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

Outline

- Theoretical Framework
  - Gebel & Pfeiffer (2010)
  - Returns to education
- Emirical framework
  - Correlated random coefficients model
  - Conditional Mean approach
  - Control funtion approach
- Replication
  - Set-up
  - Code
  - Comparison of results

Theoretical Framework

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:
  - 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

Background Information I

Increase in educational attainment

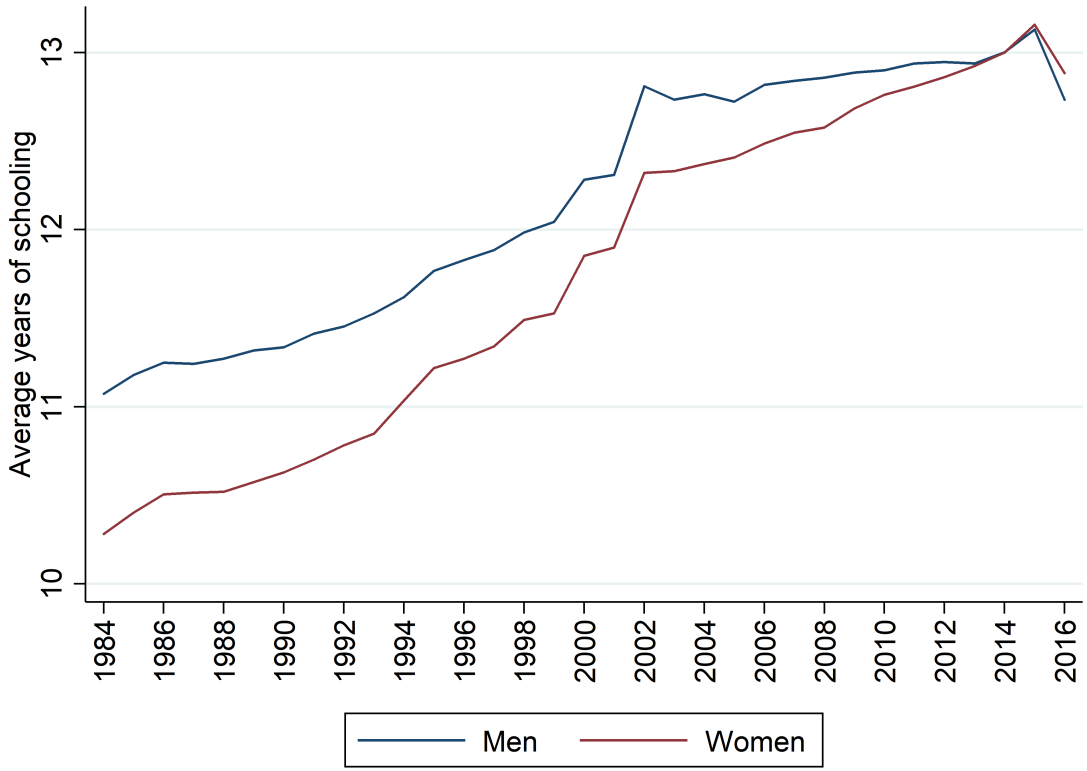


Figure 1: Source: SOEP 1984-2016, own estimations.

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** ε_{ai} and ε_{bi}.
- Do 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 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 ε_{ai} correlates with S_i

Empirical Framework (Intuition) I

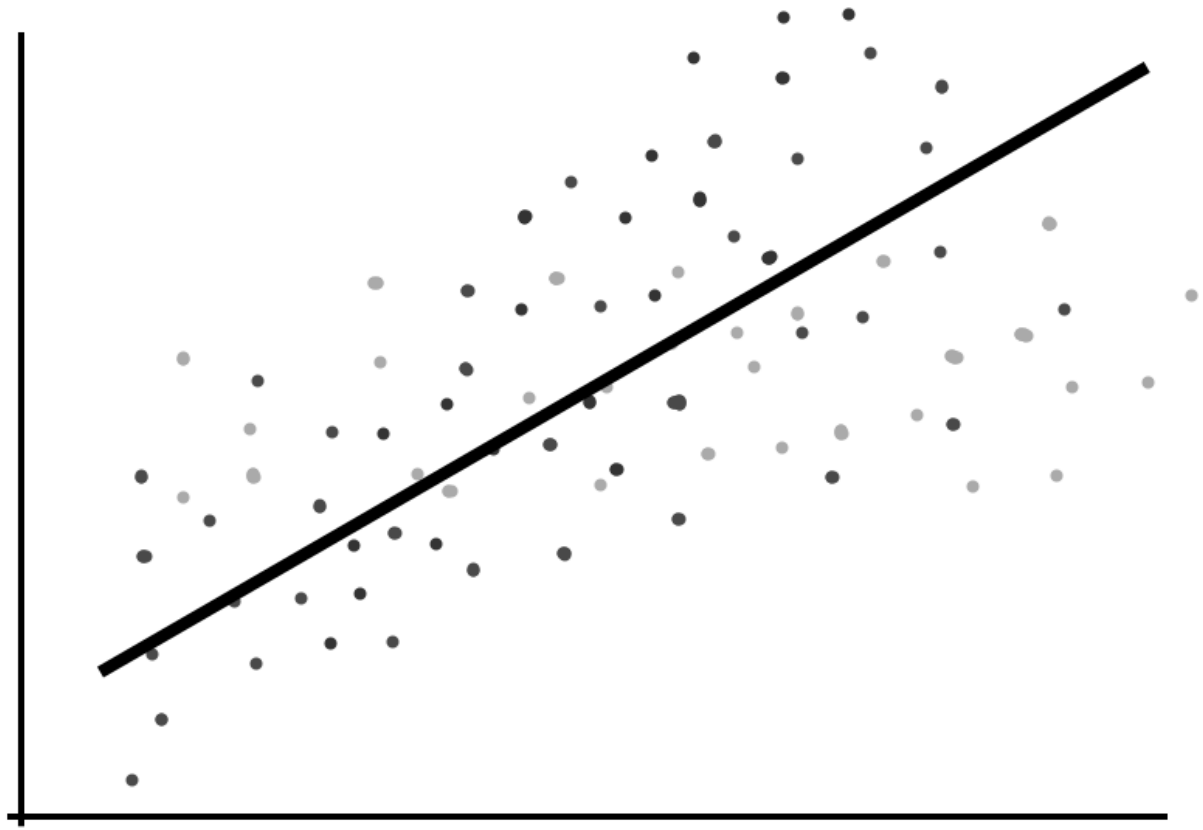


Figure 2: Simple OLS

Empirical Framework (Intuition) II

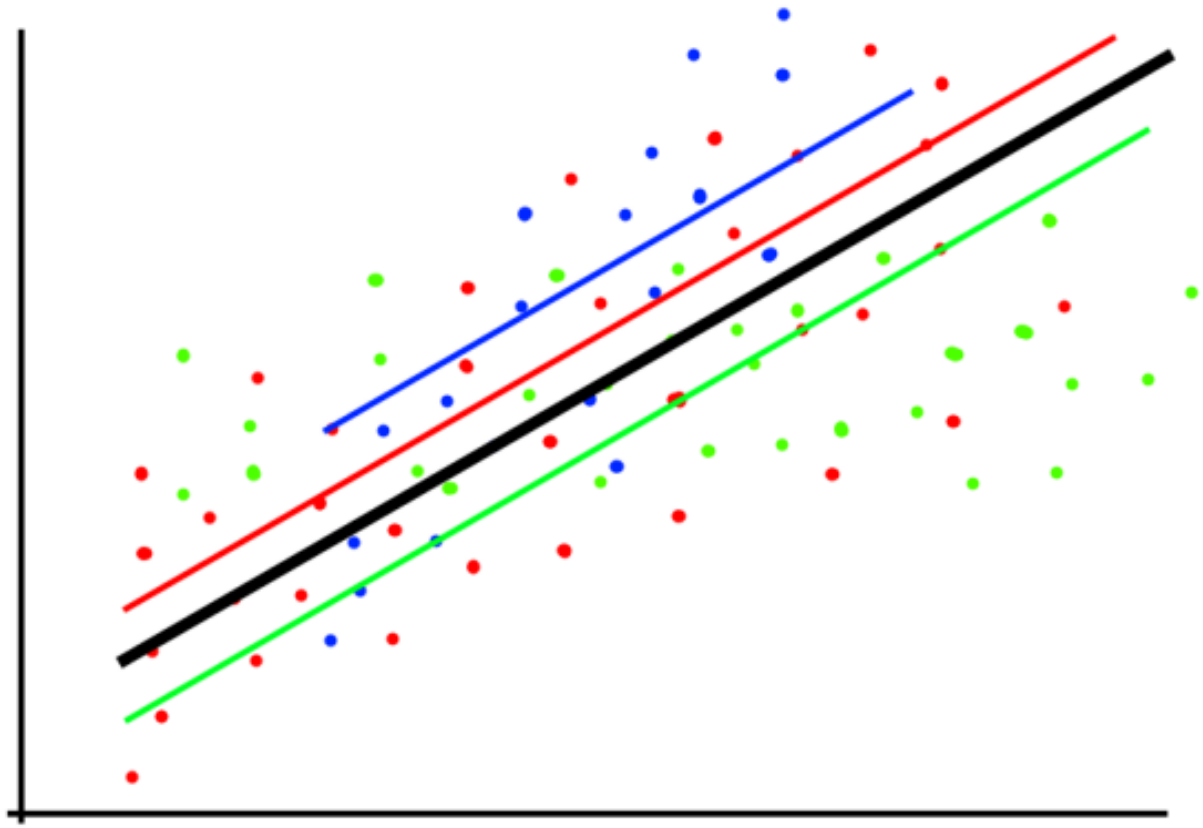


Figure 3: Multiple OLS with homogeneous return to Educ

Empirical Framework (Intuition) III

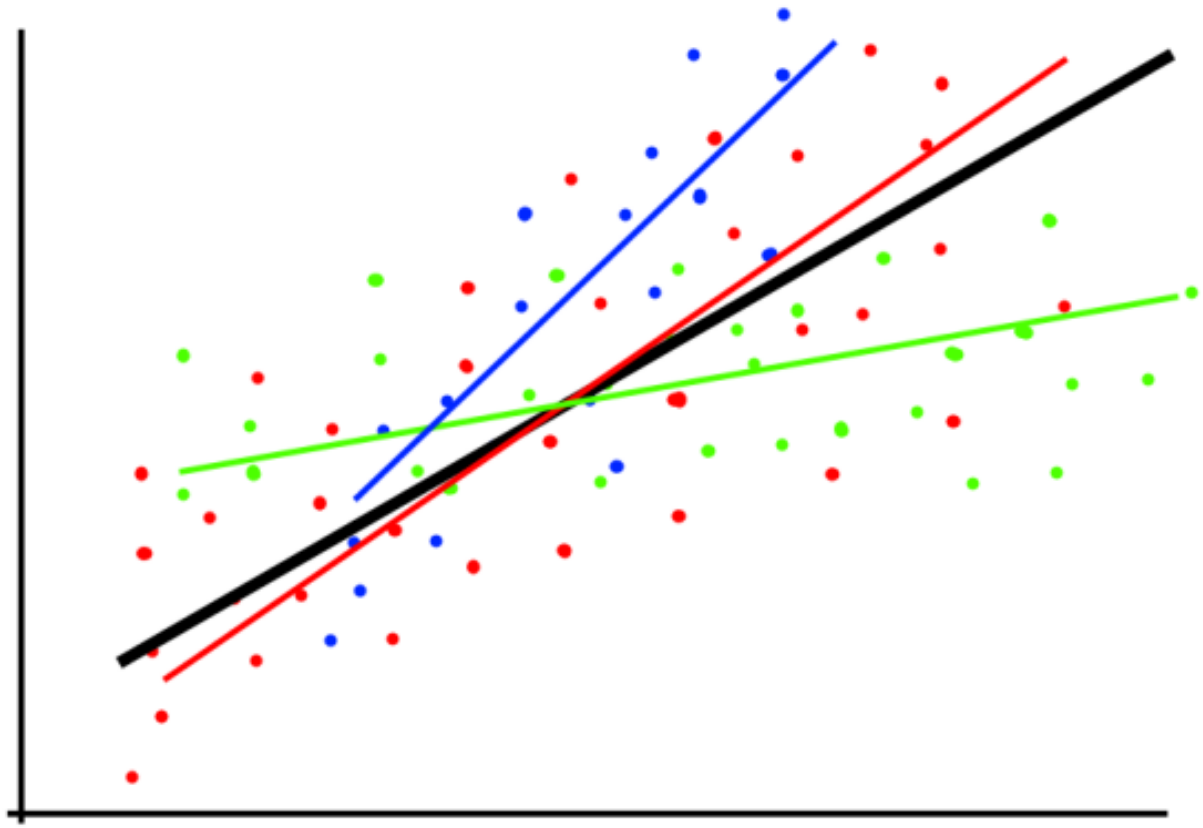


Figure 4: Correlated Random Coefficient Model

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 identity APE.
  - alternative: **L**ocal **A**verage **T**reatment **E**ffect if interested in effect of educational policy reforms.

Conditional Mean Independence

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

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

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

- 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 –> "Staying longer in Education is determined by X covariates".

Estimator for β and GLM

The **APE** can be estimated by:

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

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

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

Control Function Approach {.allowframebreaks}

- 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 step: modelling 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 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

Replication and Comparison

Replication and Comparison

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)

CODES

TODO:codes

```
*** GLM regression with Poisson distribution
glm school sex age age_sq rural edu_f occ_f edu_m, ///
    family(poisson) link(log)

*** Predict conditional mean and extract pearson residuals
predict condMean, mu
predict res_pears, pearson

*** Calculate residual
gen resid = school - condMean

*** Estimate sigma^2
egen sigma_sq_pears = mean(res_pears^2)

*** generate APE
egen bCMI = mean((resid*lnw)/ (sigma_sq_pears*condMean))
```

CODES

```
*** Generalized linear regression model with Poisson distribution
quietly glm school sex age age_sq rural edu_f occ_f edu_m, family(poisson) link(log)

*** Predict conditional mean on observable characteristics x
* (expected value of school) and pearson residuals
predict condMean, mu
predict res_pears, pearson

*** Calculate resid as difference between prediction & observed school
gen resid = school - condMean

*** Estimate sigma_sq
egen sigma_sq_pears = mean(res_pears^2)

egen bCMI = mean((resid*lnw)/ (sigma_sq_pears*condMean))
```

CODES

```
* First step: Estimate the reduced form of schooling, i.e. regre
* on all exogeneous variables including the instrument (siblings
reg school sex age age_sq rural edu_f edu_m occ_f sibl if syear=

* Obtain the residuals
predict v`n', res

* Second step:  Estimate the structural equation and include the
* the reduced form as an additional regressor
qui: reg lnw school sex age age_sq rural edu_f edu_m ///
    occ_f v`n' c.v`n'#c.school if syear==`n' // exclude instrume
```

Results comparison and conclusion

Results

Results Comparison I

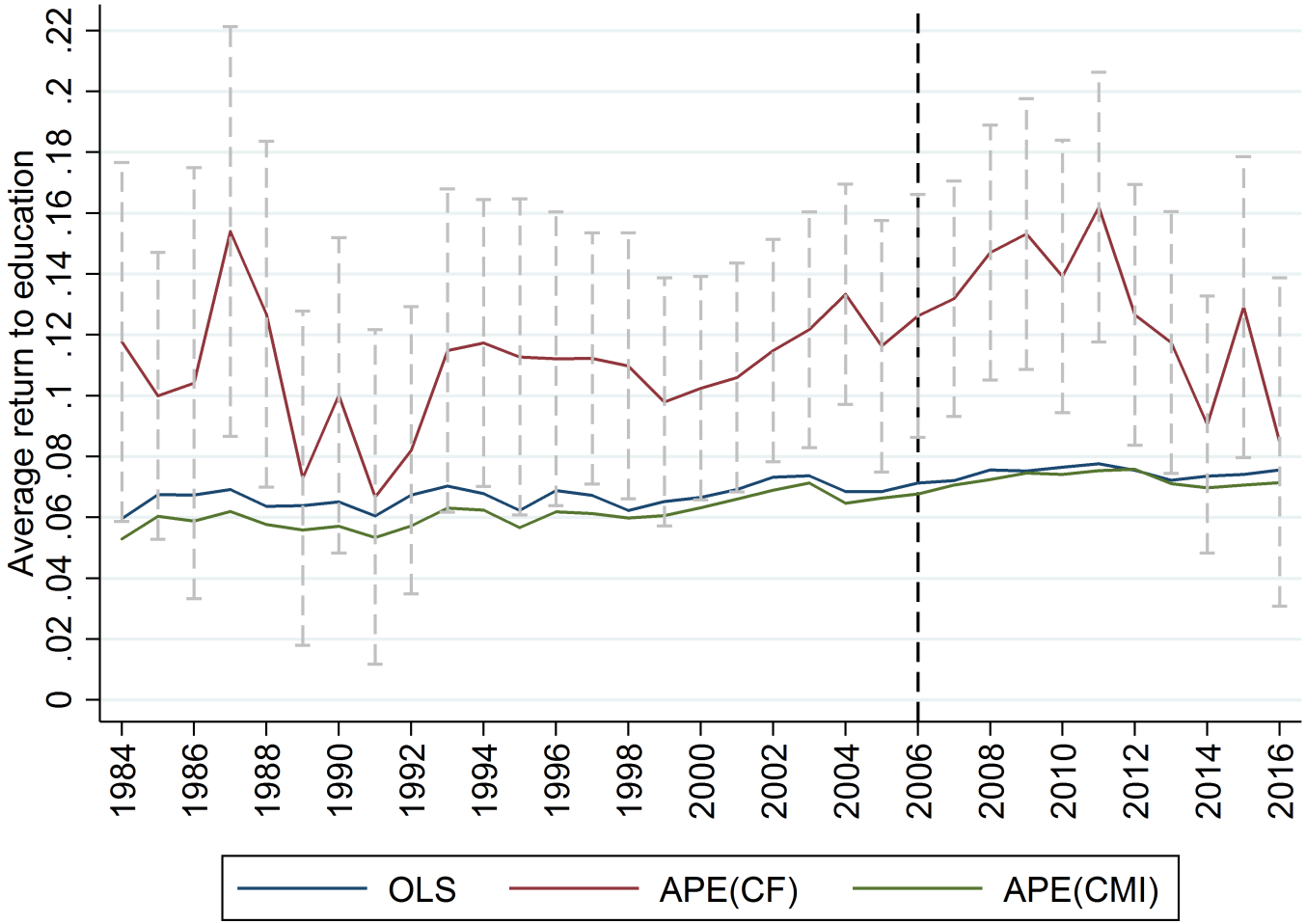


Figure 5: Replication results: Comparison between OLS, CMI and CF approaches

Results Comparison II

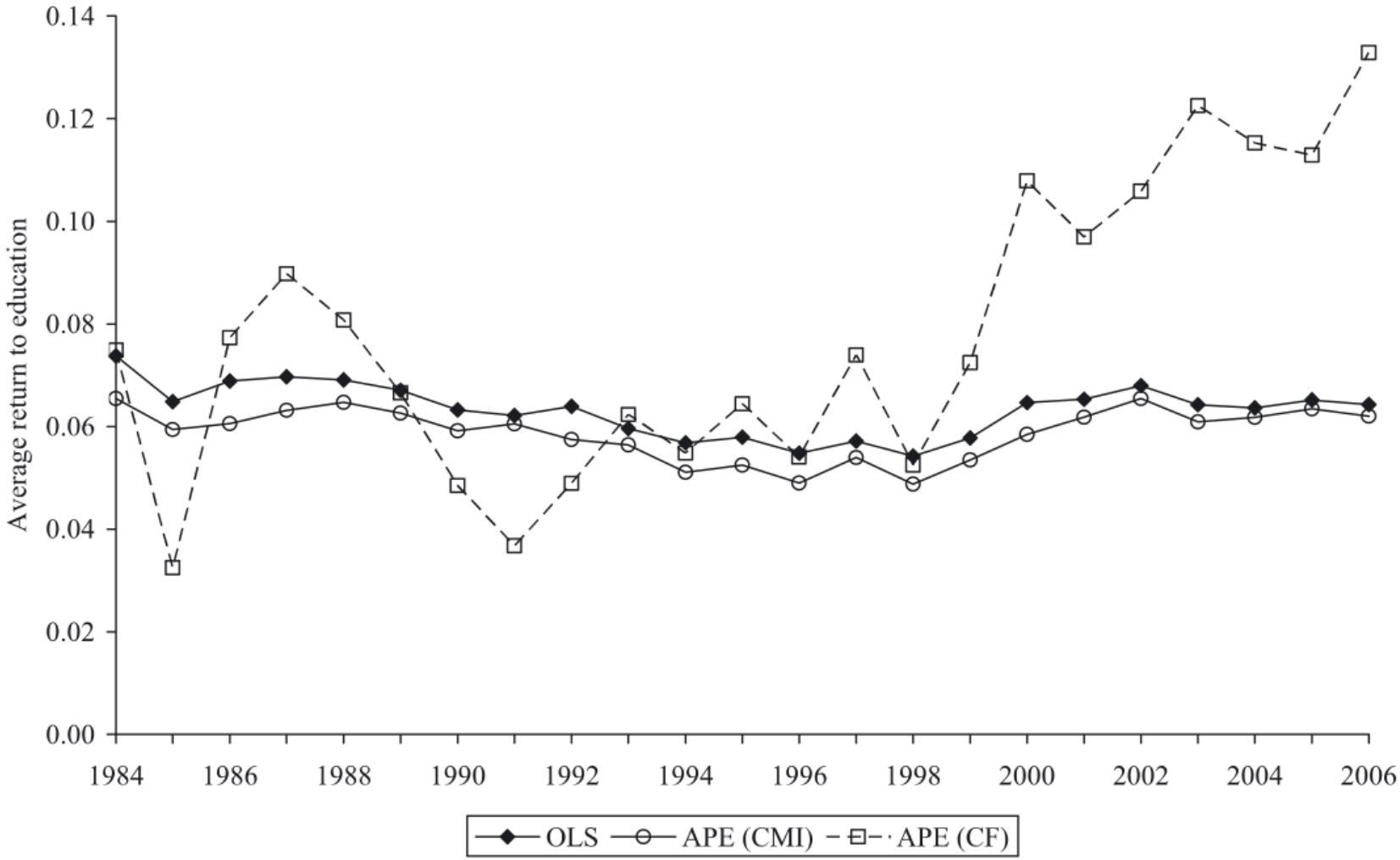


Figure 6: Original Results (GP 2010, pg.30)

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 step

- *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

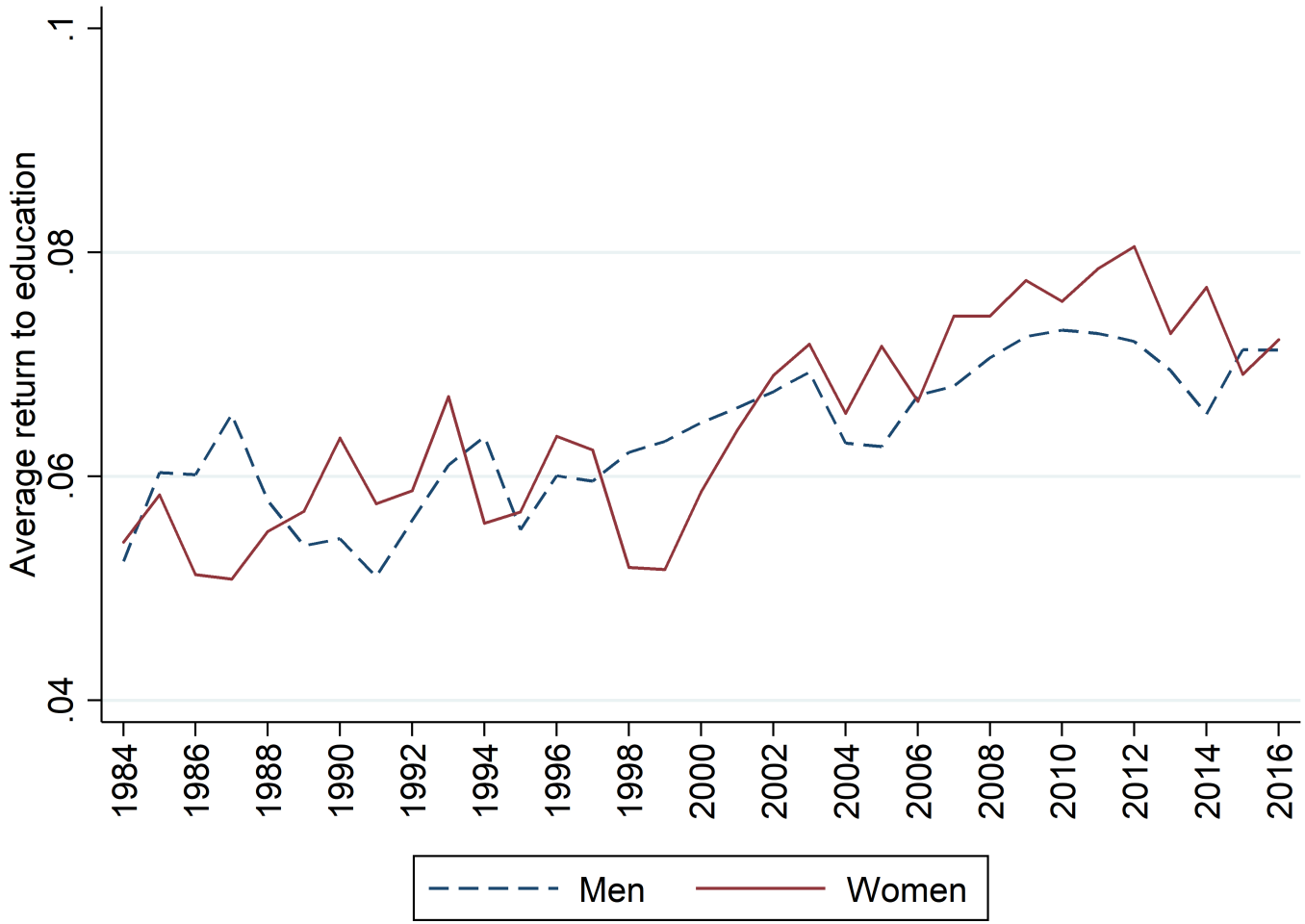
- γ_1 is negative for majority of years, yet very small and insignificant in all years
  - Gebel and Pfeiffer (2010) estimate a positive coefficient in the 1980s and 1990s - but also insignificant
- γ_2 is negative and close to zero for most years
  - Indicates that those with unexpectedly high education have lower returns to education

Control function estimates II

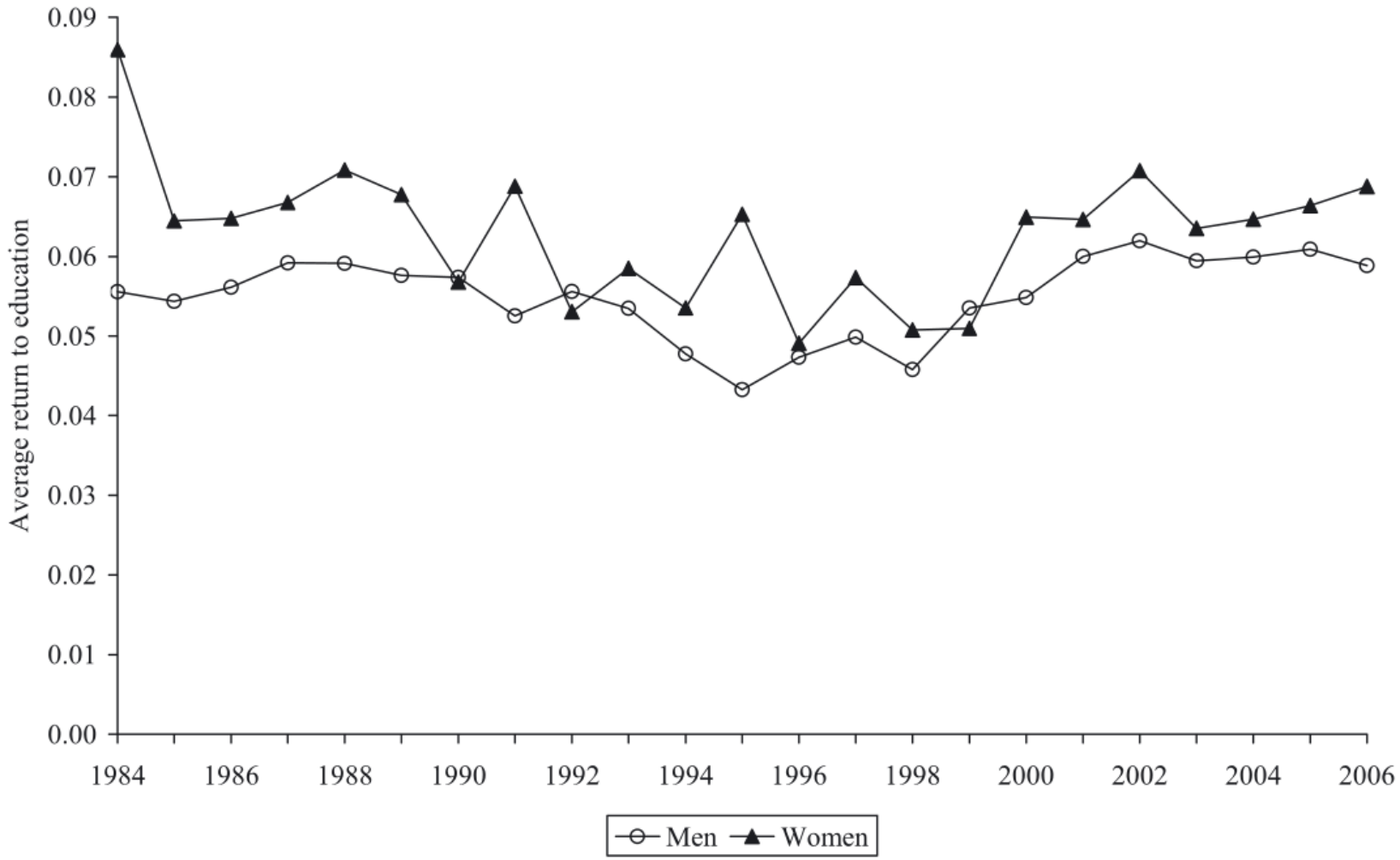
- Similarly, they are only slightly significant in the 1980s, and stronger significant in the early 2000s
- The estimates are very similar to those of Gebel and Pfeiffer (2010)

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

Heterogenous returns to education by gender I



Heterogenous returns to education by gender II



Results: Control Function (replication)

Table 1: Summary of Control Function estimates (replication)

	First Stage		Second Stage					
	IV: Nr. of Siblings		v_i			$v_i S_i$		
year	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

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