Foundations and Trends® in Accounting

Just How Sensitive are Instrumental Variable Estimates?

Suggested Citation: Peter C. Reiss (2016), "Just How Sensitive are Instrumental Variable Estimates?", Foundations and Trends® in Accounting: Vol. 10, No. 2-4, pp 204–237. DOI: 10.1561/1400000048.

Peter C. Reiss
Stanford Graduate School of Business
USA
preiss@stanford.edu

now
the essence of knowledge
Boston — Delft

Contents

Just How Sensitive are Instrumental Variable Estimates?

Peter C. Reiss

Stanford Graduate School of Business, USA; preiss@stanford.edu

ABSTRACT

Researchers regularly use instrumental variables to resolve concerns about regressor endogeneity. The existing literature has correctly emphasized that the choice of instrumental variables matter for the resulting estimates. This paper shows that researchers should also be concerned that the functional form of the instrument matters as well for the resulting estimates. For example, simply changing an instrumental variable from the level to the logarithm can change estimates directly. This article documents the problem, suggests why the problem occurs and suggests different approaches to the problem.

Peter C. Reiss (2016), "Just How Sensitive are Instrumental Variable Estimates?", Foundations and Trends® in Accounting: Vol. 10, No. 2-4, pp 204–237. DOI: 10.1561/1400000048.

1

Introduction

Many social scientists work with linear regression models in which some or all of the regressors are thought to be correlated with the regression error. Although in this case ordinary least squares (OLS) delivers the best linear predictor of the dependent variable given the right hand side regressors, this predictor differs from the best linear predictor one would obtain if there were no correlation between the regressors and the regression error. Researchers often are more interested in this second model because they see its coefficients as revealing "causal" effects. That is, they see the coefficients as suggesting how much the dependent variable would change if the corresponding right hand side variable changed by one unit and nothing else changed. Of course, the true causal model may not be linear. However, many researchers nevertheless believe that linear regressions still can reveal the signs and magnitudes of causal relations.

Instrumental variable (IV) methods are the primary means by which social scientists estimate regression models with endogenous regressors. IV methods require the researcher to identify auxiliary variables that minimally are uncorrelated with the regression error (e.g., exogeneous) and yet correlated with the right hand side endogenous regressors (i.e.,

relevant).[1] Under standard assumptions, these instruments can be used to construct consistent estimates of the coefficients. Although the use of valid IVs produces consistent estimates, it is well known that the finite and large sample distributions of the IV estimator are impacted by the choice of instruments.[2] Further, it is known that when the instruments only slightly violate the exogeneity and relevance conditions, there can be dramatic, adverse consequences for the estimator's small and large sample distribution. Violations of concern include: having too few relevant instruments; using instruments that are correlated with the regression error; and, relying on instruments that are weakly correlated with the endogenous regressors.[3]

This paper documents another issue that has received little or no attention – that seemingly irrelevant changes in the functional forms of the same instruments can lead to *vastly* different IV estimates. This is not just a sampling issue, it is present in a *given sample*. This potential sensitivity should be concerning. Two (or more) researchers could be on solid ground arguing that their IV estimates are consistent, and yet their estimates might differ dramatically. Indeed, their estimates may differ in sign! Section 2 provides such an example. Ultimately this difference in the IV estimates prompts the difficult question of which estimate(s) to report. Alternatively, how might a researcher make others aware of any sensitivity?

These issues are illustrated and addressed in what follows. Section 2 shows that an instrument's functional form can matter. It relates the sensitivity of IV estimates to an instrument relevance condition. Section 3 discusses possible approaches to the problem based on existing relevance and weak instrument diagnostics. These approaches include reporting sensitivity analyses or measures of the local variation in the estimated coefficients. Section 4 discusses possible "efficient" instrument approaches to the problem. Section 5 illustrates the problem is general. Section 6 concludes.

[1] While the term "instruments" refers to both the exogenous variables in the equation of interest and the excluded auxiliary variables, I will primarily use it to refer to the auxilary variables.

[2] See for example Phillips (1983).

[3] See for example Bound *et al.* (1995) and Murray (2006).

2

A Motivating Example

Social science researchers routinely posit linear regression relations between outcome variables y and regressors X. Sometimes these relations are motivated by theoretical models, but more often the specifications are ad hoc.

This section focuses on a regression model based on economic theory. This is done to have a clear motivation for the choice of instrument(s). The model is the simple log-log homogeneous product demand model:

$$\ln q = \beta_0 + \beta_p \ln p + \beta_I \ln y + \beta_b \ln p_b + \epsilon. \tag{2.1}$$

where q is the quantity consumed, p is the product price, I is consumer income, and p_b is the price of a substitute product. Here we focus on estimating β_p, which represents the (constant) elasticity of product demand, or the constant percentage by which quantity will fall with a one percent increase in price *holding all other variables constant*. The problem encountered with using ordinary least squares (OLS) to estimate β_p is that $\ln p$ and the error term ϵ are potentially correlated.[4]

[4]The usual explanation as to why is that prices and quantities are simultaneous determined. This means that unobservables that affect quantity demand also will impact the prices at which the market clears.

This correlation will lead OLS to produce a biased and inconsistent estimate of the conceptual quantity of interest – the elasticity of demand, β_p.[5]

Besides $\ln y$ and $\ln p_b$, which we assume are uncorrelated with the demand error, we need one or more other variables to construct a consistent estimate of β_p. These extra *instrumental* variables minimally must be: excluded from the demand equation, uncorrelated with the demand error ϵ, and correlated with the right hand side endogenous variables (here the natural logarithm of product price $\ln p$). Economic theory suggests that we can find such variables by considering the supply side of the market. Suppliers are usually thought to respond to at least some variables that do not enter consumer demands. Common candidates include the prices of production inputs.

Most demand studies proceed by identifying instruments, such as input prices, and then using them to construct strike a instrumental variable estimates of the demand parameters. But one can reasonably ask what functional form is appropriate for the instruments. For example, should one use the logarithm of the input price or the level? If one assumes that $E(\epsilon \mid \text{Input Price}) = 0$, then in theory the choice should not matter for the consistency of the IV estimator. But the zero conditional mean assumption is much stronger than is presumed by many textbooks. Typically they propose assuming that the instrument (here, the level of the input price) is uncorrelated with the structural error. This weaker assumption does not imply that the logarithms of input prices are necessarily uncorrelated with the demand error. But how does the researcher know the level versus the logarithm is uncorrelated with the demand error? When the researcher has a strong prior that only one functional form satisfies the exogeneity and relevance conditions, then that argues for using that functional form. In most cases, the researcher does not have a strong prior, and therefore should be aware that the choice could matter for the estimate(s) that they will report.

[5]Of course if the researcher is only interested in predicting quantity given these variables, then OLS will under conventional assumptions yield the best linear predictor of quantity.

To illustrate the potential sensitivity of IV estimates, we use data from Epple and McCallum (2006) to estimate the demand model (2.1). Epple and McCallum use their data to illustrate, among other things, how to model the endogenous determination of demand and supply. The data set contains annual prices and quantities of broiler chickens sold in the US from 1950 to 2001. Although Epple and McCallum consider a variety of demand and supply specifications, one is the log-log specification (2.1). Specifically, in their demand specification $\ln q_c$ is the log of the per capita quantity of chicken consumed (in pounds), $\ln p$ the log of the real price of broilers, $\ln y$ the log of real per capita income, and $\ln p_b$ the log of the real price of beef (a protein substitute). In some of their specifications, Epple and McCallum treat the natural logarithm of the price of chickens, $\ln p$, as endogenous. Included in their data is the price of corn, an input into the production of chickens. In what follows, we explore use of the real corn price, p_c.

To illustrate how the instrumental variables estimates depend on the form of the instrument used, we hold the demand equation (2.1) and the data fixed and vary the functional form of the corn price instrument in the first stage regression. A convenient way to vary the instrument's functional form is to use the Box-Cox transformation

$$p_c(\lambda) = \begin{cases} \dfrac{p_c^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln p_c & \lambda = 0 \end{cases} \tag{2.2}$$

where λ controls the functional form; e.g., it nests the real price of corn ($\lambda = 1$); the natural logarithm of the real price ($\lambda = 0$); the square root ($\lambda = 0.5$); and the inverse of the real price ($\lambda = -1$).

In practice we can construct instrumental variable estimates of β_p in two regressions (or "steps"). The first stage uses ordinary least squares to regresses the endogenous price of chicken on the exogenous instruments:

$$\ln p = \pi_0 + \pi_1 \ln p_c(\lambda) + \pi_2 \ln y + \pi_3 \ln p_b + \eta. \tag{2.3}$$

where $p_c(\lambda)$ denotes the Box-Cox transformation of the corn price p_c. The predicted values from these different first stages are then substituted for the price of chicken in the demand equation regression (the "second"

stage). The second stage regressions produce the different demand elasticity estimates, $\beta_p(\lambda)$.[6]

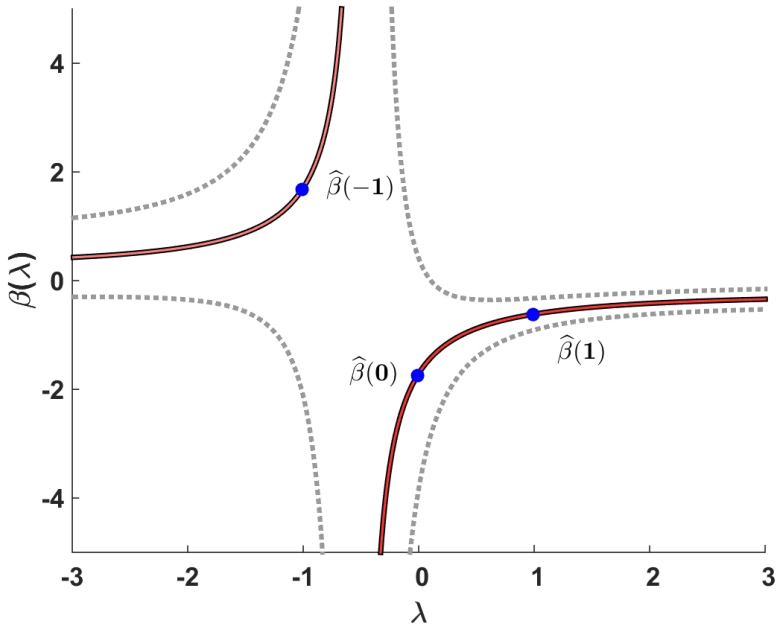

Figure 2.1: IV estimates of the price elasticity of demand versus the Box-Cox parameter. Associated 95% confidence intervals as dotted gray lines.

Figure 2.1 plots the resulting instrumental variable estimates. The dark line charts the instrumental variable estimate of $\beta_p(\lambda)$. The gray lines correspond to the upper and lower limits of a 95% confidence interval for the price elasticity estimates. What is striking about the figure is that by varying the functional form of the instrument (i.e., λ), we can get just about any value for the price elasticity. Indeed, while economic theory would predict it to be negative, there is a considerable range of transformations of the corn input price that would yield a

[6]Of course the other coefficients in the demand equation will vary with λ. Here the coefficients on log income and the substitute price exhibit similar patterns to that seen in Figure 2.1.

positive estimate.[7] Again, observe that these different estimates all are estimates of the *same* underlying population price elasticity.

The three large dots in the figure, $\hat{\beta}(-1)$, $\hat{\beta}(0)$, and $\hat{\beta}(1)$, correspond to the IV estimates one would obtain by using three common functional forms, respectively: the inverse real price of corn; the natural logarithm of the real price; or the real price of corn as instruments. Using the inverse would result in an estimated elasticity of 1.707, versus estimates of -1.711 and -0.615 for the logarithmic and linear specifications. Even restricting attention to the negative estimates, there is considerable and meaningful economic variation in the estimates. Thus, for this model and these data, the functional form of the instrument really does matter!

The functional form of the instrument is not usually something most researchers consider when using instrumental variable estimators. Instead, the focus is typically on *which* instrumental variables to use. For example, in the broiler demand model they might consider using rainfall totals or lagged consumption instead of the real price of corn. What this example demonstrates is even if one settles on specific instruments (e.g., rainfall or the real price of corn), there is still an issue of what functional form to use for that instrument in the first stage. One's choice can have a large impact on the estimated coefficient. A natural question to ask at this point is – Is this always the case? And, if it is, how might we use the data to narrow down the range of acceptable IV estimates? These questions are addressed in the remaining sections.

[7] Notice that in general $\lambda < 0$ does not cause the sign of the coefficient to change. Even when $\lambda < 0$ the instrument is increasing in the corn price.

3

Why So Sensitive?

Figure 2.1 provides a first clue as to how a researcher might go about diagnosing the sensitivity of the instrumental variable estimate to the functional form of the instrument. Specifically, it shows that the width of the estimated 95% confidence intervals for the price elasticity differ considerably with λ. The confidence intervals are widest for values of λ where the elasticity changes sign and tend to be smallest for values of λ between 2 and 4. To understand why this is so, it is useful to re-examine conditions that impact the consistency and finite sample distribution of the instrumental variable estimator.

Consider the generic linear regression model:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon = X\beta + \epsilon. \tag{3.1}$$

where $X = [\, X_1 \; X_2 \,]$ is a $N \times K = N \times (K_1 + K_2)$ matrix of regressors, ϵ is a $N \times 1$ vector of errors and $\beta = [\, \beta_1' \;\; \beta_2' \,]'$ is a $K \times 1$ vector of unknown coefficients. The coefficient vector β is not identified absent further assumptions about the data generating process. Most researchers are interested in a world where X can be manipulated independently of the other factors that determine Y. In this world, β is identified by

adding the following K moment restrictions to (3.1)

$$\underbrace{E(X_i'\epsilon_i)}_{K\times 1} = 0 \qquad i = 1, \ldots, N. \tag{3.2}$$

These moment restrictions, along with other sampling assumptions about the regressors and errors, are sufficient to prove that ordinary least squares will yield a consistent estimate of β.

As a practical matter, the only way social scientists can guarantee that $E(X_i'\epsilon_i) = 0$ holds is if they perform experiments in which they manipulate X independently of all other factors that determine Y. Since social scientists are rarely in this position, and instead must rely on observational data, questions often arise as to whether the assumption $E(X_i'\epsilon_i) = 0$ is valid. Typical concerns include the possibility of omitted variables correlated with X, simultaneity between X and Y, and measurement errors in X. There is now a large literature that explores how to test whether regressors are exogenous when the model is overidentified.[8]

Following the literature on instrumental variables, suppose that only the X_1 variables in X are suspect (i.e., potentially endogenous). Without loss of generality, we can remove the exogenous X_2's from consideration by multiplying both sides of (3.1) by the $N \times N$ projection matrix $\bar{P}_{X_2} = I - X_2(X_2'X_2)^{-1}X_2'$. This allows us to focus just on estimating β_1 in the second stage regression[9]

$$Y = X_1\beta_1 + \epsilon. \tag{3.3}$$

The instrumental variables approach presumes the existence of $L \geq K_1$ (instrumental) variables Z that do not enter equation (3.3), are correlated with X_1 and yet also are uncorrelated with ϵ. This absence of correlation can be expressed as

$$E(Z_i'\epsilon_i) = 0.$$

[8]See, for example, Davidson and MacKinnon (1993). It also has been noted that these tests are not invariant to the choice of overidentifying restrictions. A comparable point here is that they also are not invariant to the functional form of those overidentifying restrictions.

[9]Here we use the same Y and X_1 notation as in equation (3.1), however they represent $\tilde{Y} = \bar{P}_{X_2}Y$ and $\tilde{X}_1 = \bar{P}_{X_2}X_1$. This transformation does not change the points that follow and considerably simplifies the exposition.

Besides this condition, we require that the instruments be relevant, or in other words correlated with the right hand side endogenous variables (the X_1).[10] In the independent and identically distributed data case, this condition can be expressed as

$$\text{rank}\left(E\left[X_{1i}' Z_i \right] \right) = K_1 \leq L. \tag{3.4}$$

The equivalent asymptotic condition is[11]

$$\text{rank}\left[\text{plim} \frac{\sum_{i=1}^{N} X_{1i}' Z_i}{N} \right] = K_1 \leq L.$$

In practice, the researcher does not know for sure whether these relevance conditions are satisfied. However, Anderson (1984), Cragg and Donald (1993), Kleibergen and Paap (2006) and others have proposed tests for instrument relevance based on the rank of a matrix. For example, Anderson's test tests the null hypothesis that the minimum canonical correlation between X_1 and Z (given X_2) is zero (which would cause the relevance rank condition to fail). Under the null, the test statistic is asymptotically distributed central chi-squared with $L - K_1 + 1$ degrees of freedom.

In the exactly identified demand model discussed in the previous section, Anderson's test equals the sample size times the R^2 from the first-stage regression[12]

$$X_1 = Z\pi + \eta.$$

The top panel of Figure 3.1 reports how Anderson's test varies as the (Box-Cox) functional form of the instrument changes. The bottom panel reproduces Figure 2.1 for comparison. Figure 3.1 shows that Anderson's test for the relevance of the instrument is highly related to the behavior of the IV estimate. In particular, Anderson's test has a minimum at the point where the instrumental variable estimate switches sign.

[10] Recall that we have removed any correlation of X_1. Analogously we presume that we have done the same with Z. In this case, the relevance condition amounts to an assertion that the matrix of partial correlations of X_1 and Z (given X_2) is of full rank.

[11] Here X_{1i} and Z_i are row vectors corresponding to the ith observation.

[12] Again, recall that $X_1 = \bar{P}_{X_2} X_1$ and $Z = \bar{P}_{X_2} Z$.

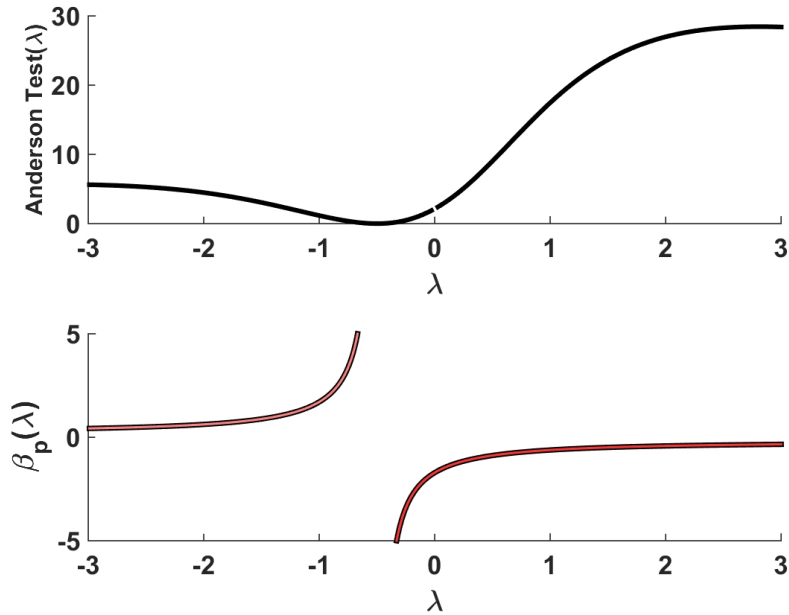

Figure 3.1: Top graph: Anderson canonical correlation LM statistic for instrument irrelevance as a function of the Box-Cox parameter. Bottom graph: IV Price elasticity estimate as a function of the Box-Cox parameter.

Given the asymptotic distribution of Anderson's test, one can examine whether ruling out IV estimates that have low test values would considerably narrow the range of IV estimates. To illustrate how such an approach might work, observe that the $\alpha = 0.9$, $\alpha = 0.95$ and $\alpha = 0.99$ critical values for Anderson's test are respectively 2.705, 3.842 and 6.635. One might choose not to accept any IV estimate that does not produce a significant Anderson test result (i.e., only choose IV estimates for which Anderson's test rejects the null hypothesis that the instruments are irrelevant).

Figure 3.2 shows how such a strategy based on a 0.9, 0.95 or 0.99 significance level would narrow the range on IV estimates produced by the Box-Cox transform. The horizontal dashed lines correspond to the chi-squared critical values 2.705, 3.842 and 6.635. The dark curve

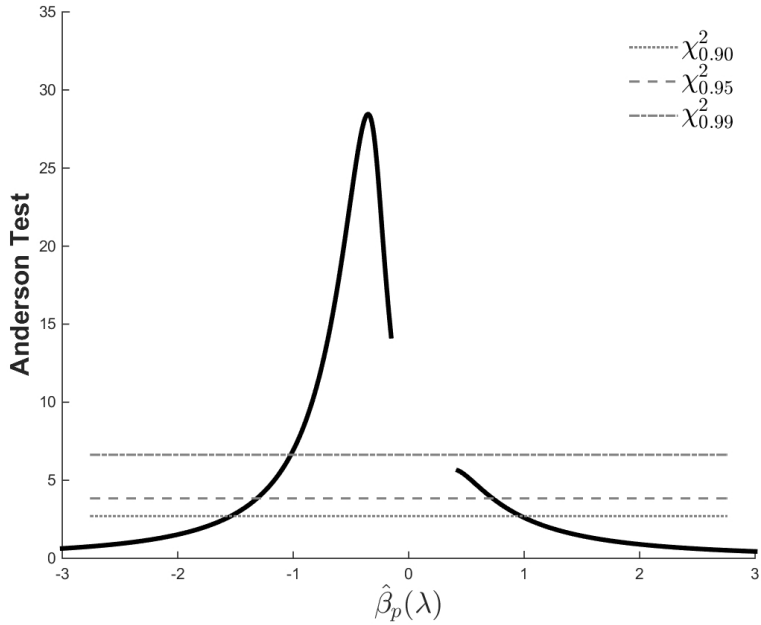

Figure 3.2: Anderson canonical correlation LM statistic for instrument irrelevance as a function of the IV price elasticity estimate.

consists of the estimated elasticities depicted in Figure 3.1 mapped to their associated Anderson canonical correlation test value. The figure shows that rejecting any estimate below the $\alpha = 0.95$ threshold (i.e., 3.842), would still generate elasticity estimates with an acceptable range from -1.30 to -0.15 and 0.43 to 0.72. Setting a higher threshold of $\alpha = 0.99$ narrows the range somewhat to just negative values, but the range of negative estimates, -1.01 to -0.15, is still wide in economic terms.

The Anderson rank test has a Lagrange Multiplier test form. Related relevance tests include the first-stage F test on the instruments that are excluded from the second stage, the related Wald test, and the Kleibergen and Paap (2006) test. The later two statistics have certain advantages, particularly when there is more than one right hand side endogenous variable or the errors are not independently and identically

distributed.[13] However, both the F test and the Cragg-Donald statistics have proven popular in this single equation model because Stock and Yogo (2005) use them as a basis for diagnosing whether the instruments are "weak".

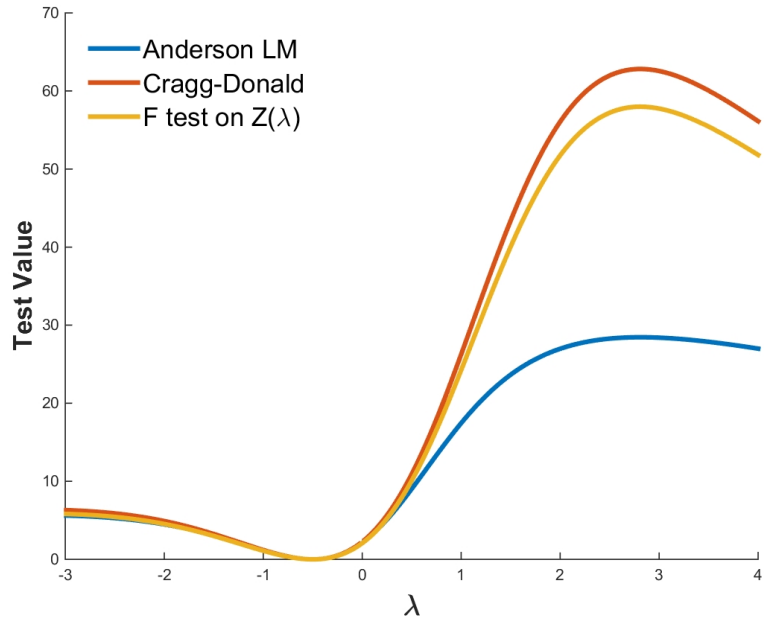

Figure 3.3: Anderson, Cragg-Donald and F tests for relevant and weak instruments as a function of the Box-Cox parameter.

Although there is no clear definition of a weak instrument, weak instruments can be thought of as a situation where the relevance condition is close to failing. In such a case, the instruments excluded from the second stage regression have little predictive power in the first stage regression. This lack of explanatory power is known to result in greater finite sample bias and increased variability of the IV estimator.[14] In some cases. the bias and variance introduced by using weak instruments

[13] See also Olea and Paueger (2013) and Sanderson and Windmeijer (2013).
[14] See for example Phillips (1983).

can result in the "cure being worse than the disease".[15] Indeed, the results in Figures 2.1 through 3.2 suggest that this is not just true from a sampling perspective, but that for the same sample, slight variations in an instrument's functional form can result in substantially different estimates. This latter point appears not to be appreciated. Another important implication is that even if the instrument(s) pass a relevance screen, one could still see a wide range of estimates produced by otherwise consistent estimators.

Staiger and Stock (1997), Stock and Yogo (2005), and Sanderson and Windmeijer (2013), among others, have proposed diagnostics for weak instruments based on first-stage regression F statistics or the Cragg and Donald statistic. The most familiar of the two is the F test applied to first-stage coefficients on the exogenous variables excluded from the second stage. Staiger and Stock suggest the informal rule that the instruments should be considered "weak", and thus that sampling issues such as bias and variance may be a concern, if this F test is less than 10. Stock and Yogo provide more detailed tables for evaluating the first-stage F test.[16]

Many econometric software packages now report test statistics for weak instruments as part of their instrumental variable routines.[17] Because these tests are related to the relevance tests, they also might provide a useful means for screening functional forms. To illustrate, Figure 3.3 plots Anderson's canonical correlation test, the Cragg-Donald statistic and the F test (here a squared t-test) on the excluded price of corn instrument, $Z(\lambda)$. The latter two tests are perhaps the most commonly used weak instrument diagnostics. The figure reveals that the tests yield similar results for values of λ less than one-half. If we use the rule of thumb that the F statistic should exceed 10 for the instrument not to be declared weak, then we would only consider values

[15] See Bound *et al.* (1995) and the original NBER Technical Working Paper No. 137, June 1993.

[16] It is important to recall that the F test may be less relevant when there is more than one right hand side endogeneous variable. See Stock and Yogo (2005) and Sanderson and Windmeijer (2013).

[17] For example, Stata's ivreg2 command has options to compute Anderson's canonical correlation coefficient, the Cragg-Donald rank statistic, F statistic and Sanderson and Windmeijer multivariate F.

of $\lambda \geq 0.5$, which implies the estimated price elasticity would lie in the range -0.88 to -0.15. This is a somewhat smaller range than implied by the $\alpha = 0.99$ cutoff for Anderson's canonical correlation statistic, but nevertheless it still is an economically wide range. Upping the F cutoff to 16.38, which according to Stock and Yogo (2005) results in a 10% maximal Wald test size distortion, narrows the range to -0.72 to -0.15, which is still a considerable range. (For example, the interval includes the OLS estimate.) Thus it appears that while weak instrument tests might prove useful at ruling out some functional form choices, they might not narrow the range considerably.

4

More on Functional Form and Efficient Instruments

To summarize the discussion to this point, the functional form of an instrument can have an important bearing on the resulting instrumental variable estimate. Absent experimentation, the researcher may be unaware that their instrumental variable estimate may depend on functional form. While tests for instrument relevance or weakness may provide useful diagnostics and screens, they do not by themselves indicate how sensitive the IV estimator may be. For instance, a specific functional form can pass, and yet other functional forms might also pass. Those two forms may yield substantially different IV estimates.

To know the potential for variation, the researcher may want to experiment with the functional form of the instrument. In the demand example, the Box-Cox transformation proved useful because all the values of the instrument were positive. Other transforms, such as the Yeo-Johnson or the modulus transformation, might also prove useful.[18] One can also experiment with instruments based on quantiles of the underlying instrumental variable. One possibility is to use an indicator

[18] See Section 5.

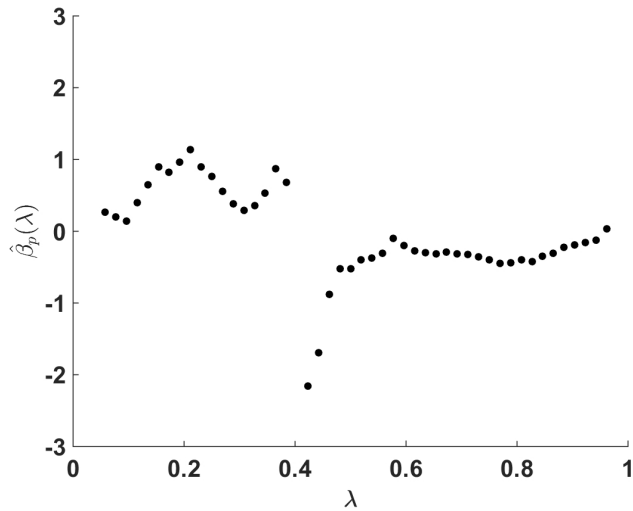

Figure 4.1: IV estimates of the price elasticity versus the quantile parameter that determines the instrument. (See equation (4.1).)

function indexed by a parameter tied to the quantiles, such as

$$h(Z, \lambda) = \begin{cases} 1 & \text{if } Z \geq \hat{F}^{-1}(\lambda) \\ 0 & \text{if } Z < \hat{F}^{-1}(\lambda) \end{cases} \tag{4.1}$$

where $\hat{F}(\cdot)$ denotes the *empirical distribution function* of Z and λ takes on a fractional value. In the demand example, this approach constructs a discrete instrument based on whether the input price is low ($Z = 0$) or high ($Z = 1$). Such a strategy is related to a strategy used by Newey (1990). Figure 4.1 shows how this choice would impact the price elasticity estimates using Epple and McCallum's data.[19] Again the pattern mimics that of Figure 2.1, even though the instrument now only is a qualitative summary of the input corn price variable.

The idea of experimenting with the functional form of the instrument raises an important issue regarding what the researcher is willing to assume about the instrument. Besides the relevance condition, the

[19]The figure varies λ discretely because the empirical distribution function has discrete steps. Not shown in the figure is an IV estimate of $\hat{\beta}_p = 33.945$ for $\lambda = 0.4$ because it distorts the scale of the figure.

instrument must also be "exogenous". A weak form of exogeneity is to assume $Cov(Z_i, \epsilon_i) = 0$. But a stronger form is needed to ensure that $Cov(h(Z_i), \epsilon_i) = 0$ for the parametric functions proposed in (2.2) and (4.1). The stronger condition $E(\epsilon_i|Z_i) = 0$ ensures this.[20]

When $E(\epsilon_i|Z_i) = 0$, the researcher in principle has many different instruments that they can use. For instance, Z works, as would Z^2, Z^3, etc. This suggests that the researcher might attempt to approximate the unknown (correct) functional form of the instrument using flexible parametric or nonparametric methods. For instance, suppose that unbenownst to the researcher the first and second stages are given by:

$$\begin{aligned} Y &= X_1\beta_1 + \xi \\ X_1 &= W\pi_0 + \omega\,. \end{aligned} \tag{4.2}$$

where W is the "true" instrument. Suppose that there exists a monotonic function $h(\cdot)$ so that $W = h(Z)$, where the Z are the instruments available (e.g., the price of corn). With conditional mean independence, we could attempt to approximate $h(Z)$ with a flexible parametric form, such as a polynomial in Z (and cross-products if there is more than one instrument). This idea follows the approach of Kelejian (1971), who explored the idea of approximating nonlinear endogenous variables with polynomials.

Table 4.1 shows what this approach would deliver using Epple and McCallum data. The columns report different IV estimates based on adding successive powers of the corn price to the first stage. For instance, the third column reports estimates based on a quadratic in the real corn price (i.e., $p_c\pi_1 + p_c^2\pi_2$). From the table we see that adding more powers moves the estimated price elasticity towards the OLS estimate of -0.265. Though this need not always happen, there are several possible explanations for why this might occur. One explanation is that if the additional terms are irrelevant, then they simply add noise to the first stage. This noise can impact the IV estimate in the second stage. One can get some sense of whether this is happening by considering whether additional terms improve the overall first-stage fit. The bottom row

[20] But in addition one may want to assume $E(\omega_i|Z_i) = 0$ as well in order to estimate $h(Z)$.

Table 4.1: Broiler Demand Equation Estimates

$$\mathbf{ln\, q_c} = \beta_0 + \beta_p \mathbf{ln\, p_c} + \beta_I \mathbf{ln\, I} + \beta_{pb} \mathbf{ln\, p_b} + \epsilon$$

		IV: (Excluded) Instruments			
Parameter	OLS	$\mathbf{p_{corn}}$	$\mathbf{p}_{corn}$ $\mathbf{p}^2_{corn}$	$\mathbf{p}_{corn}$ $\mathbf{p}^2_{corn}$ $\mathbf{p}^3_{corn}$	$\mathbf{p}_{corn}$ $\mathbf{p}^2_{corn}$ $\mathbf{p}^3_{corn}$ $\mathbf{p}^4_{corn}$
β_0	−4.680	−1.638	−4.560	−4.770	−4.647
	(0.675)	(1.343)	(0.847)	(0.806)	(0.795)
β_p	−0.265	−0.615	−0.279	−0.254	−0.268
	(0.070)	(0.149)	(0.091)	(0.086)	(0.085)
β_I	0.852	0.545	0.840	0.861	0.849
	(0.069)	(0.136)	(0.086)	(0.082)	(0.081)
β_{pb}	−0.118	−0.067	−0.116	−0.119	−0.118
	(0.084)	(0.105)	(0.084)	(0.084)	(0.084)
Anderson CC	NA	17.45	30.40	34.06	35.12
Cragg-Donald F	NA	24.24	33.08	29.12	23.42
Incremental F (First Stage)	NA	24.24	28.19	9.39	2.83

Asymptotic standard errors in parentheses.

of the table addresses this question by reporting the incremental F statistics for each new term. These statistics suggest that the first few powers of p_c matter, but starting with the quartic term more terms incrementally fail to improve the fit.[21] A second explanation is that additional terms can violate the exogeneity assumption $Cov(Z, \epsilon) = 0$, thereby leading to an inconsistent estimate. Further, it is known that the finite sample bias of the instrumental variable estimator becomes worse with more instruments ((e.g. Bound *et al.*, 1995)). A third explanation is that the more terms that are used in the reduced form, the closer the first stage fit moves to being perfect, in which case the IV estimator would return the OLS estimate.

[21]The incremental F is equal to the squared t-statistic on the last instrument added to the model. The incremental F of 2.83 for the fourth power is not statistically significant at a 5% level.

As an alternative to using a polynomial approximation, one can estimate the functional form of $W = h(Z)$ directly using nonparametric methods (provided $E(\omega|Z) = 0$). This idea follows the optimal instruments approach suggested by Newey (1990). Specifically, the idea is to recover the conditional mean using flexible semiparametric methods. For instance, one can use Robinson (1988) two-step partially linear estimator.[22] The IV estimates for the second stage using Robinson's partially-linear semiparametric approach are displayed in the last column of Table 4.2. These estimates are similar to the OLS and the cubic polynomial approximation results from Table 4.1, which are reproduced in the first two columns of Table 4.2.

Some sense of why the flexible IV results are similar can be obtained by comparing the different flexible estimates of $h(Z)$. This is done in Figure 4.2. It displays standardized estimates of $h(Z)$ for the linear and cubic IV instrument models in Table 4.1, and the semiparametric instrument model in Table 4.2.[23] The cubic and semiparametric instrument results provide similar estimates of $h(Z)$, and thus it is not surprising that they provide similar price elasticity estimates. It is also interesting to note that the Box-Cox results displayed in Figure 3.1 suggest that when $\lambda \approx 3$, and the instrument therefore is a cubic in the corn price, that the relevance and weak instrument test statistics are maximized. Interestingly, the IV results for $\lambda = 3$ are similar to those in Table 4.2.

[22] In the first step, the variables on the right hand side of the structural equation are nonparametrically regressed on the excluded instruments. For model (3.1) this amounts to constructing estimates of $E(X_1|Z)$ and $E(X_2|Z)$. The linear coefficients of the first stage are then recovered from the least squares regression

$$X_1 - \widehat{E(X_1|Z)} = (X_2 - \widehat{E(X_2|Z)})\pi_2 + \xi.$$

The estimated π_2 coefficients can then be used to construct the semiparametric estimate of $W = h(Z)$ as

$$\widehat{h(Z)} = \widehat{E(X_1|Z)} - \widehat{E(X_2|Z)}\hat{\pi}_2.$$

This estimate can then be used as an estimated "optimal" instrument in the second stage.

[23] The estimates are standardized by removing the mean from each.

Table 4.2: Broiler Demand Equation Estimates

$$\mathbf{ln\, q_c} = \beta_0 + \beta_p \mathbf{ln\, p_c} + \beta_I \mathbf{ln\, I} + \beta_{pb} \mathbf{ln\, p_b} + \epsilon$$

Parameter	OLS	IV: Excluded Instruments	
		$\mathbf{p}_{corn}$ $\mathbf{p}^2_{corn}$ $\mathbf{p}^3_{corn}$	Semiparametric $\widehat{h(Z)}$
β_0	−4.680	−4.770	−3.890
	(0.675)	(0.806)	(0.752)
β_{pc}	−0.265	−0.254	−0.356
	(0.0699)	(0.086)	(0.088)
β_I	0.852	0.861	0.772
	(0.0686)	(0.082)	(0.076)
β_{pb}	−0.118	−0.119	−0.105
	(0.0836)	(0.084)	(0.082)

Asymptotic standard errors in parentheses. The standard errors in the last column are the standard errors for 500 bootstrap replications. The first-stage estimate of $h(Z)$ is constructed using a normal kernel and least squares cross-validation to pick the bandwidth.

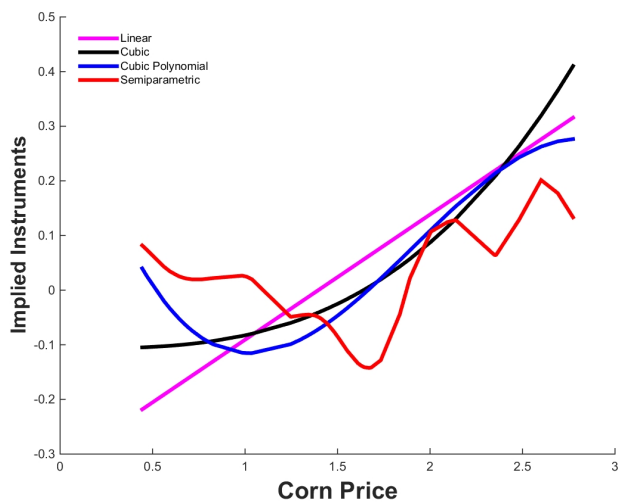

Figure 4.2: Estimates of the first-stage predictor based on different functions of the price of corn.

5

Further Evidence

So far the argument that instrument functional form matters has been largely developed using one data set, one model and one instrument. One might ask whether this sensitivity is present in other datasets and models. The answer is that the problem noted is not special. To demonstrate this, we exploit a model used by Newey (1990) to illustrate the finite sample properties of different strategies for estimating "efficient" instruments. In particular, Newey (1990) compares the use of polynomial and nonparametric first-stage estimates for the first and second stage equations

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \xi_i \\ X_{1i} &= I(\alpha_0 + \alpha_1 Z_i \geq \omega_i)\,. \end{aligned} \tag{5.1}$$

where $I(\cdot)$ is an indicator function equal to one when the condition in parentheses is true,

$$\begin{bmatrix} \xi_i \\ \omega_i \\ Z_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.7 & 0 \\ 0.7 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right),$$

and $\beta_0 = \beta_1 = \alpha_0 = 1$. This model regresses Y on a dummy variable X that is correlated with regression error ξ. Unlike the demand model, here the first stage is nonlinear in the instrument Z. Although the first stage is nonlinear, we can decompose X into its conditional expectation, $W = h(Z) = \Phi(1+Z)$, and a conditional mean zero error $\nu = X - h(Z)$ such that it resembles the reduced form equation in (4.2).[24]

To analyze the sensitivity of β_1 to the functional form of the first stage, we generate 500 simulated datasets for sample sizes of 100 and 200. We experiment with $\alpha_1 = 1$ and $\alpha_1 = 0.5$, the latter case resulting in less variation in the conditional mean of the first stage, and therefore weaker instruments. Table 5.1 reports the mean, median, and upper and lower quartiles of five different estimators of the second stage for N=100. Table 5.2 does the same for N=200. The top half of each table reports results for $\alpha_1 = 1$; the bottom reports results for $\alpha = 0.5$. The five estimators are: OLS, IV using Z as the instrument; IV using Z, Z^2 and Z^3 as instruments, IV using the optimal (first-stage conditional mean) instrument; $\Phi(1 + Z)$, and IV using a nonparametric regression estimate of the conditional mean.

The estimates in both tables suggest the following. OLS exhibits substantial bias, with the bias increasing with the weakening of the instrument. There is substantially less bias with the IV estimators, and the biases generally diminsh as the sample size increases. The bias appears worse for the flexible polynomial and nonparametric estimates. Indeed, the noise in the nonparametric estimator of the first-stage conditional mean leads to greater bias and dispersion of estimates (as measured by the interquartile range) compared to when we use the true conditional mean (Column 4). Further, the bias and variability of the flexible IV estimates appears to increase as the importance of Z in the first stage (as represented by α_1) falls. Indeed as α_1 declines to 0.5, the performance of the IV estimator using the correct conditional mean function deteriorates considerably. Thus, there appears to be some small sample evidence that estimating the first stage flexibly can lead

[24]Here $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

to greater variation in the second stage estimates. Nevertheless, looking across the results, all IV estimators perform reasonably well.

The results summarized in Tables 5.1 and 5.2 reflect statistics calculated across 500 simulated datasets. As such, they do not speak directly to the sensitivity of the IV estimates for a *given sample*. For example, in any of the 500 samples, the four IV estimates may differ substantially. To demonstrate this is the case, we follow the approach of Section 2 and explore two parametric power transformations that allow Z to take on negative values. The Yeo and Johnson (2000) transformation allows for negative values of the transformed values as follows

$$h(X,\lambda) = \begin{cases} \dfrac{(1+X)^{\lambda}-1}{\lambda} & X \geq 0,\ \lambda \neq 0 \\ \ln(1+X) & X \geq 0,\ \lambda = 0 \\ -\dfrac{(1-X)^{2-\lambda}-1}{2-\lambda} & X < 0,\ \lambda \neq 2 \\ -\ln(1-X) & X < 0,\ \lambda = 2 \end{cases} \tag{5.2}$$

Alternatively, the John and Draper (1980) modulus transformation uses

$$h(X,\lambda) = \begin{cases} \text{Sign}(X)\ \dfrac{(1+|X|)^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \text{Sign}(X)\ \ln(1+|X|) & \lambda = 0 \end{cases} \tag{5.3}$$

where $\text{Sign}(X)$ equals one if $X \geq 0$ and is minus one otherwise.

In what follows we focus on the modulus transformation results for N=100.[25] Figure 5.1 is the analog of Figure 2.1 in Section 2, except it reports statistics for different λ's across the 500 simulated data sets. The dark line is the average of the 500 estimates of β_1. It shows what one might expect based on Tables 5.1 and 5.2, that on average the choice of modulus functional form does not matter. All deliver an estimate close to one *on average*. The dotted lines in the figure are the 25th and 75th percentiles of the estimates. They appear to be narrowest somewhere in the range of λ equal to one (linear) or two (quadratic).

[25] The N=200 and the Yeo-Johnson transformation results yield largely similar general conclusions.

Table 5.1: Descriptive Statistics for the Newey Simulation Model, N=100

$$\mathbf{Y} = 1 + \mathbf{X} + \epsilon$$
$$\mathbf{X} = \mathrm{I}(\,1 + \alpha_1 \mathbf{Z}) + \omega$$

		IV: Instruments: Constant +			
	OLS	Z	Z Z^2 Z^3	$\Phi(1+\alpha_1 Z)$	$\widehat{h(Z)}$
$\alpha_1 = 1.0$					
Mean	0.158	1.006	0.990	1.028	0.903
25%tile	0.016	0.717	0.724	0.748	0.642
Median	0.149	0.992	0.966	1.004	0.890
75%tile	0.297	1.283	1.235	1.287	1.130
SE	0.207	0.430	0.406	0.428	0.382
RMSE	0.867	0.429	0.406	0.428	0.394
Weak Inst F	NA	0.96	0.38	0.97	0.99
$\alpha_1 = 0.5$					
Mean	-0.113	1.135	0.859	1.289	0.876
Lower 25%	-0.268	0.498	0.334	0.515	0.296
Median	-0.108	0.985	0.776	1.002	0.683
Upper 75%	0.029	1.566	1.343	1.544	1.226
SE	0.218	1.177	0.792	2.080	1.103
RMSE	1.134	1.183	0.804	2.098	1.109
Weak Inst F	NA	0.21	0.01	0.28	0.28

The row "Weak Inst F" reports the fraction of first-stage F tests that exceed the Stock-Yogo critical value assuming a 10% maximum Wald test size distortion threshold.

Table 5.2: Descriptive Statistics for the Newey Simulation Model, N=200

$$\mathbf{Y} = 1 + \mathbf{X} + \epsilon$$
$$\mathbf{X} = \mathrm{I}(\,1 + \alpha_1 \mathbf{Z}) + \omega$$

		IV: Instruments: Constant +			
	OLS	Z	Z Z^2 Z^3	$\Phi(1+\alpha_1 Z)$	$\widehat{h(Z)}$
$\alpha_1 = 1.0$					
Mean	0.146	1.008	0.994	1.007	0.937
25%tile	0.039	0.787	0.786	0.802	0.736
Median	0.149	1.001	0.976	0.991	0.926
75%tile	0.259	1.224	1.202	1.225	1.122
SE	0.152	0.330	0.307	0.316	0.298
RMSE	0.868	0.329	0.307	0.316	0.305
Weak Inst F	NA	1.00	0.96	1.00	1.00
$\alpha_1 = 0.5$					
Mean	-0.106	1.059	0.962	1.074	0.881
25%tile	−0.210	0.619	0.555	0.641	0.480
Median	−0.096	1.002	0.892	0.993	0.823
75%tile	0.004	1.418	1.332	1.447	1.204
SE	0.159	0.662	0.614	0.678	0.610
RMSE	1.117	0.664	0.615	0.682	0.621
Weak Inst F	NA	0.66	0.04	0.70	0.82

The row "Weak Inst F" reports the fraction of first-stage F tests that exceed the Stock-Yogo critical value assuming a 10% maximum Wald test size distortion threshold.

Figure 5.2 shows that Figure 5.1 and Tables 5.1 and 5.2 mask considerable differences in the estimates for the simulated datasets. Like Figure 2.1, it plots for a given dataset how the IV estimate varies with the λ used in the modulus transformation. The 10 lines in the figure correspond to 10 randomly generated datasets. In each case, use of an F test for weak instruments would reject the null hypothesis that the instrument is weak. Nevertheless the IV estimate can vary substantially with λ.

Figure 5.3 attempts to summarize the range of estimates one could obtain for one dataset. The summary is done across all 500 datasets. For each dataset, β_1 is estimated for λ's between -3 and 3. We take the maximum and the minimum estimates over this range and compute the difference. The empirical cumulative distribution function of these 500 differences appear in Figure 5.3 for $\alpha_1 = 0.5$ or $\alpha_1 = 1.0$ and $N = 100$. The two lines show that there can be considerable differences. For example, the median difference is 0.46 for $\alpha_1 = 0.5$ and 0.25 for $\alpha_1 = 1.0$. Again, there are substantial differences in the IV estimates even though tests suggest the instruments are not weak.

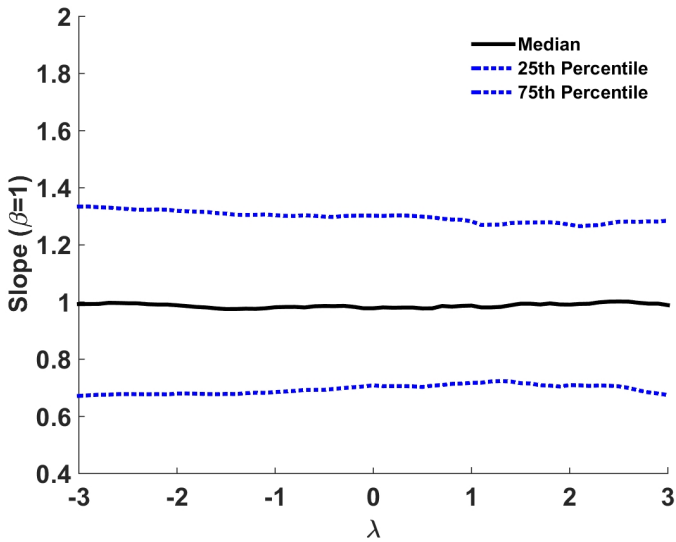

Figure 5.1: Means and average upper and lower quantiles across the 500 simulated IV estimates of $\beta_1 = 1$ (N=100).

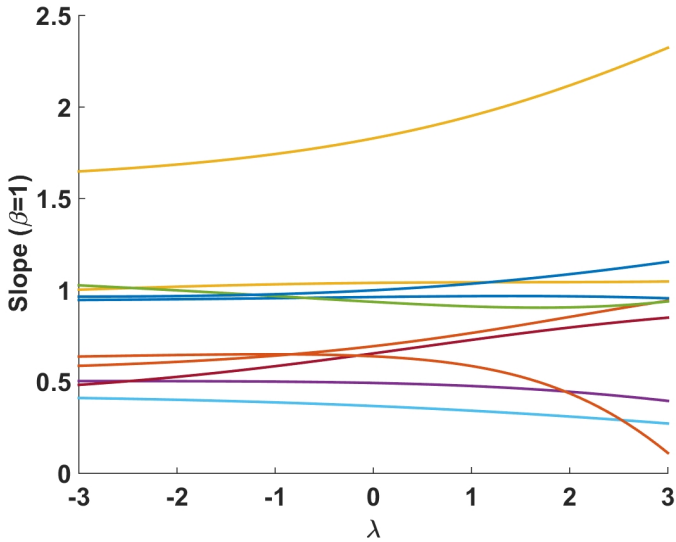

Figure 5.2: IV estimates of β_1 for ten randomly generated datasets (N=100). The parameter λ indexes the modulus transformation.

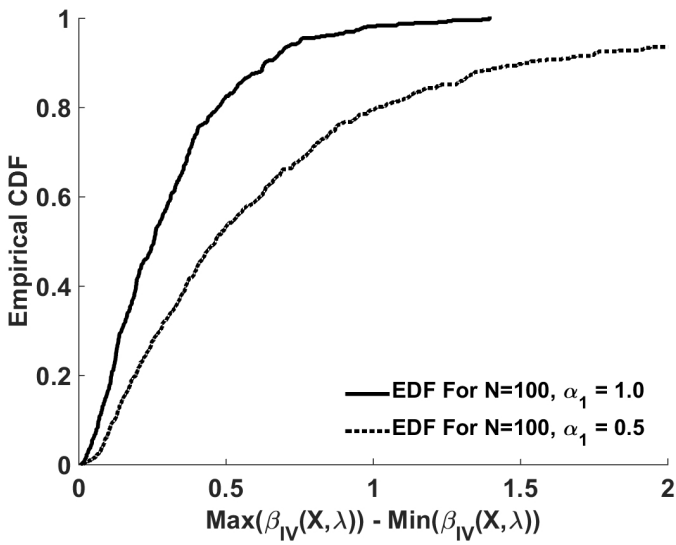

Figure 5.3: Empirical distribution functions for the difference between the maximum and minimum IV estimates of β_1 (N=100 and 200).

6

Conclusions

The main point of this article has been to urge researchers to think more carefully about not just *which* variables should serve as instruments but what *form* those variables should take as instruments. The paper has shown that the functional form of an instrument can matter for the resulting estimate. It appears to matter most when instruments would normally be judged to be weak, but even when they would not be judged weak, the instrumental variable estimates can still be sensitive to the choice of functional form in finite samples.

The potential sensitivity of the instruments to functional form could be approached in several different ways. One is for the researcher to report the range of instrumental variable estimates obtainable from different functional forms. For instance, one could experiment with different parametric transformations and report the observed range of estimates. This range could potentially be narrowed by ruling out transformations that produce weak instruments. An alternative approach is to attempt to approximate the unknown functional form of the instrument using parametric series or semiparametric methods. While these methods offer more flexibility and can help reveal nonlinearities in the first-stage conditional means, Monte Carlo evidence suggests that they

may not produce very efficient instrumental variable estimates unless the sample size is large. In practice one also has to worry that these flexible estimators presume a stronger form of exogeneity holds, and if it does not, these estimators may result in inconsistent estimates.

Acknowledgements

This paper grew out of a presentation at the Conference on Causality in the Social Sciences hosted by Stanford University's Graduate School of Business in December 2014. The author thanks John Rust for very detailed comments as well as Frank Wolak, Ali Yurukoglu and conference participants.

References

Anderson, T. 1984. *Introduction to Multivariate Statistical Analysis*. 2nd. New York: Wiley.

Bound, J., D. Jaeger, and R. Baker. 1995. "Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak". *Journal of the American Statistical Association*. 90: 443–450.

Cragg, J. and S. Donald. 1993. "Testing identifability and specifcation in instrumental variables models". *Econometric Theory*. 9: 222–240.

Davidson, R. and J. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Epple, D. and B. McCallum. 2006. "Simultaneous equation econometrics: The missing example". *Economic Inquiry*. 44(2): 374–384.

Kelejian, H. 1971. "Two-stage least squares and rconometric systems linear in the parameters but nonlinear in the endogenous variables". *Journal of the American Statistical Association*. 66: 373–374.

Kleibergen, F. and R. Paap. 2006. "Generalized reduced rank tests using the singular value decomposition". *Journal of Econometrics*. 133: 97–126.

Murray, M. 2006. "Avoiding invalid instruments and coping with weak instruments". *Journal of Economic Perspectives*. 20(4): 111–132.

Newey, W. 1990. "Efficient instrumental variables estimation of nonlinear models". *Econometrica*. 58(4): 809–837.

Olea, J. and C. Paueger. 2013. "A robust test for weak instruments". *Journal of Business and Economic Statistics*. 31: 358–368.

Phillips, P. 1983. "Exact small sample theory in the simultaneous equations model". In: *Handbook of Econometrics*. Ed. by Z. Griliches and M. D. Intriligator. Vol. I, Amsterdam: North-Holland. 449–516.

Robinson, P. 1988. "Root-N consistent semiparametric regression". *Econometrica*. 56: 931–954.

Sanderson, E. and F. Windmeijer. 2013. "A Weak Instrument F-Test in Linear IV Models with Multiple Endogenous Variables". *Working Paper* No. 13/315. University of Bristol.

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

Stock, J. H. and M. Yogo. 2005. "Testing for weak instruments in linear IV regression". In: *Identifcation and Inference for Econometric Models*. Ed. by D. W. K. Andrews and J. H. Stock. Essays in Honor of Thomas Rothenberg,. New York: Cambridge University Press. 80–108.

Yeo, I. and R. Johnson. 2000. "A new family of power transformations to improve normality or symmetry". *Biometrika*. 87: 954–959.