# Bridging the Gap Between Strength and Validity: How to increase the efficiency of weak continuous instruments

**Abstract**

Most instrumental variable analyses (IVs) employ a Two-Stage-Least-Squares (2SLS) estimator, whereby the predicted values of the instrumented variable ($X_i$) are generated by a linear regression of $X_i$ on the instrument(s) $Z_i$. Theory often dictates monotone but not necessarily linear relationships, hence the effect of $Z_i$ on $X_i$ may not be properly depicted by a linear specification. We propose a way of increasing instrument strength and thus boosting efficiency in the first stage of IV estimation. In particular we show that the predicted values obtained from a series of local non-parametric smoothing techniques are better suited to capture this effect. Monte Carlo evidence suggests that non-parametric first-stage improves efficiency without violating the orthogonality of the errors guaranteed by the OLS. Bootstrapping can account for the uncertainty of the second-level estimates. We demonstrate the usefulness of the method with three empirical applications from development economics, international political economy and political psychology.

**Keywords:** `Instrumental Variables; first-stage; Local linear regression; kernel regression; KRLS`

**Word Count:** `9,641`

Instrumental Variables estimators (IVs) are used extensively both with experimental and observational data. In experimental setups the treatment is assigned randomly but units might not abide by their treatment assignment status. In the observational world, the effect of some variable $X_i$ on some outcome $Y_i$ might not be directly identified due to selection problems. Using as-good-as-random variation of some variable $Z_i$ that is hoped to affect $Y_i$ only via its effect on $X_i$ allows researchers to draw inferences about the causal effect of $X_i$ on $Y_i$. IV estimation is used in both instances to adjust for imperfect compliance to treatment assignment ($Z_i$ does not deterministically predict $X_i$). Under a set of assumptions (see e.g. Angrist, Imbens and Rubin 1996; Bollen 2012), this procedure can produce unbiased estimates for quantities of interest, such as the average treatment effect for those who comply with their treatment assignment status.

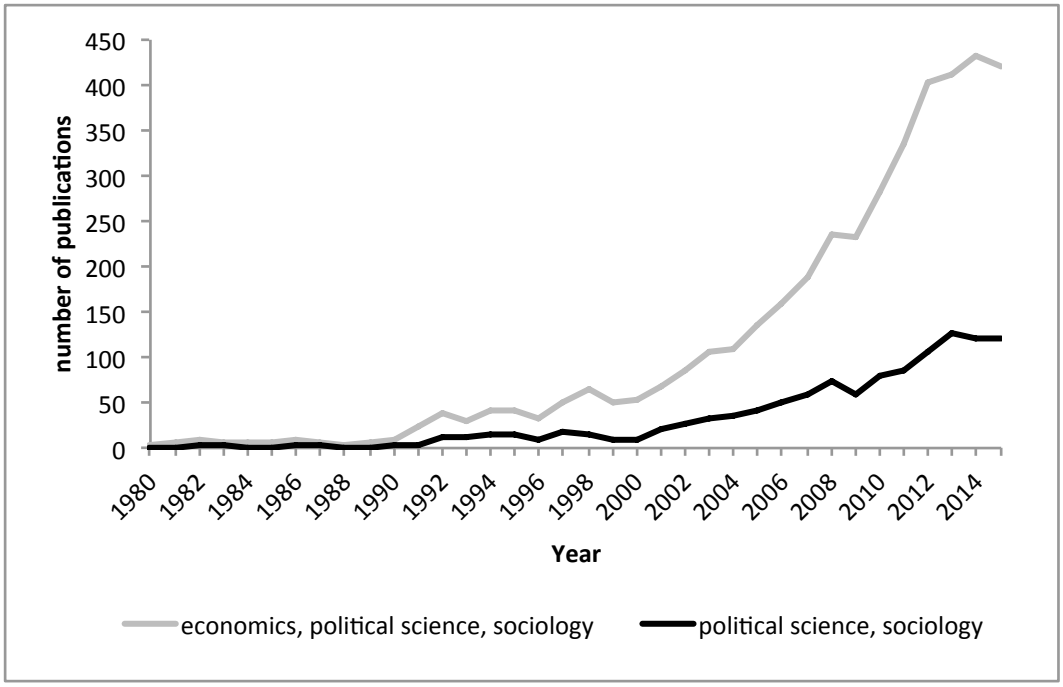
Ever-growing emphasis on clean identification designs in empirical social science research has increased the popularity of IV approaches. Figure 1 depicts this development. We have searched papers published in journals cited in the "Web of Science" which use or develop IV methods over the last three decades. We included all publications in economics, political science, and sociology. While in 1980 only a handful articles used and cited IV approaches, by 2014 more than 450 such articles were published. Since soul-searching about causality dominated economics well before other disciplines in the social sciences, it is hardly surprising it still has the lead in articles with IV designs. That said, the trend in political science and sociology remains equally impressive, with approximately 150 articles published per year using by now IVs in some part of their analysis.[1]

Despite their popularity, IVs come with a well-known problem, namely that good instruments are notoriously hard to find: typically, researchers face a trade-off between the validity

[1] For an overview of IV studies in political science see Sovey and Green (2011) who examine more than 100 articles and point to common misunderstandings in the usage of IVs in political science. Bollen (2012) also provides a very comprehensive and cogent review of IV methods as tools not only for causal inference but also for measurement and structural models. Finally, Dunning (2012, Ch. 4 and 5) offers a very intuitive introduction into both identification assumptions and estimation strategies with IV designs.

and the strength of a potential instrument. Only if an instrument is valid, that is only if the instrument shares no variance with the outcome ($Y_i$) except through its effect on the endogenous right-hand-side (RHS) variable ($X_i$), do IV approaches generate consistent estimates. Yet, valid instruments can be weak and weakness does not only lead to higher inefficiency (because the noise to signal ratio increases), but it is also likely to bias the IV estimator (Bound, Jaeger and Baker 1995). As a consequence, we have to look for instruments ($Z_i$) which are not only valid but also significantly improve our ability to predict $X_i$.[2]

**Figure 1: Number of articles in the social sciences citing IV estimation techniques**




We propose a way of increasing instrument strength and thus boosting efficiency in the first stage of the IV estimation.[3] The suggested approach is applicable when the instrument

[2] Our approach to the IV estimation is motivated by the need for unbiased causal inference, which is clearly the dominant perspective in the IV literature. Priorities might be different if for example IVs are used in a factor analytic model (see Bollen 2012).

[3] Although our attempted contribution is about efficiency, it can also help in alleviating bias due to weak instruments, as we discuss below.

can be treated as continuous. Our departure point is that most theories in social sciences are not so well specified as to predict the exact functional form of the joint distribution of $Z$ and $X$. Although this point is often neglected in applied research, one aspect that makes the IV method attractive is that there is no need to specify the structural relation between $Z$ and $X$ for the method to work (Bollen 2012; Bowden and Turkington 1990). After all, researchers are concerned with theorizing about the relationship between $X_i$ and $Y_i$. The characterization of the conditional expectation function (CEF) of $X_i$ on $Z_i$ presents a vehicle to overcome endogeneity bias.

In empirical work, however, some structure is imposed to yield an estimate of the CEF of $X$ given $Z$. More often than not, a linear specification is used to summarize this relationship. Although in some applications this linear structure may be an accurate representation of the data, it sometimes fails to capture the salient features of the mean response. Put differently, the effect of $Z_i$ on $X_i$ may not be properly depicted by a linear specification.

When the instrument is multi-valued, we suggest that researchers use a non-parametric model to obtain the predicted values of $X_i$. We present Monte Carlo evidence suggesting that non-parametric IV estimation significantly improves efficiency especially when $n$ —the number of observations— is small. We demonstrate the usefulness of this approach by employing three different non-parametric estimators in the first stage of the IV regression: the local weighted smoother (*lowess*); the kernel-based smoother (*kernel*) and kernel regularized least squares (*KRLS,* Hainmueller and Hazlett 2013).

We first provide a brief overview of IV designs, focusing mainly on the problem of weak instruments. We then present the logic underlying the use of localized smoothing techniques in IV settings. We provide Monte Carlo evidence for the efficiency gains from using a non-parametric estimate as opposed to the OLS predicted values of $X_i$ in the second stage of an IV analysis. The benefits from this approach are also illustrated with three examples from development economics, international political economy, and political psychology.

# 1 IVs and the Problem of Weak Instruments

Following Sovey and Green (2011), we refrain from presenting IV estimation using the potential outcomes notation (Angrist, Imbens and Rubin 1996). We rather stick to the simultaneous equations setting or what Bollen (2012) refers to as the auxiliary IV approach, because this framework facilitates the motivation of the key ideas. Imagine we are interested in the estimation of the effect of some explanatory factor $X_i$, on some outcome $Y_i$. The standard linear regression setup to uncover this relationship is given by equation (1):

$$Y_i = \alpha + \beta X_i + e_i \tag{1}$$

where $i$ indexes units, $\alpha$ is a constant, $\beta$ indicates the amount of change in $Y_i$ associated with a unit change in $X_i$ and $e_i$ summarizes all unmeasured influences on $Y_i$. If the data generating process stems from random assignment to different values of $X_i$ so that values of $X_i$ are independent of values of $e_i$ (e.g. $X_i$ is completely exogenous), $\beta$ would represent the causal effect of $X_i$ on $Y_i$. In observational studies however, this assumption is often difficult to sustain, even when conditioning on a series of observables.

A solution to the problem of endogenous RHS variables is provided by employing some $Z_i$, the values of which are as-good-as-randomly assigned and are also assumed to affect $Y_i$ only through their effect on $X_i$. If this is the case, $Z_i$ is a valid instrument, and thus the effect of $X_i$ on $Y_i$ can be reliably estimated by applying the IV estimand. This means starting with the first-stage equation:

$$X_i = \alpha + \gamma Z_i + u_i \tag{2}$$

Using $Z_i$ as a regressor for $X_i$, $\gamma$ is used to generate predicted values of $X_i$ which in turn are used to predict $Y_i$ in the second-stage equation:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + v_i \tag{3}$$

$\hat{\beta}_1$ provides the IV estimate of the treatment effect of $X_i$ on $Y_i$ for those units whose $Z_i$ values inform their $X_i$ values. In the absence of covariates, the IV estimand amounts to the Wald estimator, which illustrates that the magnitude of $\hat{\beta}_1$ will depend on two factors, the covariance between $X_i$ and $Y_i$ and the covariance between $Z_i$ and $X_i$:

$$\hat{\beta}_1 \equiv IV_{Wald} = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)} \tag{4}$$

It turns out that the Wald estimator will provide a consistent estimate of $\beta$ if two conditions are met. First, the researcher needs to establish that $Cov(Z_i, u_i) = 0$. This assumption may not hold if units select themselves into different values of $Z_i$ and if they do so in a way that helps also to predict their values on $Y_i$. This assumption is often referred to as the exclusion restriction, although in the potential outcomes framework it is more precisely divided into two conditions, namely ignorability (the $Z_i$ is independent of potential outcomes and treatments) and exclusion (given treatment status, $Z_i$ should not help to predict $Y_i$ in any other way than through its effect on $X_i$). Importantly, this assumption cannot be assessed empirically and it is thus left to the researcher to argue convincingly why it is likely to hold.[4] In short, the instrument $Z_i$ has to be valid.

The second condition is already apparent by looking at the denominator of equation 4. The covariance between $Z_i$ and $X_i$ needs to converge to some nonzero quantity as $n$ becomes infinite. It is necessary to emphasize that the problem here is not just a mechanic one. Even when there is a non-zero effect of $Z_i$ on $X_i$, if their correlation is small, the IV estimator may still be biased. To see why this is the case, consider the following equation, which gives the probability limit of the IV estimator (Wooldridge 2010, 514):

$$\lim_{\hat{\beta}_1} = \beta_1 + \frac{r_{Z_i e_i}}{r_{Z_i X_i}} \tag{5}$$

where $r_{ZX}$ is the Pearson correlation between $Z_i$ and $X_i$. It becomes clear that even when

[4] Pearl (2009, 274) provides a test, albeit weak, for the exclusion condition. The test is based on bounding analysis (Balke and Pearl 1997).

$Z_i$ is only trivially correlated with the unmeasured causes of $Y_i$, this correlation can create substantial bias in finite samples, if the correlation between $X_i$ and $Z_i$ is also very small.

Quite understandably, in most IV applications a lot more thinking is devoted to finding some $Z_i$ that would satisfy the first condition (validity) rather than the second. The reason is that exclusion cannot be tested empirically, whereas we can always test whether $Z_i$ exerts a significant effect on $X_i$. That said, it is still important to remember that how well $Z_i$ predicts $X_i$ does not in itself affect whether the exclusion criterion is satisfied.

Variables chosen as instruments are anticipated to have some relationship with the theoretically interesting RHS variable $X_i$. Most of the time, we only have expectations about the direction of the true CEF of $X$ given $Z$. Sometimes, our prior intuition or theory holds only locally: higher values of $Z_i$ may be associated with higher values of $X_i$ but only up to a threshold in the range of values of $Z_i$ or, reversely, after some threshold is crossed. Such patterns give rise to so-called threshold effects (Keele 2008). Such fine-grained predictions are difficult to be generated and hence a linear structure is imposed on the first stage. Sometimes this structure might be a good simplification but often it will not adequately represent the joint distribution of $X$ and $Z$. Taking these deviations from linearity into account would help our predictions of $X_i$ given $Z_i$, which in turn increases efficiency and decreases potential bias as well as boosts precision in the second stage of the estimation. The next section discusses techniques for such flexible estimation.

# 2 Non-Parametric First Stage

We suggest that instead of trying to find the proper functional form, more is to be gained by employing local estimation models, which replace a global fit between $Z_i$ and $X_i$ with a local one. One can think of the classic linear model as a special case of the family of smoothing functions. For example, in the absence of covariates, we can write the first stage as:

$$X_i = f(z_i) + v_i \tag{6}$$

where the linear fit assumes that $f = \alpha + \gamma$. Instead of imposing such a linear functional form, non-parametric smoothers allow $f$ to be estimated from the data.

There is a variety of methods for the non-parametric estimation of $f$, all of which share a common feature: instead of one parameter to summarize the relationship between $X_i$ and $Z_i$, they provide a series of such parameters, which are then connected to visualize this relationship with a plot. This strategy might be inconvenient when the ultimate goal of inference is the relationship between these two variables. It is not problematic in the context of a first-stage equation in the IV framework, where the main interest lies in obtaining the best possible predictions of $X_i$, given $Z_i$, plus a series of exogenous covariates, if needed—either to satisfy exclusion or to improve precision. Such smoothing techniques help researchers to trace the salient features of the mean response making only minimal assumptions about its distribution (Fitzmaurice, Laird and Ware 2012, 69). Without intending to provide a comprehensive account of these techniques (see Keele 2008, Ch. 2 pp.13-48), we briefly describe three of them here, kernel smoothing, local linear regression and the KRLS estimator. These are the estimators we will also use in the Monte Carlo analysis.[5]

## 2.1 Kernel Smoothing

A convenient way to think about Kernel smoothing is by imagining a scatterplot depicting the $X_i$ and $Z_i$ values of each observation. We start by defining a partially overlapping moving bin of size $h$, within which we find the $z_0$, i.e. the median observation along the range of $Z_i$. $h$ is known as the bandwidth of the kernel estimator and $z_0$ is known as the focal or evaluation point (Cleveland 1979). Within each bin, every observation is weighted according

[5] As explained also below, the list of non-parametric methods used here is far from comprehensive. The choice of smoothers is based on criteria of simplicity and easiness of exposition. We also include a more recent and more elaborate method to illustrate that the exact choice of smoother is of little practical importance.

to its distance from $z_0$, using the following function:

$$w_i = K\frac{(z_i - z_0)}{h} \tag{7}$$

where $K(\cdot)$ denotes a Kernel function that applies symmetric weights as $|z_i - z_0|$ increases. Various weights can be applied with the most common ones being the triangular (tricube) and the Gaussian kernels.[6] These weights are then used to calculate the local weighted average within each bandwidth:

$$\hat{f}(z_0) = \frac{\sum_{i=1}^{n} w_i x_i}{w_i} \tag{8}$$

Connecting each $\hat{f}(z_0)$ generates a plot that depicts the expected value of $X_i$ given $Z_i$, which can in turn be used as predictors of $\hat{Y}_i$ in Equation 3.

Similar to parametric models, there is a trade-off between bias and variance: smaller bandwidths trace the data more closely but are sensitive to noise variation (Cleveland et al. 1985). Cross-validation as well as trial-and-error methods can be useful in deciding upon the size of $h$. Although estimation varies according to the weight function and the size of the bandwidth, the differences are typically small, as shown also in the next sections.

## 2.2 Local Linear Regression

Although local weighted means allow a flexible data-driven estimation of $f(z_i)$, they do not have bias reduction properties, which can be found in least squares estimators. We can thus improve bias by replacing local means with local regression coefficients, used to generate the predicted values of $Z_i$ (see Jacoby 2000). Once again we start by defining a window width (bandwidth), which is used to slice the scatterplot into partially overlapping bins. The difference is that now instead of calculating a weighted mean, we estimate local linear

[6] The linear regression model uses a bandwidth equal to the whole range of $X_i$ and a rectangular kernel, which places the same weight to all observations.

regressions, which return the fitted values of $z_0$, the focal point.[7] Two types of weights are often added. First, similar to the kernel smoothing, observations can be weighted according to their distance from the focal point, using a tricube kernel.[8] Second, after the local linear regression predictions have been obtained, one can assess how close the fit is to each observation. Observations with small residuals are weighed more heavily than observations with large residuals. These residual-based weights are applied iteratively until no difference is found within a defined level of tolerance.[9] The final estimates produce fitted values for the $Z_0$ in each (partially overlapping) bin. The resulting plot is generated by joining the adjacent $Z_0$'s with line segments. This means that there is essentially some interpolation for observations lying between two $Z_0$'s.

Similar to kernel smoothing, a key parameter that affects the overall fit of $Z_i$ on $X_i$ is the selection of the bandwidth, which in the context of local linear regression is defined as the proportion of observations included in each bin and is known as the span of the smoother. Our Monte Carlo simulations employ a variety of different spans and the results indicate that one should avoid both under- and over-smoothing.

## 2.3 Kernel Regularized Least Squares

Kernel Regularized Least Squares (Hainmueller and Hazlett 2013, KRLS) uses machine learning methods to extend current nonlinear models by relaxing not only linearity but also addi-

[7] As discussed in Cleveland (1979), higher polynomials of $Z_i$ can be added in each local regression, although within each bin adding more than the first two polynomials does not do much to improve the estimates (Keele 2008, 28). Visual inspection might be a good guide: if the scatterplot reveals a largely monotone pattern, a linear fit is more appropriate. If local minima or maxima are observed, then a polynomial fit may be more adequate (Jacoby 2000, 587). In practical terms, all these issues will most often make little difference in the resulting predicted values obtained from this procedure.

[8] Distance-based weights are optional. If included, the estimator is known as *lowess*, whereas if excluded the smoother is known as *loess*.

[9] Non-parametric estimates are by definition ancillary statistics: their sampling distribution does not depend on the parameters of the model. This means we can try different polynomial specifications and different spans without reducing the degrees of freedom.

tivity in regression and classification problems. As Hainmueller and Hazlett (2013) suggest, *KRLS* is particularly useful when it is critical to use all the available information from one or more covariates to estimate a quantity of interest, making only minimal assumptions about the functional form. In our case, this quantity is $f(z_i)$. Similar to the local smoothing techniques we have discussed, the *KRLS* estimator allows the data to represent the relationship between outcome and covariates. However, in contrast to the previous methods it retains the characteristics of a global estimator and is thus less susceptible to the "curse of dimensionality". Regularization is then used to avoid over-fitting. $KRLS$ has been shown to perform better than other machine learning approaches for both continuous and binary outcomes. The authors suggest various ways in which $KRLS$ can be particularly useful, including propensity score estimation. We extend the list here by employing the method as a way to obtain the $E[X_i|z]$ and to use them in the second stage.

## 2.4 Other Smoothing Techniques

Our treatment of non-parametric models by no means intends to be fully encompassing. We use kernel- and regression-based local smoothing because these are the most common and simplest non-parametric approaches to summarize bivariate relationships (see e.g. Härdle 1990; Henderson and Parmeter 2015). Other smoothing techniques, such as splines, orthogonal polynomials and semi-parametric Generalized Additive models can be also applied. We choose to focus on kernel- and local regression-based estimators simply because they are the most intuitive to motivate ideas.[10] We also use $KRLS$ as a relatively new extension, which combines the logic of non-parametric inference, using machine learning-based smoothing. As is shown below, results show only minor differences but all these techniques outperform linear OLS in the first stage of an IV estimation.

[10] As Härdle (1990) points out, all smoothing methods are in an asymptotic sense equivalent to kernel smoothing. Thus, it seems logical to focus on this method while explaining how non-parametric smoothers can be used in the first stage of IV estimation.

# 3 Statistical Properties

A necessary condition for IV estimation is that the first-stage residuals be uncorrelated with the fitted values of $X_i$, i.e. $Cov(\hat{X}_i, u_i) = 0$. The need to satisfy this orthogonality criterion requires the use of OLS instead of a logit or a probit model even when $X_i$ is binary (Angrist and Pischke 2008, 190-91). The reason is that unless the true conditional expectation function is actually probit or logit, a probit or logit estimator in the first stage will not provide "clean" residuals, borrowing thus identification from the imposed functional form. Of course, we do not know what the true conditional expectation function is. This constraint results in using the 2SLS estimator, which guarantees the orthogonality assumption via the first-stage OLS estimation.[11]

It is important to emphasize that this critique does not apply to the methods discussed above. Neither kernel- nor regression-based local smoothing make any functional form assumption, thus none of these techniques gain leverage over identification at the cost of distributional assumptions about unobserved quantities. Predictions of observations are made without reference to a fixed parametric model (Härdle 1990).

The use of a locally fitted first stage may even be preferred over the 2SLS estimation on the two following grounds. First, as Bartels (1991) notes, any IV estimation incorporates modelling uncertainty. This uncertainty should be carried to the second stage as well, but this is almost never done (see Gerber, Green and Kaplan 2003). With a non-parametric fit, no modelling assumption is made during the first stage and hence there is no need to incorporate any extra uncertainty into the second stage—as the MC results on the standard errors will also reveal.

Second, it is well known that the 2SLS estimator is consistent but biased towards the OLS estimator (Bound, Jaeger and Baker 1995). A main source of this bias are weak

[11] That said, more flexible semiparametric estimators of the treatment response function are also available and can incorporate both non-linearities in the treatment response and covariates in the IV model (Abadie 2003).

instruments. Often, researchers use more than one instrument in order to collectively satisfy the first stage requirements. The problem is that the 2SLS bias is exacerbated as the number of overidentifying restrictions increases (Angrist and Pischke 2008, 154). By improving the fit of a single $Z_i$ on $X_i$, local smoothing reduces the need to use more instrumental variables. Doing so, it also spares the researcher the difficult task of justifiyng exclusion for more than one instrument. It is usually already hard enough to find a single valid instrument.

The estimation of standard errors is equally important as unbiasedness in order to draw correct inferences. In IV estimation it is crucial to adjust the standard errors of the second stage in order to incorporate the uncertainty from the first-stage predictions of $X_i$. The 2SLS estimator addresses this problem by incorporating the non-perfect match between $z_i$ and $X_i$.[12] In the absence of a single estimate to approximate the CEF of $X$ given $Z$, we need to adress this problem in a different manner. For this reason we resort to bootstrapping.

We suggest a simple, non-parametric bootstrapping procedure, wrapping the entire two-stage estimation and evaluate its relative performance in the Monte Carlo analyses. In particular, we suggest replicating the analysis across $K$ bootstrapped samples and averaging across all $\widehat{\beta_{1,k}}$'s in order to obtain the final IV estimate of the effect of $X$ on $Y$. The standard error constitutes the square root of the sum of two components: the mean error variance of the second stage (across all iterations) and, in order to incorporate the uncertainty of the first stage, the variance of the coefficients across all K samples:

$SE_{\beta_1} = \sqrt{1/k \sum_{k=1}^{K} [SE_{\beta_1}^2 + S_{\beta_1}^2]}$, where $S_{\beta_1}^2$ is the sample variance across the $k = 1, \ldots, K$ estimates.[13]

It is worth emphasizing that bootstrapping can be used even with a parametric first-stage

[12] In a typical 2SLS model this is done by adjusting the denominator of the OLS variance formula to include $\rho_{x,z}^2$, the correlation between $x_i$ and $z_i$: $Var(\hat{\beta}_1) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$. Since $0<\rho_{x,z}<1$, $Var(\hat{\beta}_{IV}) > Var(\hat{\beta}_{IV})$ (see also Wooldridge 2015, 102-3).

[13] As shown by Rubin (2004), given that the two sources of variation (the one resulting from the second-level estimation and the other stemming from the first-stage uncertainty captured through bootstrapping) are independent, the total variance is simply the sum of the two (their covariance is zero).

estimation. For instance, it might prove fruitful to take into account the first-stage uncertainty from the IV estimation in cases where a chosen instrument needs to be interacted with another covariate. An example comes from Wright (2009), who uses life expectancy, population density and a dummy for Guinea-Bissau to instrument foreign aid as a predictor of democratization. Wright hypothesizes that the aid-democratization relationship is conditioned upon the size of the dictator's support coalition. To test this hypothesis, he first obtains the predicted values of aid, based on a series of pretreatment covariates and the three variables used as instruments. Predicted aid is then interacted with the size of coalition in the second stage. The resulting standard errors stem only from the second-stage estimation and thus neglect the uncertainty from the first-stage predictions. A way to circumvent this problem would be to use the bootstrapping procedure described above.[14]

# 4 Monte Carlo Evidence

In order to substantiate our claim that using a non-parametric first-stage generates substantial efficiency gains without inducing and even reducing potential bias we run a series of Monte Carlo (MC) experiments. Clearly, it is not only important to show that our suggested approach outperforms the traditional 2SLS model in cases where the relationship between $X_i$ and $Z_i$ is non-linear or even non-monotone but that it does not perform worse than 2SLS in cases where the assumed linear relationship holds. Theoretically, this should be expected because the algorithms for lowess, as well as kernel based regression and KRLS—which we employ here—return the OLS fit if the relationship is actually linear.

In our MC experiments we compare the 2SLS model as well as the two-stage lowess, kernel and KRLS model to a simple OLS single equation model that estimates the effect

[14] Other studies have tried to mediate this problem, at the cost, however, of a potential specification error in the first-stage estimation. Indicatively, Ahmed (2012) uses the oil price in interaction with the distance of a Muslim non-oil producing country from Mecca as instruments of remittance flows. None of the two variables are included as main effects in the first-stage equation. Again, bootstrapping would enable the inclusion of both the interaction term and the main effects in the first stage regression.

of $X_i$ on $Y_i$ without taking the potential endogeneity of $X_i$ into account. Both 2SLS and the non-parametric 2-stage estimator should outperform simple OLS in case $X_i$ is indeed endogenous and $Z_i$ is a valid instrument for $X_i$.

Of course many textbooks suggest different endogeneity and over-identification tests (e.g. Durbin-Wu-Hausman test, Sargan and Hansen-Sargan statistics) in order to ensure that $X_i$ is indeed endogenous and $Z_i$ is a valid instrument for $X_i$. As has been shown elsewhere (Plümper and Troeger 2007, 2011) these tests are unfortunately notoriously powerless and only generate reliable test results in case the underlying assumptions are actually met, e.g. $Z_i$ is a valid instrument for $X_i$. We therefore refrain from using these tests in our analyses. Instead, we show how the analyzed estimators perform as we vary the degree of endogeneity of $X_i$ and $Z_i$. The results can be found in Appendix A.2.

The data generating process (DGP) used in the MC analysis follows a very simple cross-sectional set up for both stages of the model. The second stage DGP is represented by a simple linear process:

$$y_i = x_i'\beta + \epsilon, \quad i = 1, \ldots, N \tag{9}$$

While the first stage can be either linear or non-linear:

$$x_i = z_i'\gamma_1 + z_i'^2\gamma_2 + \xi_i, \quad i = 1, \ldots, N \tag{10}$$

In Equation 10 we set $\gamma_2$ either equal to zero, then the relationship between $X_i$ and $Z_i$ is linear, or non-zero, then the relationship is quadratic. We change this feature across different MC experiments. Variables are drawn as follows:

$$x_i, z_i, e_i, \xi_i \sim N(0, 1) \tag{11}$$

We are ultimately interested in estimating the effect of $X_i$ on $Y_i$ ($\beta$). To do so we use a) a simple OLS estimator of equation 9; b) a 2SLS model estimating equations 9 and 10

simultaneously; c) a non-parametric estimator (lowess, kernel, KRLS) for estimating equation 10 and using the predicted values of $X_i$ from this model to estimate equation 9. In the last set of analyses (c), we also employ the discussed bootstrapping approach and compare these SE's with those obtained parametrically from the other methods. Over the different sets of MC analyses we vary the following features of the DGP:

1. Number of observations: $n = [100, 500, 1000] \rightarrow$ efficiency for all three estimators should increase with larger number of observations.

2. Degree of endogeneity of $X : Corr(x, e) = [0.3, 0.6]$

3. Strength of the instrument $Z : Corr(x, z) = [0.3, 0.6]$

4. Validity of the instrument $Z : Corr(z, e) = [0, 0.3, 0.6]$

5. Linearity of the relationship between $X_i$ and $Z_i$: $\gamma_2 = 0$,$\gamma_2 > 0$

6. *Bandwidth*: Span of the local smoothing in the first stage (the $\alpha$ parameter): bw = [0.2, 0.5, 0.8] $\rightarrow$ the greater the bandwidth the greater the smoothing.

For each combination we run 1000 experiments by redrawing the error term and retain estimated coefficients and standard errors for each draw. We are interested in the finite sample performance of the three non-parametric estimators (Lowess, Kernel regression, KRLS) as compared to a 2SLS estimator and a single stage OLS estimator as baseline. As criterion for comparison we use the Root Mean Squared Error, which combines the bias and sampling variation of an estimator by calculating the average deviation from the true relationship:

$$RMSE = \sqrt{\frac{\Sigma(\hat{\beta} - \beta_{true})^2}{N}} = \sqrt{Var(\hat{\beta}) + [Bias(\hat{\beta}, \beta_{true})^2]} \tag{12}$$

To facilitate comparisons, we present the results for Lowess, Kernel, 2SLS and OLS below. In the Online Appendix we show the full results with the KRLS (Figures A1 to A4). To enable the presentation of the results, Table 1 displays a summary of the Figures in this

section, dividing along two dimensions: a) whether the endogeneity between $X$ and $u_i$ is weak or strong; and b) whether $Z$ is a strong or a weak instrument (i.e. whether the correlation between $X$ and $Z$ is low or high). Each permutation is implemented first without and then with covariates, as explained below. We start by first discussing the diagonal permutations and then we move to the off-diagonal scenarios.

**Table 1: Presentation of Results**

| | | Without Covariates | | With Covariates | |
|---|---|---|---|---|---|
| | | Weak Endogeneity | Strong Endogeneity | Weak Endogeneity | Strong Endogeneity |
| Instrument | Weak | Figure 2 | Figure 5 | Figure 6 | Figure 9 |
| | Strong | Figure 4 | Figure 3 | Figure 8 | Figure 7 |

## 4.1 Single Excluded Instrument

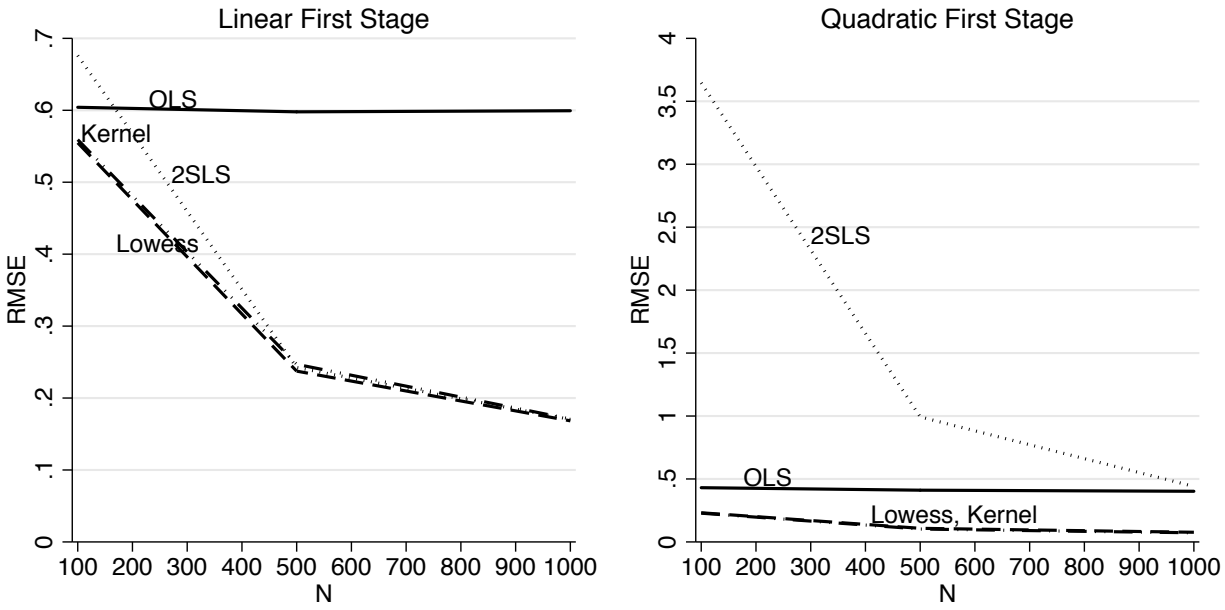
Since we focus in this study on the efficiency of the first stage estimation, i.e. the strength of the instrument, we assume in the MC analysis shown in the main text that the instrument is perfectly valid. It affects the outcome $Y_i$ only through its impact on $X_i$, thus in the DGP the correlation between $Z_i$ and $\epsilon_i$ (the error term of the second stage) is set to zero.[15]

Figure 2 depicts the root mean squared error of simple OLS, 2SLS, IV-Lowess and IV-Kernel when the instrument is valid but relatively weak and the endogeneity of $X_i$ remains relatively small as well. In case the relationship is linear (left panel of Figure 2) IV-Lowess and IV-Kernel perform at least as well as 2SLS and gain slightly in terms of RMSE if the number of observations is relatively small. Simple one stage OLS is biased as expected which translates into a higher RMSE throughout. If the relationship between the instrument $Z_i$ and $X_i$ is, however, non-linear and even non-monotone (right panel of Figure 2) IV-Lowess

[15] We also run all MC experiments for cases where the instrument is not valid and at least partially correlated with the error term of the second stage. The results, presented in the Online Appendix (Tables A1 and A2), suggest that non-parametric first-stage performs better than 2SLS even when $Z_i$ is not vaild.

and IV-Kernel clearly outperform 2SLS and gain substantially if the number of observations remains small, since in this case inefficiency of 2SLS with a weak instrument and non-capturing of the quadratic relationship between $Z_i$ and $X_i$ reinforce each other. Note the different scales of the y-axes in the two panels of Figure 2—the advantage of IV-lowess and IV-Kernel over simple OLS remains stable. If the relationship between $X_i$ and $Z_i$ is quadratic, 2SLS outperforms OLS only in the limit, if the number of observations grows very large. IV-Lowess outperforms IV-Kernel if only very slightly and IV-KRLS, especially when the true relationship between $X_i$ and $Z_i$ is linear (see Appendix Figure **??**).

**Figure 2: Weak endogeneity and weak instrument**

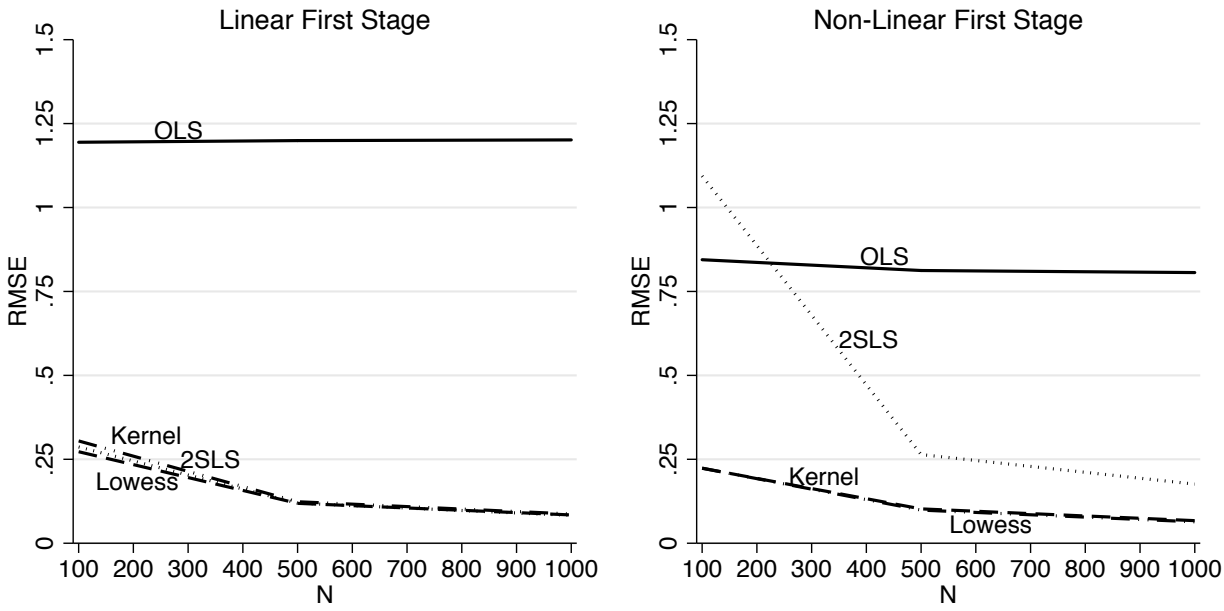



Note: Endogeneity: $corr(x, e) = 0.3$; Strength of instrument: $corr(x, z) = 0.3$; valid instrument: $corr(z, e) = 0; bw = 0.5$.

If the instrument is strong (Figure 3) and can account for most of the variation in the treatment without explaining $Y_i$ beyond $X_i$, IV-Lowess, IV-Kernel and 2SLS perform equally well and clearly outperform OLS—as theoretically expected—but only in case the relationship between the instrument and the endogenous RHS variable is linear. Once $Z_i$ exerts a non-monotone effect on $X_i$, IV-Lowess and IV-Kernel strongly outperform OLS irrespective of the number of observations and perform much better than 2SLS whenever the number

of observations remains relatively small. IV-lowess and IV-Kernel still produce significantly smaller root mean squared errors than 2SLS when the number of observations grows larger than 500 but the gains diminish with $n$ approaching infinity.

**Figure 3: strong endogeneity and strong instrument**

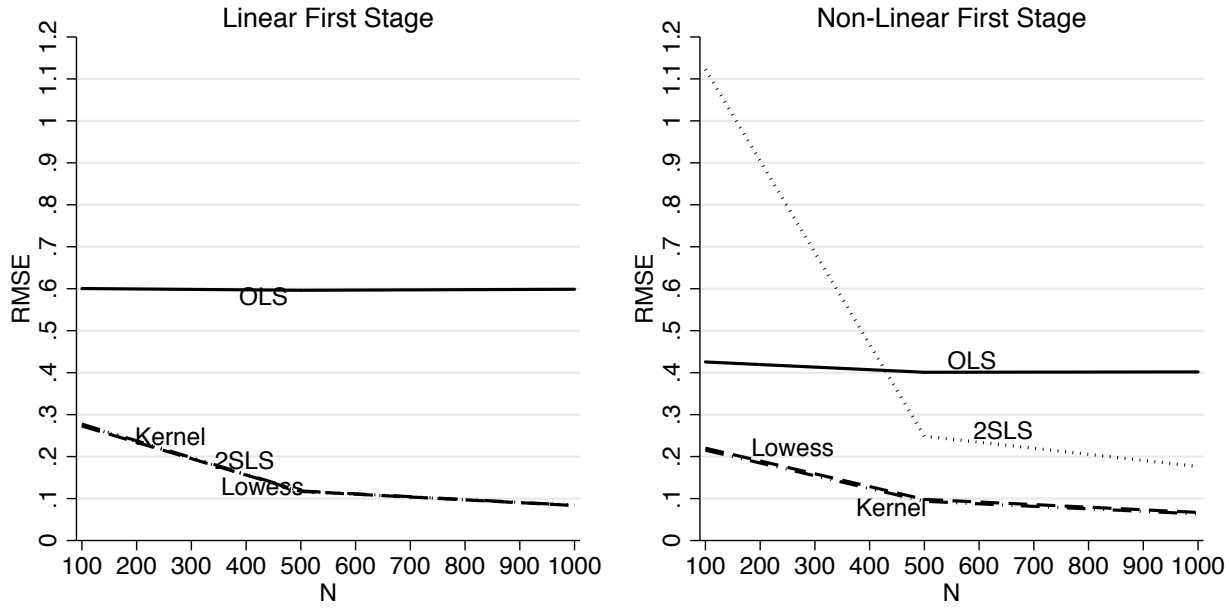



Note: Endogeneity: $corr(x, e) = 0.6$; Strength of instrument: $corr(x, z) = 0.6$; valid instrument: $corr(z, e) = 0$;$bw = 0.5$.

A similar pattern as described in Figure 3 can be detected in Figure 4, which shows the case in which the degree of endogeneity of the endogenous RHS variable ($X_i$) remains contained and the instrument is strong (see Figure 4).

The superiority of a non-parametric first stage is most obvious in cases where the treatment ($X_i$) is strongly endogenous, e.g. the correlation between $X_i$ and the error term of the second stage is large (0.6), and only a weak (but valid) instrument is available (Figure 5). This describes a common case in applied empirical research and motivates this study. The relative pattern of performance differences between the four estimators remains stable as compared to the results in Figures 2-4, but the absolute gains of IV-lowess and IV-Kernel become larger especially when the relationship between the instrument ($Z_i$) and the endogenous RHS variable ($X_i$) can be best described by a quadratic function (right panel of

**Figure 4: weak endogeneity and strong instrument**

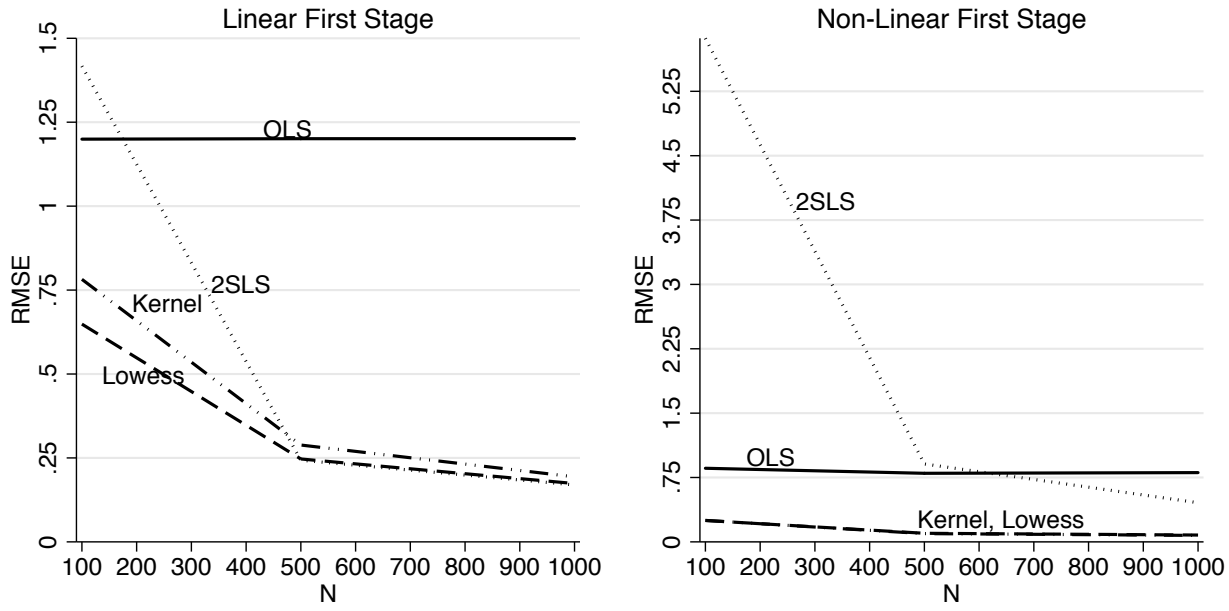



Endogeneity: corr(x,e)=0.3; Strength of instrument: corr(x,z)=0.6; valid instrument: corr(z,e)=0; bw=0.5

Figure 5). When the instrument is weak and the endogeneity of $X_i$ is strong, IV-Lowess also significantly outperforms IV-Kernel in case the relationship between $X_i$ and $Z_i$ is linear.

Overall these MC analyses seem to make a clear case for using non-parametric smoothing in the first stage of an instrumental variable estimation. There are no losses as compared to 2SLS when the relationship is indeed linear and the gains over a linear model in the first stage appear to be substantial in cases where the relationship between the instrument and the treatment become non-linear or even non-monotone. The results above also seem to put an IV-Lowess estimator ahead of a Kernel regression or KRLS because it produces simple OLS results in the case the relationship is indeed linear. A couple of qualifications have to be made though: First, this conclusion only holds if the instrument ($Z_i$) is perfectly valid. Once the instrument becomes also endogenous, the gains are much less clear. As the results in the Online Appendix (Tables **??** and **??**) show, efficiency gains of IV-Lowess over 2SLS can still be found. IV-Lowess still outperforms both simple OLS and 2SLS when the relationship is non-linear. Yet, when the relationship between the instrument and the treatment is linear and the instrument is highly invalid, a simple OLS model is the better choice though still

**Figure 5: strong endogeneity and weak instrument**




Endogeneity: corr(x,e)=0.6; Strength of instrument: corr(x,z)=0.3; valid instrument: corr(z,e)=0; bw=0.5

biased. Therefore, finding valid instruments remains one of the major tasks for instrumental variable approaches. We know that testing a potential instrument for endogeneity remains close to impossible. However, testing whether the relationship between the instrument and the treatment is non-linear can and should be done. Our preliminary results indicate that even if the instrument is not completely valid, IV-Lowess produces smaller root mean squared errors than both simple OLS and 2SLS in case $Z_i$ has a non-monotone effect on $X_i$ (Tables **??** and **??** in the Online Appendix).

Second, while Lowess has nicer bias reduction properties than Kernel smoothing and reverts to OLS if the relationship is indeed linear, overfitting can be an issue. Both Kernel smoothing and KRLS allow for optimizing the bandwidth by employing cross-validation techniques in order to resolve the trade-off between parsimony and complexity of the functional fit. We discuss the potential problem of overfitting in more detail below.

## 4.2 Instrument and Exogenous Covariates

Thus far we assumed that exclusion is satisfied without conditioning on observables. This may well not be the case. Lowess, Kernel and KRLS are equally applicable when covariates are included in both stages of the IV estimation. The difference is that now local weights, fitting windows and evaluation points are calculated within a $K$-dimensional subspace spanned by the $k$ independent variables rather than along a horizontal dimension of the bivariate scatterplot (Jacoby 2000, 599). The approach summarizes the relationship between $X_i$ and $Z_i$, controlling for the effect of $W_1, \ldots, W_k$ on $Y_i$. This can be done either with additive models (Beck and Jackman 1998; Keele 2008) or by simply running first an OLS regression on the series of covariates and then plotting the resulting residual on $Z_i$.

We perform a new set of MC's to account for the presence of exogenous covariates. Specifically our DGP now includes in addition to the endogenous $X_i$ an exogenous RHS variable $W_i$. We keep the correlation between $X_i$ and $W_i$ constant at a low level (0.2) and vary all other features as described above. The DGP of the second stage can now be described as follows:

$$y_i = x_i'\beta + w_i'\delta_1 + \epsilon_i, \quad i = 1, \ldots, N \tag{13}$$

First stage again can be either linear or non-linear:

$$x_i = z_i'\gamma_1 + z_i'^2\gamma_2 + w_i'\delta_2 + u_i, \quad i = 1, \ldots, N \tag{14}$$

For the non-parametric first stages (Lowess, Kernel, KRLS) we proceed as follows:

1. Regress the instrument $Z_i$ on $W_i$ and compute the residuals $\xi_i$.

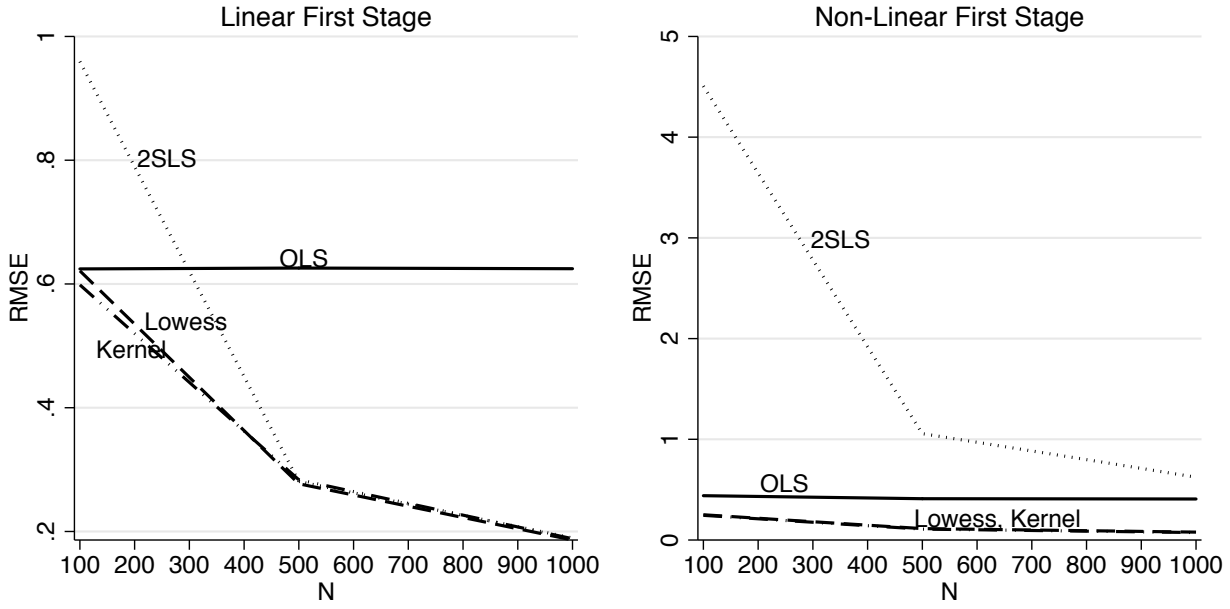
2. Lowess, kernel, KRLS of $X_i$ on residual $\xi_i$ and predict values for $X_i$ ($\hat{X}$).

3. regress $Y_i$ on $\hat{X}$ and $W_i$

In this approach the correlation between $X_i$ and $W_i$ is taken into consideration in steps 1 and 3. If $W_i$ and $Z_i$ are correlated, the residuals $\xi_i$ represent the instrument $Z_i$ cleaned from $W_i$.

Figures 6 to 9 depict the results for linear and non-linear relationships between instrument and endogenous RHS variable when an exogenous control variable exerting a linear effect on the outcome $Y_i$ is included. The results do not change much as compared to the case without control variables lending further support to our claim that using non-parametric techniques, especially a lowess smoother, in the first stage of an instrumental variable estimation robustly outperforms a two stage least squares model.

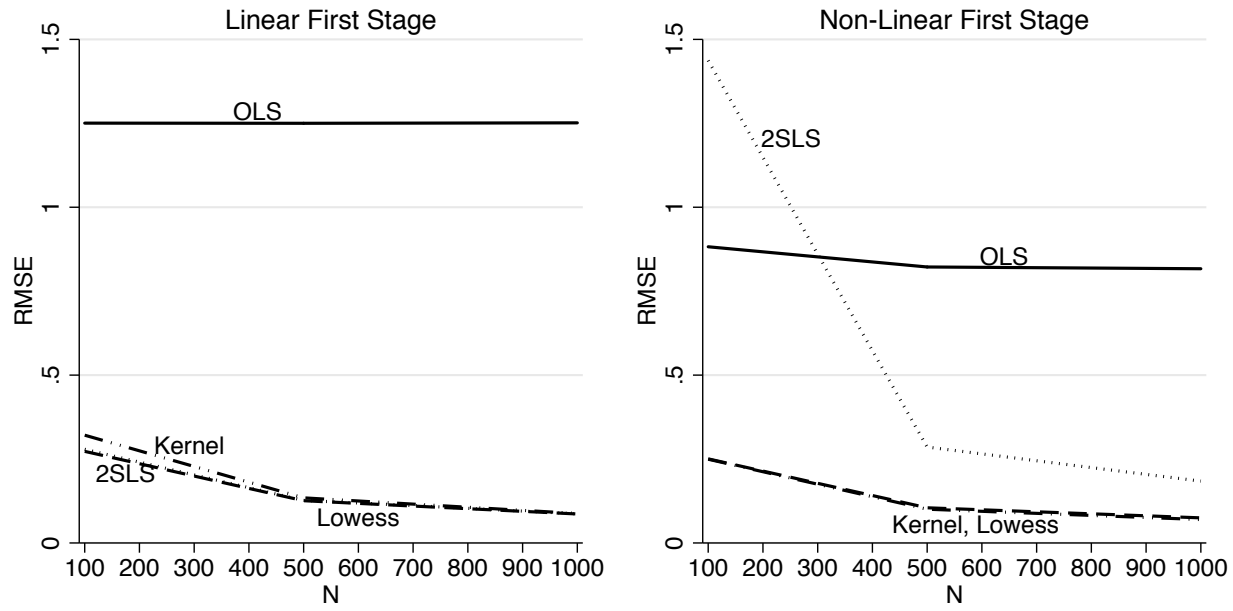
**Figure 6: Weak endogeneity and weak instrument**




Endogeneity: corr(x,e)=0.3; Strength of instrument: corr(x,z)=0.3; valid instrument: corr(z,e)=0; bw=0.5; corr(z,w)=0.2; corr(x,w)=0.2

As before, absolute gains of non-parametric first stage estimation are larger when the relationship between the endogenous RHS variable $X_i$ and the instrument $Z_i$ is non-linear and the instrument is weak.

**Figure 7: Strong endogeneity and strong instrument**



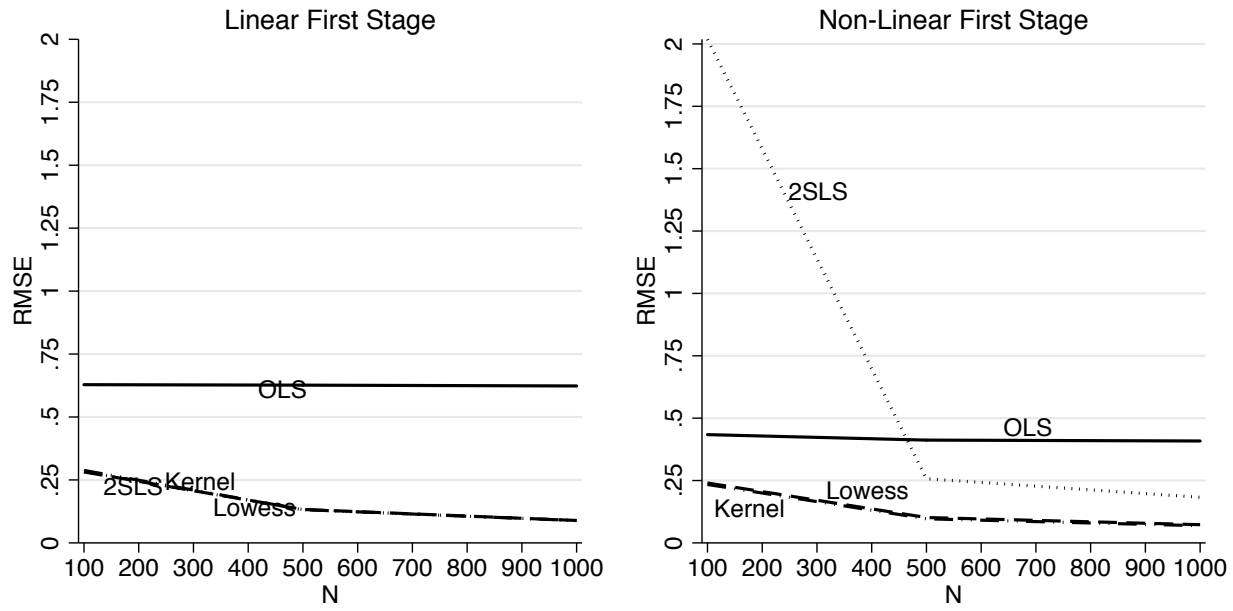

Endogeneity: corr(x,e)=0.6; Strength of instrument: corr(x,z)=0.6; valid instrument: corr(z,e)=0; bw=0.5; corr(z,w)=0.2; corr(x,w)=0.2

Whenever researchers can come up with a valid instrument that is simultaneously strong, the absolute gains of using a non-parametric technique in the first stage decrease. Yet, if the relationship between instrument and instrumented variable is non-monotone the Lowess and Kernel regression still strongly outperform a 2SLS model, especially when the number of observations remains small.

In the case where the instrument is arguably valid but only weakly related to the endogenous RHS variable, the benefit of using lowess or kernel regression in the first stage becomes massive and even in the linear case a non-parametric estimation strongly outperforms a 2SLS model when observations are limited.

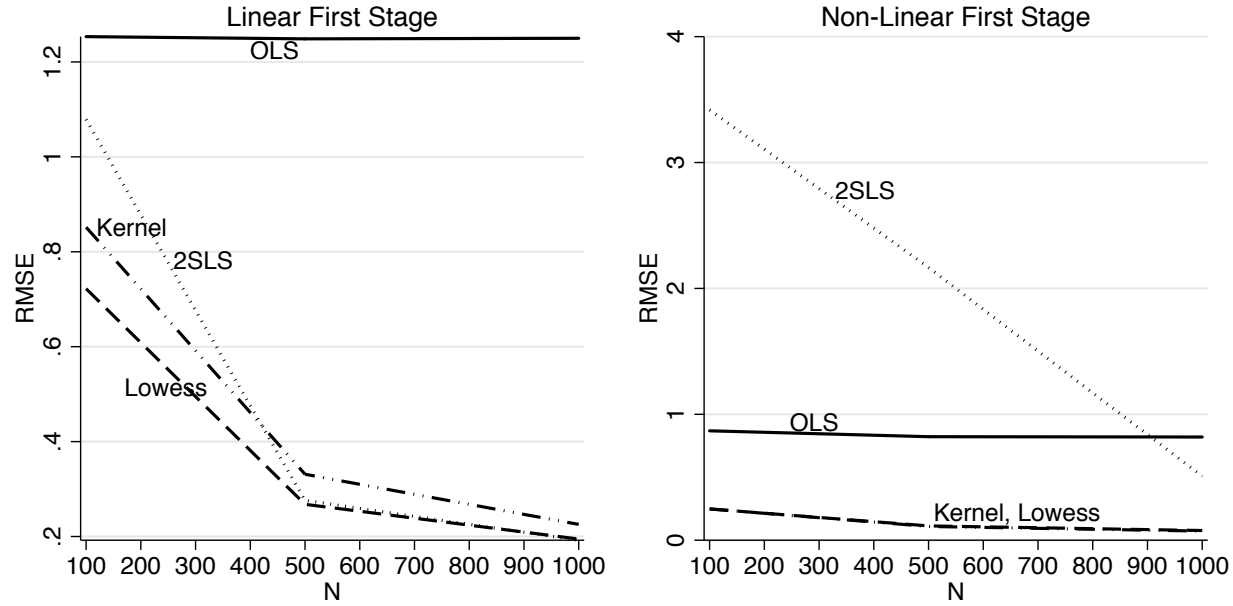
Overall, there seems to be much to gain when the first stage is non-parametrically estimated. Using a lowess or kernel smoother in the first stage produces better results than using a 2SLS estimator in most cases (and never performs worse), especially when the relationship between the instrument $Z_i$ and the endogenous RHS variable $X_i$ is non-linear. Even when the true relationship is linear, lowess and kernel perform significantly better when the

**Figure 8: weak endogeneity and strong instrument**




Endogeneity: corr(x,e)=0.6; Strength of instrument: corr(x,z)=0.6; valid instrument: corr(z,e)=0; bw=0.5; corr(z,w)=0.2; corr(x,w)=0.2

**Figure 9: strong endogeneity and weak instrument**




Endogeneity: corr(x,e)=0.6; Strength of instrument: corr(x,z)=0.3; valid instrument: corr(z,e)=0; bw=0.5; corr(z,w)=0.2; corr(x,w)=0.2

number of observations remains limited and the instrument is weak. In this case, lowess also outperforms other non-parametric estimators such as kernel regression and kernel regularized least squares.

## 4.3 A Note on Overfitting

Whenever non-parametric fitting techniques are considered, overfitting can pose a serious problem to inference. This problem equally applies to the proposed non-parametric first stage of a two stage instrumental variable model. In order to explore this potential caveat we analyze the performance of *lowess* and *kernel* when overfitting should occur with a high probability; that is when a) the correlation between instrument and instrumented variable is small and b) the chosen bandwidth is very small. We perform a series of MC simulations and display the results in the SI Appendix (Figures A.5 and A.6). We find that overfitting is a problem and when the true relationship between $Z$ and $X$ is linear it can lead to non-parametric methods performing worse than the 2SLS. In the presence of non-linearities, however, non-parametric smoothing (especially *lowess*) outperforms 2SLS, despite overfitting. In general it seems that a lowess smoother with span no smaller than 0.4 performs better or at least as good as the 2SLS estimator.

## 4.4 Standard Errors

In order to draw valid inferences from the estimation results, standard errors need to depict the variability of an estimator correctly. If estimated standard errors are too small, overconfidence is induced and leads to a higher than desirable number of Type 1 or $\alpha$ errors, rejecting the null-hypothesis even though it is correct and concluding that there is a relationship where there is none. Equally, if the computed standard errors are too big, we become under-confident in our inferences and do not reject the null as often as we should. In order to compare the performance of the examined estimators (OLS, 2SLS and IV-Lowess) with respect to their ability of correctly depicting the variability of the estimates we compute

measures of confidence for each Monte Carlo Experiment conducted above. We focus here on confidence measures for the lowess smoothing only, since it outperformed the other two non-parametric techniques in all analyzed set ups.

To assess standard errors across the estimators we calculate the underconfidence for each estimator by comparing the mean of the estimated standard errors (across the 1000 repetitions for each set of MCs) and the sampling variation of the 1000 estimated coefficients ($\beta$). We compute confidence as follows:

$$underconfidence = \frac{\sqrt{\sum_{k=1}^{K}(s.e.(\hat{\beta}^k)^2}}{\sqrt{\sum_{k=1}^{K}(\hat{\beta}^k - \bar{\hat{\beta}})^2}} \tag{15}$$

A value of 1 depicts correct standard errors, values larger than 1 indicate under-confidence (standard errors are on average too big) and values smaller than 1 point to overconfidence (estimated standard errors are too small on average).

Theoretically we expect OLS to compute correct standard errors, even though it produces biased estimates when the treatment (X) is endogenous. We also expect 2SLS to produce accurate standard errors when estimated in a full information maximum likelihood model as done here. We also compare IV-lowess SEs bootstrapped as described above: $SE_{\beta_1} = \sqrt{1/k \sum_{k=1}^{K}[SE_{\beta_1}^2 + S_{\beta_1}^2]}$. Table 2 below presents the results.

In a set up where the relationship between instrument ($Z_i$) and endogenous RHS variable ($X_i$) is linear, OLS and 2SLS generate—as expected—accurate standard errors. Quite unexpectedly, the standard errors produced by a manually computed 2-stage IV-Lowess model produce standard errors that are slightly too big and thus might induce under-confidence in estimated coefficients. Of course from a conservative statistical point of view, it is better to be under- than over-confident in estimation results. This avoids wrongly concluding that a relationship between $X_i$ and $Y_i$ exists in the underlying population. Bootstrapped IV-Lowess standard errors (last column of Table 2) differ only marginally from the original IV-Lowess standard errors and seem to slightly correct for the under-confidence.

**Table 2: Confidence of Estimated Standard Errors**

| n | Strength of Instrument (Corr($X, Z$) | Endogeneity of $X_i$: Corr(X,e) | Underconfidence OLS | 2SLS | IV-Lowess | Bootstr. Lowess |
|---|---|---|---|---|---|---|
| Linear Relationship between $Z_i$ and $X_i$ | | | | | | |
| 100 | 0.3 | 0.3 | 1.004 | 1.022 | 1.077 | 1.049 |
| 500 | 0.3 | 0.3 | 1.001 | 0.986 | 1.194 | 1.169 |
| 1000 | 0.3 | 0.3 | 0.995 | 0.995 | 1.217 | 1.201 |
| 100 | 0.6 | 0.3 | 1.005 | 1.005 | 1.192 | 1.162 |
| 500 | 0.6 | 0.3 | 1.042 | 1.02 | 1.236 | 1.229 |
| 1000 | 0.6 | 0.3 | 0.925 | 0.963 | 1.155 | 1.146 |
| 100 | 0.3 | 0.6 | 0.981 | 0.949 | 1.443 | 1.397 |
| 500 | 0.3 | 0.6 | 1.006 | 0.971 | 1.335 | 1.311 |
| 1000 | 0.3 | 0.6 | 1.018 | 1.015 | 1.381 | 1.371 |
| 100 | 0.6 | 0.6 | 0.989 | 0.959 | 1.329 | 1.266 |
| 500 | 0.6 | 0.6 | 1.004 | 1.028 | 1.362 | 1.331 |
| 1000 | 0.6 | 0.6 | 0.979 | 0.972 | 1.281 | 1.265 |
| Quadratic Relationship between $Z_i$ and $X_i$ | | | | | | |
| 100 | 0.3 | 0.3 | 0.901 | 14.98 | 1.059 | 1.031 |
| 500 | 0.3 | 0.3 | 0.963 | 4.258 | 1.119 | 1.104 |
| 1000 | 0.3 | 0.3 | 0.946 | 0.826 | 1.113 | 1.103 |
| 100 | 0.6 | 0.3 | 0.903 | 2.608 | 1.067 | 1.067 |
| 500 | 0.6 | 0.3 | 0.926 | 0.963 | 1.098 | 1.085 |
| 1000 | 0.6 | 0.6 | 0.917 | 0.941 | 1.089 | 1.093 |
| 100 | 0.3 | 0.6 | 0.770 | 93.65 | 1.096 | 1.087 |
| 500 | 0.3 | 0.6 | 0.756 | 1.780 | 1.186 | 1.178 |
| 1000 | 0.3 | 0.6 | 0.757 | 0.619 | 1.121 | 1.101 |
| 100 | 0.6 | 0.6 | 0.693 | 23.89 | 1.144 | 1.130 |
| 500 | 0.6 | 0.6 | 0.687 | 0.952 | 1.172 | 1.149 |
| 1000 | 0.6 | 0.6 | 0.733 | 0.961 | 1.159 | 1.156 |

Note: bandwidth for Lowess Smoothing is set to 0.5. Single dashed lines denote change in strength of instrument. Double dashed lines denote change in strength of endogeneity.

Once the effect of $Z_i$ on $X_i$ is non-linear (lower panel of Table 2) the IV-Lowess standard errors—both bootstrapped and not—depict the variability of the estimates correctly and remain only somewhat too large. However, simple OLS standard errors become too small and, thus, induce over-confidence and a higher probability of making Type 1 errors. This becomes increasingly so when the treatment's endogeneity grows stronger. The standard errors produced by 2SLS in the non-linear case wildly jump around, with hugely under-confident standard errors when the number of observations remains small and the instrument is weak to slight over-confidence in case $n$ grows large, endogeneity is strong and the instrument remains relatively weak.

This analysis instills additional confidence into our proposed non-parametric first-stage estimator. Not only does it outperform 2SLS in recovering MC benchmarks, but it also produces reliable SEs in the linear as well as the non-linear case.

# 5 Empirical Applications

The MC results presented above paint a very clear picture in favor of using non-parametric techniques, especially lowess smoothing in the first stage of a 2-stage IV estimator. In this section we use three empirical examples from different sub-fields in political science and economics to subject these clean findings to a plausibility test. We look at a famous example from development economics, Acemoglu, Johnson and Robinson's (2001) seminal contribution to comparative development. Further we analyze an influential study from political psychology (Krosnick and Kinder 1990) on the relationship between the timing of scandals and presidential approval. Finally we look at a recent example from International Political Economy by Dietrich and Wright (2015) studying the effects of foreign aid on democratic change in Africa.

## 5.1 Settler Mortality as Instrument for Institutional Quality

The question whether development is fostered by good political institutions or the other way around has been ubiquitous in economics and political science for several decades. The major problem faced by empirical researchers is the obvious endogeneity issue. As so often also in this case, valid instruments proved to be notoriously hard to find. In 2001 Acemoglu, Johnson and Robinson suggested to use (expected) European settler mortality as an instrument for early institutions in previously colonized countries which in turn would affect current institutions and thus levels of development. The argument is that colonizers would not invest into institutional development in areas were mortality was high due to the prevalence of diseases such as malaria. This instrument has been widely used as well as

strongly criticized because of potential endogeneity ever since.

The instrument in this case is quasi-continuous and thus our non-parametric first-stage can be easily applied. We can assume (with some certainty) that this variable is a valid instrument. However, as shown in Table 3 (bottom row) settler mortality is also a weak instrument which only explains about 5 percent of institutional quality when other geographical variables (continent dummies and latitude of capital) are taken into account. The first stage of the instrumental variable approach could therefore benefit from increasing the fit without violating the orthogonality assumption. We do this by using a lowess smoother as well as a kernel regression and KRLS in the first stage. The second stage R-squared of the lowess, kernel and KRLS IV estimation is substantially higher than that of the 2SLS. It seems that much efficiency can be gained by non-parametrically fitting the first stage of the instrumental variable model.

Figure 10 depicts graphically the first stage fit; while there seems to be a monotone relationship between settler mortality and protection against expropriation risk, this relationship is better characterized by decreasing marginal returns—a so-called threshold effect. This holds for a relatively coarse bandwidth of 0.8 (for lowess and kernel). The fit can be improved by reducing the bandwidth further.

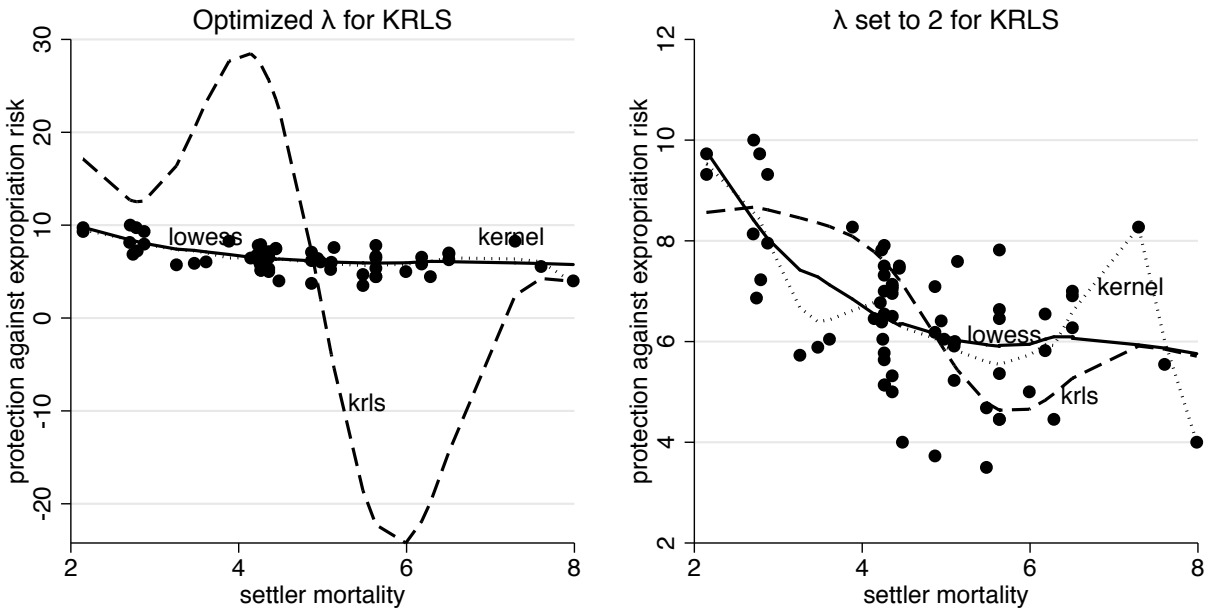
Unfortunately, in this example the optimization of $\lambda$ performed by KRLS—the parameter that governs the trade-off between complexity and parsimony of the fit—does not seem to work very well (see the left-hand panel of Figure 10). Manually setting $\lambda$ to two appears to capture the relationship better (see the right-hand panel of Figure 10). This can be also seen from the estimation results in Table 3, the coefficients for the predicted protection against expropriation seem much more in line with the estimates produced by lowess and kernel first stages when the lambda is manually chosen. We therefore recommend to always eyeball check the graphical fit of the first stage before interpreting the second stage results.

**Table 3: Settler Mortality as Instrument for Institutional Quality**

| DV of 2nd Stage: | No Exogenous Covariates | | | | | | With Exogenous Covariates | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Logged GDP 1995 | OLS | 2SLS | Lowess | Kernel | KRLS | KRLS | OLS | 2SLS | Lowess | Kernel | KRLS | KRLS |
| Instrument: Log European Settler Mortality | | | | | | | | | | | | |
| Avg. Protection against | 0.522*** | 0.944*** | | | | | 0.401*** | 1.107** | | | | |
| expropriation risk | (0.061) | (0.157) | | | | | (0.059) | (0.464) | | | | |
| Protection against | | | 0.749*** | 0.694*** | 0.031*** | 0.480*** | | | 1.078*** | 0.784*** | 5.109*** | 0.741*** |
| expropriation (predicted) | | | (0.112) | (0.097) | (0.007) | (0.073) | | | (0.227) | (0.169) | (1.073) | (0.154) |
| Lambda (KRLS) | | | | | 0.055 | 2 | | | | | 39.13 | 2 |
| Bandwidth (Lowess, | | | 0.8 | 0.5 | | | | | 0.8 | 0.5 | | |
| Kernel(optimized)) | | | | | | | | | | | | |
| Exogenous covariates | No | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Intercept | 4.660*** | 1.910* | 3.109*** | 3.511*** | 7.836*** | 4.814*** | 5.737*** | 1.44 | 1.022 | 2.985** | -25.12*** | 3.374*** |
| | (0.409) | (1.027) | (0.749) | (0.644) | (0.122) | (0.502) | (0.398) | (2.840) | (1.516) | (1.137) | (6.998) | (1.014) |
| $R^2$ | 0.54 | 0.187 | 0.418 | 0.451 | 0.257 | 0.413 | 0.714 | 0.011 | 0.631 | 0.626 | 0.631 | 0.634 |
| $n$ | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 |
| F | 72.816 | 36.394 | 44.556 | 50.990 | 21.49 | 43.62 | 28.946 | 6.847 | 19.82 | 19.376 | 19.82 | 20.07 |
| 2SLS First Stage, DV: Average protection against expropriation risk | | | | | | | | | | | | |
| Log European settler | | -0.607*** | | | -0.682** | -0.474*** | | -0.340* | | | -0.090 | -0.324 |
| mortality | | (0.126) | | | (0.239) | (0.109) | | (0.183) | | | (0.062) | (0.200) |
| Partial $R^2$ | | 0.270 | | | 0.494 | 0.395 | | 0.056 | | | 0.019 | 0.080 |

Note: columns 2 (OLS) and 3 (2SLS) exactly replicate results presented in Table 4/column 1; and columns 8 (OLS) and 9 (2SLS) exactly replicate results presented in table 4/ column 8 in Acemoglu, Johnson and Robinson (2001, 1386), *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

**Figure 10: Lowess, Kernel, and KRLS of institutional quality on settler mortality**




Endogeneity: corr(x,e)=0.6; Strength of instrument: corr(x,z)=0.6; valid instrument: corr(z,e)=0; bw=0.5; corr(z,w)=0.2; corr(x,w)=0.2

## 5.2 Timing of Scandals and Presidential Approval

Let us further empirically investigate the applicability of our suggestion by using an example from political communication. In an influential study on the importance of priming effects for presidential approval, Krosnick and Kinder (1990) use a creative design, based on the fact that the revelations about the intervention of Reagan's administration in the Iran-Contras affair took place during November 1986 and amidst the fieldwork of the American National Election Study. The authors compared respondents interviewed before November 25 and those interviewed after this date with regard to their approval of Reagan as president. The reason for this comparison is that on November 25, the Attorney General Meese announced to a national television audience that funds obtained from the sale of weapons to Iran had been given to the Contras in order to support their attempt at a coup d'état in Nicaragua. The news spread almost immediately since the issue headlined newspapers and monopolized TV broadcasts for several weeks (Krosnick and Kinder 1990, 499).

We do not provide a detailed account of the story since this can be found in the original article. What is of interest here is the potential gains from a more flexible first-stage estimation than the one given by 2SLS. Although Krosnick and Kinder use the reduced form equation to uncover priming effects, we use their design in a slightly different way. Assuming that the news caused a largely unidirectional downward shift to Reagan's electoral appeal, we want to examine the effect of emotions on Presidential affective evaluations. In particular, respondents were asked about whether Reagan evoked various different stimuli. One of them was 'angry'. We want to see whether emotions have any effect on affective presidential evaluations, measured through the typical 1-100 feeling thermometer scale. Thus, we are interested in recovering the parameter $\beta$ of Equation 1, using the thermometer scale as $Y_i$; 'angry' as $X_i$; and the date of interview as $Z_i$.

Assuming that date of interview is as good as randomly assigned, we want to compare people interviewed toward the end of the fieldwork with those who were interviewed right after the election. To be sure, we make no argument about whether exclusion here actually holds. We simply follow the reasoning driving the study by Krosnick and Kinder (1990).

The problem for us here is efficiency. Table 4 presents the results. The 3rd column uses the distinction proposed by the authors, i.e. splitting respondents into two groups according to whether they were interviewed before or after November 25. We find no difference between the two groups in their likelihood of feeling angry about Reagan. The effect is practically zero. Accordingly, it comes as no surprise that the final 2SLS estimates we get are completely uninformative (4th row of Table 4).

The problem here is that splitting respondents with regard to November 25 does not give full justice to the way events unfolded with regard to the Iran-Contras affair. In a way, the story already starts on November 3, when a Lebanese magazine reported that the US government was selling arms to Iran as part of an agreement about the return of American hostages from Lebanon. President Reagan goes public on 13 November to emphasize that 'we did not—repeat, did not—trade weapons or anything else for hostages,

**Table 4: Estimating the effect of emotions on affective Presidential evaluations, Reagan 1986.**

| DV 2nd stage: Reagan feeling thermometer | OLS | 2SLS: Before/After 25/11 | 2SLS: Linear Trend (date) | Lowess | Kernel | KRLS |
|---|---|---|---|---|---|---|
| Instrument: date of interview | | | | | | |
| Reagan makes you feel angry | -25.54*** (1.544) | | | | | |
| Angry (predicted) | | -389.05 (1603.39) | -68.41 (42.22) | -33.10** (14.99) | -29.93** (13.71) | -12.17** (5.14) |
| Intercept | 76.5*** (1.057) | 246.8 (751.4) | 96.59*** (19.81) | 79.6*** (6.881) | 78.19*** (5.952) | 69.32*** (2.507) |
| $\lambda$ KRLS | | | | | | 95.97 |
| Bandwidth (Lowess, Kernel) | | | | 0.8 | 0.8 | |
| $R^2$ | 0.223 | | | 0.005 | 0.005 | 0.003 |
| $n$ | 956 | 956 | 956 | 956 | 956 | 1924 |
| F | 273.5 | 0.059 | 2.625 | 4.875 | 4.76 | 5.61 |
| 2SLS First Stage, DV: Reagan makes you feel angry | | | | | | |
| Before/After 25/11 | | 0.009 (0.037) | | | | |
| Date of Interview | | | 0.002 (0.001) | | | 0.004*** (0.001) |
| Partial $R^2$ | | 0.001 | 0.002 | | | 0.014 |

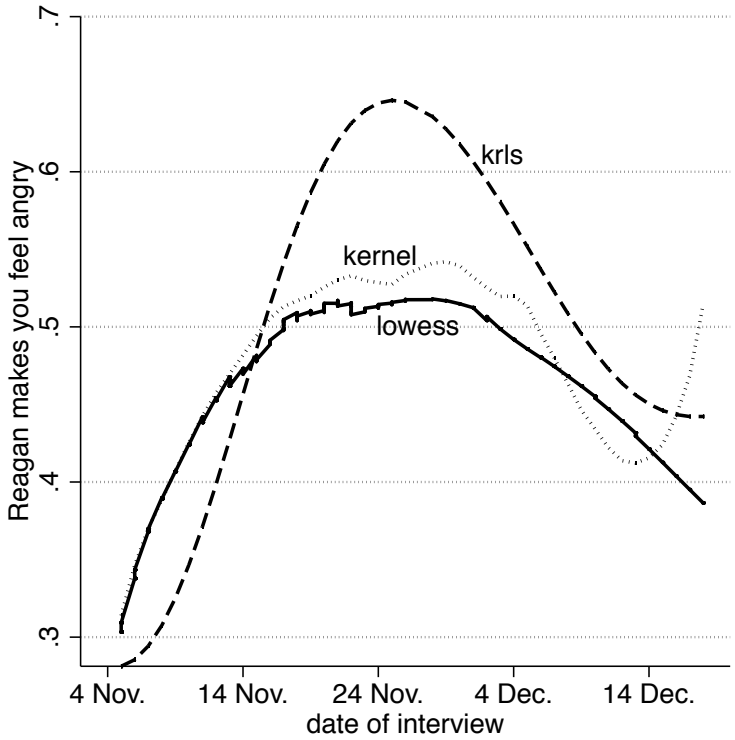
Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

nor will we' (Krosnick and Kinder 1990, 498). Therefore, the issue becomes already salient before November 25. This is also implied by the authors (Figure 1 of Krosnick and Kinder (1990)), when they discuss how and when exactly media attention to this issue escalated. The Iran-Contras affair headlined newspapers and monopolized TV broadcasts for several weeks, starting already before November 25 (Krosnick and Kinder 1990, 499). Although the sequence of the events during November makes it difficult to estimate the causal impact of a particular episode of the scandal on the decline in Reagan's popularity, it is hardly contested that when these revelations are taken together, they did affect his public image during that period. Thus, although there seems to be some relationship between date of interview and Presidential evaluations, the exact functional form of this relationship is unknown. Rather than assuming an abrupt shift after November 25, one could also consider a linear trend as a way to summarize the continuous flow of information about the Iran-Contras affair. The difference it makes in the relationship between date of interview and presidential evaluations

is shown in the 4th column of Table 4. It seems that there is some improvement over the dichotomous split used earlier. Quite expectedly, the results are more commonsensical, although still far from being informative about the effect of emotions on affective evaluations of Reagan.

The key point is that nothing stops us from trying to improve even more the fit of the instrument. For instance, one could think of a semi-curvilinear pattern, with the event becoming particularly damaging for Reagan's popularity among those respondents interviewed during the last week of November. All these are different possible scenarios. Choosing one over the other does not affect the exclusion restriction. It does affect however the efficiency with which the effect of the instrument can be estimated. Figure 11 portrays a non-monotone relationship between date of interview and feelings towards Reagan. In so doing, it showcases the need for eyeball checks in the relationship between $Z$ and $X$.

**Figure 11: Lowess, Kernel, and KRLS of interview date and feelings towards Reagan**



It seems very reasonable to assume that people get angrier the more information is disseminated about the affair, culminating with the announcement of the attorney general on 25 November. However, the impact of the scandal fades as time elapses. People interviewed more than 3 weeks later are much less affected by the scandal and revert to their long-term evaluations of Reagan.

## 5.3 Foreign Aid and Democratic Development in Africa

Whether foreign aid can effectively promote both economic as well as democratic development has been debated for decades among scholars in development and political economy. Dietrich and Wright (2015) give this question a new spin by breaking down both what is meant by foreign aid and democratic development and employ a clever IV strategy to identify their proposed mechanisms: they look at the differential effects of economic aid and democracy aid not just bluntly on general democracy scores but on whether countries allow multiple parties to stand for elections. We focus here on replicating the empirical findings for economic aid. Since the potential for endogeneity due to reversed causality is undeniable, the authors propose to use the domestic inflation rate as an instrument for economic aid. While there might be some doubt about whether this instrument meets all conditions for validity (exclusion and ignorability) we follow the authors in their argumentation and assume the instrument to be valid. In addition to proposing exogenous donor characteristics (inflation) as instruments, Dietrich and Wright propose to employ "internal" model-based instruments to identify the endogenous regressor through first-stage heteroskedasticity following Lewbel (2012). This technique allows the identification of structural parameters in regression models with endogenous regressors through selection of regressors (or part of those) as instruments that are uncorrelated with the product of heteroskedastic errors. This approach may be applied when no external instruments are available, or to supplement external instruments to improve the efficiency of the IV estimator. Dietrich and Wright clearly employ the latter option. While they are not discussing why they are employing the Lewbel model, the prob-

lem of inefficiency in the first stage of a traditional 2SLS model seems to be implicitly the driving factor.

Column 8 (2SLS with covariates) in Table 5 shows that estimating a straight forward 2SLS model with a single exogenous instrument (domestic inflation) would generate uninformative results for their main interesting variable economic aid. This is not surprising because the exogenous instrument (inflation) is extremely weak (the partial R-squared remains close to zero—0.024). The Lewbel estimator increases first stage efficiency but at the great cost of making very strong assumption about heteroscedasticity and the correlation between heteroskedastic error terms and the endogenous regressor that cannot be tested properly. Using a non-parametric first stage in this case has the great advantage of increasing first stage efficiency (in this case by factor 8) without making any further untestable assumptions.

This last example makes a strong case for using non-parametric techniques in the first stage of an IV estimation in order to increase efficiency and avoid bias. With finite samples there often exists a trade-off between unbiasedness and efficiency which is frequently resolved in favor of unbiasedness as is the case with 2SLS. Employing a non-parametric first stage estimation preserves the desirable unbiasedness and reduces the problem of inefficiency—a win-win situation.

# 6 Conclusion

Good instruments are notoriously hard to find, but crucial for drawing correct causal inferences if the regressor of interest is endogenous. In this study, we tackled one of the conditions for valid inferences in finite samples with instrumental variables, namely the strength of the instrument. We show that estimating the first stage of a two-stage IV model by non-parametric techniques can generate substantial efficiency gains over a commonly used 2SLS model, if indeed the relationship between the instrument and the endogenous regressor is non-linear. Even in the linear case this approach doesn't lose out as compared to 2SLS and

**Table 5: The Effect of Economic Aid on Move to Multiparty elections**

| DV of 2nd Stage: | No Exogenous Covariates | | | | | With Exogenous Covariates | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Multiparty politics | OLS | 2SLS | Lowess | Kernel | KRLS | OLS | 2SLS | $2SLS_{Lewbel}$ | Lowess | Kernel | KRLS |
| Instrument: Inflation | | | | | | | | | | | |
| Economic aid | 0.023 | | | | | 0.058** | | | | | |
| | (0.017) | | | | | (0.026) | | | | | |
| Economic aid | | 0.169 | 0.065 | 0.048 | 0.041 | | 0.266 | 0.108** | 0.192*** | 0.134** | 0.170*** |
| (predicted) | | (0.125) | (0.051) | (0.043) | (0.043) | | (0.18) | (0.043) | (0.069) | (0.054) | (0.066) |
| Exogenous covariates | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Lambda (KRLS) | | | | | 0.264 | | | | | | 3.526 |
| Bandwidth (Lowess, Kernel(optimized)) | | | 0.8 | 0.8 | | | | | 0.8 | 0.8 | |
| Exogenous covariates | No | No | No | No | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Intercept | 0.048 | -0.493 | -0.107 | -0.045 | -0.020 | -0.135 | -1.459 | -0.452 | -0.607 | -0.387 | -0.475 |
| | (0.066) | (0.465) | (0.189) | (0.161) | (0.159) | (0.290) | (1.174) | (0.361) | (0.385) | (0.344) | (0.362) |
| $R^2$ | 0.005 | | 0.004 | 0.003 | 0.003 | 0.039 | . | 0.029 | 0.046 | 0.042 | 0.044 |
| $n$ | 370 | 370 | 370 | 370 | 370 | 370 | 370 | 370 | 370 | 370 | 370 |
| F | 1.777 | 1.811 | 1.616 | 1.242 | 0.931 | 2.099 | 1.49 | 2.264 | 2.503 | 2.286 | 2.357 |
| 2SLS First Stage, DV: Economic Aid | | | | | | | | | | | |
| Inflation | | 0.141** | | | 0.166** | | 0.100*** | 0.156*** | | | 0.130** |
| | | (0.049) | | | (0.078) | | (0.033) | (0.028) | | | (0.060) |
| Partial $R^2$ | | 0.220 | | | 0.181 | | 0.024 | | | | 0.108 |

Note: Columns 7 (multivariate OLS) and 9 (Lewbel 2SLS) exactly replicate the empirical results in Table 1, columns 1 and 2 in Dietrich and Wright (2015), *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

even outperforms 2SLS if the number of observations remains small.

We support these claims by a series of Monte Carlo experiments taking different features of the DGP into account. Our MC results overall favor lowess smoothing in the first stage over other estimators such as Kernel regressions or KRLS. However, we do not want to push too hard for an IV-Lowess estimator but would always encourage researchers to compare the results for different non-parametric first stages. We also demonstrate the usefulness of the proposed approach by applying it to examples from political science and economics.

Increasing efficiency in an instrumental variable estimation is only one side of the medal. As our MC experiments show, it still remains crucial to find valid instruments. This is probably the harder task since it needs to be based almost exclusively on theoretical arguing because endogeneity tests are notoriously powerless and do not generate reliable guidance.

In sum, the IV-Lowess procedure cane be very useful in applications from comparative politics and political economy, which typically use quasi-continuous instruments. Importantly, increasing instrumental efficiency does not only help to satisfy the first stage assumptions. It also facilitates the exclusion restriction, not only because weak instruments bias the final IV estimates but also because this bias increases with the number of instruments used in the analysis. Aware of the difficulty to find valid instruments, we hope this non-parametric procedure will help applied researchers to use their instruments more efficiently.

# References


Abadie, Alberto. 2003. "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics* 2(2):231–63.

Acemoglu, Daron, Simon Johnson and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91(5):1369–1401.

Ahmed, Faisal Z. 2012. "The perils of unearned foreign income: Aid, remittances, and government survival." *American Political Science Review* 106(01):146–165.

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

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bartels, Larry M. 1991. "Instrumental and" quasi-instrumental" variables." *American Journal of Political Science* 35(3):777–800.

Beck, Nathaniel and Simon Jackman. 1998. "Beyond linearity by default: Generalized additive models." *American Journal of Political Science* 42(2):596–627.

Bollen, Kenneth A. 2012. "Instrumental variables in sociology and the social sciences." *Annual Review of Sociology* 38:37–72.

Bound, John, David A Jaeger and Regina M Baker. 1995. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American statistical association* 90(430):443–450.

Bowden, Roger J and Darrell A Turkington. 1990. *Instrumental variables*. Vol. 8 Cambridge University Press.

Cleveland, William S. 1979. "Robust locally weighted regression and smoothing scatterplots." *Journal of the American statistical association* 74(368):829–836.

Cleveland, William S et al. 1985. *The elements of graphing data*. Wadsworth Advanced Books and Software Monterey, CA.

Dietrich, Simone and Joseph Wright. 2015. "Foreign aid allocation tactics and democratic change in Africa." *The Journal of Politics* 77(1):216–234.

Dunning, Thad. 2012. *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press.

Fitzmaurice, Garrett M, Nan M Laird and James H Ware. 2012. *Applied longitudinal analysis*. Vol. 998 John Wiley & Sons.

Gerber, Alan S, Donald P Green and Edward H Kaplan. 2003. The illusion of learning from observational research. In *Field Experiments and their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, ed. Dawn Langan Teele. Yale University Press pp. 9–32.

Hainmueller, Jens and Chad Hazlett. 2013. "Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach." *Political Analysis* 22(3):143–68.

Härdle, Wolfgang. 1990. *Applied nonparametric regression*. Number 19 Cambridge university press.

Henderson, Daniel J and Christopher F Parmeter. 2015. *Applied nonparametric econometrics*. Cambridge University Press.

Jacoby, William G. 2000. "Loess:: a nonparametric, graphical tool for depicting relationships between variables." *Electoral Studies* 19(4):577–613.

Keele, Luke John. 2008. *Semiparametric regression for the social sciences*. John Wiley & Sons.

Krosnick, Jon A and Donald R Kinder. 1990. "Altering the foundations of support for the president through priming." *American Political Science Review* 84(02):497–512.

Lewbel, Arthur. 2012. "Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models." *Journal of Business & Economic Statistics* .

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Plümper, Thomas and Vera E Troeger. 2007. "Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects." *Political Analysis* 15(2):124–139.

Plümper, Thomas and Vera E Troeger. 2011. "Fixed-effects vector decomposition: properties, reliability, and instruments." *Political Analysis* 19(2):147–164.

Rubin, Donald B. 2004. *Multiple imputation for nonresponse in surveys*. Vol. 81 John Wiley & Sons.

Sovey, Allison J and Donald P Green. 2011. "Instrumental variables estimation in political science: A readers' guide." *American Journal of Political Science* 55(1):188–200.

Wooldridge, Jeffrey. 2015. *Introductory econometrics: A modern approach*. Nelson Education.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. MIT press.

Wright, Joseph. 2009. "How foreign aid can foster democratization in authoritarian regimes." *American Journal of Political Science* 53(3):552–571.