Instrumental Variables with Treatment Effect Heterogeneity: 
Local Average Treatment Effects

The treatment effect literature offers new insights on the estimand of an instrumental variable (IV) estimator—how we should interpret an IV estimate if we allow the effect of an endogenous regressor being instrumented for to be heterogeneous.

1 Instrumental Variable (IV) Estimator

1.1 Potential Treatment Responses

Consider a hypothetical program that randomly selects some people to receive a special education \((z=1)\); others will not receive that education \((z=0)\). Suppose that we are interested in the effect of doing exercises \((d)\) on physical health \((y)\), but not the effect of education \((z)\) on physical health \((y)\), which implies the following:

\[ z \rightarrow d \rightarrow y. \]

There is no direct effect of \(z\) on \(y\), because the education itself is unlikely to directly affect physical health (but perhaps could directly affect mental health). The variable \(z\), which affects \(d\) but not \(y\) directly, can be used as an instrument \(z\) for \(d\). In our potential outcome model an observed outcome \((y)\) can be represented by a linear combination of the untreated outcome \((y_0)\) and treated outcome \((y_1)\) by the observed treatment status \((d)\). Similarly, the observed treatment status \((d)\) can be represented by another linear combination of two potential treatment responses—\(d_1\) when \(z\) is present and \(d_0\) when \(z\) is absent—by the instrument \((z)\).

To fix this idea, let’s use the following simple model for an individual \(i\):

- **Potential treatment responses:**
  \[ d_{0i} = \alpha_1 + \epsilon_i; \]
  \[ d_{1i} = \alpha_1 + \alpha_2 + \epsilon_i; \]

- **Observed treatment status:**
  \[ d_i = (1 - z_i) d_{0i} + z_i d_{1i}; \]
  \[ y_{0i} = \beta_1 + u_i; \]
  \[ y_{1i} = \beta_1 + \beta_2 + u_i; \]

- **Potential outcomes:**
  \[ y_{0i} = \beta_1 + u_i; \]
  \[ y_{1i} = (1 - d_i) y_{0i} + d_i y_{1i}. \]

Then we can get

\[ d_i = \alpha_1 + \alpha_2 z_i + \epsilon_i; \]
\[ y_i = \beta_1 + \beta_2 d_i + u_i. \]

The potential outcomes can be used to define a treatment effect such as \((y_{1i} - y_{0i})\) for individual \(i\). Here the potential treatment responses can be used to categorize individual \(i\) into one of the four types, using the terms coined by Angrist, Imbens, and Rubin (1996):

- \(d_{0i} = 0, d_{1i} = 0\): *never-takers* (not taking the treatment, regardless of \(z_i\));
- \(d_{0i} = 0, d_{1i} = 1\): *compliers* (taking the treatment if and only if \(z_i\) present);
- \(d_{0i} = 1, d_{1i} = 0\): *defiers* (taking the treatment if and only if \(z_i\) absent);
- \(d_{0i} = 1, d_{1i} = 1\): *always-takers* (taking the treatment, regardless of \(z_i\)).

Using our previous education-exercise example, we can describe never-takers as those who would never do exercises regardless of receiving the education or not—“the lazy;” compliers do exercises if and only if
they receive the education—“the spurred;” defiers are “the rebel;” and always-takers do exercises no matter whether they receive the education or not—“the diligent.”

Note that to identify the type of individual \( i \) we need both \( d_0 \) and \( d_1 \)—how individual \( i \) responds in the presence \((z = 1)\) and in the absence \((z = 0)\) of the instrument \((z)\), which is impossible because we observe only one of \( d_0 \) and \( d_1 \) at a time for any given individual. Therefore, for individual \( i \) we cannot identify his or her type.

Substituting the \( d \)-equation into the \( y \)-structural form, we will get the \( y \)-reduced form:

\[
y_i = \beta_1 + \beta_2 (\alpha_1 + \alpha_2 z_i + \epsilon_i) + u_i = (\beta_1 + \beta_2 \alpha_1) + \beta_2 \alpha_2 z_i + (u_i + \beta_2 \epsilon_i).
\]

If \( \text{Cov}(z, \epsilon) = 0 \) and \( \text{Cov}(z, u) = 0 \), then the OLS estimator (regressing \( y_i \) on 1 and \( z_i \)) will be consistent for \((\beta_1 + \beta_2 \alpha_1)\) and \( \beta_2 \alpha_2 \). The slope parameter for \( z_i, \gamma_2 \equiv \beta_2 \alpha_2 \), represents the product of two effects: 1) \( \alpha_2 \) for \( z_i \) on \( d_i \); and 2) \( \beta_2 \) for \( d_i \) on \( y_i \). To obtain a consistent estimate of \( \beta_2 \), we can use an indirect least squares (ILS) estimator: running OLS of \( d \) on 1 and \( z \) to estimate \( \alpha_2 \) and then dividing the OLS estimate \( \gamma_2 \) by the estimated \( \alpha_2 \). Alternatively, we can estimate \( \beta_2 \) consistently in one step. Define the following

\[
\begin{align*}
z_i &\equiv (1, z_i)', d_i \equiv (1, d_i)', \alpha \equiv (\alpha_1, \alpha_2)', \beta \equiv (\beta_1, \beta_2)', \\
&\text{and we will then get} \\
d_i &= z_i' \alpha + \epsilon_i; \\
y_i &= d_i' \beta + u_i.
\end{align*}
\]

Suppose that \( \text{Cov}(\epsilon_i, u_i) \neq 0 \), which means that \( \text{Cov}(d_i, u_i) \neq 0 \). Then the OLS of \( y \) on \( d \) will be inconsistent for \( \beta \). Assuming that \( \text{Cov}(z_i, u_i) = 0 \) and \( \alpha_2 \neq 0 \), we can use \( z \) for an IV estimator to obtain a consistent estimate of \( \beta \) as follows:

\[
\left( \frac{1}{N} \sum_{i=1}^{N} z_i d_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^{N} z_i y_i \right) \xrightarrow{P} \left[ \mathbb{E}(zd') \right]^{-1} \mathbb{E}(zy) = \beta + \left[ \mathbb{E}(zd') \right]^{-1} \mathbb{E}(zu) = \beta.
\]

The last equation highlights that the instrument \( z_i \) must satisfy these two conditions:

\[
\text{Cov}(z_i, d_i) \neq 0 \text{ and } \text{Cov}(z_i, u_i) = 0.
\]

The former is sometimes called the inclusion restriction, which is indicated by \( \alpha_2 \neq 0 \). The latter is called the exclusion restriction, which will hold if there is randomization of \( z \), ensuring that \( z \) will not directly affect \( y \) (by becoming a regressor in the \( y \)-equation).

There are two important results for the IV estimator just discussed.

- First, the population version of the slope IV estimate—that is \( \beta_2 \)—can be written as

\[
\beta_2 = \frac{\mathbb{E}(y|z = 1) - \mathbb{E}(y|z = 0)}{\mathbb{E}(d|z = 1) - \mathbb{E}(d|z = 0)} \quad \text{indirect effect of } z \text{ on } y
\]

Its sample version is called the Wald estimator. Note that we also have

\[
\beta_2 = \frac{\text{Cov}(z, y)}{\text{Cov}(z, d)} = \frac{\mathbb{E}(y|z = 1) - \mathbb{E}(y|z = 0)}{\mathbb{E}(d|z = 1) - \mathbb{E}(d|z = 0)}.
\]

- Second, the Wald estimator can be interpreted as \( \mathbb{E}(y_1 - y_0|\text{compliers}) \) if the treatment effect is
heterogeneous—the effect varies across individuals:

\[
\beta_2 = \frac{\mathbb{E}(y|z = 1) - \mathbb{E}(y|z = 0)}{\mathbb{E}(d|z = 1) - \mathbb{E}(d|z = 0)} = \mathbb{E}(y_1 - y_0|\text{compliers})
\]

\[
\equiv \text{Local Average Treatment Effect (LATE)}.
\]

1.2 Sources for Instruments

The concept of IV estimators is fundamental and IV estimators have been widely used in empirical economic studies. In practice the main problem with IV estimators is not what to do with the instruments, but where to find them in the first place. For the sources for instruments, please read \textit{MHE} 4.1 “IV and Causality” carefully.

2 Wald Estimator, IV Estimator and Compliers

2.1 Wald Estimator for Homogeneous (Constant) Effect

Recall that regressing \(y\) on 1 and \(d\) produces a slope estimate (associated with \(d\)) equal to the group (indicated by \(d\)) mean difference. Although the consistency of an IV estimator for \(\beta\) can be shown easily, there is an illuminating interpretation of an IV estimator when the instrument \(z\) is binary, linking the IV estimator to group mean differences \(\mathbb{E}(y|z = 1) - \mathbb{E}(y|z = 0)\) and \(\mathbb{E}(d|z = 1) - \mathbb{E}(d|z = 0)\).

Recall the previous model:

\[
d_i = \alpha_1 + \alpha_2 z_i + \epsilon_i;
\]

\[
y_i = \beta_1 + \beta_2 d_i + u_i;
\]

\[
\Rightarrow y_i = (\beta_1 + \beta_2 \alpha_1) + \beta_2 \alpha_2 z_i + (u_i + \beta_2 \epsilon_i).
\]

Define \(z_i \equiv (1, z_i)', \ d_i \equiv (1, d_i)', \ \alpha \equiv (\alpha_1, \alpha_2)', \ \text{and} \ \beta \equiv (\beta_1, \beta_2)'.\) For an IV estimator, the slope \(\beta_2\) in

\[
\beta = [\mathbb{E}(zd^\prime)]^{-1} \mathbb{E}(zy)
\]

can be written as

\[
\beta_2 = \frac{\text{Cov}(y, z)}{\text{Cov}(d, z)} = \frac{\mathbb{E}(yz) - \mathbb{E}(y) \mathbb{E}(z)}{\mathbb{E}(dz) - \mathbb{E}(d) \mathbb{E}(z)}.
\]

Because \(\mathbb{E}(dz) = \mathbb{E}(d|z = 1) \Pr(z = 1)\) and \(d = d[z + (1 - z)]\), we can rewrite the denominator as

\[
\mathbb{E}(dz) - \mathbb{E}(d) \mathbb{E}(z) = \mathbb{E}(d|z = 1) \Pr(z = 1) - \mathbb{E}[d(z + (1 - z))] \Pr(z = 1)
\]

\[
= \mathbb{E}(d|z = 1) \Pr(z = 1) - \mathbb{E}(dz) \Pr(z = 1) - \mathbb{E}[d(1 - z)] \Pr(z = 1)
\]

\[
= \mathbb{E}(d|z = 1) \Pr(z = 1) - \mathbb{E}(d|z = 1) \Pr(z = 1)^2 - \mathbb{E}[d(1 - z)] \Pr(z = 1)
\]

\[
= \mathbb{E}(d|z = 1) \Pr(z = 1) \mathbb{E}(z = 0) - \mathbb{E}[d|z = 0] \Pr(z = 0) \Pr(z = 1)
\]

\[
= \mathbb{E}(d|z = 1) - \mathbb{E}(d|z = 0)] \Pr(z = 1) \Pr(z = 0).
\]

Similarly, we can rewrite the numerator as

\[
[\mathbb{E}(y|z = 1) - \mathbb{E}(y|z = 0)] \Pr(z = 1) \Pr(z = 0).
\]

Cancelling \(\Pr(z = 1) \Pr(z = 0)\) that appears in both the numerator and the denominator, we get

\[
\frac{\text{Cov}(y, z)}{\text{Cov}(d, z)} = \frac{\mathbb{E}(y|z = 1) - \mathbb{E}(y|z = 0)}{\mathbb{E}(d|z = 1) - \mathbb{E}(d|z = 0)} = \frac{\beta_2 \alpha_2}{\alpha_2} = \beta_2.
\]
The equality \( \beta_2 \alpha_2 / \alpha_2 \) is obtained from the numerical equivalence between regressing \( y \) (or \( d \)) on 1 and \( z \) and the group (indicated by \( z \)) mean difference for \( y \) (or \( d \)). The sample version for this is referred to as the Wald estimator:

\[
\hat{\beta}_2 = \frac{E(d|z = 1) - E(d|z = 0)}{E(y|z = 1) - E(y|z = 0)} = \frac{\sum_i y_i z_i - \sum_i y_i (1 - z_i)}{\sum_i d_i z_i - \sum_i d_i (1 - z_i)}.
\]

In contrast, the sample version for \( Cov(y, z) / Cov(d, z) \) is the so-called IV estimator. In the diagram \( z \to d \to y \), the numerator of the Wald estimator is for the indirect effect of \( z \) on \( y \), and the denominator is for the effect of \( z \) on \( d \). Doing the division, we uncover the direct effect of \( d \) on \( y \).

In a clinical trial in which \( z \) is a random assignment and \( d \) is “compliance” if \( d = z \) and “non-compliance” if \( d \neq z \), \( E(y|z = 1) - E(y|z = 0) \) is called the intent-to-treatment effect (ITT)—it means the effect of treatment intention (or assignment), but not the effect of the actual treatment received. Non-compliance to treatment can “dilute” the real effect, and the Wald estimator adjusts the diluted effect using the factor \( [E(d|z = 1) - E(d|z = 0)]^{-1} \).

### 2.2 Wald Estimator for Heterogeneous Effects

Suppose that

(a) \( Pr(d_i = 1|z_i) \) is a non-constant function of \( z \);
(b) \( (y_{0i}, y_{1i}, d_{0i}, d_{1i}) \) is independent of \( z_i \); and
(c) for all \( i \), either \( d_{1i} \geq d_{0i} \) or \( d_{1i} \leq d_{0i} \).

Condition (a) is the inclusion restriction, meaning that \( z \) is a relevant variable in the \( d \)-equation. Condition (b) implies the exclusion restriction, meaning that \( z \) enters the \( y \)-equation only through the \( d \)-equation. Condition (c) is a monotonicity restriction.

Without loss of generality, we assume \( d_{1i} \geq d_{0i} \), which precludes the existence of defiers (that is, those with \( d_{1i} = 0 \) and \( d_{0i} = 1 \)). Also note that

\[
Pr(d = 1|z = 1) = \frac{Pr(d = 1|z = 1)}{Pr(d = 1|z = 0)} > \frac{Pr(d = 1|z = 0)}{Pr(d = 1|z = 0)} = Pr(d = 1|z = 0).
\]

Using \( d = zd_1 + (1 - z)d_0 \) and \( y = dy_1 + (1 - d)y_0 \), we can get the following:

\[
E(y|z = 1) - E(y|z = 0) = E[dy_1 + (1 - d)y_0|z = 1] - E[dy_1 + (1 - d)y_0|z = 0] = E[d_1 y_1 + (1 - d_1)y_0|z = 1] - E[d_0 y_1 + (1 - d_0)y_0|z = 0] = E[(d_1 - d_0)(y_1 - y_0)].
\]

Consider the cases of \( d_1 - d_0 \) being \( \pm 1 \):

\[
E[(d_1 - d_0)(y_1 - y_0)] = E[(y_1 - y_0)d_1 - d_0 = 1] Pr(d_1 - d_0 = 1) - E[(y_1 - y_0)d_1 - d_0 = -1] Pr(d_1 - d_0 = -1) = E[(y_1 - y_0)(d_1 - d_0 = 1)] Pr(d_1 - d_0 = 1) \text{ (because \( Pr(d_1 - d_0 = -1) = 0 \) from no defiers)}
\]
Hence, we have

\[
E(y|z = 1) - E(y|z = 0) = E[(y_1 - y_0)|d_1 - d_0 = 1] \Pr(d_1 - d_0 = 1)
\]

\[
\Rightarrow E[(y_1 - y_0)|d_1 - d_0 = 1] = \frac{E(y|z = 1) - E(y|z = 0)}{\Pr(d_1 - d_0 = 1)}
\]

\[
\Rightarrow E(y_1 - y_0|d_1 = 1, d_0 = 0) = \frac{E(y|z = 1) - E(y|z = 0)}{\Pr(d_1 = 1, d_0 = 0)}.
\]

The left-hand side is the Local Average Treatment Effect (LATE). It is the average treatment effect for those “compliers” with \(d_1 = 1\) and \(d_0 = 0\). The “local” qualifier in LATE refers to the fact that LATE is specific to the instrument being used. In the expression above, it is not clear whether LATE is identified, because the right-hand side denominator involves both \(d_1\) and \(d_0\) which are not simultaneously observable.

To see why LATE actually can be identified, note that

\[
\Pr(d_1 = 1) = \frac{\Pr(d_1 = 1, d_0 = 1) + \Pr(d_1 = 1, d_0 = 0)}{\text{always-taker} + \text{complier}};
\]

\[
\Pr(d_0 = 1) = \frac{\Pr(d_1 = 1, d_0 = 1) + \Pr(d_1 = 0, d_0 = 1)}{\text{always-taker} + \text{defier}}.
\]

Taking the difference, we will get

\[
\Pr(d_1 = 1) - \Pr(d_0 = 1) = \Pr(d_1 = 1, d_0 = 0),
\]

because \(\Pr(d_1 = 0, d_0 = 1) = 0\) under the monotonicity assumption (no “defiers”). Then

\[
E(d|z = 1) - E(d|z = 0) = \frac{E(d_1|z = 1) - E(d_0|z = 0)}{\Pr(d_1 = 1|z = 1) - \Pr(d_0 = 1|z = 0)}
\]

\[
= \Pr(d_1 = 1) - \Pr(d_0 = 1) \quad \text{(because of condition b)}
\]

\[
= \Pr(d_1 = 1, d_0 = 0) \quad \text{(because of condition c)}.
\]

In words,

\[
E(d|z = 1) = \Pr(d = 1|z = 1) = \Pr(\text{always-taker or complier})
\]

\[
E(d|z = 0) = \Pr(d = 1|z = 0) = \Pr(\text{always-taker or defier})
\]

\[
\Rightarrow E(d|z = 1) - E(d|z = 0)
\]

\[
= \Pr(\text{always-taker or complier}) - \Pr(\text{always-taker or defier})
\]

\[
= \Pr(\text{complier}) - \Pr(\text{defier})
\]

\[
= \Pr(\text{complier}) \quad \text{(defiers ruled out by the monotonicity assumption)}
\]

\[
= \Pr(d_1 = 1, d_0 = 0).
\]

Thus, LATE can be represented by an identifiable form:

\[
E(y_1 - y_0|d_1 = 1, d_0 = 0) = \frac{E(y|z = 1) - E(y|z = 0)}{E(d|z = 1) - E(d|z = 0)},
\]

which coincides with the Wald estimator. This is the LATE interpretation of an IV estimator given by Angrist, Imbens, and Rubin (1996) under the three conditions (a), (b) and (c) aforementioned. It is important to keep in mind that LATE can change with different instruments being used. Thus, if IV estimates differ
when using different instruments, this can be an indication of heterogeneous treatment effects because of different compliers associated with the instruments. Also in principle covariates $\mathbf{x}$ can be allowed for by conditioning on $\mathbf{x}$ for LATE, that is

$$E(y_1 - y_0 | \mathbf{x}, \text{complier}) = \frac{E(y|\mathbf{x}, z = 1) - E(y|\mathbf{x}, z = 0)}{E(d|\mathbf{x}, z = 1) - E(d|\mathbf{x}, z = 0)}.$$ 

References
