Let \((Z_i, X_i, \delta_i)\) be IID triplets for \(i = 1, \ldots, n\); each random variable has a fourth moment, 
\[E(\delta_i) = 0, \quad \text{and} \quad E(\delta_i^2) = \sigma^2 > 0.\]
We assume \(Z_i \perp \delta_i\). Let 
\[a = E(X_i Z_i) > 0 \quad \text{and} \quad b = E(X_i \delta_i).\]
To simplify the notation, take 
\[E(Z_i^2) = 1.\]
Let 
\[Y_i = \beta X_i + \delta_i\]
Here, \(a, b, \beta, \sigma^2\) are parameters. We wish to estimate \(\beta\). In this model, \(X_i\) is endogeneous if \(b > 0\).
On the other hand, we can instrument \(X_i\) by \(Z_i\), because \(Z_i \perp \delta_i\) and \(a > 0\).

The object here is to show that the IVLS estimator differs from \(\beta\) by a random error of order \(1/\sqrt{n}\), with asymptotic bias of order \(1/n\). Based on a sample of size \(n\), the IVLS estimator is

\[
\hat{\beta}_n = \left( \sum_{i=1}^{n} Z_i Y_i \right) / \left( \sum_{i=1}^{n} Z_i X_i \right) = \beta + \eta_n
\]

where

\[
\eta_n = N_n / D_n, \quad N_n = \sum_{i=1}^{n} Z_i \delta_i, \quad D_n = \sum_{i=1}^{n} Z_i X_i
\]

Let

\[
\zeta_n = N_n / \sqrt{n}
\]

By the central limit theorem, \(\zeta_n \to N(0, \sigma^2)\): the \(Z_i \delta_i\) are IID with mean 0, and the variance is \(\sigma^2\) because 
\[E(Z_i^2 \delta_i^2) = E(Z_i^2) E(\delta_i^2) = 1.\]
Next, \(E(Z_i X_i) = a\). So \(D_n = na(1 + \xi_n)\), where

\[
\xi_n = \frac{1}{n} \sum_{i=1}^{n} (a^{-1} Z_i X_i - 1)
\]

is of order \(1/\sqrt{n}\) by the central limit theorem. Thus

\[
\hat{\beta}_n - \beta = \eta_n = \frac{\sqrt{n} \zeta_n}{na} \frac{\zeta_n}{1 + \xi_n}
\]

and—the next being a little informal—

\[
\eta_n \approx \frac{\zeta_n - \xi_n \zeta_n}{a \sqrt{n}}
\]

The step from (5) to (6) is “the delta-method,” i.e., a one-term Taylor expansion of \(1/(1 + \xi_n)\). A more rigorous argument will be given, below. We conclude that \(\beta_n - \beta\) is asymptotically normal, with mean 0 and an SE of \(1/(a \sqrt{n})\). However, there is asymptotic bias of order \(1/n\), because
\[ a^{-1}n^{-1/2}E\{\zeta_n\xi_n\} = a^{-1}n^{-1/2}E\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i \delta_i) \frac{1}{n} \sum_{i=1}^{n} \left( a^{-1}Z_iX_i - 1 \right) \right\} \]

\[ = a^{-1}n^{-2}E\left\{ \sum_{i=1}^{n} (Z_i \delta_i) \sum_{i=1}^{n} (a^{-1}Z_iX_i - 1) \right\} \]

\[ = a^{-2}n^{-2}E\left\{ \sum_{i=1}^{n} Z_i^2X_i\delta_i \right\} \]

\[ = a^{-2}n^{-1}E\{Z_i^2X_i\delta_i\} \]

For the third equality, expand the product of the two sums as a double sum

\[ \sum_{ij} (Z_i\delta_i)(a^{-1}Z_jX_j - 1) \]

When \( i \neq j \), factors are independent, and products have expectation zero because \( E(Z_i \delta_i) = E(Z_i)E(\delta_i) = 0 \). Similarly, \( E(Z_i \delta_i) = 0 \) when \( i = j \). The only terms with (possibly) non-zero expectation are \( a^{-1}Z_i^2X_i\delta_i \).

We continue with previous assumptions and notation, but give a formal theorem and proof.

**Theorem.** \( \hat{\beta}_n - \beta = \frac{\zeta_n}{a\sqrt{n}} - \Delta_n/(an) \), where \( \zeta_n \) converges in distribution to \( N(0, \sigma^2) \), and \( \Delta_n \) converges in distribution to a random variable with expectation \( k/a \), where \( k = E(Z_i^2X_i\delta_i) \) may be positive, negative, or zero.

**Proof.** Keep in mind that \( \zeta_n \) and \( \sqrt{n}\xi_n \) are asymptotically normal, with expectation 0: the notation is therefore a little misleading. Start the argument from (5) above: \( 1/(1+x) = 1 - 1/(1+x) \) unless \( x = -1 \), so

\[ a\sqrt{n}(\hat{\beta}_n - \beta) = \frac{\zeta_n}{1 + \xi_n} = \frac{\zeta_n\xi_n}{1 + \xi_n} = \zeta_n - \Delta_n \]

where

\[ \Delta_n = \frac{\zeta_n\sqrt{n}\xi_n}{1 + \xi_n} \quad (7) \]

The pairs

\[ Z_i\delta_i, \ a^{-1}Z_iX_i - 1 \]

are IID, with expectation 0 and covariance matrix

\[ K = \begin{pmatrix} \sigma^2 & k/a \\ k/a & a^{-2}E(Z_i^2X_i^2) - 1 \end{pmatrix} \quad (8) \]

The central limit theorem shows that \( (\zeta_n, \sqrt{n}\xi_n) \) converges in distribution to bivariate normal, with expectation 0 and covariance matrix \( K \). In the denominator of (7), \( \xi_n \to 0 \), so \( \Delta_n \) has the same limiting behavior as \( \zeta_n\sqrt{n}\xi_n \). QED
Remarks

(i) If $E(Z_iX_i) < 0$, replace $Z_i$ by $-Z_i$ or $a$ by $|a|$; if $E(Z_iX_i) \neq 0$, then $E(Z_i^2) > 0$ and $E(X_i^2) > 0$.

(ii) The source of the bias in IVLS is randomness in $\xi_n$, coupled with the correlation between $\xi_n$ and $\zeta_n$—that is, randomness in $\sum Z_iX_i$, coupled with the correlation between $\sum Z_iX_i$ and $\sum Z_i\delta_i$.

When $n$ is large, $\xi_n \approx 0$—the law of large numbers—and the bias is negligible. The correlation traces back to the endogeneity of $X_i$, i.e., the correlation between $X_i$ and $\delta_i$. If, e.g., $(X_i, Z_i) \perp \delta_i$, it is straightforward to show that $E(\hat{\beta}_n|X, Z) = \beta$. Then $k = 0$ in (8).

(iii) Equations (1–2) and the strong law of large numbers show that $\hat{\beta}_n \rightarrow \beta$ almost surely.

(iv) What about estimating $\sigma^2$? In our setup, if $e = Y - X\hat{\beta}_n$ is the vector of residuals, then $e_i - \epsilon_i = X_i(\hat{\beta}_n - \beta)$ so $\|e - \epsilon\|^2 = \sum_i X_i^2(\hat{\beta}_n - \beta)^2$ and $\|e - \epsilon\|^2/n \rightarrow 0$ almost surely.

(v) The usual presentation of IVLS conditions on $Z$. Then $\hat{\beta}_{IVLS} - \beta = \sum_i^n Z_i\delta_i/\sum_i^n Z_iX_i$; conditionally, the numerator is essentially normal with mean 0 and variance $\sum_i^n Z_i^2 = nE(Z_i^2)$. The denominator is essentially $\sum_i^n Z_iE(X_i|Z_i) = nE[Z_iE(X_i|Z_i)] = nE(Z_iX_i) = na$. With some more effort, the theorem can be extended to describe the limiting conditional behavior of $(\xi_n, \sqrt{n}\xi_n)$, given $Z_1, \ldots, Z_n$. In a little more detail, let $\phi(Z_i) = a^{-1}Z_iE(X_i|Z_i) - 1$, so $E(\sqrt{n}\xi_n|Z_1, \ldots, Z_n) = n^{-1/2}\sum_i^n \phi(Z_i)$. The $\phi(Z_i)$ are IID, and $E(\phi(Z_i)) = a^{-1}E(Z_iX_i) - 1 = 0$ by the definition of $a$. Moreover, 0 $\leq \operatorname{var}(\phi(Z_i)) < \infty$. If the variance is positive, the central limit theorem applies and $E(\sqrt{n}\xi_n|Z_1, \ldots, Z_n)$ converges in distribution. In any event, we can center, considering the conditional joint distribution of $\xi_n, \sqrt{n}(\xi_n - E(\xi_n|Z_1, \ldots, Z_n))$ given $Z_1, \ldots, Z_n$. Apparently, this conditional distribution converges weak-star, along almost all sample sequences of $Z_1, Z_2, \ldots$. For example, $\xi_n$ is $n^{-1/2}\sum_i^n Z_i\delta_i$, where the $\delta_i$ are IID with mean 0, and—conditionally—the $Z_i$ are (almost surely) a well-behaved sequence of constants:

$$\frac{1}{n}\sum_{i=1}^n Z_i^2 \rightarrow 1, \quad \frac{1}{n}\sum_i \{Z_i^2 : 1 \leq i \leq n \& |Z_i| > L\} \rightarrow 0 \text{ as } n \rightarrow \infty \text{ and then } L \rightarrow \infty$$

As in many other such situations, when $n$ is large, there would seem to be little difference between conditional and unconditional inference.

(vi) Suppose $(Z_i, \delta_i, \epsilon_i)$ are independent, with expectation 0, variance 1, and fourth moments. We can set $X_i = aZ_i + b\delta_i + c\epsilon_i$. Then $\operatorname{cov}(Z_i, X_i) = a$, because $E(Z_i^2) = 1$; and $\operatorname{cov}(X_i, \delta_i) = b$, because $E(\delta_i^2) = 1$. The $k$ in the theorem is $k = E(Z_i^2X_i\delta_i) = b$, because

$$Z_i^2X_i\delta_i = aZ_i^3\delta_i + bZ_i^2\delta_i^2 + cZ_i^2\epsilon_i,$$

while $E(Z_i^3\delta_i) = E(Z_i^3)E(\delta_i) = 0$, $E(Z_i^2\delta_i^2) = E(Z_i^2)E(\delta_i^2) = 1$, $E(Z_i^2\epsilon_i) = E(Z_i^2)E(\epsilon_i) = 0$.

(vii) With, say, two instruments and one endogenous variable, the proof of consistency and asymptotic normality is about the same. However, evaluating the small-sample bias is trickier. For instance, expansions like (6) can be done in the matrix domain.