# Hoping for more Wages? There is no causal relationship between one's mother's education and wages.

Dina Abu Ghosh, Anusha Champaneria,Ahmed Lugya

05/04/2020

## Abstract

We use Women's Wages dataset (R, 2018) to analyze the relationship between Wages and the years of schooling for females in the U.S. We analyze other affecting factors such as the mother's years of schooling and use Instrumental Variable estimations to observe its correlation with the years of schooling for females. The purpose of this analysis is to determine whether it is necessary to go back to school to increase hourly wage and whether the mother's level of schooling depicts the child's ability to further educate. We identify that there is no causal relationship between the mother's years of schooling and wages other than through the participants' years of schooling.

## Introduction

It is typically said that when one goes to school, they are guaranteed higher wages and better job circumstances (Williams, 2019). However, thousands of dollars are invested in university or college education yet many students struggle to find jobs due to lack of enough job opportunities, lack of experience, or the field of study fading. Faridi (2010) states that in terms of earning, it is assumed that returns to education are uniform across different levels of education. Different school years impart different skills to the workers and bring different returns. Women to this day suffer from gender inequality when it comes to the workforce where they still get paid less than men for certain jobs, therefore they tend to resort to more schooling (Faridi, 2010).

The level of education in one's family history can also potentially play a role in their child's level of education. According to Sutherland (2015), "children's educational outcomes are closely linked to their parents' level of education", higher educated mothers can engage their children in cognitively stimulating activities during developmental stages. Additionally, they are more equipped to help out with studying and homework. A reverse effect can also be played where a child might do just the opposite by learning from their mothers' experience. If a child grows up seeing their mother suffering financially due to lack of education, the child may be encouraged to attain higher education to avoid such situations they have seen. However, intelligence is not always determined by school grades. Many people are getting paid higher with fewer years of education compared to those with more years of education due to skills learned from other experiences.

In this paper, we argue that there's no way mother's education and wages could be causal except through the causal relationship of wage -> education. We construct models to determine whether the wage is explained by the years of schooling. We find that there is no causal relationship between mother's education and wages other than through the participants' years of schooling. We also discuss Instrumental Variable estimation in terms of two-stage least squares (regressions), its limits, testing procedures and we conclude with the ethical considerations and shortcomings.

**Research question:**

Do years of schooling, mother's education influence hourly wages for females in the U.S?

## Data Set

The Women's Wages dataset is based on observations for 753 women on wages, years of schooling, hours worked, and work circumstances collected from the labour force in the U.S (R,2018). Out of the total of 22 attributes, we use Wage as our dependent variable, Education (educ) as our explanatory variable and Mothers Education (motheduc) as our Instrument variable. Attribute Wage is the estimated wage earned hourly. Attribute Education is the participants' total number of years of schooling. Attribute mother's education is the mother's total years of schooling.

## Descriptive Analysis

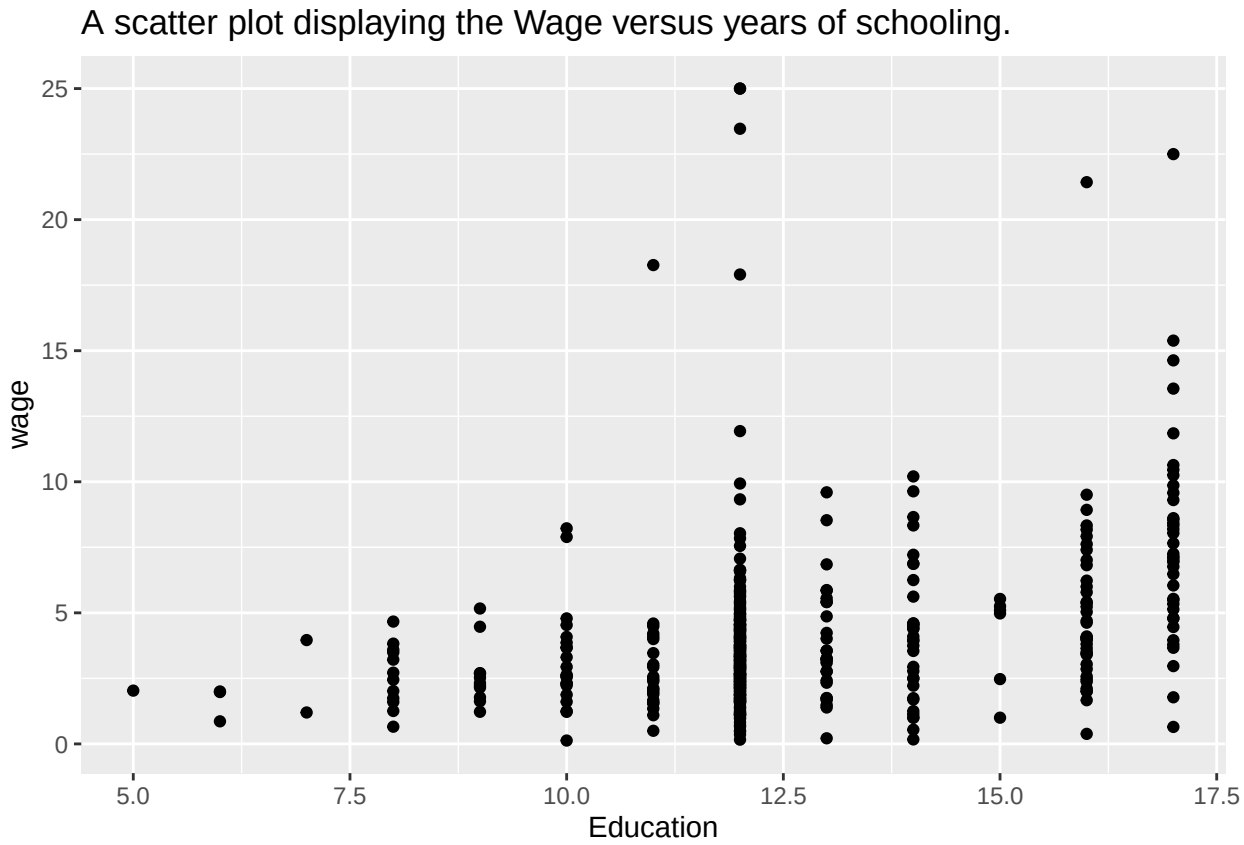


Figure 1.

Looking at Figure 1 above, you can see that as the level of education increases, the wage increases as well. You can also observe that participants with 12 years,17 years of schooling earn more wages compared to the other years.

## Data analysis: Instrumental Variable regressions

In most observational studies, the independent variables tend to correlate with the error term in the regression model, and this renders the Ordinary Least Squares estimates an inappropriate method to use. Independent variables that correlate with the error term are termed as endogenous variables, whereas independent variables that do not correlate with the error term are called exogenous. The solution to this problem associated

with the linear regression assumption is the use of instrumental variables, which are variables that do not directly influence the dependent variable but are correlated with the endogenous independent variable in question.

In the next section, we are trying to test whether years of schooling influence wages. We want to know if wage → education is an accurate causal pathway. We're going to establish causality using an instrument which is a third variable that has an association with our explanatory variable (education), but this instrument variable should have no association with our dependent variable(Wage) except through the means of our explanatory variable. In this case, our instrument is going to be a mother's education. Using a two-stage least squares process. The first stage of our process is going to be a regression between our instrument the mother's education and education and that regression has education on the left-hand side of the equation and mother's education on the right-hand side of the equation. We come up with B0 and B1, B1 being the coefficient on Mother's education or the Degree of correlation between these two things.

Stage 1: Education = B0 +B1mother's education + e

Using B0, B1 and data from our instrument (mother's education), we come up with education_hat which is different from education. Education is actual data about the participants' years of schooling. Education_hat is a prediction of the probability of the participants' years of schooling based only on their mother's education and the coefficients B0, B1.

And then the second stage of our two-stage least squares has wages our dependent variable on the left-hand side of the equation and the right-hand side of the equation has education_hat. The second regression is going to capture the causal mechanism of whether there is a correlation between the instrument(mother's education) and wages other than through the participants' years of schooling.

Stage 2: wages = B2 + B3education_hat + e

So if B3 is significant, then that shows us that there is a causal relationship between wage and education.

First, let's look at the correlation between wage and education using Ordinary least squares estimations.

Table 1: OLS output

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -2.0924 | 0.8483 | -2.4666 | 0.014 |
| educ | 0.4953 | 0.0659 | 7.5106 | 0.000 |

From Table 1 above, we can say that there is an increase of 50% in the wage for every additional year of education.

# Using Two-Stage Least Squares (2SLS)

# First stage

Education = B0 +B1mother's education + e

Table 2: First stage output

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 10.1144938 | 0.31090432 | 32.532497 | 0e+00 |
| motheduc | 0.2673697 | 0.03086304 | 8.663102 | 1e-16 |

From Table 2 above, we can say that there is an increase of 27% in the participants' years of schooling for every additional year of the mother's education.

## Second stage

wages = B2 + B3education_hat + e

Table 3: Second stage output

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 1.4737 | 2.2973 | 0.6415 | 0.5216 |
| educ_hat | 0.2136 | 0.1810 | 1.1799 | 0.2387 |

Table 3 above shows a p-value of 0.239, It is clearly evident that B3 is not significant, hence we can conclude that there is no causal relationship between mother's education and wages other than through the participants' years of schooling.

# IV Validation

Two-stages least-square relies on the chosen instrument variables being strongly correlated to the endogenous variable. We shall use the results from the first stage regression above to test the relevance of the instrument as well as test for exogeneity. We implement these tests using Standard F-test.

## a) Relevance of the Instrument

Lets begin with stating our null hypothesis, that is;

Ho: instruments are irrelevant.

Table 4: Standard_F_test_output

| res.df | df | statistic | p.value |
|---|---|---|---|
| 426 | ** | ** | ** |
| 427 | -1 | 75.04933 | 1e-16 |

Table 4 above shows that the value of the F_test is 75.05 with an extremely low p-value. Hence, we can clearly reject the null hypothesis that the instruments are irrelevant.

## b) Exogeneity Testing

In the next section we shall test for exogeneity, which is a standard assumption made when using regression analysis. Exogeneity informs us if the explanatory variables are not dependent on the dependent variable.

Step 1: We run the first stage regression of the Two Stage Least Squares.

Step 2: We then add the residuals from Step 1 into the original model. And our results are in Table 5 below;

Hausman_reg(wage = Y1 + Y2educ + first_stage$residuals)

Table 5: Hausman regression results

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 1.4737 | 2.1567 | 0.6833 | 0.4948 |
| educ | 0.2136 | 0.1700 | 1.2568 | 0.2095 |
| first_stage$residuals | 0.3313 | 0.1843 | 1.7976 | 0.0729 |

Comparing results in Table 5 above with results in Table 3, the Hausman test barely rejects the null that the variable of concern is uncorrelated with the error term, indicating that educ is marginally endogenous.

In the next step we use the Standard F-test to test the null hypothesis:

H0: The residuals are irrelevant.

Table 6: Standard_F_test_output

| res.df | df | statistic | p.value |
|---|---|---|---|
| 425 | ** | ** | ** |
| 426 | -1 | 3.2314 | 0.0729 |

From Table 6 above, we can observe a p-value of 0.0729. With an alpha of 0.05, we fail to reject the null hypothesis of residuals being exogenous.

## Shortcomings

Whereas its an interesting process to estimate causal effects, due to several reasons the explanatory variables usually turn out to be endogenous. While Instrumental Variable estimation is looked at as the general solution to this challenge because of its ability to cut correlations between the error term and explanatory variables, it has a few problems which might include:

a) The precision of Instrument Variables estimates is low compared to that of Ordinary Least Squares estimates. This can be observed in the second stage output Table 3 above where the standard error is 0.1810 which is large compared to the OLS output Table 1 where education has a standard error of 0.0659.

b) Good properties in Instrument variables estimators can only be seen when the instruments are both strong and relevant.

c) Instrument Variable estimates often have problematic finite sample properties and their estimators are usually biased.

d) Poor performance when the samples we are using are small as well as the difficulty that comes with finding good Instrument Variables to use.

And lastly, due to the scope of this assignment we had to pick three variables only which can make our analysis less conclusive, there could also be more influential variables that were not chosen that could have affected the correlation that we found between the chosen variables.

## Ethical issues

The research question explores the correlation between a female's education as a mother and hourly wage in the US however there are lots of parents who don't identify as either female or male which can make this research question and the variables chosen to be outdated. Lots of parents are gender fluid, so how would a parent's gender fluidity affect a child's education level in this case? Or how would it affect the parent's wages? Gender inclusiveness is very important to consider when looking into studies based on gender and sex. Even though sex and gender are different they should be clearly identified and acknowledged in a study as they can correlate impact the results of the study. This might be interpreted as a bias against transgender people which can affect the accuracy of the analysis. The study did not declare any limitations or biases/conflict of interest. There is no mention of obtaining the information consensually, a consent declaration should be mentioned as well. The intent of the study is listed which is an important part of research ethics (Thomas et al., 2017).

# APPENDIX

## Source Code

**This script has source code for all the analysis in Problem set 5.**

```
#### Contact details ####
# Title: Problem set 5
# Purpose: This script has source code for all the analysis in Problem set 5.
# Author: Dina, Anusha, Ahmed
# Last updated: 4 April 2020
# License: MIT License.

#### Setting up workspace ####
library(magrittr)
library(ggplot2)
library(ivpack)
library(lmtest)
library(estimatr)
library(knitr)
library(skimr)
library(broom)

#read the data
mydata <- read.csv("Mroz.csv", na.strings = ".")

#visualize the data
nrow(mydata)
ncol(mydata)

# remove observations with missing wages from dataset
mydata <- subset(mydata, is.na(wage) == FALSE)

#Visualize the data
myscatter <- ggplot(mydata, aes(x = educ, y = wage)) +
  geom_point() + xlab('education')
myscatter

##Data analysis: IV regressions

### First, out of curiosity, look at OLS:
ols <- lm(wage ~ educ,data=mydata)

# ivr_first <- lm(educ~motheduc,data=mydata)
# kable(tidy(ivr_first), digits=4, align='c',caption=
#         "OLS output")
fun1<-function(values){
  res<-c(paste(as.character(summary(values)$call),collapse=" "),
         values$coefficients[1],
         values$coefficients[2],
         length(values$model),
         summary(values)$coefficients[2,2],
```

```
         summary(values)$r.squared,
         summary(values)$adj.r.squared,
         summary(values)$fstatistic,
         pf(summary(values)$fstatistic[1],
            summary(values)$fstatistic[2],
            summary(values)$fstatistic[3],
            lower.tail=FALSE))
  names(res)<-c("call","intercept","slope",
                "n","slope.SE","r.squared",
                "Adj. r.squared",
                "F-statistic","numdf","dendf","p.value")
  return(res)}
res1<-fun1(ols)
write.csv(res1,"ols_summary.csv")

###2SLS
##First stage estimation
ivr_first <- lm(educ~motheduc,data=mydata)

fun2<-function(values){
  res<-c(paste(as.character(summary(values)$call),collapse=" "),
         values$coefficients[1],
         values$coefficients[2],
         length(values$model),
         summary(values)$coefficients[2,2],
         summary(values)$r.squared,
         summary(values)$adj.r.squared,
         summary(values)$fstatistic,
         pf(summary(values)$fstatistic[1],
            summary(values)$fstatistic[2],
            summary(values)$fstatistic[3],
            lower.tail=FALSE))
  names(res)<-c("call","intercept","slope",
                "n","slope.SE","r.squared",
                "Adj. r.squared",
                "F-statistic","numdf","dendf"
                ,"p.value")
  return(res)}
res2<-fun2(ivr_first)
write.csv(res2,"firststage_summary.csv")

educ_hat <- ivr_first$fitted.values

##Second stage ivreg estimation
ivr_second <- lm(wage~educ_hat, data = mydata)

fun3<-function(values){
res<-c(paste(as.character(summary(values)$call),collapse=" "),
         values$coefficients[1],
         values$coefficients[2],
         length(values$model),
         summary(values)$coefficients[2,2],
         summary(values)$r.squared,
```

```
          summary(values)$adj.r.squared,
          summary(values)$fstatistic,
          pf(summary(values)$fstatistic[1],
             summary(values)$fstatistic[2],
             summary(values)$fstatistic[3],lower.tail=FALSE))
  names(res)<-c("call","intercept","slope","n","slope.SE",
                "r.squared","Adj. r.squared",
                "F-statistic","numdf","dendf","p.value")
  return(res)}
res3<-fun3(ivr_second)
write.csv(res3,"secondstage_summary.csv")


##Validation
# a) Relevance of the instrument

# First Stage
first_stage <- lm(educ~motheduc,data=mydata)

instrFtest <- waldtest(first_stage,.~.-motheduc)
print(instrFtest)
# Ftest <- hux(
#   Standard_F_test_output = c("F_test","Pr(>F)"),
#   Values = c(75.049,2.2e-16),
#   add_colnames = TRUE
# )
# right_padding(Ftest) <- 10
# left_padding(Ftest)  <- 10
# bold(Ftest)[1, ]          <- TRUE
# bottom_border(Ftest)[1, ] <- 1
# Ftest %>% set_background_color(everywhere, starts_with("S"), "lightblue")

# b) Exogeneity Testing

Hausman_reg <- lm(wage~educ+first_stage$residuals,data=mydata)
print(summary(Hausman_reg))

HausWutest <- waldtest(Hausman_reg,.~.-first_stage$residuals)
print(HausWutest)

#Shortcomings
```

# References


- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Yang Jiang and Dylan Small (2014). ivpack: Instrumental Variable Estimation.. R package version 1.2. https://CRAN.R-project.org/package=ivpack
- Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/
- David Hugh-Jones (2020). huxtable: Easily Create and Style Tables for LaTeX, HTML and Other Formats. R package version 4.7.1. https://CRAN.R-project.org/package=huxtable
- Graeme Blair, Jasper Cooper, Alexander Coppock, Macartan Humphreys and Luke Sonnet (2020). estimatr: Fast Estimators for Design-Based Inference. R package version 0.22.0. https://CRAN.R-project.org/package=estimatr
- Yihui Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.26.
- Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963
- Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595
- Stefan Milton Bache and Hadley Wickham (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. https://CRAN.R-project.org/package=magrittr
- R (May 3, 2018). Retrieved from http://eclr.humanities.manchester.ac.uk/index.php/R#Data_Sets
- Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2019). skimr: Compact and Flexible Summaries of Data. R package version 2.0.2. https://CRAN.R-project.org/package=skimr
- Thomas, D., Pastrana, S., Hutchings, A., Clayton, R. and Beresford, A. (2017). Ethical issues in research using datasets of illicit origin. Cambridge Cybercrime Centre, Computer Laboratory, p.2.
- David Robinson and Alex Hayes (2019). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.3. https://CRAN.R-project.org/package=broom
- Faridi, M. (2010). The Effects of Health and Education on Female Earning. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.700.6640&rep=rep1&type=pdf
- Sutherland, A. (2015). The Many Ways Mother's Education Matters. Retrieved from https://ifstudies.org/blog/the-many-ways-mothers-education-matters