607
Consistent variance-covariance matrix estimation in relevant sample sizes

Jon Faust

http://e105.org/e607

October 4, 2016

- **The title of this lecture adds 'in relevant sample sizes' to an earlier lecture; thus, we are getting a bit more practical.**
- **Readings**
  - Same as for the earlier lecture
  - But also you might look at:

    For a discussion of these issues, see Hausman and Palmer, Heterskedasticity-Robust Inference in Finite Samples go

    http://economics.mit.edu/files/7422

- **Consistency**
  - Up to now we have explored the CAN-(WCEAVCM) framework
  - To make asymptotically justified frequentist probability statements we generally presume that our point estimators are CAN
  - Our aymptotically justified probability statements only require that we have consistent estimates of the asymptotic variance-covariance matrix.
- **Aside:: Reminder**
  - Supposing that

$$\sqrt{T}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, V)$$

- Wald test statistics such as

$$W = T(\hat{\beta} - \beta_0)'V^{-1}(\hat{\beta} - \beta_0)$$

will asymptotically be $\chi^2_{(K)}$ whether based on the true V or any consistent estimator of V.

▶ **Consistency is not a descriminating standard**

- Consistency is a weak standard, and not a very descriminating one.
- When there is one consistent estimator, there are automatically arbitrarily many.
- If $\hat{\xi}(Y)$ is consistent then so is

$$\tilde{\xi}(Y) = \hat{\xi}(Y) \times g(Y)$$

so long as $g(Y) \to_p 1$.
- There are, of course, arbitrarily many such gs.

▶ **An instance of a consistent estimator is properly viewed as a class of estimators and you should ask which member of the class you should use.**

▶ **Aside:: Example**

- One familiar example here arises in estimating the residual variance in OLS.
- Two estimator are $\hat{\varepsilon}'\hat{\varepsilon}/T$ and $\hat{\varepsilon}'\hat{\varepsilon}/(T-K)$
- These differ by $g = (T-K)/T$ which converges to 1, so when one is consistent, the other is as well.
- Best practice is generally to use $T - K$.
- This practice has some motivation in the fact the the $T - K$ version is unbiased in some cases

And we could give other justifications.

▶ **Macro**

- In macro our samples are small (in the relevant sense) and multiplying our consistent estimators by some astute $g(Y)$ that converges to 1 can often greatly improve relevant sample size performance.
- In this lecture we first illustrate this using White's HC estimator.
- And then make some more general comments.

▶ **White's variance-covariance matrix**

- The White variance-covariance matrix is,

$$\hat{\Omega} = T^{-1} \sum x_t x_t' \hat{\varepsilon}_t^2$$

And you should know conditions under which it is consistent.

- Let's see how these work in finite samples.

▶ A Monte Carlo

- DGP: $y_t = \beta^* x_t + \varepsilon_t$.
- Assume $\varepsilon \sim iidN(0,1)$s
- The xs are iid but skewed

$$x_t = N(0, 0.01) + z_t$$

  where z_t is 1 with pr. π, 0 else, indep. of all else.

▶ And so...

- OLS is consistent, asymptotically normal.
- And the conventional OLS std. err. on $\hat{\beta}$ will be consistent
- Will a White-based std. err. be consistent?

  Yes. White std. err. for β is consistent under a broader set of assumptions than the conventional std. errs.

- Suppose in our application we believe that there might be some heterosked. of the εs.

  So just to be careful, we use White std. err.

- And we use this White std. err. to form a t test of $\beta = \beta^*$.
- We use a nominal 5 percent test and check the empirical frequency with which the null is rejected using conventional standard errors and White-style..

▶ Empirical rej. freq. of a nom. 5 percent test

- rej. freq., conventional vs. White t-stat.

T	conv.	White
30	5.9	24.5
100	5.6	37.8
1000	5.1	9.7

MC replications: 10,000; share shifted: 0.01

▶ Oops

- Conventional standard error matrix does ok in all cases
- White leads to large over-rejection
- And note that the convergence to the nominal size is not monotonic in T, nor is it near complete for $T = 1000$

- **Lesson 1**
  - When we call something robust relative to some baseline, we generally mean good in some sense in a broader range of cases than the baseline.
  - But as usual, 'in some sense' needs clarification
  - White standard errors are asymptotically robust, meaning that in sufficiently large samples, they will give something close to the right answer (in the consistency sense) in a broader range of cases than the conventional standard errors.

    We might naturally call this 'asymptotically robust'
  - There is a big difference between robust in the sense of 'asymptotically robust' and robust in the sense of 'works well in relevant cases'
  - A great deal of econometric practice confuses these two
  - As we shall see, changing an estimator to make it asymptotically robust very often makes it less robust in the practical sense of 'working well in relevant cases'
- **A fix?**
  - Maybe we can make some small sample adjustment that makes the White std. err perform better in small samples

    That is, do the equivalent of multiplying by some $g(Y)$ the converges to one.
- **Finite-sample adjustments for White**
  - Folks have defined and studied a couple of adjustments
  - In STATA, for example, these are given by the hc2 and hc3 options of robust under regress
  - In hc2, we use
    $$\hat{\varepsilon}_t^2/(1-h_{tt})$$
    in place of $\hat{\varepsilon}^2$ in the White formula
  - hc3 uses,
    $$\hat{\varepsilon}_t^2/(1-h_{tt})^2$$
  - Where $h = X(X'X)^{-1}X'$, the projection matrix and h_{tt} is the t, t element.
  - So long as h_{tt} goes to 0 in the approp. sense, these adjusted estimators will be consistent when the basic White estimator is.
- **Aside:: What motivates these adjustments?**
  - Even when the OLS residuals are homoskedastic, the $\hat{\varepsilon}$s will not be

hint: $\hat{\varepsilon}_t = y_t - x_t'(X'X)^{-1}X'Y$ evaluate the variance of this term.

- To see a motivation for the hc2 and hc3 versions, examine the distribution of the $\hat{\varepsilon}$s in standard regressions with homoskedastic errors.

▶ More on the previous Monte Carlo

- rej. freq., conventional vs. White t-stat.

T	conv.	White	HC2	HC3
30	5.9	24.5	12.3	5.6
100	5.6	37.8	18.7	7.6
1000	5.1	9.7	8.2	7.1

MC replications: 10,000; share shifted: 0.01

▶ Bottom line

- HC2 is a bit better; HC3 is pretty good in this case
- Thus, the adjusted versions sometimes avoid the problem with the plane vanilla White standard errors.
- Davidson and MacKinnon's (1993) text argues that the simple version of White should never be used

  As far as I can tell, this advice is essentially ignored in macro

- I am using this example mainly to illustrate that small sample adjustments matter, but the specific lessons for White standard errors are worth remembering.

▶ Modern advanced econometrics

- Subsequent to Davidson and MacKinnon's practical advice, folks have analyzed this issue using higher order asymptotics (Edgeworth expansions)—one element of what I've called modern advanced econometrics.
- For a discussion of these issues, see Hausman and Palmer, Heterskedasticity-Robust Inference in Finite Samples go

  http://economics.mit.edu/files/7422

- Their very concise abstract is entirely in the spirit of what I am arguing here
- Hausman-Palmer abstract

  Since the advent of heteroskedasticity-robust standard errors, several papers have proposed adjustments to the original White formulation. We replicate earlier findings that each of these adjusted estimators performs quite poorly in finite samples. We propose a class of alternative heteroskedasticity-robust tests of linear hypotheses based on an Edgeworth expansions of the test statistic distribution. Our preferred test outperforms existing methods in both size and power for low, moderate, and severe levels of heteroskedasticity.

- **Aside:: Bootstrap**
  - If you are curious, their proposed solution invovles the bootstrap.
- **General lessons**
  - Every consistent estimator actually defines a family of such estimators and the reliability of finite sample work may depend a great deal on which of these you choose.
  - That is, asymptotically robust does not mean robust in a relevant sense.
  - Sometimes we have a route to doing better from higher order asymptotics
  - In other cases, some version of best practice guides us.
- **Families of HAC estimators**
  - Once we allow for time series dependence, these issues multiply
  - Take the scalar, covariance stationary w case for simplicity.
  - In this case, asymptotically we have

$$\Omega = \lim_{T\to\infty} \sigma_w^2 + 2\sum_{t=1}^{T} \frac{T-j}{T}\sigma(j)$$

  - So far we have mainly suggested replacing σs with the sample autocovariances—the $\hat{\sigma}$s.
  - At that point, we take on essentially one more practical issue: where to truncate the sum so that we don't include poorly estimated sample moments.
  - In practice, folks tend to truncate at some small share of the sample size
  - But some rule that is either deterministic or dependent on the data is sometimes used.
- **A parametric alternative**
  - An alternative is to estimate an AR, MA, or ARMA model for the ws and then use the population autocovariances for the estimated process in the general HAC formula.
  - For example, we know that if w_t is an AR(1) with AR parameter ρ, the autocovariances are

$$\sigma(j) = \sigma_w^2 \rho^j$$

  - Thus, rather than computing J sample autocovariances, we could instead compute $\hat{\rho}$ and then not truncate the sum at all in Ω instead using $\hat{\rho}^j$ for all j
  - We still have a truncation issue since we generally we won't know the appropriate AR order.

    Once again we can use data-based method to choose.

- **VARHAC**
  - In the scalar w case, Ω estimates based on a parameteric AR structure might be called AR-HAC

    You seldom hear this, however
  - In the vector w case, we can exploit a vector AR (VAR) structure giving rise to VARHAC estimators.

    this you may hear a good deal
  - VARHAC is commonly used in practice and have been shown to be better in various senses than the Newey-West estimator in some cases
- **In both Newey-West and VARHAC we have taken on a new topic: model selection. How do you choose the number of autocovariances or the order of the AR or MA. That lecture is coming next.**
- **Aside:: Note: finite-order MA cases**
  - In some cases, the structure of the problem implies that the ws will have a finite-order MA representation of known order, say, p.
  - We can exploit this information to form an estimator of Ω tailored to the appropriate $MA(p)$.
  - This case of known MA order for the errors typically arises when using overlapping data.

    For example, in quaterly data for GDP used to compute 4-quarter (that is, annual) growth rates, will generally have an MA(4) element.
  - variance-covariance matrices exploiting a known finite MA structure are often called Hansen-Hodrick variance-covariance matrices.
- **Thus,**
  - We want our variance-covariance matrix to be consistent under serial correlation of the ws of unknown form.
  - Attaining this is possible.

    And if there is one method, there are arbitrarily many
  - We also want our estimates to 'work well' in practice.

    This is a matter of choosing among the arbitrarily many
  - In the time series world, myriad approaches to model selection tend to get invoked in inventing among the myriad HAC estimators
  - Elaborate Box-Jenkins-like approaches are proposed.

See the excellent discussion in Zivot's draft chapter on GMM go

http://faculty.washington.edu/ezivot/econ583/gmm.pdf

- Ultimately, this may drive us toward higher order asymptotics and bootstrapping.