USC, Spring 2016, Economics 513

Lecture 7: Inference in the multiple linear regression model: Bootstrap.

Finding the sampling distribution of the OLS estimator

Until now we have found two methods to obtain the sampling distribution of the OLS estimator (and related statistics as the t-ratio)

1. Using the CLR assumptions and normality of the random error we derived the exact sampling distribution.
2. Under weaker assumptions on the relation between the random error and the independent variables and not assuming normality we found an approximation to the sampling distribution by using asymptotic (sample size goes to ∞) analysis.

The results were not very different if the random errors are homoskedastic: instead of the t distribution we use the standard normal and instead of the F distribution the χ^2 (for tests of more that one regression coefficient).

In this lecture we will consider a third method to find the sampling distribution: the bootstrap. This method is named after a tall tale by the (in)famous baron Von Münchhausen (1720-1797) and is just as miraculous at first sight.

The bootstrap is essentially a method to generate an asymptotic approximation without doing any math. Instead we use the computer.

The empirical cdf and the sampling distribution

For a random sample $X_1, \ldots, X_n$ we define the empirical c.d.f.

$$F_n(x) = \frac{\sharp\{X_i \leq x\}}{n}$$

This is the c.d.f. of the discrete uniform distribution that assigns probability $1/n$ to the points $X_1, \ldots, X_n$.

Because

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

with $I(.)$ the indicator function of the event between parentheses, we have by the WLLN for all x

$$F_n(x) \xrightarrow{p} \mathrm{E}[I(X_i \leq x)] = F(x)$$

with $F(x)$ the population c.d.f., i.e. the population and sample c.d.f. are the same in large samples.
Consider the sample mean of the random sample

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This is the mean of the discrete uniform distribution that assigns probability $1/n$ to $X_1, \ldots, X_n$. Expressing the mean of this discrete uniform distribution as an integral we find

$$\overline{X}_n = \int x \mathrm{d}F_n(x)$$

Here we use a different way to write the integral that corresponds to the more general definition of an integral. This will not be discussed further, and it is only used to justify the bootstrap.

Note that with this alternative expression for the integral the WLLN can be expressed as

$$\int x \mathrm{d}F_n(x) \xrightarrow{p} \int x \mathrm{d}F(x)$$

Consider the sampling variance of the sample mean $\overline{X}_n$. This is

$$\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

The numerator σ^2 can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - (\overline{X}_n)^2)^2$$

which is a biased estimator of σ^2 (to get the unbiased estimator we must divide by $n-1$).

We have

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - (\overline{X}_n)^2)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - (\overline{X}_n)^2 = \int x^2 \mathrm{d}F_n(x) - \left(\int x \mathrm{d}F_n(x)\right)^2$$

In this case the first and second moment of the empirical distribution can be computed directly, but in principle we can compute these moments by a simulation algorithm.

Estimating moments by the bootstrap

We discuss the bootstrap procedure for $S_n = \frac{1}{n}\sum_{i=1}^n X_i^2$. The procedure has the following steps:

1. Draw a sample $X_{1b}^*, \ldots, X_{nb}^*$ from the empirical distribution of $X_1, \ldots, X_n$ with replacement. For instance if $n = 3$ and $X_1 = 2, X_2 = 4, X_3 = 1$, then bootstrap samples are 2,3,2 and 4,1,3 etc.
2. Compute $S_{nb}^* = \frac{1}{n}\sum_{i=1}^n X_{ib}^{*2}$
3. Repeat this B times and compute $\hat{S}_n = \frac{1}{B}\sum_{b=1}^B S_{nb}^*$.

Note if we take the expectation with respect to the empirical distribution, then

$$E[S_{nb}^*] = \frac{1}{n}\sum_{i=1}^n E[X_{ib}^{*2}] = \frac{1}{n}\sum_{i=1}^n X_i^2$$

because

$$E[X_{ib}^{*2}] = \frac{1}{n}\sum_{i=1}^n X_i^2$$

so that the bootstrap estimator is unbiased.

Estimating a sampling variance by the bootstrap

The sample variance of the sample mean is a function of the first and second moment of F that can be estimated directly with the corresponding moments of the empirical c.d.f. Using the bootstrap to estimate these population moments is not necessary.

Now consider a statistic $T_n = g(X_1, \ldots, X_n)$ with a sampling variance that is not a simple function of moments of F. We can use the simulation algorithm to estimate its sampling variance.

1. Draw a sample $X_{1b}^*, \ldots, X_{nb}^*$ from the empirical distribution of $X_1, \ldots, X_n$ with replacement.
2. Compute $T_{nb}^* = g(X_{1b}^*, \ldots, X_{nb}^*)$
3. Repeat this B times and compute

$$\widehat{\mathrm{Var}}(T_n) = \frac{1}{B}\sum_{b=1}^{B}\left(T_{nb}^* - \frac{1}{B}\sum_{b=1}^{B}T_{nb}^*\right)^2 = \frac{1}{B}\sum_{b=1}^{B}T_{nb}^{*2} - \left(\frac{1}{B}\sum_{b=1}^{B}T_{nb}^*\right)^2$$

Why the bootstrap works

By the LLN, we have if $B \to \infty$

$$\frac{1}{B}\sum_{b=1}^{B} T_{nb}^{*2} \xrightarrow{p} E[g(X_1, \ldots, X_n)^2]$$

where the last expectation is with respect to the empirical distribution. In large samples the expected value $E[g(X_1, \ldots, X_n)^2]$ converges to the expectation with respect to the population distribution.

The bootstrap relies on the closeness of the empirical and population distribution in large samples and therefore is an asymptotic approximation. However, we use the computer and in particular a resampling algorithm to obtain an approximation to $E[g(X_1, \ldots, X_n)^2]$ that itself approximate the corresponding population expectation.

$\widehat{\mathrm{Var}}(T_n)$ is the bootstrap estimator of the sampling variance of T_n.

Bootstrap confidence intervals

Next we consider bootstrap estimation of a confidence interval. We have a random sample $X_1, \ldots, X_n$ from a distribution with cdf $F(x, \theta)$ with θ a scalar parameter.

We already have an estimator $\hat{\theta}_n$ and we want to find an $1 - \alpha$ confidence interval for θ with e.g. $\alpha = .05$.

Define $R_n = \hat{\theta}_n - \theta$ and let the (unknown) sampling distribution of R_n have c.d.f. H. Define also

$$a_n = \hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad b_n = \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)$$

Note that a_n, b_n are random variables. You can check that the population probability

$$\Pr(a_n \leq \theta \leq b_n) = 1 - \alpha$$

We do not now H but we can estimate H using B bootstrap replications $R_{nb}^* = \hat{\theta}_{nb}^* - \hat{\theta}_n$

$$\hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R_{nb}^* \leq r)$$

Hence we can estimate a_n, b_n by

$$\hat{a}_n = \hat{\theta}_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) \qquad \hat{b}_n = \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right)$$

The β quantile of the bootstrap distribution of R^*_{nb} is

$$\Pr(R^*_{nb} \leq r^*_\beta) = \beta \Leftrightarrow \Pr(\hat{\theta}^*_{nb} \leq r^*_\beta + \hat{\theta}_n) = \beta$$

so that the β quantile of the bootstrap distribution of $\hat{\theta}^*_{nb}$ is $\theta^*_\beta = r^*_\beta + \hat{\theta}_n$.

Therefore we have the alternative estimates

$$\hat{a}_n = 2\hat{\theta}_n - \theta^*_{1-\frac{\alpha}{2}} \qquad \hat{b}_n = 2\hat{\theta}_n - \theta^*_{\frac{\alpha}{2}}$$

Bootstrapping OLS

We consider the case of cross-sectional data, so that we have a random sample $y_i, x'_{1i}, i = 1, \ldots, N$ of observations on a dependent variable y and $K-1$ independent variables. Define $x'_i = (1\ x'_{1i})$ so that the OLS estimator is

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x'_i \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right)$$

The bootstrap estimator of the sampling variance can be computed as follows

1. Draw a sample $y^*_{1b}, x^{*'}_{11b}, \ldots, y^*_{nb}, x^{*'}_{1nb}$ from the empirical distribution of $y_1, x'_{11}, \ldots, y_n, x'_{1n}$ with replacement.
2. Compute
$$\hat{\beta}^*_b = \left(\sum_{i=1}^{n} x^*_{ib} x^{*'}_{ib} \right)^{-1} \left(\sum_{i=1}^{n} x^*_{ib} y^*_{ib} \right)$$
3. Repeat this B times and compute
$$\widehat{\mathrm{Var}}(\hat{\beta}) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\beta}^*_b - \overline{\hat{\beta}^*} \right) \left(\hat{\beta}^*_b - \overline{\hat{\beta}^*} \right)'$$
with
$$\overline{\hat{\beta}^*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}^*_b$$

This is called the nonparametric bootstrap
The parametric bootstrap uses the OLS residuals $e_i = y_i - x'_i \hat{\beta}, i = 1, \ldots, n$.

Step 1 in the previous algorithm is replaced by:

Draw a sample $e^*_{1b}, x^{*'}_{11b}, \ldots, e^*_{nb}, x^{*'}_{1nb}$ from the empirical distribution of $e_1, x'_{11}, \ldots, e_n, x'_{1n}$ with replacement. Compute $y^*_{ib} = x^{*'}_{ib} \hat{\beta} + e^*_{ib}$ for $i = 1, \ldots, n$ to obtain the bootstrap sample.

Because this essentially assigns residuals randomly to observations, the underlying assumption is that the random error and the independent variables are stochastically independent. We expect that the parametric bootstrap does not work well if there is e.g. heteroskedasticity.

Bootstrap confidence intervals for regression coefficients

We can compute bootstrap confidence intervals for regression coefficients as explained above.

To see whether they improve on the asymptotic confidence intervals we perform an experiment.

We use a very large dataset, the census dataset used by Angrist and Krueger in their study of the returns to education.

We treat the 329509 observations as the population. We consider the relation between log earnings and years of education. The OLS estimate (population relation) is

$$\begin{array}{rcccc} \widehat{\log(\text{earn})}_i & = & 4.9952 & + & 0.0709 \cdot \text{educ}_i \\ & & (0.00045) & & (0.0003) \end{array}$$

We draw 5,000 random samples (with replacement, although this does not matter at all given the size of the population), of size n (for $n = 20$, $n = 100$, and $n = 500$).

For each random sample we estimate the same linear regression and calculate 95% and 90% confidence intervals for the coefficient on education in four different ways: (*i*) using conventional OLS standard errors, (*ii*) using robust, i.e. heteroskedasticity-consistent OLS standard errors, (*iii*) parametric bootstrap, and (*iv*) nonparametric bootstrap.

For each computed confidence interval we check whether the "true value" (0.709) is in it and we calculate how often that happens over the 5,000 replications. The results are in Table 1.

Conclusion: With 500 observations the robust and nonparametric bootstrap based intervals are very accurate, in contrast to the conventional and parametric bootstrap based intervals. With smaller sample sizes all intervals deteriorate. The conventional intervals end up being superior to the robust intervals for small sample sizes. The nonparametric bootstrap is always better than the parametric bootstrap.

Table 1: Actual versus Nominal Coverage Rates

	95% confidence interval				90% confidence interval			
	convent	robust	par boot	nonpar boot	convent	robust	par boot	nonpar boot
$n = 20$	0.9072	0.8819	0.8898	0.9353	0.8466	0.8173	0.8274	0.8847
$n = 100$	0.9155	0.9378	0.9140	0.9437	0.8562	0.8808	0.8523	0.8903
$n = 500$	0.9284	0.9502	0.9274	0.9510	0.8693	0.9051	0.8681	0.9060