(1) **Weak Instruments Generated Data Experiment**

Consider the following three equation system:

\[
\begin{align*}
y_i &= x_i \beta + u_i \\
x_i &= \lambda u_i + \epsilon_i \\
z_i &= \gamma \epsilon_i + \nu_i, \quad i = 1, 2, \ldots, n,
\end{align*}
\]

where \(u_i, \epsilon_i\) and \(\nu_i\) are all independent draws from a \(N(0, 1)\) distribution. Finally, \(x_i\) and \(z_i\) are scalars.

(1a) If this model is the true data generating process, why must we regard \(x\) as an endogenous variable?

(1b) Show that

\[
\text{plim}(\hat{\beta}_{OLS}) = \beta + \frac{\lambda}{\lambda^2 + 1}.
\]

(1c) Show that

\[
\text{Corr}^2(x, z) = \frac{\gamma^2}{(\lambda^2 + 1)(\gamma^2 + 1)}.
\]

(1d) Under what (if any) conditions, in addition to those already given, will \(z\) be a valid instrument?

(1e) In class, we showed that the asymptotic distribution of the IV estimator is given as

\[
\sqrt{n}(\hat{\beta}_{IV} - \beta) \to N \left[0, E(z'x)^{-1}E(z'z)E(x'z)^{-1} \right].
\]

Given the three equation system above, work out the expression for this variance as a function of the parameters of our model.

(1f) Now let us consider constructing a generated data experiment using the above three equation system as the experimental design. In particular, we intend to construct this generated data experiment to illustrate how the asymptotic distribution may be a poor approximation to the (finite sample) sampling distribution, even when \(n\) is quite large, when the instruments are weak.
First, set the parameters equal to the following values: $\beta = 2$, $\lambda = 2$, $\gamma = .05$. (Note that this corresponds to the case of a weak instrument as $z$ is weakly correlated with $x$).

Now, let’s formally compare the sampling distribution to the asymptotic result in (1e). Do this as follows:

1. Set $n = 25,000$. (Note that this is a large sample!)

2. Start an outer loop of 2,500 iterations. We will calculate the IV estimate 2,500 different times to approximate the sampling distribution with $n = 25,000$.
   - Generate $n = 25,000$ observations for $y$, $x$ and $z$ according to the model and parameters defined above.
   - Calculate the IV estimate for the given sample of data. Store this estimate.
   - Repeat this process 2,500 times within the loop.

3. End the loop.

4. Plot the density of the collection of 2,500 IV estimates to approximate the sampling distribution with $n = 25,000$. I am giving you a MATLAB file called “epanech2” that does this for you. To use the m-file, suppose that your 2,500 IV estimates are saved in a vector you have called “betaiv” Then, to use the epanech2 program, just type:

   \[
   \text{[dom ran] = epanech2(betaiv)}
   \]
   \[
   \text{plot(dom,ran)}
   \]

   to plot the density.

5. On the same graph, plot the asymptotic normal distribution you obtained in (1e). Note: Because the sampling distribution may produce very large values on occasion, the default axis for the plot may be quite large. I recommend restricting the $x$ axis to the interval $(1, 3)$.

Do the two densities seem different? In particular, does the finite-sample distribution seem to have a skew? What do these results suggest about the accuracy of the normal approximation to the sampling distribution if the instruments are weak? (Though you are not required to do so, you might try the same comparison with strong and even weaker instruments).
(2) Consider the SEM model:

\[ y_1 = \gamma_1 y_2 + \beta_1 x_1 + \epsilon_1 \]
\[ y_2 = \gamma_2 y_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_2 \]

The errors are assumed independent of the \( x \)'s. A sample of 25 observations produces the following matrix of sums of squares and cross-products:

\[
\begin{array}{c|cccc}
    & y_1 & y_2 & x_1 & x_2 & x_3 \\
\hline
y_1 & 10 & 6 & 4 & 2 & 1 \\
y_2 & 6 & 10 & 3 & 6 & 7 \\
x_1 & 4 & 3 & 5 & 2 & 3 \\
x_2 & 2 & 6 & 2 & 10 & 8 \\
x_3 & 1 & 7 & 3 & 8 & 15 \\
\end{array}
\]

(2a) Estimate the parameters of the first equation by OLS.

(2b) Obtain estimates of \( \gamma_1 \) and \( \beta_1 \) using only \( x_2 \) as an instrument for \( y_2 \). Is this a reasonable instrument?

(2c) Estimate the parameters of the first equation (\( \gamma_1 \) and \( \beta_1 \)) by 2SLS.