IV Model with Heterogeneous Treatment Effects in Economic History

Federico Crippa
Department of Economics
Northwestern University

Valerio Di Tommaso
Department of Economics
Northwestern University

March 14, 2023

Abstract

Bisin and Moro (2021) analyze the implications of allowing for heterogeneity in the instrumental variable (IV) model in Economic History. They present analytical results with a binary treatment and binary instrument, but cover examples with continuous treatments and instruments. This paper shows that the generalization to the continuous setting must be done carefully. We provide a model for IV regressions with heterogeneous treatment effects where treatment and instrument are continuous variables. We show that the estimand of interest — a (weighted) average of the marginal treatment effects — is interpretable only under a set of assumptions which includes the *monotonicity* of the treatment response to the instrument. This monotonicity assumption implies a testable implication: we show how to test it as a robustness check for the model. We apply the robustness test to three empirical applications from the Economic History literature. In all cases, the model is rejected: we provide possible explanations for the rejections, highlighting the historical context of each application.

Keywords: Instrumental Variable, LATE, Economic History

1 Introduction

Instrumental variable regressions are widely used tools in applied historical economics research, with an increasing trend in their usage: Caicedo (2021) documents how the number of historical economics papers in the top 20 economics journals using an IV regression has increased more than fivefold in the last twenty years. Scholars usually leverage their knowledge of a specific historical context to find an exogenous source of variation that allows them to identify causal effects in cases that are otherwise difficult to study. Thanks to IV regressions, they can often empirically validate hypotheses made in other disciplines, such as history, sociology, or anthropology. Quantitative tools strengthen the credibility of certain theories, but they often rely on assumptions that may not seem plausible in many historical contexts.

Homogeneity of treatment effects is one of these assumptions, since historical economics regressions often range over events spread in time and space. In a recent paper, Bisin and Moro (2021) argue that a way to overcome this limitation is to allow for heterogeneous treatment effects: relying on weaker and more credible assumptions, it is possible to more robustly identify a local average treatment effect (LATE). Bisin and Moro (2021) present the IV model with heterogeneous treatment effects for Economic History focusing on a setting with binary treatment and binary instrument, *to simplify the intuition and to present results with minimal algebra*. However, they discuss examples with continuous treatments and continuous instruments (Acemoglu et al. (2001); Alesina et al. (2013); Ashraf and Galor (2013)), which are the most common setting in historical economics.

The first contribution of this paper is to formalize the IV model with heterogeneous treatment effects with continuous treatment and continuous instrument, highlighting the needed assumptions and how they may (or may not) be plausible in historical economics contexts. To identify an interpretable parameter, the shift from homogeneous to heterogeneous treatment effects must go along with additional assumptions. Researchers need to assume that the instrument marginal effects in the first stage are independent from the received instrument and *monotone*, i.e., they have the same sign for all the units. These assumptions are crucial

for the binary case, as shown in the seminal LATE result by Angrist and Imbens (1995). However, how to formulate them in the continuous treatment and instrument case seems less clear, and their importance appears underappreciated: the monotonicity assumption is mentioned by Bisin and Moro (2021) just in a footnote.[1] The correct statement of the model assumptions is necessary to discuss their validity in applications.

Once the model is set, we use the assumptions to derive an implication on observable quantities: the stochastic monotonicity of the joint distribution of the instrument and the treatment. We show how it is possible to test this implication using the test proposed by Chetverikov et al. (2021): the test can be seen as a robustness test for the IV model with heterogeneous treatment effects, easy-to-implement and easy-to-interpret. The test should not replace a careful discussion on why the model assumptions hold in a specific application since more than passing the test is needed for the assumptions to hold. Nonetheless, the empirical evidence provided by the test may strengthen the credibility of the model. When the implication is not rejected, we may assume that at least a necessary condition of the model is satisfied.

We consider three applications (Becker and Woessmann, 2009; Acemoglu et al., 2001; Gorodnichenko and Roland, 2017) to show what the assumptions for the IV model with heterogeneous treatment effects require in practice and how to use the robustness test we propose. The examples are helpful to see real case discussion of the test and the assumptions and raise some concerns related to the use of the IV model with heterogeneous treatment effects for historical economics. In fact, in all three applications, the robustness test rejects the implication. The result is remarkable: the data are rejecting the model, limiting the reliability of the results, and it is not clear how the conclusions of the three applications should be considered, as they are based on non credible premises. We aim not to contest the findings but to highlight the importance of valid assumptions and show how the robustness

[1]It seems a norm rather than an exception: Fiorini and Stevens (2021) find that half of the paper published in the AER over 2005-19 period that explicitly identifies a LATE does not even include the word "monotonicity".

test may help assess them.

1.1 Literature review

This paper addresses three strands of literature. First, it gives a methodological contribution to econometrics for historical economics. In the 1960s, the cliometric revolution bolstered quantitative approaches to studying history. More recently, the focus on historical natural experiments gained traction. By borrowing tools from applied microeconomics, researchers try to establish causal relationships in historical context leveraging on serious identification strategies (Cantoni and Yuchtman, 2021). Caicedo (2021) reports a survey of papers in historical economics resorting to IV regressions and RDD, highlighting how from 2005 to 2018, the percentage of papers using advanced econometrics in the top 5 Economic History journals has more than tripled. These methods are exploited to study a wide range of events, from the Neolithic Revolution (Ashraf and Galor, 2011) to the Protestant Reformation (Becker and Woessmann, 2009), from colonialism (Acemoglu et al., 2001) to the Great Depression (Calomiris and Mason, 2003). Critiques have been moved to this approach, as summarized in Cantoni and Yuchtman (2021): the main one is that when events are studied not because they are relevant but because they allow for clean identification strategies, the risk is of missing important aspects of historical processes. With this paper, we want to raise a different and more fundamental caveat: applying causal inference methods developed for different contexts to historical economics may not be straightforward and demands attention. Causal inference, to be meaningful, requires the correct use of its tools.

A second contribution is to the literature on interpreting instrumental variable regressions. It starts with the LATE framework proposed by Angrist and Imbens (1995) and Angrist et al. (1996). When heterogeneity in treatment effects and responses to the instrument is allowed, the IV estimates can be interpreted as an average treatment effect for a specific group, the compliers. If and how this estimand is interesting has been debated for long (Heckman, 1997; Heckman and Urzua, 2010; Deaton, 2010; Imbens, 2010), even for the specific

context of historical economics (Bisin and Moro, 2021; Casey and Klemp, 2021). We do not intend to propose a new interpretation of the LATE estimand but to stress that more assumptions than the ones usually explicitly made in applied research are required to obtain that estimand. Without these assumptions, neither the controversial LATE interpretation is valid. In this sense, our paper is similar to Blandhol et al. (2022). In the wake of Abadie (2003) and Słoczyński (2022), they show that if the two-stage least squares specification includes covariates, as done in many empirical works, then the LATE interpretation does not apply without further functional form assumptions. We focus on the assumptions needed for the LATE interpretation in case the treatment and the instrument are continuous, without dealing with the functional form and assuming instead a linear additive model.

Finally, our paper is an example of how shape restrictions used to point identify a parameter of interest give implications that can be tested to strengthen the credibility of the identification model.[2] We are not the first to consider testable implications for the IV heterogeneous treatment effect model. Considering a binary treatment and a binary instrument, Balke and Pearl (1997) and Heckman and Vytlacil (2005) derive testable implications later used by Kitagawa (2015) to build a test for instrument validity in the LATE framework. In the same binary treatment and instrument context, Machado et al. (2019) propose a test for different sets of restrictions needed to identify the sign of the average treatment effect. In this paper, we consider a continuous endogenous treatment and a continuous instrument to derive an implication similar to the one in Chetverikov and Wilhelm (2017): compared to them, we focus on the linear model and show the relevance of the implication and its test in applied research.

The rest of the paper is organized as follows. In section 2, we construct the model and derive the testable implication. In section 3, we show how the implication can be tested. In section 4, we consider three applications (Becker and Woessmann, 2009; Acemoglu et al., 2001; Gorodnichenko and Roland, 2017), discussing the assumptions required by the LATE

[2]See Chetverikov et al. (2018) for a comprehensive review on the econometrics of shape restrictions.

model and testing the derived implication. Section 5 concludes.

2 Model

2.1 Setup

Consider the following linear model, which can be considered a generalization of Angrist and Imbens (1995) and Angrist et al. (1996):

$$Y_i = \alpha + \beta_i D_i + u_i \tag{1}$$

$$D_i = \delta + \gamma_i Z_i + \epsilon_i \tag{2}$$

where Y_i is the output, D_i the continuous treatment and Z_i the continuous instrument.

The model allows for additional control variables by defining each of (Y_i, D_i, Z_i) as the residuals from a regression of each of those variables on the vector of controls (this approach is common in the literature on IV regressions, see for example Lee et al. (2022)). Unlike an otherwise standard IV regression model, here (β_i, γ_i) are not fixed quantities: they are realizations of the two *random variables* (β, γ), potentially different for different i. The first, β_i, is the *marginal effect* on the outcome Y_i of increasing the individual endogenous treatment D_i. This allows for an *heterogeneous* response to the treatment. The second, γ_i is the *marginal effect* on the endogenous treatment D_i of increasing the individual instrument Z_i. We will not make assumptions on how β and γ may be related to each other: as highlighted by Bisin and Moro (2021), in many historical applications, we expect them to be positively correlated, and our model allows for this dependency.

Consider the following assumptions instead, standard in the LATE literature with binary treatments and binary instruments:

1. Stable Unit Treatment Value Assumption (SUTVA): the outcome of unit i does not depend on the treatment status of other individuals.

2. Random assignment of the instrument Z_i: $Z_i \perp\!\!\!\perp (\gamma_i, \epsilon_i, \beta_i)$. The model with constant β and γ immediately satisfies independence. Since we want to allow for heterogeneous β and γ, independence assumptions must be added. We are asking that the amount of the instrument assigned to a certain unit is independent of the effects of the instrument on the treatment and of the treatment on the outcome. This assumption seems too strong in many applications and would probably be disputed. Later, we will discuss how it may be relaxed.

3. Exclusion restriction: $Z_i \perp\!\!\!\perp u_i$.

4. Nonzero Average Causal Effect of the instrument: $\mathbb{E}[\gamma_i] \neq 0$.

The monotonicity assumption will be discussed later. Moreover, assume $\mathbb{E}[\epsilon_i] = \mathbb{E}[u_i] = 0$.

2.2 TSLS Estimand

It is common practice to consider the model described in equations 1 and 2 and estimate a two-stage least squares regression of Y on D, using Z as the instrument. Under assumptions 1-4, the estimand β^{IV} identifies:

$$\beta^{IV} = \frac{\text{Cov}(Z,Y)}{\text{Cov}(Z,D)} = \frac{\text{Cov}(Z,\beta\gamma Z)}{\text{Cov}(Z,\gamma Z)} = \frac{\mathbb{E}[Z^2\beta\gamma] - \mathbb{E}[Z]\mathbb{E}[Z\beta\gamma]}{\mathbb{E}[Z^2\gamma] - \mathbb{E}[Z]\mathbb{E}[\gamma Z]} = \frac{\text{Var}(Z)\mathbb{E}[\beta\gamma]}{\text{Var}(Z)\mathbb{E}[\gamma]} = \mathbb{E}\left[\frac{\gamma}{\mathbb{E}[\gamma]}\beta\right].$$

Therefore, the estimand is a weighted average of marginal treatment effects β_i, where weights are proportional to the heterogeneous effects in the first stage. If the marginal effects γ_i in the first stage have the same sign for all i, i.e. $\gamma_i > 0$ (or $<$) for every i, the weights $\frac{\gamma_i}{\mathbb{E}[\gamma]}$ are positive for all the β_i. This is a desirable result: if negative weights are allowed, the interpretation of the estimand β^{IV} is almost impossible, as it could be positive even if all β_i are negative, or vice versa. To avoid this possibility, a fifth assumption is needed:

5. Monotonicity: $\forall i,\ \gamma_i > 0$. Otherwise, $\forall i,\ \gamma_i < 0$.

It is not obvious if the LATE-type parameter $\mathbb{E}\left[\frac{\gamma}{\mathbb{E}[\gamma]}\beta\right]$, even with positive weights $\frac{\gamma}{\mathbb{E}[\gamma]}$ is relevant. For the binary case, this issue has been largely debated (Heckman, 1997; Heckman and Urzua, 2010; Deaton, 2010; Imbens, 2010), but it is not our intention to take part in this discussion. We assume that applied researchers in historical economics are interested in this specific estimand β^{IV}, which requires assumptions 1-5 to be valid.

2.2.1 Relaxing the *Random assignment of the treatment* assumption

Previously, we assumed $Z_i \perp\!\!\!\perp (\gamma_i, \epsilon_i, \beta_i)$. Supposed we are interested in relaxing the assumption and allowing for the dependence of Z and γ, maintaining the assumption $Z_i \perp\!\!\!\perp (\epsilon_i, \beta_i)$. Under this new assumption, the β^{IV} estimand is:

$$\beta^{IV} = \frac{\text{Cov}(Z,Y)}{\text{Cov}(Z,D)} = \frac{\text{Cov}(Z,\beta\gamma Z)}{\text{Cov}(Z,\gamma Z)} = \frac{\mathbb{E}[Z^2\beta\gamma] - \mathbb{E}[Z]\mathbb{E}[Z\beta\gamma]}{\mathbb{E}[Z^2\gamma] - \mathbb{E}[Z]\mathbb{E}[\gamma Z]} = \mathbb{E}\left[\frac{Z^2\gamma - \mathbb{E}[Z]Z\gamma}{\mathbb{E}[Z^2\gamma] - \mathbb{E}[Z]\mathbb{E}[\gamma Z]}\beta\right].$$

The estimand is still a weighted average of marginal effects β_i, but it is now even harder to interpret the weights or to propose a meaningful assumption that guarantees them to be positive. This is the first reason that makes us particularly cautious considering this weaker version of assumption 3. The second is that we would assume dependence of β and γ, and of γ and Z, but independence of β and Z. Although possible, it is hard to imagine a context in which this assumption is more credible than assuming independence of γ and Z. Overall, the cost of weakening the assumption seems higher than the benefit.

2.3 Testable Implication

Assumptions 2-5 involve unobservable quantities (β, γ, ϵ, and u), and cannot be singularly tested. However, combined with the model in equations 1 and 2, they imply a condition on the observable joint distribution of D and Z. Crucially, this implication is testable.

To derive the implication, observe the following:

$$Pr(D_i \leq d|z) = Pr(\delta + \gamma_i Z + \epsilon_i \leq d|z) = Pr(\delta + \gamma_i z + \epsilon_i \leq d).$$

The first equality follows from equations 2. Linearity here is unnecessary: a monotone relation between Z and D would give a similar result. For z' and z'' such that $z'' > z'$

$$Pr(\delta + \gamma_i z' + \epsilon_i \leq d) \geq Pr(\delta + \gamma_i z'' + \epsilon_i \leq d)$$

and then

$$Pr(D_i \leq d|z') \geq Pr(D_i \leq d|z''). \tag{3}$$

Since the result in equation 3 holds for any d and $\{z', z''\}$ such that $z'' > z'$, it can be stated as a stochastic monotonicity condition:

$$F_{D|Z}(d|z') \geq F_{D|Z}(d|z'') \tag{4}$$
$$\forall d \quad \forall z'' > z'.$$

The implication in equation (4) involves the conditional CDF $F_{D|Z}$, which is observable. Hence, the implication can serve as the null hypothesis for a robustness test for this heterogeneous effects IV design.

What would a test of the implication actually test? In other words, for which assumptions of the model the implication in equation (4) represents a necessary condition? In section 2.2, we used assumptions 1-4 to show why assumption 5 is needed. Then, to derive the implication in this section, we only used assumptions 2 and 5. Testing the implication in equation (4) is hence equivalent to testing a necessary condition for the random assignment of the treatment and the monotonicity assumption.

Finally, it is worth noting that the (more standard) case of the homogeneous effect IV model — that is, the model in which (β, γ) are *constants* — is a particular case of the model in section 2. This implies that the testable implication derived in this section also applies to the homogeneous effect IV model, and it can be tested using the same procedure. Thus, failing the test rejects both the heterogeneous and homogeneous effect models.

3 Test

We are not the first to exploit the monotonicity assumption to derive a testable implication for the heterogeneous effects IV model. The idea is already present in the seminal paper by Angrist and Imbens (1995). In the context they consider (dummy instrument and multivalued treatment), monotonicity implies that the cumulative distribution functions (CDFs) of the treatment conditional on different instrument values should not cross, i.e., one should stochastically dominate the other. The paper does not provide any proper test (in the application, it only includes some graphical evidence of not crossing), but several tests considering the null hypothesis of dominance against the alternative of non-dominance appear in the literature, and could be applied (consider for example McFadden (1989), Barrett and Donald (2003) or Linton et al. (2005), and the comprehensive review in Whang et al. (2019)).

This approach does not work with binary treatment and binary instrument, as the CDFs of dummy variables never cross. Nonetheless, Kitagawa (2015) proposes a test to validate the design also for this setting, resulting from assumptions on instrument exclusion, random assignment, and monotonicity.

For the case with continuous treatment and instrument, the implication we derived in equation 4 is similar to a result by Chetverikov and Wilhelm (2017), although they consider different assumptions. We need a procedure to test the implication: testing stochastic monotonicity is not a new problem in economics (it usually concerns the relation between the treatment and the output), and some tests for stochastic monotonicity have been proposed:

Lee et al. (2009) developed a test based on the supremum of a rescaled second-order U-process, while Delgado and Escanciano (2012) consider a test that consists of comparing restricted and unrestricted estimates of the difference between the joint distribution functions, and Seo (2018) extends their results.

For this paper, we consider the test proposed by Chetverikov et al. (2021): the test statistic is based on a locally weighted version of Kendall's tau and can be seen as an adaptive version of the test in Lee et al. (2009). It overcomes the problem of bandwidth choice, considering a search over different values. The null hypothesis is that the joint distribution of the variables satisfies the stochastic monotonicity condition.

Before proceeding to the applications, we must emphasize that the proposed test procedure should be used as a robustness check rather than a pre-test. We recommend reporting the results of IV estimates whether or not the stochastic monotonicity is rejected. Otherwise, if estimates are reported conditionally on the test result, regular inference for the IV model would no more be valid. This concern is usual with pre-testing: proposing a post-testing inference procedure for the heterogeneous effects IV model goes beyond the scope of this paper.

4 Applications

We considered three papers in the Economic History literature using the IV model to test the monotonicity condition: Becker and Woessmann (2009); Acemoglu et al. (2001); Gorodnichenko and Roland (2017). These papers feature a continuous treatment and a continuous instrument in a linear setting. Thus, all of them fit into our linear framework of heterogeneous responses. Moreover, each paper can be in some sense illustrative for different strands of Economic History literature:

- Becker and Woessmann (2009), for the use of the distance as an instrument, which is quite common in Economic History.[3]

- Acemoglu et al. (2001), as a seminal paper in the persistence literature and discussed in Bisin and Moro (2021).

- Gorodnichenko and Roland (2017), as one example of the literature studying genetic distance and economic development.[4]

To be clear, none of the above papers directly invokes a heterogeneous effects model. The regressions in those papers are linear IV models with constant coefficients. Nonetheless, as mentioned before, the test is also valid for such homogeneous effect models. Thus, failing the test also rejects the linear IV model with constant coefficients.

4.1 Becker and Woessmann (2009)

Becker and Woessmann (2009) questioned Max Weber's idea that the higher economic prosperity of Protestant regions is due to the Protestant work ethic. The alternative theory they propose is that Protestant economies prospered because instruction in reading the Bible increased human capital and led to economic success. According to the authors, "the Protestant lead in literacy is large enough to account for practically the entire Protestant lead in economy". To test this theory, the authors consider county-level data from late-nineteenth-century Prussia, exploiting the initial concentric dispersion of the Reformation to use distance to Wittenberg as an instrument for Protestantism.

The main specification in Becker and Woessmann (2009) includes control variables. It is assumed that the instrument is exogenous conditional on them. Therefore we will use the residualized version of model equations (1) and (2), letting $(\tilde{Z}_i, \tilde{D}_i, \tilde{Y}_i)$ be the residuals from regressing each of $(\tilde{Z}_i, \tilde{D}_i, \tilde{Y}_i)$ the set of controls [5].

[3] See, for example, Dittmar (2011) for a similar paper.

[4] See also Spolaore and Wacziarg (2013) for a discussion on this literature.

[5] The set of control variables is: % people below 10 years old, % Jews, % females, % born in municipalities,

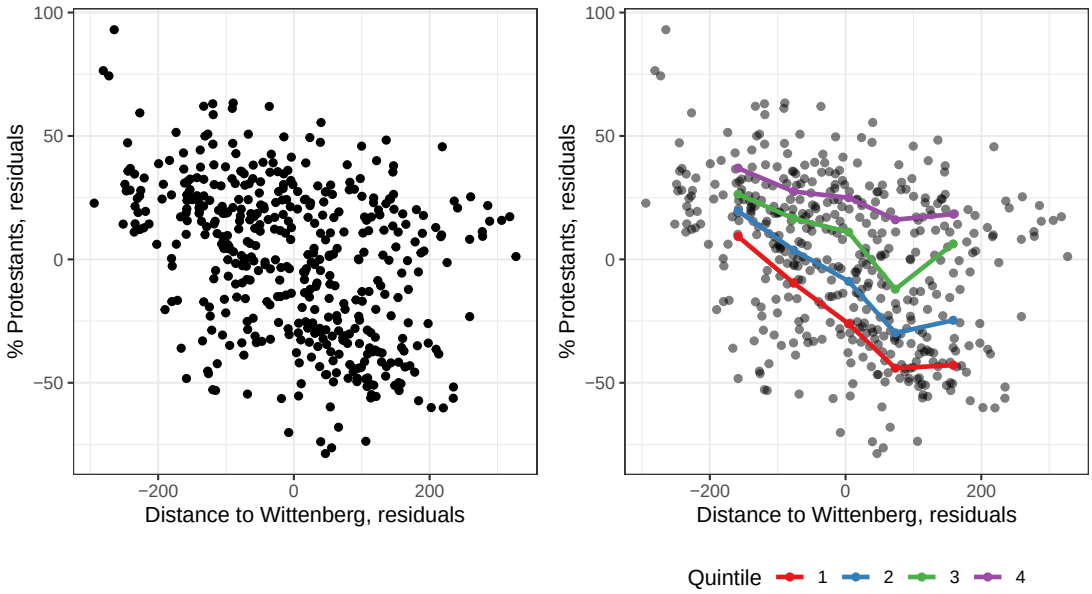

Figure 1: Scatterplots of the residuals for the distance to Wittenberg and % of Protestants. Each dot represents a Prussian county. On the right panel, colored lines show how quintiles of the treatment (% of Protestants) evolve along quintiles of the instrument (distance to Wittenberg). The counties are divided into five groups (corresponding to quintiles of the instrument). We computed quintiles of the treatment in each group and connected them by colored segments.

On the left panel, Figure 1 reports the scatterplot for $\tilde{Z}_i$ and $\tilde{D}_i$. On average, the percentage of Protestants seems to decline as we move away from Wittenberg. This is an indication of *instrument validity*, or *Nonzero Average Causal Effect of the instrument* in our model. The right panel shows how the quintiles of the treatment change moving along the quintiles of the instrument. Consider, for example, the red line: it shows how the first quintile of the treatment (% Protestants) evolves along the quintiles of the instrument (distance to Wittenberg). Under conditional stochastic monotonicity, quantiles should change monotonically. The green line in the figure suggests that it is not the case here.

% of Prussian origin, Average household size, ln(population size), Population growth 1867-1871, % missing education info.

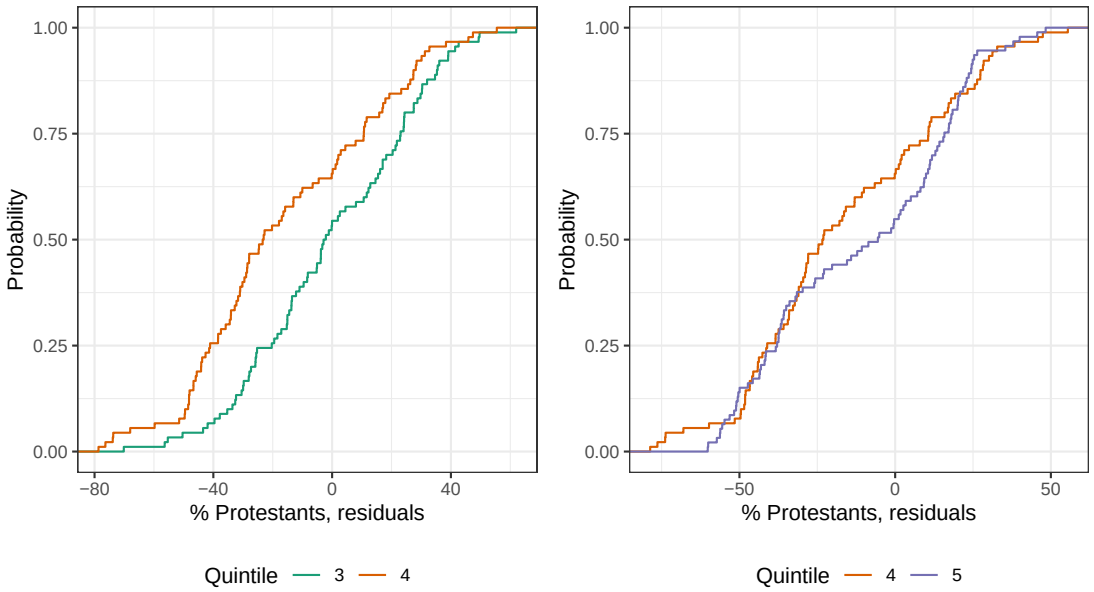

Figure 2: Plots of the empirical CDFs of the treatment (% Protestants) conditional on belonging to different quintiles of the instrument (distance to Wittenberg). The counties are divided into five groups (corresponding to quintiles of the instrument), and then the empirical CDFs are plotted for selected groups.

4.1.1 Testing the stochastic monotonicity in Becker and Woessmann (2009)

We run the test for stochastic monotonicity considering an asymptotic probability of rejection of the true null of 0.01, bootstrapping the test statistic from the empirical distribution 500 times. The null hypothesis of stochastic monotonicity is rejected.

4.1.2 Why might the test have failed?

Rejecting stochastic monotonicity means that the relation between the endogenous treatment and the instrument is not *monotone*. In the linear model, we reject the hypothesis that γ_i has only non-positive values. We focus on two possible reasons for the test rejection: differences between Prussian counties at the time of the Reform and the Euclidean distance as the instrument.

The map of Figure 3 suggests that the data do not provide sufficient evidence for the share of protestants to decrease linearly with the distance from Wittenberg. Consider the

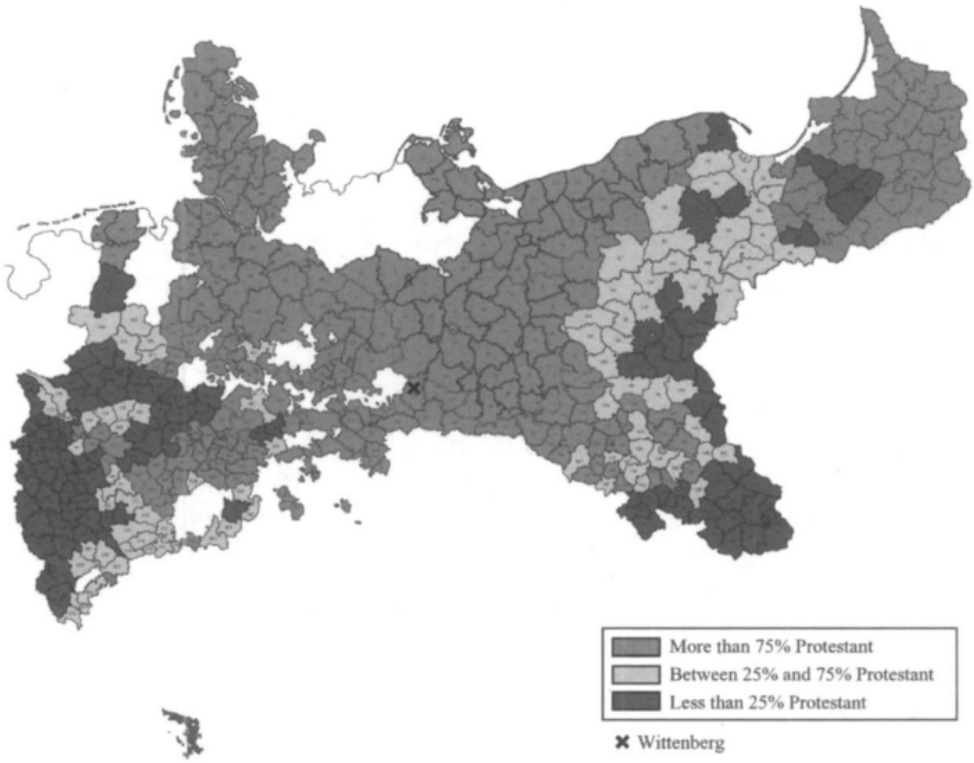

Figure 3: Figure 3 in Becker and Woessmann (2009), showing Protestantism in nineteenth-century Prussia.

territories previously part of the Prince-Bishopric of Paderborn, which are about as far to Wittenberg as the territories of the county of Kiel.[6] This county was part of the Kingdom of Denmark, a protagonist, on the Protestant side, of one of the Thirty Years' War phases, as described by the Western historiographic tradition. The difference in the share of protestants between the two regions is huge: under the monotonicity assumption, it needs to be entirely explained by substantial error term and by a *non-negative* slope — γ_i in the model — for those territories part of Prince-Bishopric of Paderborn.

To understand why the monotonicity assumption is strong in this context, consider the thought experiment of *reducing* the distance between Padeborn, part of a Catholic Diocese since 799, and Wittenberg. Under the monotonicity assumption, the number of protestants in those territories has to weakly increase. This may not happen: the decision to embrace the Reform made by local German rulers was mainly political, as demonstrated by the series

[6]The county of Paderborn is even 60 km closer to Wittenberg, to be precise.

of conflicts between Lutheran princes and the Holy Roman Empire during the mid-16th century.[7] Decreasing the distance from the center of the Reform could expose Catholic centers to a larger number of Peasants' protests and thus to harsher repression of Protestantism. In this case, the sign of the slope could be reversed, and monotonicity does not hold.

A second reason for rejecting the monotonicity assumption may be the use of the Euclidean distance between counties and Wittenberg as the instrumental variable. As the authors point out, "The main reasons for a circular dispersion around Wittenberg may have been the costs of traveling and of information diffusion through space, and these transportation and transaction costs played a crucial role at the time." However, if transportation and transaction costs played such an important role, looking at the Euclidean distance may be misleading. Traveling in Europe during the 16th century happened mainly through major roads and rivers (Braudel, 1972).[8] The Elbe, one of the major rivers in Europe, goes through Wittenberg and connects several important German cities until it flows into the North Sea near Hamburg. Looking again at Figure 3, all those territories reached by the Elbe have many protestants. Therefore, not accounting for the way travel occurred at the time can violate the monotonicity assumption.

Finally, note that by reducing the distance of a county from Wittenberg, we are also reducing its distance from major Catholic centers or territories under the direct control of the Habsburgs. This may influence the % of protestants so that monotonicity cannot be satisfied.

4.2 Acemoglu et al. (2001)

Acemoglu et al. (2001) is a seminal paper in the persistence literature. The authors aim to measure the effect of institutions on economic performance using an IV model. The instrument for the quality of institutions — measured using an index of protection against expropriation — is settler's mortality in colonies in the first half of the 19th century. The

[7]Take the example of the Schmalkaldic League, established in 1531, in which religious motives were mixed with the political ambition of detaching from the Holy Roman Empire rule (Merriman, 2009).

[8]Citing Braudel, "As one would expect, a map of the cities closely corresponds to a map of the roads" (Braudel, 1972)

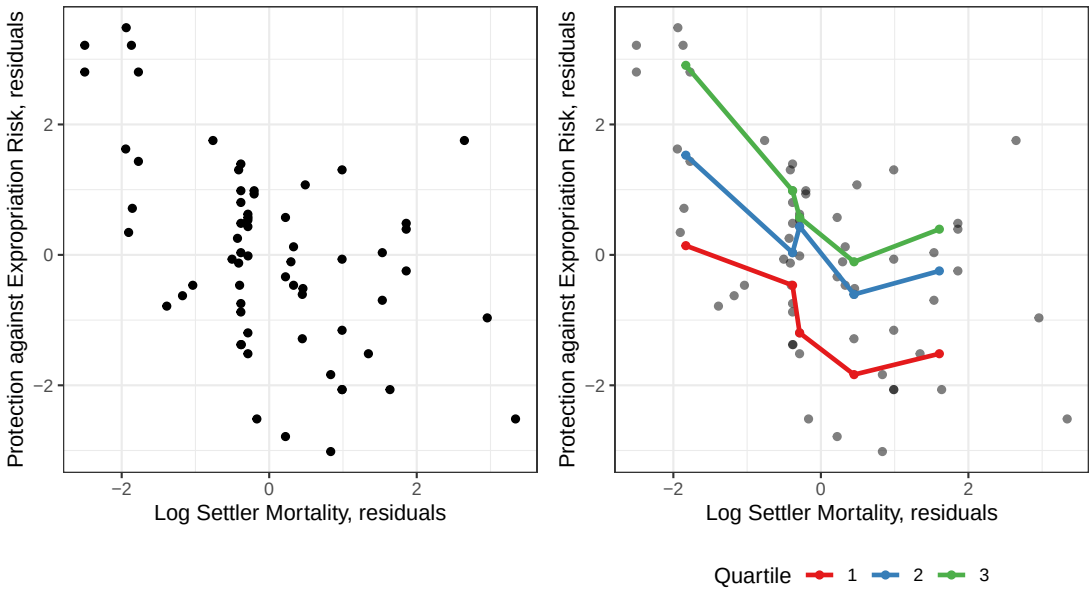

Figure 4: Scatterplots of the residuals for settlers' mortality and protection against expropriation risk. Each dot represents a country. On the right panel, colored lines show how quintiles of the treatment (protection against expropriation risk) evolve along the quintiles of the instrument (settlers' mortality). The countries are divided into four groups (corresponding to quartiles of the instrument), and quartiles of the treatment in each group are computed and connected by colored segments.

reason is that colonies with high settler mortality were more likely to be subject to extractive institutions, which could potentially harm economic prosperity in the long run.

Bisin and Moro (2021) cite this paper because the interpretation of institutional quality variable *"allows for different mechanisms connecting expropriation risk to economic performance, leading naturally to the possibility of heterogeneous effects, which depend on which mechanism is activated."*. Thus, in the same spirit as before, we will focus on the assumptions needed to allow for heterogeneous effects in this linear model with continuous variables.

Following the main specification in the paper by Acemoglu et al. (2001), we will use directly the model equations (1) and (2).

On the left panel, Figure 4 reports the scatterplot for the residualized instrument and treatment. On average, the protection against expropriation risk decreases with settlers' mortality. The right panel shows how the quartiles of the treatment change moving along

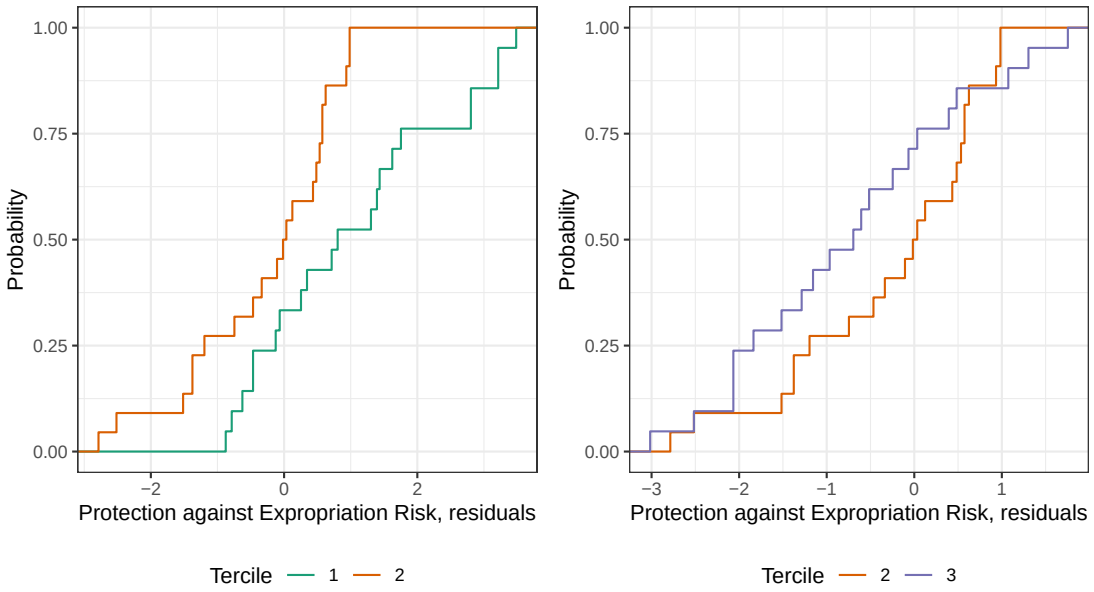

Figure 5: Plots of the empirical CDFs of the treatment (protection against expropriation risk) conditional on belonging to different terciles of the instrument (settlers' mortality). The countries are divided into three groups (corresponding to the terciles of the instrument) and then the empirical CDFs are plotted for selected groups.

the quartiles of the instrument. If the conditional stochastic monotonicity holds, quantiles should change monotonically. Again, it is not the case.

4.2.1 Testing the stochastic monotonicity in Acemoglu et al. (2001)

We run the test for stochastic monotonicity considering an asymptotic probability of rejection of the true null of 0.01, bootstrapping the test statistic from the empirical distribution 500 times. The null hypothesis of stochastic monotonicity is rejected.

4.2.2 Why might the test have failed?

To assess the validity of the monotonicity assumption, we need to ask the following: is it possible that an increase in the mortality of settlers in any country leads to a weakly-increase in institutional quality in that country? Are there any reasons why an increase in the mortality of settlers can lead to an *improvement* of institutions in some countries? We hypothesize

that this might happen, and thus the test fails, for two main reasons: first, the response of the treatment to the instrument is unlikely to have the same sign in "extractive" and "non-extractive" colonies. Second, the response may not be monotone for all the values of settlers' mortality.

Consider the following example to understand why the type of colonial regime may lead to a violation of monotonicity. The British colonial empire included both "extractive" (e.g. India, British Malaya) and "non-extractive" colonies (e.g. Canada, Australia, New Zealand, the United States). The theory presented in the paper views the nature of the colonial regime as a consequence of the (potential) settler mortality. This means that increased mortality in Australia would have led the country to a path of reduced institutional quality. However, the opposite may be true: deterioration of settler's life expectancy in Australia may have led the UK to mitigate and improve the situation by sending better colonial officers or doctors. This could cause an *increase* rather than a *decrease* of institutional quality.

On the other hand, in the "extractive" colonies the response of the colonizer to an increase in settler's mortality may have been the opposite, causing more oppressive exploitation of the resources in the colony and possibly a worsening of the institutions. The monotonicity is violated when in some places the treatment response to the instrument is positive, while in others negative.

Monotonicity does not also hold if the relationship between settlers' mortality and protection against expropriation risk is not monotone for all values of the instruments. A significant increase in settlers' mortality in a colony could lead a colonizer to back down from the original colonization plans. This can promote over time a better development, free from colonial domination. In this case, for some high levels of settlers' mortality, the effect on the protection against expropriation risk is actually *positive*.

Finally, note that, according to the theory presented by Acemoglu et al. (2001), the decision to establish an extractive versus a non-extractive colony is a *function* of the actual mortality in that territory. It means that the random assignment of the treatment assumption

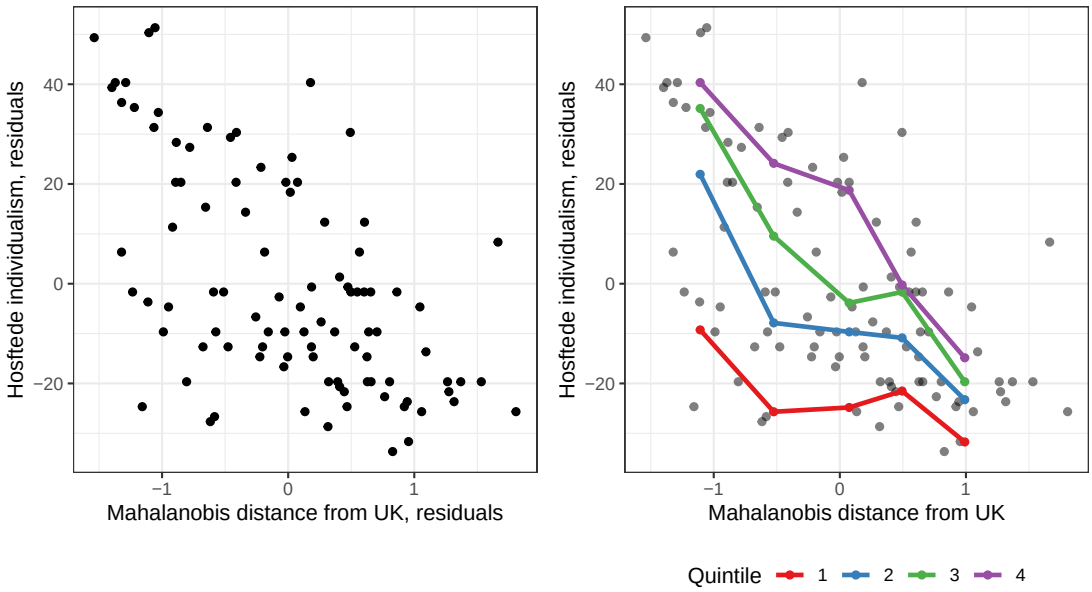

Figure 6: Scatterplots of the residuals for the Mahalanobis distance between the frequency of blood types in a given country and the frequency of blood types in the United Kingdom and Hofstede's index of individualism. Each dot represents the country. On the right panel, colored lines show how quintiles of the treatment (individualism) evolve along quintiles of the instrument (Mahalanobis distance from the UK). The countries are divided into five groups (corresponding to the instrument's quintiles), and each group's treatment's quintiles are computed and connected by colored segments.

is violated: the IV estimand is hard to interpret, as shown in part 2.2.1, and the test may reject for this reason.

4.3 Gorodnichenko and Roland (2017)

The paper by Gorodnichenko and Roland (2017) provides evidence that income disparities between countries can be attributed to *individualist culture*, because of its role in fostering innovation and thus growth. The main hypothesis of the paper is that in individualistic societies there is more innovation because, compared to collectivist societies, they provide higher social status. Thus people are more likely to spend time trying to make discoveries.

Gorodnichenko and Roland (2017) use a cross-country IV regression to test this relation between individualism and economic growth. The long-run growth is measured with the log of per capita income, while individualism is measured by Hofstede's index of individualism

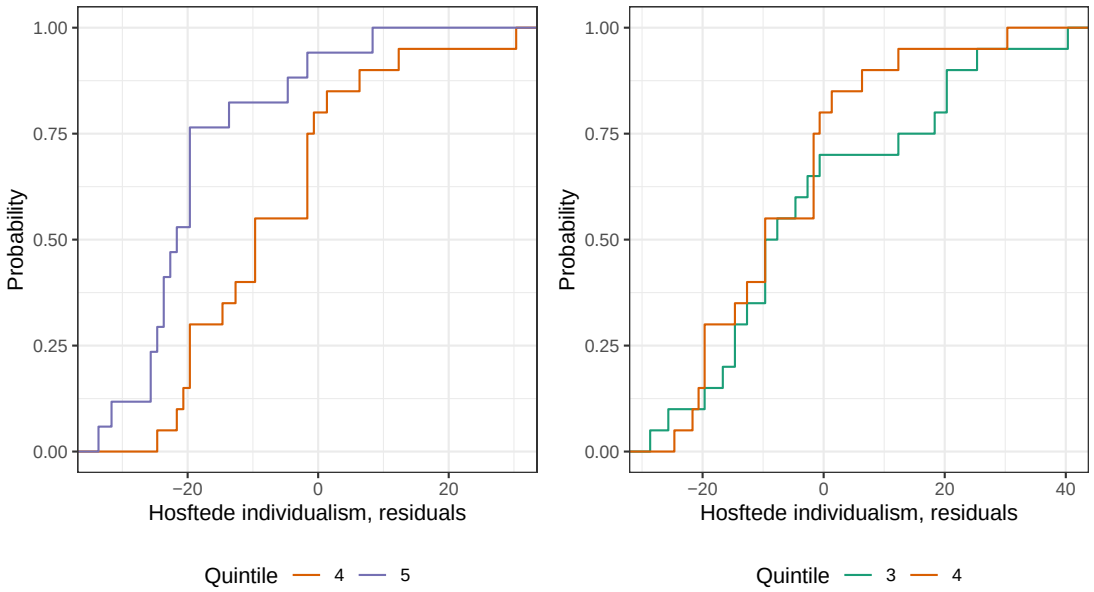

Figure 7: Plots of the empirical CDFs of the treatment (individualism) conditional on belonging to different quartiles of the instrument (Mahalanobis distance from the UK). The countries are divided into five groups (corresponding to quartiles of the instrument), and then the empirical CDFs are plotted for selected groups.

(Hofstede and Hofstede, 2001). The paper considers several instruments for individualism. In the main specification, which we replicate, the instrument is the Mahalanobis distance between the frequency of blood types in a given country and the frequency of blood types in the United Kingdom, which is the second most individualistic country in their sample.

The main specification in Gorodnichenko and Roland (2017) does not include control variables other than a constant; thus we proceed as in section 4.2. Figure 6 reports the instrument and control scatterplots. On the left panel, on average there is a negative relation between individualism and the Mahalanobis distance from the frequency of blood types in the United Kingdom. The right panel shows how the treatment's quantiles change moving along the instrument's quintiles. Again, since the relation is not monotone, we may expect a violation of the monotonicity assumption.

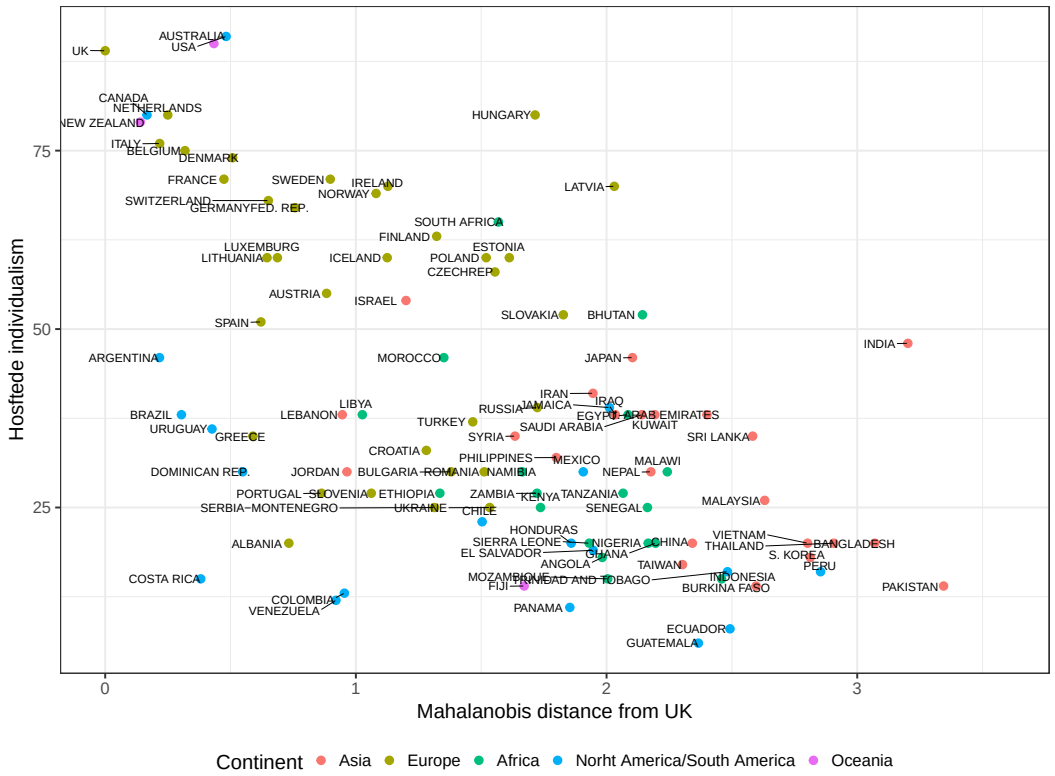

Figure 8: Scatterplot of the Mahalanobis distance between the frequency of blood types in a given country and the frequency of blood types in the United Kingdom and Hofstede's index of individualism by continent

4.3.1 Testing the stochastic monotonicity in Gorodnichenko and Roland (2017)

We run the test for stochastic monotonicity considering an asymptotic probability of rejection of the true null of 0.01, bootstrapping the test statistic from the empirical distribution 500 times. The null hypothesis of stochastic monotonicity is rejected.

4.3.2 Why might the test have failed?

As before, we will limit our discussion to the possible reasons behind the failure of the monotonicity assumption. Those reasons depend on the different colonial regimes countries were subject to, and to the connection between blood similarity and individualism.[9]

[9]However, we have serious doubts about the validity of the instruments in this case, given the high correlation of the instrument with the geographical distance from the UK or the "European" colonies of the British Colonial Empire, such as Australia, New Zealand, Canada and the USA. Running an IV regression using geographical distance from the UK returns almost the same point estimate for the second stage

Monotonicity fails when for some countries having a frequency of blood types more similar to the UK does not imply a higher level of individualism. Authors justify the instrument by saying that "parents transmit their genes as well as their cultural values to their offspring". Following this idea, it is possible that some countries became genetically closer to the UK and developed a more collectivist culture. In places where the British colonial regime has been "extractive", the response of the treatment to the instrument may be negative because the cultural values transmitted by parents could be antithetical to British individualism. As shown by Figure 8, the value of individualism is very similar in the UK and in Australia, USA, Canada and New Zealand. All of them were "non-extractive" British colonies. Other British colonies such as India, Kenya, and Nigeria have considerably lower values of individualism measure. Why in such places an increase in blood types similarity with the UK would lead to an increase of individualism?

Following again the hypothesis of the transmission of genes and cultural values connecting instrument and treatment, one can argue that the decrease of the Mahalanobis distance positively affects individualism because such a decrease is necessarily accompanied by all other values that characterize the UK. Using British colonies as an example, if we could give Nigeria the same Mahalanobis distance value as New Zealand, we would have to see an increase in individualism in Nigeria *because* it means that more European colonists are settling in Nigeria and the colonial regime becomes more similar to that of New Zealand. However, this would violate the exclusion restriction of the IV instrument, as the instrument affects other (endogenous) variables that are likely to affect income per capita directly.

5 Conclusion

Bisin and Moro (2021) argue that heterogeneity in treatment effects is a common feature

regression. We suspect the Mahalanobis distance from the frequency of blood types in the United Kingdom is capturing the "North" of the World, see Figure 8. Finally, abstracting from historical considerations, it is not very plausible from a biological point of view that a similar frequency of blood types with respect to the UK must necessarily lead to an expected increase of individualism in people and countries.

of many Economic History applications. Models with random parameters, that allow for heterogeneous effects, are hence more general and plausible than the counterpart with constant parameters. We formalized the heterogeneous treatments model for the widely used linear IV model with continuous variables, where we allowed for heterogeneous responses in both the first and second stage. Even in such a simple linear model, allowing for heterogeneous treatment effects pose problems in interpreting the regression coefficient of interest, β^{IV}. We showed that a *monotonicity* assumption on the heterogeneous effects in the first stage is needed to have a meaningful estimand. Moreover, we derived a testable implication for this model based on the cumulative density function of the treatment and the instrument.

We tested the monotonicity condition for three papers in the Economic History literature that use linear IV models with continuous variables: Becker and Woessmann (2009); Acemoglu et al. (2001); Gorodnichenko and Roland (2017). In all cases, the monotonicity assumption is rejected by a test with 1% asymptotic significance level. We provided some possible explanations for these rejections, rooted in the historical context of each application.

These results call for some caution in using the IV model with continuous variables. We advocate for a proper discussion of the model assumptions in applications. The discussion can be supported, but not substituted, by the robustness test we proposed.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics 113*(2), 231–263.

Acemoglu, D., S. Johnson, and J. A. Robinson (2001). The colonial origins of comparative development: An empirical investigation. *American economic review 91*(5), 1369–1401.

Alesina, A., P. Giuliano, and N. Nunn (2013). On the origins of gender roles: Women and the plough. *The quarterly journal of economics 128*(2), 469–530.

Angrist, J. D. and G. W. Imbens (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association 90*(430), 431–442.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Ashraf, Q. and O. Galor (2011). Dynamics and stagnation in the malthusian epoch. *American Economic Review 101*(5), 2003–2041.

Ashraf, Q. and O. Galor (2013). The'out of africa'hypothesis, human genetic diversity, and comparative economic development. *American Economic Review 103*(1), 1–46.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*(439), 1171–1176.

Barrett, G. F. and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica 71*(1), 71–104.

Becker, S. O. and L. Woessmann (2009). Was weber wrong? a human capital theory of protestant economic history. *The quarterly journal of economics 124*(2), 531–596.

Bisin, A. and A. Moro (2021). Late for history. In *The Handbook of Historical Economics*, pp. 269–296. Elsevier.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022). When is tsls actually late? Technical report, National Bureau of Economic Research.

Braudel, F. (1972). *The Mediterranean and the Mediterranean World in the Age of Philip II: Volume I* (1ST ed.).

Caicedo, F. V. (2021). Historical econometrics: instrumental variables and regression discontinuity designs. *The Handbook of Historical Economics*, 179–211.

Calomiris, C. W. and J. R. Mason (2003). Consequences of bank distress during the great depression. *American Economic Review 93*(3), 937–947.

Cantoni, D. and N. Yuchtman (2021). Historical natural experiments: Bridging economics and economic history. In *The handbook of historical economics*, pp. 213–241. Elsevier.

Casey, G. and M. Klemp (2021). Historical instruments and contemporary endogenous regressors. *Journal of Development Economics 149*, 102586.

Chetverikov, D., A. Santos, and A. M. Shaikh (2018). The econometrics of shape restrictions. *Annual Review of Economics 10*, 31–63.

Chetverikov, D. and D. Wilhelm (2017). Nonparametric instrumental variable estimation under monotonicity. *Econometrica 85*(4), 1303–1320.

Chetverikov, D., D. Wilhelm, and D. Kim (2021). An adaptive test of stochastic monotonicity. *Econometric Theory 37*(3), 495–536.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature 48*(2), 424–455.

Delgado, M. A. and J. C. Escanciano (2012). Distribution-free tests of stochastic monotonicity. *Journal of Econometrics 170*(1), 68–75.

Dittmar, J. E. (2011, 08). Information Technology and Economic Change: The Impact of The Printing Press *. *The Quarterly Journal of Economics 126*(3), 1133–1172.

Fiorini, M. and K. Stevens (2021). Scrutinizing the monotonicity assumption in iv and fuzzy rd designs. *Oxford Bulletin of Economics and Statistics*.

Gorodnichenko, Y. and G. Roland (2017). Culture, institutions, and the wealth of nations. *Review of Economics and Statistics 99*(3), 402–416.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of human resources*, 441–462.

Heckman, J. J. and S. Urzua (2010). Comparing iv with structural models: What simple iv can and cannot identify. *Journal of Econometrics 156*(1), 27–37.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica 73*(3), 669–738.

Hofstede, G. and G. Hofstede (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. SAGE Publications.

Imbens, G. W. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic literature 48*(2), 399–423.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica 83*(5), 2043–2063.

Lee, D. S., J. McCrary, M. J. Moreira, and J. Porter (2022, October). Valid t-ratio inference for iv. *American Economic Review 112*(10), 3260–90.

Lee, S., O. Linton, and Y.-J. Whang (2009). Testing for stochastic monotonicity. *Econometrica 77*(2), 585–602.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005). Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies 72*(3), 735–765.

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2019). Instrumental variables and the sign of the average treatment effect. *Journal of Econometrics 212*(2), 522–555.

McFadden, D. (1989). Testing for stochastic dominance. In *Studies in the Economics of Uncertainty*, pp. 113–134. Springer.

Merriman, J. (2009). *A History of Modern Europe: From the Renaissance to the Present* (3rd ed.). W. W. Norton & Company.

Seo, J. (2018). Tests of stochastic monotonicity with improved power. *Journal of Econometrics 207*(1), 53–70.

Słoczyński, T. (2022). When should we (not) interpret linear iv estimands as late? *arXiv preprint arXiv:2011.06695*.

Spolaore, E. and R. Wacziarg (2013, 06). How deep are the roots of economic development? *Journal of Economic Literature 51*(2), 325–69.

Whang, Y.-J. et al. (2019). Econometric analysis of stochastic dominance. *Cambridge Books*.