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

Nadav Tadelis

April 2018



**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`

# 1 Introduction

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 allows 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 we implement in the analysis.

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 attempts to show an example of what reproducibility looks like for empirical work in the social sciences, especially in applied econometrics.

This paper was written as my honors thesis for undergraduate studies in statistics at UC Berkeley. Due to deadlines for submission I was unable to spend an appropriate amount of time on the econometric analysis, and will point out some of the weak points in my models (specifically the instruments). 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 during my Junior year. 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 integration. Chitwan Kaudan created the original Makefile for running the individual Jupyter Notebooks. All aspects of those original original analyses have been altered significantly, and the history of the alterations is fully documented in the commit history of the git repo.

[A bunch more stuff needs to go here before we dive in]

# 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 [2]. 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. The data include 649 students from a Portuguese Language course of study, and 395 students from a Mathematics course.

[Need more background here]

## 2.1 Data Exploration

[Similar to data exploration notebook, but more detail]

## 2.2 Data Issues

[Go over issues with the data that I didn't originally cover (stated preferences, cross-sectional rather than panel, categorical mappings for continuous vars, etc.]

[Point out that self reported vars like 'freetime' tell us more about student's individual perceptions than the reality. Talk about the benefits and drawbacks to this kind of data]

# 3 Models

[Begin theoretical motivation here]

## 3.1 Naive OLS

[Report the two fits from the naive OLS - first one with only 'studytime' and school level characteristics, then one with individual level characteristics. Note: I think best approach is to point out that there are diminishing returns to studying at the begining, and to start with a models that have both 'studytime' and 'studytime'$^2$]

[Maybe include plot of estimated relationship between 'studytime' and 'G3']

[Lasso-flavor penalization and VIF discussion here, point out that Lasso keeps the variable of interest and also explore the other variables that it weights highly]

[Point out simultaneous causality between 'G1', 'G2', and 'studytime']

[Discuss issues with this naive approach to estimation and why it doesn't apply to causal relationships because of endogeneity, segue to 2SLS theory and explain some theory and bias]

## 3.2 Q2SLS

[Continue discussion of 2SLS theory if needed]

[Point out issue with trying to use 2SLS when you want to account for diminishing marginal returns in the endogenous variable, especially when the instruments are binary (so squaring them does nothing). Lead to discussion of Q2SLS procedure from Wooldridge]

[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]

[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]

[Report results]

[Discuss results]

# 4 Reproducibility

[Discuss importance of reproducibility in general, and especially in today's political environment. Maybe point out that HONEST might start being used non only in climate science, but in policy making as a whole, which would have massive impacts on how applied econometric research (and empirical social science research as a whole) would be viewed by policy makers]

## 4.1 Current Attempts

[Discuss current approaches to reproducibility in econometrics]

[Especially point out Gentzkow & Shapiro's guide for research methods, and its helpfulness in ensuring that research workflow makes sense, but it's lack of any open source reproducibility]

## 4.2 Basic Workflow

[Discuss git version control, environments, makefiles, using notebooks as a way to document any analysis that was not included in the final paper, but had an impact on the direction of the results (this increases confidence that there is not any unintentional p-hacking style things going on)]

## 4.3 Custom Functions

[Discuss the process of creating your own function in a way that other people can implement it without hassle: docstrings, comments, readability, etc.]

[Cover the long process of testing a new function that isnt necessarily included in the final tests.py script (function_testing.ipynb)]

[Explain Travis and CI]

## 4.4 Optional Elements

[Cover optional but useful things that the open source community has begun using: Binder, Sphinx, etc.]

[Talk to Fernando about things here that I wouldn't think of]

# 5 Conclusion

citation test [4, 3, 1]

# References


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

[2] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. *A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pages 5–12, 01 2008.

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

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