University of Illinois Department of Economics
Fall 1998 Econ 472 Roger Koenker

Lecture 6
The δ-Method and the Bootstrap
Introduction to Nonlinear Inference

Let me begin with a very simple inference problem which has a personal attraction to me, because it was one of the first interesting applied problems I faced (while writing my thesis). I had estimated a cost function of the quadratic form,

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + z_i'\beta + u_i \tag{1}$$

where y_i was log cost of firm i, x_i was log output and z_i was a vector of other characteristics of the i^{th} firm. It is easy to show that minimum average cost occurs at output level

$$\hat{q}^* = \exp\{(1-\hat{\alpha}_1)/2\hat{\alpha}_2\}.$$

It is easy enough to make a point estimate of this quantity, but the question of how to compute a confidence interval for this estimate is not quite as easy.

One approach is the δ-method, write $\theta = (\alpha, \beta)$ and $q^* = h(\theta)$, then the asymptotic normality of θ,

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, V)$$

where $\sigma^2(X'X)^{-1}$ implies that

$$\sqrt{n}(\hat{q}^* - q^*) \rightsquigarrow \mathcal{N}(0, \nabla h' V \nabla h)$$

where ∇h is quite easily computed. In effect we have pretended that the nonlinear function $h(\cdot)$ can be well approximated by the linear function

$$\tilde{h}(\theta) = h(\theta_0) + \nabla h(\theta_0) \cdot (\theta - \theta_0).$$

Obviously, this works asymptotically because for large n, $\hat{\theta}$ is concentrated very close to θ_0 and h is smooth, i.e., well-approximated by a linear function in a neighborhood of θ_0. However, we can get some idea of why the δ-method might perform badly by asking how linear is $h(\cdot)$ in some appropriately defined confidence region for θ. For example, we could draw a confidence ellipse for (α_1, α_2) based on F-theory and then compute $h(\cdot)$ for various values of (α_1, α_2) in this confidence region – would these values be well approximated by the tangent plane of $h(\cdot)$ at $\hat{\alpha}_1, \hat{\alpha}_2$, or not?

This suggestion contains the essential idea for various improvements. Let's begin by considering how we might go about computing an exact solution to the confidence interval problem. If we believed in the full classical linear model conditions for (1), *iid* Gausian errors, etc. etc., then we have already seen that

$$V_n^{-1/2}(\hat{\theta} - \theta_0) \sim \text{ Student}_{n-p}(0, I_p)$$

where he *rhs* denotes a multivariate Student-t random vector with mean 0 and dispersion matrix I_p and $n-p$ degrees of freedom with

$$V_n = \hat{\sigma}^2(X'X)^{-1}$$

Thus, in principle we could find the exact distribution of $h(\cdot)$ by the usual transformation formulae of the calculus. This is tedious and probably not worth the effort unless $h(\cdot)$ is something quite important that will be used repeatedly.

A simpler approach would be to approximate the distribution of $h(\cdot)$ by simulation. [Finally, we are getting closer to the bootstrap!]. How to do this? Let Z be a draw from $\text{Student}_{n-p}(0, I_p)$ then

$$\tilde{Z} = \hat{\theta} - V_n^{1/2} Z$$

has the distribution represented by the confidence region referred to above, in particular if we looked just at the two coordinates corresponding to (α_1, α_2) of $\tilde{Z}$, they would fall into the 95% confidence ellipse eluded to earlier with probability .95. Thus, suppose we now take a random sample of size R of such $\tilde{Z}$'s, denote the j^{th} one by $\tilde{Z}^j$ and compute R estimates of q^* from them:

$$\hat{q}^* = h(\tilde{Z}^j) \qquad j = 1, \ldots, R$$

and finally, imagine computing the standard deviation of these, or even better, computing the $\alpha/2^{\text{th}}$ and $(1 - \alpha/2)^{\text{th}}$ quantiles of these and defining a CI for q^* as

$$\{q^* : \ \ q^* \in (\hat{q}^*_R(\alpha/2), \hat{q}^*_R(1 - \alpha/2))\}.$$

As $R \to \infty$ these sample quantiles converge to the true quantiles of the distribution we could have computed analytically, but were too lazy to undertake. But now it is natural to object to the fact that we may not be sure about all of the assumptions which underlay the assertion that $\hat{\theta}$ had this exact Student-t distribution. What then?

Under the slightly weaker condition that the errors are *iid* but not necessarily Gaussian we might suggest the following strategy which brings us even closer to the bootstrap. What would be our best guess about the distribution of the errors under the conditions specified? Obviously,

$$\hat{F}_n(u) = n^{-1} \sum I(\hat{u}_i \leq u)$$

We can conveniently think of sampling from this distribution as simply drawing from the set $\{\hat{u}_1, \ldots, \hat{u}_n\}$, assigning probability $1/n$ to each element, *with replacement.* That is, on each draw we select an integer from 1 to n, say k, making sure that each integer is assigned probability $1/n$. Having done this n times we have a new vector of residuals

$$\check{u} = (\hat{u}_{k_1}, \hat{u}_{k_2}, \ldots, \hat{u}_{k_n})$$

then define a new y-vector

$$\check{y} = \hat{y} + \check{u} = y - \hat{u} + \check{u}$$

and compute a new least squares estimate

$$\check{\theta} = (X'X)^{-1} X' \check{y}$$

And now repeat this process R times each time getting a new $\check{\theta}$ and then computing a new value for

$$\check{q}^* = h(\check{\theta})$$

Again this yields a sample of R values of the quantity of interest which can then be used to estimate a standard error or construct a confidence interval.

Implementation: In `S` there are a number of functions which have built-in capability for bootstrapping. In addition, there are the functions provided in Davison and Hinkley (1997). The simplest things can be easily implemented using the sample command. To illustrate consider the following code fragment

```
fit_lm (y ~ x)
uhat_ fit$resiid
h_reg(0, R)
for (i in 1:R) {


     yh_fit$fit + sample (uhat, replace=T)
     b_lm(yh ~ x)$coef
     h[i]_exp((1-b[2])/2*b[3]))
     }


quantile(h, c(0.025, 0.975))
```

The bootstrap is a fascinating new topic which has sparked intense interest from both applied and theoretically inclined researchers since Efron's (1979) paper. There are at least a dozen recent monographs on the subject of which I would recommend Efron and Tibshirani (1993), Davison and Hinkley (1997). At an elementary level the paper of Efron and Gong (1983) is still useful, I believe.

Efron's bootstrap provides a very general approach to resampling which avoids some problems inherent in the systematic resampling of the jackknife. In German the expression *an dem eigenen Haaren aus dem Sumf ziehen* nicely captures the idea of the bootstrap – "to pull yourself out of the swamp by your own hair." The sample itself is used to assess the precision of the estimate $\hat{\theta}$.

I will conclude with a prototypical example of the use of the bootstrap. An enormous variety of other examples may be found in the books by Efron and Tibshirani (1993) and Davison and Hinkley (1997).

In regression we need not use the residual bootstrap on page 2. A more direct implementation of the bootstrap would be to "resample (x, y)-pairs" i.e., at each replication draw a random sample $\{k_1, k_2, \ldots, k_n\}$ with k_i's iid and uniform over the integers $1, \ldots, n$. The sample $\{(x_{k_i}, y_{k_i}) \;\; i = 1, \ldots, n\}$ can then be used to compute $\check{\beta}$ and a covariance matrix of $\hat{\beta}$ could be computed as

$$\hat{V} = R^{-1} \sum_{i=1}^{R} (\check{\beta}^i - \hat{\beta})(\check{\beta}^i - \hat{\beta})'$$

This approach is less sensitive to assumptions than the residual based bootstrap introduced earlier. In particular, it does not assume that the regression errors are iid so it can accommodate heteroscedasticity for example. Of course it does still assume that the observations are independent. Bootstrapping dependent observations is an inherently more difficult task which has generated its own rather large literature. Rather than using $\hat{V}$ to compute standard errors one could, of course, again use the percentile method directly on the bootstrap sample of $\check{\beta}^i$ vectors. This approach can be used effectively in M-estimation contexts to generate automatic versions of the Huber Sandwich. For OLS this approach approximates the Eicker-White formula.

References

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Stat*, 7, 1-26.

Efron, B. and R.J. Tibshirani (1993). *An Introduction to the Bootstrap*, Chapmall-Hall: New York.

Davison, A.C. and D.V. Hinkley (1997). *Bootstrap Method and their Application*, Cambridge U. Press: Cambridge.

Efron B. and G. Gong (1983). A leisurely look at the bootstrap, *Am. Statistician*, 37, 36-48.