

**Lecture 5**  
**The  $\delta$ -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,

$$(1) \quad y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + z_i^\top \beta + u_i$$

where  $y_i$  was log cost of firm  $i$ ,  $x_i = \log(q_i)$  was log output and  $z_i$  was a vector of other characteristics of the  $i^{\text{th}}$  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  $\delta$ -method, write  $\theta = (\alpha, \beta)$  and  $q^* = h(\theta)$ , then the asymptotic normality of  $\theta$ ,

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

where  $V = \sigma^2(X^\top X)^{-1}$  and  $X = [1, x_i, x_i^2, z_i^\top]$  implies that

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

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

$$\tilde{h}(\theta) = h(\hat{\theta}) + \nabla h(\hat{\theta})^\top (\theta - \hat{\theta}).$$

Note that the vector  $\nabla h(\hat{\theta})$  is fixed once the estimation is carried out, so the expression  $\nabla h^\top V \nabla h$  is just a scalar constant.

This works asymptotically because for large  $n$ ,  $\hat{\theta}$  is concentrated very close to  $\theta_0$  and  $h$  is smooth, i.e., well-approximated by a linear function in a neighborhood of  $\theta_0$ . However, we can get some idea of why the  $\delta$ -method

might perform badly by asking how linear is  $h(\cdot)$  in some appropriately defined confidence region for  $\theta$ . For example, we could draw a confidence ellipse for  $(\alpha_1, \alpha_2)$  based on F-theory and then compute  $h(\cdot)$  for various values of  $(\alpha_1, \alpha_2)$  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?

### Digression on confidence ellipses for regression coefficients

In light of these considerations, it is perhaps useful to review some basic facts about confidence regions for parameters in the classical linear regression setting. Suppose that we have a simple linear model with two covariates:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + u_i$$

we know that for  $u$  spherically normal,

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1})$$

so the variance of any linear contrast  $\alpha^\top \hat{\beta}$  is given by evaluating the quadratic form,  $\sigma^2 \alpha^\top (X^\top X)^{-1} \alpha$ . When  $x_1$  and  $x_2$  are positively correlated then  $\hat{\beta}_1$  and  $\hat{\beta}_2$  will be negatively correlated. This implies that we will be able to estimate the sum of the  $\beta$ 's well, but not their difference. To illustrate this effect consider the following example from Malinvaud's classic textbook. We have the following model of French imports:

$$y_t = \underset{(0.006)}{0.133}x_{1t} + \underset{(0.110)}{0.550}x_{2t} + \underset{(0.200)}{2.10}x_{4t} - \underset{(1.27)}{5.92}$$

where  $y_t$  is French imports,  $x_{1t}$  is gdp,  $x_{2t}$  is investment,  $x_{3t}$  is consumption, and  $x_{4t}$  is dummy variable for EC membership. All variables are in millions of French Francs in 1959 prices. In this model we are able to make reasonably precise estimates of the effect of growth of gdp and investment on imports with 95 percent confidence intervals (respectively)

$$\beta_1 \in (0.121, 0.145)$$

$$\beta_2 \in (0.33, 0.77)$$

However, if we introduce the consumption variable  $x_{3t}$ , we obtain,

$$y_t = \underset{(0.051)}{-0.021}x_{1t} + \underset{(0.087)}{0.559}x_{2t} + \underset{(0.077)}{0.235}x_{3t} + \underset{(0.16)}{2.10}x_{4t} - \underset{(1.38)}{9.79}$$

But now note that the confidence interval for  $\beta_1$  is  $(-.123, .081)$ . What happened? Roughly speaking, we will see that when, as in this example

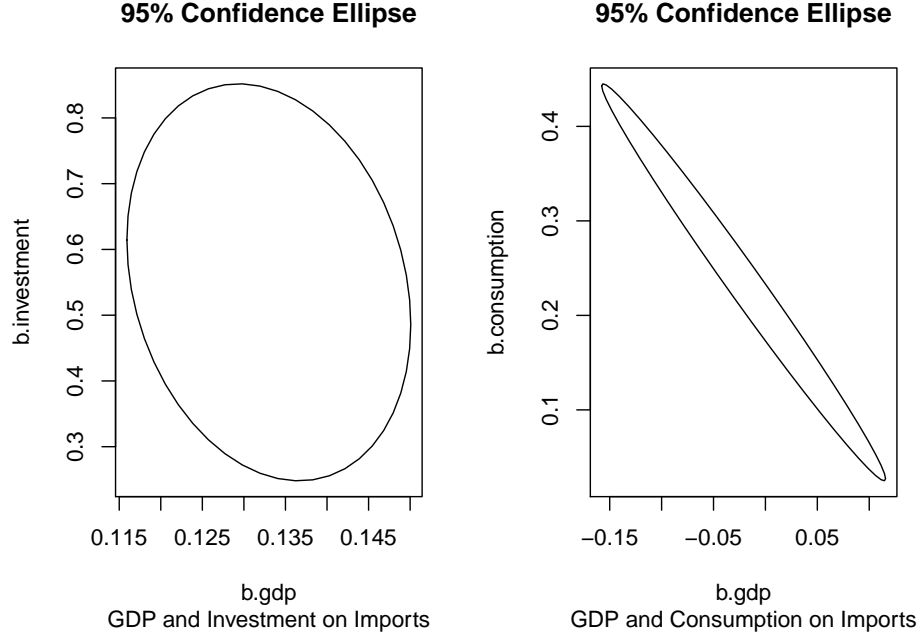


Figure 1: Confidence ellipses for two pairs of coefficients in the Malinvaud import demand equation. In the left panel the coefficients on gdp and investment are nearly independent, however in the right panel after adding consumption spending, which is quite strongly correlated with gdp, the coefficients of these two variables are very strongly negatively correlated.

the independent variables exhibit an approximately linear relationship, here  $x_3 \equiv \gamma x_1$ , then the "regression" is incapable of precisely estimating the separate effects of the two variables. This is made more explicit if we consider confidence ellipses for pairs of coefficients. Without the consumption variable we get a quite precise estimate of the gdp effect, but when we include consumption the situation changes radically – we have a very imprecise estimate of the gdp effect – even the sign of the coefficient is in doubt, and the joint confidence ellipse of the gdp and consumption coefficients is very cigar shaped. Given the orientation of the cigar it is clear that we can quite accurately estimate the effect of circumstances in which gdp and consumption move in the same direction, but we are unable to predict what would happen when they moved in opposite directions. Why?

As a second example consider the problem of jointly estimating confidence intervals for income and price elasticities of gasoline. In Figure 2 we illustrate .90 and .99 confidence ellipses for two estimated gasoline models. One is based on data from 1947-72 prior to the first oil shock, and the other is based on the entire period 1947-88. Several things are evident from the figure. First, the full data set yields much more precise estimates (smaller confidence regions). This is to be expected when there is more data, and more especially when there is more variability in the  $x$  variables, as was kindly provided by OPEC. Second, the orientation of the ellipses for the full sample is somewhat more aligned with the coordinate axes indicating that there is less correlation between the two elasticities than in the earlier period. This reflects more independent movement of prices in the OPEC period, whereas price and income were more strongly positively correlated in the earlier pre-OPEC period. Finally, and most disturbingly note that the evidence provided by the earlier period is wildly overconfident about precision of the elasticity estimates. While admitting that the price elasticity might be negative, it rules out very strongly the possibility that it could be as small as -0.50, the value obtained using the full data set. Similarly, the confidence in the lower estimate of the income elasticity is also misplaced.

Finally, to conclude this digression, let's consider the relationship between the confidence ellipses that we have seen and the conventional one dimensional confidence intervals. To fix ideas let's consider the simplest possible case: a situation in which we have a two dimensional parameter  $\beta$  that happens to be standard normal, i.e.  $\beta \sim \mathcal{N}(0, I_2)$ . This is a totally artificial situation in which we imagine that  $\hat{\beta}$  happens to take the value  $(0, 0)^\top$  and have covariance matrix,  $I_2$ . Then we have that

$$P(\beta_1^2 + \beta_2^2 < 5.99) = .95$$

since the sum of squares of the  $\beta$ 's is  $\chi_2^2$ . Thus, we get circular confidence regions and the radius of the .95 region is 2.45. Compare the area of this circle:  $\pi r^2 = 18.81$  with the area of the square formed by two .95 confidence intervals for the separate parameters: which has area  $(2 \cdot 1.96)^2 = 15.36$ . Why is this square smaller than the circle? Hint: Show that the square that contains probability .95 has area 20.05.

You can easily generalize this example to the more realistic case that  $\hat{\beta}$  is non-zero with correlated elements. In this case we can diagonalize the covariance matrix of  $\hat{\beta}$ , say  $V = PDP'$  where  $D$  is the diagonal matrix of eigenvalues and  $P$  is the matrix of eigenvectors. The eigenvectors describe the principle directions of the ellipses, and the eigenvalues represent the

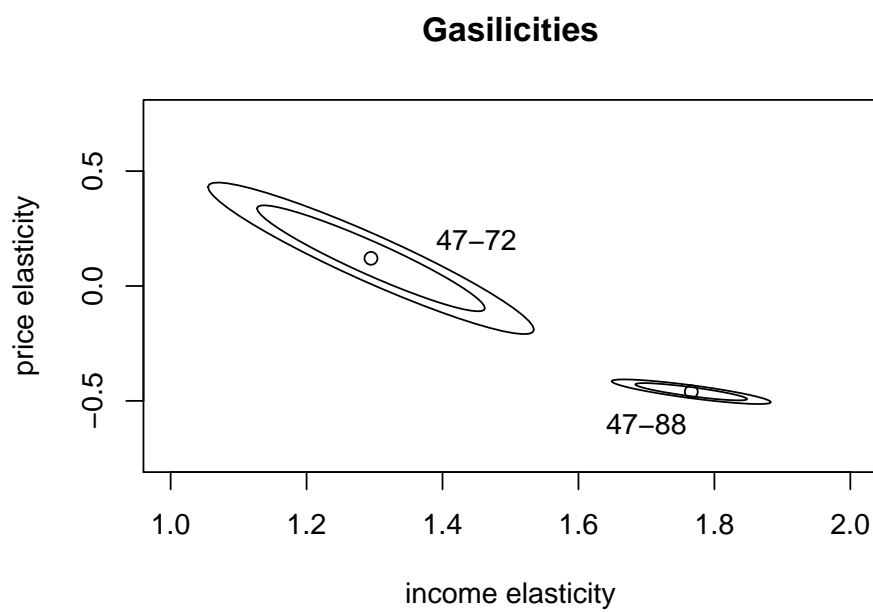


Figure 2: Confidence ellipses for income and price elasticities of gasoline in the U.S.

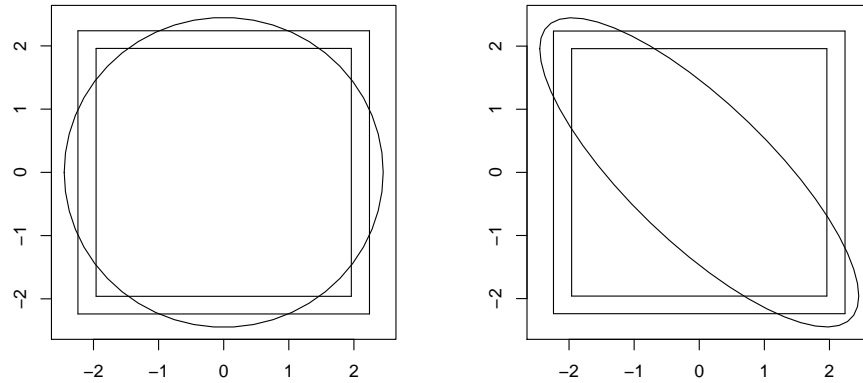


Figure 3: In the left panel the circular 0.95 confidence region for the bivariate standard normal vector is compared with two “confidence squares” – one based on univariate 0.95 intervals, the larger one on univariate 0.975 intervals. The two squares have areas 15.36 and 20.05 respectively. Which one has the same coverage probability as the circle? In the right panel the same comparison is made except that the ellipse corresponds to a 0.95 confidence region for a bivariate normal vector with unit variances, and correlation - 0.80. Now the area of the elliptical region is 15.71, and because of the unit variance assumption the squares are the same as they were in the first case, so the discrepancy between the elliptical region and the comparable square is even larger.

variability in these directions. As should be obvious at this point the discrepancy between the rectangular regions and the elliptical regions in these correlated cases can be much more extreme than in the independent case. Figure 3 illustrates these differences.

### Resampling and the Bootstrap

This suggestion at the end of the first section of this lecture 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* Gaussian 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 the *rhs* denotes a multivariate Student-t random vector with mean 0 and dispersion matrix  $I_p$  and  $n - p$  degrees of freedom with

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

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. [Now, 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  $(\alpha_1, \alpha_2)$  of  $\tilde{Z}$ , they would fall into the 95% confidence ellipse alluded to earlier with probability .95. Thus, suppose we now take a random sample of size  $R$  of such  $\tilde{Z}$ 's, denote the  $j^{\text{th}}$  one by  $\tilde{Z}^j$  and compute  $R$  estimates of  $q^*$  from them:

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

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  $R \rightarrow \infty$  these sample quantiles converge to the true quantiles of the distribution that 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  $\{\hat{u}_1, \dots, \hat{u}_n\}$ , 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

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

then define a new  $y$ -vector

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

and compute a new least squares estimate

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

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 R 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$resid
h <- rep(0, R)
for (i in 1:R) {

  yh <- fit$fit + sample (uhat, replace=TRUE)
  b <- lm(yh ~ x)$coef
  h[i] <- exp((1-b[2])/2*b[3]))
}

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



The bootstrap is a important general technique 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 among 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. It contains among other things a nice discussion of how to use the bootstrap to evaluate the fishing effect discussed in the last lecture.

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 den eigenen Haaren aus dem Sumpf zu 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  $\{k_1, k_2, \dots, k_n\}$  with  $k_i$ 's iid and uniform over the integers  $1, \dots, n$ . The sample  $\{(x_{k_i}, y_{k_i}) \mid i = 1, \dots, n\}$  can then be used to compute  $\check{\beta}$  and a covariance matrix of  $\hat{\beta}$  could be computed as

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

This is easily implemented in R in the following way:

```
bhat <- lm(y ~ x)$coef
n <- length(y)
p <- length(bhat)
R <- 500
B <- matrix(0,p,R)
for (i in 1:R) {

  s <- sample (1:n, replace=TRUE)
  B[,i] <- lm(y[s] ~ x[s,])$coef
}
```

```
Vhat <- cov(B - bhat)
```

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*, Chapman-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.