The Bootstrap in Econometrics

Russell Davidson

Department of Economics and CIREQ
McGill University
Montréal, Québec, Canada
H3A 2T7

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13236 Marseille cedex 02, France

email: **russell.davidson@mcgill.ca**

April 2010

Lecture 1

Generalities

The bootstrap is a very general statistical technique. The underlying idea is that probability distributions of which the properties depend on a data-generating process (DGP) that is unknown can be estimated using data generated by the unknown DGP. Such data give rise to an estimated DGP, called the **bootstrap DGP**. The properties of the true unknown DGP that one wants to study are then estimated as the corresponding properties of the bootstrap DGP. Thus the bootstrap can be the basis for estimating the bias, the variance, the quantiles, and so on, of an estimator, test statistic, or any other random quantity of interest. The bootstrap is most often implemented by simulation, although, conceptually, simulation is not an essential element of the bootstrap. In practice, however, simulation is almost always required.

The basic idea of **bootstrap testing** is that, when a test statistic of interest has an unknown distribution under the null hypothesis under test, that distribution can be estimated by bootstrapping – the estimated distribution of the statistic is its distribution under the bootstrap DGP. Once a bootstrap DGP has been chosen, it is not often possible to find an analytic expression for the distribution of the statistic. Simulation enters the picture at this point.

One way to define a DGP is as a recipe for simulation. The vast utility of simulation in statistics and econometrics is that it allows arbitrarily accurate characterisations of things for which no tractable analytic expression can be found. The other side of this is that it can be quite difficult to derive efficient simulation techniques for a given analytically specified distribution or DGP.

If simulation is readily possible, one can generate a large number of simulated **bootstrap samples**. Each can then be used to calculate a **bootstrap test statistic**. The empirical distribution of these bootstrap statistics is an estimate of the distribution of the test statistic. This bootstrap distribution can be used to provide critical values against which the actual test statistic can be compared. Alternatively, a P value can be computed as the probability mass in the bootstrap distribution in the region more extreme than the actual statistic.

When the bootstrap distribution is a good approximation to the unknown true distribution of the test statistic under the null hypothesis, bootstrap tests should lead to accurate inferences. In some cases, they lead to very much more accurate inferences than using asymptotic distributions in the traditional way.

Definitions

The starting point for these definitions is the concept of **DGP**, by which is now meant a **unique recipe for simulation**. A DGP generates the virtual reality that our models use as a mirror for the reality of the economic phenomena we wish to study.

A **model** is a collection of DGPs. We need a model before embarking on any statistical enterprise. This is starkly illustrated by the theorems of Bahadur and Savage (1956). We must impose some constraints on the sorts of DGP we can envisage before any valid statistical conclusions can be drawn. Let $\mathbb{M}$ denote a model. Then $\mathbb{M}$ may represent a **hypothesis**. The hypothesis is that the true DGP, μ say, belongs to $\mathbb{M}$. Alternatively, we say that $\mathbb{M}$ is **correctly specified**.

Next, we almost always want to define a **parameter-defining mapping** θ. This maps the model $\mathbb{M}$ into a **parameter space** Θ, which is usually a subset of $\mathbb{R}^k$ for some finite positive integer k. For any DGP $\mu \in \mathbb{M}$, the $k-$vector $\theta(\mu)$, or θ_μ, is the **parameter vector** that corresponds to μ. Sometimes the mapping θ is one-one. This is the case with models estimated by maximum likelihood. More often, θ is many-one, so that a given parameter vector does not uniquely specify a DGP. Supposing the existence of θ implies that no **identification problems** remain to be solved.

In principle, a DGP specifies the probabilistic behaviour of all deterministic functions of the random data it generates – estimators, standard errors, test statistics, *etc*. If $\boldsymbol{y}$ denotes a data set, or **sample**, generated by a DGP μ, then a statistic $\tau(\boldsymbol{y})$ is a realisation of a random variable τ of which the distribution is determined by μ. A statistic τ is a **pivot**, or is **pivotal**, relative to a model $\mathbb{M}$ if its distribution under DGP $\mu \in \mathbb{M}$, $\mathcal{L}_\mu(\tau)$ say, is the same for all $\mu \in \mathbb{M}$.

Asymptotics

So far, the size of the samples generated by μ has not been mentioned explicitly. We denote the **sample size** by n. An **asymptotic theory** is an approximate theory based on the idea of letting n tend to infinity. This is a mathematical abstraction of course. We define an **asymptotic construction** as a construction, in the mathematical sense, of an infinite sequence $\{\mu_n\}$, $n = n_0, \ldots, \infty$, of DGPs such that μ_n generates samples of size n. Such sequences can then be collected into an **asymptotic model**, which we still denote as $\mathbb{M}$, and which can be thought of as a sequence of models $\mathbb{M}_n$.

Statistics, too, can be thought of as sequences $\{\tau_n\}$. The distribution of τ_n under μ_n is written as $\mathcal{L}_\mu^n(\tau)$. If this distribution tends in distribution to a limit $\mathcal{L}_\mu^\infty(\tau)$, then this limit is the **asymptotic** or **limiting distribution** of τ under μ. If $\mathcal{L}_\mu^\infty(\tau)$ is the same for all μ in an asymptotic model $\mathbb{M}$, then τ is an **asymptotic pivot** relative to $\mathbb{M}$.

The vast majority of statistics commonly used in econometrics are asymptotic pivots. t and F statistics, chi-squared statistics, Dickey-Fuller and associated statistics, *etc*, and even the dreaded Durbin-Watson statistic. All that matters is that the limiting distribution does not depend on unknown parameters.

Contrary to what many people have said and thought, the bootstrap is *not* an asymptotic procedure. It is possible to use and study the bootstrap for a fixed sample size without ever considering any other sample size. What is true, though, is that (current) bootstrap *theory* is almost all asymptotic.

Monte Carlo Tests

The simplest type of bootstrap test, and the only type that can be exact in finite samples, is called a **Monte Carlo test**. This type of test was first proposed by Dwass (1957). Monte Carlo tests are available whenever a test statistic is pivotal.

Suppose that we wish to test a null hypothesis represented by the model $\mathbb{M}_0$. Using real data, we compute a realisation $\hat{\tau}$ of a test statistic that is pivotal relative to $\mathbb{M}_0$. We then compute B independent bootstrap test statistics τ_j^*, $j = 1, \ldots, B$, using data simulated using *any* DGP in $\mathbb{M}_0$. Since τ is a pivot, it follows that the τ_j^* and $\hat{\tau}$ are independent drawings from one and the same distribution, *provided* that the true DGP, the one that generated $\hat{\tau}$, also satisfies the null hypothesis.

The empirical distribution function (EDF) of the bootstrap statistics can be written as

$$F^*(x) = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}(\tau_j^* \leq x),$$

where $\mathrm{I}(\cdot)$ is the **indicator function**, with value 1 when its argument is true and 0 otherwise.

Imagine that we wish to perform a test at significance level α, where α might, for example, be .05 or .01, and reject the null hypothesis when the value of $\hat{\tau}$ is unusually large. Given the actual and simulated test statistics, we can compute a **bootstrap *P* value** as

$$\hat{p}^*(\hat{\tau}) = \Pr{}^*(\tau^* > \hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\tau_j^* > \hat{\tau}).$$

Here τ^* is the **bootstrap statistic**, that is, the random variable of which the τ_j^* are realisations. Similarly, we denote by $\Pr{}^*$ probabilities computed using the bootstrap distribution. Evidently, $\hat{p}^*(\hat{\tau})$ is just the fraction of the bootstrap samples for which τ_j^* is larger than $\hat{\tau}$. If this fraction is smaller than α, we reject the null hypothesis. This makes sense, since $\hat{\tau}$ is extreme relative to the empirical distribution of the τ_j^* when $\hat{p}^*(\hat{\tau})$ is small.

Now suppose that we sort the original test statistic $\hat{\tau}$ and the B bootstrap statistics τ_j^* in decreasing order. Define the rank r of $\hat{\tau}$ in the sorted set in such a way that there are exactly r simulations for which $\tau_j^* > \hat{\tau}$. Then r can have $B+1$ possible values, $r = 0, 1, \ldots, B$, all of them equally likely under the null. The estimated P value $\hat{p}^*(\hat{\tau})$ is then just r/B.

The Monte Carlo test rejects if $r/B < \alpha$, that is, if $r < \alpha B$. Under the null, the probability that this inequality is satisfied is the proportion of the $B + 1$ possible values of r that satisfy it. If we denote by $\lfloor \alpha B \rfloor$ the largest integer that is no greater than αB, then, assuming that αB is not an integer, there are exactly $\lfloor \alpha B \rfloor + 1$ such values of r, namely, $0, 1, \ldots, [\alpha B]$. Thus the probability of rejection is $(\lfloor \alpha B \rfloor + 1)/(B + 1)$. We want this probability to be exactly equal to α. For that to be true, we require that

$$\alpha(B + 1) = \lfloor \alpha B \rfloor + 1.$$

Since the right-hand side above is the sum of two integers, this equality can hold only if $\alpha(B+1)$ is also an integer. In fact, it is easy to see that the equation holds whenever $\alpha(B+1)$ is an integer. Suppose that $\alpha(B+1) = k$, k an integer. Then $\lfloor \alpha B \rfloor = k - 1$, and so

$$\Pr(r < \alpha B) = \frac{k - 1 + 1}{B + 1} = \frac{k}{B + 1} = \frac{\alpha(B + 1)}{B + 1} = \alpha.$$

In that case, therefore, the rejection probability under the null, that is, the Type I error of the test, is precisely α, the desired significance level.

Of course, using simulation injects randomness into this test procedure, and the cost of this randomness is a loss of power. A test based on $B = 99$ simulations will be less powerful than a test based on $B = 199$, which in turn will be less powerful than one based on $B = 299$, and so on; see Jöckel (1986) and Davidson and MacKinnon (2000). Notice that all of these values of B have the property that $\alpha(B + 1)$ is an integer whenever α is an integer percentage like .01, .05, or .10.

Bootstrap Tests

Although pivotal test statistics do arise from time to time, most test statistics in econometrics are not pivotal. The vast majority of them are, however, asymptotically pivotal. A statistic that is not an exact pivot cannot be used for a Monte Carlo test. However, approximate P values for statistics that are only asymptotically pivotal, or even nonpivotal, can still be obtained by bootstrapping. The difference between a Monte Carlo test and a bootstrap test is that for the former, the DGP is assumed to be known, whereas, for the latter, it is not. Unless the null hypothesis under test is a simple hypothesis, the DGP that generated the original data is unknown, and so it cannot be used to generate simulated data. The bootstrap DGP is an estimate of the unknown true DGP. The hope is that, if the bootstrap DGP is close, in some sense, to the true one, then data generated by the bootstrap DGP will be similar to data that would have been generated by the true DGP, if it were known. If so, then a simulated P value obtained by use of the bootstrap DGP is close enough to the true P value to allow accurate inference.

The actual implementation of bootstrap test is identical to that of a Monte Carlo test. The only difference is that we do not (usually) just choose any convenient DGP in the null model, but rather one that can be considered a good estimate of the unknown true DGP.

Lecture 2

The Golden Rules of Bootstrapping

If a test statistic τ is asymptotically pivotal for a given model $\mathbb{M}$, then its distribution should not vary too much as a function of the specific DGP, μ say, within that model. It is usually possible to show that the distance between the distribution of τ under the DGP μ for sample size n and that for infinite n tends to zero like some negative power of n, commonly $n^{-1/2}$. The concept of "distance" between distributions can be realised in various ways, some ways being more relevant for bootstrap testing than others.

Heuristically speaking, if the distance between the finite-sample distribution for any DGP $\mu \in \mathbb{M}$ and the limiting distribution is of order $n^{-\delta}$ for some $\delta > 0$, then, since the limiting distribution is the same for all $\mu \in \mathbb{M}$, the distance between the finite-sample distributions for two DGPs μ_1 and μ_2 in $\mathbb{M}$ is also of order $n^{-\delta}$. If now the distance between μ_1 and μ_2 is also small, in some sense, say of order $n^{-\varepsilon}$, it should be the case that the distance between the distributions of τ under μ_1 and μ_2 should be of order $n^{-(\delta+\varepsilon)}$.

Arguments of this sort are used to show that the bootstrap can, in favourable circumstances, benefit from **asymptotic refinements**. The form of the argument was given in a well-known paper of Beran (1988). No doubt wisely, Beran limits himself in this paper to the outline of the argument, with no discussion of formal regularity conditions. It remains true today that no really satisfying general theory of bootstrap testing has been found to embody rigorously the simple idea set forth by Beran. Rather, we have numerous piecemeal results that prove the existence of refinements in specific cases, along with other results that show that the bootstrap does not work in other specific cases. Perhaps the most important instance of negative results of this sort, often called **bootstrap failure**, applies to bootstrapping when the true DGP generates data with a heavy-tailed distribution; see Athreya (1987) for the case of infinite variance.

A technique that has been used a good deal in work on asymptotic refinements for the bootstrap is **Edgeworth expansion** of distributions, usually distributions that become standard normal in the limit of infinite sample size. The standard reference to this line of work is Hall (1992), although there is no shortage of more recent work based on Edgeworth expansions. Whereas the technique can lead to useful theoretical insights, it is unfortunately not very useful as a quantitative explanation of the properties of bootstrap tests. In concrete cases, the true finite-sample distribution of a bootstrap P value, as estimated by simulation, can easily be further removed from an Edgeworth approximation to its distribution than from the asymptotic limiting distribution.

Rules for bootstrapping

All these theoretical caveats notwithstanding, experience has shown abundantly that bootstrap tests, in many circumstances of importance for applied econometrics, are much more reliable than tests based on asymptotic theories of one sort or another. The bootstrap DGP will henceforth be denoted as μ^*. Since in testing the bootstrap is used to estimate the distribution of a test statistic under the null hypothesis, the first golden rule of bootstrapping is:

Golden Rule 1:

> The bootstrap DGP μ^* must belong to the model $\mathbb{M}$ that represents the null hypothesis.

It is not always possible, or, even if it is, it may be difficult to obey this rule in some cases, as we will see with confidence intervals.

If, in violation of this rule, the null hypothesis tested by the bootstrap statistics is not satisfied by the bootstrap DGP, a bootstrap test can be wholly lacking in power. Test power springs from the fact that a statistic has different distributions under the null and the alternative. Bootstrapping under the alternative confuses these different distributions, and so leads to completely unreliable inference, even in the asymptotic limit.

Whereas Golden Rule 1 must be satisfied in order to have an asymptotically justified test, Golden Rule 2 is concerned rather with making the probability of rejecting a true null with a bootstrap test as close as possible to the significance level. It is motivated by the argument of Beran discussed earlier.

Golden Rule 2:

> Unless the test statistic is pivotal for the null model $\mathbb{M}$, the bootstrap DGP should be as good an estimate of the true DGP as possible, under the assumption that the true DGP belongs to $\mathbb{M}$.

How this second rule can be followed depends very much on the particular test being performed, but quite generally it means that we want the bootstrap DGP to be based on estimates that are *efficient* under the null hypothesis.

Once the sort of bootstrap DGP has been chosen, the procedure for conducting a bootstrap test based on simulated bootstrap samples follows the following pattern.

(i) Compute the test statistic from the original sample; call its realised value $\hat{\tau}$.

(ii) Determine the realisations of all other data-dependent things needed to set up the bootstrap DGP μ^*.

(iii) Generate B bootstrap samples using μ^*, and for each one compute a realisation of the bootstrap statistic, τ_j^*, $j = 1, \dots B$. It is prudent to choose B so that $\alpha(B+1)$ is an integer for all interesting significance levels α, typically 1%, 5%, and 10%.

(iv) Compute the simulated bootstrap P value as the proportion of bootstrap statistics τ_j^* that are more extreme than $\hat{\tau}$. For a statistic that rejects for large values, for instance, we have

$$P_{\text{bs}} = \frac{1}{B} \sum_{j=1}^{B} \text{I}(\tau_j^* > \hat{\tau}),$$

where $\text{I}(\cdot)$ is an indicator function, with value 1 if its Boolean argument is true, and 0 if it is false.

The bootstrap test rejects the null hypothesis at significance level α if $P_{\text{bs}} < \alpha$.

The Parametric Bootstrap

If the model $\mathbb{M}$ that represents the null hypothesis can be estimated by maximum likelihood (ML), there is a one-one relation between the parameter space of the model and the DGPs that belong to it. For any fixed admissible set of parameters, the likelihood function evaluated at those parameters is a probability density. Thus there is one and only one DGP associated with the set of parameters. By implication, the only DGPs in $\mathbb{M}$ are those completely characterised by a set of parameters.

If the model $\mathbb{M}$ actually is estimated by ML, then the ML parameter estimates provide an asymptotically efficient estimate not only of the true parameters themselves, but also of the true DGP. Both golden rules are therefore satisfied if the bootstrap DGP is chosen as the DGP in $\mathbb{M}$ characterised by the ML parameter estimates. In this case we speak of a *parametric bootstrap*.

In microeconometrics, models like probit and logit are commonly estimated by ML. These are of course just the simplest of microeconometric models, but they are representative of all the others for which it is reasonable to suppose that the data can be described by a purely parametric model.

Resampling

Resampling was a key aspect of the original conception of the bootstrap, as set out in Efron's (1979) pioneering paper.

Basic Resampling

Resampling is valuable when it is undesirable to constrain a model so tightly that all of its possibilities are encompassed by the variation of a finite set of parameters. A classic instance is a regression model where one does not wish to impose the normality of the disturbances.

The parametric bootstrap DGP with normal disturbances satisfies Golden Rule 1, because the normal distribution is plainly allowed when all we specify are the first two moments. But Golden Rule 2 incites us to seek as good an estimate as possible of the unknown distribution of the disturbances. If the disturbances were observed, then the best nonparametric estimate of their distribution would be their EDF. The unobserved disturbances can be estimated, or proxied, by the residuals from estimating the null model. If we denote the empirical distribution of these residuals by $\hat{F}$, the bootstrap DGP would generate bootstrap disturbances like

$$u_t^* \sim \text{IID}(\hat{F}), \quad t = 2, \ldots, n.$$

where the notation indicates that the bootstrap disturbances, the u_t^*, are IID drawings from the empirical distribution characterised by the EDF $\hat{F}$.

The term *resampling* comes from the fact that the easiest way to generate the u_t^* is to sample from the residuals at random with replacement. The residuals are thought of as sampling the true DGP, and so this operation is called "resampling". For each $t = 2, \ldots, n$, one can draw a random number m_t from the U(0, 1) distribution, and then obtain u_t^* by the operations:

$$s = \lfloor 2 + (n-1)m_t \rfloor, \quad u_t^* = \tilde{u}_s,$$

where the notation $\lfloor x \rfloor$ means the greatest integer not greater than x. For m_t close to 0, $s = 2$; for m_t close to 1, $s = n$, and we can see that s is uniformly distributed over the integers $2, \ldots, n$. Setting u_t^* equal to the (restricted) residual $\tilde{u}_s$ therefore implements the required resampling operation.

More sophisticated resampling

But is the empirical distribution of the residuals really the best possible estimate of the distribution of the disturbances? Not always. Consider a very simple model, one with no constant term:

$$y_t = \rho y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2).$$

When this is estimated by OLS, or, if the null hypothesis fixes the value of ρ, in which case the "residuals" are just the observed values $y_t - \rho_0 y_{t-1}$, the residuals do not in general sum to zero, precisely because there is no constant term. But the model requires that the expectation of the disturbance distribution should be zero, whereas the expectation of the empirical distribution of the residuals is their mean. Thus using this empirical distribution violates Golden Rule 1.

This is easily fixed by replacing the residuals by the deviations from their mean, and then resampling these centred residuals. But now what about Golden Rule 2?

The variance of the centred residuals is the sum of their squares divided by n:

$$V = \frac{1}{n}\sum_{t=1}^{n}(\tilde{u}_t^2 - \bar{u})^2,$$

where $\bar{u}$ is the mean of the uncentred residuals. But the unbiased estimator of the variance of the disturbances is

$$s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(\tilde{u}_t^2 - \bar{u})^2.$$

More generally, in any regression model that uses up k degrees of freedom in estimating regression parameters, the unbiased variance estimate is the sum of squared residuals divided by $n - k$. What this suggests is that what we want to resample is a set of *rescaled* residuals, which here would be the $\sqrt{n/(n-k)}\tilde{u}_t$. The variance of the empirical distribution of these rescaled residuals is then equal to the unbiased variance estimate.

Of course, some problems are scale-invariant. Indeed, test statistics that are ratios are scale invariant for both the autoregressive models we have considered under the stationarity assumption. For models like these, therefore, there is no point in rescaling, since bootstrap statistics computed with the same set of random numbers are unchanged by scaling. This property is akin to pivotalness, in that varying some, but not all, of the parameters of the null model leaves the distribution of the test statistic invariant. In such cases, it is unnecessary to go to the trouble of estimating parameters that have no effect on the distribution of the statistic τ.

Weighted resampling

A way to impose the null hypothesis with a resampling bootstrap is to resample with unequal weights. Ordinary resampling assigns a weight of n^{-1} to each observation, but if different weights are assigned to different observations, it is possible to impose various sorts of restrictions. This approach is suggested by Brown and Newey (2002).

A nonparametric technique that shares many properties with parametric maximum likelihood is **empirical likelihood**; see Owen (2001). In the case of an IID sample, the empirical likelihood is a function of a set of nonnegative probabilities p_i, $i = 1, \ldots, n$, such that $\sum_{i=1}^{n} p_i = 1$. The empirical loglikelihood, easier to manipulate than the empirical likelihood itself, is given as

$$\ell(\boldsymbol{p}) = \sum_{i=1}^{n} \log p_i.$$

Here $\boldsymbol{p}$ denotes the $n-$vector of the probabilities p_i. The idea now is to maximise the empirical likelihood subject to the constraint or constraints imposed by the null hypothesis.

With very small sample sizes, it is possible that this constrained maximisation problem has no solution with nonnegative probabilities. In such a case, the **empirical likelihood ratio** statistic would be set equal to ∞, and the null hypothesis rejected out of hand, with no need for bootstrapping. In the more common case in which the problem can be solved, the bootstrap DGP resamples the original sample with observation i resampled with probability p_i rather than n^{-1}. The use of empirical likelihood for the determination of the p_i means that these probabilities have various optimality properties relative to any other set satisfying the desired constraint. Golden Rule 2 is satisfied.

The best algorithm for weighted resampling appears to be little known in the econometrics community. It is described in Knuth (1998). Briefly, for a set of probabilities p_i, $i = 1, \ldots, n$, two tables of n elements each are set up, containing the values q_i, with $0 < q_i \leq 1$, and y_i, where y_i is an integer in the set $1, \ldots, n$. In order to obtain the index j of the observation to be resampled, a random number m_i from $\mathrm{U}(0, 1)$ is used as follows.

$$k_i = \lceil nm_i \rceil, \quad r_i = k_i - nm_i, \quad j = \begin{cases} k_i & \text{if } r_i \leq q_i, \\ y_i & \text{otherwise.} \end{cases}$$

For details, readers are referred to Knuth's treatise.

Some Examples

Exact pivots, as opposed to asymptotic pivots, can be hard to find. They exist with the **classical normal linear model**, but most, like t and F tests, have distributions that are known analytically, and so neither bootstrapping nor simulation is necessary. But there exist a few cases in which there is a pivotal statistic of which the distribution under the null is unknown or else intractable.

Consider the classical normal linear regression model

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2),$$

The $1 \times k$ vector of regressors $\boldsymbol{X}_t$ is the t^{th} row of the $n \times k$ matrix $\boldsymbol{X}$. $\boldsymbol{X}$ is treated as a *fixed* property of the null model. Thus every DGP belonging to this model is completely characterized by the values of the parameter vector $\boldsymbol{\beta}$ and the variance σ^2. Any test statistic the distribution of which does not depend on these values is a pivot for the null model. In particular, a statistic that depends on $\boldsymbol{y}$ only through the OLS residuals and is invariant to the scale of $\boldsymbol{y}$ is pivotal.

The first example is the Durbin-Watson test for serial correlation. ('Nuf said!) A better example is the estimated autoregressive parameter $\hat{\rho}$ that is obtained by regressing the t^{th} residual $\hat{u}_t$ on its predecessor $\hat{u}_{t-1}$. The estimate $\hat{\rho}$ can be used as a test for serial correlation of the disturbances. Evidently,

$$\hat{\rho} = \frac{\sum_{t=2}^{n} \hat{u}_{t-1}\hat{u}_t}{\sum_{t=2}^{n} \hat{u}_{t-1}^2}.$$

Since $\hat{u}_t$ is proportional to σ, there are implicitly two factors of σ in the numerator and two in the denominator. Thus $\hat{\rho}$ is independent of the scale factor σ.

Implementation

Since the bootstrap DGP can be any DGP in the null model, we choose the simplest such DGP, with $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma^2 = 1$. It can be written as

$$y_t^* = u_t^*, \quad u_t^* \sim \text{NID}(0, 1).$$

For each of B bootstrap samples, we then proceed as follows:

1. Generate the vector $\boldsymbol{y}^*$ as an n–vector of IID standard normal variables.
2. Regress $\boldsymbol{y}^*$ on $\boldsymbol{X}$ and save the vector of residuals $\hat{\boldsymbol{u}}^*$.
3. Compute ρ^* by regressing $\hat{u}_t^*$ on $\hat{u}_{t-1}^*$ for observations 2 through n.

Denote by ρ_j^*, $j = 1, \ldots, B$, the bootstrap statistics obtained by performing the above three steps B times. We now have to choose the alternative to our null hypothesis of no serial correlation. If the alternative is positive serial correlation, then we perform a **one-tailed test** by computing the bootstrap P value as

$$\hat{p}^*(\hat{\rho}) = \frac{1}{B} \sum_{j=1}^{B} \text{I}(\rho_j^* > \hat{\rho}).$$

This P value is small when $\hat{\rho}$ is positive and sufficiently large, thereby indicating positive serial correlation.

However, we may wish to test against both positive and negative serial correlation. In that case, there are two possible ways to compute a P value corresponding to a **two-tailed test**. The first is to assume that the distribution of $\hat{\rho}$ is symmetric, in which case we can use the bootstrap P value

$$\hat{p}^*(\hat{\rho}) = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\big(|\rho_j^*| > |\hat{\rho}|\big).$$

This is implicitly a symmetric two-tailed test, since we reject when the fraction of the ρ_j^* that exceed $\hat{\rho}$ in absolute value is small. Alternatively, if we do not assume symmetry, we can use

$$\hat{p}^*(\hat{\rho}) = 2\min\left(\frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\rho_j^* \leq \hat{\rho}),\ \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}(\rho_j^* > \hat{\rho})\right).$$

In this case, for level α, we reject whenever $\hat{\rho}$ is either below the $\alpha/2$ quantile or above the $1-\alpha/2$ quantile of the empirical distribution of the ρ_j^*. Although tests based on these two P values are both exact, they may yield conflicting results, and their power against various alternatives will differ.

Power considerations

The power of a bootstrap test depends on B. If to any test statistic we add random noise independent of the statistic, we inevitably reduce the power of tests based on that statistic. Note that the bootstrap P value $\hat{p}^*(\hat{\tau})$ is an estimate of the **ideal bootstrap P value**

$$p^*(\hat{\tau}) \equiv \Pr{}^*(\tau^* > \hat{\tau}) = \operatorname*{plim}_{B\to\infty} \hat{p}^*(\hat{\tau}),$$

When B is finite, $\hat{p}^*$ differs from p^* because of random variation in the bootstrap samples. This random variation is generated in the computer, and is therefore completely independent of the random variable τ. The bootstrap testing procedure incorporates this random variation, and in so doing it reduces the power of the test.

Power loss is illustrated in Figure 1. It shows power functions for four tests at the .05 level of a null hypothesis with only 10 observations. All four tests are exact, as can be seen from the fact that, in all cases, power equals .05 when the null is true. When it is not, there is a clear ordering of the four curves, depending on the number of bootstrap samples used. The loss of power is quite modest when $B = 99$, but it is substantial when $B = 19$.

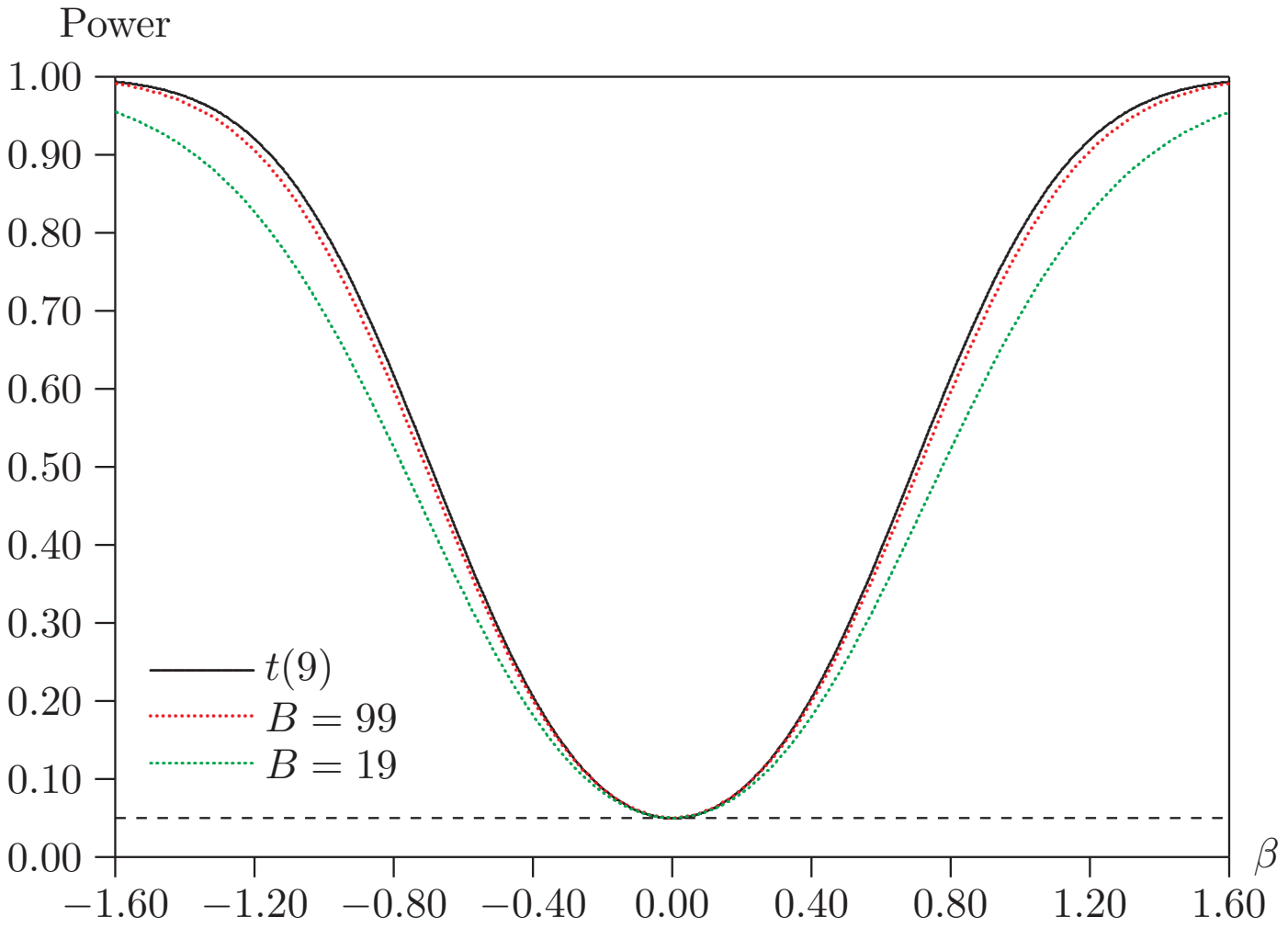

Figure 1: Power loss with finite B.

Many common test statistics for serial correlation, heteroskedasticity, skewness, and excess kurtosis in the classical normal linear regression model are pivotal, since they depend on the regressand only through the least squares residuals $\hat{\boldsymbol{u}}$ in a way that is invariant to the scale factor σ. The Durbin-Watson d statistic is a particularly well-known example. We can perform a Monte Carlo test based on d just as easily as a Monte Carlo test based on $\hat{\rho}$, and the two tests should give very similar results. Since we condition on $\boldsymbol{X}$, the infamous upper and lower bounds from the classic tables of the d statistic are quite unnecessary.

With modern computers and appropriate software, it is extremely easy to perform a variety of exact tests in the context of the classical normal linear regression model. These procedures also work when the disturbances follow a nonnormal distribution that is known up to a scale factor; we just have to use the appropriate distribution in step 1 above. For further references and a detailed treatment of Monte Carlo tests for heteroskedasticity, see Dufour, Khalaf, Bernard, and Genest (2004).

A binary choice model

Suppose that a binary dependent variable y_t, $t = 1, \ldots, n$, takes on only the values 0 and 1, with the probability that $y_t = 1$ being given by $F(\boldsymbol{X}_t\boldsymbol{\beta})$, where $\boldsymbol{X}_t$ is a $1 \times k$ vector of exogenous explanatory variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters, and F is a function that maps real numbers into the $[0, 1]$ interval. For probit, F is the CDF of the standard normal distribution; for logit, it is the CDF of the logistic distribution.

The contribution to the loglikelihood for the whole sample made by observation t is

$$\mathrm{I}(y_t = 1) \log F(\boldsymbol{X}_t\boldsymbol{\beta}) + \mathrm{I}(y_t = 0) \log\big(1 - F(\boldsymbol{X}_t\boldsymbol{\beta})\big),$$

Suppose now that the parameter vector $\boldsymbol{\beta}$ can be partitioned into two subvectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, and that, under the null hypothesis, $\boldsymbol{\beta}_2 = \mathbf{0}$. The *restricted* ML estimator, that is, the estimator of the subvector $\boldsymbol{\beta}_1$ only, with $\boldsymbol{\beta}_2$ set to zero, is then an asymptotically efficient estimator of the only parameters that exist under the null hypothesis.

Although asymptotic theory is used to convince us of the desirability of the ML estimator, the bootstrap itself is a purely finite-sample procedure. If we denote the restricted ML estimate as $\tilde{\boldsymbol{\beta}} \equiv [\tilde{\boldsymbol{\beta}}_1 \vdots \mathbf{0}]$, the bootstrap DGP can be represented as follows.

$$y_t^* = \begin{cases} 1 & \text{with probability } F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}}), \text{ and} \\ 0 & \text{with probability } 1 - F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}}). \end{cases}, \qquad t = 1, \dots, n.$$

Here the usual notational convention is followed, according to which variables generated by the bootstrap DGP are starred. Note that the explanatory variables $\boldsymbol{X}_t$ are *not* starred. Since they are assumed to be exogenous, it is not the business of the bootstrap DGP to regenerate them; rather they are thought of as fixed characteristics of the bootstrap DGP, and so are used unchanged in each bootstrap sample. Since the bootstrap samples are exactly the same size, n, as the original sample, there is no need to generate explanatory variables for any more observations than those actually observed.

It is easy to implement the above bootstrap DGP. A **random number** m_t is drawn, using a random number generator, as a drawing from the uniform U(0, 1) distribution. Then we generate y_t^* as $\mathrm{I}\big(m_t \leq F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})\big)$. Most matrix or econometric software can implement this as a vector relation, so that, after computing the $n-$vector with typical element $F(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})$, the vector $\boldsymbol{y}^*$ with typical element y_t^* can be generated by a single command.

Recursive simulation

In dynamic models, the implementation of the bootstrap DGP may require **recursive simulation**. Let us now take as an example the very simple autoregressive time-series model

$$y_t = \alpha + \rho y_{t-1} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad t = 2, \ldots, n.$$

Here the dependent variable y_t is continuous, unlike the binary dependent variable above. The model parameters are α, ρ, and σ^2. However, even if the values of these parameters are specified, we still do not have a complete characterisation of a DGP. Because the defining relation is a recurrence, it needs a starting value, or initialisation, before it yields a unique solution. Thus, although it is not a parameter in the usual sense, the first observation, y_1, must also be specified in order to complete the specification of the DGP.

ML estimation of the model is the same as estimation by ordinary least squares (OLS) omitting the first observation. If the recurrence represents the null hypothesis, then we would indeed estimate α, ρ, and σ by OLS. If the null hypothesis specifies the value of any one of those parameters, requiring for instance that $\rho = \rho_0$, then we would use OLS to estimate the model in which this restriction is imposed:

$$y_t - \rho_0 y_{t-1} = \alpha + u_t,$$

with the same specification of the disturbances u_t.

The bootstrap DGP is then the DGP contained in the null hypothesis that is characterised by the restricted parameter estimates, and by some suitable choice of the starting value, y_1^*. One way to choose y_1^* is just to set it y_1, the value in the original sample. In most cases, this is the best choice. It restricts the model by fixing the initial value. A bootstrap sample can now be generated recursively, starting with y_2^*. For all $t = 2, \ldots, n$, we have

$$y_t^* = \tilde{\alpha} + \tilde{\rho} y_{t-1}^* + \tilde{\sigma} v_t^*, \quad v_t^* \sim \text{NID}(0, 1).$$

Often, one wants to restrict the possible values of ρ to values strictly between -1 and 1. This restriction makes the series y_t **asymptotically stationary**, by which we mean that, if we generate a very long sample from the recurrence, then towards the end of the sample, the distribution of y_t becomes independent of t, as also the joint distribution of any pair of observations, y_t and y_{t+s}, say. Sometimes it make sense to require that the series y_t should be stationary, and not just asymptotically stationary, so that the distribution of every observation y_t, including the first, is always the same. It is then possible to include the information about the first observation into the ML procedure, and so get a more efficient estimate that incorporates the extra information. For the bootstrap DGP, y_1^* should now be a random drawing from the stationary distribution.

A poverty index

In some circumstances, we may wish to affect the values of more complicated functionals of a distribution. Suppose for instance that we wish to perform inference about a poverty index. An IID sample of individual incomes is available, drawn at random from the population under study, and the null hypothesis is that a particular poverty index has a particular given value. For concreteness, let us consider one of the FGT indices, defined as follows; see Foster, Greer, and Thorbecke (1984).

$$\Delta^\alpha(z) = \int_0^z (z-y)^{\alpha-1} \, \mathrm{d}F(y).$$

Here z is interpreted as a poverty line, and F is the CDF of income. We assume that the poverty line z and the parameter α are fixed at some prespecified values. The obvious estimator of of $\Delta^\alpha(z)$ is just

$$\hat{\Delta}^\alpha(z) = \int_0^z (z-y)^{\alpha-1} \, \mathrm{d}\hat{F}(y),$$

where $\hat{F}$ is the EDF of income in the sample. For sample size n, we have explicitly that

$$\hat{\Delta}^\alpha(z) = \frac{1}{n} \sum_{i=1}^{n} (z-y_i)_+^{\alpha-1},$$

where y_i is income for observation i, and $(x)_+$ denotes $\max(0, x)$.

Since $\hat{\Delta}^\alpha(z)$ is just the mean of a set of IID variables, its variance can be estimated by

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}(z - y_i)_+^{2\alpha-2} - \left(\frac{1}{n}\sum_{i=1}^{n}(z - y_i)_+^{\alpha-1}\right)^2.$$

A suitable test statistic for the hypothesis that $\Delta^\alpha(z) = \Delta_0$ is then

$$t = \frac{\hat{\Delta}^\alpha(z) - \Delta_0}{\hat{V}^{1/2}}.$$

With probability 1, the estimate $\hat{\Delta}^\alpha(z)$ is not equal to Δ_0. If the statistic t is bootstrapped using ordinary resampling of the data in the original sample, this fact means that we violate Golden Rule 1. The simplest way around this difficulty, as mentioned after the statement of Golden Rule 1, is to change the null hypothesis tested by the bootstrap statistics, testing rather what is true under the resampling DGP, namely $\Delta^\alpha(z) = \hat{\Delta}^\alpha(z)$. Thus each bootstrap statistic takes the form

$$t^* = \frac{(\Delta^\alpha(z))^* - \hat{\Delta}^\alpha(z)}{(V^*)^{1/2}}.$$

Here $(\Delta^\alpha(z))^*$ is the estimate computed using the bootstrap sample, and V^* is the variance estimator computed using the bootstrap sample. Golden Rule 1 is saved by the trick of changing the null hypothesis for the bootstrap samples, but Golden Rule 2 would be better satisfied if we could somehow impose the real null hypothesis on the bootstrap DGP. This can be achieved by the weighted resampling procedure.

Lecture 4

Heteroskedasticity

All the bootstrap DGPs that we have looked at so far are based on models where either the observations are IID, or else some set of quantities that can be estimated from the data, like the disturbances of a regression model, are IID. But if the disturbances of a regression are heteroskedastic, with an unknown pattern of heteroskedasticity, there is nothing that is even approximately IID. There exist of course test statistics robust to heteroskedasticity of unknown form, based on one of the numerous variants of the Eicker-White Heteroskedasticity Consistent Covariance Matrix Estimator (HCCME). Use of an HCCME gives rise to statistics that are approximately pivotal for models that admit heteroskedasticity of unknown form.

For bootstrapping, it is very easy to satisfy Golden Rule 1, since either a parametric bootstrap or a resampling bootstrap of the sort we have described belongs to a null hypothesis that, since it allows heteroskedasticity, must also allow the special case of homoskedasticity. But Golden Rule 2 poses a more severe challenge.

The pairs bootstrap

The first suggestion for bootstrapping models with heteroskedasticity bears a variety of names: among them the (y, X) bootstrap or the pairs bootstrap. The approach was proposed in Freedman (1981). Instead of resampling the dependent variable, or residuals, possibly centred or rescaled, one bootstraps **pairs** consisting of an observation of the dependent variable along with the set of explanatory variables for that same observation. One selects an index s at random from the set $1, \ldots, n$, and then an observation of a bootstrap sample is the pair $(y_s, \boldsymbol{X}_s)$, where $\boldsymbol{X}_s$ is a row vector of all the explanatory variables for observation s.

This bootstrap implicitly assumes that the pairs $(y_t, \boldsymbol{X}_t)$ are IID under the null hypothesis. Although this is still a restrictive assumption, ruling out any form of dependence among observations, it does allow for any sort of heteroskedasticity of y_t conditional of $\boldsymbol{X}_t$. The objects resampled are IID drawings from the *joint* distribution of y_t and $\boldsymbol{X}_t$.

Suppose that the regression model itself is written as

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \quad t = 1, \ldots, n,$$

with $\boldsymbol{X}_t$ a $1 \times k$ vector and $\boldsymbol{\beta}$ a $k \times 1$ vector of parameters. The disturbances u_t are allowed to be heteroskedastic, but must have an expectation of 0 conditional on the explanatory variables. Thus $\mathrm{E}(y_t|\boldsymbol{X}_t) = \boldsymbol{X}_t\boldsymbol{\beta}_0$ if $\boldsymbol{\beta}_0$ is the parameter vector for the true DGP. Let us consider a null hypothesis according to which a subvector of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_2$ say, is zero. This null hypothesis is not satisfied by the pairs bootstrap DGP. In order to respect Golden Rule 1, therefore, we must modify either the null hypothesis to be tested in the bootstrap samples, or the bootstrap DGP itself.

In the empirical joint distribution of the pairs $(y_t, \boldsymbol{X}_t)$, the expectation of the first element y conditional on the second element $\boldsymbol{X}$ is defined only if $\boldsymbol{X} = \boldsymbol{X}_t$ for some $t = 1, \ldots, n$. Then $\mathrm{E}(y|\boldsymbol{X} = \boldsymbol{X}_t) = y_t$. This result does not help determine what the true value of $\boldsymbol{\beta}$, or of $\boldsymbol{\beta}_2$, might be for the bootstrap DGP. Given this, what is usually done is to use the OLS estimate $\hat{\boldsymbol{\beta}}_2$ as true for the bootstrap DGP, and so to test the hypothesis that $\boldsymbol{\beta}_2 = \hat{\boldsymbol{\beta}}_2$ when computing the bootstrap statistics.

In Flachaire (1999), the bootstrap DGP is changed. It now resamples pairs $(\hat{u}_t, \boldsymbol{X}_t)$, where the $\hat{u}_t$ are the OLS residuals from estimation of the *unrestricted* model, possibly rescaled in various ways. Then, if s is an integer drawn at random from the set $1, \ldots, n$, y_t^* is generated by

$$y_t^* = \boldsymbol{X}_{s1}\tilde{\boldsymbol{\beta}}_1 + \hat{u}_s,$$

where $\boldsymbol{\beta}_1$ contains the elements of $\boldsymbol{\beta}$ that are not in $\boldsymbol{\beta}_2$, and $\tilde{\boldsymbol{\beta}}_1$ is the *restricted* OLS estimate. Similarly, $\boldsymbol{X}_{s1}$ contains the elements of $\boldsymbol{X}_s$ of which the coefficients are elements of $\boldsymbol{\beta}_1$. By construction, the vector of the $\hat{u}_t$ is orthogonal to all of the vectors containing the observations of the explanatory variables. Thus in the empirical joint distribution of the pairs $(\hat{u}_t, \boldsymbol{X}_t)$, the first element, $\hat{u}$, is uncorrelated with the second element, $\boldsymbol{X}$. However any relation between the variance of $\hat{u}$ and the explanatory variables is preserved, as with Freedman's pairs bootstrap. In addition, the new bootstrap DGP now satisfies the null hypothesis as originally formulated.

The wild bootstrap

The null model on which any form of pairs bootstrap is based posits the joint distribution of the dependent variable y and the explanatory variables. If it is assumed that the explanatory variables are exogenous, conventional practice is to compute statistics, and their distributions, conditional on them. One way in which this can be done is to use the so-called **wild bootstrap**; see Wu (1986) Liu (1988), Mammen (1993), and Davidson and Flachaire (2008).

For a regression model, the wild bootstrap DGP takes the form

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + s_t^*\tilde{u}_t$$

where $\tilde{\boldsymbol{\beta}}$ is as usual the restricted least-squares estimate of the regression parameters, and the $\tilde{u}_t$ are the restricted least-squares residuals. Notice that no resampling takes place here; both the explanatory variables and the residual for bootstrap observation t come from observation t of the original sample. The new random elements introduced are the s_t^*, which are IID drawings from a distribution with expectation 0 and variance 1.

The bootstrap DGP satisfies Golden Rule 1 easily: since s_t^* and $\tilde{u}_t$ are independent, the latter having been generated by the real DGP and the former by the random number generator, the expectation of the bootstrap disturbance $s_t^*\tilde{u}_t$ is 0. Conditional on the residual $\tilde{u}_t$, the variance of $s_t^*\tilde{u}_t$ is $\tilde{u}_t^2$. If the residual is accepted as a proxy for the unobserved disturbance u_t, then the expectation of $\tilde{u}_t^2$ is the true variance of u_t, and this fact goes a long way towards satisfying Golden Rule 2.

For a long time, the most commonly used distribution for the s_t^* was the following two-point distribution,

$$s_t^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}), \end{cases}$$

which was suggested by Mammen. A simpler two-point distribution is the **Rademacher distribution**

$$s_t^* = \begin{cases} -1 & \text{with probability } \frac{1}{2}, \\ 1 & \text{with probability } \frac{1}{2}. \end{cases}$$

Davidson and Flachaire propose this simpler distribution, which leaves the absolute value of each residual unchanged in the bootstrap DGP, while assigning it an arbitrary sign. They show by means of simulation experiments that their choice often leads to more reliable bootstrap inference than other choices.

Confidence Intervals

A confidence interval for some scalar parameter θ consists of all values θ_0 for which the hypothesis $\theta = \theta_0$ cannot be rejected at some specified level α. Thus we can construct a confidence interval by "inverting" a test statistic. If the finite-sample distribution of the test statistic is known, we obtain an **exact confidence interval**. If, as is more commonly the case, only the asymptotic distribution of the test statistic is known, we obtain an **asymptotic confidence interval**, which may or may not be reasonably accurate in finite samples. Whenever a test statistic based on asymptotic theory has poor finite-sample properties, a confidence interval based on that statistic has poor coverage.

To begin with, suppose that we wish to base a confidence interval for the parameter θ on a family of test statistics that have a distribution or asymptotic distribution like the χ^2 or the F distribution under their respective nulls. Statistics of this type are always positive, and tests based on them reject their null hypotheses when the statistics are sufficiently large. Such tests are often equivalent to two-tailed tests based on statistics distributed as standard normal or Student's t. Let us denote the test statistic for the hypothesis that $\theta = \theta_0$ by the random variable $\tau(\boldsymbol{y}, \theta_0)$.

For each θ_0, the test consists of comparing the realized $\tau(\boldsymbol{y}, \theta_0)$ with the level α critical value of the distribution of the statistic under the null. If we write the critical value as c_α, then, for any θ_0, we have by the definition of c_α that

$$\Pr{}_{\theta_0}\big(\tau(\boldsymbol{y}, \theta_0) \leq c_\alpha\big) = 1 - \alpha.$$

For θ_0 to belong to the confidence interval obtained by inverting the family of test statistics $\tau(\boldsymbol{y}, \theta_0)$, it is necessary and sufficient that

$$\tau(\boldsymbol{y}, \theta_0) \leq c_\alpha.$$

Thus the limits of the confidence interval can be found by solving the equation

$$\tau(\boldsymbol{y}, \theta) = c_\alpha$$

for θ. This equation normally has two solutions. One of these solutions is the upper limit, θ_u, and the other is the lower limit, θ_l, of the confidence interval that we are trying to construct.

A random function $\tau(\boldsymbol{y}, \theta)$ is said to be pivotal for $\mathbb{M}$ if, when it is evaluated at the true value θ_μ corresponding to some DGP $\mu \in \mathbb{M}$, the result is a random variable whose distribution does not depend on μ. Pivotal functions of more than one model parameter are defined in exactly the same way. The function is merely asymptotically pivotal if only the asymptotic distribution is invariant to the choice of DGP.

Suppose that $\tau(\boldsymbol{y}, \theta)$ is an exactly pivotal function. Then the confidence interval contains the true parameter value θ_μ with probability exactly equal to $1 - \alpha$, whatever the true parameter value may be.

Even if it is not an exact pivot, the function $\tau(\boldsymbol{y}, \theta)$ must be asymptotically pivotal, since otherwise the critical value c_α would depend asymptotically on the unknown DGP in $\mathbb{M}$, and we could not construct a confidence interval with the correct coverage, even asymptotically. Of course, if c_α is only approximate, then the coverage of the interval differs from $1 - \alpha$ to a greater or lesser extent, in a manner that, in general, depends on the unknown true DGP.

Asymptotic confidence intervals

To obtain more concrete results, let us suppose that

$$\tau(\boldsymbol{y}, \theta_0) = \big((\hat{\theta} - \theta_0)/s_\theta\big)^2,$$

where $\hat{\theta}$ is an estimate of θ, and s_θ is the corresponding standard error, that is, an estimate of the standard deviation of $\hat{\theta}$. Thus $\tau(\boldsymbol{y}, \theta_0)$ is the square of the t statistic for the null hypothesis that $\theta = \theta_0$. The asymptotic critical value c_α is the $1 - \alpha$ quantile of the $\chi^2(1)$ distribution.
The equation for the limits of the confidence interval are

$$\big((\hat{\theta} - \theta)/s_\theta\big)^2 = c_\alpha.$$

Taking the square root of both sides and multiplying by s_θ then gives

$$|\hat{\theta} - \theta| = s_\theta \, c_\alpha^{1/2}.$$

As expected, there are two solutions, namely

$$\theta_l = \hat{\theta} - s_\theta \, c_\alpha^{1/2} \quad \text{and} \quad \theta_u = \hat{\theta} + s_\theta \, c_\alpha^{1/2},$$

and so the asymptotic $1 - \alpha$ confidence interval for θ is

$$\big[\hat{\theta} - s_\theta \, c_\alpha^{1/2}, \;\; \hat{\theta} + s_\theta \, c_\alpha^{1/2}\big].$$

We would have obtained the same confidence interval if we had started with the asymptotic t statistic $\tau(\boldsymbol{y}, \theta_0) = (\hat{\theta} - \theta_0)/s_\theta$ and used the $\mathrm{N}(0,1)$ distribution to perform a two-tailed test. For such a test, there are two critical values, one the negative of the other, because the $\mathrm{N}(0,1)$ distribution is symmetric about the origin. These critical values are defined in terms of the quantiles of that distribution. The relevant ones are $z_{\alpha/2}$ and $z_{1-\alpha/2}$, the $\alpha/2$ and the $1-(\alpha/2)$ quantiles, since we wish to have the same probability mass in each tail of the distribution. Note that $z_{\alpha/2}$ is negative, since $\alpha/2 < 1/2$, and the median of the $\mathrm{N}(0,1)$ distribution is 0. By symmetry, it is the negative of $z_{1-(\alpha/2)}$. The equation with two solutions is replaced by two equations, each with just one solution, as follows:

$$\tau(\boldsymbol{y}, \theta) = \pm c.$$

The positive number c can be defined either as $z_{1-(\alpha/2)}$ or as $-z_{\alpha/2}$. The resulting confidence interval $[\theta_l,\ \theta_u]$ can thus be written in two different ways:

$$\big[\hat{\theta} + s_\theta\, z_{\alpha/2},\ \ \hat{\theta} - s_\theta\, z_{\alpha/2}\big] \quad \text{and} \quad \big[\hat{\theta} - s_\theta\, z_{1-(\alpha/2)},\ \ \hat{\theta} + s_\theta\, z_{1-(\alpha/2)}\big].$$

Asymmetric confidence intervals

The confidence intervals so far constructed are **symmetric** about the point estimate $\hat{\theta}$. The symmetry is a consequence of the symmetry of the standard normal distribution and of the form of the test statistic.

It is possible to construct confidence intervals based on two-tailed tests even when the distribution of the test statistic is not symmetric. For a chosen level α, we wish to reject whenever the statistic is too far into either the right-hand or the left-hand tail of the distribution. Unfortunately, there are many ways to interpret "too far" in this context. The simplest is probably to define the rejection region in such a way that there is a probability mass of $\alpha/2$ in each tail. This is called an **equal-tailed confidence interval**. Two critical values are needed for each level, a lower one, c_α^-, which is the $\alpha/2$ quantile of the distribution, and an upper one, c_α^+, which is the $1-(\alpha/2)$ quantile. A realized statistic $\hat{\tau}$ leads to rejection at level α if either $\hat{\tau} < c_\alpha^-$ or $\hat{\tau} > c_\alpha^+$. This leads to an **asymmetric confidence interval**.

If we denote by F the CDF used to calculate critical values or P values, the P value associated with a statistic τ should be $2F(\tau)$ if τ is in the lower tail, and $2\big(1-F(\tau)\big)$ if it is in the upper tail. In complete generality, the P value is

$$p(\tau) = 2\min\big(F(\tau), 1-F(\tau)\big).$$

Consider a one-dimensional test with a rejection region containing a probability mass of β in the left tail of the distribution and γ in the right tail, for an overall level of α, where $\alpha = \beta + \gamma$. Let q_β and $q_{1-\gamma}$ be the β and $(1-\gamma)$–quantiles of the distribution of $(\hat{\theta} - \theta)/s_\theta$, where θ is the true parameter. Then

$$\Pr(q_\beta \leq (\hat{\theta} - \theta)/s_\theta \leq q_{1-\gamma}) = 1 - \alpha.$$

The inequalities above are equivalent to

$$\hat{\theta} - s_\theta q_{1-\gamma} \leq \theta \leq \hat{\theta} - s_\theta q_\beta,$$

and from this it is clear that the confidence interval $[\hat{\theta} - s_\theta q_{1-\gamma}, \hat{\theta} - s_\theta q_\beta]$ contains the true θ with probability α. Note the somewhat counter-intuitive fact that the *upper* quantile of the distribution determines the *lower* limit of the confidence interval, and *vice versa*.

Bootstrap confidence intervals

If $\tau(\boldsymbol{y}, \theta)$ is an approximately pivotal function for a model $\mathbb{M}$, its distribution under the DGPs in $\mathbb{M}$ can be approximated by the bootstrap. For each one of a set of bootstrap samples, we compute the parameter estimate, θ^* say, for each of them. Since the true value of θ for the bootstrap DGP is $\hat{\theta}$, we can use the distribution of $\theta^* - \hat{\theta}$ as an estimate of the distribution of $\hat{\theta} - \theta$. In particular, the $\alpha/2$ and $(1-\alpha/2)$–quantiles of the distribution of $\theta^* - \hat{\theta}$, $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$ say, give the **percentile confidence interval**

$$C^*_\alpha = [\hat{\theta} - q^*_{1-\alpha/2},\ \hat{\theta} - q^*_{\alpha/2}].$$

For a one-sided confidence interval that is open to the right, we use $[\hat{\theta} - q^*_{1-\alpha},\ \infty[$, and for one that is open to the left $]-\infty,\ \hat{\theta} - q^*_\alpha]$.

The percentile interval is very far from being the best bootstrap confidence interval. The first reason is that, in almost all interesting cases, the random variable $\hat{\theta} - \theta$ is not even approximately pivotal. Indeed, conventional asymptotics give a limiting distribution of $\mathrm{N}(0, \sigma^2_\theta)$, for some asymptotic variance σ^2_θ. Unless σ^2_θ is constant for all DGPs in $\mathbb{M}$, it follows that $\hat{\theta} - \theta$ is not asymptotically pivotal.

For this reason, a more popular bootstrap confidence interval is the **percentile–*t*** interval. Now we suppose that we can estimate the variance of $\hat{\theta}$, and so base the confidence interval on the **studentised** quantity $(\hat{\theta}-\theta)/s_\theta$, which in many circumstances is asymptotically standard normal, and hence asymptotically pivotal. Let $q_{\alpha/2}$ and $q_{1-\alpha/2}$ be the relevant quantiles of the distribution of $(\hat{\theta}-\theta)/\hat{\sigma}_\theta$, when the true parameter is θ. Then

$$\Pr\left(q_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\hat{\sigma}_\theta} \leq q_{1-\alpha/2}\right) = \alpha.$$

If the quantiles are estimated by the quantiles of the distribution of $(\theta^*-\hat{\theta})/\sigma^*_\theta$, where σ^*_θ is the square root of the variance estimate computed using the bootstrap sample, we obtain the percentile–t confidence interval

$$C^*_\alpha = [\hat{\theta} - \hat{\sigma}_\theta q^*_{1-\alpha/2},\ \hat{\theta} - \hat{\sigma}_\theta q^*_{\alpha/2}].$$

In many cases, the performance of the percentile–t interval is much better than that of the percentile interval. For a more complete discussion of bootstrap confidence intervals of this sort, see Hall (1992).

Equal-tailed confidence intervals are not the only ones than can be constructed using the percentile or percentile–t methods. Recall that critical values for tests at level α can be based on the β and γ–quantiles for the lower and upper critical values provided that $1 - \gamma + \beta = \alpha$. A bootstrap distribution is rarely symmetric about its central point (unless it is deliberately so constructed). The β and γ that minimise the distance between the β–quantile and the γ–quantile under the constraint $1 - \gamma + \beta = \alpha$ are then not $\alpha/2$ and $1 - \alpha/2$ in general. Using the β and γ obtained in this way leads to the *shortest* confidence interval at confidence level $1 - \alpha$.

The confidence interval takes a simple form only because the test statistic is a simple function of θ. This simplicity may come at a cost, however. The statistic $(\hat{\theta} - \theta)/\hat{\sigma}_\theta$ is a **Wald statistic**, and it is known that Wald statistics may have undesirable properties. The worst of these is that such statistics are not invariant to nonlinear reparametrisations. Tests or confidence intervals based on different parametrisations may lead to conflicting inference. See Gregory and Veall (1985) and Lafontaine and White (1986) for analysis of this phenomenon.