# Econometric Methods Project

Mason Veilleux

January 23, 2022

## 1 Abstract

This project centers on estimating the returns to education with time-invariant variables. I follow [Hausman and Taylor, 1981] and the alternative [Amemiya and MaCurdy, 1986] to compare OLS and Random Effects estimates. My data covers prime age adults from 1981-1990 using the PSID. I find that my estimates are higher than those in HT, [Cornwell and Rupert, 1988] and [Baltagi and Khanti-Akom, 1990]. This higher estimate may be due to possibly (1) a secular rise in the returns to education or (2) differences in the accounting for wage. I find that the returns to one year of education is 29%.

## 2 Introduction

When facing the possibility of independent variables being correlated with the error term[1] in panel data, the first step is to use a fixed effects model. This method may help resolve issues such as individual effects which are time-invariant. However, so are variables like sex or education. This suggests that when estimating the returns to education, coefficients are unattainable since $S_{it} - \bar{S}_i = 0$ i.e. The HT model provides an alternative.

## 3 Literature

According to [Card, 2001], the returns to education literature studies the causal relationship of supply-side variables on the demand-side of the education market. Coined in Becker's *Human Capital* (1967) and estimated in [Griliches, 1977], the new wave of causal research attempted to solve this question using various forms of identification – finding estimates typically around 6% to 15%.

Card mentions HT which use both random effects and IV on PSID data from 1968-1972. The innovative identification strategy uses exogenous observables as instruments for endogenous variables– therefore avoiding the FE trap.

Following, papers such as AM, and [Breusch et al., 1989] have contributed more robust methods. AM adds that the IV should be in $T + 1$ while BMS argue they should be in $T - 1$. BMS is the most efficient and HT the least.

These three models are estimated and compared in both CR and BK which use PSID data from 1976-1982. CR finds efficiency gains are limited to time-invariant coefficients. BK reproduces CR's

[1] error term as $v_{it} = \mu_{it} + \alpha_i$ where $\mu_{it}$ is the idiosyncratic error term and $\alpha_i$ is the individual effects

results and finds much smaller efficiency gains. From this, it seems that the HT estimator is a first step and the AH and BMS estimators are used to increase efficiency.

For robustness, a Hausman test is not an efficient test for over-identification. According to [Ahn and Low, 1996], the test does not account for time-invariant bias with $\alpha_i$. Not only that, but when instruments outnumber endogenous variables, the Hausman test becomes even more inefficient. To compensate, I use the Sargan-Hansen statistic.

# 4 Hausman Taylor (1981) Model

When faced with panel data where we assume endogeneity from individual effects, we can identify an equation:

$$y_{it} = \beta_0 + \beta_1 X1_{it} + \beta_2 X2_{it} + \gamma_1 Z1_i + \gamma_2 Z2_i + \alpha_i + \mu_{it} \tag{1}$$

where $y_{it}$ is the log wage for person $i$ in year $t$ and $\alpha_i$ represents individual effects. Below, I have categorized the $X$ and $Z$ variables used in the analysis:

| | | $X_{it}$ | $Z_i$ |
|---|---|---|---|
| | | Time-Variant | Time-Invariant |
| 1 | Exogenous | experience, employment status$_{t-1}$ | sex, race, region, union coverage |
| 2 | Endogenous | hours worked, marriage status | education, occupation |

Representing this relationship in a DAG[2]:

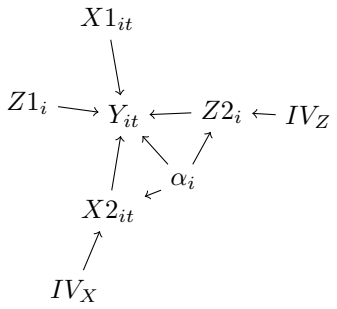


Figure 1: Directed Acyclic Graph from Hausman Taylor (1981 model)

Here the assumptions are laid out. Exogenous variables are chosen to be independent to all variables. I assume that ability has nothing to do with the quantity or quality of experience, it is given. Employment status$_{t-1}$ is a decision of one to supply their labour outside of individual effects. Sex, race, region, and union coverage all have nothing to do with $\alpha_i$. For example, boys are not more gifted than girls, and no race or region bestows any gifts to workers.

For endogeneity, hours worked is a function of one's ability. This may be because workers who decide to work less do so because they are not so good at working and therefore choose to limit consumption in exchange for more leisure. Or, for couples, one worker may be more productive

[2]IV's will be explained in section 4.1

than the other and therefore the other limits their labour supply. Additionally, in terms of marriage status, I assume males on a steep career track (measure of ability) are more likely to marry given the findings in [Ludwig and Brüderl, 2018]). For occupation, I assume that the mean in ability increases between occupations. Last is the canonical endogeneity of education where ability determines how students continue education. Note that $\alpha_i$ affects the wage since employers may be able to spot out talent and therefore give high ability workers a higher wage for their potential.

### 4.1 Between RE and FE: the IV

HT splits the treatment of variables into two groups. All exogenous (X1 and Z1) variables are treated with random effects. All endogenous (X2 and Z2) variables are given 'internal' instruments.To close the backdoor of the confounding variable, X2 variables are instrumented by deviations from the group (in the spirit of FE). Z1 variables serve as their own instrument. Group means of X1 are the instrument for Z2. These instruments make for being highly relevant and satisfy the strict exogeneity assumption since the exploited variation in the endogenous variables is channelled from FE variation[3]. In short, RE is applied to exogenous variables and FE instruments are used to exploit variation in endogenous variables.

## 5 Data

I use the PSID along with the Family Public Data Index to select the appropriate variables[4]. I follows 1,591 individuals from 1981 to 1990 (n=15,910). Ages range from 25-55 and experience is calculated as Age-Grad Year- 5 — both identical to HT. A non-positive condition is applied to education and hours worked and education. Additionally, income ranges from $1,000 to $100,000. This is to avoid any long tails and outliers. I include asset income since stock options and other types of assets are forms of remuneration. High-wage earners may have much higher ratios of asset to labour income due to remuneration at work and this may bias results. The PSID does not distinguish between asset and labour income when the individual earns both[5].

Time variance in the education variable is an issue. This is due to misrepresenting through the years (i.e in 1981, individual stated highest grade attained was 7th grade. In 1987 stated grade was 5th grade or vice versa)[6]. Not only misreporting is a possible issue, but also continuing education. Years of education increase due to a professional degree or a union offering specialized skills in a workshop. Typically, these individuals are likely to see higher returns. However, due to the chance of misreporting, I assume that all education is as it was in 1985. 1985 is assigned randomly and is the year chosen for all confidently assumed time-invariant variables[7].

## 6 Results

The education variable is statistically significant in the HT model. Here a one year increase in schooling equates to a 29.7% increase in the wage. This is more than twice the OLS estimates. Similarly, the AM reports a 22.4% increase and a much smaller standard error.

[3]HT mentioned the use of family background variables as instruments. However they broke the strict exogeneity assumption since parents can transmit individual effects (i.e high ability parent creates high ability child. See [Griliches, 1977]

[4]please see the figures file for summary statistics

[5]Using total income rather than just labour income may be why my estimates are high

[6]between 1981 and 1990, 33% reported increases in education, and 7% reported decreases

[7]In Stata, all variables are identified with '85' at the end of each variable

I compare the education coefficient to that in HT, CR, and BK in table 2. For all models, my estimates are much higher. The BK RE/GLS model is roughly the same as mine while I am more than ten points higher for both AM and HT models. These higher estimates may be because I used total income not labour income. In alternative analyses, I changed endogeneity assumptions similar to BK and could not find an estimate lower than 25%. This seems high even though I used common controls.

We also see that employment status on wages is around 12.4% - a drastic drop from 34.5% in the OLS estimate. This is likely because employment is persistent. Once controlling for year effects, the estimate drops significantly. Being employed in the previous year means that your tenure has allowed you to earn a higher wage not because of time, but because you were not unemployed. This shows the aggregate importance of keeping steady work on wages.

When controlling for $\alpha_i$ and year effects, we see that the returns to schooling gender gap increases. From [Cortes and Pan, 2020]) we understand that the differences in gender pay is mostly child-related (see Figure 3 in CP). Other reasons may be due to discrimination, productivity differences, or compensating differentials.

Lastly, we see that being single reduces your wage. This 'marriage premium' is sometimes related to a bias in the market where employers are willing to pay married workers more because they have families to feed. An alternative view from [Ludwig and Brüderl, 2018] show that this effect is not causal since "married [individuals] earn more because selection into marriage operates both on wage levels and wage growth". Therefore, those on a steep career track are especially likely to marry.

We observe that experience after controlling for $\alpha_i$, experience returns 5 percentage points more than the Pooled OLS estimate. This is puzzling since the RE/GLS estimate is similar and the Pooled OLS controlled for year effects. Experience$^2$ does not seem to have changed.

When controlling for individual and year effects, race has no effect on the returns to education. Additionally, neither does region. Occupation and union coverage have a large effect, 36.1% and 14.8% in the AM model and no other effect any the other models (see figures file for brief analysis of occupation and $\alpha_i$).

The constants from each model are inconsistent compared to CR's. someone who has no education, experience, or hours worked, and identifies with all reference categories earns a negative log(wage) at -7.87 or \$0.00038 an hour. The CR HT estimated constant is 0.1492 or \$-1.902468. This inconsistency likely means little. But I am inclined to say that my estimate makes more sense since those who do not work do not earn a wage, not a negative wage.

# 7 Robustness Check

I find that the HT model is linear in parameters, there is no heteroskedasticity, no perfect collinearity, and $E[\mu_{it}, \alpha_i | X, Z] = 0$. The RESET test finds omitted variable bias up to $\hat{y}^3$. This is problematic and unresolved. Further, I reject the null in the Hausman test between the RE and FE model ($\theta = .766$). For over-identification (order condition $m > k$), I run the Sargan-Hansen test. I fail to reject the null. Relevance assumptions are seen in table 3. Both models fail to have F-statistics greater than 10. However, the AM model's is higher. Therefore I fail to maintain the relevance assumptions.

# 8 Conclusion

I find that a one year increase in education accounts for a 29.7% increase in wages. The model fails the Ramsey test and the instruments are weakly identified. The AM model provides a more

efficient estimate of 22.4%. Some unanswered questions: how do researchers account for time-variant education. Also, why does the individual effects error term correspond positively with wages and negatively with education? Is this the discount factor effect [Card, 2001] talks about?

# 9 Appendix

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| DV:(ln)wage | OLS | Pooled OLS | RE/GLS | FE | HT | HT Education TV | AM |
| Education | 0.107*** | 0.099*** | 0.134*** | - | 0.297*** | 0.067*** | 0.224*** |
| | (0.002) | (0.002) | (0.005) | | (0.080) | (0.006) | (0.019) |
| $\text{Employed}_{t-1}$ | 0.272*** | 0.262*** | 0.116*** | 0.102*** | 0.102*** | 0.099*** | 0.102*** |
| | (0.022) | (0.022) | (0.016) | (0.016) | (0.016) | (0.016) | (0.016) |
| Hrs Wored | -0.000*** | -0.000*** | -0.000*** | -0.000*** | -0.000*** | -0.000*** | -0.000*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Male | 0.300*** | 0.293*** | 0.385*** | - | -0.347 | -1.475 | 0.194** |
| | (0.014) | (0.014) | (0.032) | | (0.920) | (2.009) | (0.088) |
| Single | -0.088*** | -0.093*** | -0.027** | -0.020 | -0.020 | -0.022 | -0.020 |
| | (0.012) | (0.012) | (0.014) | (0.014) | (0.014) | (0.014) | (0.014) |
| Experience | 0.054*** | 0.038*** | 0.086*** | 0.090*** | 0.090*** | 0.086*** | 0.090*** |
| | (0.003) | (0.003) | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) |
| $\text{Experience}^2$ | -0.001*** | -0.001*** | -0.001*** | -0.001*** | -0.001*** | -0.001*** | -0.001*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Race | -0.115*** | -0.117*** | -0.115*** | - | 0.014 | -0.177 | -0.046 |
| | (0.007) | (0.007) | (0.019) | | (0.147) | (0.305) | (0.046) |
| Region | 0.027*** | 0.027*** | 0.025*** | - | -0.046 | -0.048 | -0.004 |
| | (0.003) | (0.003) | (0.007) | | (0.062) | (0.155) | (0.017) |
| Occupation | -0.005*** | -0.006*** | -0.001 | - | 1.306 | 3.272 | 0.361*** |
| | (0.002) | (0.002) | (0.005) | | (1.579) | (3.423) | (0.121) |
| Union Covered | 0.123*** | 0.122*** | 0.122*** | - | 0.161 | 0.034 | 0.148*** |
| | (0.007) | (0.007) | (0.017) | | (0.130) | (0.303) | (0.040) |
| Year Effects | - | 0.037*** | - | - | - | - | - |
| | | (0.001) | | | | | |
| Constant | -0.482*** | -73.729*** | -1.205*** | 1.359*** | -7.699* | -10.258 | -3.735*** |
| | (0.070) | (2.751) | (0.120) | (0.054) | (4.429) | (10.574) | (0.595) |
| Observations | 15,910 | 15,910 | 15,910 | 15,910 | 15,910 | 15,910 | 15,910 |
| R-squared | 0.341 | 0.369 | - | 0.304 | - | - | - |
| Number of id | - | - | 1,591 | 1,591 | 1,591 | 1,591 | 1,591 |

Standard Errors are heteroskedastic-consistent (HC1) robust
Stata automatically clusters standard errors on id for HT and AM models
RE model: $\theta = 0.767$
*** p<0.01, ** p<0.05, * p<0.1

Table 1: Results

| | HT (1981) | CR (1988) | BK (1990) | Veilleux |
|---|---|---|---|---|
| RE/GLS | 0.068 | 0.094 | 0.137 | 0.134 |
| | (0.005) | (0.017) | (0.014) | (.005) |
| HT | 0.217 | 0.200 | 0.140 | 0.297 |
| | (0.098) | (0.078) | (0.066) | (0.080) |
| AM | — | 0.159 | 0.155 | 0.224 |
| | | (0.059) | (0.048) | (0.019) |

Table 2: Education Estimate

| | HT | AM |
|---|---|---|
| Cragg-Donald Wald F statistic | 1.41 | 7.38 |

Table 3: First Stage F-Statistic: Models Under-identified

# References


[Ahn and Low, 1996] Ahn, S. C. and Low, S. (1996). A reformulation of the hausman test for regression models with pooled cross-section-time-series data. *Journal of Econometrics*, 71(1-2):309–319.

[Amemiya and MaCurdy, 1986] Amemiya, T. and MaCurdy, T. E. (1986). Instrumental-variable estimation of an error-components model. *Econometrica: Journal of the Econometric Society*, pages 869–880.

[Baltagi and Khanti-Akom, 1990] Baltagi, B. H. and Khanti-Akom, S. (1990). On efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied econometrics*, 5(4):401–406.

[Breusch et al., 1989] Breusch, T. S., Mizon, G. E., and Schmidt, P. (1989). Efficient estimation using panel data. *Econometrica: Journal of the Econometric Society*, pages 695–700.

[Card, 2001] Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160.

[Cornwell and Rupert, 1988] Cornwell, C. and Rupert, P. (1988). Efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied Econometrics*, 3(2):149–155.

[Cortes and Pan, 2020] Cortes, P. and Pan, J. (2020). Children and the remaining gender gaps in the labor market. Technical report, National Bureau of Economic Research.

[Griliches, 1977] Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica*, 45(1):1–22.

[Hausman and Taylor, 1981] Hausman, J. A. and Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica: Journal of the Econometric society*, pages 1377–1398.

[Ludwig and Brüderl, 2018] Ludwig, V. and Brüderl, J. (2018). Is there a male marital wage premium? new evidence from the united states. *American Sociological Review*, 83(4):744–770.