University of Illinois
Spring 2017

Econ 574

Department of Economics
Roger Koenker

Lecture 13
"Monte-Carlo II – Design and Execution"

In a previous lecture we have considered some elementary aspects of random variable generation. Here we consider questions of experimental design. A general reference is DM - §21. An important aspect of simulation experiments is their *reproducibility.* If you look at my web page you will see a link to some references relevant to this; in particular there is a protocol for reproducible simulations that I've tried to follow in my own work, and there is a joint paper that I wrote with Achim Zeileis that appeared in the J. of Applied Econometrics that describes some useful tools for achieving reproducibility. We will begin by discussing general objectives of MC. First, focusing on estimation then on testing.

Estimation in Monte-Carlo simulations is usually a horse race. We have several model configurations and we want to know which of several competing estimators do best.

Q1 How to define best?

A1 MSE, often helpful to decompose bias and variance effect.

A2 MAE, sometimes useful particularly when estimators have long tails.

A3 "Pitman closeness" is another criterion, but rarely used in econometrics.

In Pitman (1937) an estimator $\hat{\theta}_n$ is defined to be a "closer" estimator of θ_0 than $\tilde{\theta}_n$ iff,

$$P[|\hat{\theta}_n - \theta_0| < |\tilde{\theta}_n - \theta_0|] > \frac{1}{2}$$

"More than half the time $\hat{\theta}$ is closer to θ_0 than $\tilde{\theta}_n$.

This criterion yields the posterior median as a Bayes estimator.

Example: Problem posed by Schrödinger to Geary in Dublin during WWII[*]

> In a town, cars are numbered 1, 2, $\dots, m$. The numbers on a sample of n cars are noted. Find the *closest* estimate of m.

Solution: $P(x_{(n)} < mc) \approx c^n$ so, let $\hat{m} = 2^{1/n} x_{(n)}$, and note

$$P(\hat{m} < m) = P(x_{(n)} < \frac{m}{2^{1/n}}) = \left(\frac{1}{2^{1/n}}\right)^n = \frac{1}{2}$$

so $\hat{m}$ is *median* unbiased for m. Half the time it is too big and half the time it is too small. This problem is related to the literature about capture-recapture models that are extensively used in biology to estimate population sizes. See for example Efron (2003) "Robbins, Empirical Bayes and Microarrays" in the *Annals of Statistics*, where the second section reviews some earlier literature on this problem and discusses methods for estimating the size of Shakespeare's vocabulary.

[*] There is a good novel about this whole episode in Schrödinger's life written by Neil Belton, called *A Sharpening of the Knives.* Bill Farebrother has written a paper with some more historical details on this problem.

Q2 How to estimate MSE?

A1 Approximate $\int(\hat{\theta}_n - \theta_o)^2 dF$ where $\hat{\theta}_n(X_1, \ldots, X_n)$ and F is the joint *df* of X, by,

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_n(x_1^i, \ldots, x_n^i) - \theta_0)^2$$

A2 But refinements are frequently worthwhile. I will discuss antithetic variables first. The idea is simple: we want to estimate something, say $\mu = E(\hat{\theta} - \theta_0)$, we have some way to compute estimates $\hat{\mu}_i$, (A1) suggests the naive estimator

$$\frac{1}{m}\sum_{i=1}^{m}\hat{\mu}_i \rightarrow \mu$$

But if we could find another estimator $\check{\mu}_i$ which was negatively correlated with $\hat{\mu}_i$ then since

$$V(\bar{\mu}) \equiv V(\frac{1}{2}(\hat{\mu}_i + \check{\mu}_i)) = \frac{1}{4}(V(\hat{\mu}_i) + V(\check{\mu}_i) + 2 \text{ Cov } (\hat{\mu}_i, \check{\mu}_i))$$

If Cov were less than zero, then this would be less than we would get by performing two independent experiments. A rather *silly example from DM*

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta_0 + u) \\ \check{\beta} &= (X'X)^{-1}X'(X\beta_0 - u)\end{aligned}$$

Thus,

$$\bar{\beta} = \frac{1}{2}(\hat{\beta} + \check{\beta}) = \beta_0$$

The estimators are perfectly negatively correlated and therefore, we have no noise whatsoever in $\bar{\beta}$. In non linear models the above technique of using IU isn't perfect but does reduce error somewhat. Another good example is adding vertical lines to the horizontal ones in the Buffon example.

A3 Control Variates – this is generally more useful. We condition on something clever and can thereby sometimes drastically reduce the variability.

Recall

$$E(E(X|W)) = EX$$

and

$$V(E(X|W)) + E(V(X|W)) = V(X)$$

so

$$V(X) \geq V(E(X|W))$$

Thus averaging $E(X|W)$ is a better way to estimate EX than just averaging X itself.

Example 1 Simon *Applied Statistics* (1976). Suppose $W \sim$ Poisson (λ) and $X \sim$ Beta (W, W^2+1) and that we would like to estimate EX. Observe that

$$E \text{ Beta } (v, w) = v/v + w$$

so

$$E(X|W) = W/(W^2 + W + 1)$$

so to estimate EX we average $E(X|W)$ i.e., compute

$$T = m^{-1} \sum_{i=1}^{m} W_i/(W_i^2 + W_i + 1)$$

To evaluate gain we would need to compute

$$V(E(X|W)) \text{ and } E(V(X|W))$$

we could estimate the former by sample variance of the $E(X|W)$'s and the latter using the fact that

$$V(X|W) = W(W^2+1)/((W^2+W+1)^2(W^2+W+2))$$

Example 2 An extremely useful example of this strategy is the normal/ind. swindle. A rather large class of problems can be handled in this way.

2.1 Let $X_1, \ldots, X_n$ be iid $\mathcal{N}(\mu, 1)$ and $\tilde{\theta} =$ median $\{X_i\text{'s}\}$. Find $V(\tilde{\theta})$. Naively, we could compute

$$T = \frac{1}{m} \sum_{i=1}^{m} (\tilde{\theta}_n^i - \bar{\theta}_n)^2$$

But here we know that the optimal estimator is $\hat{\theta}_n = \bar{X}_n$. Note

$$\tilde{\theta}_n = \hat{\theta}_n - (\hat{\theta}_n - \tilde{\theta}_n)$$

and these two pieces are independent by Problem 4 on Problem Set 3. Thus,

$$V(\tilde{\theta}_n) = V(\hat{\theta}_n) + V(\tilde{\theta}_n - \hat{\theta}_n)$$

and here $V(\hat{\theta}_n) = 1/n$ so we can compute average of the $\tilde{\theta}_n - \hat{\theta}_n$ to get more efficient estimate.

2.2 What to do in non-normal cases? Often we can condition on a GLS estimator.

A large class of *df*'s may be generated as $Y = X/Z$ where X is normal and Z is independent r.v. so $Y|Z \sim \mathcal{N}(0, Z^{-2})$.

Examples 1) $Z = \delta_{\sigma^{-1}} \Rightarrow Y \sim \mathcal{N}(0, \sigma^2)$

2) $Z = (1-\varepsilon)\delta_1 + \varepsilon\delta_{\sigma^{-1}}$ $Y \sim$contaminated normal.

3) $Z = |\mathcal{N}(0,1)|$ $\quad$ $Y \sim$ Cauchy.

4) $Z = \sqrt{x_q^2/q}$ $\qquad Y \sim$ Student.

5) $Z \sim U[0,1]$ $\qquad Y \sim$ Slash (extreme version of Cauchy)

See Efron & Olshen, *Annals* (1978), "How broad is the class of normal scale mixtures?"

Now, in general, if we say Y is Cauchy, we wouldn't know Z so we couldn't condition, but in MC if we generate Y, this way we *can know* Z and we can compute estimators conditional on Z, e.g., in the linear model

$$Y_i = x_i\beta + u_i$$

we can think of $u_i \sim v_i/w_i$ where $v_i \sim \mathcal{N}(0,1)$ and $w_i \sim \sqrt{\chi_q^2/q}$ conditional on w's we know that

$$\hat{\beta} = (X'WX)^{-1}X'Wy$$

where $W = \text{diag}\ (w_i^2)$ is efficient and has covariance matrix $(X'WX)^{-1}$ so we can use this as control variate just as in the median case, i.e., let

$$\tilde{\beta} = \hat{\beta} - (\hat{\beta} - \tilde{\beta})$$

with

$$V(\tilde{\beta}) = V(\hat{\beta}) + V(\hat{\beta} - \tilde{\beta})$$

we know $V(\tilde{\beta})$ and we estimate the 2nd bit by

$$m^{-1}\sum(\hat{\beta}_i - \tilde{\beta}_i)(\hat{\beta}_i - \tilde{\beta}_i)'$$

which, if $\tilde{\beta}_i$ is *good*, should be *small.* Actually, $E(X'WX)^{-1}$ may be nontrivial to compute, but if one is comparing several estimators for which this component is the same one need not bother with it.

MC For Testing

Inference MC is somewhat more challenging. Some general rules:

1. Contrary to common practice, correct size is rarely sufficient, power is important aspect of MC evaluation of tests.

2. Power comparisons of tests with different sizes are worthless. But it is generally difficult to find a good *general* strategy for size correction.

3. Local alternatives offer a convenient way to design power comparisons for varying n, i.e., set $\theta_n = \theta_0 + \delta/\sqrt{n}$ with fixed δ, this mimics asymptotic considerations.

Monte-Carlo Swindles for Testing

Ref.: Gross (1977, *JASA*), Koenker and Portnoy (1987, *JASA*, L-estimation)

This is really an extended example to illustrate the use of control variates for evaluation of tests and CI procedures. The setup is like before with

$$y_i = x_i'\beta + u_i \qquad\qquad u_i = z_i/v_i$$

where $z_i \sim \mathcal{N}(0,1)$ and iid, v_i is independent. Thus the efficient estimator is

$$\hat{\beta} = (X'WX)^{-1}X'Wy \qquad\qquad W = \text{ diag } (v_i^2)$$

Consider any other translation equivariant estimator $\tilde{\beta}$, i.e.,

$$\tilde{\beta}(y + X\gamma) = \tilde{\beta}(y) + \gamma$$

and scale equivariant estimator s^2, i.e.,

$$s(\sigma y) = \sigma s(y)$$

The scale estimator s will be used to studentize $\tilde{\beta}$, typically it would be an estimator of the asymptotic variance of $\tilde{\beta}$.

Let $\tilde{\alpha} = c'\beta$ be an arbitrary linear contrast, then

$$P(\tilde{a} > ks) = 1 - \Phi((ks - \tilde{\alpha} + \hat{\alpha})/\sigma_c)$$

where $\hat{a} = c'\hat{\beta}$ and $\sigma_c = c'(X'WX)^{-1}c$. Why? Because, $\alpha_0 = c'\beta$, and

$$(*) \qquad \hat{\alpha} \sim \mathcal{N}(\alpha_0, \sigma_c^2)$$

Gross Argument

$$\begin{aligned} P(\tilde{a} > ks) &= P(\tilde{a} - \hat{a} > ks - \hat{a}) \\ &= P(\hat{a} > ks - (\tilde{a} - \hat{a})) \end{aligned}$$

Now $\hat{a}$ is conditionally sufficient for α_0, so $\tilde{a} - \hat{a}$ is conditionally independent of $\hat{a}$, so we can use (*) to evaluate;

$$P(\tilde{\alpha} > kx_i) = 1 - \Phi\left(\frac{ks - (\tilde{a} - \hat{a}))}{\sigma_c}\right)$$

By symmetry,

$$P(\tilde{\alpha} > kx_i) = \Phi\left(-\frac{ks - (\tilde{a} - \hat{a})}{\sigma_c}\right)$$

Averaging these probabilities over several choices of k yields estimates $\hat{p}(k_i)$ and then regressing logit $(\hat{p}(k_i))$ on k to find k^* such that $p(k^*) = \alpha$ determines the correct critical value needed to achieve size α for the test. This is *considerably* more accurate than the naive method. As an example of the application of this method you might look at Jureckova, Koenker, Welsh *Annals of the Institute of Stat. Math.*, 1995.