The in-principle inconclusiveness of causal evidence in macroeconomics[1]

1. Introduction

Kevin Hoover's sophisticated work *Causality in Macroeconomics* begins by stating an undeniable truth: the ultimate justification for the study of macroeconomics is to provide knowledge[2] on which to base policy; policy is about influencing outcomes, about control or attempted control; and the study of the particular connections that permit control of one thing to influence another is the study of causality (cf. Hoover, 2001, p. 1). Knowledge on which to base macroeconomic policy requires that (*a*) there be aggregate quantities that can be manipulated for policy purposes, that (*b*) there be causal structures that connect these quantities with some target quantity, and that (*c*) there be evidence in support of our belief that (*a*) and (*b*) hold. Is macroeconomics capable of providing knowledge of that kind? Hoover's answer to that question is clearly positive. He holds that (i) causal structure or "causality is a feature of the world (or the economy)" (Hoover, 2001, p. 59), that (ii) the aggregate quantities that fill the positions in causal structures exist (Hoover, 2001, chap. 5), and that (iii) empirical data can be used to infer causal structure (Hoover, 2001, pp. 59, 213-4).

The present paper, by contrast, will provide a negative answer. It's going to defend that answer by arguing against (iii) that causal structure cannot be inferred. Causal structure cannot be inferred because the evidence provided by the causal inference methods that can be used in macroeconomics is in principle too inconclusive to turn the belief that *X* directly type-level causes *Y* into knowledge. This evidence is too inconclusive because it derives from the conditions of the IV method, i.e. from conditions requiring that there be no confounders of *I* and *X* and *X* and *Y* (where *I* is an instrumental or intervention variable that type-level causes *X*), because in macroeconomics, confounders that cannot be controlled for or measured are likely to be present, and because econometric causality tests can be shown to rely on the conditions of the IV method at least tacitly. The argument against (iii) has a bearing for (i): if it applies, then (i) becomes problematic. But the paper is not going to deal with the kind of causal realism expressed by (i). The paper is not going to deal with (ii) in any detailed manner either. It is just going to point out that there are macroeconomic aggregates of which we cannot know whether they are capable of manipulation for policy purposes (aggregate quantities like inflation expectations or decisions of firms to shrink or expand production).

---

[1] Acknowledgments …

[2] Hoover in fact speaks of "secure" knowledge on which to base policy. The term "secure" is suppressed here because what needs to be seen in the first place is whether in macroeconomics, there can be any knowledge on which to base policy at all.

Of the econometric causality tests that macroeconomists can use to provide evidence in support of causal hypotheses, the paper will be dealing with exactly two: with the procedure designed by Hoover (2001, chaps. 8-10) and with the procedure that comes along with the potential outcome approach that Angrist and Kuersteiner (2011) have introduced into macroeconomics more recently. It is true that on occasion, macroeconomists carry out Granger causality and super exogeneity tests to provide evidence in support of causal hypotheses: testing procedures designed by Granger (1969) and Hendry (1988) and Engle and Hendry (1993). But both procedures have received a lot of attention in the literature and do not test for causality in the strict sense of the term. Hoover (2001, pp. 151-155, 167-8) shows that Granger causality is neither necessary nor sufficient, and that super exogeneity is not necessary for what he understands by 'macroeconomic causality'.[3] And Hoover's understanding of macroeconomic causality can be shown to be adequate.[4]

The paper will begin by examining two possible instantiations of the IV-method: randomized controlled trials (RCTs) and natural experiments. Section 2 is going to look at an RCT, as it is typically conducted in microeconomics, and at the fictitious case of an RCT conducted in macroeconomics. Section 3 will analyze the famous natural experiment that Friedman and Schwartz (1963) observe to test their hypothesis that monetary changes directly type-level cause economic changes. It is going to argue that the evidence provided by that experiment is too inconclusive because it derives from the conditions of the IV method, and because confounders that violate these conditions and cannot be controlled for or measured are likely to be present in that experiment. Section 4 will defend the general case. Sections 5 and 6 are going to deal with the econometric procedures that Hoover (2001, chs. 8-10) and Angrist and Kuersteiner (2011) propose to test for causal hypotheses in macroeconomics. These sections will argue that the evidence provided by these tests (the 'Hoover test' and the 'AK test') is too inconclusive because they tacitly rely on the conditions of the IV method. The final section 7 will summarize the main argument and argue against Friedman (1953) that the inability to conduct RCTs reflects a basic difference between macroeconomics and many of the other special sciences.

---

[3] According to Granger (1980, p. 339), the basic idea of the definition of Granger causality is that "knowledge of the causal variable helps forecast the variable being discussed". And in their seminal paper on exogeneity in econometrics, Engle, Hendry, and Richard (1983, p. 384) point out that "super exogeneity is a sufficient but not a necessary condition for valid inference under policy interventions". If causality is a necessary and sufficient condition for policy analysis, these statements suggest that Granger causality and super exogeneity tests are not even meant to test for causality in the strict sense of the term.

[4] I should mention that the present paper builds substantially on a companion paper (Henschen 2018) which argues that a macroeconomic variant of Woodward's interventionist account qualifies as an adequate account of macroeconomic causality, and that this account is equivalent to Hoover's account, as long as the notion of an intervention is not restricted to parameter manipulations.

2. Randomized controlled trials (RCTs)

Imagine we would like to find out whether activation programs (programs consisting of job search activities, intensive counseling and job training) directly type-level cause the rate of exit from unemployment. We won't be able to find out about this relationship if we assign all unemployed individuals to an activation program and check, after a certain period of time, whether the exit rate has changed. The exit rate might, after all, change as a result of the influence of all sorts of causes (including an increase or decrease in economic activity). In order to find out about the relationship, we will have to proceed in roughly two steps: we will first have to select a randomizing procedure that assigns roughly half of the unemployed to a group of individuals undergoing the activation program (the treatment group) and the other half to a group of individuals not undergoing that program (the control group). The principal purpose of this randomizing procedure is to avoid bias resulting from the influence of confounders, i.e. of possible type-level causes of which we don't know anything: perhaps the long-term unemployed (men; 50+ etc.) are less inclined to exit unemployment than the newly unemployed (women; 30- etc.); results would accordingly be biased if both groups contained long-term unemployed (men; 50+) etc. in unequal numbers.

In a second step, we will have to control for type-level causes that we think are likely to bias the results of our experiment. We will have to make sure, for instance, that once individuals are assigned to the treatment and control groups, participation in the program is mandatory. Otherwise individuals that were assigned to the program and have a low inclination to exit unemployment might choose to withdraw from the program. We will also have to rule out that the unemployment insurance agencies pay special attention to the individuals in the treatment group. Otherwise they could revert to every possible means to increase the exit rate (they, after all, have a substantial interest in getting a positive evaluation for their program and in receiving more public funds). We will finally have to make sure that the individuals in the treatment and control groups believe that they belong to the same group. Otherwise the (psychologically well-understood) role obligations of being an experimental subject might bias the results of our experiment.[5]

If both steps are taken, we will be dealing with a binary intervention variable *I* that is set to 'yes' for the individuals assigned to the treatment group and to 'no' for the individuals assigned to the control group, and that is likely to satisfy the following set of conditions:

(*I*1) *I* type-level causes *X*,

[5] A study that comes close to a full implementation of that two-step procedure is an experiment that Berg and Klaauw (2006) conduct for two Dutch cities and a time-interval between 08/1998 and 02/1999. The only drawback of that study is that it fails to control for role obligations. Its main finding is that the exit rate in the monitored treatment group wasn't significantly higher than in the control group.

(*I*2) certain values of *I* are such that when *I* attains these values, *X* is no longer determined by other variables that type-level cause it but only by *I*,

(*I*3) any directed path from *I* to *Y* goes through *X*, and

(*I*4) *I* is statistically independent of any variable *Z* that type-level causes *Y* and is on a directed path that does not go through *X*.

I is likely to satisfy (*I*1) because *I* is likely to type-level cause *X* (activation program). *I* is likely to satisfy (*I*2) because *I* is likely to break the arrow that is directed into *X* and departs from a variable standing for voluntary participation. *I* is likely to satisfy (*I*3) because *I* is likely to break the directed paths from *I* to *Y* that don't go through *X* but through the variables standing for the special attention paid to the individuals in the treatment group and the role obligations of being an experimental subject. And *I* is likely to satisfy (*I*4) if the sample of the unemployed assigned to the treatment and control groups is large enough: if it is large enough, then *I* is likely to be probabilistically independent of any (nuisance) variable *Z* that type-level causes *Y* and is on a directed path that doesn't go through *X*.

Conditions (*I*1) – (*I*4) are the conditions that Woodward (2003: 98) spells out to define the term 'intervention variable'. The term 'intervention variable' figures in his definition of the term 'intervention'. And the term 'possible intervention' is used to define the term 'direct type-level cause' (cf. Woodward, 2003, pp. 55, 59). One may accordingly say that *X* directly type-level causes *Y* if *I* satisfies conditions (*I*1) – (*I*4), if there is a possible intervention on *X* that changes *Y* or its probability distribution, and if all direct type-level causes of *Y* except *X* remain fixed by intervention.[6] One may further say that *X* is likely to directly type-level cause *Y*, i.e. that we may believe with some confidence that *X* directly type-level causes *Y*, if *I* is likely to satisfy conditions (*I*1) – (*I*4), if there is likely to be a possible intervention on *X* that changes *Y* or its probability distribution, and if all direct type-level causes of *Y* except *X* are likely to remain fixed by intervention.

There are, of course, important criticisms that have been advanced against the use of RCT methodology in economics (cf. Reiss, 2013, pp. 202-206): (a) randomization ensures that treatment and control groups are identical with respect to all confounders only in the limit; (b) in economics, neither subjects nor experimenters can be blinded; (c) RCTs may introduce new confounders; (d) there is no guarantee that RCTs generalize to other settings. But while (d) relates to a difficult problem that all social policy faces (to the problem of the external validity of experiments), there are effective means of dealing with criticisms (a) – (c) in labor

[6] Woodward's definition in fact requires that all variables in a variable set *V*, except *X* and *Y*, remain fixed by intervention. But this requirement is meant to ensure that *X* is a *direct* type-level cause of *Y*. And as such, it can be replaced with the weaker requirement that all direct type-level causes of *Y* except *X* remain fixed by intervention (a requirement that can be found e.g. in Pearl, [2]2009, pp. 127-8). Cf. Henschen (2018, p. 9) for an elaboration of this point.

economics: the random sample drawn from the population of unemployed individuals can be large enough to ensure that treatment and control groups are at least nearly identical with respect to possible confounders; role obligations can be controlled for by lying to the individuals in the control group that they would undergo a treatment too (which is, of course, ethically questionable but not impossible in principle); and newly introduced confounders (like withdrawals from the program or the special attention paid to the individuals in the treatment group) can be controlled for by rendering participation mandatory, and by creating several treatment groups of which only one is monitored.

Imagine next that we would like to find out whether the real interest rate is a direct type-level cause of aggregate demand, and that the following canonical new Keynesian model (cf. Romer, [4]2012, pp. 352-3) expresses the hypothesis that we've formed about the relations of direct type-level causation that obtain among aggregate quantities in the economy in question:

(1) $y_t = E_t[y_{t+1}] - r_t/\theta + u_t^{IS}$, $\quad \theta > 0$,

(2) $\pi_t = \beta E_t[\pi_{t+1}] + \kappa y_t + u_t^{\pi}$, $\quad 0 < \beta < 1, \quad \kappa > 0$,

(3) $r_t = \varphi_\pi E_t[\pi_{t+1}] + \varphi_y E_t[y_{t+1}] + u_t^{MP}$ $\quad \varphi_\pi > 0, \quad \varphi_y \geq 0$,

where Y_t (in logarithm) represents aggregate demand, $E_t[Y_{t+1}]$ expectations in t of aggregate demand in t+1, R_t the real interest rate, Π_t (in logarithm) the rate of inflation, $E_t[\Pi_{t+1}]$ expectations in t of the rate of inflation in t+1, where U_t^{IS}, U_t^{π} and U_t^{MP} represent shocks to aggregate demand, inflation and the real interest rate, respectively, and are assumed to follow independent first-order autoregressive processes, and where the various parameters are identified in microeconomic theory: β and θ in the utility function of the representative household, κ in the price-setting behavior of the representative firm, and φ_π and φ_y in the (forward-looking) interest-rate rule followed by the central bank (cf. Romer, [4]2012, pp. 315-6, 329-31).[7] The model is obviously stylized (the dynamics of aggregate demand and inflation are very simple, the empirical performance of (2) is poor, everything is linear, all behavior is forward-looking etc.). But it is also "canonical" in the sense that it serves as a key reference point in macroeconomic dynamic-stochastic general-equilibrium (DSGE) modeling. Various modifications and extensions of it are used in central banks and other policymaking institutions.

[7] Throughout the paper, uppercase letters will be reserved for variables and lowercase letters for their values. This convention is a bit unusual for (macro-) economists but widespread in the philosophical literature on causality.

How are we supposed to find out whether the real interest rate (R_t) is a direct type-level cause of aggregate demand (Y_t)? If we were supposed to find out about that relationship by conducting an RCT, we would first draw a random sample from a population of economic systems with reserve or central banks. We would secondly use randomization techniques to assign the systems in the sample to a treatment and a control group. We would thirdly ask governments (or any other responsible and competent bodies) to control for demand and inflation expectations because we believe them to directly type-level cause aggregate demand or the real interest rate. We would fourthly ask the central banks of the systems in the treatment group to take measures that increase or decrease R_t in (3). We would finally check whether the ensuing change in the real interest rate would be followed by a change in aggregate demand only in the treatment group.

Why is it that we would never carry out any RCT like that? The first reason is that we cannot know whether we can control for demand and inflation expectations through direct human intervention. In the example from labor economics, the variables that are believed to type-level cause *X* or *Y* can be controlled for through direct human intervention. In the fictitious RCT described above, by contrast, we cannot know whether some of the variables that are believed to type-level cause *X* or *Y* (demand and inflation expectations) can be controlled for through direct human intervention. We cannot know whether demand and inflation expectations can be controlled for because the information on which their formation is based might already include the information that there is an attempt to control for them. An attempt to control for demand expectations resembles a government's attempt to sugarcoat the indicators of economic activity. Economic agents won't be taken in by such an attempt as long as their expectations are formed rationally. An attempt to control for inflation expectations, by contrast, resembles a central bank's attempt to "anchor" inflation expectations, i.e. to influence inflation expectations in a way that renders them largely invariant over time or even invariant to monetary policy interventions. And there is considerable disunity among macroeconomists whether or not a central bank can influence inflation expectations in this way.

Note that even if we were capable of controlling for demand and inflation expectations, there would still be no way to find out about that capability. There would be no way to find out about that capability because expectations variables cannot be measured. It is true that researchers sometimes use survey data to provide evidence for or against that capability.[8] But Hoover (2001, p. 137) explains why survey data cannot be taken to provide values for expectation variables. He suggests that expectations fall into the same category as

[8] Beechey et al. (2011), for instance, use survey data to show that inflation expectations are anchored in the Euro area and (less firmly) the USA. Afrouzi et al. (2015), by contrast, use survey data to show that inflation expectations aren't anchored in New Zealand.

preferences. In revealed-preference theory, a consumer's preference is reconstructed from her behavior (from her "revealed" preference). Statements about what she thinks she prefers are to be dismissed as neither verifiable nor trustworthy.[9] Similarly, expectation variables cannot be measured because a subject's statement about what she expects can neither be verified nor trusted. Expectation variables therefore need to be solved out, as in the case of the rational expectations model that Hoover discusses at various points in his work (cf. especially Hoover, 2001, pp. 64-66), or interpreted as attaining whatever value is required to render a model (like the canonical new Keynesian model above) consistent.

But there is a second reason why we would never carry out an RCT to find out whether R_t is a direct type-level cause of Y_t. In the example from labor economics, the random sample drawn from the population of unemployed individuals can be large enough to ensure that treatment and control groups are at least nearly identical with respect to possible confounders. In the case of our fictitious trial, by contrast, the random sample that we would draw from the population of economic systems with a central bank is necessarily small. Our sample might include Bolivia, Moldova, Slovakia, Thailand, the USA, and Zimbabwe, and some randomizing procedure (like flipping a coin) might assign Moldova, Thailand, and Zimbabwe to the treatment, and Bolivia, Slovakia, and the USA to the control group. Even if institutions were able and willing to fully control for demand and inflation expectations and to change nominal interest rates as requested, and even if a change in the real interest rate were followed by a change in aggregate demand only in the treatment group, there would be no guarantee that the change in aggregate demand is attributable to the change in the real interest rate. We would rather take the change in aggregate demand as an invitation to search for unknown type-level causes of aggregate demand that are present in (perhaps only one or two of the systems in) the treatment group but not in the control group. And we wouldn't be surprised if we learned that the values of variables standing e.g. for investment, government purchases or net exports in the systems of the treatment group were significantly different from the values of those variables in the systems of the control group.

3. A natural experiment in macroeconomics

The famous natural experiment that Friedman and Schwartz (1963) observe can be read as an attempt to avoid the two difficulties. A natural experiment is by definition an experiment in which control over the variables that potentially type-level cause *X* and *Y* isn't exercised by direct human intervention but by nature. And a peculiarity about Friedman and Schwartz's experiment is that it turns from a plurality of economic systems monitored during a particular

[9] Sen (1973: 242) provides a long list of quotations that testify to the worries about introspection and verifiability that motivated revealed-preference theory.

time interval to a plurality of time intervals during which one particular system is monitored. Friedman and Schwartz (1963, p. 676) argue, more specifically, that a look at the monetary history of the USA from 1867 to 1960 teaches three things: that "[c]hanges in the behavior of the money stock have been closely associated with changes in economic activity", that "[t]he interrelation between monetary and economic change has been highly stable", and that "[m]onetary changes have often had an independent origin; they have not been simply a reflection of changes in economic activity." The most important case of what Friedman and Schwartz (1963, p. 692) believe is an independent occurrence of a monetary change is the contraction of the money stock that followed the death of Benjamin Strong (the president of the Federal Reserve Bank of New York) in 1928.

With respect to Strong's death, Friedman and Schwartz (1963, p. 693) speak of a "quasi-controlled experiment." Friedman and Schwartz rarely employ the terms 'cause' or 'causes' in the *Monetary History*, and Friedman even explicitly disapproves of the use of these terms.[10] But Hoover (2009, p. 306) points out that the *Monetary History* is full of causatives (terms like 'influences', 'increases', 'engenders', 'affects' etc.). It is also clear that Friedman and Schwartz aim to derive a policy conclusion: the conclusion that contractionary (expansive) monetary policies lead to monetary contractions (expansions), and that monetary contractions (expansions) lead to economic contractions (expansions). Their quasi-controlled experiment may accordingly be interpreted as an experiment that is meant to provide evidence in support of a causal hypothesis: the hypothesis that monetary changes directly type-level cause economic changes.

A closer look at their argument (Friedman and Schwartz, 1963, especially pp. 686-695) reveals, moreover, that the experiment that is meant to provide evidence in support of that hypothesis can be regarded as an experiment in which a binary intervention variable *I* (contractionary monetary policy: yes or no) is set to 'yes' or 'no' by a procedure (Strong's death) that is at least quasi-randomizing: in which *I* is set to 'yes' in the USA for the period from 01/1929 to 03/1932, and to 'no' in the USA for most other periods,[11] and in which an important economic contraction is observed only for the period from 01/1929 to 03/1932. That contraction, Friedman and Schwartz (1963, p. 694) conclude is "strong evidence for the economic independence of monetary changes from the contemporary course of income and prices" and thus for the hypothesis that monetary changes directly type-level cause economic changes.

---

[10] Perhaps under the influence of Popper or positivism, Friedman says that he tries "to avoid the use of the word 'cause' … it is a tricky and unsatisfactory word" (cited from Hoover, 2009, p. 306).

[11] Friedman and Schwartz (1963, pp. 688-9) identify only two further short periods that were also characterized by contractionary monetary policies and associated contractions in the money stock and industrial production.

That conclusion is an overstatement, however. In order to see why, consider the case of rational expectations and the case that King and Plosser (1984) make to support their hypothesis of "reverse causation". In the case of rational expectations, agents make the best use of whatever information is available to them to form expectations of key variables (such as money supply, GDP, and prices) in a manner consistent with the way the economy actually operates. A typical rational expectations model (cf. e.g. Dornbusch, Fischer, and Startz, [7]1998, 166-168) predicts that monetary changes directly type-level cause economic changes unless agents fully anticipate the monetary policy measures leading to the monetary changes. In the case of rational expectations, one may accordingly say that *I* doesn't only type-level cause *X* (monetary contraction: yes or no) but also *Z*, while *Y* (economic contraction: yes or no) isn't only type-level caused by *X* but also by *Z*, where *Z* denotes rational expectations under full anticipation of monetary policy changes (yes or no). In this case, *I* satisfies conditions (*I*1) and (*I*2) but not conditions (*I*3) and (*I*4): *I* type-level causes *X* and breaks any other arrow that is directed into *X*; but there is also a variable *Z* that type-level causes *Y*, is on a directed path that doesn't go through *X*, and is correlated with *I* because *I* type-level causes *Z*.

King and Plosser (1984), by contrast, argue that aggregate measures of the money stock (such as *M*2) aren't set directly by the Federal Reserve but are determined by the interaction of the supply of high-powered money with the behavior of the banking system and the public, and that changes in the values of both the money stock and aggregate output result from the decision of firms to shrink production and to decrease their money holdings accordingly. If they are right, then *I* doesn't satisfy any of conditions (*I*1) – (*I*4): *I* doesn't satisfy (*I*1) because *I* doesn't type-level cause *X* at all; *I* doesn't satisfy (*I*2) because *X* isn't determined by *I* at all; and *I* doesn't satisfy (*I*3) and (*I*4) because (*I*3) and (*I*4) are vacuous (there isn't any directed path from *I* to *Y*).

The important economic contraction that Friedman and Schwartz observe for the period from 01/1929 to 03/1932 therefore cannot represent strong evidence in support of the hypothesis that monetary changes directly type-level cause economic changes. In the case of rational expectations, it's conceivable that rational agents who had a great deal at stake fully understood the "active phase of conflict" that Friedman and Schwartz (1963, p. 692) argue was unleashed by Strong's death, that these agents correctly anticipated the contractionary monetary policy that was characteristic of that phase, and that the monetary contraction caused by that policy therefore didn't cause the economic contraction. And in King and Plosser's case of "reverse causation", it is not implausible that policy measures didn't play any major role, and that it was a general pessimistic outlook that led firms to decide to shrink

production, i.e. to increase their money holdings (thereby inducing the monetary contraction) and to reduce production (thereby effectuating the economic contraction).

The economic contraction that Friedman and Schwartz observe for the period from 01/1929 to 03/1932 represents a piece of evidence that is too inconclusive to disentangle a set of competing and observationally equivalent hypotheses: the hypothesis that *X* directly type-level causes *Y*, the hypothesis that *Y* directly type-level causes *X* (King and Plosser's hypothesis of "reverse causation") and the hypothesis that there is a variable (or set of variables) *Z* that directly type-level causes both *X* and *Y* (the hypothesis that obtains in the case of rational expectations). In Friedman and Schwartz's (1963, p. 686) discussion, the observational equivalence of these hypotheses is expressed as follows: "The monetary changes might be dancing to the tune called by independently originating changes in the other economic variables; the changes in income and prices might be dancing to the tune called by independently originating monetary changes; […] or both might be dancing to the common tune of still a third set of influences." Friedman and Schwartz go on to claim that "a wide range of qualitative evidence […] provides a basis for discriminating between these possible explanations of the observed statistical covariation". The above considerations suggest, however, that the "wide variety of qualitative evidence" is wanting. And if it is wanting, then the three competing hypotheses cannot be disentangled.

The three competing hypotheses could be disentangled if *Z* could be controlled to 'no' for the period from 01/1929 to 03/1932 or to identical values for all periods between 1867 and 1960; if randomization techniques could be applied to ensure that *Z* is evenly distributed over all these periods; or if the decisions of firms to shrink or expand production could be controlled for in effective ways. It is unclear, however, whether rational expectations or decisions of firms to shrink or expand production can be controlled for through direct human intervention. And the time periods between 1867 and 1960 are probably just as diverse and small in number as the economic systems that could be assigned to treatment and control groups in a fictitious RCT like the one considered in the preceding section (remember that two world wars and the Great Depression occurred between 1867 and 1960). Therefore, the three competing hypotheses are bound to remain entangled.

Romer and Romer (1989) attempt to provide evidence along similar lines as Friedman and Schwartz. They search the records of the Federal Reserve for the postwar period to find evidence of policy shifts that were designed to lower inflation, not motivated by developments on the real side of the economy, and followed by recessions. They identify six such shifts, the most prominent being the monetary contraction that occurred shortly after Paul Volcker became chairman of the Federal Reserve Board in October 1979, and that was followed by one of the largest recessions in postwar US history. Romer and Romer (1989) argue that the

monetary contraction was motivated by a desire to reduce inflation, and not by the presence of other forces that would have caused output to decline in any event. But their argument remains open to the sort of objections that can be raised in the case of rational expectations and in King and Plosser's case of "reverse causation". In the case of rational expectations, it's conceivable that rational agents who had a great deal at stake fully understood that Volcker was going to fight inflation, that these agents correctly anticipated the contractionary monetary policy that ensued shortly after Volcker became chairman of the Federal Reserve Board, and that the monetary contraction caused by that policy therefore didn't cause the economic contraction. And in King and Plosser's case of "reverse causation", it is not implausible that the contractionary monetary policy didn't play any major role, and that it was a general pessimistic outlook that led firms to decide to shrink production, i.e. to increase their money holdings (thereby inducing the monetary contraction) and to reduce production (thereby effectuating the economic contraction).

4. The general case

The conclusion to be drawn states that the evidence deriving from these natural experiments is too inconclusive to turn the belief that monetary changes directly type-level cause economic changes into knowledge. Note that this conclusion holds in principle, and not just in the case of these experiments. It holds in principle because hidden variables operate whenever there is an attempt to control for an intervention variable *I*. A hidden variable, as I understand it, is a variable that denotes some macroeconomic aggregate, that cannot be measured, that might be incapable of manipulation through (policy or experimental) intervention, and that type-level causes *Y*.

King and Plosser (1984, p. 363) refer to their hypothesis as one of "reverse causation". But their hypothesis may also be read as one of confounding: the decisions of firms to shrink or expand production act as a common cause of both monetary and economic changes. These decisions undoubtedly cannot be measured. We might conduct a survey and ask firms how much they think they are going to produce so and so many quarters ahead. But as in the case of demand or inflation expectations (or expectation variables more generally), responses are to be dismissed as neither verifiable nor trustworthy. We therefore won't be able to find out whether these decisions are capable of manipulation through (policy or experimental) intervention.

In King and Plosser's case of "reverse causation", the hidden variable in question (i.e. decisions of firms to shrink or expand production) is assumed to be causally independent of *I* (i.e. of monetary policy interventions). But one of the lessons of the Lucas critique states that interventions on *I* always token-level cause changes in hidden variables. Considered

generally (cf. Lucas, 1976, p. 25), the Lucas critique says that the function connecting *X* and *Y* is "derived from decision rules […] of agents in the economy", that "some view of the behavior of the future values of variables of concern to them […], in conjunction with other factors, determines their optimum decision rules", and that the assumption that this view remains invariant under alternative policy rules is an "extreme assumption". If "some view of the behavior of the future values of variables of concern to them" summarizes expectations of aggregate demand, GDP, inflation and so on, if variables denoting these expectations are hidden in the sense indicated above, and if an "alternative policy rule" amounts to a policy manipulation of an intervention variable *I* that type-level causes *X*, then the Lucas critique can also be read as saying that *I* type-level causes hidden variables.

I take this reading of the Lucas critique to be rather uncontroversial. The only point I'd like to add is that it doesn't make any difference whether *I* is manipulated for policy or experimental purposes, and that the Lucas critique therefore has negative bearing for the effectiveness of the IV method in macroeconomics. In order to show that the hypothesis that *X* directly type-level causes *Y* is true, researchers need to show that there is an intervention variable *I* that satisfies conditions (*I*1) – (*I*4). They cannot show that there is such a variable unless a set *Z* of variables can be controlled for in effective ways: unless it is possible to distribute the variables in *Z* evenly over all subjects of investigation or to control for these variables through nature or direct human intervention. In many special sciences such as microeconomics or pharmacology, *Z* can be controlled for in effective ways. In macroeconomics, however, *Z* includes hidden variables that in the case of the Lucas critique (in cases like King and Plosser's) are type-level caused by *I* (are not type-level caused by *I* but type-level cause *X*). In macroeconomics, moreover, subjects of investigation (a plurality of economic systems monitored during a particular time interval or a plurality of time intervals during which one particular system is monitored) are too diverse and small in number for randomization to lead to even distributions of *Z*. In macroeconomics, the evidence that can be provided in support of conditions (*I*1) – (*I*4) is therefore in principle too inconclusive to disentangle a set of competing and observationally equivalent hypotheses, i.e. to support the hypothesis that *X* directly type-level causes *Y* (or to turn belief in the truth of that hypothesis into knowledge).

There are no less than four objections that it seems can be raised against the general case. The first objection is that in natural experiments in macroeconomics, evidence not only springs from correlations or co-occurrences of events but also from the temporal order of events and the largeness of effects, and that all pieces of evidence combine to disentangle competing and observationally equivalent hypotheses. In response to that objection, one needs to point out that temporal order or largeness of effects rarely plays any prominent role in natural experiments in macroeconomics. The monetary policy contraction that Friedman

and Schwartz (1963, pp. 688-9) observe for 10/1931 isn't followed but rather accompanied by a monetary and economic contraction in the same month; and the economic contraction that they observe for that month is relatively small because industrial production had been declining already for more than a year. But even if the order of events were such that the monetary contraction were followed by the economic contraction, neoclassical macroeconomists like King and Plosser could still defend the case of reverse causation: the monetary contraction occurred before the economic contraction because firms first decided to shrink production, then decreased their money holdings (inducing the monetary contraction), and then shrank production (effectuating the economic contraction). And even if the economic contraction were sharp and distinctive, rational expectation theorists could still deny that it was token-level caused by a monetary contraction: the monetary contraction occurred because of the contractionary monetary policy, but the economic contraction didn't occur because of the monetary contraction because economic agents fully anticipated the decisive steps that the Federal Reserve was going to take.

The second objection that it seems can be raised against the above generalization says that conditions (*I*1) – (*I*4) are too strong and should be replaced with a weaker set of conditions. Reiss (2005, pp. 973-5), for instance, claims that skipping (*I*2) "is more in line with econometric practice".[12] And he cites well-known examples from the econometric literature on instrumental variables and natural experiments in order to support his claim. But in macroeconomics, even Reiss's weakened set of conditions is unlikely to be satisfied. His set still includes conditions (*I*1), (*I*3) and (*I*4). And remember from the preceding section that conditions (*I*3) and (*I*4) are violated in the case of rational expectations under full policy anticipations, and that conditions (*I*1), (*I*3) and (*I*4) are violated in the case of reverse causation.

The third objection is directed against the in-principle modality of the above generalization: perhaps evidence supporting conditions (*I*1) – (*I*4) is currently too inconclusive to disentangle competing and observationally equivalent hypotheses, but why should we think that it is too inconclusive as a matter of principle? Shouldn't we expect scientific progress to render it sufficiently conclusive in the end? At this stage, it is impossible to surmise whether there will be any progress of that sort. Derivations of more conclusive evidence require greater numbers of or less diverse subjects of investigation or effective ways of controlling for individual expectations and decisions through direct human intervention. And at present, it is

[12] He argues, more precisely, that skipping (*I*2) and replacing (*I*4) with the condition that *I* and *Y* do not have causes in common (except those that might cause *Y* via *I* and *X*) is more in line with econometric practice. But (as he himself observes) his condition that *I* and *Y* do not have causes in common (except those that might cause *Y* via *I* and *X*) is equivalent with (*I*4) as long as the common cause principle holds.

unclear how macroeconomics could ever meet these requirements. But we cannot rule out that at one point, it will be able to meet these requirements. Behavioral economists or neuroeconomists, for instance, might at one point be able to measure the variables that currently appear hidden. Or perhaps researchers conducting surveys might at one point be able to develop the methodologies that render subjects' statements about what they expect more trustworthy or verifiable. My use of the term 'in principle' should therefore be restricted to the inconclusiveness of causal evidence in macroeconomics, as we know it today.

The fourth objection says that in order to show that *X* directly type-level causes *Y*, researchers don't need to show that there is an intervention variable that satisfies conditions (*I*1) – (*I*4); they can also carry out the Hoover or AK test. This objection is arguably the most important one. Nowadays hardly any macroeconomist doubts that RCTs cannot be conducted in macroeconomics, or that the evidence deriving from natural experiments in macroeconomics is too inconclusive to disentangle competing and observationally equivalent hypotheses. Most macroeconomists believe that causal inference must proceed from "a statistical relevance basis" (Hoover, 2001, p. 149). Since the Hoover and the AK test proceed from such a basis, they appear to represent promising alternatives to the IV method.

It is to be conceded that the generalization defended in section 4 doesn't hold as long as the Hoover or AK test can be upheld as a method of inferring causal evidence that is sufficiently strong to turn causal belief into knowledge. But the aim of the following two sections is to show that the evidence provided by these tests is in principle too inconclusive to turn causal belief into knowledge. I will argue that a statistical relevance basis determines causal structure only insufficiently, and that the additional steps that the Hoover and AK test take do not sufficiently determine that structure either. I will also argue that in order to determine that structure sufficiently one needs to assume the validity of conditions (*I*1), (*I*3) and (*I*4), i.e. the validity of conditions of which the present and the preceding two sections have shown that the evidence that can be provided in support of them is in principle too inconclusive to support the hypothesis that *X* directly type-level causes *Y*, where *X* and *Y* stand for macroeconomic aggregates. My argument is going to rely substantially on the work of Pearl ([2]2009: especially chaps. 1, 3, 5).

5. The Hoover test

The Hoover test is a testing procedure that Hoover (2001, pp. 214-7) says consists of three steps. The first step is to look at time-series data and to use non-quantitative and extra-statistical historical and institutional insights to assemble a chronology of interventions. This step is supposed to serve two purposes. The first is to divide history in periods with and without interventions so that the periods without interventions (the tranquil periods) can form

the baseline against which structural breaks (i.e. changes in any of the parameters of the process of a particular variable) can be identified. The second purpose is to provide cross-checks to statistical tests: a structural break detected at a time when no interventions can be identified may indicate econometric misspecification.

The second step of the Hoover test is to apply LSE methodology to specify a statistical model for each of the tranquil periods separately. LSE methodology operates by (i) specifying a deliberately overfitting general model, by (ii) subjecting the general model to a battery of diagnostic (or misspecification) tests (i.e. tests for normality of residuals, absence of autocorrelation, absence of heteroscedasticity and stability of coefficients), by (iii) testing for various restrictions (in particular, for the restriction that a set of coefficients is equal to the zero vector) in order to simplify the general model, and by (iv) subjecting the simplified model to a battery of diagnostic tests. If the simplified model passes these tests, LSE methodology continues by repeating steps (i) – (iv), i.e. by using the simplified model as a general model, by subjecting that model to a battery of diagnostic tests etc. Simplification is complete if any further simplification either fails any of the diagnostic tests or turns out to be statistically invalid as a restriction of the more general model.

The third step of the Hoover test is to use the simplified models for the baseline periods to identify structural breaks. By way of example, imagine that the simplified model resulting from the second step of the Hoover test is the following two-equation model (cf. Hoover, 2001, pp. 192-3):

$$y = \alpha x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma_\varepsilon^2),$$

$$x = \beta + \eta, \qquad \eta \sim N(0, \sigma_\eta^2),$$

where X and Y may stand for any of the aggregate quantities referred to above (the real interest rate and aggregate demand, respectively, the money stock and aggregate output, respectively etc.), taxes and government spending, respectively (as in the example that Hoover, 2001, section 8.1. and chapter 9, discusses), or prices and money, respectively (as in the example that Hoover, 2001, chapter 10, discusses), where $N(\cdot,\cdot)$ indicates a normal distribution characterized by its mean and variance, and where $\text{cov}(\varepsilon,\eta) = E(\varepsilon_t\varepsilon_s) = E(\eta_t\eta_s) = 0$, for $t \neq s$. The reduced form equations of that model run as follows:

$$y = \alpha\beta + \alpha\eta + \varepsilon,$$

$$x = \beta + \eta.$$

The reduced form equations describe the joint probability distribution of X and Y, $P(x,y)$, that can be partitioned into conditional and marginal distributions in two distinct ways:

$$P(x,y) = P(y|x)\cdot P(x) = P(x|y)\cdot P(y).$$

Using the reduced form equations, the conditional and marginal distributions can be calculated as follows:

$$P(y|x) = N(\alpha x, \sigma_\varepsilon^2),$$

$$P(x) = N(\beta, \sigma_\eta^2),$$

$$P(x|y) = N([\alpha\sigma_\eta^2 y + \beta\sigma_\varepsilon^2] / [\alpha^2\sigma_\eta^2 + \sigma_\varepsilon^2], [\sigma_\eta^2\sigma_\varepsilon^2] / [\alpha^2\sigma_\eta^2 + \sigma_\varepsilon^2]),$$

$$P(y) = N(\alpha\beta, \alpha^2\sigma_\eta^2 + \sigma_\varepsilon^2).$$

The third step of the Hoover test will identify a structural break in $\{\beta, \sigma_\eta^2\}$, i.e. the parameters of the *X* process, if the simplified model that results from the second step is characteristic of two adjacent tranquil periods, if the first step manages to identify interventions on *X* that occur in between these periods, and if the parameters of $P(x)$ and $P(x|y)$ break statistically in between these periods. *Mutatis mutandis*, the third step will identify structural breaks in $\{\alpha, \sigma_\varepsilon^2\}$, i.e. the parameters of the *Y* process.

The Hoover test will conclude that *X* directly type-level causes *Y* if the parameters of $P(y|x)$ remain invariant to changes in $\{\beta, \sigma_\eta^2\}$ and the parameters of $P(x)$ invariant to changes in $\{\alpha, \sigma_\varepsilon^2\}$. The parameters of $P(y|x)$ and $P(x)$ will remain invariant to changes in $\{\beta, \sigma_\eta^2\}$ and $\{\alpha, \sigma_\varepsilon^2\}$, respectively, if and only if the parameters of $P(x,y)$ are identified (or structural) and $\{\beta, \sigma_\eta^2\}$ and the parameters of $P(y|x)$, on the one hand, and $\{\alpha, \sigma_\varepsilon^2\}$ and the parameters of $P(x)$, on the other, are variation-free, i.e. mutually unconstrained.[13] A look at the above calculations of conditional and marginal distributions shows that $\{\beta, \sigma_\eta^2\}$ and the parameters of $P(y|x)$, on the one hand, and $\{\alpha, \sigma_\varepsilon^2\}$ and the parameters of $P(x)$, on the other, are indeed variation-free. But are the parameters of $P(x,y)$ identified?

Pearl (^{2}2009, pp. 149-50) points out that in order for α in $y = \alpha x + \varepsilon$ to be identified, there must be no variable (or set of variables) *Z* that *d*-separates *X* from *Y*, where *Z* is said to *d*-separate *X* from *Y* or "block" a path *p* between *X* and *Y* if and only if (i) *p* contains a chain $X \rightarrow M \rightarrow Y$ or a fork $X \leftarrow M \rightarrow Y$ such that the middle node *M* is in *Z*, or (ii) *p* contains an

[13] When defining the notion of direct type-level causation, Hoover suggests that the property of being variation-free is necessary and sufficient for structural invariance. He says, for instance, that a privileged parameterization "is the source of the causal asymmetries that define causal order", and that "a set of parameters is privileged when its members are […] variation-free" (Hoover, 2013, p. 41). When defining the notion of direct type-level causation, however, Hoover refers to structural (not statistical) parameters, i.e. to parameters that figure in a structural or (better) causal model. For a more detailed analysis of Hoover's definition, cf. Henschen (2018, section 4).

inverted fork (or collider) $X \rightarrow M \leftarrow Y$ such that the middle node M is not in Z, and such that no descendant of M is in Z (cf. Pearl, ^{2}2009, pp. 16-7). The basic problem with the Hoover test is that its three steps do not sufficiently guarantee that there is no path-blocking Z, and that it therefore cannot show that α is identified.

The same problem can be restated by noting that the three steps of the Hoover test do not sufficiently guarantee that the parameters of $P(y|x)$ remain invariant to changes in $\{\beta, \sigma_\eta^2\}$. In order for the parameters of $P(y|x)$ to remain invariant to changes in $\{\beta, \sigma_\eta^2\}$,

$$P(y \mid do(x), do(z)) = P(y \mid do(x))$$

needs to hold for all Z that d-separate X from Y, where $do(x)$ is the operator that Pearl (^{2}2009, p. 70) introduces to denote the intervention that sets X to x, and where $do(z)$ denotes the intervention that controls for any path-blocking variable Z.[14] Again, the problem with the Hoover test is that its three steps do not sufficiently guarantee that there is no path-blocking Z, and that it therefore cannot show that the parameters of $P(y|x)$ remain invariant to changes in $\{\beta, \sigma_\eta^2\}$.

Hoover might respond that interventions on Z won't escape the researcher's attention in the first step, that Z can be included in the deliberately overfitting general model that in the second step is subjected to LSE methodology and simplified to a statistical model that might be less parsimonious than the exemplary model cited above. But what if Z is a hidden variable like the ones mentioned in sections 2-4: an unobservable (and possibly uncontrollable) variable denoting inflation, demand, or GDP expectations, or the decisions of firms to shrink or expand production? If Z is a hidden variable, then the simplified model will provide a statistical relevance basis for an arbitrary number of competing causal models, i.e. then the following causal graphs will be observationally equivalent: $X \rightarrow Y$, $X \rightarrow Z \rightarrow Y$, $X \leftarrow Z \rightarrow Y$.

Perhaps Hoover believes that hidden variables are not causally relevant in all areas in macroeconomics, and that the areas in which they are relevant exclude the areas to which he applies his three-step testing procedure. It is important to see, however, that this belief would be unjustified. Hoover applies his three-step procedure to provide evidence in support of two hypotheses: the hypothesis that taxes directly type-level cause government spending (cf. Hoover, 2001, chap. 9) and the hypothesis that prices directly type-level cause money (cf. Hoover, 2001, chap. 10). Hoover is aware, of course, that it is impossible to say that these hypotheses are true *a priori*, and that it is easy to construct credible hypotheses about

[14] Pearl (2009, p. 160) in fact claims that this equation needs to hold for all Z disjoint of $\{X \cup Y\}$, but that claim is unnecessarily strong.

other causal structures. He seems to be unaware, however, that some of the aggregates that fill positions in these structures cannot be measured.

In the case of his first hypothesis, Hoover seems to underestimate the implications of the constant-share model that he analyzes at the outset of his case study (cf. Hoover, 2001, pp. 228-9). According to the constant-share model, taxes and government spending are causally independent because GNP type-level causes both taxes and government spending. But when carrying out the first step of his procedure, Hoover doesn't assemble a chronology of interventions on GNP but only a chronology of interventions in the shape of changes in military and federal spending and tax bills and tax reforms (cf. Hoover, 2001, pp. 229-31). Hoover might respond that assembling a chronology of interventions on GNP wouldn't be exceedingly difficult, and that GNP can be included in the deliberately overfitting general model that in the second step of his procedure is subjected to LSE methodology and simplified to a statistical model that might be less parsimonious than a simple bivariate model for taxes and government spending. Remember from section 3, however, that GNP is type-level caused by a hidden variable, i.e. the decisions of firms to shrink or expand production. The second of the three case studies that Lucas (1976, pp. 30-35) discusses to support his critique suggests, moreover, that this hidden variable is type-level caused by tax policy.[15]

In the case of his second hypothesis, Hoover (2001, pp. 260-1) runs money regressions with and without Federal Reserve policy instruments (reserves, the discount rate, and the Federal funds rate) in order to show that these instruments don't need to be included among the regressors of money regressions. But the rationale for including these instruments is the possible presence of a causal chain (a "Federal Reserve reaction function") that runs from prices through the Federal Reserve policy instruments and their effects on the banking system and the public to the stock of deposits. Aggregates that cannot be measured (such as inflation expectations or the decisions of firms to shrink or expand production) are likely to fill positions in that chain. In order to show that these aggregates don't need to be included among the regressors of money regressions, one would have to run money regressions that do and do not include these aggregates among their regressors. And the problem is, of course, that these aggregates cannot be included because they cannot be measured.

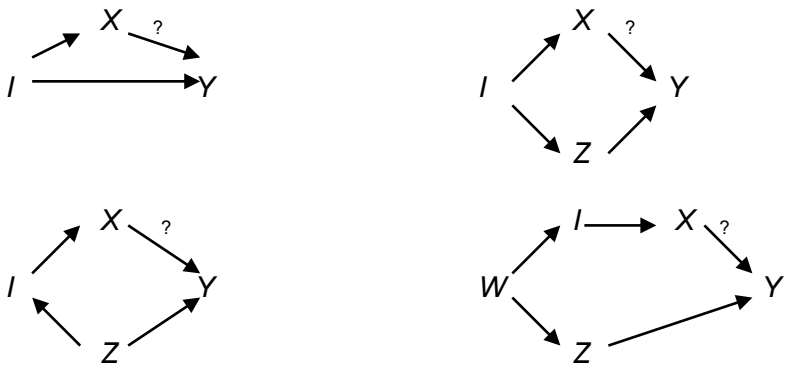
Hoover (2001, pp. 213-5, 276) is certainly aware of most of the difficulties that have been mentioned. He notes that the first step of assembling a chronology of interventions can be exceedingly difficult, and that the range of possible causal interactions could have "been expanded to include more fully the role of interest rates or the role of real variables and so forth". He also suggests that the LSE methodology to be applied in the second step has the

[15] According to Lucas, it is investment decisions that are type-level caused by tax policy. But investment decisions are likely to type-level cause production decisions.

important drawback of explicitly designing regressions that have desirable properties like normality of residuals, absence of autocorrelation etc. But he doesn't believe that his testing procedure must fail as a matter of principle, or that the evidence provided by that procedure is in principle too inconclusive to support specific causal hypotheses. All he believes is that his test is "not necessarily easy to implement", that it "may work sometimes" etc. (cf. Hoover, 2001, p. 213). And that it does work, he thinks he can show with his two case studies about the causal relations between taxes and government spending and between money and prices, respectively.

In the remainder of this section, I'd like to go a bit further than Hoover and argue that his testing procedure must fail as a matter of principle. It must fail as a matter of principle because it relies on conditions (*I*1), (*I*3) and (*I*4) from Woodward's definition of 'intervention variable', and because sections 2-4 have shown that the evidence that can be provided in support of these conditions is in principle too inconclusive to support the hypothesis that *X* directly type-level causes *Y*, where *X* and *Y* stand for macroeconomic aggregates. In order to see that the Hoover test relies on condition (*I*1), i.e. on the condition that *I* type-level causes *X*, note that a structural break in any of the parameters of the *X* process must be understood as an intervention on a parameter-intervention variable that type-level causes *X*. Hoover (2011, p. 348) agrees with this understanding when noting that with respect to condition (*I*1), there is no fundamental difference between his structural account of causality and Woodward's interventionist account.[16]

In order to see that the Hoover test also relies on conditions (*I*3) and (*I*4), i.e. on the condition that any directed path from *I* to *Y* goes through *X*, and on the condition that *I* is statistically independent of any variable *Z* that type-level causes *Y* and is on a directed path that does not go through *X*, note that these conditions are meant to rule out the following cases (cf. Woodward, 2003, pp. 101-102):



[16] For a more detailed analysis of the exact relationship between Hoover's account and a macroeconomic variant of Woodward's account, cf. Henschen (2018, especially sections 3 and 4).

These cases represent forks that contain path-blocking middle nodes: $\{I\}$, $\{I, Z\}$ or $\{I, Z, W\}$. Conditions (I3) and (I4) are therefore among the conditions that need to be satisfied in order for α to be identified, or for the parameters of $P(y|x)$ to remain invariant to changes in $\{\beta, \sigma_\eta^2\}$. And the Hoover test relies on these conditions in the sense that its three step-procedure does not sufficiently guarantee that they are satisfied.

One might wonder why the Hoover test relies on conditions (I1), (I3) and (I4) but not on condition (I2), i.e. on the condition that certain values of I are such that when I attains these values, X is no longer determined by other variables that type-level cause it but only by I. The reason is that condition (I2) is not necessary for inferring that X directly type-level causes Y. In order to assess whether X directly type-level causes Y, one might find it convenient to check whether intervening on I breaks all arrows that are directed into X and depart from variables other than I. But Reiss is right when suggesting that in general, checking whether intervening on I breaks all these arrows is not necessary for assessing whether X directly type-level causes Y (cf. section 4 above).[17]

6. The AK test

The AK test is a testing procedure that derives the hypothesis that X is a total type-level cause of Y from the following two conditions:

(a) $\sum_Z P(y|x, z)\cdot P(z) \neq 0$,

(b) potential outcomes of Y are probabilistically independent of X given Z,

where the expression in (a) measures the causal effect of X on Y, where (b) is a condition that Angrist and Kuersteiner (2011, p. 729) refer to as "selection-on-observables assumption" (or SOA, for short),[18] and where Z is an admissible (or de-confounding) set of variables. It is true that so far, the paper has been concerned with direct type-level causation, and that Angrist and Kuersteiner aim to derive a hypothesis of total type-level causation. Note, however, that the AK test would turn into an econometric procedure that tests for direct type-level causation if the following condition

(c) all direct type-level causes of Y except X remain fixed by intervention.

[17] Also note that condition (I2) doesn't figure in the definition that is central to a macroeconomic variant of Woodward's interventionist account (cf. Henschen 2018, section 3).

[18] Angrist and Pischke (2009, p. 54) refer to the same assumption as "conditional independence assumption". In the present context, however, it is more appropriate to use the term "selection-on-observables assumption" because the assumption in question might otherwise be confused with the assumption that Angrist and Kuersteiner (2011, p. 729) refer to as "the key testable conditional independence assumption", i.e. with an assumption that involves actual, not potential outcomes.

were added. This condition is meant to ensure that the relation of type-level causation between *X* and *Y* is direct (cf. Pearl, ^{2}2009, pp. 127-8).[19]

The hypothesis that Angrist and Kuersteiner (2011) aim to derive states that changes in the federal funds rate that the Federal Open Markets Committee (FOMC) intends at time *t* (ΔFF_t) directly type-level cause changes in real GDP at time *t*+*j* (ΔGDP_{t+j}), where *j* is the number of quarters ahead of *t*. It is true that they do not derive that conclusion from the inequality

(*a*') $\sum_{Z^t} P(\Delta gdp_{t+j} | \Delta ff_t, z_t) \cdot P(z_t) \neq 0$.

The causality test they use is a lot more sophisticated than a simple test for (*a*') (cf. Angrist and Kuersteiner, 2011, sections III + IV and table 3). But they are aware that their test relies on the assumption that

(*b*') potential outcomes of ΔGDP_{t+j} are independent of ΔFF_t given Z_t.

And that assumption is invalid unless Z_t is admissible (or de-confounding). Angrist and Kuersteiner believe that Z_t is admissible if the variables in Z_t figure on the right-hand side of a causal model that adequately describes the process determining ΔFF_t. Angrist and Kuersteiner (2011, p. 736) concede that they "do not really know how best to model the policy propensity score [i.e. the process determining ΔFF_t]; even maintaining the set of covariates, lag length is uncertain, for example". They therefore propose to specify a multiplicity of causal models for the process determining ΔFF_t, to submit these models to a number of diagnostic (or misspecification) tests, and then to submit these models to their causality test.

But the models they propose do not differ a lot. They all include variables standing for lagged changes in the intended federal funds rate, predicted changes in real GDP, predicted inflation and predicted unemployment innovation[20], past changes in real GDP and past inflation, and changes in predictions since the previous meeting of the FOMC. And differences between these models essentially relate to the number of lagged or predicted values that they include. The model that Angrist and Kuersteiner (2011, pp. 736-7) say performs best in terms of statistical adequacy (with respect to the diagnostic tests) runs as follows:

$$\Delta ff_t = \alpha + B\mathbf{z}_t + \varepsilon_t,$$

where α is an intercept and B a vector of parameters for the variables in Z_t. The error term ε_t represents the "idiosyncratic information" to which policymakers are assumed to react.

[19] In a probabilistic context, this condition corresponds to the condition that *Y* be 'unshieldable' from *X*, i.e. to a condition that follows from two results that Spohn (1980, pp. 77, 84) derives.

[20] By 'unemployment innovation', Angrist and Kuersteiner (2011, p. 736n) mean the unemployment rate in the current quarter minus the unemployment rate in the previous month.

Angrist and Kuersteiner (2011, p. 727) note that this information cannot be observed, and that it needs to be modeled as a stochastic shock (cf. part I, section 1). Z_t includes

- the change in the intended federal funds rate in t–1: Δff_{t-1}
- changes in real GDP, inflation and unemployment innovation in t, t+1 and t+2 that the FOMC predicts in t: ΔGDP_t, Π_t, ΔGDP_{t+1}, Π_{t+1}, ΔGDP_{t+2}, Π_{t+2}, U_t
- past changes in real GDP and inflation: ΔGDP_{t-1}, Π_{t-1}

Z_t also includes the changes in the predictions of inflation and changes in real GDP since the previous meeting of the FOMC. But I will drop these changes in order not to complicate matters too much.

So according to Angrist and Kuersteiner, Z_t is an admissible (or de-confounding) set of covariates if it includes the 10 variables listed above, i.e. if

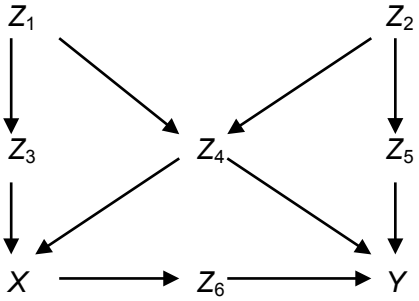
$$Z_t = \{\Delta FF_{t-1}, \Delta GDP_{t-1}, \Delta GDP_t, \Delta GDP_{t+1}, \Delta GDP_{t+2}, \Pi_{t-1}, \Pi_t, \Pi_{t+1}, \Pi_{t+2}, U_t\}$$

If Z_t is an admissible (or de-confounding) set of covariates, then the conditional independence assumption (*b*') will hold: then we will be allowed to ignore changes in the intended federal funds rate if we wish to determine potential outcomes of changes in real GDP. Intuitively, this makes a lot of sense: if the 10 variables in Z_t sufficiently determine the FOMC's decision to set the federal funds rate at a particular level, then we don't need to look at that level if we are interested in potential outcomes of changes in real GDP. The interesting and somewhat counterintuitive point is that we may nonetheless be allowed to say that changes in the intended federal funds rate directly type-level cause changes in real GDP. If changes in the intended federal funds rate are ignorable, conditionally on the 10 variables in Z_t, and if Angrist and Kuersteiner's causality test leads to positive results, then changes in the intended federal funds rate can be said to directly type-level cause changes in real GDP.

Now, Angrist and Kuersteiner's causality test does lead to positive results. In the case of their preferred model (the "baseline Romer model"), these results say that seven to twelve quarters ahead, there is a causal effect of changes in the intended federal funds rate on changes in real GDP at a significance level of 1 or 5%: that the probability that real GDP_{t+j} changes as a result to a change in FF_t lies somewhere between 0.040 and 0.092 for $j = 7 \ldots 12$ (cf. Angrist and Kuersteiner, 2011, table 3). But the problem with these results is that they are likely to be biased. They are likely to be biased because Angrist and Kuersteiner are likely to be misguided in their choice of covariates.

Pearl ([2]2009, pp. 79-80) emphasizes that Z doesn't qualify as admissible unless it satisfies the "back-door criterion". He says that Z satisfies the backdoor criterion relative to an ordered pair of variables (X, Y) in a causal graph if Z (i) doesn't include any descendants of X and (ii)

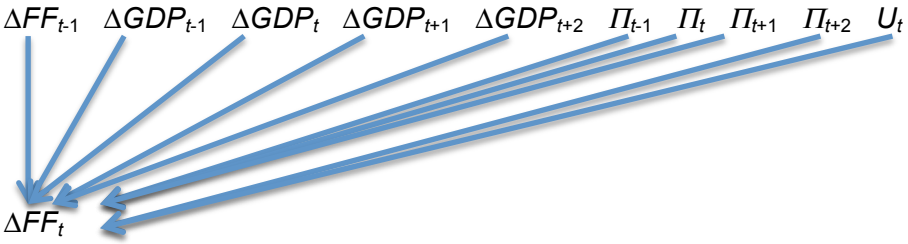
blocks every path between X and Y that contains an arrow into X. Remember from the preceding section that Z is said to "block" a path p if p contains at least one arrow-emitting node that is in Z or at least one collision node that is not in Z and has no descendants in Z (cf. Pearl, ^{2}2009, pp. 16-7). Pearl (^{2}2009, p. 80) uses the following graph to illustrate the backdoor criterion:



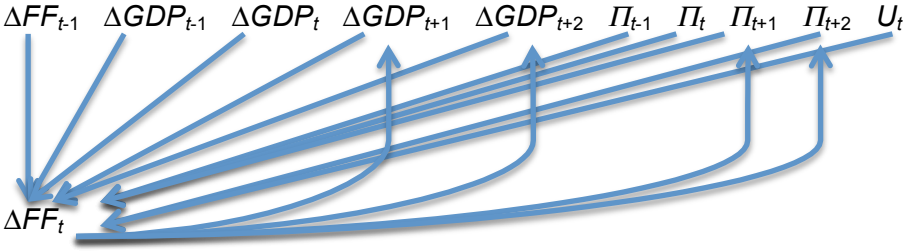
In this graph, $\{Z_3, Z_4\}$ and $\{Z_4, Z_5\}$ satisfy the backdoor criterion because they do not include any descendants of X, and because they block every path between X and Y that contains an arrow into X. It is clear that $\{Z_6\}$ does not satisfy the backdoor criterion because it is a descendent of X. It is worth noting, however, that by itself $\{Z_4\}$ doesn't satisfy the backdoor criterion either: it blocks the path $X \leftarrow Z_3 \leftarrow Z_1 \rightarrow Z_4 \rightarrow Y$ because the arrow-emitting node is in Z; but it does not block the path $X \leftarrow Z_3 \leftarrow Z_1 \rightarrow Z_4 \leftarrow Z_2 \rightarrow Z_5 \rightarrow Y$ because none of the arrow-emitting nodes (Z_1, Z_2) is in Z, and because the collision node Z_4 is not outside Z. Pearl (^{2}2009, pp. 80-81) then proves the proposition that $P(y\,|do(x)) = \sum_z P(y\,|x,\, z)\cdot P(z)$ if and only if Z satisfies the backdoor-criterion, where $P(y\,|do(x))$ is the probability that $Y = y$ if X is set to x by intervention.

For Angrist and Kuersteiner's analysis of monetary policy shocks this means that Z_t is an admissible (or de-confounding) set of covariates if and only if Z_t satisfies the backdoor criterion, i.e. if and only if Z_t (i) doesn't include any descendants of ΔFF_t and (ii) blocks every path between ΔFF_t and ΔGDP_{t+j} that contains an arrow into ΔFF_t. It is true that Pearl proves his theorem for variables that aren't time-indexed. But no matter if time-indexed or not, variables represent sets of potential values that are measurable or quantifiable. When time-indexed, they just represent sets of ordered pairs that assign each possible value to each possible point in time. There is accordingly no reason why Pearl's theorem shouldn't be applicable to time series data.

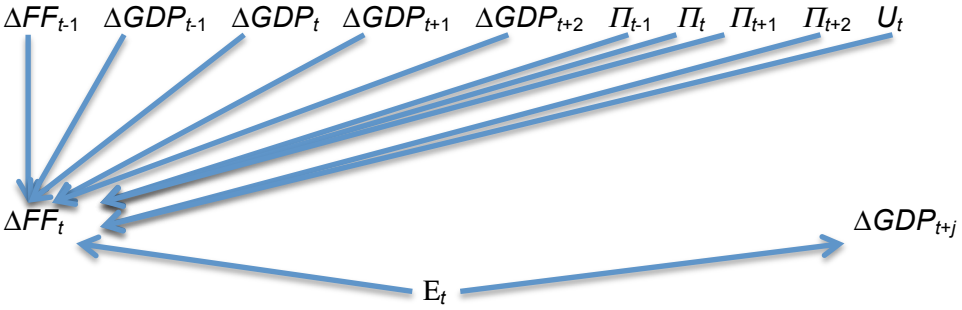
Angrist and Kuersteiner are likely to be misguided in their choice of covariates because Z_t is unlikely to satisfy the backdoor criterion. Here is the causal graph that corresponds to Angrist and Kuersteiner's preferred model:



If the members on the FOMC believe that monetary policy has real-economy effects (as most of them do), then their predictions of changes in real GDP and inflation will depend on the federal funds rate that they intend to set now: then there will be causal arrows from ΔFF_t (changes in the intended federal funds rate) to changes in real GDP and inflation in t+1 and t+2 (though perhaps not in t) that the FOMC predicts in t (violating condition (i) of the backdoor criterion)[21]:



And if the idiosyncratic information to which policymakers are assumed to react is also the sort of information that also makes firms shrink or expand production (a general pessimistic or optimistic outlook on the economy), then there will be arrows from that kind of information to ΔFF_t and ΔGDP_{t+j} (violating condition (ii) of the backdoor criterion):



More generally speaking, the problem with the AK test is that results from probabilistic causality tests will be biased unless the set of covariates satisfies the backdoor criterion, that the backdoor criterion implies Woodward's conditions (*I*1) – (*I*4), and that sections 2-4 have shown that the evidence that macroeconomists can provide in support of these conditions is too inconclusive in principle.

[21] It would be implausible to say that the real effects of ΔFF_t would materialize suddenly in t+j, and not continuously over j quarters.

Why does the backdoor criterion imply Woodward's conditions (*I*1) – (*I*4)? For Pearl ([2]2009, pp. 70-71), *do*(*x*) amounts to setting *X* to *x* by manipulating an intervention variable *I* and by breaking all arrows directed into *X* and departing from variables other than *I*. For Pearl, that is, *do*(*x*) requires that Woodward's conditions (*I*1) and (*I*2) be satisfied. Condition (ii) of the backdoor criterion, moreover, rules out the same cases as Woodward's conditions (*I*3) and (*I*4) (cf. Fig. 1 in section 5 above). These are cases in which *Z* blocks the paths between *X* and *Y* that contain an arrow into *X*. Therefore, conditioning on *Z* (knowing the value of *Z*) rules out the same cases as conditions (*I*3) and (*I*4). The backdoor criterion is a bit stronger than Woodward's conditions (*I*1) – (*I*4) since condition (i) of the backdoor criterion also rules out cases in which arrows are directed into *Z* and depart from *X*.[22] But the backdoor criterion represents a set of conditions that includes Woodward's conditions (*I*1) – (*I*4). One may accordingly say that the backdoor criterion implies these conditions.

As long as *I*, *Z* and *W* are known and measurable, there will be no problem: one of *I*, *Z* and *W* can simply be added to the set of admissible (or de-confounding) covariates. But in macroeconomics, hidden variables (i.e. variables that are causally relevant, though unobservable and possibly incapable of manipulation through direct human intervention) are always likely to be present: variables standing for decisions of firms to shrink or expand production, idiosyncratic information that guides these decisions, inflation expectations etc. And if a hidden variable is present, then causal inference based on the potential-outcome approach will be defective. The evidence that the AK test can provide is therefore too inconclusive to disentangle competing and observationally equivalent causal hypotheses in macroeconomics.

7. Conclusion

In sections 2-6 I have argued that the evidence that macroeconomists can provide in support of the hypothesis that *X* directly type-level causes *Y* is too inconclusive in principle. Sections 2-4 have shown that the evidence provided by the IV method is too inconclusive because it derives from conditions (*I*1) – (*I*4) of Woodward's definition of 'intervention variable', and because in macroeconomics, hidden variables that violate these conditions, are likely to be present. Section 5 has argued that the evidence provided by the Hoover test is too inconclusive because it cannot show that the parameters of $P(x,y)$ are identified, or that $P(y|x)$ remains invariant to changes in the parameters of the *X* process, and because conditions (*I*1), (*I*3) and (*I*4) are among the conditions that need to be satisfied in order for

[22] Condition (i) of the backdoor criterion is meant to ensure acyclicity (cf. Pearl, [2]2009, p. 339). Henschen (2018, section 5) argues that a potential-outcome approach to macroeconomic causality reduces to a macroeconomic variant of Woodward's interventionist account if condition (*I*2) and condition (i) of the backdoor criterion are dropped.

the parameters of $P(y,x)$ to be identified, or for the parameters of $P(y|x)$ to remain invariant. Finally, section 6 has tried to show that the evidence provided by the AK test is too inconclusive because the results of that test will be biased unless the backdoor criterion guides the choice of covariates, and because the backdoor criterion implies conditions (*I*1) – (*I*4). An important conclusion to be drawn from sections 2-6 states that macroeconomics cannot do justice to its ultimate justification if the ultimate justification for the study of macroeconomics is to provide knowledge on which to base policy, and if knowledge on which to base policy is causal knowledge (cf. section 1).

By way of conclusion, I'd like to point out that another conclusion that needs to be drawn from sections 2-6 conflicts with a claim that Friedman makes in a famous passage from his 1953 paper. In that passage, Friedman (1953/^{2}1994, p. 185) claims that "[t]he inability to conduct so-called 'controlled experiments' does not […] reflect a basic difference between the social and physical sciences […] because the distinction between a controlled experiment and uncontrolled experience is at best one of degree". It is true that the inability to conduct RCTs doesn't reflect a basic difference between the social and physical sciences: in many social sciences (including microeconomics), RCTs are conducted on a large scale; and in some physical sciences (such as astronomy), RCTs cannot be conducted. It is also true that the distinction between a controlled experiment and uncontrolled experience is "at best one of degree".

Note, however, that the inability to conduct RCTs does reflect a basic difference between macroeconomics and many of the special sciences (including microeconomics and pharmacology). Unlike researchers in many of the special sciences, macroeconomists cannot provide conclusive evidence in support of hypotheses of direct type-level causation. They cannot provide that evidence because they cannot conduct RCTs; and they cannot conduct RCTs because in macroeconomics, hidden variables (variables that are causally relevant, though unobservable and possibly incapable of control through direct human intervention) are always likely to be present.

Note further that the immediate context suggests that Friedman's claim is primarily concerned with macroeconomics. It is true that most of the more detailed examples he discusses are drawn from microeconomics (and especially the field of industrial organization). But in the more immediate context of his claim, Friedman (1953/^{2}1994, p. 186) refers to "the hypothesis that a substantial increase in the quantity of money within a relatively short period is accompanied by a substantial increase in prices". He also maintains that "experience casts up […] direct, dramatic, and convincing […] evidence" in support of that hypothesis. His claim is, moreover, repeated almost literally in Friedman and Schwartz (1963, p. 688). If the claim is that the inability to conduct RCTs doesn't reflect a basic

difference between macroeconomics and many of the special sciences (including microeconomics and pharmacology), then that claim conflicts with an important conclusion that needs to be drawn from sections 2-6 of this paper.

References:

Afrouzi, H. et al. (2015). "Inflation Targeting Does Not Anchor Inflation Expectations. Evidence From Firms in New Zealand." NBER Working Paper 21814.

Angrist, J. D. and Kuersteiner, G. M. (2011). Causal effects of monetary shocks: semi-parametric conditional independence tests with a multinomial propensity score. *The Review of Economics and Statistics* 93(3), 725-747.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton, NJ: PUP.

Beechey, M. J. et al. (2011). "Are Long-Run Inflation Expectations Anchored More Firmly in the Euro Area than in the United States?" *American Economic Journal: Macroeconomics* 3(2), 104-129.

Berg, G. J. v. d. and Klaauw, B. v. d. (2006). "Counseling and monitoring of unemployed workers: theory and evidence from a controlled social experiment." *International Economic Review* 47(3), 895-936.

Dornbusch, R., Fischer, S., and Startz, R. ([7]1998). *Macroeconomics*. Boston: McGraw-Hill.

Engle, R. F. and Hendry, D. F. (1993). Testing Super Exogeneity and Invariance in Regression Models. *Journal of Econometrics* 56, 119-139.

Engle, R. F., Hendry, D. F. and Richard, J. F. (1983). Exogeneity. *Econometrica* 51(2), 277-304.

Friedman, M. (1953/[2]1994). The Methodology of Positive Economics. In Hausman, D. M. (ed.), *The Philosophy of Economics. An Anthology*. Cambridge, MA: CUP.

Friedman, M. and Schwarz, A. J. (1963). *A Monetary History of the United States, 1867-1960*. Princeton, NJ: PUP.

Granger, C. W. J. (1969). "Investigating Causal Relations By Econometric Models and Cross-Spectrum Methods." *Econometrica* 37(3), 424-438.

Granger, C. W. J. (1980). Testing for Causality: A Personal Viewpoint. *Journal of Economic Dynamics and Control* 2(4), 329-352.

Hendry, D. F. (1988). The Encompassing Implications of Feedback versus Feedforward Mechanisms in Econometrics. Oxford Economic Papers 40(1), 132-149.

Henschen, T. (2018). What is macroeconomic causality? *Journal of Economic Methodology* 25(1), 1-20.

Hoover, K. D. (2001). *Causality in Macroeconomics*, Cambridge: CUP.

Hoover, K. D. (2009). Milton Friedman's Stance: The Methodology of Causal Realism. In Mäki, U. (ed.), *The Methodology of Positive Economics: Milton Friedman's Essay Fifty Years Later*. Cambridge, MA: CUP, 303-320.

Hoover, K. D. (2011). Counterfactuals and Causal Structure. In McKay Illari, P., Russo, F. and Williamson, J. (eds.), *Causality in the Sciences*. Oxford: OUP, 338-360.

Hoover, K. D. (2013). Identity, Structure, and Causal Representation in Scientific Models. In Chao, H.-K., Chen, S.-T. and Millstein, R. (eds.). *Towards the Methodological Turn in the Philosophy of Science: Mechanism and Causality in Biology and Economics*. Dordrecht: Spinger, 35-60.

King, R. G. and Plosser, C. I. (1984). Money, Credit and Prices in a Real Business Cycle. *American Economic Review*, 74, 363-380.

Lucas, R. E. (1976). "Econometric Policy Evaluation: A Critique." *Carnegie-Rochester Conference Series on Public Policy* 1, 19-46.

Pearl, J. (^{2}2009). *Causality: Models, Reasoning, and Inference*. Cambridge, MA: Cambridge University Press.

Reiss, J. (2005). Causal Instrumental Variables and Interventions. *Philosophy of Science* 72, 964-976.

Reiss, J. (2013). *Philosophy of Economics*. London/New York: Routledge.

Romer, C. D. and Romer, D. H. (1989). "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz." *NBER Macroeconomics Annual* 4: 121-170.

Romer, D. (^{4}2012). *Advanced Macroeconomics*. New York: McGraw-Hill.

Sen, A. (1973). "Behavior and the Concept of Preference." *Economica* 40(159), 241-259.

Spohn, W. (1980). Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic* 9: 73-99.

Woodward, J. (2003). *Making Things Happen: a Causal Theory of Explanation*. Oxford: OUP.