OLS and Haavelmo bias

- Generally in each equation of the structural form we have several endogenous variables
- OLS applied to separate equations will give inconsistent estimates (simultaneity bias)
- The bias is the result of the correlation between explanatory variables and dependent variable.

Example 1. *(Keynes consumption function and GDP identity)*

$$C = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 Y + \boldsymbol{\varepsilon}$$

$$Y = C + I + G$$

Assume that investments and government expenditures are exogenous. C *is correlated with* ε *and* Y *is a function of* $C \Longrightarrow Y$ *correlated with* ε

Indirect OLS

- Estimation of the reduced form
  - on the right hand side we only have the exogenous variables
  - on the left hand side endogenous variables
  - assume the homoscedasticity and no autocorrelation

- For each equation of the reduced form we can use the OLS and obtain unbiased estimates of multipliers

- Is it possible to obtain the estimates of structural parameters on the basis of estimates of multipliers?

- Elements of the matrix $\mathbf{\Pi}$ are functions of elements of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ as $\mathbf{\Pi} = \boldsymbol{A}^{-1}\boldsymbol{B}$

- When it is possible to derive $\boldsymbol{A}$ and $\boldsymbol{B}$ from the system of equations $\widehat{\boldsymbol{\Pi}} = \widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}$?

- If structural model is:
  - not identified: no unique solution
  - exactly identified: one solution
  - overidentified: several contradictory solutions

Example 2. *(demand and supply - original overidentified model)*

$$\widehat{\boldsymbol{\beta}}_1 = \frac{\widehat{\boldsymbol{\pi}}_{11}}{\widehat{\boldsymbol{\pi}}_{21}},\ \widehat{\boldsymbol{\beta}}_1 = \frac{\widehat{\boldsymbol{\pi}}_{12}}{\widehat{\boldsymbol{\pi}}_{22}}, \widehat{\boldsymbol{\beta}}_2 = -\widehat{\boldsymbol{\pi}}_{23}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\alpha}}_1\right)$$

$$\widehat{\boldsymbol{\alpha}}_1 = \frac{\widehat{\boldsymbol{\pi}}_{13}}{\widehat{\boldsymbol{\pi}}_{23}},\ \widehat{\boldsymbol{\alpha}}_2 = \widehat{\boldsymbol{\pi}}_{22}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\alpha}}_1\right),\ \boldsymbol{\alpha}_0 = \widehat{\boldsymbol{\pi}}_{21}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\alpha}}_1\right)$$

Two alternative estimators for $\boldsymbol{\beta}_1$ parameter!

- Is there something wrong with indirect OLS? If the model is correctly specified the alternative estimators should ultimately converge to the parameter value.

- However: in small samples the estimates obtained from them will generally be different.

- Is it then better to have exactly identified model rather than overidentified one?
  - The overidentifing restrictions represent the theoretical knowledge (a priori knowledge - not coming from data) which should improve the quality of estimates
  - But: indirect OLS is not well suited as an estimation method for an overidentified model

Instrumental variable method - single equation approach

- With instrumental variable method we can estimate the regression equation even if the explanatory variables are correlated with the error term

- Instruments $\boldsymbol{W}_t$
  - are correlated with $\boldsymbol{X}_t$
  - are not correlated with ε_t

- Fitted values from regression of $\boldsymbol{X}_t$ on $\boldsymbol{W}$:

$$\widehat{\boldsymbol{X}} = \boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\boldsymbol{X} = \boldsymbol{P}_W\boldsymbol{X}$$

where $\boldsymbol{P}_W = \boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'$.

Remark 3. *In the context of simultaneous equation model, $\boldsymbol{W}$ is the matrix of exogenous variables, and $\widehat{\boldsymbol{X}}$ is the matrix of fitted values of endogenous (and exogenous) variables obtained from the reduced form.*

- $\boldsymbol{P}_W$ is symmetric and idempotent.

- Regression $\boldsymbol{y}$ on fitted values $\widehat{\boldsymbol{X}}$ gives $2SLS$ estimator:

$$\begin{aligned}\widetilde{\boldsymbol{b}} &= \left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{y} = \left(\boldsymbol{X}'\boldsymbol{P}_W\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_W\boldsymbol{y} \\ &= \left(\boldsymbol{X}'\boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\boldsymbol{y}\end{aligned}$$

Remark 4. *Number of instrumental variables must be larger than the number of explanatory variables. In the context of simultaneous equation models this condition reduces to necessary condition for identification as $K \geq G_j + K_j - 1$ where G_j is the number of endogenous variables in the equation including dependent variable, K_j the number of exogenous variables, and K the total number of available instruments (exogenous variables).*

- Hausman test for endogeneity of explanatory variables: we check whether it is necessary to use the IV method:
  - IV estimators and OLS estimators are consistent if explanatory variables are exogenous
  - only IV is consistent if some explanatory variables are endogenous,
- So: test for endogeneity just by testing the equality of estimators:

$$\sqrt{n}\left(b_{IV} - b_{OLS}\right) \xrightarrow{D} N\left(0, \mathbf{\Sigma}_{b_{IV}-b_{OLS}}\right)$$

- It is also possible to test the *overidentifing* restrictions in the equation and the structural model. The test is called the Sargan test

Two Stage Least Squares ($2SLS$) - multiple equation approach)

- Equations of the structural model cannot be estimated with OLS because of the (endogeneity/simultaneity) problem
- Consistent estimates can be obtained in this case with $2SLS$
- Instrumental variables: all exogenous variables in the model
- Necessary condition for feasibility of the $2SLS$: the number of instruments larger or equal to number explanatory variables
- It is equivalent to the identification condition!

- If the equation of the model is identified it can be estimated with $2SLS$

- Two stages of estimation with $2SLS$

1. Calculate fitted values from regression of endogenous variable on the instruments (exogenous variables) - these fitted values can be obtained from reduced form

2. Calculate $2SLS$ regression for each equation of the structural form (endogenous variables are replaced with the fitted values calculated at the first stage)

- Observations for j equation of the structural form can be written as

$$\boldsymbol{y}_j = \boldsymbol{Z}_j \boldsymbol{\beta}_j + \boldsymbol{u}_j$$

where $\boldsymbol{Z}_j$ is the vector of endo and exogenous explanatory variables in j equation.

- Then all the observations for all the equations in the model can be stacked into

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_G \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \boldsymbol{Z}_2 & & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & & \boldsymbol{Z}_G \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_G \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_G \end{bmatrix}$$

- Which can also be written as standard linear model:

$$\overline{\boldsymbol{y}} = \overline{\boldsymbol{Z}}\,\overline{\boldsymbol{\beta}} + \overline{\boldsymbol{u}}$$

- Fitted values $\widehat{\boldsymbol{Z}}_j$ from the regression of $\boldsymbol{Z}_j$ on the exogenous variables $\boldsymbol{X}$

are equal to

$$\widehat{\boldsymbol{Z}}_j = \boldsymbol{X}\widehat{\boldsymbol{\Pi}}_j = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}_j = \boldsymbol{P}_X\boldsymbol{Z}_j.$$

- And $2SLS$ estimator for all the equations can be written as standard $2SLS$ estimator for stacked model

$$\boldsymbol{b}_{2MNK} = \left(\overline{\widehat{\boldsymbol{Z}}}'\overline{\widehat{\boldsymbol{Z}}}\right)^{-1}\overline{\widehat{\boldsymbol{Z}}}'\overline{\boldsymbol{y}}$$

where

$$\overline{\widehat{\boldsymbol{Z}}} = \begin{bmatrix} \widehat{\boldsymbol{Z}}_1 & \cdots & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \widehat{\boldsymbol{Z}}_G \end{bmatrix}$$

- $2SLS$ estimator is inefficient because they when estimating i equation we are not using the information related to restrictions imposed on other equations

- On the other hand with this method it is possible to estimate one equation without estimating or even specifying the other equations - it suffices to specify which are the endogenous and exogenous variables (single equation approach)

- Estimation methods based on such limited information are called *Limited information* methods

Example 5. *(demand and supply - original overidentified model) -* $2SLS$ *estimation. Demand equation*

$$Q_D = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 P + \boldsymbol{\alpha}_2 Y + u_1$$

is estimated wit IV *methods with instruments* $1, Y,\ P_M$*. Similarly supply* *equation*

$$Q_S = \boldsymbol{\beta}_1 P + \boldsymbol{\beta}_2 P_M + u_2$$

is estimated with the use of instruments $1, Y,\ P_M$*. In second equation we have more instruments than needed (equation is overidentified).*