UCI Paul Merage School of Business

Leadership for a Digitally Driven World™

MFIN 290: Financial Econometrics

Lecture 2-2

But what if it's not true?

- Violations of GM Assumptions

- Textbook references: 2.3-2.9, 5.1-5.5

- Remember:
- Students can use Stata and MATLAB for free through a virtual computer lab service setup by the group that supports computing for the entire campus, OIT. That can be accessed here: https://www.oit.uci.edu/labs/apporto/.

Violations of GM Assumptions

We saw that we used a subset of the GM assumptions to prove different properties of the OLS estimator.

It's important to think this through when reviewing model's test performance. What if we fail normality tests? What about non-sphericality?

What if there is a relationship between the errors (or functions of the errors) and the X's?

Violations of GM Assumptions: Normality

We have the central limit theorem if we are taking averages (usually!)… so clearly some violations of the normality assumptions have different consequences.

For example, sometimes, normality becomes asymptotic, so we know our inference is valid only if we have a lot of data… sometimes the form of variance is different.

There are different remedies for each, and it is likely that these effects have different consequences for your model.

Violations of GM Assumptions: Sphericality

If $E(\epsilon\epsilon'|X) = \mathbf{\Omega} \neq \sigma^2 I$, as we saw last time our unbiasedness proof is unaffected, though our variance of the estimator (and thus our confidence intervals on any hypothesis testing) will change!

- $E\big((b-\beta)(b-\beta)'\big) = \mathrm{E}\big((A\epsilon)(A\epsilon)'|X\big)$
- $= E(A\epsilon\epsilon' A'|X)$
- $= AE(\epsilon\epsilon')\mathrm{A}' = \mathrm{A}\mathbf{\Omega}A' \ldots \neq \sigma^2 AA'$
- $= (X'X)^{-1}X'\mathbf{\Omega}X(X'X)^{-1}$

- This is used in the derivation of BLUE and in the construction of confidence intervals => our inference may be incorrect and there may be other, lower variance linear unbiased estimators!

Violation of Rank: Comments on Multicollinearity

- You will often hear discussion and concerns about multicollinearity, and outliers when you build models.
- Why? "High" correlation is not any of our Gauss Markov assumptions?

- Note: Perfectly correlated X's are very rare (that's likely a failure of the *model or modeler)*, and would cause $X'X$ to not be invertible. It is far more often that there is concern that the X's be not *too* correlated.

- If that's the case, then the inverse of X'X will still exist but may not be very well behaved
- Intuition: dividing by a number very close to zero

Comments on Multicollinearity

- Variance of beta $= \sigma^2 (X'X)^{-1}$, can show that the variance of the k^{th} beta $= \frac{\sigma^2}{(1-R_k^2)\sum(x_{ik}-\overline{x_k})^2}$
- Where R_k^2 is the R^2 from a regression of x_k on all of the other variables. So there are three things going on:
-

1. greater correlations of x_k with other variables (R_k^2 higher), higher the variance will be,
2. greater variation in x_k, the higher the variance will be,
3. better fit/ lower σ^2, lower the variance will be.

Comments on Multicollinearity

- People characterize this with the VIF (variance inflation factor) = $\frac{1}{(1-R_k^2)}$ for each variable.

- Some rules of thumb here, though this is less valuable when some variables are constructed to be correlated…

Comments on Multicollinearity

- People characterize this with the VIF (variance inflation factor) $= \frac{1}{(1-R_k^2)}$ for each variable.

- Some rules of thumb here, though this is less valuable when some variables are constructed to be correlated…

- Perhaps the effect of a certain variable varies by some categorical information?
- We think that energy stocks are more exposed to Oil price movements? Then we create a variable $Energy$ equal to one if in the energy industries, 0 otherwise, and another equal to $Energy * OilPrice$.
- $Expect$ this to be correlated with Oil and $Energy$ … would not want to remove these variables if they ought to be in the model.

Comments on Multicollinearity

- Some shrinkage estimators (ridge/L2 regularization) are also an option here, but introduce bias by design.

- In sum: multicollinearity is not a problem from the perspective of the Gauss Markov assumptions. Diagnosing it is trying to separate a bad model from bad data. We OFTEN have bad models *and* bad data in the real world. It is a symptom of a different issue.

- "A finding that suggests multicollinearity is adversely affecting the estimates seems to suggest that but for this effect, all the coefficients would be statistically significant and of the right sign. Of course, this need not be the case…. 'remedies' for multicollinearity might well amount to attempts to force the theory on the data." – Greene, *Econometric Analysis*

Violations of $E(\epsilon|X) = 0$

Treatment Effects

Define the indicator variable d_i as $d_i = \begin{cases} 1 \; if \; in \; treatment \\ 0 \quad\quad if \; in \; control \end{cases}$

The model is then:

$$y_i = \beta_1 + \beta_2 d_i + \epsilon_i, i = 1, \ldots, N$$

And the regression functions are:

$E(y_i) = \beta_1 + \beta_2$ if individual is in treatment group

$E(y_i) = \beta_1$ if individual is in control group

Difference Estimator

Using Least Squares,

$$\widehat{\beta_2} = \frac{\sum_{i=1}^{N}\left(d_i - \overline{d}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{N}\left(d_i - \overline{d}\right)^2} = \overline{y}_1 - \overline{y}_0$$

That is, the coefficient is the difference in the means across groups.

This is called a Difference Estimator

Can rewrite as

$$\beta_2 + \frac{\sum_{i=1}^{N}\left(d_i - \overline{d}\right)\left(e_i - \overline{e}\right)}{\sum_{i=1}^{N}\left(d_i - \overline{d}\right)^2} = \beta_2 + \left(\overline{e}_1 - \overline{e}_0\right)$$

Difference Estimator

For the estimated treatment effect to be unbiased, we need

$$E(\bar{e_1} - \bar{e_0}) = E(\bar{e_1}) - E(\bar{e_0}) = 0$$

If we allow individuals to "self-select" themselves into the treatment and control groups, then $E(\bar{e_1}) - E(\bar{e_0})$ is the selection bias in the estimation of the treatment effect.

Natural Experiments

Randomized controlled experiments are rare in economics because they are expensive and involve human subjects

Natural experiments, also called quasi-experiments, rely on observing real-world conditions that approximate what would happen in a randomized controlled experiment

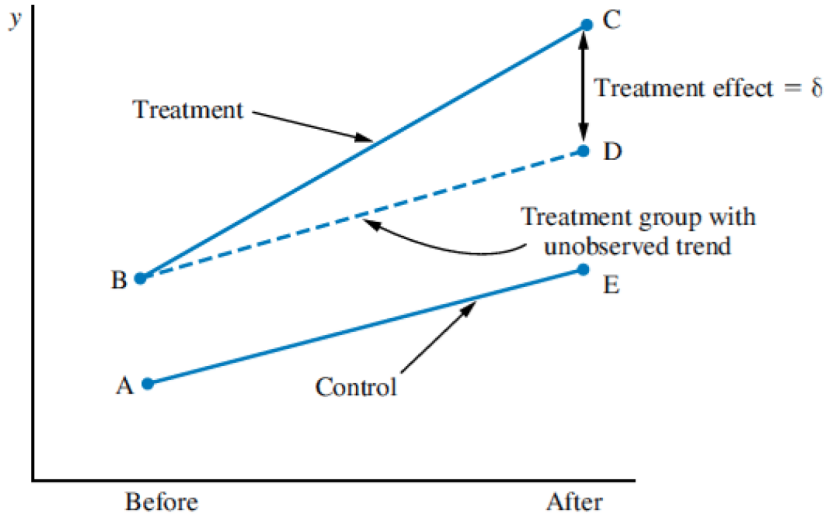
In these cases, treatment appears as if it were randomly assigned

Validity of experiment hinges on how “natural” it is!

Difference in Difference

Imagine we wanted to know the impact of school sizes. We couldn't just look at two schools with different sized courses…

But imagine we were following two schools. Midway through the first grade, School A suddenly cut their class sizes in half (with econometrically convenient identical teachers). We can look for a change in the trend that each class was following…

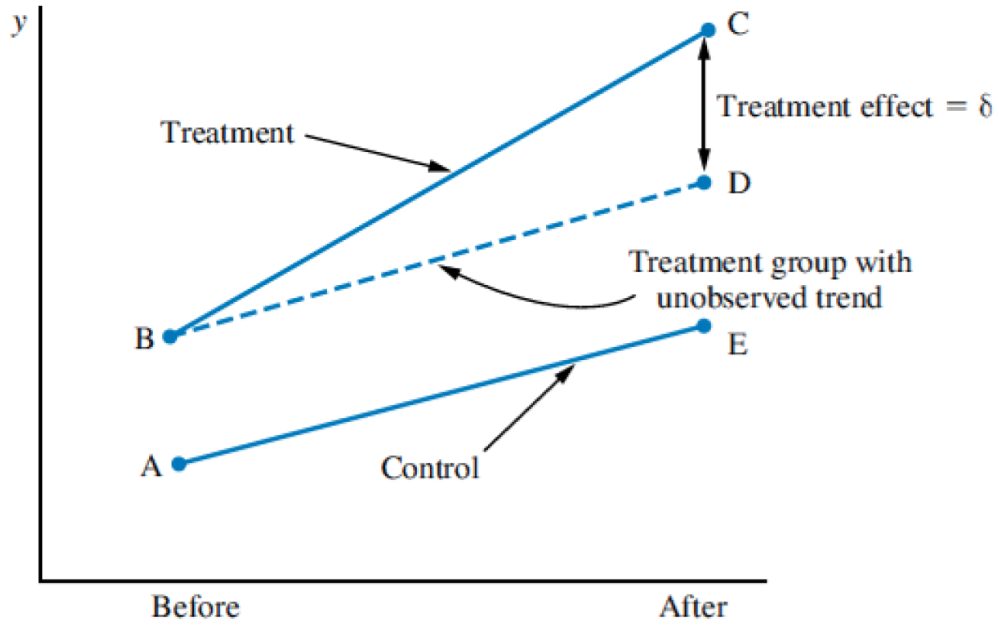


This is called a "difference in difference"

Requires that the groups would have otherwise followed a parallel trend (why?)

Difference in Difference

Estimation of the treatment effect is based on data averages for the two groups in the two periods



$$\left(\hat{C} - \hat{E}\right) - \left(\hat{B} - \hat{A}\right)$$

$$= \left(\bar{y}_{\text{Treatment, After}} - \bar{y}_{\text{Control, After}}\right)$$

$$- \left(\bar{y}_{\text{Treatment, Before}} - \bar{y}_{\text{Control, Before}}\right)$$

Estimating a Difference in Difference

Consider the following model:

We th

$$y_{it} = \beta_1 + \beta_2 TREAT_i + \beta_3 AFTER_t + \delta (TREAT_i \times AFTER_t) + e_{it}$$

$$E(y_{it}) = \begin{cases} \beta_1 & TREAT = 0,\ AFTER = 0 \ (\text{group } A) \\ \beta_1 + \beta_2 & TREAT = 1,\ AFTER = 0 \ (\text{group } B) \\ \beta_1 + \beta_3 & TREAT = 0,\ AFTER = 1 \ (\text{group } E) \\ \beta_1 + \beta_2 + \beta_3 + \delta & TREAT = 1,\ AFTER = 1 \ (\text{group } C) \end{cases}$$

Estimating a Difference in Difference

$$E(y_{it}) = \begin{cases} \beta_1 & TREAT = 0,\ AFTER = 0 \ \text{(group } A\text{)} \\ \beta_1 + \beta_2 & TREAT = 1,\ AFTER = 0 \ \text{(group } B\text{)} \\ \beta_1 + \beta_3 & TREAT = 0,\ AFTER = 1 \ \text{(group } E\text{)} \\ \beta_1 + \beta_2 + \beta_3 + \delta & TREAT = 1,\ AFTER = 1 \ \text{(group } C\text{)} \end{cases}$$

$$\begin{aligned} \delta &= (C - E) - (B - A) \\ &= [(\beta_1 + \beta_2 + \beta_3 + \delta) - (\beta_1 + \beta_3)] \\ &\quad - [(\beta_1 + \beta_2) - \beta_1] \end{aligned}$$

Violations of Exogeneity

If $E(\epsilon|X) \neq 0$ our unbiasedness proof is affected. This has consequences for the other items as well, as the variance derivation relied on $E(\mathrm{b}) = \beta$

In this portion of lecture, we discuss several common examples of endogeneity, as well as a conceptual solution.

Omitted Variables Bias

- Suppose that the "true" model is given by:
- $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$
-
- But instead, we estimate:
- $Y = X_1\widehat{\beta_1} + u$

Omitted Variables Bias

- $\widehat{\beta_1} = (X_1'X_1)^{-1}X_1'Y$
- $\widehat{\beta_1} = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)$

Omitted Variables Bias

- $\widehat{\beta_1} = (X_1'X_1)^{-1}X_1'Y$
- $\widehat{\beta_1} = (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)$
- $\widehat{\beta_1} = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon$
- Bias $= E(\widehat{\beta_1} - \beta_1) = E((X_1'X_1)^{-1}X_1'X_2\beta_2) + E((X_1'X_1)^{-1}X_1'\varepsilon)$
-
- Unless $E(X_1'X_2) = 0$ (they are not correlated at all) or
- $\beta_2 = 0$ (it's not actually omitted!)
- $\widehat{\beta_1}$ will be biased!

Omitted Variables Bias

- More than that, it will be biased in a very particular way:
- $(X_1'X_1)^{-1}X_1'X_2\beta_2$ should look familiar…

- It's the slope(s) of the regression of the omitted variable X_2 on each of the columns of the included X_1.

- Omitted variables happens essentially *all the time*. Sometimes we can't observe something plausibly important, sometimes we are using a noisy approximation, etc.

- Key is being able to think through the issue, plausibly test for important effects.

Omitted Variables Bias

- Suppose now that the "true" model is given by:
- $Y = X_1\beta_1 + \epsilon$

- But instead, we estimate:
- $Y = X_1\widehat{\beta_1} + X_2\widehat{\beta_2} + u$

- **Q: Will $\widehat{\boldsymbol{\beta_1}}$ be biased?**
- No! What's $E(\widehat{\beta_1})$ in this model?
-

Omitted Variables Bias

- **Q: Will $\widehat{\beta_1}$ be biased?**
- No! What's $E(\widehat{\beta_1})$ in this model?

- Why don't we just include everything? Power!
- **Why might we include something (statistically) insignificant**? Omitted variables!

Measurement Error

- Let's go back to single variables for a moment. Suppose the true model is:

- $y = x^*\beta + \varepsilon$
- But we observe $x = x^* + u$
- What do we get if we regress y on x?
- $\hat{\beta} = cov(x, y)/var(x)$

Measurement Error

- Let's go back to single variables for a moment. Suppose the true model is:

- $y = x^*\beta + \varepsilon$
- But we observe $x = x^* + u$
- What do we get if we regress y on x?
- $\hat{\beta} = cov(x, y)/var(x)$

Let's substitute $x^* = x - u$ into $y = x^*\beta + \varepsilon$

Measurement Error

- Let's substitute $x^* = x - u$ and write
- $y = x\beta + \varepsilon - \beta u = x\beta + w$

- We know that our OLS estimate $\hat{\beta}$ will be biased if the regressors covary with the error term.

- $Cov(x, w) = Cov(x^* + u\ , \varepsilon - \beta u) = \ -\beta\sigma_u^2$

- Which is always nonzero if there is any measurement error (and if β is non-zero)

Measurement Error

- So if it's biased, what is the expectation?
- $\frac{cov(x,y)}{var(x)} = \frac{1/n \sum(x^*+u)(x^*\beta+\varepsilon)}{1/n \sum(x^*+u)^2} = \frac{\beta Var(x^*)}{Var(x^*)+Var(u)}$
-
- Where in the second step, we use the fact that $u, x^*, and\ \varepsilon$, are all independent.

- $\frac{\beta Var(x^*)}{Var(x^*)+Var(u)} = \beta \frac{1}{1+\frac{\sigma_u^2}{\sigma_{x^*}^2}} < \beta$
- Called "attenuation bias", "iron law of econometrics".

Multivariate Measurement Error

- $Y = X\beta + W$
- $\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + W)$
- $\hat{\beta} - \beta = (X'X)^{-1}X'W$

- We have:
- $X' = X^{*\prime} + U'$;
- $E(X'X) = X^{*\prime}X^* + U'U$

Multivariate Measurement Error

- $\hat{\beta} - \beta = (X^{*'}X^{*} + U'U)^{-1}[X^{*'}(W - U\beta) + U'(W - U\beta)]$
- $E(X^{*'}W) = E(X^{*'}U) = E(U'W) = 0 \Rightarrow$
- $\hat{\beta} - \beta = -(X^{*'}X^{*} + U'U)^{-1}U'U\beta$

- Since the inverse applies to every estimated coefficient, even when certain columns of X are measured correctly ($X_k^* = X_k$), measurement error in one will bias all of the coefficients, not just the one measured incorrectly

Multivariate Measurement Error

- The easy example of measurement error is errant data: GDP was mismeasured, income was reported with noise, etc.
- However, this understates the issue. Often the true model we have in mind has concepts for the X's that we approximate with data… we want "Education" so we use "Years in School". We want "Economic Strength", so we use "Real GDP" (and/or maybe Unemployment rates). We may have measurement error even when the underlying data is correctly measured!
- Examples:
- We regress income on education.
- What are the omitted variables? Quality of school? Ability?
- "Market" return is NOT S&P 500! "Expected return" is not the realized return!

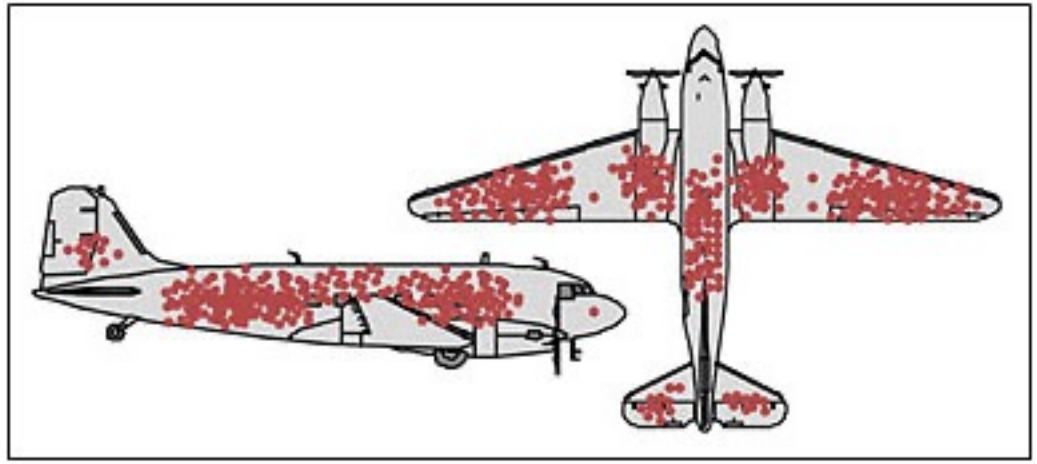
General Endogeneity

Common Example: Selection Bias

Selection

Where to armor planes?

Reviewed returning planes for where they were damaged.



Credit: Cameron Moll

General Endogeneity

Avoid the faulty line of reasoning known as post hoc: One event's preceding another does not necessarily make the first the cause of the second. An alternative way to state this is that "correlation is not the same as causation"

One example of the problem we face with economic data is that the data generally exhibit a selection bias. That is, people chose (or self-select) to do something specifically because they would have favorable outcome. When membership in the treated group is in part determined by choice, the sample cannot be viewed as a random sample.

Identifying and correcting these issues is (to me) what defines an econometrician. The software will always execute what you tell it, but ensuring it is interpretable as intended? That's where we earn our keep

Selection Bias

In general, this endogeneity (selection bias) interferes with a straightforward examination of the data, and makes more difficult our efforts to measure a causal effect.

We would like to randomly assign items to a treatment group, with others being treated as a control group and we could then compare the two groups

Unfortunately (or fortunately depending on your point of view!), the ability to perform randomized controlled experiments in economics is limited because the subjects are people, and their economic well-being is at stake

Selection Bias

What's wrong with each of these:

1. Living by the Beach is good for your health
2. Eating yogurt helps you live longer
3. People who get Flu shots get sick just as often as people who do not
4. People who switch to Insurance Company A save 35% on car insurance

General Endogeneity

This term comes from simultaneous equations models: It means "determined within the system"

Insights surrounding endogeneity are a key contribution from Econometrics just as insights related to equilibrium are a (the?) key contribution of Economics

Imagine we are estimating the model

$$y = x\beta + \epsilon$$

But are concerned that $E(X'\epsilon) \neq 0$

Instrumental Variables

Suppose that there is another variable, z, such that:

- z does not have a direct effect on y, and thus it does not belong on the right-hand side of the model as an explanatory variable
- z is not correlated with the regression error term … variables with this property are said to be "exogenous" variables ($z'\epsilon = 0$)
- z is correlated (enough) with x, the endogenous explanatory variable ($z'x \neq 0$)

Then the variable z is called an "instrumental variable" and we can use this to identify the true effect!

Instrumental Variables

These are not easy to think of!

- What is correlated with your years of education, but not your ability?
- What is correlated with going to the doctor, but not your health status?
- What shifts the demand curve, but not the supply curve (or vice versa?)

Instrumental Variables

If such a variable z exists, then it can be used to form a moment condition:

$$E(z'\epsilon) = 0 \Rightarrow E\left(z'(y - \beta_1 - \beta_2 x)\right) = 0$$

With sample moments:

$$\frac{1}{N}\sum_i \left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i\right) = 0$$

$$\frac{1}{N}\sum_i \left(y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i\right) z_i = 0$$

Introduction to Instrumental Variables

- $y = x^*\beta + \varepsilon$
- But we observe **two** noisy signals of the true value
- $x = x^* + u$
- $z = \phi x^* + v$
- Where the usual independence assumptions apply to $\varepsilon, u \; and \; v$

Introduction to IV

- $cov(z, x) = E(\phi x^* + v)(x^* + u)$
- $cov(z, x) = (\phi \sigma^2_{x^*})$
- $cov(z, y) = E(\phi x^* + v)(x^* \beta + \varepsilon)$
- $cov(z, y) = (\phi \beta \sigma^2_{x^*})$

- $\Rightarrow \frac{cov(z,y)}{cov(z,x)} = \beta$

- Which we can recover directly from our earlier moment condition... Though there are more options…

Two Stage Least Squares

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e$$

Let x_K be an endogenous variable, that is $E\ (x_K{}'\ e) \neq 0$

Assume that the first K − 1 variables are exogenous variables that are uncorrelated with the error term e, that is, $E\ (x_K{}'\ e) = 0$ for $k = 1, \ldots K - 1$

Then, we can estimate this equation in two steps using the least squares estimation procedure in each step

Two Stage Least Squares: First Stage

The first stage regression has the endogenous variable on the left-hand side, and on the right-hand side all the exogenous **and** instrumental variables

$$x_K = \gamma_1 + \gamma_2 x_2 + \cdots \gamma_{K-1} x_{K-1} + \theta_1 z_1 + \cdots + \theta_L z_L + v_K$$

This yields fitted values:

$$\hat{x}_K = \hat{\gamma}_1 + \hat{\gamma}_2 x_2 + \cdots \hat{\gamma}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + \cdots + \hat{\theta}_L z_L$$

Two Stage Least Squares: Second Stage

The second stage regression is based on the original specification:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_K \hat{x}_K + e^*$$

The least squares estimators from this equation are the instrumental variables (IV) estimators. Because they can be obtained by two least squares regressions, they are also commonly called the two-stage least squares (2SLS) estimators

The IV/2SLS estimator of the error variance is based on the residuals from the original model (use x_K, not $\widehat{x_K}$):

$$\hat{\sigma}^2_{IV} = \frac{\sum_i \left(y - \hat{\beta}_1 + \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_K x_K\right)^2}{N - K}$$

Two Stage Least Squares: Interpretation

The interpretation of the coefficient in the second stage will be based on the instrument.

Sometimes instruments can be valid, but too specific/small (rainfall may lower demand, sure, but maybe that's different than a systematic shift…)

Two Stage Least Squares: Second Stage

Regress x_k on $Z = [z, x_1, \dots, x_{k-1}]$

Get coefficient of $\hat{\delta} = (Z'Z)^{-1}Z'x_k$, predicted value of

$X\hat{\delta} = X(Z'Z)^{-1}Z'X$ (for all of the x's… for the non instrumented ones this will just be x_n)

$$\left(X\hat{\delta}\right)' = X'Z'(Z'Z)^{-1}X'$$

Use this in a regression of y on $X\hat{\delta}$, will get

$$\left[\left(X\hat{\delta}\right)' X\hat{\delta}\right]^{-1} \left(X\hat{\delta}\right)' y = (X'Z'(Z'Z)^{-1}X' \, X(Z'Z)^{-1}Z'X)^{-1}(X'Z'(Z'Z)^{-1}X'\mathrm{y})$$

Two Stage Least Squares: Inference

The first stage regression is a key tool in assessing whether an instrument is "strong" or "weak" in the multiple regression setting. A very weak instrument retains non-trivial bias in finite samples and very low power.

Would like the first stage to have a large R^2 value / large F statistic. Can introduce bias otherwise (Angrist Kreuger…).

Note that while we can estimate this in two steps, one should always use a stored routine. The coefficients will be the same, but the standard errors in the second stage need to incorporate the fact that the x_K is an estimate… which it will not do if you manually run the second regression.

Specification Tests

So, can we test for whether x is correlated with the error term? This is the key assumption that drives whether or not our coefficients are unbiased / if our model is well identified….

If we can it would be useful to have a guide of when to use least squares and when to use IV estimators (as only one possible application!)

Can we test if our instrument is valid, and uncorrelated with the regression error, as required?

The hypothesis testing in this case would be

$$H_0 : cov(x, \epsilon) = 0 \; vs. \; H_1 : cov(x, \epsilon) \neq 0$$

Specification Tests: Intuition

$$H_0 : cov(x, \epsilon) = 0 \; vs. \; H_1 : cov(x, \epsilon) \neq 0$$

If the null hypothesis is true, then both the least squares estimator and the instrumental variables estimator are consistent (that is, they should be the SAME)

Specification Tests: Intuition

$$H_0 : cov(x, \epsilon) = 0 \; vs. \; H_1 : cov(x, \epsilon) \neq 0$$

If the null hypothesis is true, then both the least squares estimator and the instrumental variables estimator are consistent (that is, they should be the SAME)

If the null hypothesis is false, then the least squares estimator is not consistent, and the instrumental variables estimator is consistent (that is, they should be DIFFERENT)

Specification Tests: Intuition

Naturally, if the null hypothesis is true, then one should want to use the more efficient = lower variance estimator (least squares in this case), which is the least squares estimator

If the null hypothesis is not true, then it would be wrong to use the least-squares because of bias/inconsistency. The only (potentially!) consistent estimator is the instrumental variables estimator

There are several forms of the test, usually called the Hausman Specification test

"Specification Tests in Econometrics" (1977)

Specification Tests: Intuition

Need two estimators that go to different limits under H_1 and the same under H_0

If the estimators produce "different" coefficients $=>$ reject H_0

Form of the Hausman Test

If we have an efficient and inefficient estimator with the same limit (β) under H_0

$\widehat{\beta_{eff}}$ and $\widehat{\beta_{ineff}}$

$\widehat{\beta_{eff}}$ and $\widehat{\beta_{ineff}}$ are both asymptotically normal, mean β, which means

$\widehat{\beta_{eff}} - \widehat{\beta_{ineff}}$ is normal with mean 0… but what is $var\left(\widehat{\beta_{ineff}} - \widehat{\beta_{eff}}\right)$?

Form of the Hausman Test

If we have an efficient and inefficient estimator with the same limit (β) under H_0

$\widehat{\beta_{eff}}$ and $\widehat{\beta_{ineff}}$

It has to be the case that $cov(\widehat{\beta_{eff}}, \widehat{\beta_{ineff}} - \widehat{\beta_{eff}}) = 0$ or we could construct a more efficient = lower variance estimator with the same limit… but we know that's impossible since $\widehat{\beta_{eff}}$ is the efficient estimator.

$$cov(\widehat{\beta_{eff}}, \widehat{\beta_{ineff}} - \widehat{\beta_{eff}}) = -var(\widehat{\beta_{eff}}) + cov(\widehat{\beta_{eff}}, \widehat{\beta_{ineff}}) = 0$$

$$\Rightarrow var(\widehat{\beta_{eff}}) = cov(\widehat{\beta_{eff}}, \widehat{\beta_{ineff}})$$

Form of the Hausman Test

$$var(\widehat{\beta_{ineff}} - \widehat{\beta_{eff}}) = var(\widehat{\beta_{eff}}) + var(\widehat{\beta_{ineff}}) - 2cov(\widehat{\beta_{eff}}, \widehat{\beta_{ineff}})$$

Substitute in $var(\widehat{\beta_{eff}}) = cov(\widehat{\beta_{eff}}, \widehat{\beta_{ineff}})$

$$var(\widehat{\beta_{ineff}} - \widehat{\beta_{eff}}) = var(\widehat{\beta_{ineff}}) - var(\widehat{\beta_{eff}})$$

$$(\widehat{\beta_{ineff}} - \widehat{\beta_{eff}}) \sim \mathrm{N}(0, var(\widehat{\beta_{ineff}}) - var(\widehat{\beta_{eff}}))$$

Can conduct inference using this distribution (square and use χ^2)

Form of the Hausman Test

$$(\widehat{\beta_{ineff}} - \widehat{\beta_{eff}})' \left(var(\widehat{\beta_{ineff}}) - var(\widehat{\beta_{eff}})\right)^{-1} (\widehat{\beta_{ineff}} - \widehat{\beta_{eff}}) \sim \chi_k^2$$

Can conduct inference using this distribution (square and use χ^2)