Tutorial 2: Instrumental Variables

Alexander Wintzéus*

February 7, 2024

In this tutorial, we will discuss estimation using instrumental variables in the context of the linear regression model. To introduce instrumental variables methods in `Stata`, we will continue with the replication of the main findings in Acemoglu, Johnson, and Robinson (2001). As before, the exercises are collected in the first part of this document. A brief theoretical summary is presented in the second part.

1 Application

Recall that Acemoglu et al. (2001) are interested in quantifying the effect of institutions on income per capita by estimating the following linear equation:

$$y_i = \mu + \alpha R_i + \mathbf{x}_i'\gamma + \varepsilon_i \tag{1}$$

where y_i denotes the log of income per capita in country i, R_i is a measure of protection against expropriation of private property, $\mathbf{x}_i$ is a vector of covariates, and ε_i is a random error term. The parameter of interest is α, the effect of institutions on (log) income per capita.

As discussed in the first tutorial, there are reasons to suspect that the exogeneity assumption is violated. For example, there may be a number of different reasons why countries differ both in their institutions and income levels. If we are unable to control for all of these factors, our OLS estimate of α may be suffering from *omitted variable bias*. It is also quite likely that richer economies can choose or afford better institutions. In this case, our estimate of the effect of institutions on income may partly be capturing the effect of income on institutions. As such, *reverse causality* may also be a problem to take serious in this context.

To overcome these issues, the authors employ an Instrumental Variables (IV) strategy. To this end, Acemoglu et al. (2001) propose a theory of institutional differences among countries colonized by Europeans to derive a possible source of exogenous variation. European countries implemented different colonization policies which created different sets of institutions. In some countries, Europeans set up *extractive states* (e.g., Belgian colonization of Congo). These institutions did not

*KU Leuven, Department of Economics. alexander.wintzeus@kuleuven.be

introduce much protection for private property as their main purpose was to transfer as much resources of the colony to the colonizer. In other colonies, however, many Europeans migrated and settled (e.g., Australia, Canada, and the U.S.). Here, the colonizers tried to replicate their European institutions, with a strong emphasis on private property protection.

The colonization strategy employed by the European powers was influenced by the feasibility of settlements. In countries where the disease environment was not favorable, the possibility of creating a sustainable settlement was limited. As a consequence, the formation of an extractive state in these places was more likely. Assuming that the colonial state and its institutions persisted, the authors claim they can use mortality rates expected by the *first* European settlers in the colonies as an instrument for *current* institutions. To obtain such a measure, the authors use data on the mortality rates of soldiers, bishops, and sailors stationed in the colonies between the seventeenth and nineteenth century.

Load the data file `AJR2001-AER.dta` into `Stata` and use the `browse` and `describe` commands to get an overview of its contents.

1. For the proposed instrument to be valid, two conditions must be met. First, the instrument must be *relevant*. That is, it must be (strongly) correlated with the endogenous independent variable. Second, it must satisfy an *exclusion restriction*. This boils down to the requirement that the instrument affects the dependent variable only through the endogenous independent variable.[1]

   (a) Compute the correlation coefficient between `setmort` (log settler mortality rate) and `risk` (average protection against expropriation risk) using the `correlate` or `pwcorr` command. Use the `sig` option to let `Stata` provide significance levels. Do you think the relevance condition is met?

   (b) Plot the empirical relationship between `loggdp` and `setmort` using the `scatter` command. Add a linear fit using the `lfit` command. Can this graph inform us about the validity of the exclusion restriction? If not, what can we learn from this graph?

   (c) Why is the exclusion restriction fundamentally untestable?

2. Let us now revisit the estimation of the effect of institutions on income per capita by OLS.

   (a) Create a global macro containing a list of the control variables included in the regression using the `global` command.

   (b) Estimate equation (1) by OLS using the `regress` command. Perform the estimation with and without heteroskedasticity-robust standard errors and store the estimates using the `estimates store` command.

[1] Strictly speaking, the exclusion restrictions imposes the instrument to be uncorrelated with the error term. A corollary of this condition and the relevance condition is that the instrument only affects the dependent variable through the endogenous independent variable. Often, the latter is referred to as the exclusion restriction.

(c) Under the assumption that the proposed instrumental variable is valid, how do you think the estimate of α would change if we would move from OLS to IV?

3. We will now turn to the estimation of the effect of institutions on income per capita by Two-Stage Least Squares (2SLS).

   (a) Estimate equation (1) by 2SLS using the `ivregress` command. Make use of the `2sls` estimator and supply the `first` option to let `Stata` show the first-stage results in the output window. Again, perform the estimation with and without heteroskedasticity-robust standard errors and store the estimates using the `estimates store` command.

   (b) What can we learn from the first-stage regressions?

   (c) Compare the IV (or 2SLS) estimate of α to the one obtained by OLS. Is it in line with your expectations? If not, can you think of reasons why this might be the case?

   (d) Compare the IV (or 2SLS) estimate of the coefficient on the variable `latitude` to the one obtained by OLS. What does the result suggest?

4. The assumption that the exclusion restriction holds is crucial to be able to interpret the estimate of the effect of institutions on income per capita as causal. Can you think of other channels through which expected mortality of European settlers at colonization could affect current income per capita?

   (a) Construct a global macro containing variables reflecting the current disease environment using the `global` command. In particular, include `malfal94` (share of population living in area where malaria is endemic), `imr95` (infant mortality rate), and `leb95` (life expectancy at birth).

   (b) Estimate equation (1) again by 2SLS using the `ivregress` command. Perform the regressions including the additional controls and with heteroskedasticity-robust standard errors. You can leave out the continent dummies. Does the estimate of α change?

5. In the previous exercise, we treated the variable `risk` as endogenous. In this case, the OLS estimator is generally biased and inconsistent. As an alternative, we resorted to the IV (or 2SLS) estimator, which is known to be consistent in this case. If, however, the variable `risk` is in fact *exogenous*, using IV (or 2SLS) is inefficient compared to OLS. Hence, it is of interest to *test* whether the variable `risk` is truly endogenous.

   (a) Use the `estimates restore` command to reload the stored initial 2SLS regression with *classical* standard errors. Afterwards, use the `estat endogenous` post-estimation command to perform a Durbin and Wu-Hausman test for endogeneity.[2] What are the null

[2] The difference between the Durbin and Wu-Hausman test is that the former uses an estimate of the error variance based on the model assuming all variables are *exogenous*. The Wu-Hausman test, on the other hand, uses an estimate of the error variance based on the model assuming the tested variable is *endogenous*. Often, this distinction is ignored and a single test is referred to as the Durbin-Wu-Hausman test.

hypotheses H_0 of these tests? Why does this imply that we cannot use the stored regression with heteroskedasticity-robust standard errors?

(b) What do you conclude based on the results of the tests?

6. The theory proposed by Acemoglu et al. (2001) underlying their choice of instrument builds on the following chain of relatedness: Mortality faced by Europeans at colonization (`setmort`) affected European settlements (`euro1900`). In turn, these influenced the institutions established by the European colonizers (`cons00a`) and, consequently, the institutions present in the ex-colonies today (`risk`). To further investigate the validity of the authors' approach, it is possible to estimate the model using `euro1900` and `cons00a` as additional instruments and test for overidentifying restrictions.

(a) Estimate equation (1) using the `ivregress` command with *classical* standard errors. First, include `euro1900` (measure of European settlements in 1900) as an additional instrument for `risk`. Second, redo the exercise with `cons00a` (measure of institutional quality in 1900) as an additional instrument.[3]

(b) After each regression, use the `estat overid` command to test for overidentifying restrictions. What can we learn from this test? What can you conclude from its results?

[3] You can leave out the continent dummies. However, it may also be interesting to redo this exercise including the continent dummies.

2 Theoretical summary

In this section, we briefly summarize the main theoretical concepts regarding Instrumental Variables (IV) estimation in the context of the linear regression model and the potential outcomes framework.

2.1 Instruments in the model-based approach

To present estimation using instrumental variables, we start by maintaining the linearity assumption. That is, the relationship between the dependent variable and the independent variables is linear:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2}$$

However, we will now assume that the exogeneity condition is violated. In other words, at least one independent variable is *not* orthogonal to the error term. In this case, the OLS estimator of the parameter vector will be biased and generally inconsistent.

To introduce how we can exploit instrumental variables in order to obtain consistent estimates of the parameters, we will start by looking at the case of one endogenous regressor (x_i) and a single instrument (z_i). Under what conditions can this be achieved? If we have one endogenous regressor and one instrument, these conditions are well-known. First, the instrument must *relevant*. That is, $\text{Cov}(z_i, x_i) \neq 0$. Second, the instrument must satisfy an *exclusion restriction*. That is, the instrument is orthogonal to the error term $\text{Cov}(z_i, \varepsilon_i) = 0$.

In general, it might be the case that we have more than one endogenous regressor and multiple instruments. Let $\mathbf{X}$ be the $n \times K$ matrix of possibly endogenous regressors and let $\mathbf{Z}$ be the $n \times L$ matrix of instruments. The *relevance* condition and *orthogonality* (or *exclusion*) condition translate in this general setting to the following three conditions:

1. **Orthogonality condition:** The instruments $\mathbf{z}_i$ are predetermined in the sense that they are orthogonal to the contemporaneous error term ε_i. This provides us with the following L orthogonality conditions:[4]

$$E[\mathbf{z}_i \varepsilon_i] = \mathbf{0}. \tag{3}$$

[4] At this point, it is worth noting that the regressors $\mathbf{x}_i$ and the instruments $\mathbf{z}_i$ may overlap. All predetermined regressors are included as instruments. Hence, if no regressors are endogenous, the set of instruments will be identical to the regressors. Note that this further implies that if the regression contains a constant term, it will automatically appear both as a regressor and as an instrument.

2. **Rank condition:** The $L \times K$ matrix $E[\mathbf{z}_i \mathbf{x}_i']$ has full column rank:

$$\text{rank}(E[\mathbf{z}_i \mathbf{x}_i']) = K. \tag{4}$$

This is nothing more than a rank condition for identification. It is both necessary and sufficient. If satisfied, the parameter vector $\boldsymbol{\beta}$ is uniquely identifiable. Note that this condition generalizes the *relevance* condition. Condition (4) can only be satisfied if the determinant of said matrix is not zero. For this to hold, it is relatively easy to show that for each endogenous regressor there must be at least one instrument for which the covariance between regressor and instrument is not zero.

3. **Order condition:** The number of instruments is at least as large as the number of regressors: $L \geq K$. Equivalently, the number of instruments excluded from (2) must be at least as large as the number of endogenous regressors. Note that the order condition is a necessary condition for identification. If it is not met, the rank condition is automatically not satisfied.

Suppose that the number of instruments L is **exactly equal** to the number of regressors K. Under the preceding assumptions, we can exploit the sample analogue to the orthogonality conditions to obtain the Instrumental Variables (IV) estimator:

$$\boldsymbol{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{5}$$

It can be shown that this estimator consistently estimates the true parameter vector $\boldsymbol{\beta}_0$.[5] Now, suppose that the number of instruments at hand L **exceeds** the number of regressors K. In this case, the model is **over-identified**. We could in principle use any subset of instruments of dimension K and estimate the parameters using (5), which would imply a possibly large number of different estimates. This is neither practical nor efficient as we would not be using the information carried by the orthogonality conditions of the remaining $L - K$ instruments. An alternative estimator that uses all instruments at once is known as the Two-Stage Least Squares (2SLS) estimator:

$$\boldsymbol{\beta}_{2SLS} = (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \tag{6}$$

$$= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y} \tag{7}$$

where $\mathbf{P}_\mathbf{Z}$ is the projection matrix corresponding to the orthogonal projection onto the column space of $\mathbf{Z}$. It is symmetric and idempotent. Algebraically, the 2SLS estimator can be envisaged as proceeding in two stages. First, we project the columns of the data matrix $\mathbf{X}$ onto the column space of $\mathbf{Z}$. The resulting matrix $\widehat{\mathbf{X}} = \mathbf{P}_\mathbf{Z}\mathbf{X}$ contains the components of $\mathbf{X}$ that can be *explained* by the instruments. In a second stage, we regress the dependent variable $\mathbf{y}$ onto the columns of $\widehat{\mathbf{X}}$. Note that the standard errors obtained in the second stage are incorrect as they are based on the *wrong* residual. Hence, one should correct the obtained standard errors (`Stata` does this automatically).

[5] Note that if all regressors are exogenous, then $Z = X$ and the IV estimator boils down to the OLS estimator.

It can be shown that the IV estimator is a 2SLS estimator in the **just-identified** case (i.e., $L = K$). More generally, the 2SLS estimator can most easily be understood in the context of the Generalized Method of Moments (GMM). In particular, under the assumption that the error term is conditionally homoskedastic, it can be shown that the optimal or efficient GMM estimator coincides with the 2SLS estimator. However, we will not discuss this further. For more details, see for example Hayashi (2000).

2.2 Instruments in the design-based approach

The traditional instrumental variables approach does not translate well into the standard potential outcomes framework. However, in the 1990's, a lot of work has pushed towards making this happen (Imbens & Angrist, 1994; Angrist, Imbens, & Rubin, 1996).

For ease of exposition, we will assume in what follows that we have a binary treatment D_i and binary instrument Z_i. Now, let $Y_i\big(D_i(Z_i), Z_i\big)$ denote the potential outcomes. The traditional instrument validity conditions can now be expressed as follows:

1. **Exclusion restriction:** The instrument Z_i only affects the outcome Y_i through its effect on the treatment D_i:

$$Y_i\big(D_i(Z_i), Z_i\big) = Y_i\big(D_i(Z_i)\big). \tag{8}$$

   Note that the treatment effect can thus be different for each individual i, whereas in the model-based approach we assumed constant effects ($\boldsymbol{\beta}$ did not vary across i).

2. **Relevance condition:**

$$P(z) = E[D_i|Z_i = z] \tag{9}$$

   varies across values z.

We will maintain the assumption throughout the rest of this section that the instrument Z_i is completely randomly assigned relative to both D_i and Y_i. Nevertheless, you should be wary of the fact that random assignment of Z_i does *not* guarantee that the exclusion restriction holds!

A key point made by Imbens and Angrist (1994) is that in this setup it may not be possible to identify the Average Treatment Effect (ATE). The intuition is that we are only identifying the treatment effect for those individuals whose behavior (or treatment uptake) changed due a change in the instrument Z_i. Furthermore, this effect can be zero or negative even if the true causal effect is positive. It is important to note that in a constant effects world or in the case of one-sided compliance,[6]

[6] One-sided compliance is the case where $\Pr[D_i(1) - D_i(0) = -1] = 0$.

this problem does *not* exist.

However, as noted by Imbens and Angrist (1994), if one is willing to assume *monotonicity*, then one can identify the Local Average Treatment Effect (LATE). That is, assume for all observations $i = 1, \ldots, n$

$$D_i(1) \geq D_i(0). \tag{10}$$

Then the Wald ratio estimates the LATE:

$$\tau_{LATE} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \tag{11}$$

$$= E[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1]. \tag{12}$$

Two remarks are in order. First, note that the monotonicity assumption is also fundamentally untestable. Second, note that our IV *only* identifies the effect of a treatment for those individuals that responded to the instrument.

Finally, while we considered a very easy setup with a binary treatment and instrument, it is relatively straightforward to generalize to settings with multi-valued instruments. It is also possible to generalize to settings with a multi-valued treatment, but this is more challenging. Interested students can refer to Angrist and Imbens (1995).

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, *91*(5), 1369-1401.

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

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

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467-475.