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

Nadav Tadelis

May 2018

Abstract

In this paper I establish a framework for reproducible empirical research. I 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 of the causal impact of studying on grades. 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. An instrumental variables approach is presented 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 the 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. I will point out some of the weak points in my models and assumptions that can be improved upon with more time and effort. Any comments or suggestions would be greatly appreciated and can be directed to 'ntadelis [at] berkeley [dot] edu'.

The basic 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 changes is fully documented in the commit history of the git repo.

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.

2.1 Data Exploration

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

I 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 the 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.

For a more detailed description of the data cleaning and exploration process, see the data exploration notebook (linked to above). It explains all the steps taken in the cleaning phase,

figures showing distributions across different variable splits, and a discussion of the mapping schemes for the studying time variable.

2.2 Data Issues

As is often the case with education data, the data have some limitations that must be considered. First, much of the data are stated preferences (as opposed to observed). Fortunately, the test scores and number of absences are provided by the schools, so I 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 I am 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 is information 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. The issue of self reported data becomes more problematic when it comes to the variable of interest - hours of studying per week. While the central research goal is identifying a relationship between hours of studying and academic success, the data on hours of studying cannot be trusted as accurate. One approach could be to interpret this research as an analysis on the returns to perceived studying time rather than actual studying time. In the rest of the paper I choose to assume that the students are accurately reporting their average weekly studying times for the sake of clarity and simplicity.

Second, the 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, I

explore two different re-mapping schemes[1]. But for the other variables with this format I treat them as categorical variables and include indicators for each level (thus allowing for some nonparametric specifcation).

Lastly, the data are cross-sectional rather than longitudinal, and the specific sampling time is unspecified. This is especially troubling in this setting because the original paper provides 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]. In particular, if students respond to low grades by studying more in the next time period, and respond to high grades by studying less, then this creates a "mean reversion" endogeneity problem that would work against identifying the effect of studying on grades, that can only be corrected for by using a richer panel data structure.

[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

3 Models

First, I set up the 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 the model is:

$$g_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \boldsymbol{x}_{i,3:k-1} \cdot \boldsymbol{\beta}_{3:k-1} + \beta_k a_i + \varepsilon_i$$

Where $\boldsymbol{x}_{i,3:k-1}$ is a vector of school and course level characteristics and ε_i captures some unobserved heterogeneity and disturbances. I assume that grades are nonlinear in weekly study time. While this nonlinearity complicates the model, it seems necessary because marginal returns to studying must vary depending on the initial level of studying.

Clearly, there are issues with this model. How am I 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, I 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 am I defining ability? By its very nature, ability is unobserved. I will partially proxy for ability using other individual level characteristics (IQ scores, age, parents' education, etc.). But I cannot fully capture ability because it includes many skills and traits, some of which are difficult to quantify in a meaningful way. Hence, the model is never going to be fully specified regardless of what variables are observed.

I think of a student's utility maximization problem as being some function of grades, free time (consider everything that is not studying, sleeping, or class as "free time"), and studying time. I would expect that the impact of grades and free time on utility would be positive, and the impact of study time would be negative, with magnitudes of 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 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 the estimate of the marginal effect of studying, and would be collinear with the amount of weekly studying. However, when observations are thought of 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. The intuition here is that the model estimation must try to account for the unobserved decision making process, which is assumed to include ability. So it must proxy for ability and include any other variables that might be included in the unobserved decision making process.

Note that the decision function must be nonlinear. There has to be some concavity, otherwise the optimality solution is a corner solution and students will either spend no time studying, or all their time studying. Since I am not estimating the actual decision function, these nonlinearities are not included in the estimation, but they are an implicit assumption that should be pointed out.

3.1 Naive OLS

In this section I discuss results from the naive model specifications estimated in the model fitting 1 notebook - nbviewer, git. This analysis explores simple models based on the structural

[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. I simplify the specification here dramatically.

equation defined in section 3. I include results for both discrete (Table 1) and continuous (Table 2) mappings of studying time, with the continuous mapping including a squared term. For both mappings I run six OLS regressions. First the data are split into three samples: one including the full dataset with all students, one with only students studying the Portuguese Language Course, and one with only students studying the Mathematics Course. For each of these groups I run two regressions, one with only school and course level characteristics, and one including proxies for ability. Note that both of these models include an extra indicator for the sample that includes students from both classes: Course_math is a binary variable indicating whether the student is in the Mathematics Course. School_GP is an indicator for which of the two schools the student is enrolled in. These two variables make up the school and course level controls[4] ($\boldsymbol{x}_{i,3:k-1}$ in the structural equation). For more information on which variables are included as proxies, see the model fitting 1 notebook linked to above. The regression results are reported in the tables below.

Table 1: Naive OLS, Discrete Mapping

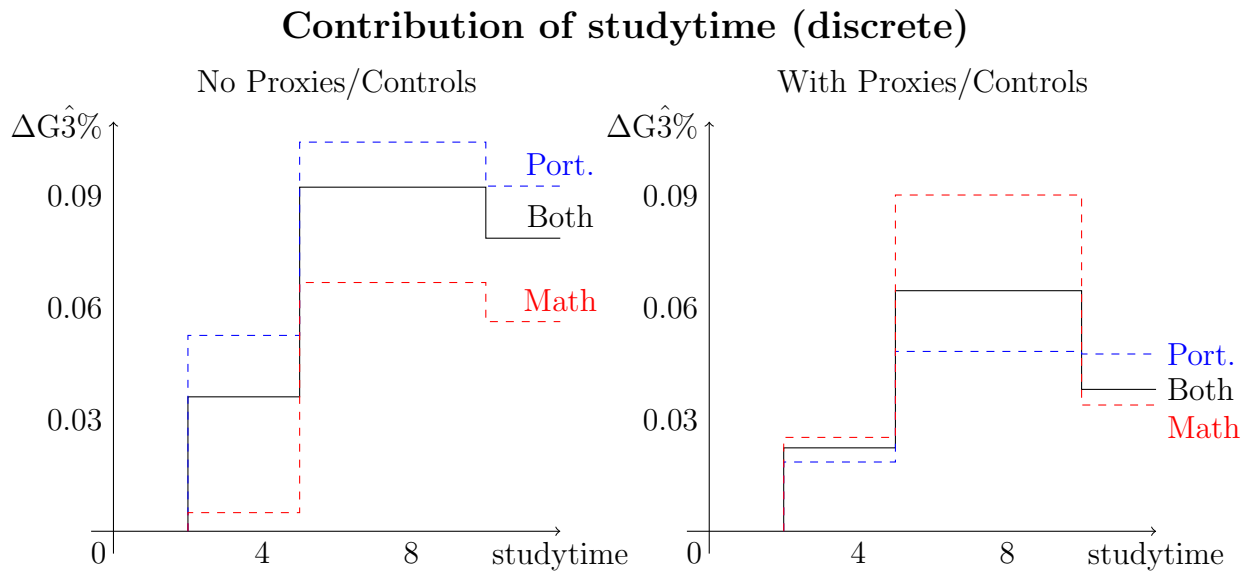
	Dependent variable: G3 Grade (Percent)					
	Both	*Portuguese*	*Mathematics*	*Both*	*Portuguese*	*Mathematics*
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.5120***	0.4942***	0.4784***	0.4492***	0.3136***	0.8002***
	(0.0137)	(0.0143)	(0.0366)	(0.1106)	(0.1099)	(0.2827)
Studytime $\in [2,5)$	0.0361***	0.0526***	0.0050	0.0213	0.0181	0.0264
	(0.0136)	(0.0143)	(0.0285)	(0.0142)	(0.0147)	(0.0382)
Studytime $\in [5,10)$	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 $\in [10,20)$	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	0.09	0.13	0.02	0.33	0.410	0.35
Adjusted R^2	0.09	0.13	0.01	0.28	0.34	0.22
F Statistic	24.62***	22.95***	1.47	7.23***	5.91***	2.74***
Note:						*p<0.1; **p<0.05; ***p<0.01

[4]Ideally the models would include interaction effects of time and these dummies, but the data are cross-sectional.

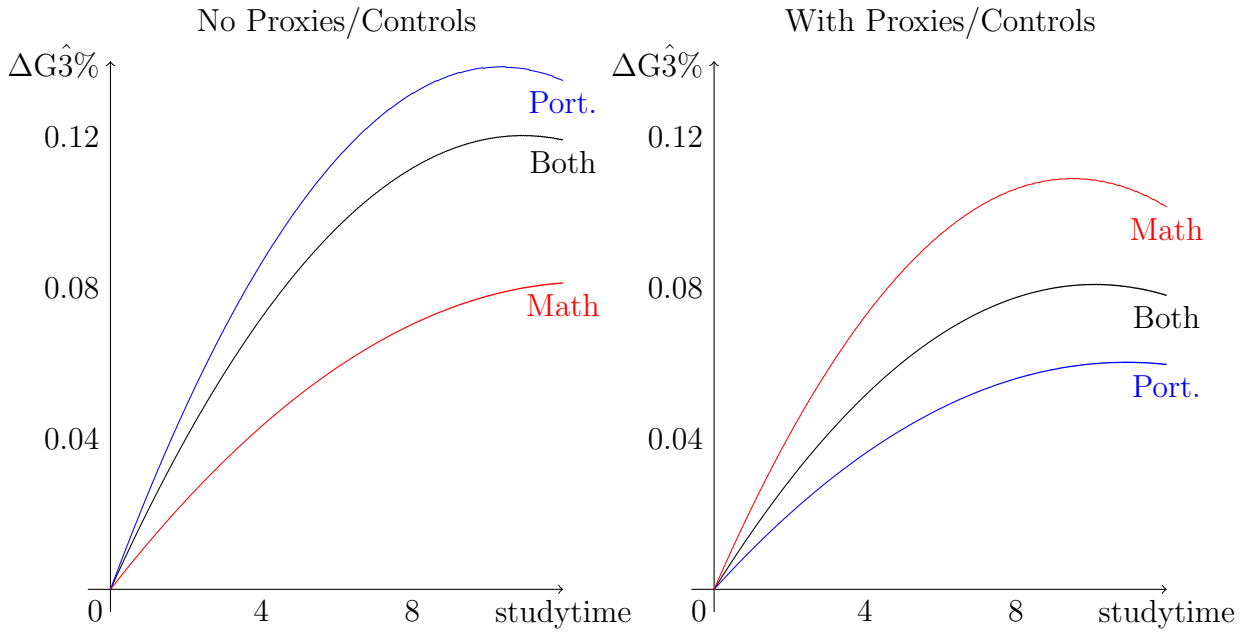
Table 2: Naive OLS, Continuous Mapping

	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

In all twelve models the estimated coefficients on weekly study time report decreasing marginal returns. Studying time's estimated contribution to the final score (in percentage points) is plotted below (these plots are simply visual representations of the coefficients on study time in each of the models).



Contribution of studytime (continuous)



These plots allow clearer visual representation of the estimated relationship between studying time and final grade in the naive OLS models. As expected the behavior of the continuous and discrete estimates are similar. The decrease in size of the estimated relationship between studying and grades for the Portuguese Language students suggests that students with higher ability are studying more, and getting better grades, and that once ability is controlled for the predicted returns to studying are smaller. I would like to emphasize that this discussion does not describe any causal relationships, these models are purely predictive. Interestingly the estimated relationship between studying and grades in the Mathematics students is flipped; this seems to suggests that in mathematics, students with higher ability study less to achieve the same grade (i.e. that studying is less important than ability). However, these intuitive conclusions may be invalid for two reasons. First, studying time is endogenous in this set up, and second, there is no information as to the direction of the bias term on the endogenous variables (see section 3.1.3 for more detail).

The next section describes the results from some additional analyses related to these naive regressions.

3.1.1 Additional Analysis

In the model fitting 1 notebook I also include some additional explorations of the estimated models. Specifically these analyses use the sample including both courses of study. I compute the variance inflation factor to check for multicollinearity (O'brien [2007], Mansfield and Helms [1982]). I find slight collinearity between studying time and its square, but this collinearity is not too concerning because it is on the variable of interest, and the VIF is low enough to ignore. The two variables have VIF's of 13 and 12, which are right on the threshold of what could be considered an issue for multicollinearity. But there is an expectation for some correlation between the two variables (they are directly dependent on one another) and as they are the variables of interest, PCA or another dimension reduction method are not usable because the coefficients on these variables need to be interpretable. The rest of the collinearity that the VIF detects is so low that it can be ignored.

I also include a brief exploration of a LASSO-flavored penalization for variable selection (Cameron [2017]) . If a penalization parameter that shrinks all but seven of the coefficients to zero is selected, then 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. I 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 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[5]. 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 perhaps their studying decisions did not 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 only the single survey results with no time stamp are included in the data, determining this relationship is intractable. With this limited information I 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 I add many covariates as proxies representing unobserved ability (parental education, individual characteristics, and others detailed in the model fitting 1 notebook - nbviewer, git). Section 3 pointed out that since ability is unobservable and undefinable it cannot be expected to be fully capture through the proxies. Let q_i be the unobserved portion of ability that is not captured through the imperfect proxies. Then:

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

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. The main coefficients of interest are β_1 and β_2 the coefficients on studying time. It is reasonable to assume that $\mathrm{E}[\nu_i | \boldsymbol{x}_i, q_i] = 0$ unfortunately there is no choice but to

[5]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.

stick q_i into the error term which gives 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 $\varepsilon_i = \gamma q_i + \nu_i$. Clearly 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. The next step is to look at the consistency of this set up (starting with the linear projection of q_i onto $\boldsymbol{x}_i$):

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

Where (1) and (2) follow from the definition of linear projections. Now the original model can be represented 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 OLS will achieve consistent estimates of $(\beta_j + \gamma\delta_j)$, but the object of interest is actually $\boldsymbol{\beta}$. OLS does not allow recovery of $\boldsymbol{\beta}$ in this case because:

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

If $\gamma \neq 0$ and $\delta_j \neq 0$ then getting a consistent estimate of β_j using OLS is impossible. Rather, OLS returns an endogenous estimator with a bias of $\gamma\delta_j$ (notice that the better the prox-

ies 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 the studying and grades setting this is actually a very reasonable assumption; $x_{i,5}, ..., x_{i,k}$ have already been defined 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[6] the endogeneity problem so that the only endogenous variables are s_i and s_i^2 ($\text{Cov}[s_i, \varepsilon_i],\ \text{Cov}[s_i^2, \varepsilon_i] \neq 0$). The next sections first give a brief overview of the instrumental variables approach to addressing endogeneity, and then explain why standard two stage least squares is intractable in the grades and studying setting. I then explore an instrumentation approach with nested generated regressors that may solve the problem.

3.2 Q2SLS

First, a brief refresher of the instrumental variables approach to addressing endogeneity. The last section showed that for an endogenous variable $x_{i,j}$ OLS is no longer a consistent estimator for β_j. The instrumental variables approach provides a solution to the endogeneity problem. It requires 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 the instrument is nonzero: $\theta_1 \neq 0$
Note that this is equivalent to saying: $\text{Cov}[x_{i,j}, z_i] \neq 0$

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

If (1) – (3) hold, then z_i is called 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. The first stage of

[6] If there were only one endogenous variable this could be simplified further because $\hat{\beta}_j \xrightarrow{p} \beta_j + \gamma \frac{\text{Cov}[x_j, q]}{\text{Var}[x_j]}$. In this case guesses could be made about the direction of the bias.

2SLS regresses the endogenous variable $x_{i,j}$ on the exogenous variables and the instruments to get the fitted values $\hat{x}_{i,j} = \sum_{k \neq j}\{\hat{\phi}_k x_{i,k}\} + \hat{\theta}_1 z_{i,1} + \cdots + \hat{\theta}_M z_{i,M}$. Then, $\hat{x}_{i,j}$ is used as an estimate of the exogenous part of $x_{i,j}$ and the second stage regresses the dependent variable on the 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 there are as many first stage regressions as there are endogenous variables). However, with multiple endogenous variables, there must be at least as many instruments as endogenous variables, otherwise the fitted values would be collinear and the second stage estimators would lose consistency. The next section discusses issues with the 2SLS method in the grades and studying setting.

3.2.1 Motivation

Section 3.1.3 assumed that the causal relationship between grades and studying time is quadratic. This would usually not be a barrier to using 2SLS; the standard approach[7] 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 there are 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 (when compared to squaring and interacting continuous variables). Happily, Wooldridge

[7] Assuming only one instrument is available.

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[8] $x_{i,j}$ on the exogenous variables and the instruments to get fitted values $\hat{x}_{i,j}$

   (b) Regress the endogenous variable squared[9] $x_{i,j}^2$ on the exogenous variables, the instruments, and the squared fitted values $(\hat{x}_{i,j})^2$. This gives fitted values for $\hat{x^2_{i,j}}$

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

The intuition behind this model is that while using $(\hat{x}_{i,j})^2$ will not produce consistent coefficient estimates[10] 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 $x_{i,j}^2$. This should allow identification of $\boldsymbol{\beta}$ even with the quadratic endogeneity and binary instrumentation; in fact it is possible to achieve identification with a single binary instrument.

3.2.3 Properties of Q2SLS

Due to the unusual nested regressor set up of the Q2SLS procedure it can be difficult to obtain intuition for the asymptotic behavior of the model. In order to explore some of those asymptotic properties I use a Monte Carlo simulation with a toy model. I run the simulation in the function testing notebook, which simultaneously tests the custom function `Quadratic2SLS()` from `quadratic2SLS.py` - nbviewer, git. For a more detailed discussion of the simulation set up and results see section 6.2.

[8] In the previously defined model specification this is $s_i = x_{i,2}$

[9] In the previously defined model specification this is $s_i^2 = x_{i,3}$

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

The Monte Carlo simulation involved repeatedly drawing from a toy DGP with artificially introduced endogeneity. I drew from two separate DGP's, one with strong instruments, and one with weak instruments[11]. Note that in the Q2SLS setting, the strength of instruments is a bit tricky to think about; in part (1.a) of the Q2SLS procedure instrument strength is determined as usual by looking at how strongly the instruments are correlated with the endogenous variable, in part (1.b) of the procedure there is an additional generated instrument, the strength of which is determined by how well the first instruments identify the endogenous variable. In essence this means that if the instruments are strong in part (1.a) then the added generated instrument will also be a strong in part (1.b). The main results from the simulation are reported in Figure 1.

[11]In standard 2SLS the "strength" of instruments measures how much of the variation in the endogenous variable the instruments capture. This can be measured by looking at the correlation of the instruments and the endogenous variable, or by constructing a first stage F statistic (Angrist and Pischke [2009])

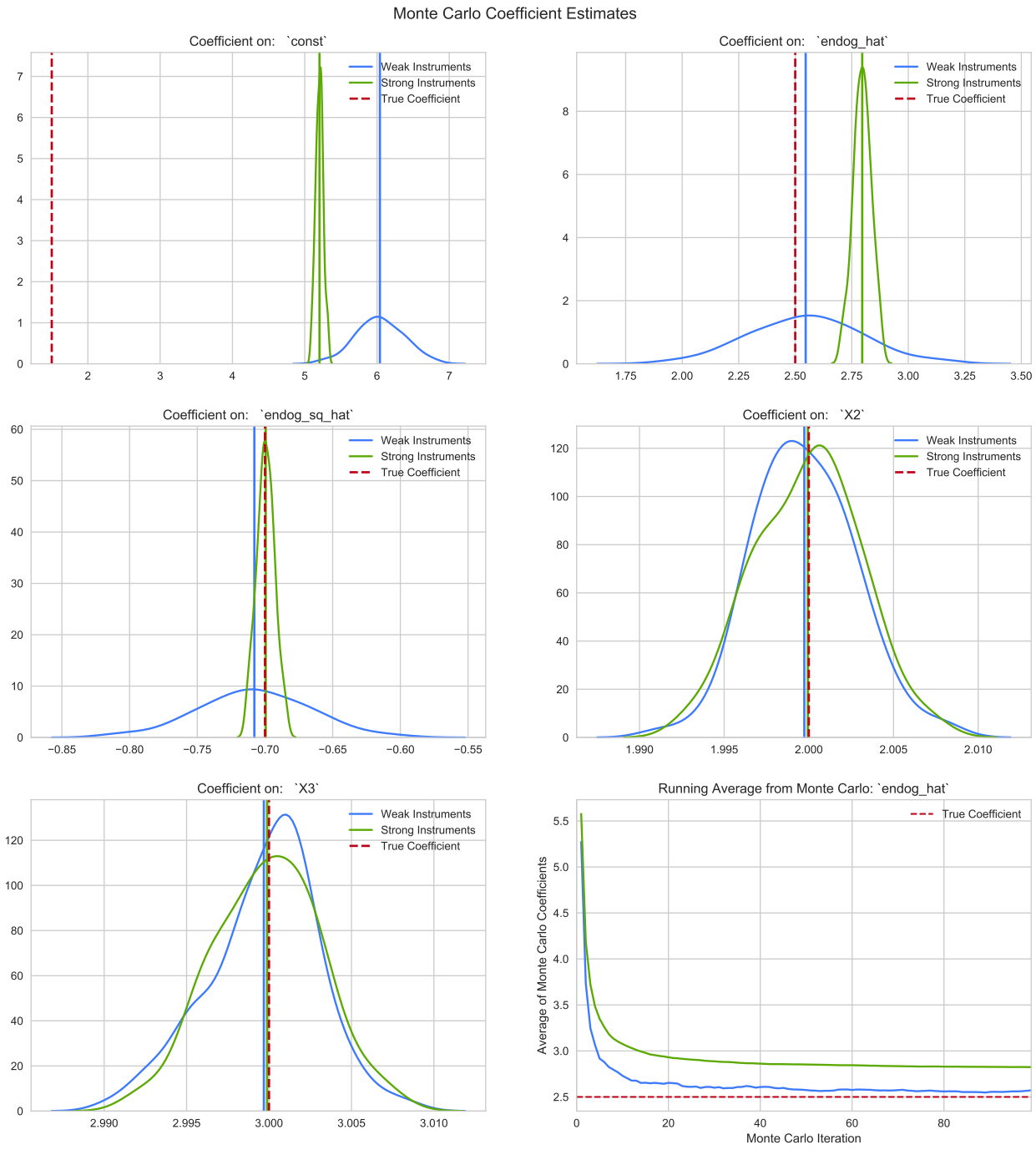


Figure 1: Monte Carlo simulation coefficient estimate distributions

The distribution of the estimated coefficients on the exogenous variables X2 and X3 are unbiased and consistent for both strong and weak instruments (as expected). It also looks as though the estimate on 'endog_sq_hat' (in section 3.2.2 I call this $\hat{x}^2_{i,j}$) is unbiased and

consistent, although with weak instruments the variance is much larger[12]. The truly baffling behavior is the estimated coefficient on 'endog_hat' (in section 3.2.2 I call this $\hat{x}_{i,j}$). With strong instruments it achieves consistency, but severe bias - the true coefficient is not even in the support of the estimated coefficients. And yet with weak instruments it appears to achieve unbiasedness and consistency (albeit a much larger variance again). At first it may seem as though there is a funky collinearity problem between the two generated regressors, but in both the strong and weak instruments the Pearson correlation coefficient is around 0.98 (which leads to the question of why there isn't a collinearity problem in the weak instruments approach). While this suggests there is collinearity present, it is unreasonable to think that this collinearity is causing the consistent bias because it is present at equal levels in both the strong and weak instrument settings[13]. Due to time constraints I was unable to find a satisfactory explanation of this behavior, however I plan to continue researching possible sources of the bias under strong instrumentation and will update this document with the progress.

An additional issue with this nested generated regressor approach is that computing the variance of the coefficient estimates is nontrivial. Due to submission deadlines, there was not enough time to determine whether a consistent estimator for variance of the coefficient estimates exists. I resort to bootstrapping for variance. Currently the function uses the "pairs bootstrap", also called "nonparametric bootstrapping". This involves drawing a sample of the same size (as the original data) from the original data B times, and using the results to estimate the distribution of the estimates. Usually one would report the average bootstrapped estimate as the final estimate (this is standard in prediction settings), but in this causal setting I report the coefficient from running the model on the full sample, and estimate the coefficient variances using the variance from the bootstrapping. These approaches should yield the same result (we would expect that when bootstrapping with regressions

[12]This is unsurprising - the weaker instruments have more noise introduced into the first stage, which would increase the variance of the coefficient estimates on the instrumented regressors.

[13]Additionally, if the only issue were multicollinearity then it is unlikely that the bias would be consistent and isolated to only of the instrumented regressors under strong instruments

the average estimates would approach the full sample estimates as B $\longrightarrow \infty$). The standard testing procedure[14] is then used to test for significance.

One direction of further development is adding different bootstrapping schemes. I would like to build functionality for a Wild Restricted Efficient Bootstrap approach to coefficient variance estimation. The WRE bootstrap preforms well under endogeneity, especially in the case of weak instruments (MacKinnon [2012]).

3.2.4 Implementation

In order to implement the Q2SLS procedure, the first step is to find valid instruments for the endogenous variable. One instrument that I am very confident in is an artificial binary instrument z_i which is a plausible exogenous shifter of studying time. The artificial instrument combines two categorical variables such that $z_i \equiv$ [student i chose school based on distance *and* student i's commute is longer than 15 minutes]. Intuitively, this instrument is indicating whether a student has chosen their school based on its proximity to their home (as opposed to its reputation, the courses it offers, or "other"), and this student has a commute time that is long enough to plausibly detract from the time available for studying[15]. I believe that this instrument is fairly well justified. The other instruments are less convincing; I include different levels of "going out" measuring how often a student spends time with friends. These instruments are less compelling because it is plausible that students who spend more time with their friends could be labelled as more "social" and might put more weight on their social life and less weight on their grades in their time allocation decisions. The other issue with using going out as an instrument is that if students that go out also

[14]Note that by using t-test statistics to test for significance, I implicitly assume that the distribution of $\boldsymbol{\beta}$ is approximately multivariate normal. The results from the Monte Carlo simulation suggest that this is not an unreasonable assumption when the data are nice. If I did not assume normality, I could use confidence intervals from the bootstrapped distribution to test for significance, unfortunately there was not enough time to implement this into the estimation procedure.

[15]The reason I do not simply use commute time as an instrument is that it is possible that students choose further schools because of their reputation or course offerings. This would invalidate commute time as an instrument because the longer commute would also indicate higher motivation/ability, causing commute time to be endogenous.

drink more and get less sleep then their testing score might be lower regardless of how much they study. Fortunately I can control for students daily and weekly alcohol consumption, so the detrimental effects of alcohol on grades can be ignored. I can also assume that students are getting enough sleep because at their ages it is plausible that their parents are setting and enforcing their bed times. Once these assumptions are made, the only further assumption for including going out as an instrument is sociability does not affect grades independently of studying time or alcohol consumption. This is a strong assumption, but one that I feel reasonably comfortable making. One note to keep in mind is that these instruments are binary and weak. In the Monte Carlo simulation (see section 3.2.3) there was some evidence suggesting that weak instruments yield unbiased estimates of $\boldsymbol{\beta}$ but that they have very high variance, which means that any result should be taken with skepticism. The results from a standard 2SLS and Q2SLS are reported in Table 3 and visual representations are included below.

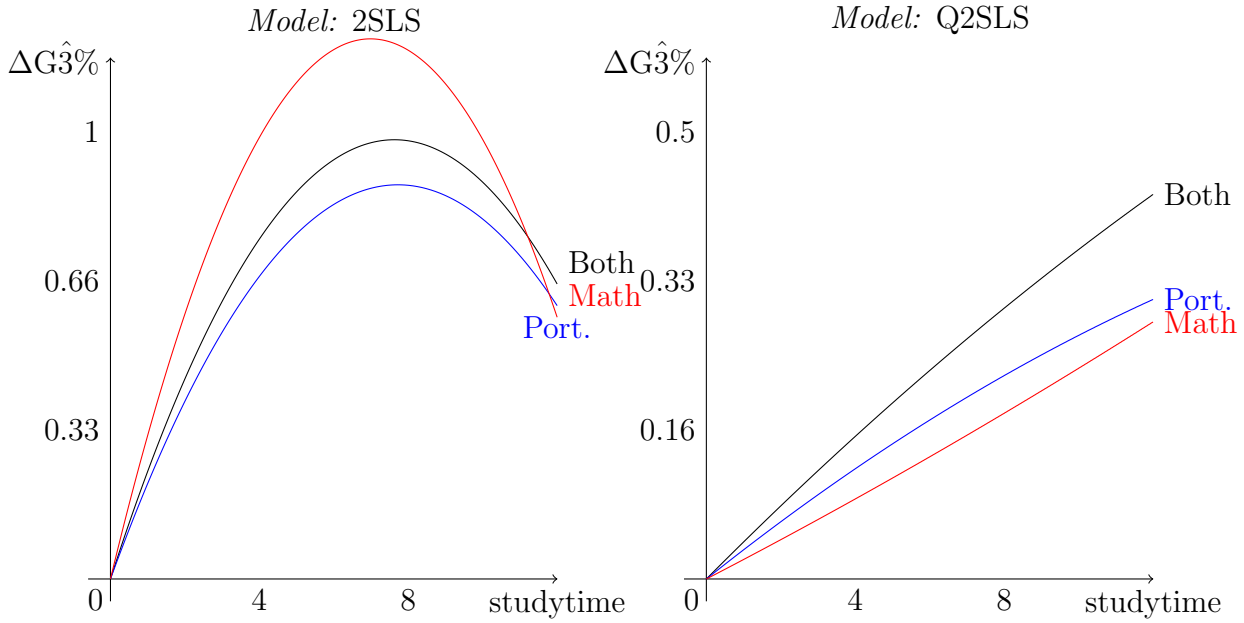
Table 3: Addressing Endogeneity

	Dependent variable: G3 Grade (Percent)					
	Model: 2SLS			*Model:* Q2SLS		
	Both	*Portuguese*	*Mathematics*	*Both*	*Portuguese*	*Mathematics*
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.6097***	0.5062**	0.6186	0.4984***	0.3490***	0.9174***
	(0.2122)	(0.2448)	(0.5787)	(0.1304)	(0.1227)	(0.3161)
Studytime	0.2578**	0.2285*	0.3462	0.0419*	0.0332	0.0216
	(0.1130)	(0.1346)	(0.2366)	(0.0235)	(0.0220)	(0.0340)
Studytime^2	−0.0169**	−0.0148	−0.0248	−0.0005	−0.0006	0.0002
	(0.0085)	(0.0096)	(0.0182)	(0.0006)	(0.0006)	(0.0008)
School_GP	0.0096	0.0214	0.0479	0.0132	0.0462*	−0.0487
	(0.0332)	(0.0368)	(0.1124)	(0.0265)	(0.0241)	(0.0601)
Course_math	−0.1046***			−0.0976***		
	(0.0233)			(0.0172)		
Observations	1,044	649	395	1,044	649	395
Df	56	55	55	56	55	55
R^2	−1.27	-1.39	−2.35			
Adjusted R^2	−1.4	-1.61	−2.88			
F Statistic	177***	130***	58			

Note: *p<0.1; **p<0.05; ***p<0.01

Studytime and Studytime^2 are instrumented for in both models.

SE's for Q2SLS are bootstrapped. R^2 and F stats have not been computed for Q2SLS.



Interestingly the coefficient estimates for the 2SLS and Q2SLS approaches are very different. The standard 2SLS approach estimates that the causal effect of studying on grades (across both courses) is $+0.258s - 0.017s^2$ which would mean that a student studying 7.6 hours per week would maximize their contribution of studying to grade at 98 points (out of 100). The Q2SLS approach estimates this same causal effect as $+0.042s - 0.0005s^2$ which would estimate that studying 7.5 hours increases your grade by 28 points (compared to not studying), and that the maximum contribution would be 88 points achieved by studying 42 hours. This is a fantastic result demonstrating that Q2SLS yields much more believable estimates than standard 2SLS. It is wholly unreasonable to think that all students could achieve a near perfect final score by studying 65 minutes per day. However it seems plausible that studying 65 minutes per day could increase a student's final score by 28 points. While the maximum reached in Q2SLS seems a bit extreme (I would not expect increasing returns to studying all the way up to 6 hours a day), since the range of observed studying times goes up to a maximum of around 10-20, the out of range values can largely be ignored.

I would not suggest that the Q2SLS estimates are truly correct because the instruments are not overly convincing, I still do not fully understand the asymptotic properties of the

Q2SLS procedure, and the samples are fairly small. However, these results clearly demonstrate that there are situations in which Q2SLS will yield far more plausible identification of endogenous variable effects than 2SLS. A further direction of study I intend to pursue is exploring Monte Carlo simulations with different DGPs and evaluating how 2SLS and Q2SLS estimates compare.

4 Reproducibility

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[16], 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

[16]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.

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 wider degree, research need not be impeded[17].

4.1 Current Resources

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 research 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 sections outline 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

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

[17] Ignoring the issue of data confidentiality.

which can be a useful source for reviewers interested in how a final analysis was settled on[18]. 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 to 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 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 these are run out of order they will just return errors. This is where a Makefile comes in. The original researchers create a Makefile with different so called "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. This project has 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 defeats the purpose of the Makefile. Anaconda facilitates creation of custom environ-

[18] 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.

ments 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 this project's 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 Professor 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

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 [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. I 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[19] 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.

Once 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 (I 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. I use Travis CI to check the 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. These precautions take little time, and provide assurances that no careless errors are being made when implementing in-house developed functions.

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

4.4 Optional/Additional Elements for Reproducibility

There are some tools that the open source community has begun using which are useful for empirical reproducibility, but not imperative. Here I cover only two tools from a massive pool of resources. Binder allows anyone to view and edit Jupyter Notebooks in an executable environment on a remote server through their browser. This means that even if a reader/reviewer does not want to clone the entire repository to their local machine, they can still run and edit the Jupyter Notebooks used in the analysis by using Binder. Researchers can link to their Binder pages by simply including a badge in their repository's README, or they can link to it directly.

Sphinx is also quite useful, although it has not been used in this project due to time constraints. Sphinx allows a researcher to render their public git repository as a webpage without too much hassle. While this does not sound all that useful, it is tremendously helpful for readers unfamiliar with git repositories who want to navigate a complicated project with many subdirectories.

There are many other useful tools that have recently been gaining traction. For example there are many third party apps that will automatically check a repository to see what portion of its functions have CI testing implemented, which can be useful for making sure that all functions are tested in large projects with many researchers. There are new tools like these coming out weekly, and keeping somewhat in the loop can help researchers find which tools work best with their workflow.

I would like to point out that even in research that is meant to stand as an example of a reproducible work, there may be compromises. For example, this project involved creating LaTeXtables to report the regression results from the analysis in the Jupyter Notebooks. Unfortunately there are no simple ways to port regression results directly from Python[20] into a formatted LaTeXtable. Instead I resorted to copy and pasting .tex code from Jupyter Notebooks, or filling in the tables manually in the .tex file. While this should not be a

[20] R has an excellent package, stargazer, which can be used in this fashion.

problem for exact reproduction (none of the values change), it would cause issues if the regression outputs were changed but the .tex file was not edited accordingly. These types of lapses in reproducibility are sometimes difficult to overcome, and in cases such as those it is important to point out any possible issues that they might cause.

5 Conclusion

This paper aims to provide an example of an interesting econometric study researched in a reproducible and responsible manner. I believe that through the discussion and implementation of various reproducible elements, it has succeeded in this goal. However, there are always new contributions to the reproducibility community, as well as techniques which I may not be familiar with. If any readers have suggestions to improve this paper's reproducibility, I would be excited to hear and apply them.

The econometric analysis in this paper demonstrates an unusual two stage approach to identification of non linear endogenous variables through instrumentation. I believe that the results clearly show that the Q2SLS procedure yields more plausible estimates in certain settings where 2SLS might be more commonly used. I plan to continue developing this research in the future, and welcome any contributions. If you would like to contribute, please issue a pull request to my git repo. I intend to add regression summary analysis to the Q2SLS script (R^2, F-statistic, tests for over identification in the first stages, a better object class for the regression results, and more). I am also planning on continuing the exploration of Q2SLS and its asymptotic properties. I will approach this using both theory and simulations, and will update this paper and repository as progress is made. Lastly, I plan to clarify the situations in which a researcher should use Q2SLS rather than 2SLS; this will most likely involve simulating different DGPs and comparing the performance of the two models, and later finding the theoretical motivation.

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 (slides)'.
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. [2012], 'Inference based on the wild bootstrap (slides)'.
URL: *https://www.math.kth.se/matstat/gru/sf2930/papers/wild.bootstrap.pdf*

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/*

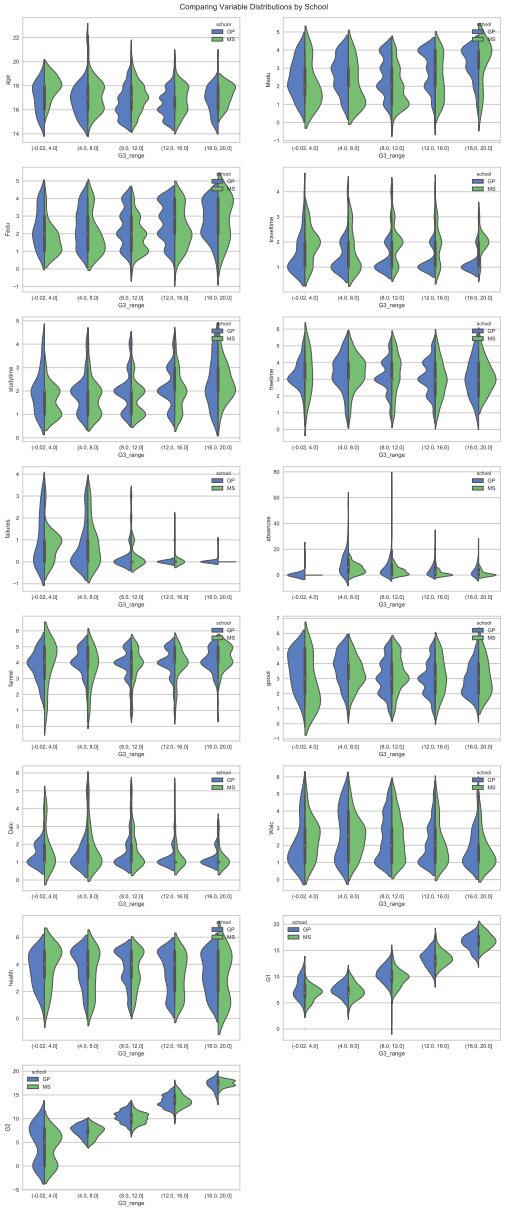
Tibshirani, R. [1996], 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

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

6 Appendix

6.1 Data Exploration Figures

[ONLY ONE FIGURE INCLUDED FOR NOW, NEED TO ADD OTHER FIGURES AND SCALE NICELY] - or would this make the doc too long? it would add at least 10 pages, i could just link to the /figures directory



6.2 Q2SLS Function Testing Procedure and Monte Carlo

This section explains in greater detail[21] how the `Quadratic2SLS()` function from `quadratic2SLS.py` was tested/explored in the function testing notebook - nbviewer, git. The notebook served two purposes, the first goal was to make sure that the function worked as expected with data from a toy data generating process (DGP) simulated through Monte Carlo, and the second was to analyze the asymptotic properties of the estimation procedure in this toy model. The first step was to set up a DGP that was simple, and fit the use case for Q2SLS; i.e. the DGP is quadratic in the endogenous variable.

I first present some motivation for the DGP: I model a toy causal relationship between score on an academic metric Y_i and various student characteristics. Let $X_{1,i}$ be student i's weekly hours of studying, then it is reasonable to believe that there are diminishing marginal returns to studying, and that score will be quadratic in studying time. Let $X_{2,i}$ and $X_{3,i}$ be some exogenous shifters of score, unrelated to studying time. Let $X_{4,i}$ be "ability", an unobserved feature. As $X_{4,i}$ increases (higher ability), a student will study more, and perhaps their returns to studying will also be greater, hence $X_{4,i}$ is both correlated with $X_{1,i}$ and it is used in construction of Y_i. Omission of $X_{4,i}$ in the estimation procedure is where the endogeneity comes from - once it is omitted $X_{1,i}$ becomes correlated to the error term. The instruments $Z_{1,i}$ and $Z_{2,i}$ are some exogenous shifters of studying time unrelated to score Y_i. These instruments are correlated with $X_{1,i}$ but are not used in constructing Y_i. The Monte Carlo simulation pulls from the following DGP:

[21] This section re-uses a lot of the comments in the notebook, and for a more complete run through of the Monte Carlo procedure please refer to the notebook.

$$\begin{pmatrix} \mathrm{X}_{1,i} \\ \mathrm{X}_{2,i} \\ \mathrm{X}_{3,i} \\ \mathrm{X}_{4,i} \\ \mathrm{Z}_{1,i} \\ \mathrm{Z}_{2,i} \end{pmatrix} \sim N\left[\begin{pmatrix} 3 \\ -1.5 \\ 1.1 \\ 2.3 \\ -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0.75 & \phi_1 & \phi_2 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 1 & 0 & 0 \\ \phi_1 & 0 & 0 & 0 & 1 & 0 \\ \phi_2 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}\right]$$

$$\varepsilon_i \sim N(0,1)$$

$$\mathrm{Y}_i = 1.5 + 2.5\mathrm{X}_{1,i} - 0.7\mathrm{X}_{1,i}^2 + 2\mathrm{X}_{2,i} + 3\mathrm{X}_{3,i} + 2\mathrm{X}_{4,i} + \nu_i$$

The estimation procedure is as follows: in each iteration I pull n observations of $[\mathrm{Y}, \mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3, \mathrm{Z}_1, \mathrm{Z}_2]$. In this case it is clear that omission of X_4 causes X_1 to be endogenous in an estimation of the DGP ($\mathrm{Cov}[\mathrm{X}_{1,i}, \mathrm{X}_{4,i}] \neq 0 \Rightarrow \mathrm{Cov}[\mathrm{X}_{1,i}, \varepsilon_i] \neq 0$). This process is repeated 100 times, pulling n = 500,000 observations in each iteration. This entire estimation procedure is run using two DGP's, one with strong[22] instruments ($\phi_1 = 0.8$, $\phi_2 = 0.6$) and one with weak instruments ($\phi_1 = 0.25$, $\phi_2 = 0.2$). Below are plots of some of the results from the simulation, for more figures see the notebook or the /figures directory.

[22] See section 3.2.3 for a note on the strength of instruments in the Q2SLS setting.

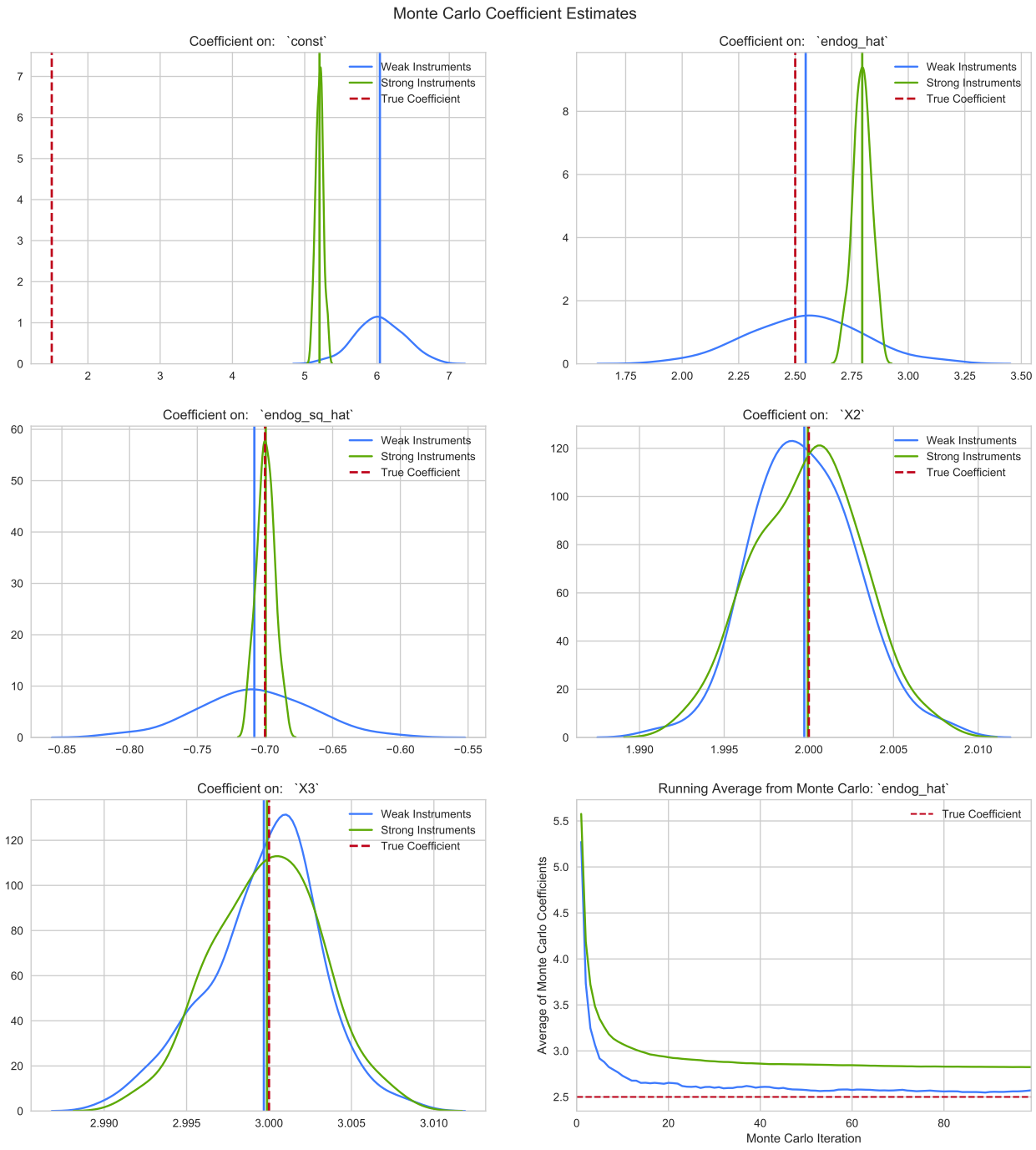


Figure 2: This is the same plot from section 3.2.3 which shows the distributions of the coefficient estimates

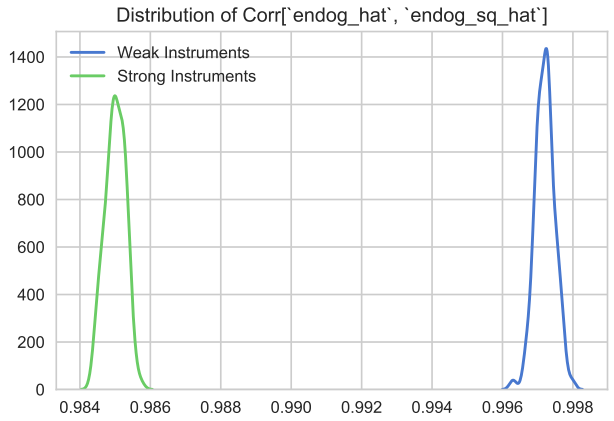


Figure 3: Correlation between the two generated regressors across the Monte Carlo simulations

This figure helped dismiss the idea that the consistent bias in the strong instruments simulation was caused by multicollinearity. The plot shows that in both strong and weak instrumentation, the generated second stage regressors are collinear, in fact the weaker instrumentation has more collinearity. While this collinearity is not great in and of itself, observing it in both instrumentation settings puts the claim that the bias arises from collinearity into question (otherwise one would expect to see the same bias in both the strong and weak instruments, rather than just the strong).

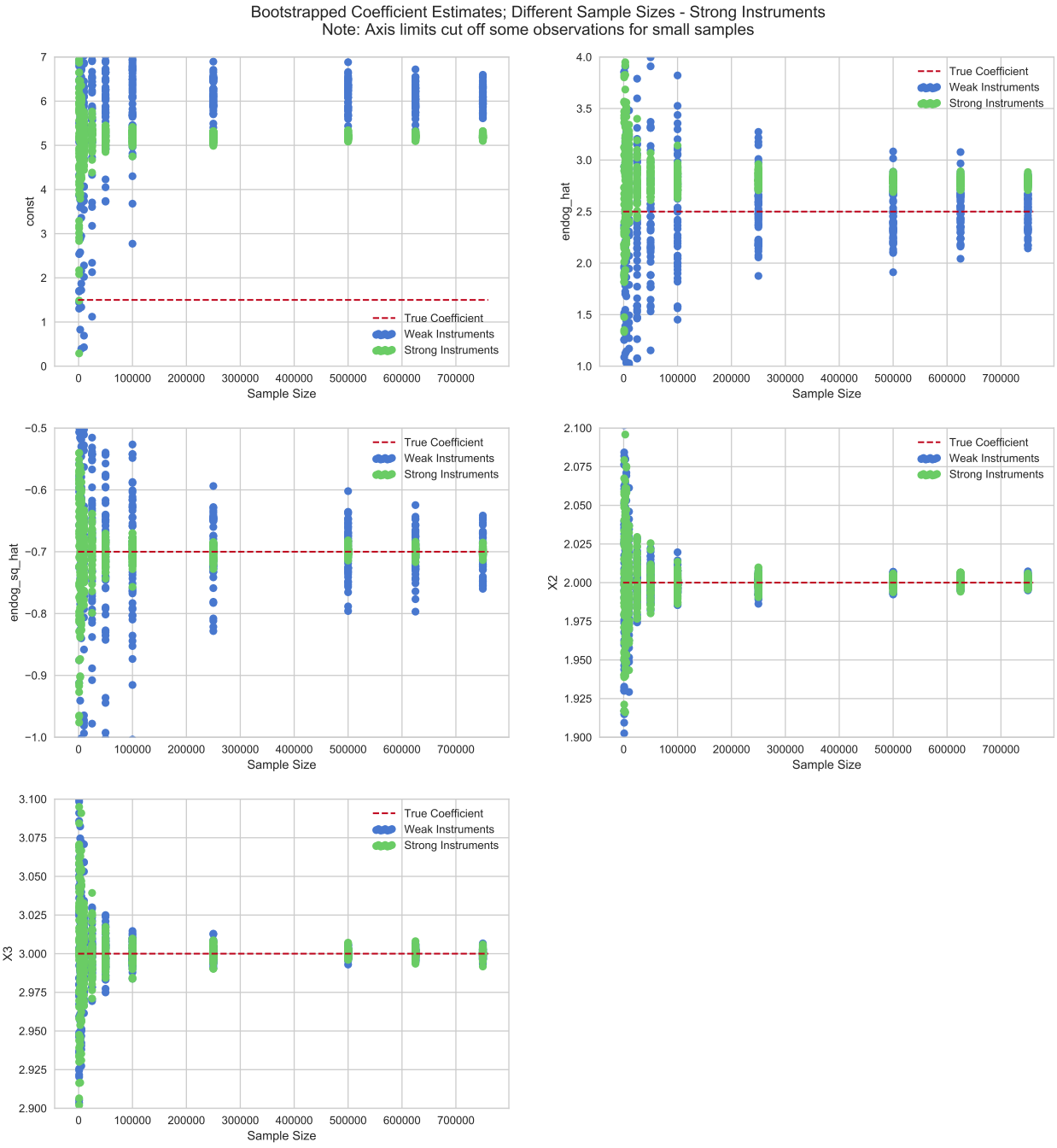


Figure 4: Some support that the estimator is consistent

The results from the simulation reassures that the function was written correctly, but as mentioned previously, the asymptotic behavior is somewhat troubling.

6.3 Q2SLS Extra

[Put bootstrap coefficient distributions here]