Lecture 24

11-17-2021

Announcements

There is no section on 11/24

There is no class on 11/29

WP10 and Checkpoint 13 are due on Friday

PS5 will be available on Thursday

Instrumental Variables

Imagine we have constant effects.

Our true model of the world is $Y_i = \alpha + \delta D_i + \beta X_i + \epsilon_i$

Our observed model of the world is $Y_i = \alpha + \delta D_i + \gamma_i$

What is the bias?

Instrumental Variables

What is the bias?

Our estimate $\hat{\delta} = \frac{E[YD]-E[Y]E[D]}{V[Y]}$

The true value

$$\hat{\delta} = \frac{E[\alpha D + D^2\delta + \beta XD + \epsilon D] - E[\alpha + \delta D + \beta X + \epsilon]E[D]}{V[Y]}$$

$$\hat{\delta} = \delta + \beta\frac{C(X,D)}{V(D)}$$

Instrumental Variables

Suppose we have an instrument Z that is independent of X and the error term

The covariance of Y and Z

$$
\begin{aligned}
C(Y,Z) &= C(\alpha + \delta D + \beta X + \epsilon, Z) \\
C(Y,Z) &= \delta C(D,Z) + \beta C(X,Z) + C(\epsilon, Z)
\end{aligned}
$$

What is the best estimate of δ?

Two Stage Least Squares

Consider a sample with Y, D, and Z. Assume the DGP

$$Y_i = \alpha + \delta D_i + \epsilon_i$$

$$D_i = \gamma + \beta Z_i + \epsilon_i$$

We can rewrite our IV estimate:

$$\frac{C(Y,Z)}{C(D,Z)} = \frac{\frac{1}{n}\sum_i^n (Z_i - \bar{Z})Y_i}{\sum_i^n (Z_i - \bar{Z})D_i}$$

Two Stage Least Squares

When we substitute in appropriately, we sub in the fitted values of the endogenous predictors from the first stage regression.

Recall that $D = \gamma + \beta Z + \epsilon$.

That means that our estimate is $\hat{D} = \hat{\gamma} + \hat{\beta} Z$

Then our 2SLS estimator becomes: $\frac{C(\hat{\beta}Z, Y)}{V(\hat{\beta}Z)}$

Thus: The 2SLS estimator uses only the fitted values based on all variables in the model including the instrument. Because the instrument is exogenous this makes the fitted values themselves exogenous!

Instrumental Variables: Bias and Consistency

2SLS is a ratio estimator (specifically a Wald Estimator). Ratio estimators are biased.

In general for a ratio $E[\frac{A}{B}] = \frac{E[A]-C(\frac{A}{B},B)}{E[B]}$. The Bias becomes negligible as our sample size increases

The ratio estimator is a consistent estimator. As sample size grows the estimated ratio converges to the true value of the LATE.

Weak Instruments

Our IV 2SLS estimate of the treatment effect with a constant is $\frac{C(Y,Z)}{C(D,Z)}$.

This function is clearly undefined when C(D,Z) = 0. What happens when it is very small?

Small perturbations in the denominator will lead to massive swings in our coefficient

The bias can be quantified as $\frac{C(\epsilon,D)}{V(D)} \frac{1}{F+1}$

The Problems with IVs in modern practice

Lal *et. al* (2021) suggest there are three major problems with IVs in political science:

1. Many studies do not report the appropriate first stage F statistic
2. Many studies underestimate the uncertainty of their estimates
3. 2SLS estimates are often much bigger in magnitude suggesting failures of exclusion restrictions

What should we do in research?

Lal *et. al* suggest:

1. Think about the design hard before beginning
2. After running the first stage regression, plot our actual treatment against our first stage predictions and check the IV strength by eye
3. Consider the expectation of the OLS bias. If OLS is expected to have upward bias be considered if the 2SLS are even bigger than the OLS ones.

What does `iv_robust` do?

Explicit notes are here

In short:

1. `iv_robust()` estimates a two stage least squares regression.
2. Variance estimation is the same as `lm_robust()` except the pacakges replace our endogenous estimator with the second stage regressors, $\hat{D}$, and we replace the residuals are from the final coefficients and the endogenous, uninstrumented regressor.