# **Impact of Education on Wages**

Meho Porobic

School of Management, Technical University of Munich

Advanced Seminar: Quantitative Economic Research

Michael Mühlegger

Prof. Dr. Helmut Farbmacher

May 24th 2024

**Author Note**

I have slightly digressed from the classic APA structure in order to provide a better overview for the given assignment and to align with the given instructions regarding formatting.

# **Introduction**

Human capital theory in a broad sense encompasses the central idea that human capital and thus labor output varies across individuals due to some underlying characteristics. These intangible characteristics include skills, training, ability, formal education, and other that impact economic productivity and with that the earnings of an individual. Some of the mentioned characteristics fall into the internal locus of control for an individual which then implies the possibility, along with the implied incentive, of both the individual and the environment (state, employer) to treat the controllable factors as a form of long-term investments. In that regard, the focal point of this research is going to be a relationship between *education* and the returns on it in the form of future *wages*.

This relationship is among the “most robust findings in social science” (Deming, 2022, p. 75). This is not surprising when we take into account the explosive expansion of the field of human capital since the nineties of the past century (Deming, 2022, p.76). Consequently, the expansion of research itself on this topic is hardly surprising, when the observed increase in education of the world population has skyrocketed since the 1950s; the proportion of those with secondary education went from 13% to 51%, and for those with tertiary 2.2% to 14.6% (Lee & Lee, 2016, as cited in Deming, 2022, p.75). In the same manner public spending on education around the world has increased with a slightly steeper upsurge in developed countries (Roser & Ortiz, 2016 as cited in Deming, 2022, p.75).

With this in mind, it is only natural ask why is this so? Well, it is nothing but a sensible response of both individuals and states to the underlying incentives empirically revealed by the studies. The studies emphasize education and its ability to explain the wage variance both within and between countries, accounting for up to one-third and one-half of variance respectively (Deming, 2022, p.82).

The field of human capital has come a long way since its early inception with the work of Becker (1961; see Deming, 2022). This was not without the criticisms on numerous grounds. From the very terminology that might insinuate workers as assets of the capital owner, or the failure to address inequality and the socio-economic obstacles that come with it, to the overemphasis on formal education and the economic returns, while ignoring the quality as an important aspect of said education, non-economic benefits of learning, and the possible effects of social capital factors on wages. These critiques are all sound and some have compelled research to new and uncharted areas, but they all fall out of scope for this work.

This work zeroes in on the wages regressed on the education, along with other controls and instrumental variables to account for possible endogeneity. To that end further structure will consist of the main body, discussion, and conclusion. The main body will introduce relevant literature as well as the assumptions, methods, and the dataset, with the variables within it in parallel to the research pertinent to them. The part on discussion will wrap up the findings, review possible gaps that might sprout further research, and rationalize the policy relevance of the findings before the final conclusion.

**Impact of Education on Wages**

The “Mincer equation” will form the basic structure of our models (Mincer, 1974, p.83-96). It formalizes the relationship between log earnings, linear years of education, and quadratic experience in the model below:

There were some unknowns with regard to the dataset provided. Among them, was the exact point in time at which the data was collected (‘recent years’ was given) which forces an assumption of data collection at the point of analysis in order to calculate age and experience. This might create a problem with how experience relates to wages. If the data was collected at an earlier point in time, the quadratic effect might ‘stretch’ to a longer interval based on my faulty assumption, and thus provide an inadequate estimate of quadratic experience. This is delved upon further in the section of analysis.

Lemieux (2003, p.4-6, p.8-9) has proposed some additions to the formal model mentioned above. First, he proposes quartic function for experience to better capture the understatement of the initial experience returns in the first 15 years of the career. This does make sense, but using higher order polynomials here would output an error, possibly due to limits in floating-point computations, which would require data manipulation in such a way that would obfuscate interpretations. Here, the data fits the quadratic form of experience, and possibly even the linear. Second, he proposes quadratic term for years of education to account for the convexity, but this convexity was not captured by this dataset. Similar proposals are made by David Card (1999, p.1817-1826), however this paper will rely on Mincer equation as a foundation.

**Dataset, Variables, and Methods** The dataset in question contains fifty thousand observations, and is assumed to be randomly sampled so we assume data points to be independent and identically distributed. Assumed year of data collection is 2024. It contains seven variables out of which we can extract additional information. Additionally, the dataset contains no missing values.

***Wage and logwage***– is the sole dependent variable of interest here. Following the Mincer model, regressand will be in the logarithmic form. The literature does point out possibilities of better fits with different transformations in different cases of observed form on wages; hourly, monthly, etc. However, this dataset provides weekly earnings of an individual and the plain log function is easier to interpret.

***Edu*** – refers to years of education an individual has acquired and the dataset does not provide level of educational attainment in such a way in which we could create distinct categories. Having this would possibly allow to examine significant jumps after completed different levels, or perhaps varying slopes across different levels. Examining these categories would be important especially important if we were to place weight on the assumptions of Signaling Theory; which claims that employers use education to raise wages since it signals a higher level of ability (See Spence, 1974 for more on signaling theory). Nonetheless, we must not discount the effects years of education have in and of themselves. Numerous studies have shown that additional years of study result in higher wages even when no degree is earned and thus no signal is present (see Deming, 2022, p.80 for an overview).  
 Yet before continuing, we have to address the problems that arise around it. First, we have a specification bias, which points to omitted variable in the model related to both the dependent variable and included independent variable, thus forcing the double burden on the included one. With this context we have to assume any rational agent will seek to maximize education given that he is aware of the returns on it. With that we have to assume some measure of individual’s ability to include it to control for this (Deming, 2022, p77; Griliches, 1977, p.5). In this case, we do risk measurement error, when assuming the test validity of IQ as a measure of ability, but psychometric discussions are beyond the scope of this work. Second, endogeneity of education in the presence of omitted ability could be corrected for through the use of instruments such as distance to nearest university, family background, or extended years of education due to peculiarity of state laws. These will be discussed in further paragraphs.

***Edu\_par*** – refers to years of education attained by parents. The literature shows a long-standing tradition of using family background, and in particular parental education, to either directly control for unobserved ability or to instrumentalize education (Griliches, 1979; Siebert, 1989; Ashenfelter & Rouse, 1998, as cited in Card, 1999, p.1822). This is due to high correlation between education and parental education, so much so that twin studies would show up to sixty percent of explained variation in education using observed and unobserved family background factors, while thirty percent of the variation in education in the US adults is explain by education of the parents. However, Card (1999, p.1825) emphasizes that unless all of the unobserved ability is absorbed by the family background controls, there is of it being an illegitimate instrument in a way in which it might reduce bias, but still lead to upward-biased returns to education.

***IQ*** – refers to the results of a recently taken IQ test. Among the potential sources of heterogeneity across the individual’s education we have already reviewed family background, and we are left with school quality, and the aforementioned ability (Card, 1999, p.1852). School quality is an interesting avenue to observe and account for, especially since we know that majority of the educational expansion in the US, in the previous decades, has occurred in the less prestigious institutions (Deming, 2022, p.78). Further, some argue for much higher estimates of returns to education if we were to control for quality of the institution (Carneiro & Lee, 2011, as cited in Deming, p.78). Regardless, with no data on which institutions were attended, this route of investigation falls out of scope. Finally, we are left with an IQ score that acts as a proxy for ability in our case. Altonji and Dunn (1995, 1996a ,1996b as cited in Card, 1999, p.1853-1854) produced a number of papers where they fit log earnings onto education along with controls and their interactions, among which they used IQ which yielded likely imprecise estimates. Additionally, harsher criticisms claim IQ has negligible effects on wages, and thus should not be used as a measure of ability (Deng, 2010; not formally published). This data-set has shown found some relevance in the use of IQ and it will be investigated further.

***Dist*** – refers to distance to the nearest university city. Data description provided with the dataset phrases this as “distance to the next city/town that offers higher education”. This might mean that for some individuals which reside in the university city, the data would point to a distance to the nearest next/other university city. This would greatly confound the data, so it is assumed here that the data shows distance to the first/nearest university city. Using distance to university as an instrument to education is logical since proximity will certainly affect decision making directly by providing opportunity, or indirectly by observing others taking up those opportunities, but we should keep in mind the risk with this data; potentially residence at the time of data collection was used to calculate distance, and is thus not the same as the residence at the time the participants were making the decision to go to university (Card, 1993, p.16). It is assumed here that the distance refers to the one at the time of making the decision (18-20 years old). Additionally, dummies could be constructed in to create categorically segment the sample (e.g. very near, near, far, very far), there was no need for it in this case and the variable will be treated solely as continuous.

***State*** – refers to the state of residence. It ranges from 1 to 56, and the mapping to which states it refers to was not provided. Some literature did show the use segments as dummy such as states of the south or similar, but here, since it is not possible to map exact states, we can factorize the state variable and use it as a set of dummies to observe how the intercept might change; i.e. we can see the wage difference across different states and control for it.

***Born*** – provides us with the date of birth (DOB). This can be further used to extract age and experience, under the assumption data was collected in 2024. A notable use of the exact dates was in the work of Angrist and Krueger (1991, p.979) where they noticed a peculiarity with state mandated schooling where individuals born in the first quarter of the year would start at an older age and can therefore drop out after less completed years overall. This can be observed by using a dummy for all observations born in the first quarter. This is not as relevant in this sample since these effects diminish with further education. The next two variables are constructed on the basis of this one.

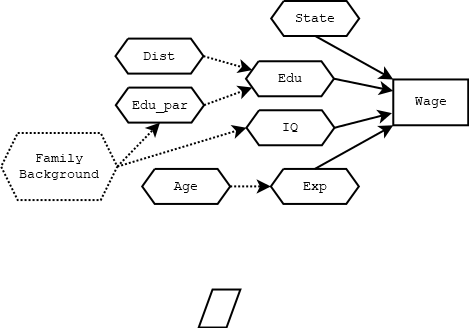
***Age*** – is calculated by taking the day at which the analysis is done and subtracting the given DOB, and it should be kept in mind that age is floored in these calculations.

***Exp*** ***and Exp2*** – refers to the experience of the individual and is calculated by subtracting years of education from the age, and then subtracting addition six years, to satisfy the assumption that primary education starts at the age of seven. Mincers model suggests using the quadratic form which will be ‘Exp2’, but when modelling linear form will be considered too. Age will be used as a natural instrument to experience.

The methods used here will rely on linear regressions with the use of instrumental variables. Although, there could be a place and time for regression discontinuity (RDD) if there was a number of distinct categories made out of a running variable (e.g. with distance), and those categories had visible shifts, or perhaps it might be possible to find a more adequate fit using varied slopes RDD. That is not considered here.

There are more general limitations to be considered before moving on. We should not lose sight of the possibility that returns on education might fluctuate over time. Furthermore, same could be said for all other variables mentioned and possible interactions among them. This paper assumes constant effects across time. Additionally, fine grained geographical observation (along with their ethnic and class composition) could be better at revealing more appropriate shapes of the relationships. This cannot be done here. Bearing this in mind, the goal here is to find a simple model that would correspond to the previous findings that estimate returns on education for each additional year in the range of 6-18%, with the median in the range 10-12% (Deming, 2022, p. 77). Before proceeding with the analysis and modeling the graphical representation of the relationships mentioned is laid out below in Figure 1.

**Figure 1**  
*Diagram of Assumed Underlying Relations*



*Note*. The diagram shows how concept of family background can be accounted for through the use IQ as a proxy for ability. In addition, it shows how distance and parental education satisfy relevance criteria, but also exclusion since their effects are only through education and they do not affect ability.

**Analysis, Results, and Discussion**

With the diagram above to guide the reasoning, the following paragraphs will present the models that result from these assumptions. It should be noted that full code can be reviewed in the Appendix (R Quarto document). First of all, we observe the correlational matrix on the next page.

**Table 1***Correlation Matrix of used Variables*

| X | wage | logwage | edu | edu\_par | age | age2 | exp | exp2 | iq | dist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| wage | 1.00 | 0.79 | 0.56 | 0.05 | 0.01 | 0.01 | -0.43 | -0.40 | 0.15 | 0.03 |
| logwage | 0.79 | 1.00 | 0.68 | 0.06 | -0.00 | -0.00 | -0.53 | -0.52 | 0.20 | 0.04 |
| edu | 0.56 | 0.68 | 1.00 | 0.12 | -0.04 | -0.04 | -0.80 | -0.79 | 0.06 | 0.08 |
| edu\_par | 0.05 | 0.06 | 0.12 | 1.00 | 0.00 | 0.00 | -0.09 | -0.09 | -0.00 | 0.00 |
| age | 0.01 | -0.00 | -0.04 | 0.00 | 1.00 | 1.00 | 0.63 | 0.64 | 0.00 | -0.00 |
| age2 | 0.01 | -0.00 | -0.04 | 0.00 | 1.00 | 1.00 | 0.63 | 0.64 | 0.00 | -0.00 |
| exp | -0.43 | -0.53 | -0.80 | -0.09 | 0.63 | 0.63 | 1.00 | 0.99 | -0.04 | -0.07 |
| exp2 | -0.40 | -0.52 | -0.79 | -0.09 | 0.64 | 0.64 | 0.99 | 1.00 | -0.04 | -0.07 |
| iq | 0.15 | 0.20 | 0.06 | -0.00 | 0.00 | 0.00 | -0.04 | -0.04 | 1.00 | -0.00 |
| dist | 0.03 | 0.04 | 0.08 | 0.00 | -0.00 | -0.00 | -0.07 | -0.07 | -0.00 | 1.00 |

*Note*: state and born variables are left out. Continuity was assumed with the use of Pearson.

It is important to note any multicollinearity, along with checks for possible exclusion and relevance criteria, which will both be checked for in the first stage of 2SLS. The table shows correlation of ‘edu’ with both the ‘wage/logwage’, while it is low for ‘edu\_par’ and ‘dist’. The instruments do not correlate with ‘iq’ and thus show promise as instruments. Same can be observed for ‘exp’ and ‘age’ and their correlation with wage/logwage. ‘Exp’ and ‘edu’ show possibly concerning negative relationship for multicollinearity, but it is only logical to expect someone who spent longer in education has less experience.

There are four models to review, each have shown significant F-test on first stage 2SLS demonstrating relevance of the instruments, each have resulted in significant Wu Hausman tests showing stronger consistency of estimators over their OLS counterparts, and each have non-significant Sargan test for overidentification (two instruments used for educ) and thus we can assume exogeneity for all of our instruments. The models are presented in Table 2 on the next page.

**Table 2***Table of Four Final Models*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  | *Dependent variable: log(wage)* | | | |
|  |  | | | |
|  | Quadratic Experience | | Linear Experience | |
|  | (No States) | (With States) | (No States) | (With States) |
|  | | | | |
| poly(exp, 2, raw = TRUE)1 | 0.017 | 0.020 |  |  |
|  | (0.016) | (0.016) |  |  |
|  |  |  |  |  |
| poly(exp, 2, raw = TRUE)2 | -0.0002 | -0.0003 |  |  |
|  | (0.0003) | (0.0003) |  |  |
|  |  |  |  |  |
| exp |  |  | 0.006\*\*\* | 0.006\*\*\* |
|  |  |  | (0.001) | (0.001) |
|  |  |  |  |  |
| edu | 0.110\*\*\* | 0.110\*\*\* | 0.110\*\*\* | 0.109\*\*\* |
|  | (0.005) | (0.005) | (0.005) | (0.005) |
|  |  |  |  |  |
| iq | 0.010\*\*\* | 0.010\*\*\* | 0.010\*\*\* | 0.010\*\*\* |
|  | (0.0002) | (0.0002) | (0.0002) | (0.0002) |
|  |  |  |  |  |
|  |  |  |  |  |
| Constant | 3.429\*\*\* | 3.271\*\*\* | 3.570\*\*\* | 3.459\*\*\* |
|  | (0.217) | (0.216) | (0.079) | (0.078) |
|  |  |  |  |  |
|  | | | | |
| Observations | 50,000 | 50,000 | 50,000 | 50,000 |
| R2 | 0.447 | 0.459 | 0.447 | 0.459 |
| Adjusted R2 | 0.447 | 0.458 | 0.447 | 0.458 |
| Residual Std. Error | 0.632 (df = 49995) | 0.626 (df = 49945) | 0.632 (df = 49996) | 0.626 (df = 49946) |
|  | | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | |

*Note:* for ‘with states’ models, variable states are used as a factor to create dummies. The values are not present here (see Appendix). All models use instruments ‘edu\_par’, ‘age’, and ‘dist’.

For clarification purposes model with 56 states dummies and with quadratic experience corresponds to the equation noted below:

From the table we can observe that models using quadratic form of experience are insignificant, even though it should be mentioned that they result in a plausible estimate of a quadratic function opening downwards with a vertex at around 33, implying rising returns on experience up to 33 years of experience. Linear estimation of experience suggests 0.6% increase of wage for each year. Across all models we see an increase in 1% of weekly wages for every additional IQ point, while we also see that all four models converge around 11% increase in weekly salary for each additional year of education. Which matches aforementioned range of estimations from previous literature; ‘6-18%, with the median in the range 10-12%’. Finally, dummies for each state show intercept changes, which allow for inter-state comparison of weekly wages. However, values and their significance levels can be found in the Appendix.

Additionally, each model has normally distributed residuals (plots in Appendix), but the plotted fitted values against residuals show slight sphericity. This could mean that heteroskedasticity is present. However, in my case, robust estimation took too long on my computer so I had to accept this as a limitation of my analysis.

Further limitations of this study include unmapped states which could allow for better geographical segmentation; lack of ethnic composition of participants, class composition along with familial incomes; quality of education in the form of tuition costs, teacher/student ratios, or other; year when the institutions were built. There is also room for temporal investigation of the changes to returns on education and experience, as well as for those changes across different industries. Methodologically, avenues for better ability measurements could be made, but also for more adequate modelling which would better capture the changes across lifetime. In addition, this study did not find the use for RDD, and it could be used if some theoretically sound shifts can be observed, perhaps from high school to bachelor’s graduate, or to model returns with a different curve for PhDs. All of these limitations can act as novel paths for further research. This study is of critical importance to policy makers that focus on multiplicative effects education with technology and its contribution to a healthy growth of an economy.

## Conclusion

In summary, this work attempts to provide a coherent estimate of returns to education placing its foundations upon Mincers Equation and uses theoretically sound instruments to face well known issue of endogeneity present with the use of years of education. It accounts for ability bias through the use of recent IQ tests, and instrumentalizes distance to nearest university and parental education, in order to find a more robust estimate, and controls for state differences in salary and years of experience. It neglects the possibility of present heterogeneity with omitted robust estimation, but it does come up with four sound models that converge upon the estimate of 11% returns on weekly wages for each additional year of education.

## References

Altonji, J. G., & Dunn, T. A. (1996a). The effects of family characteristics on the return to education. *The Review of Economics and Statistics*, *78*(4), 692. https://doi.org/10.2307/2109956

Altonji, J. G., & Dunn, T. A. (1996b). Using siblings to estimate the effect of school quality on wages. *The Review of Economics and Statistics*, *78*(4), 665. https://doi.org/10.2307/2109953

Altonji, J., & Dunn, T. (1995). *The Effects of School and Family Characteristics on the Return to Education*. https://doi.org/10.3386/w5072

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Ashenfelter, O., & Rouse, C. (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *The Quarterly Journal of Economics*, *113*(1), 253–284. https://doi.org/10.1162/003355398555577

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of Political Economy*, *70*(5, Part 2), 9–49. https://doi.org/10.1086/258724

Card, D. (1993). *Using Geographic Variation in College Proximity to Estimate the Return to Schooling*. https://doi.org/10.3386/w4483

Card, D. (1999). The causal effect of education on earnings. *Handbook of Labor Economics*, 1801–1863. https://doi.org/10.1016/s1573-4463(99)03011-4

Carneiro, P., & Lee, S. (2011). Trends in quality-adjusted skill premia in the United States, 1960–2000. *American Economic Review*, *101*(6), 2309–2349. https://doi.org/10.1257/aer.101.6.2309

Deming, D. J. (2022). Four facts about human capital. *Journal of Economic Perspectives*, *36*(3), 75–102. https://doi.org/10.1257/jep.36.3.75

Deng, Binbin, (2010). "Schooling and Wage Revisited: Does Higher IQ Really Give You Higher Income?" *MPRA Paper* 23206, University Library of Munich, Germany.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica*, *45*(1), 1. https://doi.org/10.2307/1913285

Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *The Bell Journal of Economics*, *10*(1), 92. https://doi.org/10.2307/3003321

Lee, J.-W., & Lee, H. (2016). Human capital in the Long Run. *Journal of Development Economics*, *122*, 147–169. https://doi.org/10.1016/j.jdeveco.2016.05.006

Lemieux, T. (n.d.). The “mincer equation” thirty years after schooling, experience, and earnings. *Jacob Mincer A Pioneer of Modern Labor Economics*, 127–145. https://doi.org/10.1007/0-387-29175-x\_11

Mincer, J. (1974). *Schooling, experience, and earnings*. National Bureau of Economic Research.

Roser, Max, and Esteban Ortiz-Ospina (2016). “Financing Education.” Our World in Data. https://ourworldindata.org/financing-education (accessed May 25, 2024)

Spence, M. (1974). Competitive and optimal responses to signals: An analysis of efficiency and distribution. *Journal of Economic Theory*, *7*(3), 296–332. https://doi.org/10.1016/0022-0531(74)90098-2

Siebert, W. S. (1985). Developments in the economics of human capital. Carline, Derek: *Labour Economics*, Longman, London and New York, 5-77.

**Software:**

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Posit team (2024). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. http://www.posit.co/.

Packages: Tidyverse, flextable, ggthemes, lubridate, psych, stargazer, ivreg, modelsummary, car

Dia Diagram