Quantitative Economics for the Evaluation of the European Policy

Dipartimento di Economia e Management

Co-funded by the Erasmus+ Programme of the European Union

Project funded by

European Commission Erasmus + Programme –Jean Monnet Action

Project number 553280-EPP-1-2015-1-IT-EPPJMO-MODULE

Irene Brunetti   Davide Fiaschi   Angela Parenti[1]

31/10/2016

[1] ireneb@ec.unipi.it, davide.fiaschi@unipi.it, and aparenti@ec.unipi.it.

From Theory to Empirics in Solow model

To derive the specification of a statistical model of cross-country growth differences from which the effects on growth of various factors, including initial conditions, may be identified we need to specify:

From Theory to Empirics in Solow model

To derive the specification of a statistical model of cross-country growth differences from which the effects on growth of various factors, including initial conditions, may be identified we need to specify:

1. a theoretical (neoclassical) framework for understanding growth dynamics;

From Theory to Empirics in Solow model

To derive the specification of a statistical model of cross-country growth differences from which the effects on growth of various factors, including initial conditions, may be identified we need to specify:

1. a theoretical (neoclassical) framework for understanding growth dynamics;
2. the relationship between this theoretical model of growth dynamics and the specification of a growth regression.

Growth dynamics: basic ideas

For economy i at time t, let $Y_{i,t}$ be the output, $L_{i,t}$ the labour force and $A_{i,t}$ the level of technology.

Growth dynamics: basic ideas

For economy i at time t, let $Y_{i,t}$ be the output, $L_{i,t}$ the labour force and $A_{i,t}$ the level of technology.

Assume that $L_{i,t}$ and $A_{i,t}$ grow exogenously at rates n_i and g_{A_i} respectively, that is:

$$\begin{aligned} L_{i,t} &= L_{i,0} e^{n_i t}; \\ A_{i,t} &= A_{i,0} e^{g_{A_i} t}. \end{aligned} \tag{1}$$

Growth dynamics: basic ideas

For economy i at time t, let $Y_{i,t}$ be the output, $L_{i,t}$ the labour force and $A_{i,t}$ the level of technology.

Assume that $L_{i,t}$ and $A_{i,t}$ grow exogenously at rates n_i and g_{A_i} respectively, that is:

$$\begin{aligned} L_{i,t} &= L_{i,0} e^{n_i t}; \\ A_{i,t} &= A_{i,0} e^{g_{A_i} t}. \end{aligned} \tag{1}$$

Let define the variables per efficiency unit of labour input, that is:

$$\begin{aligned} k_i &\equiv \frac{K_i}{A_i L_i}; \\ y_i &\equiv \frac{Y_i}{A_i L_i}; \\ f &\equiv F\left(\frac{K}{AL}, h\right). \end{aligned} \tag{2}$$

Log-linearization around the equilibrium

The generic one-sector growth model implies, to the first-order approximation, that:

$$\log(y_{i,t}) = (1 - e^{-\lambda_i t})\log(y_i^{EQ}) + e^{-\lambda t}\log(y_{i,0}), \tag{3}$$

where y_t is the level of GDP per worker in efficiency units at time t, and the parameter λ_i measures the **rate of convergence** of economy i to its steady state.

Log-linearization around the equilibrium

The generic one-sector growth model implies, to the first-order approximation, that:

$$\log(y_{i,t}) = (1 - e^{-\lambda_i t})\log(y_i^{EQ}) + e^{-\lambda t}\log(y_{i,0}), \tag{3}$$

where y_t is the level of GDP per worker in efficiency units at time t, and the parameter λ_i measures the **rate of convergence** of economy i to its steady state.

A *steady state* of the economy is defined as any level of y_i^{EQ} such that, if the economy starts with $y_{i,0} = y_i^{EQ}$, then $y_{i,t} = y_i^{EQ}$ for all $t \geq 1$.

Log-linearization around the equilibrium

The generic one-sector growth model implies, to the first-order approximation, that:

$$\log(y_{i,t}) = (1 - e^{-\lambda_i t})\log(y_i^{EQ}) + e^{-\lambda t}\log(y_{i,0}), \tag{3}$$

where y_t is the level of GDP per worker in efficiency units at time t, and the parameter λ_i measures the **rate of convergence** of economy i to its steady state.

A *steady state* of the economy is defined as any level of y_i^{EQ} such that, if the economy starts with $y_{i,0} = y_i^{EQ}$, then $y_{i,t} = y_i^{EQ}$ for all $t \geq 1$.

Naturally, we are interested to know whether the economy will converge to the steady state if it starts away from it.

Log-linearization around the equilibrium (cont.)

Eq. (3) is expressed in terms of the **unobservable** outcome per efficiency unit of labour, $y_{i,t}$.

Log-linearization around the equilibrium (cont.)

Eq. (3) is expressed in terms of the **unobservable** outcome per efficiency unit of labour, $y_{i,t}$.

In order to describe the dynamics in terms of the **observable** GDP per worker, $\frac{Y_{i,t}}{L_{i,t}}$, we get:

$$\log\left(\frac{Y_{i,t}}{L_{i,t}}\right) - g_{A_i} t - \log(A_{i,0}) = (1 - e^{-\lambda_i t})\log(y_i^{EQ}) + \\ + e^{-\lambda_i t}\left[\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) - \log(A_{i,0})\right]. \tag{4}$$

Log-linearization around the equilibrium (cont.)

Rewrite Eq. (4) we get:

$$\log\left(\frac{Y_{i,t}}{L_{i,t}}\right) = g_{A_i}t + e^{-\lambda_i t}\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + (1-e^{-\lambda_i t})\log(y_i^{EQ}) + (1-e^{-\lambda_i t})\log(A_{i,0}) \tag{5}$$

Log-linearization around the equilibrium (cont.)

Rewrite Eq. (4) we get:

$$\log\left(\frac{Y_{i,t}}{L_{i,t}}\right) = g_{A_i} t + e^{-\lambda_i t}\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + (1 - e^{-\lambda_i t})\log(y_i^{EQ}) + (1 - e^{-\lambda_i t})\log(A_{i,0}) \tag{5}$$

or, equivalently:

$$\begin{aligned}\log\left(\frac{Y_{i,t}}{L_{i,t}}\right) - \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) &= g_{A_i} t + (1 - e^{-\lambda_i t})\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \\ &+ (1 - e^{-\lambda_i t})\log(y_i^{EQ}) + (1 - e^{-\lambda_i t})\log(A_{i,0})\end{aligned} \tag{6}$$

Log-linearization around the equilibrium (cont.)

Rewrite Eq. (4) we get:

$$\log\left(\frac{Y_{i,t}}{L_{i,t}}\right) = g_{A_i}t + e^{-\lambda_i t}\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + (1-e^{-\lambda_i t})\log(y_i^{EQ}) + (1-e^{-\lambda_i t})\log(A_{i,0}) \tag{5}$$

or, equivalently:

$$\begin{aligned}\log\left(\frac{Y_{i,t}}{L_{i,t}}\right) - \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) &= g_{A_i}t + (1-e^{-\lambda_i t})\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \\ &+ (1-e^{-\lambda_i t})\log(y_i^{EQ}) + (1-e^{-\lambda_i t})\log(A_{i,0})\end{aligned} \tag{6}$$

Define the average growth rate of GDP per worker, $\overline{g_{iY/L}}$, as:

$$\overline{g_{iY/L}} \approx \frac{\log\left(\frac{Y_{i,t}}{L_{i,t}}\right) - \log\left(\frac{Y_{i,0}}{L_{i,0}}\right)}{t} \tag{7}$$

Log-linearization around the equilibrium (cont.)

Moreover, define:

$$\beta_i \equiv -\frac{(1 - e^{-\lambda_i t})}{t} \tag{8}$$

Log-linearization around the equilibrium (cont.)

Moreover, define:

$$\beta_i \equiv -\frac{(1 - e^{-\lambda_i t})}{t} \tag{8}$$

Therefore we ge:

$$\overline{g_{iY/L}} = g_{A_i} t + \beta_i \left(\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \log(y_i^{EQ}) + \log(A_{i,0}) \right) \tag{9}$$

Log-linearization around the equilibrium (cont.)

Moreover, define:

$$\beta_i \equiv -\frac{(1 - e^{-\lambda_i t})}{t} \qquad (8)$$

Therefore we ge:

$$\overline{g_{iY/L}} \quad = \quad g_{A_i} t + \beta_i \left(\log \left(\frac{Y_{i,0}}{L_{i,0}} \right) + \log(y_i^{EQ}) + \log(A_{i,0}) \right) \qquad (9)$$

Eq. (9) decomposes the growth rate in region i into two distinct components:

Log-linearization around the equilibrium (cont.)

Moreover, define:

$$\beta_i \equiv -\frac{(1 - e^{-\lambda_i t})}{t} \tag{8}$$

Therefore we ge:

$$\overline{g_{iY/L}} \quad = \quad g_{A_i} t + \beta_i \left(\log \left(\frac{Y_{i,0}}{L_{i,0}} \right) + \log(y_i^{EQ}) + \log(A_{i,0}) \right) \tag{9}$$

Eq. (9) decomposes the growth rate in region i into two distinct components:

1. g_{A_i}: measuring growth due to technological progress;

Log-linearization around the equilibrium (cont.)

Moreover, define:

$$\beta_i \equiv -\frac{(1 - e^{-\lambda_i t})}{t} \tag{8}$$

Therefore we ge:

$$\overline{g_{iY/L}} = g_{A_i} t + \beta_i \left(\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \log(y_i^{EQ}) + \log(A_{i,0}) \right) \tag{9}$$

Eq. (9) decomposes the growth rate in region i into two distinct components:

1. g_{A_i}: measuring growth due to technological progress;
2. $\beta_i \left(\log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \log(y_i^{EQ}) + \log(A_{i,0}) \right)$: measuring growth due to the gap between initial output per worker and the steady-state value

Log-linearization around the equilibrium (cont.)

Moreover, define:

$$\beta_i \equiv -\frac{(1 - e^{-\lambda_i t})}{t} \tag{8}$$

Therefore we ge:

$$\overline{g_{iY/L}} = g_{A_i} t + \beta_i \left(\log \left(\frac{Y_{i,0}}{L_{i,0}} \right) + \log(y_i^{EQ}) + \log(A_{i,0}) \right) \tag{9}$$

Eq. (9) decomposes the growth rate in region i into two distinct components:

1. g_{A_i}: measuring growth due to technological progress;
2. $\beta_i \left(\log \left(\frac{Y_{i,0}}{L_{i,0}} \right) + \log(y_i^{EQ}) + \log(A_{i,0}) \right)$: measuring growth due to the gap between initial output per worker and the steady-state value

The second source of growth is what is meant by “catching up” in the literature. As $t \to \infty$ the importance of the catch-up term, which reflects the role of initial conditions, diminishes to zero.

Constant parameters across regions

Under the additional assumptions that the rates of technological progress, and the λ_i parameters are **constant across regions**, i.e. $g_{A_i} = g_A$, and $\lambda_i = \lambda \, \forall i$, Eq. (9) may be rewritten as:

$$\overline{g_{iY/L}} = g_A t + \beta \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \beta \log(y_i^{EQ}) + \beta \log(A_{i,0}) \qquad (10)$$

Constant parameters across regions

Under the additional assumptions that the rates of technological progress, and the λ_i parameters are **constant across regions**, i.e. $g_{A_i} = g_A$, and $\lambda_i = \lambda \, \forall i$, Eq. (9) may be rewritten as:

$$\overline{g_{iY/L}} \;=\; g_A t + \beta \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \beta \log(y_i^{EQ}) + \beta \log(A_{i,0}) \qquad (10)$$

Therefore, in a cross-section of regions, we should observe a **negative relationship** between average rates of growth and initial levels of output over any time period.

Constant parameters across regions

Under the additional assumptions that the rates of technological progress, and the λ_i parameters are **constant across regions**, i.e. $g_{A_i} = g_A$, and $\lambda_i = \lambda \, \forall i$, Eq. (9) may be rewritten as:

$$\overline{g_{iY/L}} = g_A t + \beta \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \beta \log(y_i^{EQ}) + \beta \log(A_{i,0}) \qquad (10)$$

Therefore, in a cross-section of regions, we should observe a **negative relationship** between average rates of growth and initial levels of output over any time period.

Regions that start out below their balanced growth path must grow **relatively quickly** if they are to **catch up** with other regions that have the same levels of steady-state output and initial conditions.

Cross-region growth regressions

Eq. (10) provides the motivation for the standard cross-region growth regression.

Cross-region growth regressions

Eq. (10) provides the motivation for the standard cross-region growth regression.

Typically, these regression specifications start with Eq. (10) and append a random error term ν_i so that:

$$\overline{g_{iY/L}} \quad = \quad g_A t + \beta \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \beta \log(y_i^{EQ}) + \beta \log(A_{i,0}) + \nu_i \quad (11)$$

Cross-region growth regressions

Eq. (10) provides the motivation for the standard cross-region growth regression.
Typically, these regression specifications start with Eq. (10) and append a random error term ν_i so that:

$$\overline{g_{iY/L}} = g_A t + \beta \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \beta \log(y_i^{EQ}) + \beta \log(A_{i,0}) + \nu_i \quad (11)$$

Implementation of Eq. (11) requires the development of empirical analogs for $\log(y_i^{EQ})$ and $\log(A_{i,0})$.

Mankiw, Romer and Weil (1992)

Assuming a Cobb-Douglas function $F(K_{i,t}, A_{i,t}H_{i,t}) = K_{i,t}^{\alpha}(A_{i,t}H_{i,t})^{1-\alpha}$ then:

$$\dot{k}_{i,t} = s_i k_{i,t}^{\alpha} h_i^{1-\alpha} - (\delta + g_A + n_i)\, k_{i,t}; \tag{12}$$

$$k_{i,t}^{EQ} = h_i \left(\frac{s_i}{\delta + g_A + n_i} \right)^{1/(1-\alpha)}; \tag{13}$$

and:

$$y_{i,t}^{EQ} \equiv \frac{Y_{i,t}}{A_{i,t}L_{i,t}} = h_i \left(\frac{s_i}{\delta + g_A + n_i} \right)^{\alpha/(1-\alpha)}. \tag{14}$$

Mankiw, Romer and Weil (1992)

Substituting Eq. (14) into Eq. (11) we get a cross-region growth regression of the form:

$$\begin{aligned}\overline{g_{iY/L}} &= g_A + \beta \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \beta\frac{\alpha}{1-\alpha}\log(n_i + g_A + \delta) + \\ &- \beta\frac{\alpha}{1-\alpha}\log(s_i) - \beta\log(h_i) - \beta\log A_{i,0} + \nu_i \qquad (15)\end{aligned}$$

Mankiw, Romer and Weil (1992)

Substituting Eq. (14) into Eq. (11) we get a cross-region growth regression of the form:

$$\begin{aligned} \overline{g_{iY/L}} &= g_A + \beta \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \beta \frac{\alpha}{1-\alpha} \log(n_i + g_A + \delta) + \\ &- \beta \frac{\alpha}{1-\alpha} \log(s_i) - \beta \log(h_i) - \beta \log A_{i,0} + \nu_i \end{aligned} \quad (15)$$

Initial condition on TFP not observable ($A_{i,0}$) $\Rightarrow$ we need proxies!

Assumption on initial TFP

Mankiw, Romer and Weil (1992) argue that $A_{i,0}$ should be interpreted as reflecting not just technology, which they assume to be **constant across regions**, but **region-specific** influences on growth such as resource endowments, climate and institutions.

Assumption on initial TFP

Mankiw, Romer and Weil (1992) argue that $A_{i,0}$ should be interpreted as reflecting not just technology, which they assume to be **constant across regions**, but **region-specific** influences on growth such as resource endowments, climate and institutions.

They assume these differences vary randomly in the sense that:

$$\log A_{i,0} = \log A + e_i,$$

where e_i is a region-specific shock distributed independently of n_i, s_i, h_i.

Assumption on initial TFP

Mankiw, Romer and Weil (1992) argue that $A_{i,0}$ should be interpreted as reflecting not just technology, which they assume to be **constant across regions**, but **region-specific** influences on growth such as resource endowments, climate and institutions.

They assume these differences vary randomly in the sense that:

$$\log A_{i,0} = \log A + e_i,$$

where e_i is a region-specific shock distributed independently of n_i, s_i, h_i.

Substituting this into Eq. (15) and defining $\epsilon_i = \nu_i - \beta e_i$, we have the regression relationship:

$$\begin{aligned} \overline{g_{iY/L}} &= g_A - \beta \log A + \beta \log\left(\frac{Y_{i,0}}{L_{i,0}}\right) + \beta \frac{\alpha}{1-\alpha} \log(n_i + g_A + \delta) + \\ &- \beta \frac{\alpha}{1-\alpha} \log(s_i) - \beta \log(h_i) + \epsilon_i \end{aligned} \tag{16}$$

Other issues for the estimation

- **Intercept** is an estimate of the growth rate of exogenous technological progress

Other issues for the estimation

- **Intercept** is an estimate of the growth rate of exogenous technological progress
- **Constrained on parameters**

Other issues for the estimation

- **Intercept** is an estimate of the growth rate of exogenous technological progress
- **Constrained on parameters**
- **Conditional convergence** occurs when $\beta < 0$ and $\beta > -1$ and depends on $t \Rightarrow$ speed of convergence

Other issues for the estimation

- **Intercept** is an estimate of the growth rate of exogenous technological progress
- **Constrained on parameters**
- **Conditional convergence** occurs when $\beta < 0$ and $\beta > -1$ and depends on $t \Rightarrow$ speed of convergence
- All the **determinants are exogenous** and **no relevant variable are omitted** $\Rightarrow$ OLS are unbiased estimators!

Cross-region estimation

$$\overline{g_{Y/L}}_i = intercept + \beta_0 \log\left(Y/L_{i,1991}\right) + \beta_1 \bar{s}_i + \beta_2 \bar{n}_i + \beta_3 \bar{h}_i + \epsilon_i \qquad (17)$$

	Estimate	Std. Error	t-Stat.	P-value
(Intercept)	-0.0929	0.0123	-7.53	0.0000
β_0	-0.0154	0.0011	-14.57	0.0000
β_1	0.0027	0.0029	0.93	0.3532
β_2	-0.0146	0.0034	-4.31	0.0000
β_3	0.0204	0.0024	8.57	0.0000
	Res.se=0.008956 (255) DF			
	R-squared=0.6411, Adj.R-squared=0.6354			
	F-stat.=112.6 (1,255) DF, p-value=$< 2e^{-16}$			

Endogeneity in cross-region regression

Simultaneity Problem

The fact that the right-hand-side variables are not exogenous, but are **jointly determined with the growth rate** (for example the level of investment is highly correlated with growth).

- *Estimation issue*: estimates can be biased.
- *Identification issue*: the value of β can fail to illustrate how initial conditions affect expected future income differences if the saving rate is itself function of income. Hence, $\beta \geq 0$ may be compatible with at least partial convergence, while $\beta < 0$ with economic divergence if physical and human capital accumulation for rich and poor are diverging across time.

Endogeneity in cross-region regression

Measurement Error

In this case we would like to measure the (partial) effect of a variable but we can **observe only an imperfect measure** $\Rightarrow$ we introduce measurement error.

Omitted Variables

Omitted variables appear when we would like to control for one or more additional variables but, usually because of data unavailability, we cannot include them in a regression model. $\Rightarrow$ one way to represent this situation

is to write the regression equation considering the omitted variable as part of the error term.

Instrumental Variables and Two-Stage Least Squares

Consider the linear model:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + u \tag{18}$$

$$E(u) = 0,\ Cov(x_j, u) = 0,\ j = 1, 2, ..., K - 1 \tag{19}$$

therefore x_K might be correlated with u. In other words, $x_1, ..., x_{K-1}$ are exogenous while x_K is potentially endogenous

$\Rightarrow$ OLS estimation generally results in **inconsistent** estimators of all the β_j if $Cov(x_K, u) \neq 0$

Instrumental Variables and Two-Stage Least Squares (2)

The method of instrumental variables (IV) provides a general solution to the problem of an endogenous explanatory variable. To use the IV approach with x_K endogenous, we need an observable variable, z_1, not in equation (19) that satisfies two conditions:

- z_1 must be uncorrelated with u: $Cov(z_1, u) = 0 \Rightarrow z_1$ is exogenous
- The second requirement involves the relationship between z_1 and the endogenous variable, x_K. Consider the regression of x_K on *all* the exogenous variables:

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K \tag{20}$$

  where $E(e_K) = 0$ and e_K is uncorrelated with $x_1, ..., x_{K-1}$ and z_1 $\Rightarrow \theta_1 \neq 0$

z_1 is an **instrumental variable** candidate for x_K!

Two-stage least squares (2SLS) estimator

Under certain assumptions, the two-stage least squares (2SLS) estimator is the most efficient IV estimator:

1. Obtain the fitted values $\hat{x}_K$ from the regression:
$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K \tag{21}$$
This is called **first-stage regression**.
2. Run the OLS regression
$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_K \hat{x}_K + u \tag{22}$$
This is called the **second-stage regression**, and it produces the $\hat{\beta}_j$

Control Function

- For handling endogeneity in nonlinear models we must use a different approach, i.e. the **control function** (CF).

Control Function

- For handling endogeneity in nonlinear models we must use a different approach, i.e. the **control function** (CF).
- CF uses extra regressors to break the correlation between endogenous explanatory variables and unobservables affecting the dependent variable.

Control Function

- For handling endogeneity in nonlinear models we must use a different approach, i.e. the **control function** (CF).
- CF uses extra regressors to break the correlation between endogenous explanatory variables and unobservables affecting the dependent variable.
- The method still relies on the availability of exogenous variables that do not appear in the structural equation

Control Function (2)

Consider the linear model:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + u \quad (23)$$

$$E(u) = 0, \; Cov(x_j, u) = 0, \; j = 1, 2, ..., K - 1 \quad (24)$$

therefore x_K might be correlated with u. In other words, $x_1, ..., x_{K-1}$ are exogenous while x_K is potentially endogenous.

Control Function (2)

Consider the linear model:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + u \tag{23}$$

$$E(u) = 0,\ Cov(x_j, u) = 0,\ j = 1, 2, ..., K-1 \tag{24}$$

therefore x_K might be correlated with u. In other words, $x_1, ..., x_{K-1}$ are exogenous while x_K is potentially endogenous.

The *reduced form* of x_K is the linear projection of x_K onto the exogenous variables (and the instruments):

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K \tag{25}$$

with $Cov(x_j, e_K) = 0,\ \ j = 1, 2, ..., K-1$ and $Cov(z_1, e_K) = 0$.

Control Function (3)

Endogeneity arises *if and only if* u is correlated with e_K!

Control Function (3)

Endogeneity arises *if and only if* u is correlated with e_K!

Write the linear projection of u on e_K as:

$$u = \rho e_K + \epsilon \tag{26}$$

By definition, $Cov(e_K, \epsilon) = 0$, $Cov(x_j, \epsilon) = 0$ and $Cov(z_1, \epsilon) = 0$ because u and e_K are both uncorrelated with x_j $j = 1, ..., K_1$ and z_1.

Control Function (3)

Endogeneity arises *if and only if* u is correlated with e_K!

Write the linear projection of u on e_K as:

$$u = \rho e_K + \epsilon \tag{26}$$

By definition, $Cov(e_K, \epsilon) = 0$, $Cov(x_j, \epsilon) = 0$ and $Cov(z_1, \epsilon) = 0$ because u and e_K are both uncorrelated with x_j $j = 1, ..., K_1$ and z_1.

Pluggin (9) in (6) we get:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \rho e_K + \epsilon \tag{27}$$

where now e_K can be viewed as an explanatory variable in the equation, and $Cov(y, \epsilon) = 0$.

Control Function (3)

Endogeneity arises *if and only if* u is correlated with e_K!

Write the linear projection of u on e_K as:

$$u = \rho e_K + \epsilon \tag{26}$$

By definition, $Cov(e_K, \epsilon) = 0$, $Cov(x_j, \epsilon) = 0$ and $Cov(z_1, \epsilon) = 0$ because u and e_K are both uncorrelated with x_j $j = 1, ..., K_1$ and z_1.

Pluggin (9) in (6) we get:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \rho e_K + \epsilon \tag{27}$$

where now e_K can be viewed as an explanatory variable in the equation, and $Cov(y, \epsilon) = 0$.

$\Rightarrow$ run OLS of y on x_j $j = 1, ..., K_1$, z_1 and e_K using random sample.

Control Function (4)

Problem: e_K **not observable**!!

Control Function (4)

Problem: e_K **not observable**!!

- From $x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K$ we get:

$$e_K = x_K - (\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1) \tag{28}$$

Control Function (4)

Problem: e_K **not observable**!!

- From $x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K$ we get:

$$e_K = x_K - (\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1) \qquad (28)$$

- given that we observe $\mathbf{x}, z_1$ we can estimate the model (28) by OLS $\Rightarrow$ replace e_K with $\hat{e}_K$.

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \rho \hat{e}_K + error \qquad (29)$$

where:
$error = \epsilon + \rho(x_1, ..., x_{K-1}, z_1) \left[(\hat{\delta}_0, \hat{\delta}_1, ..., \hat{\delta}_{K-1}, \hat{\theta}) - (\delta_0, \delta_1, ..., \delta_{K-1}, \theta) \right]$
depends on the sampling error.

Control Function (4)

Problem: e_K **not observable**!!

- From $x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K$ we get:

$$e_K = x_K - (\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1) \tag{28}$$

- given that we observe $\mathbf{x}, z_1$ we can estimate the model (28) by OLS $\Rightarrow$ replace e_K with $\hat{e}_K$.

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \rho \hat{e}_K + error \tag{29}$$

where:
$error = \epsilon + \rho(x_1, ..., x_{K-1}, z_1) \left[(\hat{\delta}_0, \hat{\delta}_1, ..., \hat{\delta}_{K-1}, \hat{\theta}) - (\delta_0, \delta_1, ..., \delta_{K-1}, \theta) \right]$
depends on the sampling error.

$\Rightarrow$ OLS estimator of (29) will be consistent!!

Control Function (5)

- The OLS estimation of (29) is an example of CF estimator.

Control Function (5)

- The OLS estimation of (29) is an example of CF estimator.
- The inclusion of residuals $\hat{e}_K$ "controls" for the endogeneity of x_K in the original equation.

Control Function (5)

- The OLS estimation of (29) is an example of CF estimator.
- The inclusion of residuals $\hat{e}_K$ "controls" for the endogeneity of x_K in the original equation.
- The OLS estimate of β_j and $j = 1, ..., K$ are *identical* to the 2SLS.

Control Function (5)

- The OLS estimation of (29) is an example of CF estimator.
- The inclusion of residuals $\hat{e}_K$ "controls" for the endogeneity of x_K in the original equation.
- The OLS estimate of β_j and $j = 1, ..., K$ are *identical* to the 2SLS.
- Test of endogeneity: $\rho = 0$.

Control Function (5)

- The OLS estimation of (29) is an example of CF estimator.
- The inclusion of residuals $\hat{e}_K$ "controls" for the endogeneity of x_K in the original equation.
- The OLS estimate of β_j and $j = 1, ..., K$ are *identical* to the 2SLS.
- Test of endogeneity: $\rho = 0$.
- Problem: $\hat{e}_K$ is a *generated regressor* $\Rightarrow$ we need bootstrap for right standard errors!

Control Function: summaryzing

1. Obtain the **residuals** $\hat{e}_K$ from the regression:

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K \tag{30}$$

Control Function: summaryzing

1. Obtain the **residuals** $\hat{e}_K$ from the regression:

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K \tag{30}$$

2. Run the OLS regression

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \rho \hat{e}_K + error \tag{31}$$

Control Function: summaryzing

1. Obtain the **residuals** $\hat{e}_K$ from the regression:

$$x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + e_K \tag{30}$$

2. Run the OLS regression

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \rho \hat{e}_K + error \tag{31}$$

3. Test $\hat{\rho} = 0$.

References

- Wooldridge: Econometric Analysis of Cross Section and Panel Data; Chapter 5 and Chapter 6
- R file: endogeneity_EUregions.R