14.385 Recitation 7

Paul Schrimpf

October 31, 2008

1 Midterm

Average was 97.62.

1.1 Bootstrap

In part (a), the best answers to the question about whether the bootstrap is a good idea was: yes, especially if you bootstrap a pivotal statistic; or maybe: yes, but it would be better to use parametric bootstrap.

1.2 Test Scores

You want to evaluate the effect of an after school program on test scores. You have a data set with the following information: test scores, y, whether the child attended the after school program, d, and some other covariates, x, for example, family income and whether the child lives with one or both parents. Participation in the after school program was completely voluntary, but before the program began, the school sent a random subset of students detailed information extolling the benefits of the program. You observe an indicator for whether a student received this information, z.

(a) (15 min) Describe how you could estimate the effect of the program. Propose a test for the hypothesis that the program had zero effect.
An answer is IV.

(b) (15 min) Due to no child left behind, adminstrators especially care about the effect of the program on the low end of test scores. Describe how you could estimate this effect. Briefly discuss how you could compute your estimator.
The intended answer was quantile IV. Most people got that, but not everyone wrote down the objective function correctly. The moment condition for quantile IV is:

$$E\left[\tau - \mathbf{1}(y \leq x\beta(\tau) + \alpha(\tau)d)|x, z\right] = 0$$

which gives us the objective function

$$Q(\alpha, \beta) = \frac{1}{n}\sum \left(\tau - \mathbf{1}(y_i \leq x_i\beta(\tau) + \alpha(\tau)d)\right)' w_i' A w_i \left(\tau - \mathbf{1}(y_i \leq x_i\beta(\tau) + \alpha(\tau)d)\right)$$

The estimates can be computed by using quasi-Bayesian methods as on the last problem set.

(c) (15 min) Suppose a large portion of students received a perfect score on the test. How would you modify your estimator(s)?

The way to modify quantile IV is tricky. The invariance principle that works for quantile regression does exactly not apply here. The reason is that we must condition on z instead of d, as we would in exogenous quantile regression. Let's start from the first order condition.

$$\tau = P(y^* \leq x\beta(\tau) + \alpha(\tau)d)|x, z)$$

where y^* is the uncensored test score. The censored test score, $y = \min\{y^*, 100\}$ is less than or equal to y^*, so

$$\tau \leq P(y \leq x\beta(\tau) + \alpha(\tau)d|x, z)$$

Similarly, we know that

$$1 - \tau = P(y^* > x\beta(\tau) + \alpha(\tau)d|x, z)$$

and

$$1 - \tau \geq P(y > x\beta(\tau) + \alpha(\tau)d|x, z)$$

These two conditional moment inequalities can be combined into an objective function and set inference can be done.

2 HAC

To estimate the asymptotic variance of GMM or do efficient GMM, we need to estimate $\Omega = \lim Var(\sqrt{n}\hat{g}(\beta_0))$. When the data is iid estimation of Ω is straightforward, we can just use its sample analog. When the data is autocorrelated, estimation is more complicated. Newey and West (1987) developed a heteroskedasticity and auto-correlation consistent (HAC) covariance estimator.

We have a series $\{z_t\}$, and we want to estimate its long-run variance, $\mathcal{J} = \lim var\left(\frac{1}{\sqrt{T}}\sum z_t\right)$. If we assume that z_t is covariance stationary, so that $cov(z_t, z_{t+k})$ only depends on k and not t, and denote the kth autocovariance as $\gamma_k = cov(z_t, z_{t+k})$, then we have $\mathcal{J} = \sum_{-\infty}^{\infty} \gamma_k$.

2.1 A naïve approach

$\mathcal{J}$ is the sum of all auto-covariance. We can estimate $T - 1$ of these, but not all. What if we just use the ones we can estimate, *i.e.*

$$\tilde{\mathcal{J}} = \sum_{k=T-1}^{T-1} \hat{\gamma}_k \,,\, \hat{\gamma}_k = \frac{1}{T}\sum_{j=1}^{T-k} z_j z_{j+k}$$

It turns out that this is very bad.

$$\begin{aligned}
\tilde{\mathcal{J}} =& \sum_{k=T-1}^{T-1} \hat{\gamma}_k \\
=& \frac{1}{T}\sum_{k=T-1}^{T-1}\sum_{j=1}^{T-k} z_j z_{j+k} \\
=& \frac{1}{T}(\sum_{t=1}^{T} z_t)^2 \\
=& (\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t)^2 \\
\Rightarrow& N(0, \mathcal{J})^2
\end{aligned}$$

so $\tilde{\mathcal{J}}$ is not consistent; it converges to a distribution instead of a point. The problem is that we're summing too many imprecisely estimated covariances.

2.2 Truncated sum of sample covariances

What if we don't use all the covariances?

$$\tilde{\mathcal{J}}_2 = \sum_{k=-S_T}^{S_T} \hat{\gamma}_k$$

where $S_T < T$ and $S_T \to \infty$ as $T \to \infty$, but more slowly.

This estimator is consistent, but it has poor small sample properties. In particular, it may lead to $\tilde{\mathcal{J}}_2 < 0$ (or in vector case, $\tilde{\mathcal{J}}_2$ not positive definite)

Example 1. Take $S_T = 1$, then $\tilde{\mathcal{J}}_2 = \hat{\gamma}_0 + 2\hat{\gamma}_1$. In small samples, we may find $\hat{\gamma}_1 < -1/2\hat{\gamma}_0$.

2.3 Weighted, truncated sum of sample covariances

$$\hat{J} = \sum_{j=-S_T}^{S_T} k_T(j)\hat{\gamma}_j$$

We need conditions on S_T and $k_T()$ to give us consistency and positive-definiteness.

Remark 2. $k_T()$ is called a kernel.

Behavior of S_T gives consistency – need $S_T \to \infty$, but slower than $T \to \infty$.
$k_T()$ guarantees positive-definiteness by down-weighting high lag covariances. Also need $k_T() \to 1$ for consistency.

2.3.1 Consistency

Theorem 3. *$\hat{\mathcal{J}}$ is consistent if we assume:*

- $\sum_{-\infty}^{\infty} |\gamma_j| < \infty$
- $k_T(j) \to 1$ *as* $T \to \infty$ *and* $|k_T(j)| < 1 \ \forall j$
- $\xi_{t,j}$ *(defined below) are sationary for all* j *and* $\sup_j \sum_k |Cov(\xi_{t,j}, \xi_{t+k,j}| < C$ *for some constant* C *(limited dependence)*
- $S_T \to \infty$ *and* $\frac{S_T^3}{T} \to 0$

Proof. This is an informal "proof" that sketches the ideas, but isn't completely rigorous.

$$\hat{\mathcal{J}} - \mathcal{J} = -\sum_{|j|>S_T} \gamma_j + \sum_{j=-S_T}^{S_T} (k_T(j) - 1)\gamma_j + \sum_{j=-S_T}^{S_T} k_T(j)(\hat{\gamma}_j - \gamma_j$$

We can interprete these three terms as follows;

1. $\sum_{|j|>S_T} \gamma_j$ is truncation error
2. $\sum_{j=-S_T}^{S_T}(k_T(j) - 1)\gamma_j$ is error from using the kernel
3. $\sum_{j=-S_T}^{S_T} k_T(j)(\hat{\gamma}_j - \gamma_j$ is error from estimating the covariances

Terms 1 and 2 are non-stochastic. They represent bias. The third term is stochastic; it is responsible for uncertainty. We will face a bias-variance tradeoff.

We want to show that each of these terms goes to zero

1. Disappears as long as $S_T \to \infty$, since we assumed $\sum_{-\infty}^{\infty} |\gamma_j| < \infty$.
2. $\sum_{j=-S_T}^{S_T}(k_T(j) - 1)\gamma_j \leq \sum_{j=-S_T}^{S_T} |k_T(j) - 1||\gamma_j|$ This will converge to zero as long as $k_T(j) \to 1$ as $T \to \infty$ and $|k_T(j)| < 1 \ \forall j$.
3. Notice that for the first two terms we wanted S_T big enough to eliminate them. Here, we'll want S_T to be small enough.

   First, note that $\hat{\gamma}_j \equiv \frac{1}{T}\sum_{k=1}^{T-j} z_k z_{k+j}$ is not unbiased. $E\hat{\gamma}_j = \frac{T-j}{T}\gamma_j = \tilde{\gamma}_j$. However, it's clear that this bias will disappear as $T \to \infty$.

Let $\xi_{t,j} = z_t z_{t+j} - \gamma_j$, so $\hat{\gamma}_j - \tilde{\gamma}_j = \frac{1}{T}\sum_{\tau=1}^{T-j} \xi_{\tau,j}$. We need to show that the sum of $\xi_{t,j}$ goes to zero.

$$\begin{aligned} E(\hat{\gamma}_j - \tilde{\gamma}_j)^2 =& \frac{1}{T^2}\sum_{k=1}^{T-j}\sum_{t=1}^{T-j} \text{Cov}(\xi_{k,j}, \xi_{t,j}) \\ \leq& \frac{1}{T^2}\sum_{k=1}^{T-j}\sum_{t=1}^{T-j} |\text{Cov}(\xi_{k,j}, \xi_{t,j})| \end{aligned}$$

We need an assumption to guarantee that the covariances of ξ disappear. The assumption that $\xi_{t,j}$ are sationary for all j and $\sup_j \sum_k |\text{Cov}(\xi_{t,j}, \xi_{t+k,j})| < C$ for some constant C implies that

$$\frac{1}{T^2}\sum_{k=1}^{T-j}\sum_{t=1}^{T-j} |\text{Cov}(\xi_{k,j}, \xi_{t,j})| \leq \frac{C}{T}$$

By Chebyshev's inequality we have:

$$P(|\hat{\gamma}_j - \tilde{\gamma}_j| > \epsilon) \leq \frac{E(\hat{\gamma}_j - \tilde{\gamma}_j)^2}{\epsilon^2} \leq \frac{C}{\epsilon^2 T}$$

Then adding these together:

$$\begin{aligned} P(\sum_{-S_T}^{S_T} |\hat{\gamma}_j - \tilde{\gamma}_j| > \epsilon) \leq& \sum_{-S_T}^{S_T} P(|\hat{\gamma}_j - \tilde{\gamma}_j| > \frac{\epsilon}{2S_T + 1}) \\ \leq& \sum_{-S_T}^{S_T} \frac{E(\hat{\gamma}_j - \gamma_j)^2}{\epsilon^2}(2S_T + 1)^2 \\ \leq& \sum_{-S_T}^{S_T} \frac{C}{T}(2S_T + 1)^2 \approx C_1 \frac{S_T^3}{T} \end{aligned}$$

so, it is enough to assume $\frac{S_T^3}{T} \to 0$ as $T \to \infty$.

□

2.3.2 Positive Definiteness

Under appropriate conditions on the kernel, $k_T(j)$, the estimate of the long run variance is guaranteed to be positive definite.

Assume $k_T(j)$ is an inverse Fourier transform of $K_T(l)$, *i.e.*

$$k_T(j) = \sum_{l=-(T-1)}^{T-1} K_T(l) e^{-i\frac{2\pi jl}{T}}$$

Lemma 4. *$\hat{\mathcal{J}}$ is non-negative with probability 1 if and only if $K_T(l) \geq 0$ and $K_T(l) = K_T(-l)$*

Common Kernels

Definition 5. *Bartlett kernel* $k_T(j) = k(j/S_T)$ where $k(x) = \begin{cases} 1 - |x| & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$

Newey-West (1987) (this is one of the most cited papers in economics)

Definition 6. *Parzen kernel* $k(x) = \begin{cases} 1 - \sigma x^2 - \sigma|x|^2 & 0 \leq x \leq 1/2 \\ 2(1 - |x|)^3 & 1/2 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Keifer & Vogelsang Keifer & Vogelsang (2002) consider setting $S_T = T-1$. This gives $\hat{\mathcal{J}}$ inconsistent (it converges to a distribution). However, $\hat{\mathcal{J}}$ usually isn't what we care about. We care about testing $\hat{\beta}$, say by looking at the t statistic. We can use $\hat{\mathcal{J}}$ with $T = S_T$ to compute $t = \frac{\hat{\beta}}{se(\hat{\beta})}$, which will converge to some (non-normal) distribution without any nuisance parameters, and we can use this distribution for testing. The motivation for doing this is that Newey-West often works poorly in small samples.

2.4 Parametric HAC Estimation

Assume z_t is AR(p)
Estimate by OLS

$$z_t = a_1 z_{t-1} + ... + a_p z_{t-p} + e_t$$

then use $\hat{a}(L) = 1 - \hat{a}_1 - .. - \hat{a}_p$ to construct $\hat{\mathcal{J}}$,

$$\hat{\mathcal{J}} = \frac{\hat{\sigma}^2}{\hat{a}(1)^2}$$

where $\hat{\sigma}^2 = \frac{1}{T}\sum \hat{e}_t^2$

Two questions:

- What p? – model selection criteria, BIC (Bayesian informaiton criteria)
- What if z_t is not AR(p)?

The second question is still an open question. Den Haan and Levin (1997) showed that if z_t is AR(p), then the convergence of the parameteric estimator is faster than the kernel estimator described below.

2.5 Prewhitening

Nonparametric HAC performs poorly when the series is persistent. Parametric HAC performs poorly if the model is wrong. Prewhitening combines the two. From the above we know that if e_t is white noise with variance Σ, then when $A(L)z_t = B(L)e_t$, the long-run variance of z_t is

$$\mathcal{J}_z = A(1)^{-1}B(1)\Sigma B(1)'A(1)^{1'}$$

Similarly if e_t is not white noise, but has long-run variance $\mathcal{J}_e$, then

$$\mathcal{J}_z = A(1)^{-1}B(1)\mathcal{J}_e B(1)'A(1)^{1'}$$

The prewhitened nonparametric estimate of $\mathcal{J}_z$ is then simply:

$$\hat{\mathcal{J}}_z = \hat{A}(1)^{-1}\hat{B}(1)\hat{\mathcal{J}}_e\hat{B}(1)'\hat{A}(1)^{1'}$$

where $\hat{A}$ and $\hat{B}$ are estimated by OLS or Kalman filtering, and $\hat{J}_e$ is estimated by doing nonparametric HAC hat $\hat{e}_t$.

Practical Advice This summer, Mark Watson gave a lecture on HAC
http://nber15.nber.org/c/2008/si2008/TSE/Lecture9.pdf
and this is a short summary of what he recommended. When doing HAC, you have to choose which of the three methods to use, and then if you choose ARMA, the lag lengths, or if you choose nonparametric, the kernel and bandwidth. In this discussion, the goal is to do inference on $\hat{\beta}$

- Simulations show large size distortions for all methods (reject at 5% level far more than 5% of time). Tests work worse when
  - Sample size is smaller
  - Data is more persistent (e.g. an AR(1) with coefficient near one)

- If it is the correct model, parametric ARMA works best. Sometimes theory suggests an ARMA (den Haan and Levin 1997).
- Kiefer-Vogelsang leads to smaller size distortions, but has less power than kernel methods
- For kernel methods:
  - The theoretically optimal[1] kernel is called the quadratic-spectral (QS) kernel. In practice, all common kernels perform similarly.
  - For inference, it is not necessarily best to minimize MSE of $\hat{\mathcal{J}}$
    * See Sun, Philips, and Jin (2008) for a more formal discussion
    * Intuition: suppose $z \sim N(\mu, \sigma^2)$ (think of z as $\sqrt{n}(\beta - \hat{\beta}_0)$) and $\hat{\sigma}^2$ is an estimate of σ^2. For testing $H_0 : \mu = 0$ at level α, we would compute a critical value, c, from the normal distribution such that $P(|z/\sigma| < c) = \alpha$. If we don't know σ, then this how well this test would work depends on how close $P\left(\frac{z^2}{\hat{\sigma}^2} < c^2\right)$ is to $P\left(\frac{z^2}{\sigma^2} < c^2\right)$. Very loosely:

$$P\left(\frac{z^2}{\hat{\sigma}^2} < c^2\right) = E\left[\mathbf{1}(z^2 < \hat{\sigma}^2 c^2)\right] = E\left[g(\hat{\sigma}^2)\right] \tag{1}$$

$$\approx E\left[g(\sigma^2) + (\hat{\sigma}^2 - \sigma^2)g'(\sigma^2) + \frac{1}{2}(\hat{\sigma}^2 - \sigma)^2 g''(\sigma^2)\right] \tag{2}$$

$$\approx Eg(\sigma^2) + Bias(\hat{\sigma}^2)g' + \frac{1}{2}MSE(\hat{\sigma}^2)g'' \tag{3}$$

      So the error in the test depends on a combination of the bias and MSE of $\hat{\mathcal{J}}$. The best choice of S_T for testing shouldn't minimize MSE; it should minimize this combination of bias and MSE. Since bias decreases with S_T, the best S_T for testing is greater than the best S_T for MSE.
    * Andrews (1991) gives formula for minimal MSE choice of S_T.
      · For inference: use larger S_T
      · For a GMM weighting matrix, minimal MSE seems like a good choice
    * Similar reasoning suggests (maybe) using a longer lag length for an ARMA model than suggested by BIC or AIC.

[1] In the sense that it minimizes MSE of $\hat{\mathcal{J}}$