# Problem Set 6: IV/2SLS

## Review the Concepts and Proofs

1. Suggest a method to check the relevance of a set of instruments. Can you find a test for the exogeneity assumption?
2. Give an intuitive explanation of the order criterion of 2SLS.
3. Derive the 2SLS estimator.
4. Show that in the case of exact identification IV and 2SLS are the same.
5. Explain why the Sargan test of overidentifying orthogonality restrictions lead to problematic test results. Hint: Which error is controlled for?

## Exercises

1. Consider a linear model with mean adjusted data:

$$\begin{aligned} y_i &= \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i, \quad i = 1, \ldots, N, \\ \text{for short:} \quad \mathbf{y} &= X\boldsymbol{\beta} + \boldsymbol{e}. \end{aligned} \tag{1}$$

The following sample moments are known:

$$(X'X) = \begin{pmatrix} 20 & 10 \\ 10 & 20 \end{pmatrix}, \ (X'\mathbf{y}) = \begin{pmatrix} 12.5 \\ 10.5 \end{pmatrix},$$

Assume that $x_{1,i}$ is endogenous and that $z_i$ is a valid instrument with following sample moments:

$$\sum_{t=1}^{N} z_i x_{1,i} = 10, \quad \sum_{i=1}^{N} z_i x_{2,i} = 15 \quad \text{and} \quad \sum_{i=1}^{N} z_i y_i = 8.5.$$

(a) Explain the principle of instrument variable (IV) estimation. In which sense can the OLS estimator be regarded as a special case of IV estimation?

(b) Estimate the model by means of the OLS and the IV method.

2. Consider a model for the health of an individual:

$$health = \beta_0 + \beta_1 age + \beta_2 weight + \beta_3 height + \beta_4 male + \beta_5 work + \beta_6 exercise + u,$$

where *health* is some quantitative measure of the person's health, *work* is weekly hours worked, *exercise* is the hours of exercise per week, and *age*, *weight*, *height*, and *male* are self-explanatory.

(a) Why might you be concerned about *exercise* being correlated with the error term $u$?

(b) Suppose you can collect data on two additional variables, *disthome* and *distwork*, the distances from home and from work to the nearest health club or gym. Discuss whether these are likely to be uncorrelated with $u$.

(c) Now assume that *disthome* and *distwork* are uncorrelated with $u$, as are all variables in equation (2) with the exception of *exercise*. Write down the reduced form for *exercise*, and state the conditions under which the parameters of equation (2) are identified.

(d) Derive the conditions for identification above. How can they be tested?

3. Consider the structural model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + v$$

where $v$ has zero mean and is uncorrelated with $x_1, \ldots, x_{K-1}$ but correlated with $x_K$. In addition, $x_K$ is thought to be correlated with at least one of the other regressors. Furthermore, you are given a set of variables $z_1, z_2, \ldots, z_M$ that are redundant in the structural equation above.

(a) Suggest an IV method for consistently estimating the parameters $\beta_0, \ldots, \beta_K$. Under which conditions is the model identified? How many overidentifying restrictions do you have? Suggest a test procedure to test for overidentification.

(b) If the equation is a log(*wage*) equation, $x_K$ is education and $z_1, \ldots, z_M$ are family background variables, such as parents' education and family income, describe the economic assumptions needed for consistency of the the IV procedure in part (a).

4. Consider IV estimation of the simple linear model with a single, possibly endogenous, explanatory variable, and a single instrument:

$$y = \beta_0 + \beta_1 x_1 + u,$$

$$\mathrm{E}(u) = 0, \qquad \mathrm{Cov}(z_1, u) = 0, \qquad \mathrm{Cov}(z_1, x_1) \neq 0, \qquad \mathrm{E}(u^2|z_1) = \sigma^2.$$

   (a) Under the preceding (standard) assumptions, show that Avar $\sqrt{N}(\hat{\beta}_1 - \beta_1)$ can be expressed as $\sigma^2/(\rho^2_{z_1x_1}\sigma^2_{x_1})$, where $\sigma^2_{x_1} = \mathrm{Var}(x_1)$ and $\rho_{z_1x_1} = \mathrm{Corr}(z_1, x_1)$. Compare this result with the asymptotic variance of the OLS estimator under Assumptions OLS.1-OLS.3.

   (b) Comment on how each factor affects the asymptotic variance of the IV estimator. What happens as $\rho_{z_1x_1} \to 0$?

5. Consider the model $y = \boldsymbol{x\beta} + u$, where $x_1$ and $x_2$ are the (potentially) endogenous explanatory variables and $x_3$ and $x_4$ are the exogenous variables. (We assume a zero intercept just to simplify the notation; the following results carry over to models with an unknown intercept.) Let $z_1$ and $z_2$ be the instrumental variables available from outside the model. Let $\boldsymbol{z} = (z_1, z_2, x_3, x_4)$ and assume that $\mathrm{E}(\boldsymbol{z}'\boldsymbol{z})$ is nonsingular, so that instruments are not perfectly correlated with each other.

   (a) Show that a necessary condition for the rank condition (which is the relevance condition) is that for each endogenous regressor ($x_1$ and $x_2$) at least one instrument must appear in the reduced form.

   (b) Give a simple example showing that the condition from part a is not sufficient for the rank condition.

   (c) Show that a sufficient condition for the rank condition is that only one instrument appears in the reduced form of one endogenous regressor, respectively.

6. A researcher is interested in the dynamic relationship between an exogenous variable, $x$, and a dependent variable, $y$. She specifies a partial adjustment model which posits that $y_{i,t}$ is not only a function of $x_{i,t}$ but also of the lagged dependent variable, $y_{i,t-1}$. In addition, she suspects that a fraction $\gamma$ of previous period's disturbance, $e_{i,t-1}$, affects current period's dependent variable. Altogether, she obtains the model

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \beta_2 y_{i,t-1} + e_{i,t} + \gamma e_{i,t-1}, \quad t = 1, 2, \ldots, \; i = 1, \ldots, N.$$

For a random sample of individuals, she collects the two observations of periods $T$ and $T-1$, i.e., $(x_{i,T}, y_{i,T})$ and $(x_{i,T-1}, y_{i,T-1})$. She assumes that (1) $x$ is strictly exogenous with respect to $e$, i.e., $E(e_{i,t}|x_{i,T-1}, x_{i,T}) = 0$ for $t = T-1, T$, and (2) $e_{i,t} \overset{iid}{\sim} (0, \sigma^2)$ for $t = T-1, T$ and $i = 1, \ldots, N$.

(a) Show that the assumptions imply both $E(e_{i,T}e_{i,T-1}) = 0$ and $E(e_{i,T}y_{i,T-1}) = 0$, $i = 1, \ldots, N$.

(b) Show that defining the regression error $u_{i,T} = e_{i,T} + \gamma e_{i,T-1}$ and estimating the equation

$$y_{i,T} = \beta_0 + \beta_1 x_{i,T} + \beta_2 y_{i,T-1} + u_{i,T}, \quad i = 1, \ldots, N,$$

by OLS is inconsistent. Which regressor causes the problem?

(c) Propose an appropriate instrumental variable. Does it satisfy the conditions for consistent IV estimation?