Instrumental Variable Regression

Problem and solution

- The standard equation for the OLS regression coefficient (without the intercept) in terms of variances and covariances is (see Hamilton p 294) is:

$$y = \beta \cdot x + \varepsilon$$

$$Cov(y, x) = \beta \cdot \underbrace{Cov(x, x)}_{Var(x)} + \underbrace{Cov(\varepsilon, x)}_{=0 \text{ by assumption}}$$

  Thus the regression coefficient under the assumption of $Cov(x, \varepsilon) = 0$ becomes

$$\beta_{OLS} = Cov(y, x)/Var(x).$$

- If the independence assumption between x and the disturbances ε breaks down the variable x becomes an ***endogenous regressor***, due to its relationship with the random disturbances.
- The OLS estimate for β becomes biased:

$$\beta_{OLS}^{biased} = \beta + \underbrace{Cov(\varepsilon, x)/Var(x)}_{bias}.$$

- However, if another variable z, which is called an ***instrumental*** variable, that is independent of ε can be found then we still can estimate β as long as $Cov(x, z) \neq 0$:

$$Cov(y, z) = \beta \cdot Cov(x, z) + \underbrace{Cov(\varepsilon, z)}_{=0 \text{ by assumption}}$$

  Preferably, the relationship between x and the instrumental variable z is strong.

- Thus the instrumental estimator becomes

$$\boxed{\beta_{IV} = Cov(y,z)/Cov(x,z)}$$

The regression parameter β_{IV} still measures the influence of x on y and not z on y.

- See the R script **`Wooldridge01.r`** and the Ballentine figure in Kennedy p 147.

Underlying idea of two-stage least squares

- In instrumental variable estimation we need to distinguish between 3 groups of regressors:
    1. The set endogenous regressors $\mathbf{X}_{EN}$, which cause problems because they are influenced (i.e., correlated) with the disturbances
    2. The set exogenous regressors $\mathbf{X}_{EX}$, which are regular regressors that are uncorrelated with the disturbances
    3. The set of instrumental variables $\mathbf{X}_{IV}$, which are correlated with $\mathbf{X}_{EN}$ but uncorrelated with the disturbances.
- These three groups can be pooled together into a matrix purely exogenous variables and our original independent variables:
    1. $\mathbf{Z} = [\mathbf{X}_{IV}|\mathbf{X}_{EX}]$
    2. $\mathbf{X} = [\mathbf{X}_{EN}|\mathbf{X}_{EX}]$
- Notes:
    1. The exogenous regressor $\mathbf{X}_{EX}$ functions as its own instrumental variable, i.e., it leads to a one-to-one prediction.
    2. There needs to be at least as many instrumental variables as there are endogenous variables in order to make the regression system identifiable.
- An unbiased estimator for $\boldsymbol{\beta}$ is achieved with the help of instrumental variables in a two-stage estimation procedure:

1. At the 1st stage a set of instrumental variables $\mathbf{Z}$, which are assumed to be independent of the error terms $\boldsymbol{\varepsilon}$ but correlated with the endogenous regressor, are used to model with linear regression the endogenous regressors:

   $\widehat{\mathbf{X}} = \mathbf{Z} \cdot \boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma} = (\mathbf{Z}^T \cdot \mathbf{Z})^{-\mathbf{1}} \cdot \mathbf{Z}^T \cdot \mathbf{X}$ being a set of simple OLS estimators for each variable in $\mathbf{X}$. This is also called the ***reduced form***.

2. At the 2nd stage the predicted endogenous regressors $\widehat{\mathbf{X}}$ are used to model the dependent variable by $\mathbf{y} = \widehat{\mathbf{X}} \cdot \widehat{\boldsymbol{\beta}}_{IV} + \boldsymbol{\varepsilon}$. This is also called the ***structural form***.

   This is possible because as long as $Cov(\mathbf{Z}, \boldsymbol{\varepsilon}) = \mathbf{0}$ so is $Cov\left(\widehat{\mathbf{X}}, \boldsymbol{\varepsilon}\right) = \mathbf{0}$.

- The separate two-stage estimation approach, however, leads to ***biased*** standard error of $\boldsymbol{\beta}_{IV}$. This problem can be overcome by pooling both stages together:
  - In compact notation the estimator becomes:

    $$\widehat{\boldsymbol{\beta}}_{IV} = (\mathbf{X}^T \cdot \mathbf{H} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{H} \cdot \mathbf{y}$$

    where $\mathbf{H}$ the is the hat matrix $\mathbf{H} = \mathbf{Z} \cdot (\mathbf{Z}^T \cdot \mathbf{Z})^{-1} \cdot \mathbf{Z}$. Thus $\widehat{\mathbf{X}} = \mathbf{H} \cdot \mathbf{X}$. The hat matrix $\mathbf{H}$ is idempotent with $\mathbf{H} = \mathbf{H} \cdot \mathbf{H}$.

  - The proper covariance matrix of $\widehat{\boldsymbol{\beta}}_{IV}$ is

    $$Cov\left(\widehat{\boldsymbol{\beta}}_{IV}\right) = \hat{\sigma}^2 \cdot (\mathbf{X}^T \cdot \mathbf{H} \cdot \mathbf{X})^{-1}$$

Test $\mathbf{X}_{IV}$ for Instrument Relevance

- At the first stage compare the models $\widehat{\mathbf{X}}_{EN} = [\mathbf{X}_{IV}|\mathbf{X}_{EX}] \cdot [\boldsymbol{\beta}_{IV}^T|\boldsymbol{\beta}_{EX}^T]^T$ against the restricted model $\widehat{\mathbf{X}}_{EN} = \mathbf{X}_{EX} \cdot \boldsymbol{\beta}_{EX}$ with the partial *F*-test.
- Does the model with the additional instrumental variables $\mathbf{X}_{IV}$ improve substantially the model fit of $\widehat{\mathbf{X}}_{EN}$?

- The hypotheses are
  - $H_0: \boldsymbol{\beta}_{IV} = \mathbf{0}$ against
  - $H_1: \boldsymbol{\beta}_{IV} \neq \mathbf{0}$.

  A rejection of H_0 indicates that $\mathbf{X}_{IV}$ are ***strong*** instruments.

Test $\mathbf{X}_{EN}$ for exogeneity (modified Hausman test)

- The residuals $\mathbf{E} = \mathbf{X}_{EN} - \widehat{\mathbf{X}}_{EN}$ at the first stage, i.e., $\widehat{\mathbf{X}} = \mathbf{Z} \cdot \boldsymbol{\Gamma}$, are no longer be correlated $\widehat{\mathbf{X}}_{EN}$ and comprise of the ***unique*** variation of $\mathbf{X}_{EN}$ and perhaps the variation that $\mathbf{X}_{EN}$ shares with $\boldsymbol{\varepsilon}$.
- Thus an augmented OLS regression $\hat{\mathbf{y}} = \mathbf{X} \cdot \widehat{\boldsymbol{\beta}} + \mathbf{E} \cdot \widehat{\boldsymbol{\beta}}_E$ should give
  - $H_0: \widehat{\boldsymbol{\beta}}_E = \mathbf{0}$ if $\mathbf{X}_{EN}$ is uncorrelated with the disturbances $\boldsymbol{\varepsilon}$. Thus IV estimation is not necessary.
  - $H_1: \widehat{\boldsymbol{\beta}}_E \neq \mathbf{0}$ if $\mathbf{X}_{EN}$ is correlated with the disturbances $\boldsymbol{\varepsilon}$ and IV regress should be performed.
- Note that $\widehat{\boldsymbol{\beta}}$ in the augmented model is equal to instrumental variable estimator $\widehat{\boldsymbol{\beta}}_{IV}$ because the residuals $\mathbf{E}$ in the augmented model control for the potential endogeneity.

Sargan test for instrument $\mathbf{X}_{IV}$ validity

- The regression residuals of the instrumental variable model $\mathbf{e}_{IV} = \mathbf{y} - \left(\mathbf{X} \cdot \widehat{\boldsymbol{\beta}}_{IV}\right)$ should be uncorrelated with exogenous regressors $\mathbf{X}_{EX}$ and the instrumental variables $\mathbf{X}_{IV}$.
- Therefore, a regression of $\mathbf{e}_{IV}$ on $\mathbf{Z}$ should result in an $R^2 = 0$. The Sargan statistic $n \cdot R^2 \sim \chi^2(df)$ with $df = \#\text{ of instruments} - \#\text{ of endogenous regressors}$.
- The hypotheses are
  - $H_0: n \cdot R^2 = 0$ and all instruments $\mathbf{Z} = [\mathbf{X}_{IV} | \mathbf{X}_{EX}]$ exogenous.
  - $H_0: n \cdot R^2 \neq 0$ at least one instrument in $\mathbf{Z}$ endogenous. Therefore, the IV estimates still will be biased.
- There are several additional assumptions that the Sargan test makes. These are discussed in Kennedy.

- Note: as the sample size n increases the Sargan test will become more and more significant. This problem is highlighted in the script **`ChiSquareSampleSizeEffect.r`** .

Literature:

Woolridge, J. M. **2009**. *Introductory Econometrics. A Modern Approach.* Cengage. Chapter 15: "*Instrumental Variables Estimation and Two-Stage Least Squares"*

Kennedy, P. **2008**. *A Guide to Econometrics*. Wiley-Blackwell. Chapter 9: *"Violating Assumption Four: Instrumental Variable Estimation"*

Instrumental variables software

- The function **`ivreg( )`** in the package **`AER`**.
- The **`simex`** and **`mcsimex`** packages, which aim at modelling the biases in the regression coefficients with a parametric function.
- The package **`ivmodel`**.