Applying open science practices in empirical research on child development

Abstract:

Understanding the role of human capital investments made by parents is important for the study of intergenerational mobility. Frequently, however, the effects of these investments can only be determined using non-experimental techniques—often on existing secondary datasets—that aim at establishing causal links between investments and long-run outcomes. The mechanisms and parameters that result from this research are used by policy makers to inform the implementation and funding of social programs. However, current researcher degrees of freedom in conducting these studies too easily allow detrimental research practices, such as specification searching and data mining, that pose a potential threat to the credibility of this research. The goal of this paper is to argue for applying open, transparent, and reproducible science practices in regression-based causal-inference methods in child development research that use existing secondary datasets. First, we discuss the importance of credible evidence used to inform policy-making and threats to their credibility. Then, we discuss the open, transparent, and reproducible science practices proposed as a (partial) solution to these credibility threats. We provide an example of applying these in an analysis on intergenerational mobility that uses administrative data from Norway to empirically examine the relationship between (the timing of) parental income and child educational outcomes. Making commonplace the use of open, transparent, and reproducible science practices will increase the credibility of causal evidence from the regression-based econometric techniques frequently used to inform policy analysis and decision-making in child development.
I – Introduction

Understanding the role of human capital investments made during childhood is important for the study of intergenerational mobility, as it can inform policies that lead to more efficient labor markets and reduced inequality. Parental income and its timing have an effect on the quantity and quality of parent’s investments in children, which in turn have an effect on children’s outcomes (Mayer, 2010). For example, Solon (2004) finds that wealthy parents invest more in their children’s human capital and education, which results in higher earnings for these children later in life. As in other fields of study in economics and the social sciences, the literature of intergenerational mobility has moved from measuring correlations and elasticities to identifying and understanding casual mechanisms (Black and Devereux, 2010). Many of the studies that aim at finding these casual mechanisms in this field typically rely on already existing secondary datasets that can match parents and their children over decades to track long-term educational and labor market outcomes. Understanding the mechanisms underlying observed intergenerational correlations, and whether they are due to differential human capital investments, could suggest a role for public provision or financing of education as a policy to equalize opportunities between rich and poor children (Black and Devereux, 2010). In other words, rigorous empirical evidence can and should be used to inform policy decisions related to child development.

Using studies that rely on empirical analysis of secondary data to estimate the effect of child investments on later outcomes and intergenerational mobility is often the only way to inform empirically policy decisions that may have large impacts on child development. Using randomized trials or experiments to evaluate the impact of policies or programs on the long-term outcomes of children may be impossible to accomplish for ethical and practical reasons. Because
of these constraints, the production of credible causal evidence requires a careful design-based approach (Angrist and Pischke, 2010). Particularly for policy research that uses existing data to estimate parameters of causal relationships between variables of models used in policy simulations, it is important to ensure the validity of these casual claims. Therefore, replication and independent verification of the parameters used in these types of models is critical.

In this paper, we argue for applying open, transparent, and reproducible science practices in regression-based causal-inference methods in child development research that uses existing secondary datasets. First, we discuss the importance of credible estimates for model parameters used in policy-making and threats to their credibility identified through the current reproducibility crisis in social sciences more generally. Then, we discuss the open, transparent, and reproducible science practices proposed as a (partial) solution to these credibility threats. We demonstrate the application of these practices to part of a previous analysis (Carneiro et al. 2015) on intergenerational mobility and the role of parental income: specifically, a regression-based econometric analysis examining the causal effects of the timing of parental income on child educational outcomes. Promoting the use of these open science practices that facilitate the transparency and reproducibility of empirical research will increase the credibility of causal evidence from the regression-based econometric techniques frequently used to inform policy analysis and decision-making in child development.

II – The Reproducibility Crisis

There is growing concern about the credibility of empirical social science research (Ioannidis, 2005 and Galiani et al. 2017). As discussed in Ioannidis (2005), the “increasing concern that most current published research findings are false” is distorting the literature and
body of evidence with false positives, leading to issues with publication bias or the file-drawer problem (Franco et al. 2014). The result of this “distorted body of evidence with too few null effects and many false positives” is an exaggeration of the effectiveness of programs and policies, diverting public funds and efforts to potential useless or insignificant programs (Miguel et al., 2014). Two key sources of concern include: researcher misconduct and questionable /detrimental research practices. The main difference between them is the degree to which researchers intend to or knowingly present invalid results as true. The Office of Research Integrity, part of the U.S. Department of Health and Human Services,¹ defines research misconduct as “fabrication, falsification, or plagiarism in proposing performing, or reviewing research, or in reporting research results”, which includes making up or manipulating data or processes to present desired results. In this paper, we focus on the often unintentional or subconscious “detrimental” research practices that are commonplace in the social sciences. These include usually accepted research practices that have the potential to manipulate study results as desired and are facilitated by “Researcher Degrees of Freedom” (Simmons et al., 2011). Simmons defines “Researcher Degrees of Freedom” as the set of decisions researchers make in the process of collecting and analyzing data (Simmons et al., 2011). For example, p-hacking involves the misuse of data analysis to find (desired) patterns in data that can be presented as statistically significant when in fact there is no real underlying effect (Brodeur et al., 2016). One way of achieving this is through data mining or manipulating the data by adjusting sample sizes, or selectively dropping or including observations. Another way is “specification searching”, which involves adjusting model specifications through variable selection or operationalization, and selectively reporting only those tests that have significant effects.

¹ For more information see: https://ori.hhs.gov/definition-research-misconduct
Publication bias, inability to replicate, and specification searching remain widespread in social sciences disciplines such as economics (Christensen and Miguel, 2017). In fact, in the last six years, only 11 replication studies were published in the top 11 empirical economics journals (Galiani et al. 2017). All of these concerns also apply to studies on intergenerational mobility, as the incentives for research in this field are no different than those of other fields in economics or the social sciences. Therefore, empirical studies in this field would benefit from open, transparent, and reproducible research practices.

III – The Open Science Movement: Standards and Guidelines

There is a growing movement in the social sciences to improve the openness, transparency, and reproducibility of empirical research to improve its credibility (Miguel et al., 2014). Recently, this movement and its implications for the credibility of empirical evidence have also gained the attention of those working in policy making and analysis (Hoces de la Guardia et al., 2018). In an effort to align researcher incentives with open, transparent, and reproducible research, the Center for Open Science led a multi-disciplinary and multi-stakeholder group in the development of Transparent and Openness Promotion Guidelines (TOP)² for scientific journals’ publication policies and procedures (Nosek et al., 2015).

There are eight modular, yet complimentary, standards in the TOP guidelines: (1) citation standards, (2) data transparency, (3) analytic methods (code) transparency, (4) research materials transparency, (5) design and analysis transparency, (6) preregistration of studies, (7) preregistration of analysis plans, and (8) replication. A description of each standard, along with an example of journal guidelines which would and would not meet the standard is provided in

² For more information, please see: https://cos.io/our-services/top-guidelines/
Table 1. Furthermore, to accommodate for differences between journals and disciplines, the Committee developed levels for each standard (Nosek et al., 2015). Each level indicates an increase in the transparency of the standard. Journals would therefore be able to select at which level to hold each standard based on the readiness of authors and researchers to adopt each transparency standard (Nosek et al., 2015).

Table 1: Transparency Standards Description and Examples (Nosek et al., 2015)

<table>
<thead>
<tr>
<th>Standard</th>
<th>Description</th>
<th>Does not meet Standard (Level 0)</th>
<th>More Stringent Standard (Level 2 or 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Citation Standards</strong></td>
<td>Extends current article citation norms to data, code, and research materials.</td>
<td>Journal guidelines say nothing, or encourage citation of data, code, and materials.</td>
<td>Journal guidelines require and enforce appropriate citation for data, code, and materials.</td>
</tr>
<tr>
<td><strong>Data Transparency</strong></td>
<td>Incentivize authors to make data available in trusted repositories.</td>
<td>Journal guidelines say nothing, or authors may choose to share data.</td>
<td>Journal requires data to be posted in a trusted repository.</td>
</tr>
<tr>
<td><strong>Analytic Methods Transparency</strong></td>
<td>Incentivize authors to make code available in trusted repositories.</td>
<td>Journal guidelines say nothing, or authors may choose to share code.</td>
<td>Journal requires code to be posted to a trusted repository.</td>
</tr>
<tr>
<td><strong>Research Materials Transparency</strong></td>
<td>Encourage authors to provide materials used in the research.</td>
<td>Journal guidelines say nothing, or authors may choose to share material.</td>
<td>Journal requires materials to be posted to a trusted repository.</td>
</tr>
<tr>
<td><strong>Design and Analysis Transparency</strong></td>
<td>Increase transparency about the research process to reduce vague or incomplete reporting of methods.</td>
<td>Journal guidelines say nothing, or encourage design and analysis transparency.</td>
<td>Journal requires and enforce adherence to design and analysis transparency for review and publication.</td>
</tr>
<tr>
<td><strong>Preregistration of Studies</strong></td>
<td>Consists of registering a study in a public registry prior to conducting it.</td>
<td>Journal guidelines say nothing.</td>
<td>Journal requires pre-registration of studies and provides a link to it in the article.</td>
</tr>
<tr>
<td><strong>Preregistration of Analysis Plans</strong></td>
<td>Consists of registering an analysis in a public registry prior to conducting it.</td>
<td>Journal guidelines say nothing.</td>
<td>Journal requires a pre-analysis plans and provides a link to it in the article.</td>
</tr>
<tr>
<td><strong>Replication</strong></td>
<td>Consists of independent verification of research results.</td>
<td>Journal guidelines say nothing, or discourage submission of replication studies.</td>
<td>Journal encourages submission of replication studies, with (blind) peer review of the results.</td>
</tr>
</tbody>
</table>
IV – The Open Science Movement: Practices and Tools

In addition to these guidelines, a series of open science practices are being promoted and accompanying tools created to increase the replicability and reproducibility of empirical research. These practices and tools are categorized based on which step of the research they belong to: (1) design (which includes pre-registration, pre-analysis plans, and power planning), (2) conduct (which includes data management, version control, and open notebooks), (3) dissemination (that includes transparent reporting and disclosure, pre-prints, open access, and data visualization), and (4) archiving (which includes data repositories and dynamic documents). Below, we summarize each of these practices, how they can improve open research, and how could they be applied in this field and type of analysis. Ultimately, the goal of these practices and tools, especially if used in combination, is to give credibility to the results and to evidence of empirical social science research.

Design:

Pre-registration consists of stating research hypotheses and planned research design before the study takes place in a public registry. Pre-registration improves research openness, transparency, and reproducibility in several ways. It creates a record of the hypotheses that have been tested, regardless of the outcome that sometimes make publication impossible (e.g. null or negative results). It also reduces the risk of data mining, allows for a more careful identification of sample sizes needed for statistical test, and more generally helps researchers think through their research questions and methods more carefully. Ideally, it should occur before the start of data collection. When studies use secondary data, researchers can make open and public the intended analysis that was conducted using this existing data.
A pre-analysis plan is a step-by-step plan that describes how a researcher will analyze the data. As with pre-registration, it is ideal to do this before collecting the data. In studies that use readily available secondary data, this may be impossible. However, the researcher could still write a pre-analysis plan before having looked at the data (or in some cases even before the most recent wave of a public data is made available). The purpose of a pre-analysis plan is to prevent issues associated with data mining, which is particularly important in studies that look at multiple outcomes or have multiple hypothesis testing. The researcher has an incentive to try to look at a subset of outcomes that support their hypothesis. Pre-analysis plans prevent this from happening by setting out in advance the intended analysis. Pre-analysis plans are also useful in studies that collect data because they help researchers define the logic model behind their assumptions and allow them to think carefully about what variables they need and how these should be measured.

When designing studies testing a hypothesis, it is important to keep in mind the power of a study: that is, the probability of detecting an effect when there is an effect to detect. In most fields, 80% power is a lower bound threshold used for acceptable power. Issues with power are more relevant when researchers make decisions regarding how much data to collect (i.e. what is the correct sample size needed) and the smallest effect size of interest. For studies that use already collected secondary data, where sample sizes are already set, power planning can be used to estimate the minimum detectable effect size given the sample size.

**Conduct:**

Data management is a key component for open, transparent, and reproducible research. It is especially important during the collection of data and must balance issues of privacy (e.g.
personally identifiable data) with transparency and reproducibility. Version control plays an important part of data management as well as research openness, transparency, and reproducibility by maintaining a readable and sharable version of the code. Keeping track of the changes in the code ensures transparency in the way the analysis is conducted and the replicability of the study. By sharing this version controlled code with other researchers, the research is also made open and easier to reproduce. A popular tool for version control is Git. There are also personal reasons why researchers benefit from using version control, including helping keep one’s own files organized and make collaboration with others easier and cleaner.

Similar to the way that version control helps keep track of changes made in the code, open documents allow the same tracking for the written component of a study. In general, economists often use latex to write their paper, and this can also be combined with version control. Another useful tool is Jupyter. One important component related to data management, version control, and open notebooks is Docker. Docker is a tool that allows you to set the environment in which code can run that can be shared (as an image) with others. This allows for easy replication, as well as increasing transparency and openness. Another example is Code Ocean. A dockerfile is a file that allows the environment to run certain code to be set up, including building the dependencies (e.g. operating system needed, software programs and the packages needed) and pulling data available online, if possible.

**Dissemination:**

A key component for dissemination is transparent reporting and disclosure of the entire research process to the audience. Using previous data management and version control tools, as well as open notebooks would improve transparent reporting and disclosure, without adding
much more work for the researchers. Preprints allow research to be disseminated without having to wait for peer-reviewed publication. As with pre-registration, they help record the current research and reduce issues associated with publication bias. Open access would be improved, in general, by using tools that are open source (e.g. R) instead of more commonly used software (e.g. Stata or Matlab) which were used in this study. Newer tools of disseminating results could also be used to improve the way data and results are presented using data visualization tools. They are especially useful for sensitivity analysis and robustness checks that require multiple similar graphs.

Archiving:

Access to data and data availability are key components of open, transparent, and especially replicable and reproducible research. For other researchers to be able to replicate a study they require access to the data. Examples of publicly accessible data repositories include: Dataverse, Figshare, and Amazon S3. However, data privacy concerns, such as de-identification, should be addressed before making the data publicly available to ensure the safety of study participants.

Dynamic documents track changes made in a document and facilitate collaboration between researchers as well as easily share needed codes for data cleaning, analysis, and visualization. They also improve the replicability of a study by combining narrative text and code into a single document. Many of these tools work well many programing languages, interact well with open data repositories, and allow for interactive output for data visualization. They can also help structure the workflow and analysis, leading to more transparency. Examples of dynamic documents include: LaTeX, R-Markdown, and Jupyter Notebook.
V – Example: Study Description

In their paper on “International Mobility and the Timing of Parental Income”, Carneiro et al. (2015) extend the standard intergenerational mobility literature by using data from Norway to model an individual’s adult educational outcomes as a function of parental income throughout their childhood. Specifically, they examine the effects of parental income at three stages: early childhood (ages 0 to 5 years), middle childhood (ages 6 to 11), and late childhood (ages 12 to 17). The outcomes they focus on are child level of education, final educational achievement, and other outcomes later in life. The analysis is divided into three components. First, they describe the “association between different patterns of timing of parental income and the education (and other adult outcomes) of children” (Carneiro et al 2015). Second, they assess whether this association is causal using local linear regressions. Third, they examine whether the empirical patterns observed can be explained using economic models of parental investment in children with multiple periods of childhood. The results find that timing of parental income matters, specifically that “conditional on permanent income, education is maximized when income is balanced between the early childhood and middle childhood years” (Carneiro et al 2015). Furthermore, the empirical results match the results from a model with “income uncertainty, partial insurance, and complementarity between investments in children across periods” (Carneiro et al 2015). To answer their research questions, the researchers needed detailed and longitudinal administrative and demographic data of parents and children. The main data source used in this study comes from Norway’s census, the Norwegian Registry maintained by Statistics Norway, and other administrative data. This census data for the years 1971 to 2006 was linked across generations as well as to administrative datasets, including Military data.
V – Example: Applications and Limitations

The use/application of the open science tools/practices apply differently depending on the type of research, methodology, and data. Some of them are more easily applied to this type of analysis while some would be nearly impossible. The following section describes the application of these guidelines and practices to a subset of the analysis in the study described previously. For each, we discuss their importance, benefits, challenges, and limitations of applying them to our example. The materials created for this example are in an OSF Project Page https://osf.io/384bf/ and include: (1) a registration form, (2) an analysis plan, (3) the data that is available from the website, (3) code that cleans and prepares the data for the reproduction of part of the analysis, (4) code that reproduces part of the analysis, (5) code that outputs/summarizes the results, and (6) this paper as a preprint.

During the design phase, the analysis in this study would have benefited from pre-registration, pre-analysis plan, and power planning. These, however, would differ from the traditional design practices as outlined for experimental studies with primary data collection. Because this study uses previously collected secondary data, pre-registration in this case could not have occurred before data collection. However, pre-registration could still have been possible before the data were accessed, merged into the final data set, and prepared for analysis. Any form of study registration is still useful/necessary to address the “denominator problem” by creating a record of the research question and hypothesis of this study. Additionally, it informs other researchers in what capacity has this secondary data been examined and used. For this example, the pre-registration would include a description of the research questions, hypotheses, the data intended to be used, and power planning. Power planning is a common first step in
studies that collect primary data. However, even when using secondary data sets, power planning can provide researchers with information on the probability of finding a desired effect. Specifically, it provides researchers before beginning their study on the minimum detectible effect-size given the size of their sample.

A pre-analysis plan would have been an important contribution to the credibility and reproducibility of this study because they help prevent hypothesis hacking, decrease researcher degrees of freedom, specification searching, and other forms of data mining. For this analysis, a pre-analysis plan could have been written that including a description of the methodological strategy to be used and how the model would be specified. More specifically, it would have specified which educational outcomes were to be evaluated and how to measure parental income (e.g. pre/post taxes) at which stages (e.g. age ranges for early childhood). To provide an example of how these design practices could have been applied to a subset of the analysis in our study, a mock registration form and analysis plan are included as appendices.

While conducting this analysis, data management, version control, and open notebook would have facilitated future reproducibility and enhanced the quality of the analysis itself. Even though no primary data was collected, data management remains an important component. For this study, data management consist of keeping a record of the steps necessary to create the final sample used for the regression. This includes the source of the data, data cleaning steps taken (e.g. sub setting, manipulations, transformations, dealing with missing values, etc.), and how the different databases were linked and merged together into the final sample. The paper was not entirely detailed about the other sources of data or how this data was linked across datasets. Version controlled code and open notebooks could have facilitated this record keeping of data management. For this study, version controlled code or open document could have been
provided detailing all steps in the analysis, from obtaining the data (if possible through the API), to processing the data for analysis, and conducting the econometric analysis itself. Furthermore, both codes and open documents could go further to make replication easy by using a Docker file that specifies the dependencies (e.g. operating system, software programs and the packages) needed to run the code.

Open tools and practices could have also been used in the dissemination of this study. The paper tries to explain their analysis in a transparent way, but details are still missed. With such a large and complex analysis, it is difficult to disclose all of the steps, such as details on the data sources and data manipulation. One aspect of dissemination this paper does follow is preprints. This paper is a preprint because it is a “working paper” (Discussion Paper Series). Open access in this study could have also been improved by putting the final version of the data in a public data repository (e.g. Dataverse) and publishing in an open access journal. Furthermore, data visualization tools could also have been used to increase the general transparency. For example, plots in the appendix that show the robustness of their results to different specifications could have been made more transparent using data visualization tools such as R-shiny. Using R-shiny in an online appendix would allow other researchers to vary the parameters used as desired.

Archiving the data used for this study is a key component of replicability and reproducibility. Because the main data source for this study, the Norwegian Registry that is maintained by Statistics Norway, is only available to the public in aggregated tables, replication of this study was not possible. Statistics Norway, however, does have an open data API, where researchers can extract and create the data (in aggregate) they need. If more disaggregated data were available, the process of sharing the data used in this study could be accomplished by
sharing the code that uses this API to recreate the data used in the study, along with a Docker file that creates the correct environment to run the code. Furthermore, being able to link to live data in conjunction with the use of dynamic document would facilitate updating the results of this analysis are more (newer) data comes in (e.g. a new wave of the census). However, given the sensitive nature of this data, especially the military administrative data which contains IQ and health scores, important de-identification steps must be taken before sharing the data publicly.

VI – Conclusion

The main limitation to applying open, transparent, and reproducible research practices and standards to these types of study consist of: (1) developing a detailed pre-analysis plan and (2) privacy concerns with certain personally identifiable data. If researchers do not have enough information about the data structure then creating a detailed enough data analysis plan would be difficult. Furthermore, many of the model specification used in this paper relied on a detailed understanding of the data structure, knowledge about the number of missing observations or variables with few observations, for example, incomes in very low or high percentiles. Ideally, despite the sensitive nature of the data, a final version of the data that follows the FAIR Data Principles (that is, data that is findable, accessible, interoperable and reusable) could have been shared after ensuring that the data is secure and de-identified.

Applying these open, transparent, and reproducible research practices, as shown in our example, will enhance the credibility of causal evidence obtained from secondary-data sources. Particularly in the field of child development, increasing the accuracy and reliability of evidence will improve the design, implementation, and evaluation of policies and programs designed to improve child wellbeing.
References


Appendix A – Registration Form

The following is a mock registration form for a subset of the regression-based analysis using secondary datasets conducted in the paper “International Mobility and the Timing of Parental Income” (No. 9479, Institute for the Study of Labor Discussion Paper) by Carneiro et al (2015). This paper includes further research questions, hypotheses, and variables than described below, but these are not included or addressed. Furthermore, because the paper uses already existing secondary data, the pre-registration for this study would have ideally been done before any of the collaborators accessed, prepared, and analyzed the data.

Study Information

Title: Intergenerational Mobility and the Timing of Parental Income

Authors: Pedro Carneiro, Italo Lopez Garcia, Kjell G. Salvanes, Emma Tominey

Research Question(s):

Is there an association between the timing of parental income and a child’s education?

Is this association between child’s education and the timing of parental income causal?

Hypotheses:

The authors hypothesize that the timing of parental income matters more than parental income. Furthermore, the timing of parental income has a causal effect on child’s education consistent with economic models.

Sampling Plan

Data:

The data source for this study is the Norwegian Registry maintained by Statistics Norway, for the years ranging from 1971 up to 2006. This is a linked administrative dataset that covers the population of Norway. Some of the information it provides includes: month and year of birth, educational attainment, labor market status, earnings, a set of demographic variables, and parent’s marital status. Furthermore, it links individuals to their parents and obtain labor market information for both fathers and mothers.
The study also combines this data with administrative data of IQ and health scores of military tests taken upon entry to the compulsory military service for males between the ages of 18 and 20.

**Sample Size:**

The authors selected all births between 1971 and 1980 for which they could obtain (construct) annual parental taxable earnings by linking the individual data with their parents’ data and have data of the outcomes of interest, for a total final sample size of 522,490 children.

**Power Planning:**

Using the sample size above, we can estimate what would be the minimum detectable effect size of parental income on a child’s educational attainment. Using an 80% power and a 5% significance level, given our sample size of 522,490, and assuming the null hypothesis is 0, the minimum detectable effect size is: 0.0078.

**Variables:** To answer the research questions, the following variables are obtained and/or constructed.

- **Main outcome of interest:** individual’s adult educational outcomes, which includes years of education for each individual, an indicator for dropping out of high school at the age of 16, and an indicator for college enrollment. Individuals are at least 26 years of age when educational attainment is observed.

- **Other outcomes of interest:** IQ scores (on a scale of 0 to 9) obtained for males ages 17 and 20 from their compulsory military entry exam, health scores (on a scale of 0 to 9) obtained for males ages 17 and 20 from their compulsory military entry exam, grades achieved at the end of 10th grade when individuals are aged 16 (for cohorts born on or after 1986), and teenage pregnancy as an indicator for whether an individual female has a child between the ages of 16 and 20. Also, average annual earnings at age 30.

- **Main “treatment” variable:** parental income, which is constructed linking individuals to their parent’s annual taxable earnings data for each year, starting from three years preceding the child’s birth until their 20th birthday. Earnings include wages and income from business activity, as well as unemployment and sickness benefits.

- **Other “treatment” variable:** stages of an individual's life in which parental income may have an effect, divided into childhood (ages 0 to 5 years), middle childhood (ages 6 to 11), and late childhood (ages 12 to 17).

- **Control variables:** age, gender, parental years of education, parental age at birth, marital status of the parents, family size, and municipality of residence (for each year of the child's life).
Appendix B – Analysis Plan

The following is a mock analysis plan for a subset of the regression-based analysis using secondary datasets conducted in the paper “International Mobility and the Timing of Parental Income” (No. 9479, Institute for the Study of Labor Discussion Paper) by Carneiro et al (2015). This paper includes further analysis and modeling that are not included or addressed in the methods and models below. Ideally, a pre-analysis plan would have been prepared before any of the actual analysis for this study occurred.

Overview

The first component of the analysis will be a descriptive analysis of the association between different patterns of timing of parental income and educational (and other adult) outcomes of children. The second component will be to assess the extent to which this association can be interpreted as causal. To make this claim, two issues must be addressed. First, the timing of parental income may be endogenous and partially correlated with investments in children (e.g. taking maternity leave). Second, heterogeneity across parents can simultaneously be related with income profiles and investments in children (e.g. parents with steep income profiles may provide different parenting than parents with flat income profiles). We will attempt to address these issues by using only father’s income, controlling for permanent father’s income (measured either during the child’s life or the entire life of the father, depending on the available data), parental education interacted with parental age and the slope of the individual specific income profile. We will also perform certain robustness checks described below.

Statistical Models:

First, we will perform parametric regressions of the timing of parental income and educational outcomes. We will begin examining only the effect of paternal permanent income ($PI_i$), constructed as $PI_i = I_{i1} + I_{i2} + I_{i3}$ where $I_{it}$ is the paternal income in each period (t=1 if age is between 0 and 5, t=2 if age is between 6 and 11, and t=3 if age is between 12 and 17), on years of education. We will allow for a more flexible specification, by constructing indicator variables for each percentile of the permanent income distribution, $q_{Pl,i}^k$, which takes a value of 1 if the paternal permanent income of individual $i$ is in percentile $k$ of the distribution and 0 otherwise:

$$Y_i = \sum_{k=1}^{100} \phi_{Pl}^k q_{Pl,i}^k + \epsilon_i$$
Then, we will introduce income during two periods of childhood (period t=2 and t=3, leaving period t=1 out due to collinearity) by assuming an additively separable function for \( m(PI_t, l_{i2}, l_{i3}) \), which is a function that links the timing of parental income to outcomes, as such:

\[
m(PI_t, l_{i2}, l_{i3}) = m_1(PI_t) + m_2(l_{i2}) + m_3(l_{i3})
\]

Again, we will approximate each component using dummies for each percentile of the distribution, and add controls, as shown below:

\[
Y_i = \sum_{k_1=1}^{100} \phi_{k_1} q_{k_1} + \sum_{k_2=1}^{100} \phi_{k_2} d_{k_2} + \sum_{k_3=1}^{100} \phi_{k_3} q_{k_3} + Z_i \delta + \epsilon_i
\]

Where \( Y_i \) is the outcome of interest (i.e. education, high school completion, college attendance, earnings, IQ, health, teenage pregnancy, grades in high school), \( Z_i \) are a set of covariates (i.e. paternal years of education interacted with paternal age at birth, paternal income growth based on pre-birth and post-17 income, maternal years of education interacted with paternal age at birth, year of birth, and gender), and \( \epsilon_i \) is the unobserved heterogeneity that is left after controlling for permanent income. As before, \( q_{k_1}, q_{k_2}, q_{k_3} \) are indicators that take the value 1 if the father of child \( i \) has permanent income, income in period t=2, or income in period t=3, in \( k \) percentile and 0 otherwise. From this specification, we can test and reject that \( m_2(l_{i2}) \) and \( m_3(l_{i3}) \) does not vary with \( l_{i2} \) and \( l_{i3} \). Particularly, we test whether the coefficients \( \phi_{k_2} \) and \( \phi_{k_3} \) are all equal to each other across different values of \( k_2 \) and \( k_3 \).

**Robustness Checks:**

One of the assumptions of this model is that \( l_{i2} \) and \( l_{i3} \) are uncorrelated with \( \epsilon_i \) after conditioning on \( PI_t \) and the controls \( Z_i \). This implies that pre-birth investments should be uncorrelated with the timing of income, but may still affect child outcomes. To test this, we will examine the relationship between having a low birth weight baby and subsequent timing of parental income. Second, we will also control additionally for marital status and the number of children in the households, which will be excluded from the main specification because they are potentially endogenous to family income. Lastly, we will check the construction of our measures for per period income and permanent income by checking our assumption of which fixed discount rate to use to construct the parental income measures. We will test different fixed rates (0%, 2%, 4%, 6%, 10%, 15%), as well as time-varying discount rates using official real interest data for Norway between 1971 and 1998.

**Inference Criteria:**

For our significance tests, we will use a statistical significance level of \( \alpha = 0.05 \) (i.e. the coefficients will be statistically significant if their p-value < 0.05).