Proofs in EC221: A Cheat Sheet

Tom Glinnan

February 2022

This is a list of ideas which you can try to use when coming up with your own proofs in EC221. I'll keep adding on to the list if I think of more

A Sidenote: Learning Proofs

In both this course and in others, try your best not to revise by just memorising proofs. A much better idea is to look through a proof and see that *you yourself* can explain what's being done at every step. Equally, when giving answers to proof-based questions on problems sets or on the exam, make sure you explain why you can do what you're doing at each step (unless it's something very obvious, like rearranging terms) - that way it's clear that you know what you're doing. And remember: if you're looking at a proof and have no idea at all about a particular step, it might be a good idea to step back and re-look at the math itself!

Some Ideas

- **Using the LIE**. Something that comes up often is the idea that we can sometimes do our proofs in two steps: by first 'treating the x_i as fixed'[1] and taking expectations. This works because of the LIE: the fact that $\mathbb{E}(f(X,Y)) = \mathbb{E}(\mathbb{E}(f(X,Y)|X)) = \mathbb{E}(\mathbb{E}(f(X,Y)|Y))$. The intuition of the LIE is just that if we have a table, we can add up all of its entries in one sum, or alteratively we can add up all of the rows, then add up all the totals, or add up all the columns then add up all the totals. See below for an application of how we might use it
- **(*Key*) Plugging in the model, and relating the estimator to the true parameter**. Think back to the big picture of what we're trying to do in the LT: we make some assumptions about the DGP[2], and prove that under those assumptions the DGP has certain properties (unbiasedness, efficiency, consistency, etc). In order to prove these properties, we therefore have to use our assumptions about the DGP. Because of this, many of these proofs start in the same way:

    1. Write down the equation for the estimator. Whether the matrix or vector form is easier to work with depends on the question

[1] Some terminology. *Deterministic* means fixed / non-random, *stochastic* means random

[2] For example, the Gauss-Markov assumptions

2. Plug in the true model
3. Isolate the true parameter that's being estimated. Then, you'll be able to apply your assumptions

For example:

Lemma 1. *If the DGP can be written as $y_i = \beta' x_i + \varepsilon_i$ with $\mathbb{E}(\varepsilon_i|x_i) = 0$ and data is iid, then the OLS estimator $\hat{\beta}$ is unbiased for β*

Proof. Start with the (here: vector) expression for $\hat{\beta}$, which is $\hat{\beta} = (\frac{1}{n}\sum x_i x_i')^{-1}\frac{1}{n}\sum x_i y_i$. Now plug in the true model: as $y_i = \beta' x_i + \varepsilon_i$ we can write

$$\hat{\beta} = \Big(\frac{1}{n}\sum x_i x_i'\Big)^{-1}\frac{1}{n}\sum x_i(x_i'\beta+\varepsilon_i) = \Big(\frac{1}{n}\sum x_i x_i'\Big)^{-1}\frac{1}{n}\sum x_i x_i'\beta+\Big(\frac{1}{n}\sum x_i x_i'\Big)^{-1}\frac{1}{n}\sum x_i\varepsilon_i$$

Since $(\frac{1}{n}\sum x_i x_i')^{-1}\frac{1}{n}\sum x_i x_i' = I$, this can be written as $\hat{\beta} = \beta + (\frac{1}{n}\sum x_i x_i')^{-1}\frac{1}{n}\sum x_i\varepsilon_i$. This isolates the true parameter β: now that we have an expression where we have $\hat{\beta}$ on one side and β on another side, it's easy to see how $\hat{\beta}$ relates to β. Specifically for unbiasedness we need to show that $\mathbb{E}\hat{\beta} = \beta$. Using the two step procedure mentioned above, we first take $\mathbb{E}(\cdot|x_i)$ on both sides, giving:

$$\mathbb{E}(\hat{\beta}|x_i) = \beta + \Big(\frac{1}{n}\sum x_i x_i'\Big)^{-1}\frac{1}{n}\sum x_i\mathbb{E}(\varepsilon_i|x_i)$$

since the expectation of the (constant vector) β is β and since we're conditioning on x_i we can take all of the x_i out of the expression. Now our assumed property $\mathbb{E}(\varepsilon_i|x_i) = 0$ means we have $\mathbb{E}(\hat{\beta}|x_i) = \beta$, and so $\mathbb{E}(\hat{\beta}) = \beta$ follows by taking expectations on both sides and applying the LIE □

If you look for it throughout the course, you'll see that this structure of proof comes up a lot: for example, when talking about s^2 (which is a function of the residuals $\hat{\varepsilon}$) we use the fact that $\hat{\varepsilon} = M\varepsilon$ to get an expression for s^2 which involves the true errors ε instead of the residuals, because our assumptions are about errors, not residuals

- **Taking transposes to swap order**. We know that $(AB)' = B'A'$. As we've seen in problem sets, this is often a useful thing to look out for, especially if we're trying to make a quadratic form. Remember though that $AB = BA$ isn't true in general!
- **$X'X$ is symmetric and psd**. Can be useful to know (if you want to you can redo the proof for yourself) - remember there's lots we can do if a matrix has these properties
- **Relationships between assumptions**. $\mathbb{E}(x_i\varepsilon_i) = 0$ iff $cov(x_i, \varepsilon_i) = 0$ iff $corr(x_i, \varepsilon_i) = 0$, where the equivalences follow as $\mathbb{E}(\varepsilon_i) = 0$. Try to remember how strong each of the typical assumptions we make are: x_i independent of $\varepsilon_i \implies \mathbb{E}(\varepsilon_i|x_i) = 0 \implies \mathbb{E}(x_i\varepsilon_i) = 0$, but the reverse implications don't hold[3]

[3]The first implication follows as if x_i is independent of ε_i then knowing about x_i tells me nothing about ε_i, and in particular tells me nothing about its mean, so $\mathbb{E}(\varepsilon_i|x_i)$ is the same for all x_i, and so must be 0 (as $\mathbb{E}(\varepsilon_i) = 0$). The second implication comes from the LIE: $\mathbb{E}(x_i\varepsilon_i) = \mathbb{E}(x_i\mathbb{E}(\varepsilon_i|x_i))$

- **The 'Trace Trick'**. This is in the proof of the variance of some estimators. The trace of a matrix A is the sum of its diagonal elements, which turns out to always equal the sum of the eigenvalues (when it's diagonalisable). There are two very useful facts about the trace:
  - The trace of a scalar is that same scalar (of course)
  - $tr(AB) = tr(BA)$. More generally, if there are 3 elements ABC then $tr(ABC) = tr(CAB)$ - we can take the last element and put it at the front

  If you like, you can prove tat $tr(AB) = tr(BA)$ by writing out AB sum notation. Also, the eigenvalue thing comes straight from the second property: $tr(A) = tr(PDP^{-1}) = tr(P^{-1}PD) = tr(D)$, and of course $tr(D)$ is the sum of the eigenvalues. Have a look at the proof to show that s^2 is unbiased from tthe end of lecture 2 for an example!