Identification with Models and Exogenous Data Variation

R. Jay Kahn
Stephen M. Ross School of Business
University of Michigan

Toni M. Whited
Stephen M. Ross School of Business
University of Michigan

Ross School of Business Working Paper Series
Working Paper No. 1323
July 2016

Identification with Models and Exogenous Data Variation

R. Jay Kahn
University of Michigan

Toni M. Whited*
University of Michigan and NBER

June 11, 2016

Abstract

We distinguish between identification and establishing causality. Identification means forming a unique mapping from features of data to quantities that are of interest to economists. Establishing causality is synonymous with finding sources of exogenous variation. These two issues are often confused. However, exogenous variation is only sometimes necessary and never sufficient to identify economically interesting parameters. Instead, even for causal questions identification must rest on an underlying economic model. We illustrate these points by examining identification in two recent papers: one causal study relying on an entirely verbal model and one non-causal study relying on a formal mathematical model.

*We would like to thank an anonymous referee for helpful comments.

1. Introduction

In terms of its pure statistical definition, identification is simple. An applied econometrician defines an objective function over parameters and a data population, and her goal is to select parameters that minimize this objective function, in which the population has been replaced by a specific sample. A parameter is identified if there is a unique minimum for the objective function at its true value in the population. Yet discussion of this pure statistical issue of identification is not of particular interest to applied economists because the parameter at the minimum of the objective function may or may not be of interest from an *economic* point of view. For example, if we say a regression of price on quantity does not identify demand, we are not arguing that the regression itself is not well-formed. Ordinary least squares produces an unbiased estimate of the slope coefficient on price. However, we are stating that this estimation has not identified an economic parameter, typically a utility parameter, that we find interesting. The true problem of identification is then using an econometric objective function to form a mapping from observed data to relevant economic parameters. Unfortunately, identifying an economically interesting parameter is far more difficult than the sheer statistical definition of identification might suggest.

The purpose of this paper is to delineate the relationship between estimating a causal effect and the more general issues of identification. Since at least Alfred Marshall's *Principles of Economics*, economists have understood causality as a *ceterus paribus* comparison: the causal effect of variable A on variable B is the change in B that results from altering A while holding all other features of the world constant. See, for instance, the discussion in Heckman and Pinto (2015). Causal effects are simply elasticities, but they are difficult to estimate because econometricians rarely observe occasions where one variable is altered while others are held constant, that is, where there is genuine exogenous variation in a variable.

This exogenous variation forms the focus of how many economists think about identifica-

tion. However, the general issue of identification is broader in scope than the establishment of exogenous variation. We want to make three points on this subject. First, identification relates to parameters of an econometric model. In some applications, these parameters are the elasticities that define causal effects, but, as we show below, they need not be. Relatedly, not all interesting questions are causal in nature, so not all identification issues revolve around establishing causality. For example, one might be able to run an experiment to establish that a causal effect exists, but the experiment alone cannot establish what the effect means, and often the economic forces that are behind the causal effect are at least as interesting as the effect itself. Second, finding exogenous variation in a variable is never sufficient for identification of an economically interesting parameter, and for some questions, exogenous variation may not be necessary. In addition, neither the presence of random variation nor the establishment of causality necessarily fulfills the goal of answering interesting questions. Third, identification of an economically interesting parameter is always based on a verbal or mathematical theory. Thus, identification can never be free of assumptions or even light on assumptions.

None of these points are entirely new. In fact, the last can be traced back at least as far as Koopmans (1949), who pointed out:

> Where statistical data are used as one of the foundation stones on which the equation system is erected, the modern methods of statistical inference are an indispensable instrument. However, without economic "theory" as another foundation stone, it is impossible to make such statistical inference apply directly to the equations of economic behavior which are most relevant to analysis and to policy discussion.

We will illustrate these points by looking at two different papers and examining how each identifies an economic parameter. Along the way, we also relate each of these identification

strategies back to the statistical definition of identification.

Because the problem of identification is really a problem of identifying interesting parameters, it matters which parameters are deemed to be interesting. In practice, parameters are going to be interesting if they address important questions being asked in a body of research. For example, there are many questions for which economists are ultimately concerned with the effect of limited government interventions on economic variables: how an increase in the minimum wage affects employment (Neumark and Wascher 1992; Card and Krueger 1994; Dube et al. 2010; Sorkin 2015), how class size affects achievement (Angrist and Lavy 1999; Krueger 1999; Krueger and Whitmore 2001; Chetty et al. 2011), or how training affects earnings (Ashenfelter 1978; Ashenfelter and Card 1985; Heckman et al. 1997). In all these cases the parameter of interest is a simple elasticity. The government has a lever at its disposal, and we want to know what happens to some outcome when it pulls that lever. This simple *ceteris paribus* comparison surrounding a specific and limited government intervention makes identification relatively straight-forward. In such a situation, it is common to see researchers adopt an experimental approach. Yet even in this straightforward context, the average treatment effect that comes from such an approach is limited in its applicability. It represents only an estimate of the average causal effect of a variable under a particular, historical intervention. Without additional assumptions, it is difficult to extrapolate any such results to predictions about future interventions of a similar type. Nonetheless, if the goal of a study is only to establish the average effect of a previous intervention, then we can answer this question by estimating the relevant elasticity, as long as there is exogenous variation along the same dimension as the lever in the specific intervention. In this case, if the causal link to a government policy has been identified, the problem is solved. The link between causality and identification in these popular quasi-experimental studies makes it easy to confuse identification with the establishment of causality through exogenous variation. In fact, Angrist and Pischke (2008) present the issue of identification entirely as a

search for an approximation to an ideal experiment.

However, not all questions of interest can be phrased in experimental terms. In particular, in corporate finance, we are rarely confronted with the strong policy levers that have made the estimation of treatment effects one of the central activities of many areas of applied microeconomics. Yet we do have interesting questions to answer.

2. Identification with exogenous variation

For example, Bennedsen et al. (2007) studies Danish family firms, asking whether in-family succession of CEOs hurts performance. Given that most firms in the world are family firms, this question is clearly of interest. This question can also be phrased as a *ceteris paribus* comparison: how would the performance of the company have been different if an outside hire had been chosen as a CEO instead of a family member? In contrast to other comparisons, where the concern is with the impact of a government policy, here the parameter of interest is really related to an underlying agency problem. If an in-family CEO is appointed, he is drawn from a limited pool of family members. Within this limited pool, candidates are unlikely to be as proficient as if they were drawn from the broader market outside the family. So while the inside hire creates a non-pecuniary benefit to the family, it could hurt the performance of the firm. To understand the magnitude of this agency problem, we want to estimate the loss in performance that is due to the choice of a family member over an outsider. The loss in performance is a consequence caused by the choice of a family member over an outside candidate, but the parameter that measures the performance loss is of interest not because it represents an average, presumably causal, estimated effect: it is interesting because it represents a deeper agency friction.

Identifying the parameter that represents this agency friction is difficult because demand for a family CEO is endogenous to the performance of the company. Poor performance may force the family to choose an outside CEO instead of a relative, and good performance may

make a family insouciant about the specter of an incompetent family CEO. In this case, in a simple regression of performance on the choice of CEO:

$$\text{performance} = \alpha + \beta(\text{In-family succession}) + u$$

the coefficient on in-family succession, β, does not identify agency costs, because it is a function of both agency costs and the unobserved economic variables affecting demand for family CEOs. While the statistical parameter is well defined, the OLS objective function does not have a unique minimum at the true value of agency costs, as the model of the underlying economics of the question does not allow for a mapping of this regression slope, β, to the agency parameter of interest.

In order to identify the agency friction, the model needs more structure. With data from Danish firms, the authors can observe the demographics of controlling families. They choose the gender of the first-born child as an instrument to determine the causal impact of in-family succession. The argument the authors present is that families without male first-born children will be less likely to choose a family CEO, yet families with and without male first-born children should have ex-ante identical performance, as the biology of child gender is genuinely random. Crucially, while biology buys randomness, the power of the instrument in identifying the agency frictions comes from two additional assumptions added to the model. For the agency parameter to be identified, the reader has to believe the following. First, female CEOs will be no different from male CEOs, otherwise the instrument does not satisfy the exclusion restriction. Second, Danish families have a preference for primogeniture, otherwise the instrument will be weak (Staiger and Stock 1997). If both of these assumptions hold, then the instrumental variables objective function formed from the gender of the eldest child has a unique minimum at the agency cost friction, which is then identified.

These identifying assumptions are relatively mild, but they are still assumptions, and family preference for primogeniture is less innocuous than it seems. To see the importance

of this assumption, note that most completely random variables, for example, the inches of rainfall in Kansas in a year, are useless as instruments for identifying the effects of agency problems on firm performance because Danish family firms do not react them. The gender of a first-born child makes a good instrument not only because it is random but because Danish firms react to the instrument. Thus, in order to use the exogenous variation from the gender of the first-born, we must also assume controlling families are somewhat sexist. In the absence of sexism, the instrument has no bearing on the succession decision, and the parameter is again unidentified. If we are willing to assume that sexism exists, then the exclusion restriction is that sexism affects firm performance *only* through the choice of a family CEO, and this assumption is non-trivial. For instance, imagine that sexism causes controlling families to raise first-born boys and girls differently. Boys are groomed to lead the family firm, while girls are encouraged to pursue other professions. In this case, the gender of the (potential) family CEO would affect (potential) firm performance, and the exclusion restriction would not hold. The data provide no evidence for or against the possibility of grooming, so identification requires that one assume away this possibility. This example thus illustrates that it is not exogenous variation alone that allows the agency parameter to be identified, it is the assumptions made in the verbal theory of behavior this paper advances.

3. Identification without exogenous variation

For some questions, exogenous variation is not even necessary to identify an economically interesting parameter. In fact, causal inference in general may not be the point. To illustrate this point, we turn outside corporate finance and examine Davis et al. (2014), which asks the extent to which agglomeration externalities impact aggregate consumption growth. Agglomeration externalities are the productivity gains that occur when workers and firms locate in the same area. In this case, the identification of the impact of agglomeration on consumption growth is more difficult, in large part because the question cannot be phrased

as a *ceteris paribus* comparison. For example, one would want to compare consumption growth in Chicago as it currently is with consumption growth in a counter-factual "city" in which Chicago's population is spread out over Illinois (but no other changes are made). Of course, such a situation is difficult to envision, and impossible to observe.

Nonetheless, the process used to establish identification carries many similarities with the process used to establish identification in Bennedsen et al. (2007). Davis et al. (2014) starts with an explicit set of assumptions that underlie a dynamic general-equilibrium model of agglomeration in cities. In the model, firms do not take into account the positive productivity spillovers that they generate when they hire extra workers within a city. This externality then affects consumption growth as long as land prices are rising. In this case, an increase in the forecast of land prices leads firms to economize on space now. This reaction in turn leads to an increase in productivity via the externality because more workers and firms are clustered onto a smaller space. The set of assumptions that leads to this behavior in the model in turn implies that the correlation between a forecast of land prices and the growth of total factor productivity is a function of the agglomeration externality and some other easily-estimated parameters. This result means that the OLS objective function in a regression of total factor productivity on forecasted land prices has a unique minimum at the true value of the agglomeration externality parameter, without the need for any exogenous source of variation.

This identification is not assumed. Just as in Bennedsen et al. (2007), it is the result of a careful argument extended from a set of assumptions. However, there are two important differences between identification in the structural study and identification in the reduced form study. First, the arguments in Davis et al. (2014) are phrased using mathematics, and the arguments in Bennedsen et al. (2007) are verbal. Second, all of the assumptions needed for identification are contained in Davis et al. (2014). In contrast, some of the identifying assumptions in Bennedsen et al. (2007) are not contained in the paper, even though the paper

is quite explicit in stating that the gender of the first-born can only affect CEO succession via the choice of a family CEO, and even though the paper is as careful as it can be to convince the reader that all possibilities for violation of the exclusion restriction have been exhausted. Despite the cleanness of the natural experiment and the high level of care taken in its execution, because the identifying assumptions are verbal, there is always room to consider new alternatives, such as the priming we describe above, that could violate the exclusion restriction.

4. Conclusion

Both our examples illustrate the importance of a model for identification by pointing out that the model allows for the identification of an interesting quantity with a statistical parameter by advancing an internally consistent set of assumptions. A mathematical model is not essentially better or worse for this purpose than a verbal one. But careful advancement of a set of assumptions is difficult to accomplish verbally. Lawyers and researchers in the humanities practice for years to make their verbal arguments internally consistent. Economists are rarely so well prepared when they venture into a verbal model, but we have a great deal of experience with making mathematical arguments. So while mathematical models are not essentially better, they are often easier for economists to apply. Still, the theory is what allows us to identify structural parameters from statistical quantities.

Both examples also illustrate that the question to be addressed comes before the model or the natural experiment. The model, with all of its assumptions, or the natural experiment, with its accompanying assumptions, is only a tool. It would be difficult to address the in-family CEO succession question with the estimation of a dynamic model, which would need to incorporate both product market conditions and family dynamics. In addition, as pointed out above, it would be impossible to assess the affects of agglomeration externalities with a natural experiment. More generally, there is no one approach that will be useful for

answering all questions. But for whatever kind of question one asks, the key to identifying relevant parameters is to proceed guided by (either verbal or mathematical) theory and conscious of the necessary assumptions.

References

Angrist, J. D. and Lavy, V. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *Quarterly Journal of Economics*, 114:533–575.

Angrist, J. D. and Pischke, J.-S. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ashenfelter, O. 1978. "Estimating the effect of training programs on earnings." *Review of Economics and Statistics*, 60:47–57.

Ashenfelter, O. and Card, D. 1985. "Using the longitudinal structure of earnings to estimate the effect of training programs." *Review of Economics and Statistics*, 67:648–660.

Bennedsen, M., Nielsen, K. M., Perez-Gonzalez, F., and Wolfenzon, D. 2007. "Inside the family firm: The role of families in succession decisions and performance." *Quarterly Journal of Economics*, 122:647–691.

Card, D. and Krueger, A. B. 1994. "Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania." *American Economic Review*, 84:772–793.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *Quarterly Journal of Economics*, 126:1593–1660.

Davis, M. A., Fisher, J. D. M., and Whited, T. M. 2014. "Macroeconomic implications of agglomeration." *Econometrica*, 82:731–764.

Dube, A., Lester, T. W., and Reich, M. 2010. "Minimum wage effects across state borders: Estimates using contiguous counties." *Review of Economics and Statistics*, 92:945–964.

Heckman, J. and Pinto, R. 2015. "Causal analysis after Haavelmo." *Econometric Theory*, 31:115–151.

Heckman, J. J., Ichimura, H., and Todd, P. E. 1997. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme." *Review of Economic Studies*, 64:605–654.

Koopmans, T. C. 1949. "Identification problems in economic model construction." *Econometrica*, 17:125–144.

Krueger, A. B. 1999. "Experimental estimates of education production functions." *Quarterly Journal of Economics*, 114:497–532.

Krueger, A. B. and Whitmore, D. M. 2001. "The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR." *Economic Journal*, 111:1–28.

Neumark, D. and Wascher, W. 1992. "Employment effects of minimum and subminimum wages: Panel data on state minimum wage laws." *Industrial and Labor Relations Review*, 46:55–81.

Sorkin, I. 2015. "Are there long-run effects of the minimum wage?" *Review of Economic Dynamics*, 18:306–333.

Staiger, D. and Stock, J. H. 1997. "Instrumental variables regression with weak instruments." *Econometrica*, 65:557–586.