Introductory Econometrics

Based on the textbook by WOOLDRIDGE: *Introductory Econometrics: A Modern Approach*

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

January 23, 2013

Outline

- Heteroskedasticity
- Regressions with time-series observations
- Asymptotics of OLS in time-series regression
- Serial correlation in time-series regression
- Instrumantal variables estimation
  - Basic issues of IV estimation
  - Two-stage least squares
  - Hypothesis tests for IV estimation

The problem of endogeneity

In some regression problems, there are good reasons to surmise that some explanatory variables X and the errors u are correlated. This situation implies biased and inconsistent OLS estimation.

$$\mathrm{E}\hat{\beta} = \beta + \mathrm{E}(X'X)^{-1}(X'u) \neq \beta$$
$$\hat{\beta} = \beta + (X'X)^{-1}(X'u) \not\to \beta$$

Examples for potential endogeneity

- **Feedback**: The explanatory variable x may, by construction or by its concept, depend on the dependent variable y. y depends on u, thus there will be endogeneity. [Aggregate consumption depends on income depends on consumption]
- **Individual characteristics**: Unobserved features may be correlated with x and also affect the reaction of y to x. [Smart persons not only have more years of education but their education also pays more wage]
- **Errors in variables**: x as well as y may be observed with errors, and these observation errors may be correlated.

Endogeneity is a conceptual problem

OLS automatically considers a regression $y = X\beta + u$ with u and X uncorrelated. It consistently estimates this β (the slope of a fitted line or hyperplane, the correlation), but this β does not represent the theoretical underlying relation.

This β does not represent the marginal response of households to changed income or of individual wages to more education.

Because endogeneity is a conceptual problem, there are no direct statistical tests.

Instrumental variables: the idea

Suppose we have a variable z such that **instrument relevance**

$$\text{corr}(x, z) \neq 0$$

and **instrument exogeneity**

$$\text{corr}(z, u) = 0$$

hold, such that z approximates x but not the critical features. Then,

$$\hat{\beta}_{IV} = (Z'X)^{-1}(Z'y) = \beta + (Z'X)^{-1}(Z'u) \to \beta$$

is a consistent estimator for β. It is called the **instrumental variables estimator** or **IV estimator**, and z is called an **instrument**.

Instrumental variables: the implementation

- Separate instruments must be found for each regressor variable x_i. More instruments than regressors are possible: **generalized IV** estimator;
- Of course, relevance (non-zero correlation with the corresponding x_i) and exogeneity (zero correlation with errors) must hold for all $i = 1, \ldots, k$;
- There is no automatic rule for finding instruments. This search is based on conceptual issues and can be difficult;
- If a variable does not cause problems and is exogenous, it can be its own instrument.

The variance of the IV estimator

Assume all Gauss-Markov assumptions **MLR.1–MLR.5** hold except **MLR.4** (otherwise no IV is needed). A variant of **MLR.4** must hold for the instruments:

$$\mathrm{E}(u|z_1, \ldots, z_k) = 0.$$

Then, the variance of the IV estimator is

$$\mathrm{var}(\hat{\beta}_{IV}) = \sigma_u^2 (X'Z(Z'Z)^{-1}Z'X)^{-1},$$

and σ_u^2 can be estimated by

$$\hat{\sigma}_u^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2.$$

This variance can become large when relevance is low ('weak instruments').

R^2 in IV regression

By construction, OLS regression involves minimizing squared errors or, in other words, maximizing R^2. R^2 will always be lower for other estimators, such as IV. R^2 cannot be used as a guideline whether to use IV rather than OLS.

Remember the source of the endogeneity bias problem: OLS estimates a different model. However, OLS estimates that unwanted model efficiently.

Two-stage least squares: the idea

The efficiency of an IV estimator is not guaranteed. When there are many valid instruments (relevance and exogeneity hold for all of them), there are many IV estimates for the same problem.

One may consider constructing a quite efficient IV estimator as follows:

1. Perform k OLS regressions of each x_j on all available instruments $z_1, \ldots, z_m$ with $m > k$ and keep the systematic values $\hat{x}_j$;
2. Perform an IV regression of y on $x_1, \ldots, x_k$ with instruments $\hat{x}_j$ used for x_j.

This estimator is called the **two-stage least squares** estimator (TSLS, 2SLS).

2SLS calculated literally in two stages

- 2SLS estimates can be calculated if only an OLS program is available. First run k first-stage regressions, then regress y on $\hat{x}_1, \ldots, \hat{x}_k$. It can be shown that this yields the same values for $\hat{\beta}_{IV}$;
- This needs to be done for the endogenous or 'suspicious' variables only. The exogenous regressors are their own instruments;
- This algorithm, however, delivers invalid standard errors, as the program does not know that indeed the second stage is not simply OLS.

Remarks on 2SLS

- 2SLS essentially purges the 'bad' variables from their suspicious parts and keeps their exogenous parts;
- Historically, 2SLS has been popular in estimating single equations within large macroeconometric models: endogenous regressors of an equation are instrumented by all exogenous variables in the whole model;
- Currently, 2SLS is even used increasingly in microeconometric problems.

The order condition for 2SLS

Regressors can be classified into two sets:

- The good exogenous regressors;
- The bad endogenous regressors.

The **order condition** says that there must be at least as many additional exogenous instruments—i.e. exogenous variables that are not in the first set—as endogenous regressors. This order condition (counting rule) is only necessary, not sufficient.

The necessary and sufficient **rank criterion** that guarantees identifiability is quite complex.

Hypothesis tests with IV estimation

Two types of test help in assessing whether the instruments are correctly chosen and whether endogeneity is a problem:

1. Hausman-type tests on endogeneity
2. Overidentification tests

Remember that endogeneity is basically an issue beyond statistical testing, thus the verdict of these tests may be of limited value.

Hausman-type endogeneity tests: the idea

The basic idea of a **Hausman test** is to compare an estimator that is efficient under restrictive assumptions (and otherwise inconsistent) and an inefficient estimator that is consistent even when the assumptions are violated.

OLS is efficient if there is no endogeneity problem. 2SLS is consistent even when there is endogeneity.

Hausman-type endogeneity tests: the algorithm

Consider a suspicious x_j among the regressors $1 \leq j \leq k$ that may be endogenous:

1. Determine the best exogenous approximant $\hat{x}_j$ by regressing x_j on all exogenous variables, keeping the residuals v_j;
2. Include v_j as an additional regressor in the equation

$$y = X\beta + \gamma v_j + \text{error},$$

which is estimated by OLS. If there is endogeneity, $\gamma \neq 0$, which is tested by the corresponding t–value.

Overidentification tests

If the number of instruments m exactly matches the number of endogenous regressors g, there is nothing to test. If $m > g$, 2SLS implicitly uses restrictions (moment conditions) that can be tested statistically. The following algorithm for an LM–type test can be used:

1. Estimate the main regression by 2SLS and keep the residuals $\hat{u}_i$;
2. Regress $\hat{u}$ on all exogenous variables and note the R^2 of this auxiliary regression;
3. Under the null of instrument validity for all instruments, nR^2 is distributed χ^2_{m-g}.