Economic Policy Analysis (EC406)

Problems in Applied Econometrics

John Van Reenen

Centre for Economic Performance

London School of Economics

* These lecture notes are incomplete. Reading the lecture notes without attending lectures can be very misleading.
Motivation

How does one estimate econometrically the effects of economic policies on outcomes of interest?

Examples:

- What is the impact of employment protection legislation on employment, productivity and economic growth?
- What are the effects of government policies to promote primary, secondary or tertiary education on levels of educational attainment and wages?
- What are the effects of policies to alleviate malaria, AIDS and other diseases on economic development?

What are the econometric problems in estimating the effects of such policies?
Structure of Subsequent Lectures

This lecture
- Assumptions needed to consistently estimate parameters of interest using Ordinary Least Squares (OLS)
- Problems when these assumptions are not satisfied
  - Omitted variable bias
  - Unobserved heterogeneity and panel data methods
  - Endogeneity and Instrumental Variables

Subsequent lectures
- The policy evaluation problem
- Econometric methods for policy evaluation
Ordinary Least Squares Estimation

Suppose that we have a cross-section of data on, for example, individuals \( i \in \{1, \ldots, n\} \) for a single year.

Consider the following simple linear regression model:

\[
y = \beta_0 + \beta_1 x_{1i} + u_i
\]

One key assumption in estimating this equation using OLS is

\[
E(u_i \mid x_{1i}) = E(u_i) = 0 \quad \text{and} \quad Cov(x_{1i}, u_i) = 0
\]

Many applied econometric problems relate to concerns that this assumption may not be satisfied.
Omitted Variables Bias

Suppose that the true economic model is:

\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \quad \text{and} \quad E(u_i | x_{1i}, x_{2i}) = 0 \]

Suppose that the econometrician estimates the mis-specified model:

\[ y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \quad \text{and} \quad E(\varepsilon_i | x_{1i}) \neq 0 \]

The OLS estimate of \( \beta_1 \) from the mis-specified model is:

\[ \tilde{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1) y_i}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2} \]
Omitted Variables Bias

Substituting for \( y \) using the true model and taking expectations conditional on the value of the independent variables:

\[
E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1) x_{2i}}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}
\]

Where the term after \( \beta_2 \) is the estimated slope coefficient from an OLS regression of the omitted variable \( x_2 \) on \( x_1 \)

\[
E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1
\]

If \( \beta_2 \neq 0 \) and \( \delta_1 \neq 0 \), OLS is biased (and inconsistent). This is precisely the case where \( E(\varepsilon_i | x_{1i}) \neq 0 \) in the mis-specified model.
### Direction of Bias

<table>
<thead>
<tr>
<th>$\beta_2 &gt; 0$</th>
<th>$\text{Corr}(x_1, x_2) &gt; 0$</th>
<th>$\text{Corr}(x_1, x_2) &lt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive bias [away from zero]</td>
<td>Negative bias [towards zero]</td>
<td></td>
</tr>
<tr>
<td>$\beta_2 &lt; 0$</td>
<td>Negative bias [away from zero]</td>
<td>Positive bias [towards zero]</td>
</tr>
</tbody>
</table>

In the more general case, where the mis-specified model has multiple included independent variables and multiple omitted independent variables, the direction of bias is harder to sign.
Unobserved Heterogeneity

Suppose that we have panel (longitudinal) data on individual states $i \in \{1, \ldots, I\}$ over time $t \in \{1, \ldots, T\}$.

Consider the following linear regression model:

$$ y_{it} = \beta_0 + \beta_1 x_{1it} + \delta_t d_t + v_{it} $$

$$ v_{it} = \eta_i + u_{it} $$

$$ \text{Cov}(x_{1it}, \eta_i) \neq 0 $$

$$ \text{Cov}(x_{1it}, u_{is}) = 0 \text{ for all } t, s $$

$\eta_i$: unobserved effect capturing time-invariant determinants of $y$

$d_t$: omit dummy for first time period (normalization)

Pooling cross-sections across time and estimating this model by OLS:

- Yields incorrect standard errors (OLS assumes that $v_{it}$ is independently and identically distributed across $i$ and $t$)
- Yields biased and inconsistent estimates of $\beta_1$ since

$$ \text{Cov}(x_{1it}, \eta_i) \neq 0 $$
Differenced Estimators

Taking differences across time eliminates the unobserved effect:

\[ y_{it} = \beta_0 + \beta_1 x_{1it} + \delta_t d_t + \eta_i + u_{it} \]

\[ \text{Cov}(x_{1it}, u_{is}) = 0 \text{ for all } t, s \]

\[ \Delta y_{it} = \beta_1 \Delta x_{1it} + \delta_t \Delta d_t + \Delta u_{it} \]

Suppose \( x_i \) subject to classical measurement error:

- Estimated coefficients biased towards zero (attenuation bias)
- Differencing typically reduces the amount of variation due to true variation in the explanatory variables relative to measurement error
- Longer the time interval across which differences are taken, the greater the relative importance of true variation in the explanatory variables
Fixed Effects (Within Groups) Estimation

Consider the following linear regression model:

\[ y_{it} = \beta_1 x_{1it} + \eta_i + u_{it} \]

\[ \text{Cov}(x_{1it}, \eta_i) \neq 0 \quad \text{Cov}(x_{1it}, u_{is}) = 0 \text{ for all } t,s \]

Take time means for each \( i \) in the above equation:

\[ \bar{y}_i = \beta_1 \bar{x}_{1i} + \eta_i + \bar{u}_i \]

Take differences between each variable and its time mean. Estimating the transformed model using pooled OLS yields an unbiased and consistent estimate of \( \beta_1 \):

\[ (y_{it} - \bar{y}_i) = \beta_1 (x_{1it} - \bar{x}_{1i}) + (u_{it} - \bar{u}_i) \]

From the estimated value of \( \beta_1 \), obtain estimates of the fixed effect:

\[ \hat{\eta}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{1i} \]
Fixed Time and Group Effects

Include full set of fixed effects and time dummies:

\[ y_{it} = \beta_0 + \beta_1 x_{1it} + \delta_t d_t + \eta_i + u_{it} \]

Slope coefficient estimated from transformed model using pooled OLS:

\[
\begin{align*}
(y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}) &= \beta_1 (x_{it} - \bar{x}_{1i} - \bar{x}_t + \bar{x}) + (u_{it} - \bar{u}_i - \bar{u}_t + \bar{u}) \\
\end{align*}
\]

Obtain estimated fixed effects from:

\[
\begin{align*}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\eta}_i &= (\bar{y}_i - \bar{y}) - \hat{\beta}_1 (\bar{x}_{1i} - \bar{x}) \\
\hat{\delta}_t &= (\bar{y}_t - \bar{y}) - \hat{\beta}_1 (\bar{x}_t - \bar{x})
\end{align*}
\]
Endogeneity and Instrumental Variables
(drop i sub-script for notational simplicity)

Consider the following true economic model:

\[ y = \beta_0 + \beta_1 x + u \]
\[ \text{Cov}(x,u) \neq 0 \]

Search for instrumental variables \( z_j, j \in \{1, \ldots, J\} \), that satisfy:

- **Order condition**: \( \text{Cov}(z_j,u) = 0 \)
- **Rank condition**: \( \text{Cov}(z_j,x) \neq 0 \)

First-stage regression (if 2 IV’s):

\[ x = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \nu \]

Yields predicted values:

\[ \hat{x} = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 \]
\[ x = \hat{x} + \hat{\nu} \]

Second-stage regression:

\[ y = \beta_0 + \beta_1 \hat{x} + \beta_1 \hat{\nu} + u \]
\[ \text{Cov}(\hat{x},u) = 0 \]
\[ \text{Cov}(\hat{x},\hat{\nu}) = 0 \]
Endogeneity and Instrumental Variables

Estimating the second-stage regression using OLS yields consistent estimates of $\beta_1$.

OLS standard errors are incorrect and need to be adjusted, because they do not take into account the fact that second-stage predicted variables have an prediction error (*generated regressors*).

Reduced form equation:

\[ y = \beta_0 + \beta_1 \pi_1 z_1 + \beta_1 \pi_2 z_2 + \beta_1 v + u \]

**Weak instruments problem**: important to test for the statistical significance of the instruments in the first-stage regression (“Rank condition”)

If instruments are weakly correlated with the endogenous variable, 2SLS is biased towards OLS (e.g. Bound, Jaeger and Baker 1995)
Testing for Endogeneity

**OLS versus 2SLS**
- If $x$ is endogenous, OLS is inconsistent and 2SLS is consistent
- If $x$ is exogenous, both OLS and 2SLS are consistent, but 2SLS is inefficient

**Two approaches for testing for endogeneity**
- Hausman (1978) specification test, compares the change in the estimated coefficients and standard errors between OLS and IV
- Include the residuals from the first-stage regression in the second-stage equation and test the statistical significance of the estimated coefficient on the residuals

\[ y = \beta_0 + \beta_1 x + \gamma_1 \hat{v} + \omega \]

Null hypothesis: $\gamma_1 = 0$. If the null hypothesis is rejected, conclude that $x$ is endogenous
Testing for Instrument Validity

A key assumption in instrumental variable estimation is:

\[ \text{Cov}(z_j, u) = 0 \]

Suppose that, as in our example, we have more instruments than endogenous variables (the model is \textit{over-identified}).

Under the assumption that one of the instruments is exogenous, we can test the validity of the other instrument.

Intuition: estimate the preceding model using 2SLS and only one of the two instruments. Compute the following residuals:

\[ \hat{u} = y - \hat{\beta}_0 - \hat{\beta}_1 x \]

Under the null hypothesis that this instrument is exogenous, the other instrument if valid should be uncorrelated with these residuals.
Sargan Test of the Model’s Overidentifying Restrictions

(1) Estimate the second-stage equation by 2SLS and obtain the 2SLS residuals
(2) Regress the 2SLS residuals on all exogenous variables - the instruments (excluded exogenous variables) and the other exogenous variables in the second-stage equation. Obtain the $R^2$
(3) Under the null hypothesis that the excluded exogenous variables are uncorrelated with the second-stage error, $nR^2 \sim \chi^2(q)$

where $q$ is the number of excluded exogenous variables minus the total number of endogenous explanatory variables

If $nR^2$ exceeds, for example, the 5% critical value in the $\chi^2(q)$ distribution, we reject the null hypothesis and conclude that at least some of the excluded variables are not exogenous
Empirical Application: Labour Regulation in India


- India = federal democracy
- Industrial Disputes Act (1947) – Central act
- State governments given right to amend Central act under Indian Constitution
- > 120 amendments up to 1992. Classify each as pro-worker (+1), pro-employer (-1), or neutral (0)
- Look at net direction of change in each year
- Pro-employer, Pro-worker, and Control states

Example: West Bengal 1980 – Prior payment of compensation to the worker is a condition precedent to the closure of an undertaking. Under the central act payment of compensation does not need to be made prior to closure
Figure 1: Labour Regulation in India: 1958-1992
Figure 2: Registered Manufacturing Output Per Capita: 1958-1992
Figure 3: Registered Manufacturing Employment: 1958-1992
Attractive Features of this Empirical Application

Unobserved heterogeneity in other institutions and policies less of a concern across states within a country than across countries

Direct policy-based measure of labor regulation

Cross-section and time-series variation
  – Inclusion of state fixed effects means that parameters of interest identified from time-series variation

Labor regulation only applies to a specific sector of the economy – registered manufacturing as opposed to unregistered

India was a largely closed economy during 1947-91
Empirical Specification

Consider panel data regressions of the form:

$$y_{st} = \alpha_s + \beta_t + \mu r_{st-1} + \xi x_{st} + \varepsilon_{st}$$

Where

- $y_{st}$ is a logged outcome variable
- $r_{st-1}$ is the measure of labor regulation (lagged one year)
- $x_{st}$ are exogenous variables of interest that help explain the outcomes
- $\alpha_s$ is a state fixed effect
- $\beta_t$ is a year effect
Outcomes and Control Variables

**Outcomes of interest**
- State output disaggregated by sector (registered and unregistered manufacturing)
- Registered manufacturing workers and employees
- Wages per worker and productivity per employee
- Poverty measures

**Control variables**
- Population growth
- Development expenditure per capita
- Installed electricity capacity per capita
Table 3

<table>
<thead>
<tr>
<th>Method</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log state output per</td>
<td>OLS</td>
<td>OLS</td>
<td>OLS</td>
<td>OLS</td>
<td>OLS</td>
<td>OLS</td>
<td>OLS</td>
</tr>
<tr>
<td>capita</td>
<td>(0.14)</td>
<td>(1.81)</td>
<td>(1.69)</td>
<td>(0.29)</td>
<td>(2.05)</td>
<td>(2.90)</td>
<td>(2.46)</td>
</tr>
<tr>
<td></td>
<td>−0.002</td>
<td>0.019*</td>
<td>−0.034*</td>
<td>−0.019</td>
<td>−0.073**</td>
<td>−0.186***</td>
<td>0.086**</td>
</tr>
<tr>
<td>State effects</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Year effects</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.93</td>
<td>0.84</td>
<td>0.95</td>
<td>0.76</td>
<td>0.93</td>
<td>0.93</td>
<td>0.75</td>
</tr>
<tr>
<td>Observations</td>
<td>509</td>
<td>509</td>
<td>509</td>
<td>509</td>
<td>509</td>
<td>508</td>
<td>509</td>
</tr>
</tbody>
</table>

**Note:** Absolute $t$-statistics calculated using robust standard errors clustered at the state level are reported in parentheses. *significant at 10 percent, **significant at 5 percent, ***significant at 1 percent. Total, nonagricultural, agricultural, total manufacturing, registered manufacturing, and unregistered manufacturing output figures are all components of state domestic product and are expressed in log real per capita terms. State amendments to the Industrial Disputes Act are coded 1 = pro-worker, 0 = neutral, −1 = pro-employer and then cumulated over the period to generate the labor regulation measure. The data are for the sixteen main states for the period 1958–1992. Haryana split from the Punjab in 1966, and, after this date, we include Haryana as a separate observation. We therefore have a total of 862 possible observations with deviations accounted for by missing data. See Appendix 1 for details on the construction and sources of the variables.
TABLE V


<table>
<thead>
<tr>
<th>Method</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log registered manufacturing employment</td>
<td>Log daily employment in registered manufacturing</td>
<td>Log earnings per worker in registered manufacturing</td>
<td>Log fixed capital per capita</td>
<td>Log number of factories per capita</td>
<td>Log value added per employee</td>
</tr>
<tr>
<td>Labor regulation</td>
<td>−0.072*</td>
<td>−0.285***</td>
<td>0.008</td>
<td>−0.120**</td>
<td>−0.234**</td>
<td>−0.127**</td>
</tr>
<tr>
<td>[r − 1]</td>
<td>(1.70)</td>
<td>(3.48)</td>
<td>(0.09)</td>
<td>(2.49)</td>
<td>(3.44)</td>
<td>(2.16)</td>
</tr>
<tr>
<td>Log development expenditure per capita</td>
<td>0.076</td>
<td>0.327*</td>
<td>0.207</td>
<td>0.594***</td>
<td>0.229</td>
<td>0.262***</td>
</tr>
<tr>
<td>(0.84)</td>
<td>(1.82)</td>
<td>(1.52)</td>
<td>(2.93)</td>
<td>(1.50)</td>
<td>(2.09)</td>
<td></td>
</tr>
<tr>
<td>Log installed electricity capacity per capita</td>
<td>0.073</td>
<td>0.111</td>
<td>0.019</td>
<td>0.232*</td>
<td>0.037</td>
<td>−0.034</td>
</tr>
<tr>
<td>(1.34)</td>
<td>(1.51)</td>
<td>(0.34)</td>
<td>(1.82)</td>
<td>(0.95)</td>
<td>(0.45)</td>
<td></td>
</tr>
<tr>
<td>Log state population</td>
<td>−0.099</td>
<td>2.122</td>
<td>1.116</td>
<td>−1.130</td>
<td>1.18</td>
<td>−1.19</td>
</tr>
<tr>
<td>(0.09)</td>
<td>(1.14)</td>
<td>(0.93)</td>
<td>(0.61)</td>
<td>(0.42)</td>
<td>(0.81)</td>
<td></td>
</tr>
<tr>
<td>Congress majority</td>
<td>0.008</td>
<td>−0.009</td>
<td>−0.037*</td>
<td>0.008</td>
<td>−0.006</td>
<td>0.009</td>
</tr>
<tr>
<td>(0.61)</td>
<td>(0.39)</td>
<td>(1.66)</td>
<td>(0.43)</td>
<td>(0.36)</td>
<td>(0.73)</td>
<td></td>
</tr>
<tr>
<td>Hard left majority</td>
<td>−0.028</td>
<td>−0.124***</td>
<td>0.0004</td>
<td>0.001</td>
<td>−0.044*</td>
<td>0.019</td>
</tr>
<tr>
<td>(1.43)</td>
<td>(3.93)</td>
<td>(0.01)</td>
<td>(0.05)</td>
<td>(1.81)</td>
<td>(0.90)</td>
<td></td>
</tr>
</tbody>
</table>
Table 8

<table>
<thead>
<tr>
<th>Method</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall head count</td>
<td>Urban head count</td>
<td>Rural head count</td>
<td>Urban head count</td>
<td>Urban head count</td>
<td>Urban head count</td>
<td>Urban head count</td>
</tr>
<tr>
<td>Labor regulation ( \ell - 1 )</td>
<td>-0.008</td>
<td>2.288***</td>
<td>-0.821</td>
<td>2.070**</td>
<td>-0.270</td>
<td>2.251**</td>
<td>1.916**</td>
</tr>
<tr>
<td>(0.01)</td>
<td>(3.31)</td>
<td>(0.48)</td>
<td>(2.52)</td>
<td>(0.30)</td>
<td>(2.52)</td>
<td>(1.99)</td>
<td></td>
</tr>
<tr>
<td>Log development expenditure per capita</td>
<td>-3.468</td>
<td>-0.983</td>
<td>-2.900</td>
<td>-4.044</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.82)</td>
<td>(0.32)</td>
<td>(0.79)</td>
<td>(0.94)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log installed electricity capacity per capita</td>
<td>0.242</td>
<td>1.260</td>
<td>1.058</td>
<td>0.875</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.28)</td>
<td>(1.60)</td>
<td>(1.02)</td>
<td>(1.27)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log state population</td>
<td>-5.448</td>
<td>38.74</td>
<td>-3.717</td>
<td>-10.42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.29)</td>
<td>(1.28)</td>
<td>(0.19)</td>
<td>(0.56)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congress majority</td>
<td>0.418**</td>
<td>0.206</td>
<td>0.464**</td>
<td>0.452**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1.98)</td>
<td>(0.63)</td>
<td>(2.36)</td>
<td>(1.99)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

Applied economic problems in estimating effects of economic policies?
- Omitted variable bias
- Unobserved heterogeneity
- Endogeneity

Econometric methods for addressing these concerns
- Instrumental Variables
- Differenced and Fixed Effects Estimators

Importance of identifying assumptions. What sources of variation in the data are being used to identify parameters of interest?

Empirical application
- Labor regulation in India
- Fixed effects estimation and differences across sectors