Problem set 2
ANSWERS
607: Applied Macroeconometrics
Fall 2017
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.

This problem set continues our exploration of the fact that most of advanced econometrics comes down to analyzing quadratic forms that are the result of some quadratic approximation (often, a linear approximation to a first order condition). A key to understanding quadratics is the GLS principle.

For the entire problem set, 2.4. A small digression: generalized distance and the GLS principle may be useful background.

For the SUR question, 3.2. All linear models are single equation, and therby are easy may be useful background.

1. Review. Define...

   (a) Delta Method

   Answer/comment

   This question continues our emphasis on the point that advanced econometric theory comes down, in many cases, to a dogged pursuit of linear or quadratic approximations.

   Aside: Linear or quadratic? We often take a linear approximation to some moment condition. That moment condition can often, itself, be seen as a first order condition. Thus, our linear approximation to the moment condition is a quadratic approximation to the original objective.

   Given a statistic with an asymptotic distribution, the delta method is a way to establish the asymptotic distribution of functions of that statistic. Assume, θ a vector and $g(\theta)$ is a vector function. If we have

   $$\sqrt{T}(\hat{\theta} - \theta) \rightarrow_d N(0, V)$$

then

$$\sqrt{T}(g(\hat{\theta}) - g(\theta)) \rightarrow_d N(0, G(\theta)VG(\theta)')$$

where G is a matrix first derivative of g. This holds where the derivatives are continuous. To sketch a proof, take a Taylor series approx:

$$g(\hat{\theta}) \approx g(\theta) + G(\theta)(\hat{\theta} - \theta)$$

Waving our hands and rearranging, we have

$$\sqrt{T}(g(\hat{\theta}) - g(\theta)) \approx \sqrt{T}G(\theta)(\hat{\theta} - \theta)$$

which we could imagine would give the result stated above.

Like all asymptotic results, one must ask whether the approximation will be good in your case. Suppose we know that the the asymptotic normality of $\hat{\theta}$ describes the behavior of $\hat{\theta}$ well in the case at hand, then we can do some reasonable thinking about how accurate the Delta method approximation will be by investigating g in a region of what we believe to be the true θ to see how good the first order Taylor series approx. is. Basically, the approx. will be no better than the first order Taylor series approx. to g.

(b) Likelihood ratio test

Answer/comment

Here it is important to investigate two cases: first is when we have a point null and alternative hypothesis and second is when we have composite hypotheses instead. Of course, the Neyman-Pearson lemma shows that a likelihood ratio-based test will be most powerful under certain assumptions in the case of point hypotheses.

In the point hypotheses case, suppose $L(\theta|X)$ is the likelihood, null is $H_0 : \theta = \theta_0$, an alternative of interest is $H_K : \theta = \theta_K$. The LR test would reject H_0

$$\frac{L(\theta_0|X)}{L(\theta_K|X)} < \eta$$

for some fixed η.

Of course, any monotonic transform of the statistic gives an equivalent test, and we often consider minus twice the log of the likelihood ratio and reject for large values.

When the two hypotheses are composite instead, null is $H_0 : \theta \in \Theta_0$, an alternative of interest is $H_K : \theta \in \Theta_K = \Theta - \Theta_0$ (where $\theta \in \Theta$ is maintained), the LR test is generally computed as

$$\frac{\sup_{\theta \in \Theta_0} L(\theta|X)}{\sup_{\theta \in \Theta} L(\theta|X)}$$

and once again minus twice the log is the often the stat. reported.

In this part of the course, we are hammering home the importance of quadratic approximation and the GLS principle: every $X'X$ deserves to have a Σ^{-1}: $X'\Sigma^{-1}X$, where Σ is the variance-covariance matrix of X.

To relate these ideas to the LR test, take a quadratic approximation to the log-likelihood under the null hypothesis that $\theta = \theta^*$:

$$L(\theta^*) \approx L(\hat{\theta}_{MLE}) + q(\hat{\theta}_{MLE}) + \frac{1}{2}(\theta^* - \hat{\theta}_{MLE})' Q(\hat{\theta}_{MLE})(\theta^* - \hat{\theta}_{MLE})$$

since $q(\hat{\theta}_{MLE}) = 0$ by construction, we re-arrange to get,

$$-2(L(\hat{\theta}_{MLE} - L(\theta^*)) \approx (\hat{\theta}_{MLE} - \theta^*)' Q(\hat{\theta}_{MLE})(\hat{\theta}_{MLE} - \theta^*)$$

Since minus Q goes to the information matrix, which is the inverse of the variance-covariance matrix of the MLE, the RHS is a quadratic form asymptotically obeying the GLS principle. When the MLE is also asymptotically normal, we have the χ^2 result.

This result, I believe, is properly attributed to S.S. Wilks in (1938) (https://projecteuclid.org/euclid.aoms/1177732360) and is a nice illustration of the fact that statisticians had nicely scratched the surface of the relation between locally quadratic stuff, and normality, and extremum estimators by the 1930s. It took most of the rest of the century to nail down the stuff we're discussing.

Finally, you should know that the Neyman-Pearson Lemma does not apply in the case of a composite null hypothesis: different tests will be most powerful against different θs in Θ_K. Various

asymptotic optimality can be proven, however. For example, the LR test will often be *locally most powerful invariant*. See Engle's Chapter in the Handbook of Ecmet (Vol. 2, ch. 13). The modifiers 'locally' and 'invariant' limit the class of alternatives and of test statistics we are considering, respectively.

---

2. χ^2 distribution.

   (a) Define a χ^2 random variable in terms of gaussian random variables.

   Answer/comment
   If $x_j \sim iidN(0,1)$, then

$$\sum_{j=1}^{p} x_j^2 \sim \chi^2_{(p)}$$

   (b) If $z \sim N(0,\Sigma)$ is a $(k \times 1)$ vector and Σ is full rank, sketch the argument as to why a quadrative form satisfying the GLS principle, $z'\Sigma^{-1}z$, is $\chi^2_{(k)}$.

   Answer/comment
   This uses the ideas in our disucssion of generalized distance and quadratic forms. Take a square L such that $L\Sigma L' = I$. Then

$$w = Lz \sim iidN(0,1)$$

   and

$$w'w = z'L'Lz = z'\Sigma^{-1}z$$

   is the sum of iid N(0,1) variables.

   (c) Sketch the $\chi^2_{(k)}$ density for $k = 1$. How does this differ from the density for higher k?

   Answer/comment
   The support for the key square is nonnegative. The key here is that with 1 degree of freedom, the global maximum is at zero and the density declines monotonically. For all other degress of freedom, the mode occurs at a strictly positive point, that point

moving rightward as the degrees of freedom grow.

(d) State the mean and variance of the $\chi^2_{(k)}$ as a function of k.

Answer/comment
Using your deep knowledge of $N(0,1)$ variables and sums thereof, we have that the expectation of a $\chi^2_{(k)}$ is k and the variance is $2k$.

(e) If $x \sim \chi^2_{(k)}$, what happens to the mean and variance of x/k as $k \to \infty$.

Answer/comment
Notice that x/k is the sample mean of k iid variables that have mean 1 and variance 2. Thus, the mean is 1 for all k and the variance is $2/k$, and

$$\sqrt{k}(x/k - 1) \to_d N(0, 2)$$

---

3. 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:

$$Y_n = X_n\boldsymbol{\beta}_n + \varepsilon_n$$

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 $e_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Nt})'$, $Ee_te_t' = \Omega$, 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,

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

where $\mathbf{Y}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ 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.

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
The key here is that that the $\hat{\varepsilon}_{nt}$s from equation-by-equation OLS are orthogonal to the Xs in the relevant equation. Thus, if we stack them into $\hat{\varepsilon}$, we will also have

$$\mathbf{X}'\hat{\varepsilon} = 0$$

since this stacked system will just give the sum of the $X_n'\varepsilon_n$s across the N equations. Thus, the least squares solution for the βs in the larger system is just as in equation-by-equation OLS.

(c) This new system satisfies,

$$E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \Omega \otimes I_T$$

where I_T is a $T \times T$ identity matrix. 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; re-shape the $\boldsymbol{\varepsilon}$ vector into A, a $T \times N$ matrix with the εs for the n^{th} equation block in column n. Form

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

Then form

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

This estimator is Zellner's SUR estimator.

(d) Give a simple condition when SUR gives numerically the same answer as OLS—even without restricting Ω.

Answer/comment

When the RHS of each equation is the same, SUR=GLS=OLS. One reason to imagine this result might be true goes like this: Take the 2 equation case. Zellner's insight in SUR is that even though the regressions seem to be unrelated, the εs in equation 1 might tell us something about the correct βs in another equation. But our main way of learning about the β in the LS family is in the orthogonality of the ε and Xs in the equation. When the RHS's are the same in each equation, then by assumption, the εs in equation 2 are orthogonal to the Xs in equation 1. In this case, we could say that the εs in equation 2 don't tell us anything about equation 1 we didn't already know from the ε in equation 1.

That's the yacky version. To see this in the formulae is mainly a puzzle in Kronecker algebra.

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

where the last is equation-by-equation OLS.

I suppose that it is in the third line where we learn that $\hat{\Omega}$ is doing the same thing in each equation and thereby not helping.

This result happens to be much more than a curiosity in macroeconometrics and time series. Much work uses vector autoregressions, which have the form of a vector of variables on the LHS and p lags of each variable on the RHS of each equation: thus the

system has the SUR form with the regressors the same in each equation.

One (weak but practical) reason to like VARs relative to other time series systems of equations is that OLS is efficient whereas some SUR or GLS approach would, more generally, be called for.

---

4. OLS, SUR, and GLS again. In ps1, we discussed the fact that OLS is asymptotically efficient, and thereby as efficient as GLS, when computing the sample mean of a covariance stationary process.

   (a) Show that OLS and GLS correspond exactly when the k regressors are an exact linear combination of k eigenvectors of Σ. (Notes: Puzzling over this is probably worth some time, but feel free to look up this result if it doesn't jump out. Working through the linear algebra in this case is of some value. You might ponder how this is related to the fact that OLS and SUR are the same when the regressors are the same in each equation.)

   Answer/comment

   Note: Googling 'OLS equivalent to GLS' would have led you quickly to several sources. These would lead to lots of proofs and partial proofs, many of which involve citations to Amemiya's Advanced Econometrics text. Amemiya (famous JHU alum) gives several related equivalent expressions for when the two are identical. His proof in the text is incomplete and for the particular bit we need here just says its easy (wasn't for me), and notes that T.W. Anderson does a full proof in his classic text The Statistical Analysis of Time Series. You could get this free online on the Hopkins network (try Chapter 2, http://onlinelibrary.wiley.com/doi/10.1002/9781118186428.ch2/pdf Studying this text would reward you handsomely if you want to understand time series.

   Doing this proof requires wading through some linear algebra, but at the heart are great insights about the nature and role of orthogonality in econometrics.

   I'll spell this out a bit more fully than Anderson. I think I have it right, but there are more primes and conformability issues than I can reliably get right without more proofreading than I'm up to right now. Please proofread and get back to me.

Our equation is

$$Y = X\beta + \varepsilon$$

with K columns in X and

$$E\varepsilon\varepsilon' = \Sigma$$

We can always decompose a pos. def. variance-covariance matrix in terms of eigenvectors and eigenvalues as,

$$\Sigma = W\Lambda W'$$

where $W'W = I$ so that $W' = W^{-1}$.

The problem asks us to consider the case in which

$$X = W_k C$$

where C is square and of full rank K (the number of regressors), and where W_k is K columns of W.

For simplicity, we're going to assume that W_k is the first K columns of W so that

$$W = [W_k : \tilde{W}_k]$$

We'll have a bunch of items floating around related to W and also have their analogs related to W_k. These analogs will all have a k subscript.

One key to the problem is that $W'W = I$, $W_k'W_k = I$ and

$$W_k'W = I_k$$

where I_k $(K \times T)$ is the first k rows of a $(T \times T)$ identity matrix.

Just as $\Sigma W = W\Lambda$,

$$\Sigma W_k = W_k\Lambda_k$$

where Λ_k the $K \times K$ upper left block of Λ.

The OLS $\hat{\beta}$ is:

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \\
&= (C'W_k'W_kC)^{-1}C'W_k'y \\
&= (C'C)^{-1}C'(W_k'Y) \\
&= C^{-1}C'^{-1}C'(W_k'Y) \\
&= C^{-1}W_k'Y
\end{aligned}$$

(Note that the third equation line says that regressing Y on X is equivalent to a regression of $W_k'Y$ on C. Huh? Much of what regression is doing is orthogonalizing the Xs. When we wrote $X = W_k c$ where the W_ks are orthogonal, and of unit length, we've simplified the problem a great deal.)

Similarly, GLS is

$$\hat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

Let's to the two terms separately:

$$(X'\Sigma^{-1}X)^{-1} = (C'W_k'W\Lambda^{-1}W'W_kC)^{-1}$$

We can either jump a few steps or spell it out being careful with the various ranks of items. I'll try to do the tedious way right:

$$\begin{aligned}(X'\Sigma^{-1}X)^{-1} &= (C'I_k\Lambda^{-1}I_k'C)^{-1} \\ &= C^{-1}\Lambda_k C'^{-1}\end{aligned}$$

The second term in $\hat{\beta}_{GLS}$ is,

$$\begin{aligned}X'\Sigma^{-1}Y &= C'W_k'W\Lambda^{-1}W'Y \\ &= C'I_k\Lambda^{-1}W'Y \\ &= C'I_k\Lambda^{-1}[W_k : \tilde{W}_k]'Y \\ &= C'\Lambda_k^{-1}W_k'Y\end{aligned}$$

Now putting the two terms back together we have that

$$\hat{\beta}_{GLS} = C^{-1}\Lambda_k C'^{-1}C\lambda_k^{-1}W_k'Y$$

which reduces the OLS estimator.

(b) Note: corrected. Suppose that y_t follows an MA process of order p, $MA(p)$:

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_p\varepsilon_{t-p}$$

where ε_t is mean zero and constant variance.

Show that $Y = (y_1, \ldots, y_T)'$, has only p nonzero autocovariances.

Answer/comment
It should be pretty clear that for $q > p$, $Ey_t y_{t-q} = 0$, and otherwise not. Both ys are collections of εs. If $q \leq p$ some εs are in common, otherwise not.

(c) Take $T > 2p$ and consider Σ, the variance-covariance matrix of Y. Ignoring the first and last p rows of Σ, give a formula for the sum of the elements on any other row.

Answer/comment
Excluding the noted rows,

$$s \equiv \sigma(0) + 2\sum_{j=1}^{p} \sigma(j)$$

This, of course, is a version of a formula that is ubiquitous in time series.

(d) Why does the previous part imply a sense in which for large T, i (a vector of 1s) is *almost* an eigenvector of Σ?

Note: Following on the earlier part of the problem, this is a crude way of seeing why, for MA processes at least, OLS is equivalent to GLS asymptotically.

Answer/comment
The rowsum result above can be written:

$$\Sigma i \approx si$$

for scalar s, where the approximate equality is exact excluding the first and last p elements of the RHS vector.

If this held with strict equality for all elements, then this formula is precisely the definition of i being (proportional to) an eigenvector of Σ. Note: 'proportional to' because by convention, we scale eigenvectors to have unit length.

Using this logic, or more direct verification, you can show that the GLS estimate of the mean has the same asymptotic distribution as the sample mean, hence, both are efficient. Another way of saying this, is that when estimating the mean of MA processes,

the GLS estimator may be thought of as a 'small sample correction' to the sample mean, where the correction mainly affects the first and last p observations.

Going one step further, we can see that this same result holds for AR processes. There are two ways to see this. First, a linear algebra buff could show that the INVERSE variance-covariance matrix of an AR process has that same banded structure as the variance-covariance matrix of an MA— that is, the inverse variance-covariance matrix a nonzero band $(1+2p)$ along the main diagonal. There is much insight in seeing why the MA variance-covariance matrix has the same banded structure as the inverse AR variance-covariance matrix.

Alternatively, and with even more hand-waving, we'll later see that every AR can be approximated arbitrarily well by an MA. Thus, the efficiency result for the MA family carries over.

More generally, Grenander (1954) has a lovely article on efficiency of the sample mean in the covariance-stationary case.

I have deliberately drug you into time series in these problem sets without explicitly giving the classical time series introduction. This may or may not be good pedagogy, but my intention is to stretch your minds a bit before we settle into the standard results.

Finally, some history of thought on this topic.

As recently as the late 1960s, the best minds in the field were working out expressions for the relative efficiency of OLS and GLS in the simplest of cases such as the AR(1), see Chipman, et al. (1968).

One nice thing about these papers is that they reveal much about the structure of covariance stationary processes in cases that can be easily (by modern standards) understood. For example, take the AR(1) case with parameter ρ. The structure of the implied Σ^{-1} implies that the GLS estimate of the mean is

$$\bar{y}_{GLS} = \frac{(1-\rho)^{-1}(y_1 + y_T) + \sum_{t=2}^{T-1} y_t}{T + \frac{2\theta}{1-\rho}}$$

As T goes to ∞, the denominator goes to T and all observations except the first and last are treated just as in the sample mean.

The route we have traced to this result was pointing out that i is nearly an eigenvector of Σ and then appealing to the result that OLS and GLS correspond in this case. The eigenvector result is not a time series result at all, but it seems to come in handy most in time series and it was time series folks who first noted it. Specifically, Puntanen and Stayen (1989) say that the great time series econometrician T.W. Anderson first established the result, and that his establishment of this led to the famous papers by Durbin and Watson deriving, among other things, the Durbin-Watson statistic.

Side note: Although I didn't grasp much of the stuff just discussed above at the time, one of my thesis papers is in this tradition (When are variance ratio tests for serial dependence optimal, https://www.jstor.org/stable/2951545).

Note: The answer set for this question will discuss the following references.

John S. Chipman, Koteswara Rao Kadiyala, Albert Madansky and John W. Pratt, Efficiency of the Sample Mean when Residuals Follow a First-Order Stationary Markoff Process, Journal of the American Statistical Association, Vol. 63, No. 324 (Dec., 1968), pp. 1237–1246.

http://www.jstor.org/stable/2285880

Grenander, Ulf. On the Estimation of Regression Coefficients in the Case of an Autocorrelated Disturbance. Ann. Math. Statist. 25 (1954), no. 2, 252–272.

http://projecteuclid.org/euclid.aoms/1177728784

Simo Puntanen and George P. H. Styan. The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator, The American Statistician, Vol. 43, No. 3 (Aug., 1989), pp. 153-161

http://www.jstor.org/stable/2685062