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

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 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 implemented in the analysis.

This paper was written as my honors thesis for undergraduate studies in statistics at UC Berkeley. I would like to thank my thesis advisor, Professor Fernando Pérez, for his generosity with his time and deep knowledge of the reproducibility movement. 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`[Maybe link to script in repo], and the associated tests and Travis integration. 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 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 (Cortez & 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 background here?]

2.1 Data Exploration

There is some pre-analysis data cleaning and exploration that we run, which can be found in the data exploration notebook here; some figures from this notebook are included in the Appendix. 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 the 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.

[Very Rough]

2.2 Data Issues

As is often the case in the education space, 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. [Expand more here, talk about why this is bad when studytime changes through the academic year. This might be a good place to include the simultaneous causality issue, or at least allude to it so that it is more familiar when we go over it later (?)]

[Very Rough]

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

3 Models

We need to first set up our structural equation [Maybe this isn't actually a structural equation? Need to double check this term] defining the relationship between grades and studying (Card & 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:

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

Where X_i is a matrix 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

negative coefficient on studying time[2]. 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, there is an argument for including free time in the final model estimation procedure.

3.1 Naive OLS

[Report the four fits from the naive OLS - first one with only 'studytime' and school level characteristics, then one with individual level characteristics (for each course of study)]

[Maybe include plot of estimated relationship between 'studytime' and 'G3' (nonlinear, one line for math, one for Portuguese, line type can denote individual level controls - or will that make the graph to crowded?)]

3.1.1 Additional Analysis

[Note that the model fitting 1 notebook includes running the model on both mapping schemes for 'studytime']

[Lasso-flavor penalization and VIF discussion here (maybe no numbers, just general takeaways and link to the relevant notebook), point out that Lasso keeps the variable of interest and also explore the other variables that it weights highly], note: in lasso section cite: Cameron (2017) note: in VIF section cite: O'brien (2007) and Mansfield & Helms (1982)

3.1.2 Simultaneous Causality

There is an additional issue with this model specification that we did not explore in the original discussion. The data include three test scores: G1, G2, and G3. G1 and G2

[2]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.

are midterms, and G3 is the final. It is plausible that part of how students inform their study allocation decision comes from how well they did on their previous exam and how many hours they studied in the time leading up to the exam. The data is taken at a single undefined point in time, so it is unclear at what point the students are reporting their weekly amount of studying. This complicates things because if the survey was administered after the midterms, then there is clear simultaneous causality between study time, G1 and G2. This is the motivation for excluding G1 and G2 from the model estimation. There may be a novel way of

[Need to edit this section, right now its really rough]

[Maybe move this to right before the Naive OLS section?]

Actually: [Maybe move this to the Data Issues as its own subsection?]

3.1.3 Endogeneity

[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]

3.2.1 Motivation

[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]

3.2.2 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), and the simulation seems to suggest that in the case of weak instruments the estimates are still okay (maybe) but that the SE's are going to be too 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]

3.2.3 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]

Note: Include discussion of validity of exclusion restriction - `http://econ.lse.ac.uk/staff/spischke/ec533/Weak%20IV.pdf`

[Report results]

[Discuss results]

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]

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 (n.d.). 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[3], 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 & 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[4].

[Need to clean up this section a lot]

[3]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.

[4]Ignoring the issue of confidential data.

4.1 Current Resources

[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 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)] [Definitely link to the stat 159 page because we don't want to go into the nitty gritty details of everything, just an overarching view of the tools and how they're used together. also explain the benefits of all of these tools (ability to see edit history, ease of remote collaboration, ensuring compatibility on multiple machines, automatic function testing, etc.)]

4.3 Custom Functions

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

[Cover the long process of testing a new function that isn't 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.]

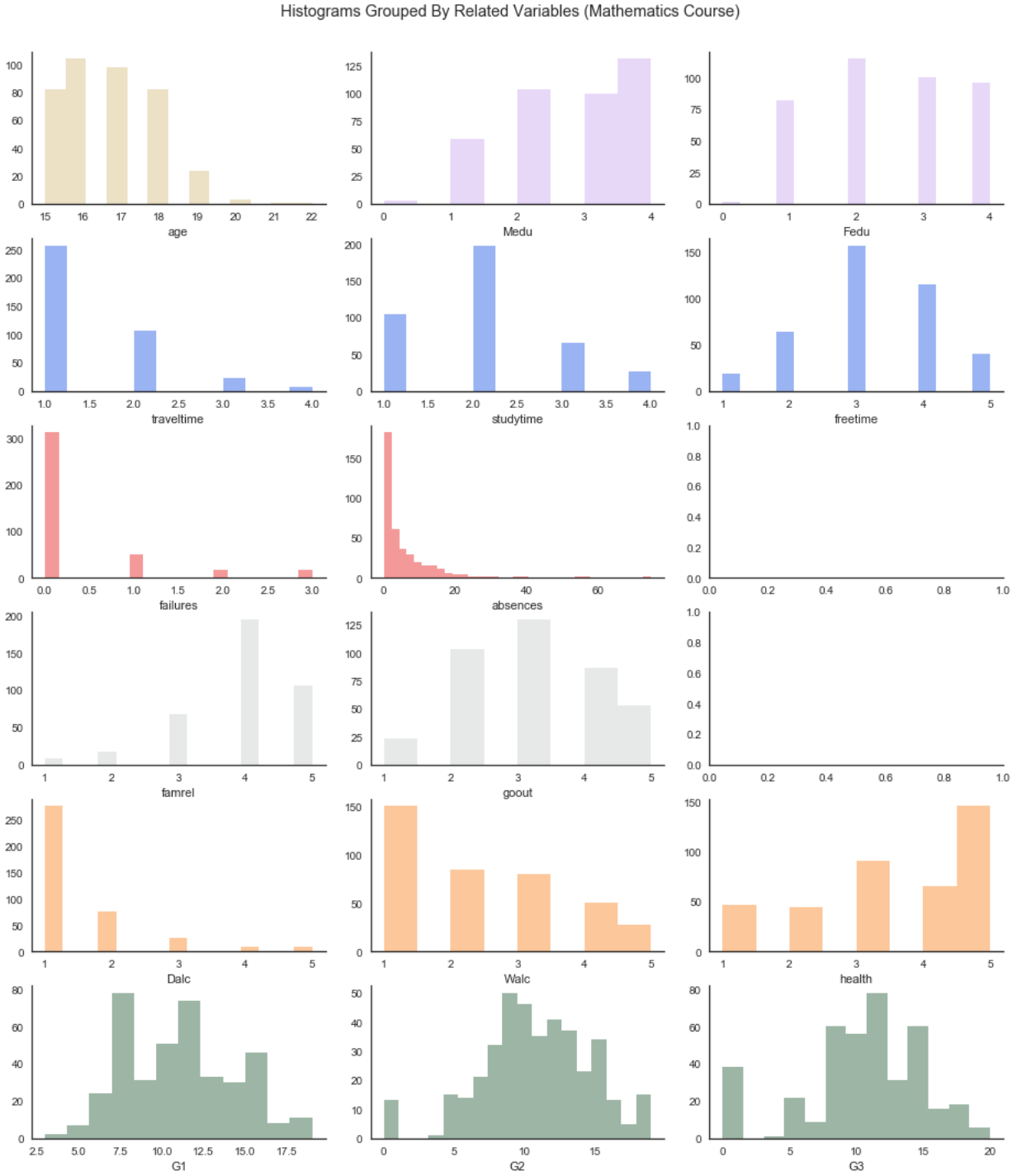
[Talk to Fernando about things here that I wouldn't think of]

5 Conclusion (? - maybe not needed)

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

6 Appendix

6.1 Data Exploration Figures



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***	0.5611***	0.3397**	1.0857**
	(0.0137)	(0.0143)	(0.0366)	(0.1392)	(0.1381)	(0.4946)
Studytime_2	0.0361***	0.0526***	0.0050	0.0224	0.0186	0.0252
	(0.0136)	(0.0143)	(0.0285)	(0.0142)	(0.0149)	(0.0327)
Studytime_3	0.0924***	0.1045***	0.0668*	0.0646***	0.0483***	0.0903**
	(0.0177)	(0.0166)	(0.0380)	(0.0183)	(0.0184)	(0.0426)
Studytime_4	0.0787***	0.0927***	0.0563	0.0381	0.0476*	0.0339
	(0.0285)	(0.0278)	(0.0575)	(0.0293)	(0.0269)	(0.0672)
School_GP	0.0742***	0.0856***	0.0283	0.0371**	0.0605***	−0.0309
	(0.0133)	(0.0142)	(0.0337)	(0.0150)	(0.0160)	(0.0453)
Course_math	−0.0954***			−0.0958***		
	(0.0128)			(0.0150)		
Individual Controls	*no*	*no*	*no*	*yes*	*yes*	*yes*
Observations	1,044	649	395	1,044	649	395
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	7.04***	5.72***	2.834***
Note:						*p<0.1; **p<0.05; ***p<0.01

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***	0.5497***	0.3312**	1.0605**
	(0.0161)	(0.0165)	(0.0402)	(0.1384)	(0.1365)	(0.4971)
Studytime_continuous	0.0220***	0.0268***	0.0128	0.0160***	0.0110**	0.0229*
	(0.0053)	(0.0052)	(0.0113)	(0.0056)	(0.0056)	(0.0126)
Studytime_continuous_sq	−0.0010***	−0.0013***	−0.0005	−0.0008**	−0.0005	−0.0012
	(0.0004)	(0.0004)	(0.0007)	(0.0004)	(0.0004)	(0.0008)
School_GP	0.0737***	0.0856***	0.0262	0.0372**	0.0605***	−0.0312
	(0.0133)	(0.0142)	(0.0335)	(0.0150)	(0.0161)	(0.0453)
Course_math	−0.0956***			−0.0957***		
	(0.0128)			(0.0150)		
Individual Controls	*no*	*no*	*no*	*yes*	*yes*	*yes*
Observations	1,044	649	395	1,044	649	395
R^2	0.09	0.13	0.01	0.33	0.41	0.35
Adjusted R^2	0.09	0.13	0.00	0.28	0.34	0.22
F Statistic	30.11***	30.61***	1.41	7.09***	5.84***	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

6.5 Miscellanea

References

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

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

Card, D. & 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. & 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. & 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. & White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics* **29**(3), 305–325.

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

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

Schulson, M. (n.d.), '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.