Problem set 6
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. Thinking about Monte Carlos, samples, EDFs, and urns. Throughout this problem set, drawing from a urn is taken to mean drawing uniformly from the items in the urn.

   Given a sample of data, Y, $(T \times 1)$, the empirical distribution function (EDF) is defined as

   $$EDF(c) = \frac{\#(y_t < c)}{T}$$

   where # means 'number of.' That is, the EDF is the share of the ys that are less than c.

   (a) Give some conditions under which the EDF will be consistent for the true CDF. That is, $EDF(c)$ converges to the true $F(c)$ with sample size for any fixed c.

   Answer/comment

   I have not been specific about what sort of convergence we are looking for, so let's take an easy one: $EDF(c) \rightarrow_p F(c)$. Define the random variable $z_{c,t}$ to be 1 if $y_t < c$, 0 otherwise. $EDF(c)$ is the sample mean of $z_{c,t}$ and $F(c)$ is the population mean. Thus, we only need that z_t satisfies a WLLN. Note: Given a process for y, if the implied z satisfies a WLLN for some c, it will satisfy it for all c.

   (b) Suppose we have a sample from an iid random variable Y $(T \times 1)$. We write each value on a piece of paper, throw them in an urn and form new samples by drawing with replacement from the urn.

What is the relation between the EDF and the exact distribution function for a single draw from this urn?

Answer/comment

They are the same. The distribution induced by the urn described places mass $1/T$ on each sample value, as the EDF does.

(c) Given the vector `Y`, how would you compute $EDF(c)$ in Matlab?

Answer/comment

```
function p=EDF(c,Y); p = mean(Y<c);
```

(d) You find yourself in Matlab with only acess to the `mean` and `randn` functions. Write a single command that returns the value of the standard normal CDF at c.

Answer/comment

```
p = mean(randn(10000,1)<c)
```

The point of these last two questions is to emphasize how simple theory and Matlab programs become if one views probabilities as the mean of indicator variables.

---

2. More urns.

Our data are $Y \sim P_\theta$, where Y is $(T \times 1)$ and the elements of Y are iid. We are interested in the distribution of a scalar statistic $\psi(Y)$, but P_θ is unknown. Specifically, we are looking for a good approximation to

$$\pi = P_{\theta^*}(\psi(Y) > c)$$

under an unknown P_{θ^*}.

We follow the urn procedure desribed in problem 1, and draw a zillion samples of size T. We compute $\psi(Y)$ on each sample, and use the EDF of the resulting ψs as our proxy for the true distribution.

Formally or informally give an argument as to why $\hat{\pi} = 1 - EDF(c)$ will be consistent for π?

Answer/comment

We are attempting to assess

$$\pi = \mathrm{pr}_\theta(\{Y | \psi(Y) > c\})$$

where Y is $T \times 1$.

Define the indicator variable: $z = 1$ if $\psi(Y) > c$ and zero otherwise.

Suppose we have a large number, J of samples of size T drawn according one particular EDF. You should be able to make the argument that $\hat{\pi}$, the sample mean of the z across the J samples, will converge to truth under the EDF, call it $\pi(EDF)$:

$$\hat{\pi} \rightarrow_p \pi(EDF)$$

In other words, so long as we do enough Monte Carlo repetitions, our estimate of π will be arbitrarily close to the true π under the DGP generating the data.

Suppose we had a large number, J, of samples of size T drawn according to θ^*. By the same argument,

$$\hat{\pi} \rightarrow_p \pi(\theta^*)$$

Thus, we just need to figure out an argument that $\pi(EDF) \rightarrow \pi(\theta^*)$. This convergence will happen as $T \rightarrow \infty$. Of course, $\pi()$ may change with T, so let's write $\pi_T(\theta^*)$ and $\pi_T(EDF_T)$. We need to show that *for some sense* of $|.|$ and $\rightarrow$:

$$|\pi_T(\theta^*) - \pi_T(EDF(Y_T)| \rightarrow 0$$

Remember that the EDF is given by the sample and that we need the result to hold with arbitrarily high probability. That is, as T gets large, we want that the probability of the set of samples, Y_T for which $|\pi_T(\theta) - \pi_T(EDF(Y_T))|$ is small to go to 1. Define the event,

$$\mathcal{Y}_T = \{Y_T || \pi_T(\theta) - \pi_T(EDF(Y_T))| < \delta\}$$

For all δ AND all $\theta \in \Theta$, we'd like

$$\text{pr}_\theta(\mathcal{Y}_T) \rightarrow_T 1$$

Note that this must hold for all θ because we don't know which one is true. Thus,no matter which θ is true, the result holds.

In short, our desired result will be achieved if as when the sample size gets large, it is very odd to get a sample with an EDF that doesn't look like the true underlying distribution function, F_{θ^*}. Formal proofs show that this condition holds fairly generally.

I have deliberately discussed this in an informal way without emphasizing particulars about forms of convergence. See the texts referenced for more detailed arguments.

The one thing I wanted you to puzzle over is the role of growing T and how the EDF, which is chosen based on the sample, 'converges' to the true distribution function. Starting from any distribution function Y_Ts looking like $Y_T \in \mathcal{R}^T$ *could* be drawn. As the sample size gets large, the mass on Ys with implied EDFs that do not look like F_{θ^*} falls. All mass is concentrated on Ys with EDFs closely matching the F_{θ^*}.

---

3. More urns, again. $Y \sim P_\theta$, and we want to know the distribution of, $\psi(Y)$, where Y is $(T \times 1)$. We know $Ey_t = 0$ and that y_t follows a covariance stationary MA(1) driven by iid shocks, but we don't know the shock distribution.

   Consider the following 'urn' procedure. Partition the T observations into B blocks of b contiguous observations. (Suppose that we are only interested in samples of sizes satisfying $T = Bb$ for integer B, b. This integer-constraint issue is a bit of a pain in practice, but not for the basic theory argument).

   Write each block on a piece of paper, throw them in the urn, and form new samples of size T by drawing B blocks with replacement.

   Once again, draw a zillion such samples, compute ψ on each, and use the resulting EDF as a proxy for the true CDF of ψ under the unknown P_θ.

   Formally or informally, give an explanation as to why the EDF of $\psi(Y)$ will be a good approximation to the true distribution of ψ under P_θ as T, B, and b get large.

   Answer/comment
   First, you get bonus points for remembering that there is an observational equivalence problem with MA(1)s. That is, the root of θ and $1/\theta$ generate identical data. Thus, ψ may not be identified if it depends on the true value of the root. Let's set aside this issue, say, by assuming that ψ is the same for any two observationally equivalent θs. That is, $\psi(\theta_1) = \psi(\theta_2)$ whenever θ_1 and θ_2 are observationally equivalent. O.K., let's ignore this detail henceforth,.

   This problem pushes the previous problem just a bit further. We now

need to show that we can make the same argument as above that $|\pi_T(BRS_T) - \pi(\theta^*)|$ goes to zero, where BRS_T is the distribution function implied by the Block Resampling Scheme.

I mainly wanted you to think about the senses in which BRS, the distribution function implied by the block resampling scheme will look like the true distribution function. Try the following intuition.

In the true DGP, any two nonadjacent observations are independent and have the same unconditional distributions. Any two adjactent observations under the true DGP have whatever dependence is implied by the MA(1) structure and shock distribution.

Let's see how this is approximately true in the block resample DGP, as B, b, and T get large.

Take any two nonadjacent observations, say observation t_1 and t_2. By definition, these will be, say, the i_1^{th} observation of the j_1^{th} block and the i_2^{th} observation of the j_2^{th} block. Clearly, unless $j_1 = j_2$, that is, unless we happen to pick two observations from different instances of the same block, the two observations will be independent. As the number of blocks gets large, the share of pairs of observations coming from the same block gets small. Of course, our two observations will only be dependent if $j_1 = j_2$ AND $|i_1 - i_2| < 2$—that is the observations are from the same location in the same block. As the blocksize gets large, this also becomes unlikely. Informally, then, any two nonadjacent items drawn at random will be nearly independent.

Take any pair of adjacent observations from the block resampling distribution. So long as these do not span a block boundary, the two observations will will have been adjacent in the original sample. Thus, they will bear the same dependence as any two observations in the original sample. Ignoring the integer constraint, the number of observation pairs spanning boundaries is approximately T/B. The share of pairs spanning boundaries is $(T/B)/(0.5T) = 1/(2B)$. So long as this goes to zero (which we have assumed), almost all pairs of obervations will behave like pairs from the original sample.

The key here more generally is that the blocksize gets large relative to the span of dependence in the DGP. In the MA(1) case, that span is 2 observations. In the AR case, dependence never dies entirely. We need that the blocksize gets large relative to the span of 'nontrivial' dependence. So long as the data are stationary, dependence must die out and we can get a blocksize big enough. Thus, we want, have T,

sample size, b block size, and B number of blocks composing a samples. And (ignoring integer constraints) $T = Bb$. We want,

1) A blocksize, b, that is large relative to the span of nontrivial dependence.

Our resampling scheme will destroy any dependence that spans periods larger than b, so this dependence that is destroyed must be trivial.

2) A number of blocks that compose a sample, B, that is large.

We are drawing B blocks from the urn. We always want a large number of draws from the urn in order to apply LLNs.

3) A number of blocks composing the sample, B that is small relative to T.

Block boundaries will be a problem since obervations in our new samples that span boundaries will be (nearly) independent and, thus, won't look like the true DGP. Thus, we need boundaries to become rare in our re-sampled samples.

If we have all these properties, then we might expect that the re-samples will look like they are drawn from the original distribution, and the probability of samples Y such that $|\pi(BRS(Y)) - \pi(\theta^*)|$ is small will be converging to 1.

---

4. One more urn. Suppose the model explains $Z = [Y\,X] \sim P_\theta$ where Y and X are $(T \times 1)$ and the data have no time dependence (the rows of Z are mutually independent). You'd like to understand the distribution of the OLS $\hat{\beta}$ from a regression of Y on X.

   You write each observation (each row of Z) on a piece of paper. Put them in an urn and draw a zillion samples of size T, drawing with replacement. You compute $\hat{\beta}$ on each and take the EDF of the resulting $\hat{\beta}$s as a proxy for the exact distribution of $\hat{\beta}$.

   Formally or informally give an argument as to why the EDF of the $\hat{\beta}$ will converge to the distribution of $\hat{\beta}$ under the true DGP.

   Answer/comment

   More of the same as for the last two problems. We are interested in estimating $\text{pr}_{\theta^*}(\hat{\beta} < c)$ for arbitrary c. We can define the event $\mathcal{Z} = \{Z|\hat{\beta} < c\}$, and follow the argument from question xxxx. The only subtlety is that our data were iid and $T \times 1$ in that question. Here, the data are iid and $T \times 2$. This difference has no effect on

the argument. So long as the item resampled (whether scalar or more complicated) is iid in the original DGP, the argument will go through.

---

5. The subject of the classroom presentation: Suppose you have $Z = [YX]$ $(T \times 2)$ where $(y_t, x_t) \sim iidN(0, \Omega)$ with Ω full rank. Define the correlation of x and y to be ρ and $\hat{\rho}$ to be the natural sample estimate—the sample covariance divided by the square root of the product of the two sample variances.

   (a) Show that $\hat{\rho}$ is consistent.

   Answer/comment

   You should be able to show that both the sample covariance and the the sample variances are consistent. Since $\hat{\rho}$ is a smooth function of these, it also consistent.

   (b) Show that $\sqrt{T}(\hat{\rho} - \rho) \overset{a}{\sim} N(0, \sigma^2)$ for some σ^2 and give an expression for σ^2 in terms of the parameters driving the Zs.

   Answer/comment

   Note that $\hat{\rho}$ is a function of sample sums, $(\sum x, \sum y, \sum xy \sum x^2, \sum y^2)$ and you know how to find the joint distribution of sample moments (noralized sample sums). The normalized vector just given is jointly normal under mild assumptions. Thus you just need to derive the variance-covariancematrix of the vector and then use the delta method to get the sample variance. This is just a little bit tedious, but it once may be worthwhile. Or just look it up. The variance is $(1 - \rho^2)^2$.

   (c) Run a Monte Carlo for various sample sizes and values of ρ. For an interesting assortment of T and ρ show a histogram approximating the distribution of $\hat{\rho}$. (Print a matrix of histograms, each row a different T; each column a different ρ). Print analogous tables in which each cell gives the mean, variance, and skewness of $\hat{\rho}$ for different (T, ρ) pairs.

   (d) Show that the t-statistic, $\tau = \sqrt{T}(\hat{\rho} - \rho_0)/\hat{\sigma}$, is asymptotically $N(0, 1)$ when constructed with ρ_0 equal to the true ρ.

   Answer/comment

   This is trivial based on the asymptotic normality from above.

(e) Using the same Monte Carlo as above, compute the distribution of the marginal significance of the t-test just described. More specifically, on each Monte Carlo sample, compute the marginal significance of both one-tailed tests (left tail and right tail) and the two-tailed test. Produce a matrix of histograms of those p values.

Note: On each sample, compute the statistic $\hat{\tau}$ basing it on the true ρ for that sample. The marginal significance under the three different tests are then,

$$\text{pr}(z < \hat{\tau}), \text{pr}(z > \hat{\tau}), \text{ and } \text{pr}(|z| > |\hat{\tau}|)$$

where $\hat{\tau}$ is the value of the test statistic on the sample and z is a standard normal random variable. In these xpressions, z is stochastic and $\hat{\tau}$ is taken as fixed at the sample value. Intuition: for the right-tail test, you compute the probability that an $N(0,1)$ variable would be bigger than the value of the statistic you observe on the sample.

What do you see?

Answer/comment

Note: For any test with properly calibrated size, under the null hypothesis, the distribution of the p-values should be uniform on $[0,1]$. You should ponder this until it makes sense.

Consider the following. Take any distribution function, F. If $u \sim U[0,1]$ (uniform on 0, 1), then

$$z = F^{-1}(u) \sim F$$

where $\sim F$ means, 'is distributed according to the distribution described by F'. t

Indeed, this is one standard way of generating random variables on the computer. For example, in STATA a common way to generate standard normals is `z = invnorm(uniform())` .

By the same token, if z F, then

$$u = F(z) \sim U[0,1]$$

Of course, when z is a test statistic and F is its distribution, then this says that p values should be uniformly distributed. (Assuming that the p-value is either $F(z)$ or $1 - F(z)$—that is, we either reject for large or for small values.)

Thus, one way to represent the calibration of a test is to run Monte Carolos under one or more DGPs consistent with the null, compute the p-value for each draw and then check the histogram of the p-values against a uniform.

Alternatively you could check the empirical distribution function of the $p-values$ against a 45 degree line on the box from 0,0 to 1,1.

(f) What is Fisher's-z transform in this context; describe informally why it makes sense.

Notes: The distribution of estimates of correlation has been discussed immensely since the early days of statistics. This may be the simplest case that gives rise to versions of many of the complexities that drive us to advanced methods. It provides a nice context in which to think about the stochastic expansions we will be discussing and their practical importance. If you are into theory, mastering this literature will give you sense of the topics we'll be grappling with.

A very nice summary of the highlights in this area is provided in the introduction of Ogasawara (2006) cited below. If you read the article, you'll also note how explicit expressions in the stochastic-expansions literature get pretty messy even in simple cases.

In the next problem set, we'll show that the bootstrap can at times capture higher order terms in expansions using Monte Carlo methods of the type you played with in the earlier questions. The Monte Carlo methods are conceptually straightforward and easy to implement, which explains why the bootstrap has become such an important technique in practice.

Haruhiko Ogasawara, Asymptotic expansion of the sample correlation coefficient under nonnormality, Computational Statistics & Data Analysis, Volume 50, Issue 4, 24 February 2006, Pages 891-910,

http://www.sciencedirect.com/science/article/pii/S0167947304003251

Answer/comment

Let me make just one comment here.

We derived that

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

where $V(\rho) = (1-\rho^2)^2$.

Let's take an arbitrary function g. By the delta method, we have that $g(\hat{\rho})$ will have distribution,

$$\sqrt{T}(g(\hat{\rho}) - g(\rho)) \overset{a}{\sim} N(0, G(\rho)^2 V(\rho))$$

where $G(q)$ is a scalar equal to $\partial g/\partial\rho|_{\rho=q}$.

Suppose we want to pick a g such that the asymptotic variance does not vary with ρ. We'll need a function with first derivative satisfying,

$$G^2(\rho)V(\rho) = const.$$

You'll see that Fisher's-z transform fills the bill. Thus, it answers the question, 'what simple transform of ρ will lead to a statistic with asymptotic variance independent of ρ.

There was a time when folks spent time thinking about which monotonic transformation g would most improve the quality of our asymptotic approximations. This has largely been supplanted by other methods and you don't hear a lot about these any more.
