLM

The **APE** can be estimated by:

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

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

Where  $\sigma^2$  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).
- CF approach can identify APE in heterogeneous returns while standard IV approach may not.
- Similar to Heckman two-step estimator.
- Models schooling choice explicitly in first step

**First stage:** modelling schooling choice

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

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

- $v_i$ ,  $\varepsilon_{ai}$  and  $\varepsilon_{bi}$  are normally distributed with zero means and positive variances, that are possibly correlated
- $v_i$  is positive if an individual acquires higher education than expected conditional on observed characteristics

**Second stage:** augmented wage equation

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

where:

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

# CONTROL FUNCTION APPROACH III

Interpretation of the coefficients of the control functions

- $\gamma_1$  measures the effect of those unobserved factors that led to over- or under-achievement in education on the wage
  - Thus, if  $\gamma_1$  is positive, the unobserved factors affect schooling *and* wages positively
- $\gamma_2$  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

# REPLICATION AND COMPARISON

# REPLICATION AND COMPARISON

# CODES

```
*** Generalized linear regression model with Poisson distributio  
    glm school sex age age_sq rural edu_f occ_f edu_m, family(po
```

TODO: play with codes

# SET-UP

- We use the same sample: West Germans (not foreign-born or self-employed) between 25 and 60 years who work full-time
- We have less observations than Gebel and 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
- control variables: age and age squared, gender, father's education, mother's education, father's occupation, rural or urban household, number of siblings (as instrument)

# RESULTS COMPARISON AND CONCLUSION



# RESULTS

# ESTIMATED RETURNS ON EDUCATION

- Estimates from OLS and 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 and original estimations - Our OLS estimates are on average larger than those of Gebel and Pfeiffer (2010) by 0.004 percentage points - Our CMI estimates are on average larger than those of Gebel and 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 and -0.23)
- We would assume that the instrument does not directly affect the error term in the wage equation

## Coefficients of the control functions

- $\gamma_1$  is negative for majority of years, yet very small and insignificant in all years
  - Gebel and Pfeiffer (2010) estimate a positive coefficient in the 1980s and 1990s - but also insignificant
- $\gamma_2$  is negative and close to zero for most years
  - Indicates that those with unexpectedly high education have lower returns to education

# CONTROL FUNCTION ESTIMATES II

- 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 and Pfeiffer (2010)
- that both coefficients are (mostly) negative hints that educational expansion caused more  $\hat{\alpha}$  less able  $\hat{\alpha}$  to achieve higher education

# DIVERGENCES BETWEEN REPLICATION AND GEBEL AND PFEIFFER (2010)

- sample not the same
- large variance of CF approach

# RESULTS COMPARISON I

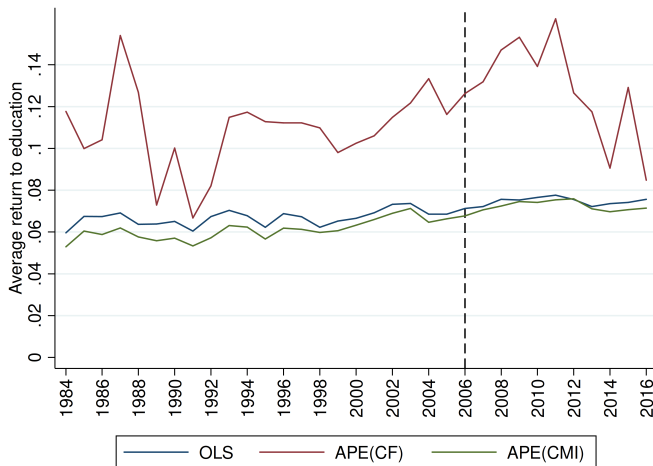


FIGURE 5: Replication results: Comparison between OLS, CMI and CF approaches

# RESULTS COMPARISON II

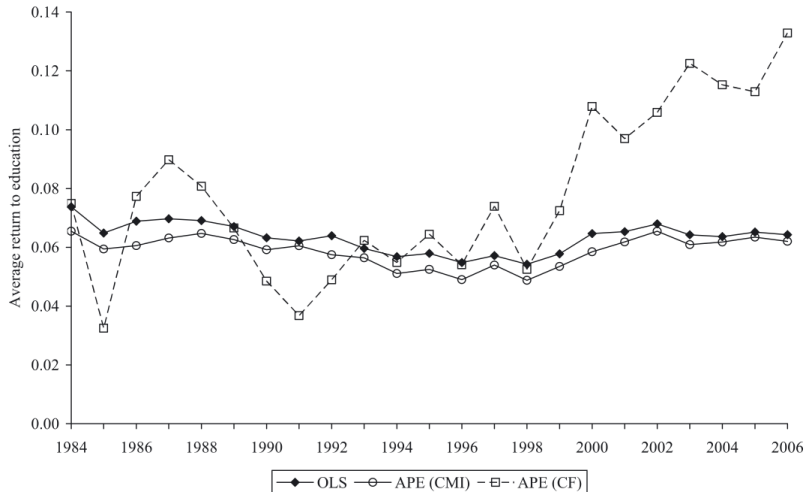


FIGURE 6: Original Results (GP 2010, pg.30)

# RESULTS COMPARISON III

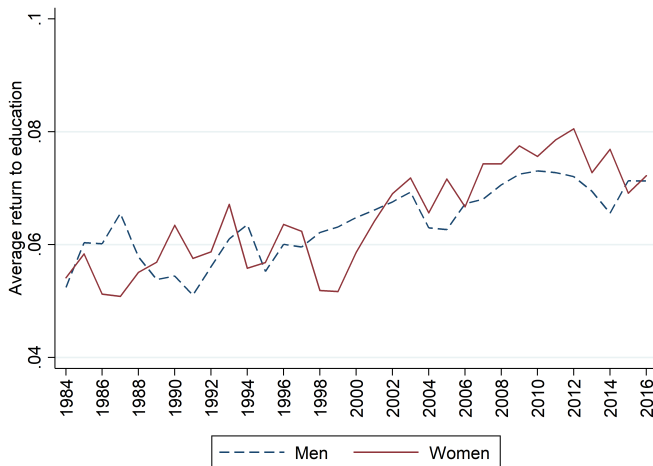


FIGURE 7: Replication results: **APE** by gender (CMI approach)



# RESULTS COMPARISON IV

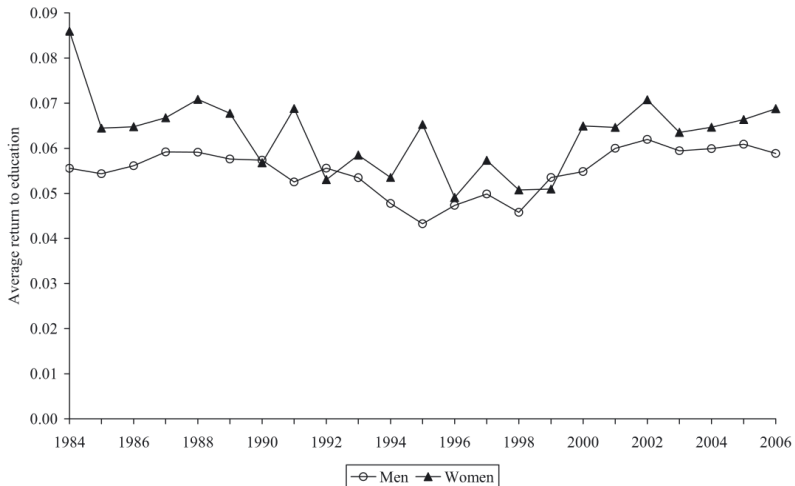


FIGURE 8: Original Results (GP 2010, pg.34)

# RESULTS: CONTROL FUNCTION (REPLICATION)

year	First Stage		Second Stage					
	IV: Nr. of Siblings		$v_i$			$v_i S_i$		
	coef.	s.e.	coef.	s.e.	p	coef.	s.e.	p
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

# RESULTS: CONTROL FUNCTION (REPLICATION)

# CONCLUSION

# PRO'S AND CON'S OF ESTIMATION METHODS

- CMI
  - no analytical standard errors
- CF
  - requires further distributional assumptions on error terms
  - valid and relevant “instrument”

# THE END I

TODO: add