Problem set 4
ANSWERS
607: Applied Macroeconometrics
Fall 2016
Jon Faust

The following is due at the beginning of next class. You can turn in any paper in my mailbox or in class; email me and requested computer work. You may work in groups; hand in a single submission for the group. The submission should list those who contributed.

1. Terms

   (a) Size. Given a model $Y \sim P_\theta, \theta \in \Theta, Y \in \mathcal{Y}$, a conventional test is completely defined by a hypothesis, $H_0 : \theta \in \Theta_0$, where Θ_0 is a subset of Θ, and a rejection region, $\mathbf{Y}^c$, which is a subset of $\mathcal{Y}$. If we observe $Y \in \mathbf{Y}^c$, we consider H_0 to be rejected. In this notation, what is the size of the test? And what is the power of the test against $\theta = \theta_k$.

   Answer/comment

   Here we discuss size, which for emphasis we might call exact size. When the null is not simple, size is the maximum probability you reject the null when it is true over all the ways the null could be true:

   $$\sup_{\theta \in \Theta_0} P_\theta(Y \in \mathbf{Y}^c)$$

   Power under θ_k:

   $$P_{\theta_k}(Y \in \mathbf{Y}^c)$$

   Here we have taken the rejection region as a primitive. But we often choose that region by first picking the size, and then choosing the particular region of that size in order to maximize power. And then we can talk of power of a size-α test. Smaller size leads to smaller power. That is, when you lower your probability of falsely rejecting the null, you also lower your probability of rejecting it when it is false.

(b) Define *consistent test*

Answer/comment

A test is consistent if for every $\theta_k \notin \Theta_0$, power goes to 1 with sample size. No matter how the null is false, you'll figure it out using the test if the sample size gets large enough.

(c) Define *uniformly most powerful test*.

Answer/comment

This ties back to our discussion of the Neyman-Pearson Lemma in the previous problem set.

Sticking with the same setup we have used above, for any test we can asses the power against each $\theta_k \in \Theta_K$. Suppose that for each θ_K there is a most powerful test. If one and the same test attains this maximum power for each $\theta_K \in \Theta_K$, then this test naturally deserves the title Uniformly Most Powerful.

Some details: Since we are comparing power across tests, we have to hold the size constant. Thus, we should call the test 'the UMP size-α test' where we have limited the universe to all size-α tests.

Second, a natural way to find UMP tests when they exist is to look at the form of the point optimal test (see the earlier Neyman-Pearson discussion) and see that the form of the test does not vary with $\theta_k \in \Theta_k$.

---

2. Zellner's Seeminly Unrelated Regressors (SUR) estimator. We have an N-equation system:

$$y_{nt} = X_{nt}\beta_n + \varepsilon_{nt},$$

or in vector notation:

$$\mathbf{Y}_n = \mathbf{X}_n\beta_n + \varepsilon_n$$

Suppose that all of the εs are independent of all of the Xs, mean zero and iid. The εs are, however, contemporaneously correlated across equations. That is, if $e_t = (\varepsilon_{1t}, \dots, \varepsilon_{nt})'$, $Ee_te_t' = \Omega$, where Ω need not be diagonal.

(a) Would OLS on each equation separately give rise to CAN $\hat{\beta}$s?

Answer/comment

Under mild additional regularity conditions on the Xs (full rank

$X'X$, etc.) and εs, each equation satisfies the assumptions for OLS to be CAN.

(b) We can stack our system into a single equation,

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where $\mathbf{Y}$, β and ε are $NT \times 1$ vectors formed by stacking the N underlying column vectors, 1 through N. For, example, for $\mathbf{Y}$:

$$\mathbf{Y}' = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_N')'$$

The matrix $\mathbf{X}$ is block diagonal with $\mathbf{X}_j$ as the j^{th} diagonal block:

$$\mathbf{X} = I_K \otimes X$$

where remember X is $T \times K$.

Under the same conditions as discussed in part (a), OLS on this equation will give rise to CAN estimates. Show that OLS on the stacked system gives the same answer for the βs (and hence $\hat{\varepsilon}$s) as equation-by-equation OLS.

Answer/comment

Because the $\mathbf{X}$ matrix is block diagonal, you should easily be able to verify this. You'll see much of what you need in the FGLS==OLS problem below.

(c) This new system satisfies,

$$(\Omega \otimes I_T) = E\varepsilon\varepsilon' = \Omega \otimes I_T$$

where I_T is a $T \times T$ identity matrix and Ω was defined above. The GLS estimator is then:

$$\hat{\beta}_{GLS} = (\mathbf{X}'(\Omega \otimes I_T)^{-1}\mathbf{X})^{-1}\mathbf{X}'(\Omega \otimes I_T)^{-1}\mathbf{Y}$$

Describe the natural FGLS estimator.

Answer/comment

Estimate the equation by OLS. Get the vector ε. Reshape it into a matrix with the $\hat{\varepsilon}$s for the j^{th} equation in column j. Call the $T \times N$ matrix A. Form

$$\hat{\Omega} = A'A/T$$

Then form

$$\hat{\beta}_{FGLS} = (\mathbf{X}'(\hat{\Omega} \otimes I_T)^{-1}\mathbf{X})^{-1}\mathbf{X}'(\hat{\Omega} \otimes I_T)^{-1}\mathbf{Y}$$

where

$$(\hat{\Omega} \otimes I_T) = \hat{\Omega} \otimes I_T$$

This estimator is Zellner's SUR estimator.

(d) Show that SUR is numerically equal to OLS when each equation has the same regressors.

<u>**Answer/comment**</u>
Let's define $\tilde{\beta} = \hat{\beta}_{FGLS}$ to get rid of that FGLS subscript:

$$\begin{aligned}
\hat{\beta}_{FGLS} &= (\mathbf{X}'(\hat{\Omega} \otimes I_T)^{-1}X)^{-1}X(\hat{\Omega} \otimes I_T)Y) \\
\tilde{\beta} &= ((I_K \otimes X')(\hat{\Omega} \otimes I_T)^{-1}(I_K \otimes X))^{-1}(I_K \otimes X')(\hat{\Omega} \otimes I_T)^{-1}\mathbf{Y})
\end{aligned}$$

We need the following Kronecker identities, the first of which we applied in the second line above,

$$\begin{aligned}
(A \otimes B)' &= (A' \otimes B') \\
(A \otimes B)^{-1} &= (A^{-1} \otimes B^{-1}) \\
(A \otimes B)(C \otimes D) &= (AC \otimes BD)
\end{aligned}$$

Repeatedly apply the latter two to the expression above, and you will find Ω cancelling so that,

$$\hat{\beta}_{FGLS} = (X'X^{-1}X \otimes I_k)Y$$

which is OLS.

(e) Our first vector autoregression (VAR). Stay with the SUR system just described, but suppose that the X variables in each equation are a constant and p lags of each of the y variables.
Define $\tilde{X}$ as the natural RHS for each equation. That is $\tilde{X}$ is a matrix with t^{th} row equal to:

$$(1, y_{1t-1}, \ldots, y_{1t-p}, \ldots, y_{Nt-1}, \ldots, y_{Nt-p})$$

(where that 1 is for the constant).

In this case, we call the system a VAR, vector autoregression.

For the VAR, then,

$$\mathbf{X} = I \otimes \tilde{X}$$

and FGLS and OLS are equivalent.

Question: Why do lazy time series workers like VARs?

Answer/comment

Because they are working with a system of equations, but it is cleverly designed so that equation-by-equation OLS is efficient.

More generally, you might start to notice that sometimes the notation gets a bit dense (until you are used to it), but fundamentally this problem just added 2 things to our simplest understanding of OLS and GLS.

1) In the linear case, you should never think of a multi-equation case, you stack the system and analyze it using single equation tools.

2) We have one more case where it seems that GLS is appropriate, but GLS reduces to OLS (so that OLS is efficient). This arises when the equation has the form of a stacked system of equations with the RHS in each equation the same so that $\mathbf{X} = I \otimes_N \tilde{X}$ for some matrix of regressors $\tilde{X}$.

---

3. Class presentation. We are moving toward the realization that all our magical asymptotic results are magical until we remember that these are asymptotic approximations. And whenever you use an approximation you should find some reasonable method to evaluate the adequacy of the approximation.

   In statistics, Monte Carlo is one essential tool in this regard. The key idea behind Monte Carlo work is that if you want to know the distribution of some random item AND if you know how to generate independent replicates of that random variable, then the following trivial recipe always works.. Generate a zillion replicates and take the histogram of these to be a pretty good guide as to the true distribution. (Of course, we can summarize the collection with more formal measures than a histogram as well, such as approximating $F(c)$ by the share of replicates less than c.)

   Here is the simplest Monte Carlo recipe for understanding the distribution of a statistic:

- Pick a DGP (remember that a DGP a recipe for simulating a data sample from the implied process)
- Generate a zillion independent samples of size T using the DGP.
- Calculate your statistic of interest on each of the zillion samples.
- Then look at the histogram or empirical distribution of the zillion replicates to get a sense of the distribution of the statistic in a sample of size T.

In this problem, we'll use Monte Carlo to see how well 'robust standard errors' of the variety proposed by White perform.

Maintained model:

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

$\varepsilon_t \sim N(0, \sigma^2_{\varepsilon,t})$ and ε_t uncorrelated with x_t. Thus, the model allows heterskedasticity of unknown form in ε. Further,

$$\begin{aligned} x_t &= u_t + v_t \\ u_t &\sim N(0, \sigma^2_u) \end{aligned}$$

where v_t is 1 with prob. π and 0 with prob. $1-\pi$. Also: u, v, and ε are mutually independent.

The observables are x and y, the parameters are $\alpha, \beta, \sigma^2_{\varepsilon,t}, \sigma^2_u, \pi$.

(a) Will the OLS estimate of β be consistent? Efficient? Asymptotically normal? (Sketch proofs)

Answer/comment

$$\hat{\beta} - b = \bar{Q}^{-1}\bar{w}$$

where $Q_t = x_t x_t'$ and $w_t = x_t \varepsilon_t$.

We have $Ew_t = 0$ and w_t is xxx so a WLLN applies. Thus, $\bar{w} \to_p 0$.

Similarly, $EQ_t = Q \neq 0$, and since Q_t is iid, a WLLN applies, and $\bar{Q} \to_p Q \neq 0$

Hence, $\hat{\beta} \to_p \beta$, and OLS is consistent.

So long as ε_t satisfies the Lindberg condition, w_t will as well (e.g., the variance of ε is bounded above and below. Then a CLT will apply to $\sqrt{T}\bar{w}$:

$$\sqrt{T}\bar{w} \overset{a}{\sim} N(0, \Omega)$$

for some Ω, and OLS will satisfy,

$$\sqrt{T}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, Q^{-1}\Omega Q^{-1})$$

OLS need not efficient under heteroskedasticity, for example, some form of FGLS will be. Given that we have said nothing about the form of the heterosked., we cannot be sure if there is a feasible estimator that is more efficient.

(b) Supposing that we choose to use the OLS estimator, we need to use an estimator of the variance-covariancematrix of the coefficients that is appropriate under heteroskedasticity. This are called HC for heteroskedasticity consistent. We'll consider several versions of White's variance-covariancematrix. There are several asymptotically equivalent versions of the White HC estimator. STATA, for example, offers 3 versions of robust standard errors (the `robust`, `robust hc2`, and `robust hc3` options of the regress command, see `help regress`).

Problem: Give the formula for these three different versions and explain briefly why they are asymptotically equivalent.

(c) One main use of the standard errors of a coefficient is creating t-statistics. We'll assess the properties of White's estimators by checking the distribution under the null hypothesis of t statistics under particular instances of the maintained model.

We will follow the recipe: pick a DGP, generate a zillion replicates of the relevant t statistic. Then we'll summarize the distribution by asking how often the statistics exceed the, say, asymptotically justified 5 percent critical value. Of course, we'd like this to be 5 percent.

So, we'll try various instances of the DGP. None will have heteroskedasticity, the both OLS and all the versions of White should give valid results asymptotically (that is the empirical rejection frequency will be 5 percent in a sufficiently large sample).

What will differ across the DGPs is the sample size and π: the probability with which 'outliers' happen in the xs. I won't bother to state the specifics of the DGPs here: you'll see them explicitly in the provided computer code.

Specifically, nearly complete code for doing this Monte Carlo in Matlab is provided with the problem set. There are a number of instances of ??? in the code where you need to do something. I think there may be one or two other errors. Study the code until you understand what's going on, then complete it run it and answer the following questions:

- Overall, which estimator seems to be most robust in the sense of delivering approximately proper size across a wide range of T and π?
- Does it matter which version of White's standard errors you use?
- What lessons do you think might generalize from this exercise?

Note 1: I would like you to learn good programming habits (if you don't already possess these). This is why I start you off with partially completed code.

Note 2: The code as written computes the results and puts them in a matrix, but doesn't print them. You need to figure out how to print the tables. I have included a simple table printing routine (`prtTable.m`). It was written for a slightly different purpose, but it will give you an idea about how to simply print rough, but nice looking tables. If you are more into elegance, I have included a more complicated routine that will generate latex code that can be compiled into nicer looking tables. These are just provided to give a sense of where you might go in making it easy to do stuff like this.

Note 3: We will continue this problem in a subsequent problem set, illustrating the glory of writing good, re-usable computer code.