Lecture 19. Endogenous Regressors and Instrumental Variables

In the previous lecture we consider a regression model (I omit the subscripts

\[(1) \quad Y = \beta_1 + \beta_2 D + u\]

The problem is that the dummy variable \(D\) is endogenous, i.e. there is a relation between \(D\) and \(u\), e.g. because both are related to the same unobserved variable. If we express the relation between \(D\) and \(u\) as a linear regression

\[u = \kappa_1 + \kappa_2 D + v\]

then the OLS estimator of the regression coefficient of \(D\) in (1) estimates \(\beta_2 + \kappa_2\) and NOT the structural regression coefficient \(\beta_2\).
Solution: Find variable \( Z \) such that

1. \( Z \) is related to \( D \).
2. \( Z \) has no direct effect on \( Y \), i.e. \( Z \) only has an effect on \( Y \) through \( D \).

Such a variable \( Z \) is called an instrumental variable.

Consider the linear regression models

(1) \[ Y = \beta_1 + \beta_2 D + u \]

(2) \[ D = \delta_1 + \delta_2 Z + \nu \]

(3) \[ Y = \gamma_1 + \gamma_2 Z + \epsilon \]

By substitution of (2) in (1) we see that the following relation exists between the \( \beta \)'s, \( \gamma \)'s, and \( \delta \)'s

\[ \gamma_1 = \beta_1 + \beta_2 \delta_1 \quad \gamma_2 = \beta_2 \delta_2 \]
In particular

\[ \beta_2 = \frac{\gamma_2}{\delta_2} \]

Because \( Z \) has no direct effect on \( Y \), it is not an omitted variable in (1) and hence not related to \( u \). Both (2) and (3) are perfectly fine regressions in which there is no relation between \( Z \) and the error. Hence the OLS estimators \( \hat{\gamma}_2 \) and \( \hat{\delta}_2 \) in (2) and (3) estimate \( \gamma_2 \) and \( \delta_2 \).

Conclusion: An estimator of \( \beta_2 \) is

(4) \[ \hat{\beta}_2 = \frac{\hat{\gamma}_2}{\hat{\delta}_2} \]

In the case that \( Z \) is also a dummy variable we have

\[ \hat{\beta}_2 = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{D}_{Z=1} - \bar{D}_{Z=0}} \]
There is an alternative method to obtain the estimator (4)

Consider again the linear regression models

(1) \[ Y = \beta_1 + \beta_2 D + u \]

(2) \[ D = \delta_1 + \delta_2 Z + v \]

The procedure consists of 2 steps

1. Estimate the regression model in (2) and obtain the predicted values

\[ \hat{D}_i = \hat{\delta}_1 + \hat{\delta}_2 Z_i \]

2. Estimate the linear regression model with dependent variable \( Y \) and independent variable \( \hat{D} \). The estimate of the regression coefficient of \( \hat{D} \) is the estimate of the structural regression coefficient \( \beta_2 \).

This estimator is called the Two-Stage Least Squares (2SLS) estimator.
To see that the 2SLS estimator indeed estimates the structural regression coefficient $\beta_2$ we write the regression equation estimated in the second step

\[ Y = \alpha_1 + \alpha_2 \hat{D} + w \]

Substitute the predicted value $\hat{D} = \delta_1 + \delta_2 Z$ to obtain

\[ Y = (\alpha_1 + \alpha_2 \delta_1) + \alpha_2 \delta_2 Z + w \]

Hence estimating a regression with $\hat{D}$ as the independent variable is the same as estimating a regression with $\delta_2 Z$ as the independent variable.
Compare this to

\[ Y = \gamma_1 + \gamma_2 Z + \varepsilon \]

OLS gives the estimator \( \hat{\gamma}_2 \). If we multiply the variable \( Z \) by a constant \( c \), then the OLS estimator of the regression coefficient of \( cZ \) is \( \hat{\gamma}_2 / c \) (exercise in homework). Hence the OLS estimator with \( \hat{\delta}_2 Z \) as independent variable is

\[ \frac{\hat{\gamma}_2}{\hat{\delta}_2} = \hat{\beta}_2 \]

which is the estimator of the structural regression coefficient that we found before.

**Conclusion:** The 2SLS estimator estimates the structural regression coefficient \( \beta_2 \).
Advantage of 2SLS procedure is that it can be used if there is more than one instrumental variable.

Warning: The standard errors and the t-values reported in the second stage regression of $Y$ on $\hat{D}$ are not correct. There is a special formula that is used in all computer packages that have 2SLS as an option.
Application: Loss of labor market experience due to participation in Vietnam war

Duration of service in Vietnam was on average 36 months. Was this lost time or did that experience contribute to the earnings potential of draftees?

Economic model for earnings derives from human capital theory

\( Y_{ti} = \delta_t + \gamma_1 S_i + \gamma_2 E_{ti} + \gamma_3 E_{ti}^2 + u_{ti} \)

with \( Y_{ti} \) earnings of individual \( i \) in year \( t \), \( S_i \) years of schooling of \( i \), and \( E_{ti} \) the labor market experience of \( i \).

- Business cycle effects \( \delta_t \)
- Linear in schooling and quadratic in experience
Experience is not directly observed. Instead we use

\[(6) \quad E_{ti} = A_{ti} - S_i - 6 - D_i \lambda \]

with \(A_{ti}\) is the age of \(i\) in year \(t\) and \(D_i\) is the indicator of participation in Vietnam war. The parameter \(\lambda\) is the loss of labor market experience due to the Vietnam war and this is what we want to estimate.

If we substitute (6) in (5) we get

\[
Y_{ti} = \delta_t + \gamma_2(A_{ti} - 6) + \gamma_3(A_{ti} - 6)^2 - \gamma_2 \lambda D_i + \gamma_3 \lambda^2 D_i - 2\gamma_3 \lambda (A_{ti} - 6) D_i + \gamma_1 S_i + \gamma_3 S_i^2 - 2\gamma_3 (A_{ti} - 6) S_i + 2\gamma_3 \lambda S_i D_i + u_{ti}
\]

We make the assumptions

- Schooling and age are independent
- Schooling independent of participation in Vietnam war (more controversial)
Under these assumptions

\[ \gamma_1 S_i + \gamma_3 S_i^2 - 2\gamma_3 (A_{ti} - 6) S_i + 2\gamma_3 \lambda S_i D_i \]

can be added to the error term.

If we define \( X_{ti} = A_{ti} - 6 \), the regression equation is

\[ Y_{ti} = \delta_t + \gamma_2 X_{ti} + \gamma_3 X_{ti}^2 + (-\gamma_2 \lambda - \gamma_3 \lambda^2) D_i - 2\gamma_2 \lambda X_{ti} D_i + v_{ti} \]

This is a linear regression

\[ Y_{ti} = \delta_t + \gamma_2 X_{ti} + \gamma_3 X_{ti}^2 + \pi_1 D_i + \pi_2 X_{ti} D_i + v_{ti} \]

with \( X_{ti} \) an exogenous and \( D_i \) an endogenous variable.

We use the lottery number \( Z_i \) as the instrumental variable and we replace \( D_i \) by the predicted value \( \hat{D}_i \). We find

\[ \hat{\lambda} = -\frac{\pi_2}{2\hat{\gamma}_2} \]
Result: See Table 5, first column.

Conclusion: Loss experience is 2 years which is less than the average service time which is 3 years, or 3 years Vietnam is 1 year US work experience.
Table 5.—Earnings-Function Models for the Veteran Effect, White Men Born 1950–52

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model (5): Loss of Experience (1)</th>
<th>Model (6): Loss of Experience, Reduced Growth Rate (2)</th>
<th>Model (7): Unrestricted Reduced Form (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>experience slope, $\beta_0$</td>
<td>0.1022</td>
<td>0.1016</td>
<td>0.1010</td>
</tr>
<tr>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td>(0.0077)</td>
<td>(0.0077)</td>
</tr>
<tr>
<td>Experience Squared, $\gamma$</td>
<td>-0.0027</td>
<td>-0.0025</td>
<td>-0.0025</td>
</tr>
<tr>
<td>(0.0003)</td>
<td>(0.0003)</td>
<td>(0.0003)</td>
<td>(0.0003)</td>
</tr>
<tr>
<td>Veteran Effect on Slope, $\beta_1$</td>
<td>0.0055</td>
<td>0.0055</td>
<td></td>
</tr>
<tr>
<td>(0.0022)</td>
<td>(0.0022)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Veteran Loss of Experience, $l$</td>
<td>2.08</td>
<td>1.84</td>
<td></td>
</tr>
<tr>
<td>(0.38)</td>
<td>(0.43)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\omega_1 = -[\beta_1 - \gamma^2 + \beta_2]$</td>
<td></td>
<td>-0.180</td>
<td></td>
</tr>
<tr>
<td>(0.052)</td>
<td></td>
<td>(0.006)</td>
<td></td>
</tr>
<tr>
<td>$\omega_2 = -[2\gamma - \beta_1]$</td>
<td></td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>(0.004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at Which Reduced Form</td>
<td>50.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Veteran Effect ($\omega_1 + \omega_2 \gamma$) = 0</td>
<td></td>
<td>(15.9)</td>
<td></td>
</tr>
<tr>
<td>$\chi^2$(dof)</td>
<td>1.41(1)</td>
<td>813.37(1247)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard errors in parentheses.
The table reports estimates of experience-earnings profiles that include parameters for the effect of veteran status. Estimates are of equations (5), (6) and (7) in the text. The estimating sample includes FICA taxable earnings from 1975–84 for men born 1950, 1976–84 earnings for men born 1951, and 1977–84 earnings for men born 1952. The estimation method is optimally weighted Two-Sample Instrumental Variables for a nonlinear model in columns (1) and (2), and for a linear model in column (3).

experience. Excluding the time-varying intercept, model (6) contains four structural parameters; $\beta_0, \beta_1, \gamma$, and $l$, and four reduced form parameters; $\beta_0, \gamma, \pi_1$ and $\pi_2$. Model (5) imposes one testable restriction on the reduced form by setting $\beta_1 = 0.15$. Table 5 shows results from nonlinear GLS estimation of (5) and (6), and results from Linear GLS estimation of (7), using data on the real FICA earnings of white men born from 1950 to 1952. The weighting matrix used in estimation was derived in a manner similar to the weighting matrix used to construct the estimates in Table 4.16 Because earnings functions are commonly fit in logs, the dependent variable is taken to be the log of mean earnings for each cell. The log of the mean is not the same as the mean of the log, but the CWHS data set does not contain the mean of log earnings. If earnings are approximately log-normally distributed, use of the log of the mean will provide a reasonable approximation. In practice, estimates of earnings by cohort $c$ and suppose that each time-series is of length $T$. Then the second term in the optimal weighting matrix has the following block corresponding to the time series of earnings for lottery number cell $j$ of cohort $c$:

$$(1/n)\Pi_{c}\left(\frac{1}{n}n_{cj}\right)\Sigma_{j}$$

$s_{T}\pi_{T}$

where $s_{T}$ is a vector of $T$ 1's. In practice, $\Pi_{c}$ is replaced by weighted least squares estimates (weights are the inverse sampling variances of $\hat{\beta}_c$) of the reduced form equation (7). Estimates of models for the log of earnings replace $s_{T}\pi_{T}$ with data-derived estimates of the covariance matrix of log $\gamma$.