2018

Intro to Econometrics

LITERATURE REVIEW AND EMPIRICAL ANALYSIS

HAYES, MASON R.

TABLE OF CONTENTS

PART I

AJR (2001)

The main assumptions behind the AJR (2001) paper are that the expected mortality rates of European settlers in each colony are indicative of how the colonists would create institutions in each colony. Importantly, we assume that the institutions remained after the colonies became independent. Better institutions in a colony are theorized to be predictive of a higher level of economic development in the present.

The researchers *'regress current performance* (GDP per capita, PPP, 1995) *on current institutions* ("risk of expropriation" index from Political Risk Services) *and instrument the latter by settler mortality rates.'* In other words, settler mortality rates are theorized to have a high correlation with current institutions. This is confirmed by the fact that SMR explains more than 25% of variation in current institutions.

In their discussion of early settler mortality rates, the authors discuss various problems that could occur with using this as an instrument for institutions. To discredit some of the possible problems and to strengthen their conclusions, they demonstrate that settler mortality rate is unlikely to be a proxy for geographical or climactic features within a country.

The main econometric model of the paper proposes that economic level (GDP per capita, PPP, 1995) depends on quality of current institutions (instrumented by settler mortality rates). In this model, log GDP per capita is the outcome variable and log settler mortality rate is the explanatory variable. Various difficulties arise when trying to estimate this model with the standard OLS method, including collecting data, estimating the validity of proxy variables, correcting for possible endogenous variables, and establishing the plausibility of the proposed assumptions. It is difficult to test run an OLS regression with current institutions vs GDP per capita because GDP per capita and current institution may suffer from simultaneity which would bias and ultimately discredit any simple OLS regression.

The data collection and construction are the primary problems with this type of historical research, especially this paper; as their data relies entirely on sources that are sometimes incomplete, they 'construct' data to fill in the blanks. These problems are discussed at length in Albouy (2012).

Settler mortality rate is used as an instrument for current institutions because current institutions is an endogenous variable (it is influenced by many factors that are difficult to control for) and settler mortality rate is highly correlated with current institutions, but not correlated with other possible explanatory factors such as disease or geographical factors. For example, it is not known whether rich countries choose better institutions, or whether better institutions lead to richer countries. By using settle mortality rates and the proposed channels through which it works (settler mortality rates determine settlements, settlements determine early institutions, strong correlation between early institutions and current institutions), the problem of endogeneity is resolved.

The exclusion restriction (settler mortality rate is uncorrelated with the error term, i.e. past settler mortality rates have no effect on GDP per capita except through their effect on institutions) is argued by the researcher to be valid for the primary reason that 'diseases are unlikely to be the reason why many countries in Africa and Asia are poor today' (AER 2001, p 1371) because indigenous people suffered less

from the diseases, and this is the most likely way SMR could be correlated with the error term. They also show that neither the fact that a country is African or its distance from the equator is significant after institutions are controlled for. This shows that Africa does not suffer from an inherent disadvantage beyond the effect of institutions (through the channel of settler mortality, settlements, past institutions, current institutions).

The exclusion restriction could also be invalidated if there are significant omitted variables; the researchers checked for the most plausible omitted variables (identity of the main colonizer, legal origin, climate, religion, geography, natural resources, soil quality, and measures of ethnolinguistic fragmentation) and found that their estimates changed very little. Albouy (2012) also challenges this analysis in his comments on the research. They also controlled for current 'disease environment . . . and the current fraction of the population of European descent' – and the results were robust to these controls.

In summary, the researchers showed that they had a relevant IV, persuasively explained why they believed their instrumental variable to be exogenous in theory, and then tested the inclusion of possible significant omitted variables which ultimately led to an insignificant change in the results of their initial regression; they then used overidentification tests to confirm the exogeneity of the instrument. The theoretical argument as well as the numerous econometric tests build a very compelling case for the validity of the exclusion restriction which, in addition to the relevance restriction that was earlier demonstrated, shows that settler mortality rate is an acceptable instrumental variable and can be useful in analyzing differences among countries' current GDP per capita.

ALBOUY (2012) RESPONSE TO AJR (2001)

David Y. Albouy argues in his 2012 response to the AJR (2001) paper that the methods used by AJR to form their data set, as well as the data itself, are flawed. His primary argument is that the sources of the mortality rate data and the extrapolation techniques ('conjectured data') used by AJR lead to a weak empirical relationship between expropriation risk and settler mortality rates. If these claims are valid, settler mortality rate would lose its robustness as an instrumental variable and could invalidate any econometric research that used the conjectured data. Albouy claims that the relationship between expropriation risk and settler mortality rates is 'an artifact of the data's construction' (Albouy 2012, p. 3060).

One of the main problems with the conjectured data is that AJR assigns settler mortality rates to neighboring countries, which turns out to be 'not just unreliable, but often deeply flawed, generating rates that may be far too high or too low' (p. 3064). Another important flaw is that AJR (2001) systematically use mortality rate data for campaigning soldiers (instead of soldiers in barracks, who would have lower disease and mortality rates) for countries with high risk of capital expropriation and low GDP per capita, which biases the results of the regression and artificially supports the AJR hypothesis by reinforcing patterns in the data. This is a form of selection bias and has the potential to harm not only the results of the AJR (2001) analysis but also the reputations of the authors.

AJR (2012) Reply to Albouy (2012)

The primary argument in AJR (2012) is that Albouy (2012) discarded relevant data from Africa and Latin America and that, by doing so, his results are skewed from the outset to confirm his hypothesis. By including the data, AJR's data construction and its analysis becomes less problematic than Albouy's paper suggests – as the authors claim, "Albouy needs to discard almost 60 percent of our original sample in order to undermine our results" (p. 3078).

AJR argue that Albouy's claim that there are significant differences in campaign data and peacetime data is exaggerated since "most of these campaigns did not involve much actual fighting" (p. 3079). They further argue that it is too difficult, sometimes impossible, and usually "makes little sense" to classify mortality rates as from a campaign or from peacetime (p. 3078).

Bonus section

It is difficult to form an argument against or in favor of either paper without devoting a large amount of time to thoroughly analyzing all the given data, the data sources, the statistical analyses and the methods that the authors use.

The fact that Albouy discards the data from 36 countries from the outset, with little convincing explanation for the reason, makes me doubt all conclusions that he draws. AJR showed the robustness of their data in AJR (2000) and they point out how Albouy largely ignored this. Albouy's other argument, that AJR selectively used campaign data and barracks data to reinforce the data trend, is not compelling; as AJR pointed out, mortality rate over a certain number should theoretically have no additional effect on settlements (AJR 2012, p. 3080). They showed that this is indeed the case, as mortality rate can be capped at 250 per 1000 and the results still stand.

Based on these few reasons, I believe the data and results in AJR (2001) to stand up to the criticism of Albouy (2012), and AJR (2012) needed to do little more than restate the robustness checks that they did years before Albouy published his criticism of their data.

Part II

Section A: Proposed Models

The goal of this part of the project is to examine various econometric models for two primary questions:

A) What are the factors that affect a woman's decision to enter the labor force? Can a model be shown that explains women's labor force participation rate, or to what extent they participate in the labor force (number of hours worked)?
B) If a woman enters the labor force, what are her returns to schooling (how does a one-year increase in education affect the log of her wage)?

For question A, there are various factors that could explain a woman's decision to enter the labor force. Firstly, I will differentiate between presence in the labor force and level of participation in the labor force; for the first models I consider, labor force participation will be a binary dependent variable –

either a woman is in the labor force, or she is not. In later models I will explore the returns to education and other relevant variables for women in the labor force.

Possible factors to consider in a decision to join the labor force:

1. Level of education
2. Family background
3. Age
4. Number of children
5. Other sources of income (parents, siblings, spouse, etc.)
6. Availability of work (unemployment level)
7. Whether her female friends/neighbors are in the labor force or not
8. Ability to work

For some of these possible factors, we have no data. There are also many possible factors for which we do not or cannot have data.

Using the provided data set, there are a few important models that can explain labor force participation and the wage effects of education. The first model I will explore is:

Model 1: Joining the Labor Force

Equation A: $$P(inlf = 1) = G(\beta_0 + \beta_1 * exper + \beta_2 * expersq + \beta_3 * age + \beta_4 * educ + \beta_6 * kidslt6 + \beta_6 * nwifeinc$$

I chose this model because:

- Participation in the labor force is binary; one is either in the labor force or not in the labor force.
- One would expect past experience in the labor force to be highly correlated with current participation in the labor force – experience squared is included because there might be diminishing returns to this relationship.
- Women may leave the labor force as they get older.
- Education should be highly correlated with labor force participation; however, education might be endogenous (For example, if a woman wants to join the labor force, she might pursue an education for that purpose. Also, joining the labor force may allow a woman to pursue an education).
- Women are more often than men caretakers of children, especially when they are young; this was especially true at the time of the data collection (1976).
- If a household has a high level of income independent of the woman's wages, she might see employment as unnecessary. This may suffer from bias – if a woman plans to never enter the labor force, she may not marry a man with a very low wage. Further, people with significantly different wages may not be as likely to marry.

Model 2: Returns of Education

The second question to explore is how education affects a woman's wages if she is in the labor force. If she is not in the labor force, naturally wages will equal zero, which eliminates the need for a dummy variable (women with wages of zero are discarded from the sample).

I will construct a model with this as one explanatory variable and *log(wage)* as the outcome variable.

Equation B1: $educinstru = \alpha_0 + \alpha_1 * motheduc + \alpha_2 * fatheduc + \alpha_3 * huseduc$

Equation B2: $\log(wage) = \beta_0 + \beta_1 * educinstru + \beta_2 * nwifeinc + \beta_4 * exper$

The reasoning behind this initial model is:

- It has been well-established that in general people with a higher level of education have higher wages. However, education exhibits endogeneity; it is highly correlated with experience, intelligence, ability, etc. We can solve for this by using 2SLS, using mother's, father's and husband's educational levels to provide a *source* of exogenous variation in education.
- Non-wife income would be expected to be highly correlated with the wife's wage. If the non-wife income is very high, her wage is probably lower (the husband or family would pay for a larger share of expenses).
- More experience should mean higher wages

I will exclude marginal tax rate from this model because it would be expected to be simultaneous; as wage increases, marginal tax rate increases. As more taxes are paid, wage decreases.

SECTION B: EMPIRICAL RESULTS

Equation A: $P(inlf = 1) = G(\beta_0 + \beta_1 * exper + \beta_2 * expersq + \beta_3 * age + \beta_4 * educ + \beta_6 * kidslt6 + \beta_6 * nwifeinc)$

Model 1A: Probit, obs 1-753
Dependent variable: inlf, QML standard errors

	Coefficient	*Std. Error*	*z*	*Slope*[*]
const	0.463352	0.449337	1.031	
exper	0.122110	0.0186951	6.532	0.0476835
expersq	−0.00188279	0.000597920	−3.149	−0.000735218
age	−0.0553178	0.00783762	−7.058	−0.0216013
educ	0.128694	0.0255899	5.029	0.0502542
kidslt6	−0.880900	0.115838	−7.605	−0.343987
nwifeinc	−0.0118298	0.00531759	−2.225	−0.00461947

The adjusted pseudo-R^2 for this regression is 0.206, and 73.8% of results are correctly predicted. The results are generally as expected: prior work experience is very predictive of whether a woman will currently be in the labor force. For the *average* (white, married) woman, one additional year of experience leads to a 4.8% increase in the probability that she is in the labor force. However, there is diminishing return to experience (the coefficient and the slope of expersq are negative).

For the average woman, adding one additional year of education yields a 5% increase in the probability that she is in the labor force.

Significantly, for the average woman, having one child less than six (average = 0.23, rounded to 0) decreases her probability of being in the labor force by 34.4%.

There are a few problems with my initial model. I supposed from the beginning that education is endogenous, but I did not account for this in this model. By running the regression again using *educinstru* as defined in Equation B1, the results are as follows:

Model 1B: Probit, using observations 1-753
Dependent variable: inlf, QML standard errors

	Coefficient	*Std. Error*	*z*	*Slope*[*]
const	0.800904	0.586809	1.365	
exper	0.126369	0.0184192	6.861	0.0494004
expersq	−0.00191717	0.000585646	−3.274	−0.000749468
age	−0.0552987	0.00780455	−7.085	−0.0216176
educinstru	0.0902834	0.0382708	2.359	**0.0352939**
kidslt6	−0.826630	0.114344	−7.229	**−0.323149**
nwifeinc	−0.00804594	0.00553411	−1.454	−0.00314534

In this regression, *educinstru* becomes less significant that *educ*, which is an expected result for an instrumental variable, as *educ* might be misleadingly significant from its correlation with ϵ. The adjusted pseudo R^2 for this regression is 0.185, slightly less than in Model 1A. 71% of the cases are still 'correctly predicted.'

Importantly, the null hypothesis that $H_0: \epsilon \sim N(\mu, \sigma^2)$ cannot be rejected at the 5% significance level.

Using the IV for education, we can see that an additional year of education has a smaller impact than was predicted in Model 1A. (+5% vs +3.5%, respectively). Furthermore, having one child reduces the probability of being in the labor force by 32.3% as opposed to the 34.4% in Model 1A. The *ceteris paribus* returns to experience and age remain relatively unaffected.

Surprisingly, non-wife income becomes less significant when IV is used for education (the magnitude of z value drops from 2.22 to 1.45, and the *ceteris paribus* return drops from -4.6% to -3.1%). This could be that husband education shares variation with husband income (included in non-wife income), but further analysis is necessary to make any meaningful conclusion.

Ceteris Paribus Effect (in percent, %) of One Unit Increase on Probability of Being in Labor Force (taken from the average)

	Model 1A	*Model 1B*
exper	+4.770	+4.94
expersq	−0.0735	-0.0749
age	−2.16	-2.16
educ or educinstru, respectively	+5.03	+3.53
kidslt6	−34.4	-32.3
nwifeinc	−.462	-0.315

Equation B2: $\log(wage) = \beta_0 + \beta_1 * educinstru + \beta_2 * nwifeinc + \beta_3 * exper$

Model 2A: OLS, using observations 1-428
Dependent variable: lwage

Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*
const	−0.0154772	0.303765	−0.05095	0.9594
educinstru	0.0640578	0.0245680	2.607	0.0094
nwifeinc	0.00899887	0.00285080	3.157	0.0017
exper	0.0185619	0.00432208	4.295	<0.0001

From *Model 2A*, the projected returns to education for an employed woman is 6.4% for each additional year of education. Although still significant even at the 1% level, *educinstru* is much less significant than *educ*, as is expected from 2SLS. The coefficient is also smaller. This could be explained by the possible endogeneity of *educ*: because education was expected to be correlated with the error term ϵ, I have instrumented education with mother's, father's and husband's education to remove any endogenous source of variation and isolate the exogenous sources of variation. This requires an exogenous assumption for each of those variables, which is most likely a good approximation.

The R^2 of this regression is 0.0736, which we must consider; this model explains only a fraction of the variation in *log(wage)*. The R^2 doubles when *educ* is used instead of *educinstru*. However, the F-statistic of *Model 2A* is 11.94, meaning that we cannot accept the null hypothesis that the model is insignificant.

The constant is completely insignificant (even at the 95% level), which is a large source of error in the model. Although the constant is skewed, the predicted marginal effects should not be too distorted from this, so it does not warrant a new model. The constant itself is not the focus of the analysis; rather, we are more interested in the *ceteris paribus* effects of factors such as education and experience on wage.

Wage Effects of 1 Unit Increase for Working Women

Variable	*Model 2A*	*Model 2B (replace educinstru with educ)*
const	-0.015 ± 0.303	-0.42 ± 0.183
educinstru	6.4%	10.2%
nwifeinc	0.9%	0.54%
exper	1.9%	1.7%

The results from this model suggest that perhaps the instruments used for education are weak. Statistical tests for weak instruments, however, suggest that all three IV for education (*motheduc*, *fatheduc*, and *huseduc*) are significant. The results of the model suggest that the returns to education are not so high as expected, but the results nonetheless conform to the initial expectations: education,

experience, and non-wife income are strongly correlated with *log(wage)* and an additional unit of each results in a higher wage.

Model 2B: OLS, using observations 1-428
Dependent variable: lwage
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*
const	−0.427723	0.182621	−2.342	0.0196
educ	0.102400	0.0137231	7.462	<0.0001
nwifeinc	0.00542396	0.00271112	2.001	0.0461
exper	0.0167906	0.00411494	4.080	<0.0001

CONCLUSION

Model 1B and *Model 2A* are my preferred models, each with a different goal. *Model 1B* is useful in predicting if a woman is in the labor force, and *Model 2A* is useful in demonstrating the wage effects of education.

In all models, I have shown that it is necessary to use IV for education, as the variable *educ* itself results in significantly higher estimations for wage effect and change in probability of labor force participation.

In general, the results conform to what should be expected in theory with a high level of confidence; the wage returns to education and experience are demonstrated to be positive with a high level of confidence.

Overall, I am most confident in the results of *Model 1B*, and I believe that it is the most powerful explanatory model. The wage returns to education and experience are, in many cases, 'obvious' and conform to what one would expect without any prior knowledge or analysis. The idea that having a child decreases the probability of labor force participation so significantly is more important because it can have important consequences in public policy towards women in or out of the labor force as well as women with and without children.

Model 1B opens the questions of how a child in the family affects men's participation in the labor force, or if men also are less likely to be in the labor force as they age. *Model 2A* will allow for the comparison between the returns to education for men and women. All these possible questions can be explored in future research.