Master 2

Econometrics I

Nour Meddahi (Toulouse School of Economics)

Instrumental Variable Estimation and HAC

I. Instrumental Variable Estimation

Assume that we want to estimate the simple model $y = \beta\tilde{x} + \varepsilon$, $E[\varepsilon\tilde{x}] = 0$, $E[\varepsilon] = 0$. However, the variable $\tilde{x}$ is measured with an error, $x = \tilde{x} + v$ with v independent with $\tilde{x}$ and with ε, and $E[v] = 0$. For simplicity, we assume $E[\tilde{x}] = 0$. What are the properties of the OLS estimator $\hat{\beta}$?

$$\hat{\beta} = ArgMin_b \sum_{t=1}^{n}(y_t - bx_t)^2 = \frac{\sum_{t=1}^n x_t y_t}{\sum_{t=1}^n x_t^2} = \frac{\sum_{t=1}^n x_t(\beta\tilde{x}_t + \varepsilon_t)}{\sum_{t=1}^n x_t^2} = \frac{\sum_{t=1}^n x_t(\beta(x_t - v_t) + \varepsilon_t)}{\sum_{t=1}^n x_t^2}$$

$$= \beta - \beta\frac{\sum_{t=1}^n (\tilde{x}_t + v_t)v_t}{\sum_{t=1}^n x_t^2} + \frac{\sum_{t=1}^n x_t\varepsilon_t}{\sum_{t=1}^n x_t^2}$$

$$= \beta - \beta\frac{\frac{1}{n}\sum_{t=1}^n \tilde{x}_t v_t}{\frac{1}{n}\sum_{t=1}^n x_t^2} - \beta\frac{\frac{1}{n}\sum_{t=1}^n v_t^2}{\frac{1}{n}\sum_{t=1}^n x_t^2} + \frac{\frac{1}{n}\sum_{t=1}^n x_t\varepsilon_t}{\frac{1}{n}\sum_{t=1}^n x_t^2}.$$

Hence,

$$\text{Plim}\hat{\beta} = \beta - \beta\frac{E[\tilde{x}_t v_t]}{E[x_t^2]} - \beta\frac{E[v_t^2]}{E[x_t^2]} + \frac{E[x_t\varepsilon_t]}{E[x_t^2]}$$

$$= \beta - \beta\frac{Var[v_t]}{Var[\tilde{x}_t] + Var[v_t]} = \beta\frac{Var[\tilde{x}_t]}{Var[\tilde{x}_t] + Var[v_t]}.$$

Hence, the OLS estimator is biased and inconsistent when $Cov[x_t, \varepsilon_t] \neq 0$. The same inconsistent problem happens when the variable $\tilde{x}$ is endogenous.

Solution: Instrumental variable.

Assume that we have a variable z_t such that $Cov[z_t, \tilde{x}_t] \neq 0$, $Cov[z_t, v_t] = 0$ and $Cov[z_t, \varepsilon_t] = 0$. Define $\hat{x}_t$ as $\hat{x}_t = \gamma z_t$ where $x_t = \gamma z_t + \eta_t$ with $Cov[\hat{z}_t, \eta_t] = 0$ (regression of x_t on z_t). Then,

$$\begin{aligned}
\hat{\beta}_{IV} = ArgMin_b \sum_{t=1}^{n} (y_t - b\hat{x}_t)^2 &= \frac{\sum_{t=1}^{n} \hat{x}_t y_t}{\sum_{t=1}^{n} \hat{x}_t^2} = \frac{\sum_{t=1}^{n} \hat{x}_t(\beta\tilde{x}_t + \varepsilon_t)}{\sum_{t=1}^{n} \hat{x}_t^2} = \frac{\sum_{t=1}^{n} \hat{x}_t(\beta(x_t - v_t) + \varepsilon_t)}{\sum_{t=1}^{n} \hat{x}_t^2} \\
&= \frac{\sum_{t=1}^{n} \hat{x}_t(\beta(\hat{x}_t + \eta_t - v_t) + \varepsilon_t)}{\sum_{t=1}^{n} \hat{x}_t^2} \\
&= \beta + \beta \frac{\frac{1}{n}\sum_{t=1}^{n} \hat{x}_t \eta_t}{\frac{1}{n}\sum_{t=1}^{n} \hat{x}_t^2} - \beta \frac{\frac{1}{n}\sum_{t=1}^{n} \hat{x}_t v_t}{\frac{1}{n}\sum_{t=1}^{n} \hat{x}_t^2} + \frac{\frac{1}{n}\sum_{t=1}^{n} \hat{x}_t \varepsilon_t}{\frac{1}{n}\sum_{t=1}^{n} \hat{x}_t^2}. \\
\text{Hence,} \qquad \text{Plim}\hat{\beta}_{IV} &= \beta + \beta \frac{E[\hat{x}_t \eta_t]}{E[\hat{x}_t^2]} - \beta \frac{E[\hat{x}_t v_t]}{E[\hat{x}_t^2]} + \frac{E[\hat{x}_t \varepsilon_t]}{E[\hat{x}_t^2]} \\
&= \beta.
\end{aligned}$$

Hence, the IV estimator is a consistent estimator of β. In practice, be sure that $Cov[z_t, x_t] \neq 0$, otherwise one faces the problem of weak instruments.

General approach: A consistent estimator of the general model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, when $Cov[\mathbf{X}, \varepsilon] \neq 0$, can be obtained though if we could find a matrix $\mathbf{Z}$ of order $n \times l$, with $l \geq k$ (more instruments than variables) such that: 1) the variables in $\mathbf{Z}$ correlated with those in $\mathbf{X}$ and plim $\mathbf{Z}'X/n = \Sigma_{\mathbf{Z}X}$ finite and full rank. 2) plim $\mathbf{Z}'\varepsilon/n = 0$.

Pre-multiplying the regression model by $\mathbf{Z}'$ yields

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\beta + \mathbf{Z}'\varepsilon, \ \mathrm{var}(\mathbf{Z}'\varepsilon) = \sigma^2(\mathbf{Z}'\mathbf{Z}).$$

This suggests using GLS yielding the so-called instrumental variable estimator:

$$\hat{\beta}_{IV} = (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'\mathbf{y} = (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{P_Z}\mathbf{y},$$

setting $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}'$. The covariance matrix is

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

and disturbance variance may be estimated by

$$\hat{\sigma}^2_{IV} = (\mathbf{y} - \mathbf{X}\hat{\beta}_{\mathbf{IV}})'(\mathbf{y} - X\hat{\beta}_{IV})/n.$$

Special case when $l = k$. Then $\mathbf{Z}'\mathbf{X}$ non-singular yielding

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-\mathbf{1}}\mathbf{Z}'\mathbf{y} \text{ with } \mathrm{var}(\hat{\beta}_{IV}) = \sigma^2(\mathbf{Z}'\mathbf{X})^{-\mathbf{1}}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-\mathbf{1}},$$

II. Serial Correlation in the Disturbances: The HAC Estimator

While one does not use the GLS estimator when Ω is unknown, one has to estimate consistently $Var[\hat{\beta}^{OLS}]$. Under heteroskedasticity, one should use the Eicker-White estimator. However, the Eicker-White estimator is not consistent when the disturbances u_t are serially correlated. There are two leading examples:

1) Multi-horizon forecasting: $r_{t+1:t+k} = x_t'\beta + \varepsilon_{t+k}$. Due to the overlapping of periods, the disturbances ε_{t+k} are correlated. The OLS estimator is still consistent, biased in finite sample and not asymptotically. We need to estimate $Var[\hat{\beta}]$.

2) We want to estimate the mean of the short term interest rate r_t, $\bar{r}$, and a variance of $\bar{r}$. The problem is that the short term interest rate is highly correlated with unknown correlation (if we do not specify a model).

Let us focus on the second example.

$$Var[\bar{r}] = Var\left[\frac{1}{n}\sum_{t=1}^{n} r_t\right] = \frac{1}{n^2}\sum_{1\leq i,j\leq n} Cov[r_i, r_j] = \frac{1}{n}Var[r_t] + \frac{2}{n}\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)Cov[r_t, r_{t+l}],$$

under the assumption $E[r_i] = E[r_{i+h}]$ and $Cov[r_i, r_j] = Cov[r_{i+h}, r_{j+h}]$ for any i, j, h. In this case, we will say that the process r_t is a second order stationary process.

One can show that $lim_{n\to\infty} nVar[\bar{r}] = Var[r_t] + 2\sum_{l=1}^{\infty} Cov[r_t, r_{t+l}]$.

A potential estimator of $Var[\sqrt{n}\bar{r}]$ is $\hat{V}ar[r_t] + 2\sum_{l=1}^{\infty} \hat{C}ov[r_t, r_{t+l}]$, where $\hat{C}ov[r_t, r_{t+l}] = \frac{1}{n-k}\sum_{t=1}^{n-l}(r_t - \bar{r})(r_{t+l} - \bar{r})$. There are three problems. First, we have finite sample, se will not be able to estimate an infinite number of parameters. Second, we should estimate a small number of parameters, otherwise the quality of the estimators is poor. Finally, we have to be sure that the estimator is positive (univariate case) or positive definite (regression case).

A solution has been proposed by Newey and West. They show that the following estimator is positive and consistent (under some assumptions)

$$\hat{V}ar[\sqrt{n}\bar{r}] = \hat{V}ar[r_t] + 2\sum_{l=1}^{L}\left(1 - \frac{l}{L}\right)\hat{C}ov[r_t, r_{t+l}].$$

Such estimator is called a Heteroskedasticity and Autocorrelation Consistent (HAC) estimator of the standard errors. The parameter L is called the truncation parameter of the HAC estimator. L must be chosen such that it is large in large samples, although still much less than n. A good guideline is $L = 0.75n^{1/3}$.

In the regression case, form the formula $\hat{\beta} = \beta + [\mathbf{X'X}]^{-\mathbf{1}}\mathbf{X'}\varepsilon$, one gets

$$\begin{aligned} Var[\hat{\beta} \mid \mathbf{X}] &= [\mathbf{X'X}]^{-\mathbf{1}} Var[\mathbf{X'}\varepsilon][\mathbf{X'X}]^{-\mathbf{1}} \\ &= \left[\sum_{t=1}^{n} \mathbf{x_t x_t'}\right]^{-1} Var\left[\sum_{t=1}^{n} x_t \varepsilon_t\right] \left[\sum_{t=1}^{n} \mathbf{x_t x_t'}\right]^{-1} \\ &= \frac{1}{n}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{x_t x_t'}\right]^{-1} Var\left[\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t \varepsilon_t\right] \left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{x_t x_t'}\right]^{-1}. \end{aligned}$$

The Newey and West estimator of $Var\left[\sqrt{n}\sum_{t=1}^{n} x_t\varepsilon_t\right]$ is given by

$$\hat{\Sigma}_{x\varepsilon} = \hat{V}ar[x_t\varepsilon_t] + \sum_{l=1}^{L}\left(1 - \frac{l}{L}\right)\left(\hat{C}ov[x_t\varepsilon_t, x_{t+l}\varepsilon_{t+l}] + \hat{C}ov[x_t\varepsilon_t, x_{t+l}\varepsilon_{t+l}]'\right),$$

where

$$\hat{C}ov[x_t\varepsilon_t, x_{t+l}\varepsilon_{t+l}] = \frac{1}{n-k}\sum_{t=1}^{n-l}(x_t\varepsilon_t - \bar{x\varepsilon})(x_{t+l}\varepsilon_{t+l} - \bar{x\varepsilon})'.$$

Then, a positive definite estimator of the variance of $\hat{\beta}$ is

$$Var[\hat{\beta}] = \frac{1}{n}\left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{x_t x_t'}\right]^{-1} \hat{\Sigma}_{x\varepsilon} \left[\frac{1}{n}\sum_{t=1}^{n} \mathbf{x_t x_t'}\right]^{-1}.$$

Observe that the Eicker-White estimator is a special case of the HAC estimators.