1 Assumptions for Instrumental Variable (IV) Models

1.1 IV Model

Consider the following model:

$$Y = \beta_0 + \beta_1 \tilde{X}_1 + ... + \beta_k \tilde{X}_k + \beta_{k+1} \tilde{W}_1 + ... + \beta_{k+r} \tilde{W}_r + U$$

- Y is the outcome variable
- $\left(\tilde{X}_1, ..., \tilde{X}_k\right)$ are the **endogenous** regressors (potentially correlated with U)
- $\left(\tilde{W}_1, ...\tilde{W}_r\right)$ are the included **exogenous** regressors (uncorrelated with U)
- $\left(\tilde{Z}_1, ...\tilde{Z}_m\right)$ are the excluded exogenous regressors (uncorrelated with U) or instruments
- $(\beta_1, ..., \beta_{k+r})$ are the parameters of interest

Define $(r + m + 1) \times 1$ dimensional vector $\mathbf{Z} = \left(1, , \tilde{Z}_1, ..., \tilde{Z}_m, \tilde{W}_1, ..., \tilde{W}_r\right)'$ This vector collects all the exogenous variables (whether in the regression or not). Define $(r + k + 1) \times 1$ dimensional vector $\mathbf{X} = \left(1, \tilde{X}_1, ..., \tilde{X}_k, \tilde{W}_1, ..., \tilde{W}_r\right)'$ collecting all the included regressors (whether exogenous or not).

1.2 Assumptions and Properties

1. Assumption 1: Error term U is mean independent of the exogenous regressors $\mathbf{Z}$ $\mathbb{E}(U|\mathbf{Z}) = 0$. Recall, this implies $\mathbb{E}(\mathbf{Z}U) = 0$ and since $\mathbf{Z}$ includes a constant, this means that the exogenous regressors are uncorrelated with U.
2. Assumption 2: We observe an i.i.d. sample

$$\left\{Y_i, \tilde{X}_{1i}, ..., \tilde{X}_{ki}, \tilde{W}_{1i}, ...\tilde{W}_{ri}, \tilde{Z}_{1i}, ...\tilde{Z}_{mi}\right\}_{i=1}^{n}$$

3. Assumption 3: The $(r + m + 1) \times (r + k + 1)$ dimensional matrix $\mathbb{E}\left(\mathbf{Z}\mathbf{X}'\right)$ has rank $(r + k + 1)$. The square matrix $\mathbb{E}\left(\mathbf{Z}\mathbf{Z}'\right)$ has rank $(r + m + 1)$. This assumption requires that $k \leq m$, which means that "we need as many instruments as endogenous regressors". When there is only one endogenous variable, if $m = 1$ and $r = 0$, this assumption requires $Cov(X, Z) \neq 0$, if $m > 1$, this assumption requires that at least one of the coefficients on the instruments from the regression of X on $\mathbf{Z}$ not equal to zero. When $k = m$, the model is **just identified**. When $k < m$, the model is **overidentified** (can do some instrument checking). When $k > m$ the model is not identified (need more instruments).
4. Assumption 4: The matrix $\mathbf{S} \equiv \mathbb{E}\left(\mathbf{Z}U^2\mathbf{Z}'\right)$ is strictly positive definite.
5. Assumption 5: The elements of $\left\{Y, \tilde{X}_1, ..., \tilde{X}_k, \tilde{W}_1, ...\tilde{W}_r, \tilde{Z}_1, ...\tilde{Z}_m\right\}$ all have finite fourth moments (large outliers are therefore unlikely).

As before, from Assumption $A1$ through $A4$, we get that $\hat{\beta}$ is consistent for β, and in large samples $\hat{\beta}$ is approximately normally distributed with mean β and population variance matrix $V/\sqrt{n}$. With $A1$ through $A5$, we also get an expression we can estimate for the asymptotic variance of the IV estimators $Avar(\hat{\beta})$.

1.3 Exercise 1 (Stock and Watson Exercise 12.9)

A researcher is interested in the effect of military service on human capital. He collects data from a random sample of 4000 workers aged 40 and runs the OLS regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, where Y_i is a worker's annual earnings and X_i is a binary variable that is equal to 1 if the person served in the military and is equal to 0 otherwise.

1. Explain why the OLS estimates are likely to be unreliable. (Hint: Which variables are omitted from the regression? Are they correlated with military service?)

2. During the Vietnam War, there was a draft in which priority for the draft was determined by a national lottery. (The days of the year were randomly reordered 1 through 365. Those with birth dates ordered first were drafted before those with birth dates ordered second, and so forth.) Explain how the lottery might be used as an instrument to estimate the effect of military service on earnings. (For more about this issue, see Joshua D. Angrist (1990).)

2 Two Stage Least Squares (2SLS)

What if we have multiple candidate IVs for the same endogenous variable? We use a process called Two Stage Least Squares: **2SLS**. Essentially, we use multiple IVs to construct a single, stronger IV to use in our estimation.

Our structural model, as before:

$$Y = \beta_0 + \beta_1 X + \beta_2 W_1 + \epsilon$$

but we have two exogenous variables Z_1 and Z_2 that 1) are both correlated with X, and 2) satisfy the exclusion restrictions (they do not appear in the structural model and are uncorrelated with the error ϵ). So both are candidates as IVs for endogenous X. Thus

$$X = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 W_1 + \nu$$

Here the best IV for X will make use of all exogenous variables. The intuition for this is that including more valid instruments helps explain more of the exogenous variation in X (variation in X that is not correlated with ϵ) and thus generate more precise estimates.

For IV to meet the relevance restriction, one needs at least $\gamma_1 \neq 0$ or $\gamma_2 \neq 0$, which one can test via an F-test. The larger our F-statistic, the stronger our instruments (and the less concern we have about potential bias from weak instruments).

Using these IVs, we can isolate the part of X that is not correlated with ϵ by construction, since none of the Z variables are correlated with ϵ:

$$\hat{X} = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 W_1$$

We can then use this exogenous part of x to identify the causal relationship between x and y.

The procedure to carry out 2SLS is as follows:

- Regress X on Z_1, Z_2, and W_1 and obtain the fitted values $\hat{X}$: these will be uncorrelated with ϵ. This is called the **first stage**.
- Verify that Z_1 and Z_2 are jointly significant in the **reduced form** (the regression of Y on all exogenous variables) with an F-test.
- Then to estimate the **structural equation** use $\hat{X}$ as IV for X, for the regression of Y on $\hat{X}$ and W_1,

Note: you will want to use a command in Python or Stata to do this automatically, since standard errors and test statistics obtained in this way are not valid. You can actually run a 2SLS model using the

IV2SLS command in Python or ivreg2 command in Stata, but you won't need to know how to do that for this course. In practice, we'll often use 2SLS to generate IV estimates even when we only have one IV.

3 Weak Instrument

3.1 Definition

Weak instruments, which are instruments that are only marginally valid, can cause many problems. Consider the simplest case

$$Y = \beta_0 + \beta_1 X + u$$

and $\mathbb{E}\left(u(1, Z)'\right) = 0$ so that

$$\beta_1 = \frac{Cov\left(Y, Z\right)}{Cov\left(X, Z\right)}$$

If $Cov(X, Z)$ is close to zero, we call Z a weak instrument. If $Cov(X, Z)$ is small, small changes in s_{xz} (from one sample to the next) can induce big changes in $\hat{\beta}_1$. If $Cov\left(X, Z\right)$ is close to (or equal to) zero, large n distribution of $\hat{\beta}_1$ no longer normal. If instruments are weak, the usual methods of inference are unreliable – potentially very unreliable.

3.2 Checking for weak instruments with a single X

1. Compute the first-stage F-statistic (Recall, F=Wald/#restrictions)
2. *Rule-of-thumb*: If the first stage F-statistic is less than 10, then the set of instruments is weak.
3. If so, the 2SLS estimator will be biased, and statistical inferences (standard errors, hypothesis tests, confidence intervals) can be misleading.
4. Note that simply rejecting the null hypothesis of that the coefficients on the Z's are zero isn't enough – you actually need substantial predictive content for the normal approximation to be a good one.
5. There are more sophisticated things to do than just compare F to 10 but they are beyond this course.

3.3 Exercise 2

Suppose you wish to measure the impact of smoking on the weight of newborns. You are planning to use the following model,

$$log(bw_i) = \beta_0 + \beta_1 male_i + \beta_2 order_i + \beta_3 log(income_i) + \beta_4 cig_i + \epsilon_i$$

where *bw* is the birth weight, *male* is a dummy variable assuming the value 1 if the baby is a boy or 0 otherwise, *order* is the birth order of the child, $log(income)$ is the log income of the family, *cig* is the amount of cigarettes per day smoked during pregnancy, i indexes the observation and the β's are the unknown parameters.

1. What could be the problem in using OLS to estimate the above model?

2. Suppose you have data on the average price of cigarettes in the state of residence. Would this information help to identify the true parameters of the model?

3. Suppose you use average price of cigarettes, $cigprice$ and cigarette tax rate, $cigtax$, as instruments in the estimation of the above model. What are the conditions under which the parameters of the model above are identified?

4. How would you test whether the instruments are relevant? Are they strong instruments?

4 Instrument Validity: Testing Exogeneity

4.1 The (Wald) J-test of overidentifying restrictions

Assumption A1 implied that the instruments are uncorrelated with the error term. If the instruments are correlated with the error term, the first stage of TSLS doesn't successfully isolate a component of X that is uncorrelated with the error term, so is correlated with U and TSLS is inconsistent. If there are more instruments than endogenous regressors, it is possible to test – partially – for instrument exogeneity.

Again, consider the simplest case:

$$Y = \beta_0 + \beta_1 X + U$$

Suppose there are two valid instruments: Z_1 and Z_2. Then you could compute two separate TSLS estimates. Intuitively, if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid. The **(Wald) J-test of overidentifying restrictions** makes this comparison in a statistically precise way. This can only be done if $m > k$ (overidentified).

If we make the additional assumption that $\mathbb{E}\left(U^2|\mathbf{Z}\right) = \sigma^2$, we can use the following test construction.

1. First estimate the equation of interest using TSLS and all m instruments; compute the predicted values , using the actual X's (not the ones used to estimate the second stage)
2. Compute the residuals $Y_i - \hat{Y}_i$
3. Regress against $\tilde{Z}_1, ..., \tilde{Z}_m, \tilde{W}_1, ..\tilde{W}_r$
4. Compute the Wald statistic testing the hypothesis that the coefficients on $\tilde{Z}_1, ..., \tilde{Z}_m$ are all zero.

5. The Wald statistic for this test is known as the J-statistic

Under the null hypothesis that **all the instruments are exogenous** , and conditional homoscedasticity, J has a chi-squared distribution with $m - k$ degrees of freedom. If some instruments are exogenous and others are endogenous, the J statistic will be large, and the null hypothesis that all instruments are exogenous will be rejected.

4.2 Exercise 3 (Stock and Watson Exercise 12.7)

In an IV regression model with one regressor, X_i, and two instruments, Z_{1i} and Z_{2i}, the value of the J-statistic is J = 18.2.

1. Does this suggest that $\mathbb{E}[u_i|Z_{1i}, Z_{2i}] \neq 0$? Explain.

2. Does this suggest that $\mathbb{E}[u_i|Z_{1i}] \neq 0$? Explain.