Reproducibility and Applied Econometrics - The Effect of Studying on Grades*

Nadav Tadelis

May 2018

DRAFT

Abstract

In this paper we establish a framework for reproducible empirical research. We use a non-standard 2SLS model to estimate the marginal effects of studying on grades. The paper is split into two distinct sections. The first part is the econometric analysis on the causal impact of studying. The second part details the steps taken to ensure reproducibility and suggests how to easily integrate these methods into a researcher's future projects.

Git Repository: `https://github.com/nadavtadelis/Reproducible_Metrics`

*I would like to thank my thesis advisor, Professor Fernando Pérez, for his support and insights, and Professor Maximilian Auffhammer for his generosity with his time.

1 Introduction

In recent years there has been a strong push to increase the reproducibility and replicability of scientific research. Unfortunately this movement seems to have been centered on the hard sciences and has not yet become standard practice in the social sciences. It is possible that this is partly due to a lack of reproducibly researched papers in these fields. This paper explores what responsible and reproducible research practices look like in applied econometrics. We present an instrumental variables approach to estimating the causal effect of studying on grades and develop custom python scripts to implement an unusual 2SLS set up that is specialized for nonlinearity in our endogenous predictor. The latter part of this work discusses the current state of reproducibility in econometric research, and explains in detail the techniques implemented in the analysis.

This paper was written as my honors thesis for undergraduate studies in statistics at UC Berkeley. Due to deadlines for submission I have been unable to spend an appropriate amount of time on the econometric analysis, and will point out some of the weak points in my models and assumptions. Any comments or suggestions would be greatly appreciated.

The rough idea for the econometric analysis in the paper is adapted from an independent project I completed in Spring 2017. In a subsequent class the project was modified and rebuilt in a reproducible fashion. Sarah Johnson created the original intermediate functions in `p3functions.py`, and the associated tests and Travis CI. Chitwan Kaudan created the original Makefile for running the individual Jupyter Notebooks. All aspects of those original analyses have been altered significantly, and the history of the alterations is fully documented in the commit history of the git repo.

[Maybe more background here before we dive in?]

DRAFT

2 Data

The data being used are from the public archive of UCI's machine learning repository and were collected by Paulo Cortez of the University of Minho, Portugal in the 2005 - 2006 academic year (Cortez and Silva [2008]). The data were collected in two secondary schools in the Alentejo region of Portugal, using school reports and questionnaires. The data were cleaned to only include students for which all the variables are known - and a further 111 students were discarded because of mismatched information between the surveys and the school reports. The data come with a file containing attribute information which can be found here; these include school, course, and many individual level characteristics. The data include 649 students from a Portuguese Language course of study, and 395 students from a Mathematics course.

[Need more about data collection methods here? There isn't much description in the original paper, maybe this is enough?]

2.1 Data Exploration

[This section is very rough, needs work!]

There is some pre-analysis data cleaning and exploration that we run, which can be found in the data exploration notebook - nbviewer, git; some figures from this notebook are included in appendix 6.1 [MAKE SURE TO INCLUDE THESE]. The notebook contains the full data exploration process, as well as explanations and commentary detailing the figure generation process.

We draw attention to two attributes of special import: there are stark differences in the distributions of the data between the two schools, and across the two courses of study. The school level differences seem to be fairly consistent; perhaps the two schools have slightly different grading policies. However, the within school course level differences are highly variable, this is perhaps capturing some heterogeneity between students who choose Portuguese Language and Mathematics. This strongly suggests that even our simplest models should include controls for school to account for different policies, and should estimate the two courses as separate entities to account for different types of students in each course.

2.2 Data Issues

[Very Rough, needs a lot of work]

DRAFT

As is often the case with education data, the data have some large issues that must be considered. As any undergraduate taking their first econometrics course would point out, much of the data are stated preferences (as opposed to observed). Happily, the test scores and number of absences are provided by the schools, so we can be certain in their accuracy. The issue with self-reported data is the potential for inaccuracy. Even if a student is not maliciously providing misinformation, it is likely that personal biases are affecting the response. With variables like weekly hours of free time, this self reported data might actually be a benefit, as we are getting information about the student's perception of reality rather than the truth. If a student reports that they have very little free time then that tells us something about how they view their current time allocations. As such, these variables might be useful as controls for student level heterogeneity, but should be considered with a healthy dose of skepticism, and estimated coefficients should be interpreted with care. [Think about this more.] The issue of self reported data becomes more problematic when it comes to our variable of interest - hours of studying per week. While our central research goal is identifying a relationship between hours of studying and academic success, our data on hours of studying cannot be trusted as accurate [expand more here].

In addition to the issue of self reporting, our data on studying time (and other variables) suffers from another issue; categorical mappings of quantitive measures. Many of the quantitive variables in the data are reported as categorical bins. For example, weekly studying time is coded as four distinct levels (0-2 hours, 2-5 hours, 5-10 hours, and 10+ hours). In the case of studying time, we explore two different re-mapping schemes[1]. But for the other variables with this format we treat them as categorical variables and include indicators for each level (thus allowing for some nonlinearity). [expand more here]

Lastly, the data are cross-sectional rather than longitudinal, and the specific sampling time is unspecified. This is especially troubling in this setting because we are provided no information on when the student surveys were administered. Weekly studying time may change throughout the year, and may be partially determined by grades of previous examinations. If this is the case, and students update their studying allocation based on exam results, there is clear simultaneous causality between studying time and exam grades[2]. [Do I need more here? Or is this enough of a brief intro to the later section?] - or should i just move the simultaneous causality section to here?

[1] Mapping both as discrete and continuous, detailed in the data exploration notebook - nbviewer, git

[2] A point explored more in section 3.1.2

DRAFT

3 Models

[Do I need to focus more on the proxy model here?]

We need to first set up our structural equation defining the relationship between grades and studying (Card and Krueger [1992]). Let an individual's grade be g_i and weekly hours of studying be s_i and their "ability" be a_i. Then our model is:

$$g_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 a_i + \boldsymbol{x}_{i,4:k}\boldsymbol{\beta}_{4:k} + \varepsilon_i$$

Where $\boldsymbol{x}_{i,4:k}$ is a vector of school and course level characteristics and ε_i captures some unobserved heterogeneity and disturbances. Note that grades are nonlinear in weekly study time. While this nonlinearity complicates our model, it seems necessary because assuming that marginal returns to studying must vary depending on the initial level of studying.

Clearly, there are issues with this model. How are we defining grades? The ideal set up would have a course specific set of simultaneous equations, where the number of equations is equal to the number of classes. The next best set up would involve estimating one equation for each type of class (quantitative, literary, historical, etc.) within each course. Another alternative would be to define grades as cumulative GPA. In this analysis, due to the limitations of the data, we define g_i as the student's score on the final test for their course of study (G3 in the data).

Another issue with this model is: how are we defining ability? By its very nature, ability is unobserved. We can proxy for ability using other individual level characteristics (intelligence, age, parents' education, etc.) but we cannot fully capture ability because it does not have a clear measurable meaning. Hence, we must keep in mind that the model is never going to be fully specified.

We can think of a student's utility maximization problem as being some function of grades, free time (let's consider everything that is not studying, sleeping, or class as "free time"), and studying time. We would expect that the coefficients on grades and free time would be positive, and the coefficient on study time would be negative, with magnitudes of these coefficients being determined by an individual's preferences. For example, a student who cares very little about grades, enjoys constantly partying, and hates studying, would have a small positive coefficient on grades, a large positive coefficient on free time, and a large

DRAFT

negative coefficient on studying time[3]. This student would maximize their utility and choose how to allocate their time, and would probably end up spending very little time studying. Notice that before maximizing their utility, an individual would replace grades with the previously defined model for grades (dependent on studying and ability); so someone who heavily values grades could end up having a positive coefficient on studying after including the model for grades into their utility function, even if they do not intrinsically value studying.

Establishing this utility function gives motivation for including variables that might introduce multicollinearity. For example, weekly amount of free time might not improve our estimate of the marginal effect of studying, and would be collinear with the amount of weekly studying. However, when we think of our observations as realizations of a decision making process that involves utility maximization over unobserved ability, there is an argument for including free time in the model estimation as a proxy for ability.

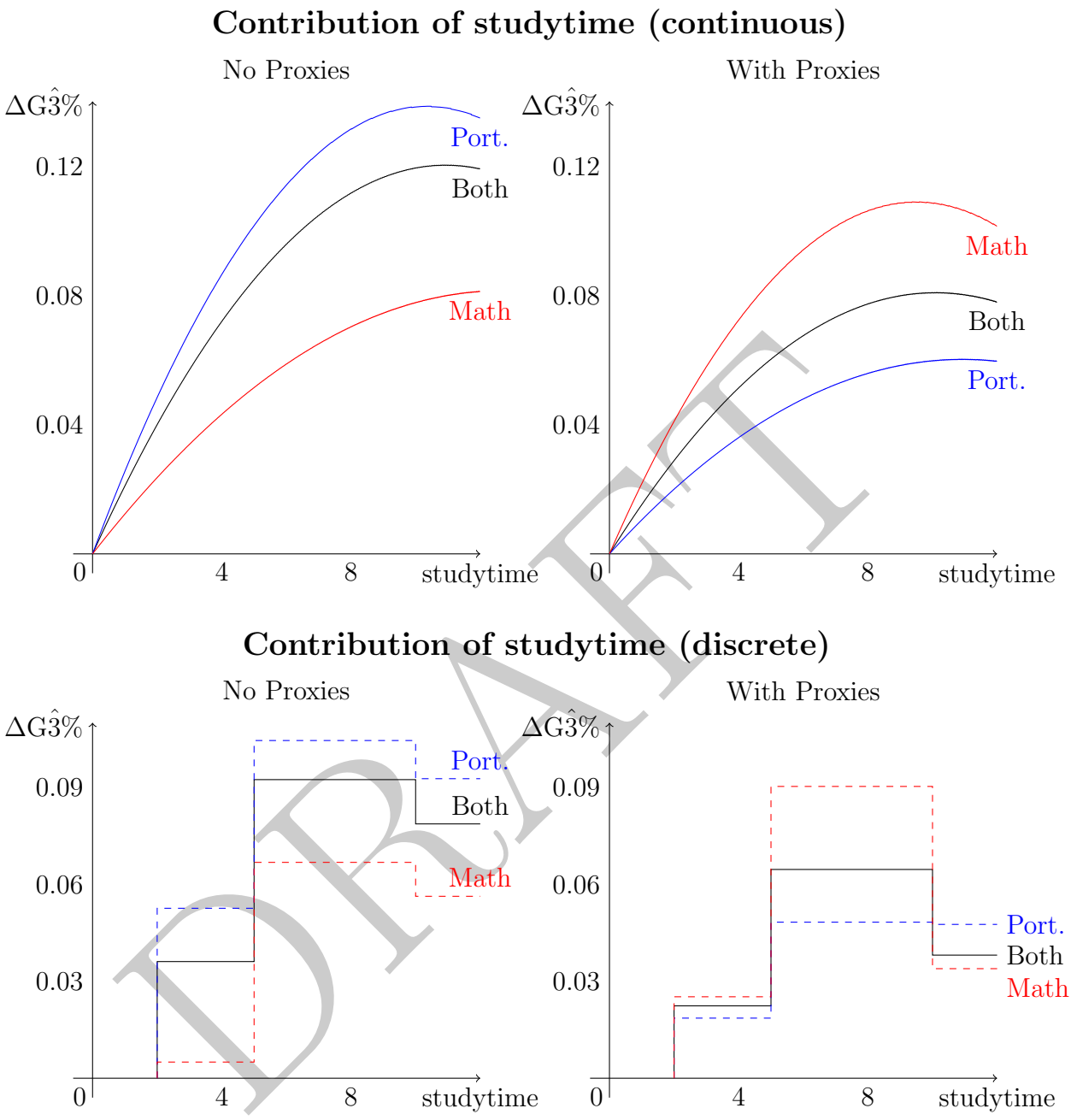
3.1 Naive OLS

[Plots need finishing (better labels, axis names, titles for with/without controls, choose line type and color, etc.] [Maybe only include the first two plots in this section, and include the discontinuous mapping only in appendix?]

In this section we discuss results from the naive model specifications we estimate in the model fitting 1 notebook - nbviewer, git. Below we plot studying time's estimated contribution to the final score (in percent).

[3]Of course, for some people and some subject matters of study, the coefficient on studying time may be positive with a decreasing marginal utility. We simplify the specification here dramatically.

DRAFT



DRAFT

3.1.1 Additional Analysis

In the model fitting 1 notebook we also include some additional explorations of the estimated models. We compute the variance inflation factor to check for multicollinearity (O'brien [2007], Mansfield and Helms [1982]). We find slight collinearity between studying time and its square, but we are not worried about this collinearity because it is on our variable of interest, and the VIF is low enough to ignore. The rest of the collinearity that the VIF detects is so low that we don't need to consider it.

Additionally, we include a brief exploration of a LASSO-flavored penalization for variable selection (Cameron [2017]) . We find that if we select a penalization parameter that shrinks all but seven of the coefficients to zero, the remaining coefficients are on studying time, school, course of study, number of failures, interest in higher education, mother's education, and whether the student is in the lowest bin of alcohol consumption. It is reassuring to see that the coefficient on studying time is not shrunk to zero. It is also nice to see that lasso shrinkage retains mother's education, as mother's education is often used as a proxy for ability, and this result seems to support it's importance. We discuss the implications of the other non-zero coefficients in more detail in the model fitting 1 notebook - nbviewer, git.

3.1.2 Simultaneous Causality

There is an additional issue with this model specification that was alluded to in section 2.2. The data include three test scores: G1, G2, and G3. G1 and G2 are midterms, and G3 is the final. It is plausible that part of how students inform their study allocation decision comes from their results previous exams, and the how much time they studied for those exams[4]. The data is taken at a single undefined point in time, so it is unclear when the students are reporting their weekly amount of studying. This complicates things; if the survey was administered after both midterms, then we may not expect their studying decision to change between the survey and the final, but if the survey was administered before either of the midterms then there is clear simultaneous causality between studying time, G1 and G2. Since we have only the single survey results with no time stamp, determining this relationship is intractable. With this limited information we must move forward either assuming that the reported studying times do not vary with mid term test results, or that the students were surveyed after both mid term results were released. These assumptions motivate the decision to exclude G1 and G2 from the model estimation.

3.1.3 Endogeneity

In the second run of naive regressions we add many covariates as proxies representing unobserved ability (parental education, individual characteristics, and others detailed in the model fitting 1 notebook - nbviewer, git). We pointed out that since ability is unobservable and undefinable we cannot expect to fully capture it through our proxies. Let q_i be the

[4]For example, if a student studied 5 hours a week leading up to the exam and received an 80% she may choose to study for longer in the future in the hopes of increasing her grades.

DRAFT

unobserved portion of ability that we are not capturing through our imperfect proxies. Then we have:

$$g_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 school_i + \beta_4 course_i + \beta_5 x_{i,5} + \cdots + \beta_k x_{i,k} + \gamma q_i + \nu_i$$

Where $x_{i,5}, ..., x_{i,k}$ are proxies for ability and ν_i is the structural error term. Let $x_{i,j}$ be the j'th covariate in the design matrix ($x_{i,1} = 1$, $x_{i,2} = s_i$, etc.) and let $\boldsymbol{x}_i$ be the vector of covariates for i. We are mainly interested in β_1 and β_2 the coefficients on studying. It is reasonable to assume that $\mathrm{E}[\nu_i | \boldsymbol{x}_i, q_i] = 0$ unfortunately we have no choice but to stick q_i into the error term which gives us the new structural equation:

$$g_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 school_i + \beta_4 course_i + \beta_5 x_{i,5} + \cdots + \beta_k x_{i,k} + \varepsilon_i$$

Where $\varepsilon_i = \gamma q_i + \nu_i$. We know that ν_i is well behaved (uncorrelated with $x_{i,j} \forall j \in \{1, k\}$) but ε_i is uncorrelated with the covariates iff q_i is uncorrelated with the covariates. Lets look at the consistency of this set up (starting with the linear projection of q_i onto $\boldsymbol{x}_i$):

$$\begin{aligned}
&g_i = \delta_0 + \delta_1 s_i + \delta_2 s_i^2 + \delta_3 school_i + \delta_4 course_i + \delta_5 x_{i,5} + \cdots + \delta_k x_{i,k} + \eta_i \\
&\mathrm{E}[\eta_i] = 0 \qquad (1) \\
&\mathrm{Cov}[x_{i,j}, \eta_i] = 0 \qquad (2)
\end{aligned}$$

Where (1) and (2) follow from the definition of linear projections. Now we can represent our original model in terms of this linear projection:

$$\begin{aligned}
g_i = \;& \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \beta_3 school_i + \beta_4 course_i + \beta_5 x_{i,5} + \cdots + \beta_k x_{i,k} \\
& + \gamma\delta_0 + \gamma\delta_1 s_i + \cdots + \gamma\delta_k x_{i,k} + \gamma\eta_i + \nu_i \\
g_i = \;& (\beta_0 + \gamma\delta_0) + (\beta_1 + \gamma\delta_1) s_i + (\beta_2 + \gamma\delta_2) s_i^2 + \cdots + (\beta_k + \gamma\delta_k) x_{i,k} + \gamma\eta_i + \nu_i \\
\Rightarrow\; & \mathrm{E}[\gamma\eta_i + \nu_i] = 0 \\
& \mathrm{Cov}[x_{i,j}, \gamma\eta_i + \nu_i] = 0
\end{aligned}$$

This shows that we can run OLS and get consistent estimates of $(\beta_j + \gamma\delta_j)$, but we are actually interested in $\boldsymbol{\beta}$. OLS does not allow us to recover $\boldsymbol{\beta}$ in this case because:

$$\hat{\beta}_j \xrightarrow{p} \beta_j + \underbrace{\gamma\delta_j}_{\text{omitted variable bias}}$$

DRAFT

If $\gamma \neq 0$ and $\delta_j \neq 0$ then we cannot get a consistent estimate of β_j using OLS. Rather, we get an endogenous estimator with a bias of $\gamma\delta_j$ (notice that the better our proxies identify ability, the smaller the bias term will be). Often times in practice researchers will assume that all δ_j's are zero except for the ones on the variables of interest. The implicit assumption behind setting $\delta_j = 0$ is that the unobserved q_i and $x_{i,j}$ are independent (i.e. $\text{Cov}[x_{i,j}, q_i] = 0$). In our setting this is actually a very reasonable assumption; we have already defined $x_{i,5}, ..., x_{i,k}$ as proxies for ability, meaning that the unobserved portion of ability should be orthogonal to each of these variables. It is also reasonable to assume that there is independence between school and course level characteristics and the unobserved q_i. This simplifies[5] our endogeneity problem so that the only endogenous variables are s_i and s_i^2 ($\text{Cov}[s_i, \varepsilon_i]$, $\text{Cov}[s_i^2, \varepsilon_i] \neq 0$). In the next sections we first give a brief overview of the instrumental variables approach to addressing endogeneity, and explain why standard two stage least squares is intractable in our setting. We then explore an instrumentation approach with nested generated regressors that may solve the problem.

3.2 Q2SLS

Lets first give a brief refresher of the instrumental variables approach to addressing endogeneity. In the last section we showed that for an endogenous variable $x_{i,j}$ OLS is no longer a consistent estimator for β_j. The instrumental variables approach allows us to solve the endogeneity problem. We require an instrumental variable z_i which is observable and satisfies the following conditions:

(1) The instrument is uncorrelated with disturbances: $\text{Cov}[z_i, \varepsilon_i] = 0$

(2) In the reduced form of $x_{i,j}$: $x_{i,j} = \phi_0 x_{i,1} + \phi_1 school_i + \cdots + \phi_{k-2} x_{i,k} + \theta_1 z_i + \vartheta_i$ the coefficient on our instrument is nonzero: $\theta_1 \neq 0$

(3) And z_i is not one of our exogenous variables in the original estimation

If (1) – (3) hold, then we call z_i a valid instrument for $x_{i,j}$. In settings with multiple valid instruments $\boldsymbol{z_1}, \boldsymbol{z}_2, ..., \boldsymbol{z}_M$ the standard procedure for identifying $\boldsymbol{\beta}$ is 2SLS. In the first stage of 2SLS we regress the endogenous variable $x_{i,j}$ on the exogenous variables and the instruments to get the fitted values $\hat{x}_{i,j} = \sum_{k \neq j} \{\hat{\phi}_k x_{i,k}\} + \hat{\theta}_1 z_{i,1} + \cdots + \hat{\theta}_M z_{i,M}$. We use $\hat{x}_{i,j}$ as an estimate of the exogenous part of $x_{i,j}$ and in the second stage we regress our dependent

[5] If we had only one endogenous variable we could simplify further because $\hat{\beta}_j \xrightarrow{p} \beta_j + \gamma \frac{\text{Cov}[x_j, q]}{\text{Var}[x_j]}$. In this case we would be able to make some guesses about the direction of the bias.

DRAFT

variable on our exogenous variables and $\hat{x}_{i,j}$ to identify $\hat{\boldsymbol{\beta}}$. This approach can be easily extended to cases with multiple endogenous variables (where you would have as many first stage regressions as there are endogenous variables). However, with multiple endogenous variables, you must have at least as many instruments as endogenous variables, otherwise the fitted values would be collinear and the second stage estimators would lose consistency. In the next section we discuss issues with the 2SLS method in our setting.

3.2.1 Motivation

In section 3.1.3 we established that the causal relationship between our grades and studying time is quadratic. This would usually not be a barrier to using 2SLS; the standard approach[6] is to use the original instrument and its square as instruments in the two first stage equations (Angrist and Pischke [2009]). However, this approach is unusable when the only instrument available is binary. In fact, even if we have multiple binary instruments it is not convincing to think that the exogenous portion of the squared endogenous variable will be correctly identified because interacting the binary instruments gives very little additional variation [THINK ABOUT THIS MORE]. Happily, Wooldridge suggests a variant of 2SLS that seems to address this issue of identifying a nonlinearly transformed variable with linear projection on binary variables.

3.2.2 Estimation Procedure

The procedure that Wooldridge suggests is as follows (Wooldridge [2001]):

(1) *First Stage:*

   (a) Regress the endogenous variable[7] $x_{i,j}$ on the exogenous variables and the instruments to get fitted values $\hat{x}_{i,j}$

   (b) Regress the endogenous variable squared[8] $x_{i,j}^2$ on the exogenous variables, the instruments, and the squared fitted values $(\hat{x}_{i,j})^2$. This gives fitted values for $\hat{x}_{i,j}^2$

(2) *Second Stage:*
Regress the dependent variable on the exogenous variables and the two fitted values from the first stage ($\hat{x}_{i,j}$, $\hat{x}_{i,j}^2$)

[6] Assuming only one instrument is available.

[7] In our model specification this is $s_i = x_{i,2}$

[8] In our model specification this is $s_i^2 = x_{i,3}$

DRAFT

The intuition behind this model is that while using $(\hat{x}_{i,j})^2$ will not produce consistent coefficient estimates[9] for β_j; $(\hat{x}_{i,j})^2$ is still a square of the exogenous portion extracted from $x_{i,j}$ in part a of the first stage, and as such can be used as a valid instrument for $x_{i,j}^2$ [NEED TO MAKE THIS MORE CLEAR]. This should allow us to identify $\boldsymbol{\beta}$ even with the quadratic endogeneity and binary instrumentation.

3.2.3 Properties of Q2SLS

[Explain the testing that I did for Q2SLS, mention that it looks like the coeff.s on endog_sq_hat and the exog vars are consistent and unbiased, but that the coeff on endog_hat is consistent but biased. Point out the oddness of this behavior and note that I plan on researching this further.] Note: pretty sure my instruments are weak (need to check), the simulation suggests that in the case of weak instruments the estimates are good but that the SE's are going to be large for the endogenous vars.

[Discuss difficulty with finding the asymptotically consistent estimator for variance in Q2SLS because of the nested generated regressors]

[Give brief overview of bootstrapping and explain how its being used here to estimate variance of coeff. estimates in 2nd stage] I think I'll need to explain that we report the coefficient estimates built on the full sample, rather than the average coefficients from the bootstrapping step - maybe link this back to the question of why the coefficient estimates are biased in the simulation, its not the same thing because here we are re-drawing and in simulation they are new draws, but it seems like there may be some connection here (maybe). Also add note about future integration of a wild bootstrap approach to variance estimation (stricter assumptions, but some really nice properties for endogenous models - see slide 15 from the link below)

Mention that the current method of bootstrapping is called "pairs bootstrap" in regression setting, and "nonparametric bootstrap" more generally.

FOR BOOTSTRAPPING COV MATRIX: See 8th slide of LINK. It also has notes on hypothesis testing and p values for bootstrapping with regression coefficients. SLIDE 14 HAS DISCUSSION OF BOOTSTRAPPING WITH ENDOGENOUS VARS!! (Wild Restricted Efficient Bootstrap (Davidson and MacKinnon 2010) is good for endogeneity and is good for hypothesis testing when you have weak instruments)

[9]Because the linear projection of the square is not the square of the linear projection, this mistake is known as the forbidden regression

DRAFT

3.2.4 Application

[Discuss chosen instruments and give motivation for why they might be okay to use, give strong disclaimer that even if they are valid, they are probably weak instruments] Justification for using 'going out' as instrument is stronger if we explain that we have controls for weekly and daily alcohol consumption (which is one thing that could affect grades through going out independently of studying time) so its a semi reasonable claim to say that going out is exogenous variation in available studying time once the detrimental effects of going out (namely drinking) are controlled for. not a super strong argument, but at least a little justification. Point out that in the monte carlo simulation, weak instruments resulted in (seemingly) unbiased consistent estimates for the true coeff, but that the variance was massive - which suggests that we should expect a massive variance on our coeff.s because we have weak instruments.

[Report results] Include plots like the naive ols plots, one for 2SLS and one for Q2SLS

[Discuss results] Note: seems like we should expect that after identifying the causal effect we would see a decrease in the size of the coefficient on studying time because before that coefficient could have been positively biased because people who study more might also have a higher level of motivation of 'ability', so there might be a plausible intuitive argument that the size of the coeff should go down. need to think about this a bit more, but it might be worth mentioning.

Maybe: [Reference appendix, and in the appendix include the results from Q2SLS with only 'essential' controls in the extra results subsection. Similar to the approach for the naive OLS at the beginning]

DRAFT

4 Reproducibility

[Need to clean up this section a lot]

The importance of reproducible research has been understood for many years. In 1980 Thomas Mayer said "Neither originality, logical rigor or any other criterion is ranked as 'essential' by so many natural scientists as is replicability". Mayer was focusing on replicable research; unfortunately in the social science replicability is often far harder than reproducibility. When it takes five years and tens of thousands of dollars to collect the data for a single study, it is a difficult proposition to attempt to replicate the study from scratch. Additionally, it is unlikely that Mayer could have foreseen the degree to which data wrangling and programming would penetrate into academic research. This reliance on complex data pipelines and analysis with hundreds of dependencies makes replication a beast of a whole new nature. In the current state of empirical research, reproducing the results is half the battle. This opaqueness has led to many published papers that included results that were 'p-hacked' or just included mistakes. Ensuring reproducibility would help alleviate many of these issues, and is a big step on the road to replicability.

Reproducibility has become especially important in the current political environment. The National Association for Scholars' recently published a report in which they advocate for use of the Secret Science Reformation Act (renamed the HONEST Act), which would forbid the Environmental Protection Agency from using any research that is not "substantially reproducible" Schulson [2018]. The NAS report goes even further and suggests that the bill should encompass all federal agencies and courts. Putting aside the various issues with the act[10], if HONEST were to be expanded to other federal agencies, then policy-oriented research in the social sciences would be forced to adapt. Policy makers would need to see "substantial reproducibility" before allowing a study to influence their policy decisions. Currently, very few empirical papers could be labeled as even somewhat reproducible, and part of the reason why might be that researchers believe implementing reproducible methods into their projects would be too costly in time and resources Feigenbaum and Levy [1993]. Hopefully this paper provides a strong argument showing that conducting empirical research using reproducible methods is not only easy, but beneficial. And that were HONEST to be implemented to a wide degree, research need not be impeded[11].

[10]The act does not define "substantially reproducible" with any rigor. It is possible that the flexibility of the act could lead to climate change deniers (and others) to call almost any study not substantially reproducible and bar federal agencies from acting on good research.

[11]Ignoring the issue of data confidentiality.

DRAFT

4.1 Current Resources

[Very Rough]

There are currently some good sources for empirical researchers looking to set up an organized and scalable workflow. One of the best of these is provided by the Gentzkow & Shapiro lab (link, Matthew Gentzkow [2017]), it is meant to be an internal reference source for their researcher assistants, but also serves as an excellent guide for other researchers looking to ensure good internal programming practices (they also provide a more general guide here). It covers Computing environments, project management, version control (through GitHub), and coding principles. These are all indispensable tools from the researcher's point of view as they allow for smooth and standardized workflows that are easily interpreted and can be reused later with ease. The one omission from this guide is reproducible and open sourced research practices. This omission is not surprising because of the lack of a large reproducibility movement within the empirical social sciences. The next section outlines a couple basic suggestions for establishing a workflow that not only ensure good internal practices, but also includes elements that help with reproducibility/replication and open source access.

4.2 Basic Workflow

[Very Rough]

One of the most useful tools for empirical researchers is GitHub. GitHub allows for seamless version control and easy collaboration. Version control is extremely useful, allowing researchers to easily view their editing history and past analyses that may have been changed or deleted, this also provides a public ledger documenting the history of the research analysis which can be a useful source for reviewers interested in how a final analysis was settled on[12]. Version control also means that researchers can "checkout" to a new branch and explore alternative analyses without changing the original files. The collaborative aspect of GitHub allows multiple researchers can work on the same files in tandem and later merge the changes together. GitHub also provides an online hosting service, local repositories can be pushed to their public site with a one line command. This is the most effective way of creating an open access research project, by pushing to a public repository you theoretically allow

[12] An additional step that helps reviewers and readers is including annotated analyses (Jupyter Notebooks or Rmarkdown files are excellent for this) that were discarded or not included in the final report. This transparency can help ease suspicion of p-hacking or other unintentional mistakes with analysis or data handling.

DRAFT

anybody to clone your repository and run it locally on their own machine. In practice there are certain limitations that can cause problems with running a public repository on different local machine.

One of the most common issues is that the order of analysis is unclear. With large project there are often multiple scripts that clean data, then run different types of analyses, or create figures. If run out of order they will just return errors. This is where a Makefile comes in. The original researchers create a Makefile with different "phony" commands, and anybody that downloads the repository can then run these one line commands that are actually doing multiple things in a specific order. In this project we have multiple Makefile commands, for example, `make all` runs all of the python Jupyter Notebooks in the correct order and typesets this pdf from .tex and .bib files and `make clean` deletes all the intermediate data and figures. The Makefile allows someone to clone the repository from GitHub and then simply run on line in terminal `make all` to reproduce the results. Unfortunately this doesn't always work because of the different package versions on different machines, this is where environments come in.

If two machines run different versions of python, or even just different versions of specific packages (like scipy or matplotlib), then the same commands may produce different outputs, which betas the purpose of the Makefile. Anaconda facilitates creation of custom environments to solve this issue. In this project the environment is specified in the `environment.yml` file. It creates a local environment with only the packages that are used in the analysis and their specific versions at the time of the analysis. For example our environment uses python 3.6.4 and pandas 0.22.0. Even if there are updates to pandas in the future that renders the current scripts unusable, they will still produce identical results within this environment. By combining these tools the analysis can be identically reproduced even years after the fact.

For additional information, the notes from Fernando Pérez's course "Reproducible and Collaborative Data Science" (Fall 2017) are an excellent resource exploring many useful tools and practices (link, Pérez [2017]).

4.3 Custom Functions

[Very Rough]

Many papers in applied econometrics involve developing custom functions for novel analyses. In some cases the paper is an exploration of this new method of analysis, in which case the researchers usually build a public package with the functions to allow others to use them; a good example of this is `causalTree` package documented here (Athey and Imbens

DRAFT

[2016]). However this level of documentation and package creation is not necessarily worth the amount of time it takes when the custom analysis functions are not the main point of the research project. In this case it is still useful to use basic documentation tools and function testing, these steps allow the researcher to go back and easily reuse their previous functions, and makes them more understandable for others who might want to implement them. This is the approach taken in this project. We create multiple custom functions (both for data cleaning and for the unusual Q2SLS procedure) and they can hopefully serve as an example of how to create custom functions that can be used by others without too much hassle.

One of the easiest steps to take that helps both the researcher and future users is good documentation and commenting. Most programming languages have support for function docstrings[13] which should include a description of what the function does, the parameters it takes, the output(s), and possibly an example. This provides a future user with all the necessary information on how to implement the function. Additional inclusion of comments within the function helps explain what specific steps are doing, which is useful for understanding the mechanics and for further development or customization.

One the researcher has created a nicely formatted function script, the next step is to check that it runs without issues. This is where testing comes in. The most basic method of testing is to simply create a separate script that tests the function in a scenario where the researcher knows exactly what the output should be. This approach is useful for preliminary testing and for testing complicated functions (we use this method for testing the Q2SLS function in the function testing notebook - nbviewer, git). Once the function developer has a test case (or cases) that they know the function should always pass, they can incorporate continuous integration (CI). CI usually involves writing a testing script that tests the function and asserts that the output is as expected. A third party service will then check that the function passes the tests every time edits are pushed to the public git repo. We use Travis CI to check our custom functions; the testing script is `tests.py`, the environment the tests run in is `.travis.yml`, and the build status can be monitored through the badge at the top of the README or on Travis's platform here. CI is expected in the reproducibility community, and empirical research would greatly benefit from its adoption. It is an easy way of verifying that there are no tricky mistakes being made in any of the analysis steps within a larger research project, as well as continuously checking that all functions pass the predefined tests.

[13]For example, python supports docstrings formatted as : `'''function docstring'''`.

DRAFT

4.4 Optional/Additional Elements

[Cover optional but useful things that the open source community has begun using: Binder, Sphinx, etc.] and discuss things I didn't have time to implement

[FOR SURE: mention that I did not have time to set up a way of porting the coeff.s and SE's directly from the notebooks to the tex file, and that ideally this would be fully automatic, so if the results changed in the Jupyter Notebook, then they would change in the tex file and the pdf (although since we are careful with the environments and version control, the coeffs should never change unless the model specifications change)]

DRAFT

5 Conclusion (? - maybe not needed)

Maybe include a conclusion in the Q2SLS section, but not here.

DRAFT

6 Appendix

6.1 Data Exploration Figures

[ONLY ONE FIGURE INCLUDED FOR NOW, NEED TO ADD OTHER FIGURES AND SCALE NICELY]

6.2 Naive OLS Extra Models

[PUTTING ALL TABLES HERE FOR NOW BECAUSE HAVING TROUBLE PUTTING THEM IN SPECIFIC SECTION]

QUESTION: [Should we report these tables in the main paper and then report tables with all the covariates in the appendix? These would be extra long, and are viewable in the model fitting 1 notebook, so maybe it'd be better to just link to them?]

Table 1: Naive OLS, Discrete Mapping

	Dependent variable: G3 Grade (Percent)					
	Both	*Portuguese*	*Mathematics*	*Both*	*Portuguese*	*Mathematics*
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.5120***	0.4942***	0.4784***	0.4492***	0.3136***	0.8002***
	(0.0137)	(0.0143)	(0.0366)	(0.1106)	(0.1099)	(0.2827)
Studytime_2	0.0361***	0.0526***	0.0050	0.0213	0.0181	0.0264
	(0.0136)	(0.0143)	(0.0285)	(0.0142)	(0.0147)	(0.0382)
Studytime_3	0.0924***	0.1045***	0.0668*	0.0631***	0.0474***	0.0923**
	(0.0177)	(0.0166)	(0.0380)	(0.0182)	(0.0180)	(0.0424)
Studytime_4	0.0787***	0.0927***	0.0563	0.0384	0.0481*	0.0383
	(0.0285)	(0.0278)	(0.0575)	(0.0290)	(0.0264)	(0.0623)
School_GP	0.0742***	0.0856***	0.0283	0.0371**	0.0602***	−0.0267
	(0.0133)	(0.0142)	(0.0337)	(0.0150)	(0.0160)	(0.0445)
Course_math	−0.0954***			−0.0958***		
	(0.0128)			(0.0150)		
Proxies/Controls	*no*	*no*	*no*	*yes*	*yes*	*yes*
Observations	1,044	649	395	1,044	649	395
Df	5	4	4	68	67	67
R^2	0.094	0.131	0.015	0.328	0.410	0.352
Adjusted R^2	0.090	0.126	0.005	0.282	0.342	0.219
F Statistic	24.62***	22.95***	1.469	7.23***	5.91***	2.74***
Note:						*p<0.1; **p<0.05; ***p<0.01

DRAFT

Table 2: Naive OLS, Continuous Mapping

	Dependent variable: G3 Grade (Percent)					
	Both	*Portuguese*	*Mathematics*	*Both*	*Portuguese*	*Mathematics*
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.4879***	0.4686***	0.4572***	0.4378***	0.3049***	0.7802***
	(0.0161)	(0.0165)	(0.0402)	(0.1103)	(0.1091)	(0.2819)
Studytime_continuous	0.0220***	0.0268***	0.0128	0.0155***	0.0107*	0.0232*
	(0.0053)	(0.0052)	(0.0113)	(0.0055)	(0.0055)	(0.0126)
Studytime_continuous_sq	−0.0010***	−0.0013***	−0.0005	−0.0008**	−0.0004	−0.0012
	(0.0004)	(0.0004)	(0.0007)	(0.0004)	(0.0003)	(0.0008)
School_GP	0.0737***	0.0856***	0.0262	0.0371**	0.0603***	−0.0270
	(0.0133)	(0.0142)	(0.0335)	(0.0150)	(0.0160)	(0.0445)
Course_math	−0.0956***			−0.0959***		
	(0.0128)			(0.0150)		
Proxies/Controls	*no*	*no*	*no*	*yes*	*yes*	*yes*
Observations	1,044	649	395	1,044	649	395
Df	5	4	4	68	67	67
R^2	0.09	0.13	0.01	0.33	0.41	0.34
Adjusted R^2	0.09	0.13	0.00	0.28	0.34	0.22
F Statistic	30.11***	30.61***	1.41	7.29***	6.03***	2.97***
Note:						*p<0.1; **p<0.05; ***p<0.01

6.3 Q2SLS Function Testing Procedure

[Can use some of the markdown from the functions testing notebook here, and obv.s link to the notebook as well]

6.4 Q2SLS Extra Models

[PUTTING ALL TABLE HERE FOR NOW BECAUSE HAVING TROUBLE PUTTING THEM IN SPECIFIC SECTION] Note: this Q2SLS table does not have the correct coefficient values, it is just here as a placeholder and to set up the formatting.

DRAFT

Table 3: Addressing Endogeneity - WRONG VALUES (JUST PLACEHOLDER FOR NOW)

	Dependent variable: G3 Grade (Percent)					
	Both	*Portuguese*	*Mathematics*	*Both*	*Portuguese*	*Mathematics*
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.4879***	0.4686***	0.4572***	0.4378***	0.3049***	0.7802***
	(0.0161)	(0.0165)	(0.0402)	(0.1103)	(0.1091)	(0.2819)
Studytime_continuous	0.0220***	0.0268***	0.0128	0.0155***	0.0107*	0.0232*
	(0.0053)	(0.0052)	(0.0113)	(0.0055)	(0.0055)	(0.0126)
Studytime_continuous_sq	−0.0010***	−0.0013***	−0.0005	−0.0008**	−0.0004	−0.0012
	(0.0004)	(0.0004)	(0.0007)	(0.0004)	(0.0003)	(0.0008)
School_GP	0.0737***	0.0856***	0.0262	0.0371**	0.0603***	−0.0270
	(0.0133)	(0.0142)	(0.0335)	(0.0150)	(0.0160)	(0.0445)
Course_math	−0.0956***			−0.0959***		
	(0.0128)			(0.0150)		
Model	*2SLS*	*2SLS*	*2SLS*	*Q2SLS*	*Q2SLS*	*Q2SLS*
Observations	1,044	649	395	1,044	649	395
Df	5	4	4	68	67	67
R^2	0.09	0.13	0.01	0.33	0.41	0.34
Adjusted R^2	0.09	0.13	0.00	0.28	0.34	0.22
F Statistic	30.11***	30.61***	1.41	7.29***	6.03***	2.97***

Note: SE's for Q2SLS are bootstrapped. *p<0.1; **p<0.05; ***p<0.01

6.5 Miscellanea (?)

DRAFT

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. [1996], 'Identification of causal effects using instrumental variables', *Journal of the American Statistical Association* **91**, 444–455.

Angrist, J. D. and Pischke, J.-S. [2009], *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Athey, S. and Imbens, G. [2016], 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences* **113**(27), 7353–7360.

Cameron, A. C. [2017], 'Machine learning for microeconometrics'.
URL: *http://cameron.econ.ucdavis.edu/e240f/trmachinelearningseminar.pdf*

Card, D. and Krueger, A. B. [1992], 'Does school quality matter? returns to education and the characteristics of public schools in the united states', *Journal of Political Economy* **100**(1), 1–40.

Cortez, P. and Silva, A. [2008], 'Using data mining to predict secondary school student performance', *A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)* pp. 5–12.
URL: *http://www3.dsi.uminho.pt/pcortez/student.pdf*

Feigenbaum, S. and Levy, D. M. [1993], 'The market for (ir)reproducible econometrics', *Social Epistemology* **7**(3), 215–232.

Greene, W. H. [2011], *Econometric Analysis*, 7 edn, Prentice Hall.

MacKinnon, J. G. and White, H. [1985], 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics* **29**(3), 305–325.

Mansfield, E. R. and Helms, B. P. [1982], 'Detecting multicollinearity', *The American Statistician* **36**(3a), 158–160.

Matthew Gentzkow, Jesse Shapiro, A. R. [2017], 'Gslab ra manual'.
URL: *https://github.com/gslab-econ/ra-manual/wiki/Getting-Started*

O'brien, R. M. [2007], 'A caution regarding rules of thumb for variance inflation factors', *Quality & Quantity* **41**, 673–690.

Pérez, F. [2017], 'Reproducible and collaborative data science'.
URL: *https://berkeley-stat159-f17.github.io/stat159-f17/*

Schulson, M. [2018], 'Science's "reproducibility crisis" is being used as political ammunition'.
URL: *https://www.wired.com/story/sciences-reproducibility-crisis-is-being-used-as-political-ammunition/*

DRAFT

Tibshirani, R. [1996], 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Wooldridge, J. M. [2001], *Econometric Analysis of Cross Section and Panel Data*, MIT Press Books, 2 edn, The MIT Press.

DRAFT