Intro to Econometrics: Exam 3 Review

Matthew Alampay Davis
July 17, 2023

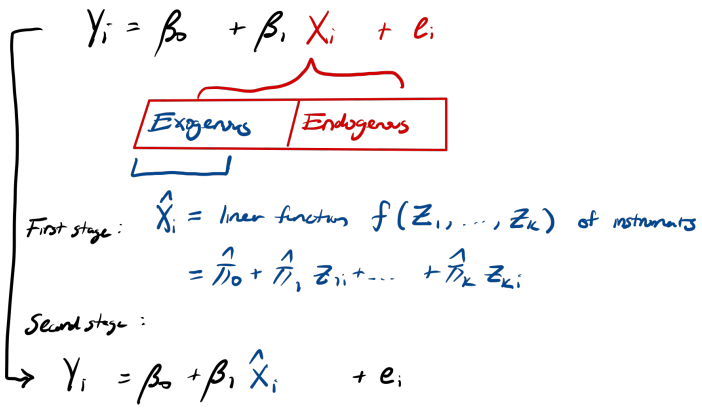
I've prepared some notes that I think may be helpful in guiding and structuring your revision heading into next week's exams. The attached is not meant to be a sufficient summary of each topic; you'll still be best served going through the problem sets and practice problems (my recitation folders sometimes contain additional practice problems, solutions, and notes) and visit topics you are uncertain about in the lecture slides and textbook. That said, students have appreciated similar notes in the past so I hope you'll find them helpful too.

I'll also flag that I have not seen the final exam so the points I address here contain no information about what to expect. Mainly, I just hate when I have to take points off for an error or misunderstanding that could be pre-empted by placing emphasis on certain aspects that may be easy to gloss over or forget. The intention here is to provide more intuition for topics that I think can be confusing and to raise questions that you may find it fruitful to investigate yourelf. Good luck with the exam and thanks for being a great class this year!

1 Instrumental Variables

- There are many estimators that use instrumental variables. The one that we focus on in this course is the two-stage least squares (2SLS or TSLS) because it has the advantage of being able to combine multiple instruments and control variables very easily.

- This is how I explained 2SLS in my office hours. If it is not easy to follow, feel free to forget it.



1. The first equation gives the regression we'd like to estimate

2. Endogeneity of X_i (correlation with e_i) biases estimation of β_1

3. We can think of decomposing the variation X_i into its exogenous variation (blue, independent of the error term) and its endogenous variation (red, covaries with the error term)

4. Ideally, we'll have k instruments $Z_1, ..., Z_k$ that covary with X_i as measured in the by the second equation, the first stage. The greater the proportion of the variation in X_i the instruments can explain, the better can the first-stage prediction $\hat{X}_i$, which is just a linear function of the instruments, approximate X_i. This is the relevance condition for a valid instrument.

5. Otherwise, the instruments do not contain enough information about X_i to be useful, which we call the problem of weak instruments. The textbook goes over the implications are for inference, but it can be understood as a lack of information about X_i, just like how having too small a sample doesn't allow you to learn anything.

6. Even with sufficient covariation with X_i, the instruments need to covary with X_i in the right way. In particular, we want these instruments to covary with *only* the exogenous variation (labeled in blue). If the instruments also covary with enough of the endogenous component (the red part), then the proxy $\hat{X}_i$ will increasingly be a linear function of (an) endogenous variable(s) and thus itself be increasingly endogenous. This violates the exogeneity condition for a valid instrument.

7. As a simplification, if you look at the blue line below the box, we want this to be as wide as possible (relevant) but entirely contained in the left side of the box (exogenous). We want to capture as much of the exogenous variation as possible using a linear combination of the instruments: $\hat{X}_i = f(Z_1, ..., Z_k)$.

- We can test whether instruments are relevant/strong using the familiar F test of joint significance in the first-stage regression. The F statistic measures how much variation in X is captured by variation in the instruments, but says nothing about whether this covariation is exogenous. We essentially cannot measure exogeneity of instruments and mainly only establish exogeneity by through theoretical arguments about the relationship between the instruments and the outcome variable Y.

- So then what does the J statistic tell us?

- Here's the basic logic of the overidentifying restrictions test:

  - Suppose we have one endogenous variable and two possible instruments (just for simplicity)
  - We know how to use one instrument to create a "proxy" for an endogenous variable so that we arrive at a 2SLS estimator
  - Let us do this for both instruments to get a 2SLS estimator for each

- If the two resulting estimators are different enough from one another, they can't both be unbiased. The more different the estimators are, the larger is the J statistic

- For review: what exactly are the null hypotheses of both these tests? Make sure you can express them both in terms of words and in terms of coefficients. Also make sure you know which regression those coefficients correspond to!

- Given the requirements of relevance and exogeneity, it might be important to keep in mind how they affect the resulting estimators. Not just whether they become biased or inconsistent but whether they're biased or inconsistent in a particular direction (in the simple case of one endogenous regressor).

2 Experiments and quasi-experiments

- The difference-in-differences with repeated cross-section is given by the following equation

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 T_i + \beta_3 D_t + u_{it} \quad (1)$$

where

  - T_i is an individual binary indicator for whether observation i is assigned to the treatment group (very important: if an individual i is in the treatment group, they will still have a value of $T_i = 1$ even before the treatment period)
  - D_t is a time binary indicating whether the treatment has been administered (very important: if an individual i is in the control group and doesn't receive treatment, they will still have a value of $D_t = 1$ during the treatment period)
  - X_{it} is the interaction $T_i \times D_i$. It only equals one if individual i is in the treatment group *and* period t takes place after the treatment group has received their treatment.

The coefficient β_1 is the desired difference-in-differences estimate. Why? First consider the treatment group ($T_i = 1$):

  - Post treatment: $\mathbb{E}[Y_{it}|T_i = 1, D_t = 1] = \beta_0+\beta_1+\beta_2+\beta_3$
  - Pre treatment: $\mathbb{E}[Y_{it}|T_i - 1, D_t = 0] = \beta_0 + \beta_2$
  - Take their difference: $\mathbb{E}[\Delta Y^{treatment}] = \beta_1 + \beta_3$

Then consider the control group ($T_i = 0$):

  - Post treatment: $\mathbb{E}[Y_{it}|T_i = 0, D_t = 1] = \beta_0 + \beta_3$
  - Pre treatment: $\mathbb{E}[Y_{it}|T_i - 0, D_t = 0] = \beta_0$
  - Take their difference: $\mathbb{E}[\Delta Y^{control}] = \beta_3$

Then the differences in their differences:

$$\mathbb{E}[\Delta Y^{treatment}] - \mathbb{E}[\Delta Y^{control}] = \beta_1 + \beta_3 - \beta_3 = \beta_1 \quad (2)$$

3 Big data

- I didn't have as much time for this chapter but the main thing to flag is that between regular regression for causal inference, this big data section, and the time series sections, we've come across three different notions of "prediction" that are often confused for one another. These predictions are evaluated by how they minimize the following:

  1. In-sample regression: the mean squared error
  2. Big data: the mean squared prediction error
  3. Time series: the mean squared forecast error

  What are the differences between these? What are their objectives? What are the implications for inference and interpretation? Why do they not lead to the same estimates?

- In all cases, a model $Y = f(X)$ is estimated using some sample data. Then they each use this model to produce some prediction $\hat{Y}$ to be evaluated against some true value Y which gives their respective errors named above. Some key differences between these concepts in this prediction step:

  - whether the true value Y is "in sample" or "out of sample", i.e., whether the true values were used to estimate the model
  - distinguishing big data prediction and time series forecasts/predictions: whether the prediction $\hat{Y}$ is a function of in-sample X or out-of-sample Y

- You should be able to explain the conceptual differences between lasso, ridge, and principal component analysis in words. How do these methods affect inference/interpretability compared to regular unpenalized regressions? What are the benefits?

4 Time series and dynamic causal effects

- We can estimate dynamic multipliers by running regressions of the following form (for simplicity, we are here assuming just one independent variable of interest X):

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + ... + \beta_p X_{t-p} + u_t \quad (3)$$

or we can estimate it through the differenced regression:

$$Y_t = \gamma_0 + \gamma_1 \Delta X_i + \gamma_2 \Delta X_{t-1} + ... + \gamma_p X_{t-p} + u_t \quad (4)$$

- In the first regression, β_1 is the estimated contemporaneous effect of an increase in X by one unit. β_2 is the estimated effect of a one-unit increase in $X_{i,t-1}$. Since we cannot change the past value of X today, you can interpret this as the estimated effect of a one-unit increase in the previous period's level of X.

- Suppose we only experience a one-unit impulse shock in X at time t and otherwise X is zero in all other periods. If

the model is correctly specified, this would imply this one-off shock increases Y_t by β_1, increases Y_{t+1} by β_2, ..., increases Y_{t+p} by β_p

- This motivates consideration of the cumulative effect of this one-off shock, captured by the cumulative multiplier. Two years after this one-off shock, the total effect is a β_1 increase in Y_t and a β_2 increase in Y_{t+1}. This results in a two-period cumulative multiplier of $\beta_1 + \beta_2$, the total impact on Y that the one-off shock has over two years. Over p periods, the cumulative effect is of course $\beta_1 + \beta_2 + ... + \beta_p$. According to our model, a one-off shock does not have a measurable impact after p periods so this sum represents the long-run cumulative multiplier of a one-off shock.
- These cumulative effects are immediately given by estimating the second regression by the following relations:
  - $\gamma_0 = \beta_0$
  - $\gamma_1 = \beta_1$
  - $\gamma_2 = \beta_1 + \beta_2$
  - $\gamma_3 = \beta_1 + \beta_2 + \beta_3$
  - ...
  - $\gamma_p = \sum_i^p \beta_i$

  So you can back out the coefficients of the first regression from the coefficients of the second regression and vice versa. However, only the second regression can give you the appropriate standard errors for cumulative multipliers and only the first regression can give you the appropriate standard errors for the dynamic multipliers
- Importantly: this second equation contains differenced regressors *except for the non-differenced term representing the pth lag!*. Not realizing this is a very common mistake.

5 General

- Read questions closely the first time
  - Subquestions often ask you for multiple things. People lose unnecessary points for getting the econometrics correct but then failing to notice the "Discuss" follow-up question
  - The other side of this is that people lose time by answering questions that aren't asked just to be overly safe.
- If a question asks you to interpret a coefficient, write an interpretation in words describing the implied relationship between the relevant variables, including units for all. You can almost never go wrong with "a one-(unit? percent? percentage point? standard deviation?) increase in X is associated with a $\hat{\beta}$ (unit? percent? percentage point? standard deviation?) increase in Y". That said, keep interpretations in terms of the underlying units (e.g., "a 1% increase in X" is better than "a one-unit increase in log X")
- When arguing for (non-)significance, saying "it is (in)significant" is not enough: what is "it"? and how do you know? For example, cite a p-value or confidence interval or test statistic and indicate which coefficient(s) it corresponds to. This can still be done in one or two sentences.
- On units, the textbook tells you how to interpret a log-log, log-linear, linear-log, and linear-linear regression. What about the case in the practice problem where we were regressing the first difference of log GDP against the first difference of log M2? For example

$$\Delta \log Y_t = \beta_0 + \beta_1 \Delta \log M_t + u_t \tag{5}$$

  - First note that when you're taking differences in logs of some unit, the result is still in log units. So $\log(GDP_{2021}) - \log(GDP_{2020})$ is still in units of log GDP. Thus, we're still talking about log variables and thus we can interpret the coefficients in terms of percentages
  - Also note that the difference in logs is a growth rate. So you are regressing a percentage against a percentage and from our binary dependent variables chapter, we know we can thus interpret the coefficients in terms of percentage points (or we should have; lots of people made this mistake in problem set 6)
  - Both interpretations are permissible because both of the following are equivalent, just note the difference in units and language used:
    1. "A one percent increase in **the money supply** is associated with a $\hat{\beta}_1$ percent increase in **GDP**."
    2. "A one percentage point increase in **the growth rate of the money supply** is associated with a $\hat{\beta}_1$ percentage increase in **the growth rate of GDP**."
- We've encountered several different standard errors:
  1. heteroskedasticity-robust standard errors
  2. cluster-robust standard errors
  3. heteroskedasticity and autocorrelation-robust standard errors (aka Newey-West standard errors)

  Keep in mind what problem they intend to address, i.e. what are they robust to? what models do they correspond to? what assumption is violated if we don't use them? does using them make it more likely or less likely to reject the null hypothesis relative to IID standard errors?
- Might be helpful to revise the basic logarithm rules
- At least two topics lend themselves to essay/understanding-based questions: discussion of whether an analysis has internal/external validity and discussion of whether a particular variable is exogenous or endogenous. These often require some creativity to apply it to a new context and thus a good understanding of what those concepts entail.