Contents

1 The Jackknife **3**

1.1 The framework and some intuition for the Jackknife estimators 3

1.2 Examples of Application 6

1.3 Consistency results 7

1.3.1 Consistency of the Jackknife variance estimator 8

1.3.2 Consistency of the Jackknife bias estimator 11

1.4 Extensions 15

2 The General Theory of the Bootstrap **17**

2.1 The framework and some intuition for the Bootstrap 17

2.2 Consistency of the Bootstrap 22

2.3 Bootstrap Confidence Intervals 23

2.4 Bootstrap hypothesis tests 27

2.5 The Bootstrap Delta Method 30

3 The Pairwise Bootstrap **35**

3.1 Construction of a Probability in the Pairwise Bootstrap Scheme 35

3.2 The Pairwise Bootstrap CLT 37

3.3 Applications to the Main Examples 39

3.4 Inference for a Non-Parametric Density (optional) 41

3.5 Pairwise Bootstrap for M Estimators 46

3.6 Pairwise Bootstrap for GMM 47

3.7 When the Pairwise Bootstrap is Inconsistent 48

4 The Residual, the Wild, and the Weighted Bootstrap **51**

4.1 The Residual Bootstrap 51

4.1.1 The residual bootstrap with heteroskedasticity 55

4.2 The Wild Bootstrap 56

A Background Statistical Results **61**

Rappel: Let $(Z_n)_n$ be a sequence of r.v. and let $(r_n)_n$, $r_n \in \mathbb{R}$

(I) we say that $\{Z_n\}_n$ is bounded in probability if

$$\forall \varepsilon > 0, \exists M \text{ s.t. } P(|Z_n| > M) < \varepsilon$$

we write $Z_n = O_p(1)$

(II) we write $Z_n = O_p\left(\frac{1}{r_n}\right)$ if $Z_n \cdot r_n = O_p(1)$

(III) we write $Z_n = o_p(1)$ if $Z_n \to 0$ in probability

that is: $\forall \varepsilon > 0$, $P(|Z_n| > \varepsilon) \to 0$

(IV) we write $Z_n = o_p\left(\frac{1}{r_n}\right)$ if $Z_n \cdot r_n = o_p(1)$

<u>Properties</u>.

Let $(a_n)_n$, $(b_n)_n$ be two sequences of real numbers.

(I) $O_p(a_n) \cdot O_p(b_n) = O_p(a_n b_n)$

$Z_n = O_p(a_n)$, $U_n = O_p(b_n)$ then $Z_n U_n = O_p(a_n b_n)$

(II) $O_p(a_n) \cdot o_p(b_n) = o_p(a_n b_n)$

$Z_n = O_p(a_n)$, $U_n = o_p(b_n)$ then $Z_n U_n = o_p(a_n b_n)$

(III) If $Z_n = o_p(a_n)$ with $a_n \to 0$, then $Z_n = o_p(1)$

(IV) If $Z_n = O_p(a_n)$ with $a_n \to 0$, then $Z_n = o_p(1)$

(V) $O_p(1) + o_p(1) = O_p(1)$

(VI) $z_n = o_p(1)$ then $z_n = O_p(1)$

But the converse is not necessarily true.

Chapter 1

The Jackknife

The Jackknife is a tool for estimating the bias and variance of a statistic. It is based on the principle of repeatedly computing the statistic by removing at each time one observation. In this chapter, we introduce the theory for the Jackknife and prove consistency results for variance and bias estimation. The statistics considered are smooth functions of the sample mean. Such a class covers a wide range of estimators encountered in statistic. The organization of this chapter is as follows: Section 3.1 introduces the statistical framework and provides the intuition behind the bias and variance jackknife estimators; Section 1.2 contains some examples of applications; Section 1.3 proves the consistency of the jackknife variance and bias estimators; finally, Section 1.4 discusses some extensions.

1.1 The framework and some intuition for the Jackknife estimators

We observe an iid sample $\{X_i\}_{i=1}^n$ and want to make inference on an unknown population parameter θ. To this end, we use a statistic (an estimator)

$$T_n := T_n(X_1, .., X_n) \ .$$

The Jackknife for bias reduction. It often turns out that T_n is a biased estimator of θ and the Jackknife can be used to reduce its bias. To see why, let us start by introducing the bias of T_n

$$Bias(T_n) := \mathbb{E}\{T_n\} - \theta \ .$$

The Jackknife estimator of the bias of T_n is formally defined as

$$b_{jack} := (n-1)\left(\overline{T}_n - T_n\right), \text{ where } \overline{T}_n := \frac{1}{n}\sum_{i=1}^{n} T_{n-1,i} \tag{1.1}$$

and $T_{n-1,i}$ is the statistic computed by dropping the ith observation from the sample, i.e.

$$T_{n-1,i} := T_{n-1}(X_1, .., X_{i-1}, X_{i+1}, .., X_n) .$$

An estimator with the above structure is often called "leave-one-out" estimator. The *bias-corrected jackknife estimator* is therefore

$$T_{jack} := T_n - b_{jack} . \tag{1.2}$$

T_{jack} will admit a lower bias compared to T_n. To heuristically justify this, assume that the bias of the statistic admits the following expansion:

$$\text{Bias}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) . \tag{1.3}$$

The above expansion can be often derived after imposing smoothness conditions on the functional form of T_n. Since our present purpose is only to give a heuristic justification for using b_{jack}, we will not enter such details. By (1.2) we obtain

$$\begin{aligned}\text{Bias}(T_{jack}) =& \text{Bias}(T_n) - \mathbb{E}\{b_{jack}\} \\ =& \text{Bias}(T_n) - (n-1)\left[\mathbb{E}\{\overline{T}_n\} - \mathbb{E}\{T_n\}\right] .\end{aligned}$$

The iid assumption ensures that $\mathbb{E}\{\overline{T_n}\} = \mathbb{E}\{T_{n-1,1}\}$, so $\mathbb{E}\{\overline{T}_n\} - \mathbb{E}\{T_n\} = \text{Bias}(T_{n-1,1}) - \text{Bias}(T_n)$. Since Equation (1.3) also holds for the statistic $T_{n-1,1}$, the quantity in the square brackets on the RHS equals

$$\begin{aligned}\text{Bias}(T_{n-1,1}) - \text{Bias}(T_n) =& \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right) - \frac{a}{n} - \frac{b}{n^2} - O\left(\frac{1}{n^3}\right) \\ =& \frac{a}{n(n-1)} + b\frac{2n-1}{n^2(n-1)^2} + O\left(\frac{1}{n^3}\right) .\end{aligned}$$

Gathering together the previous two displays gives

$$\begin{aligned}\text{Bias}(T_{jack}) &= \text{Bias}(T_n) - \frac{a}{n} - b\frac{2n-1}{n^2(n-1)} + O\left(\frac{1}{n^3}\right) \\ &= \frac{b}{n^2} - b\frac{2n-1}{n^2(n-1)}O\left(\frac{1}{n^3}\right) \\ &= O\left(\frac{1}{n^2}\right) .\end{aligned}$$

By comparing the above display with (1.3) we obtain that the bias of T_{jack} is of lower order than the bias of T_n. This justifies T_{jack} as a bias-reduced version of T_n and b_{jack} as an estimator of $\text{Bias}(T_n)$.

Remark 1.1.1. The argument provided here is heuristic. In practice, the superior finite sample bias properties of T_{jack} compared to T_n are shown in simulations.

Remark 1.1.2. Notice that the b_{jack} does not require to know the analytic expression of $\text{Bias}(T_n)$ and can be computed in a straightforward way.

The consistency of b_{jack} as an estimator of $\text{Bias}(T_n)$ will be proved in Section 1.3

The Jackknife for variance estimation. The jackknife can also be used to estimate the (asymptotic) variance of T_n. It is often the case that $\sqrt{n}(T_n - \theta) \rightsquigarrow \mathcal{N}(0, V)$, where V denotes the asymptotic variance of $\sqrt{n}T_n$. In some cases, the estimation of V is difficult to implement as V has a difficult expression, or the available estimators of V do note perform well in finite sample, see Section 1.2. The Jackknife provides a valuable alternative by giving an *automatic* method for the estimation of V.
To construct the Jackknife variance estimation, we start by expressing the bias reduced estimator T_{jack} as a sample mean. From the definitions of T_{jack} in (1.2) and b_{jack} in (1.1) we get

$$\begin{aligned} T_{jack} &= T_n - (n-1)\left[\overline{T_n} - T_n\right] \\ &= nT_n - (n-1)\overline{T_n} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[nT_n - (n-1)T_{n-1,i}\right] \\ &=: \frac{1}{n}\sum_{i=1}^{n}\widetilde{T}_{n,i} \ . \end{aligned}$$

The construction of the Jackknife variance estimator starts from the following idea: the variance of T_{jack} can approximate the variance of T_n. So, if $\widetilde{T}_{n,i}$ with $i = 1, \ldots, n$, can roughly be thought as being iid, then $Var(T_{jack}) \approx (1/n)Var(\widetilde{T}_{n,1})$. Thus, the jackknife variance estimator is defined as

$$\begin{aligned} \upsilon_{jack} &:= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\widetilde{T}_{n,i} - \frac{1}{n}\sum_{j=1}^{n}\widetilde{T}_{n,j}\right]^2 \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[nT_n - (n-1)T_{n-1,i} - \frac{1}{n}\sum_{j=1}^{n}\left(nT_n - (n-1)T_{n-1,j}\right)\right]^2 \\ &= \frac{n-1}{n}\sum_{i=1}^{n}\left[T_{n-1,i} - \frac{1}{n}\sum_{j=1}^{n}T_{n-1,j}\right]^2 \ . \end{aligned} \tag{1.4}$$

Similarly to b_{jack}, υ_{jack} does not require knowing the analytic expression of the (asymptotic) variance of $\sqrt{n}T_n$ and can be straightforwardly computed.

1.2 Examples of Application

OLS inference with heteroskedasticity. We observe $\{Y_i, Z_i\}_{i=1}^n$, with Z_i valued in $\mathbb{R}^k$ and Y_i real valued. We are interested in making inference on the coefficients of the linear projection of Y onto X. So, consider the regression

$$Y_i = Z^T\beta_0 + \varepsilon \text{ with } \mathbb{E}\varepsilon Z = 0$$

Assume that $\mathbb{E}ZZ^T$ is invertible (i.e. there is no collinearity between the explanatory variables Z). Then, $\beta_0 = (\mathbb{E}ZZ^T)^{-1}\mathbb{E}ZY$ and the OLS estimator is

$$\widehat{\beta} := \left(\overline{ZZ^T}\right)^{-1}\overline{ZY}\,.$$

When the errors are heteroskedastic, $\mathbb{E}\{\varepsilon^2|Z\}$ is not constant, and from basic statistical textbooks we know that

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}(0, \Sigma) \text{ with } \Sigma = (\mathbb{E}ZZ^T)^{-1}(\mathbb{E}\varepsilon^2 ZZ^T)(\mathbb{E}ZZ^T)^{-1}\,,$$

where $\rightsquigarrow$ denotes convergence in distribution (i.e., weak convergence). "Dividing" by the asymptotic standard deviation yields an asymptotically pivotal statistic[1]

$$\widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}(0, I)$$

where $\widehat{\Sigma}^{-1/2}$ is the unique matrix satisfying $\widehat{\Sigma}^{-1/2}\widehat{\Sigma}^{-1/2} = \widehat{\Sigma}^{-1}$ and

$$\widehat{\Sigma} = \left(\overline{ZZ^T}\right)^{-1}\left(\overline{ZZ^T\widehat{\varepsilon}^2}\right)\left(\overline{ZZ^T}\right)^{-1} \text{ with } \widehat{\varepsilon}_i = Y_i - Z_i^T\widehat{\beta}\,.$$

$\widehat{\Sigma}$ is called "Heteroskedasticity Consistent Covariance matrix Estimator", in the sense that it estimates consistently the asymptotic variance of $\widehat{\beta}$ with and without heteroskedasticity, and it hence allows a correct inference on β_0. Such an estimator is also known as "Huber-White" estimator, from the names of the authors who proposed it. Despite this robusteness feature, a problem with such an estimator is that is might behave poorly in finite samples. The jackknife provides a valuable alternative. From Equation (1.4), the Jackknife variance estimator in this context writes as

$$\frac{n-1}{n}\sum_{i=1}^n\left[\widehat{\beta}_i - \frac{1}{n}\sum_{j=1}^n\widehat{\beta}_j\right]^2$$

[1]Recall that a statistic is asymptotically pivotal when its asymptotic distribution does not depend on unknown parameters.

where $\widehat{\beta}_j$ represents the OLS estimator based on all observations except the jth. Then, inference on β_0 can be based on

$$\frac{\widehat{\beta} - \beta_0}{\sqrt{\upsilon_{jack}}}\,.$$

Inference when the variance has a difficult expression. The jackknife can also be employed when the asymptotic variance has a difficult expression. Several examples are reported in Shao and Tu (1995), Chapter 2.

Bias-improved 2SLS estimator. Two-stages least squares estimators (2SLS) are used in quasi-experimental studies to estimate treatment effects with imperfect compliance, i.e. Local Average Treatment Effects (LATE). Consider a linear model with endogenous regressors

$$Y = Z^T\beta_0 + \varepsilon \text{ with } \mathbb{E}W\varepsilon = 0 \text{ and } \mathbb{E}Z\varepsilon \neq 0\,,$$

where Y is a scalar, Z, W are two vectors and ε is unobserved. W is an instrumental variable. The coefficient β_0 is identified as

$$\beta_0 = \left[(\mathbb{E}ZW^T)(\mathbb{E}WW^T)^{-1}(\mathbb{E}WZ^T)\right]^{-1}(\mathbb{E}ZW^T)(\mathbb{E}WW^T)^{-1}(\mathbb{E}WY)$$

and the 2SLS estimator is

$$\widehat{\beta} = \left[(\overline{ZW^T})(\overline{WW^T})^{-1}(\overline{WZ^T})\right]^{-1}(\overline{ZW^T})(\overline{WW^T})^{-1}(\overline{WY})\,.$$

Differently from the OLS, the 2SLS estimator is biased. The jackknife can be used to reduce its bias. The bias formula in (1.1) adapted to $\widehat{\beta}$ is

$$b_{jack} := \frac{n-1}{n}\sum_{j=1}^{n}\left(\widehat{\beta}_j - \widehat{\beta}\right)$$

where $\widehat{\beta}_j$ denotes the 2SLS estimtor that uses all the observations except the jth. The bias-corrected 2SLS estimator based on the jackknife is then $\widehat{\beta} - b_{jack}$.

1.3 Consistency results

In this section we provide the consistency results for υ_{jack} and b_{jack}. We will assume that the population parameter θ takes the following form

$$\theta = g(\mu) \text{ where } \mu = \mathbb{E}\{X\}$$

and g is a fixed function. The estimator T_n is a function of the sample mean

$$T_n := g(\overline{X}_n) \text{ , where } \overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \text{ .}$$

Assumption A. *$\mathbb{E}XX^T$ is well defined and finite.*

Assumption B. *g is twice continuously differentiable in a neighborhood of μ.*

1.3.1 Consistency of the Jackknife variance estimator

Before turning to the consistency of υ_{jack} we show how such an estimator can be used to construct confidence intervals about θ. Under Assumption A and B, using a second-order Taylor expansion (see Proposition A.0.11) yields that with probability approaching one (wpa1)

$g(a) = g(b) + \nabla g(b)^T \cdot (a-b) + \frac{(a-b)^T \cdot \nabla^2 g(\xi)(a-b)}{2}$

$$\begin{aligned} T_n - \theta =& \nabla g(\mu)^T(\overline{X}_n - \mu) \\ &+ \frac{1}{2}(\overline{X}_n - \mu)^T \nabla^2 g(\mu)(\overline{X}_n - \mu) \\ &+ \frac{1}{2}(\overline{X}_n - \mu)^T \left[\nabla^2 g(\xi_n) - \nabla^2 g(\mu)\right](\overline{X}_n - \mu) \end{aligned} \tag{1.5}$$

$\xi \in [a,b]$

$\mu = \mathbb{E}(\overline{X}_n)$

$\theta = g(\mu) = g(\mathbb{E}(\overline{X}_n))$

CLT: $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$

with $\xi_n \in [\overline{X}_n, \mu]$. By the imposed assumptions, the CLT yields that $\overline{X}_n - \mu = O_P(n^{-1/2})$. By the LLN and the Continuous Mapping Theorem $\nabla^2 g(\xi_n) - \nabla^2 g(\mu) = o_P(1)$. Thus, the above display ensures that

Recall: if $z_n \xrightarrow{d} \xi$ then $z_n = O_P(1)$

$\overline{X}_n \xrightarrow{d} \mu$

$$\sqrt{n}(T_n - \theta) = \nabla g(\mu)^T \sqrt{n}(\overline{X}_n - \mu) + o_P(1) \rightsquigarrow \mathcal{N}(0, V), \text{ with } V := \nabla g(\mu)^T Var(X) \nabla g(\mu)$$

covariance matrix

$\sqrt{n}(\overline{X}_n - \mu) \to \mathcal{N}(0, Var(X))$ $\quad \nabla g(\mu)^T \cdot \sqrt{n}(\overline{X}_n - \mu) \to \mathcal{N}(0, V)$

and

$$\frac{T_n - \theta}{\sigma_n} \rightsquigarrow \mathcal{N}(0,1), \text{ where } \sigma_n^2 := n^{-1} \nabla g(\mu)^T Var(X) \nabla g(\mu) \text{ .} \tag{1.6}$$

If υ_{jack} is consistent for σ_n^2, i.e. $\upsilon_{jack}/\sigma_n^2 \xrightarrow{P} 1$, then by Slutsky's theorem

$$\frac{T_n - \theta}{\sqrt{\upsilon_{jack}}} = \frac{\sigma_n}{\sqrt{\upsilon_{jack}}} \cdot \frac{T_n - \theta}{\sigma_n} \rightsquigarrow \mathcal{N}(0,1) \text{ .}$$

Hence confidence intervals for θ can be easily constructed. The following Theorem shows the consistency of υ_{jack}.

Proposition 1.3.1. *Under Assumptions A and B*

$$\boxed{\frac{\upsilon_{jack}}{\sigma_n^2} \xrightarrow{P} 1}$$

Proof. We first prove some preliminary results that will be used multiple times in the

present proof. Let $\overline{X}_{n-1,i} = \sum_{j\neq i} X_j/(n-1)$. Notice that

$$\begin{aligned}\overline{X}_{n-1,i} - \overline{X}_n &= \left(\frac{1}{n-1} - \frac{1}{n}\right)\sum_j X_j - \frac{X_i}{n-1} \\ &= \frac{\overline{X}_n}{n-1} - \frac{X_i}{n-1} \ . \qquad \checkmark \end{aligned} \tag{1.7}$$

So, by Assumption A (see the comments below)

$$\begin{aligned}(n-1)\sum_i ||\overline{X}_{n-1,i} - \overline{X}_n||^2 &= \frac{1}{n-1}\sum_i ||\overline{X}_n - X_i||^2 \\ &= \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^T(X_i - \overline{X}_n) \\ &= \text{Trace}\left(\frac{1}{n-1}\sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right) \\ &= \text{Trace}\left(\text{Var}(X)\right) + o_P(1) \ , \end{aligned} \tag{1.8}$$

where the third equality has used the fact that $a^T a = \text{Trace}(a^T a) = \text{Trace}(a\ a^T)$ for any vector a.

We now enter the core of the theorem. The LLN ensures that $||\overline{X}_n - \mu|| = o_P(1)$. By the above display, $\max_{i=1,..,n} ||\overline{X}_{n-1,i} - \overline{X}_n||^2 \leq \sum_{i=1}^n ||\overline{X}_{n-1,i} - \overline{X}_n||^2 = O_P(n^{-1})$. This implies that $\max_{i=1,..,n} ||\overline{X}_{n-1,i} - \mu|| = o_P(1)$. Hence, Assumption B and a 1st order Taylor expansion yield that wpa1

$$\begin{aligned} T_{n-1,i} - T_n =& g(\overline{X}_{n-1,i}) - g(\overline{X}_n) \\ =& \nabla g(\overline{X}_n)^T(\overline{X}_{n,i} - \overline{X}_n) + R_{n,i} \ , \\ \text{where } R_{n,i} :=& \left[\nabla g(\xi_{n,i}) - \nabla g(\overline{X}_n)\right]^T(\overline{X}_{n-1,i} - \overline{X}_n) \end{aligned} \tag{1.9}$$

and $\xi_{n,i} \in [\overline{X}_{n-1,i}, \overline{X}_n]$. By Equation (1.7) $\sum_i(\overline{X}_{n-1,i} - \overline{X}_n) = 0$, so using (1.9)

$$\frac{1}{n}\sum_i (T_{n,i} - T_n) = \frac{1}{n}\sum_i R_{n,i} =: \overline{R}_n \tag{1.10}$$

We thus obtain

$$
\begin{aligned}
v_{jack} =& \frac{n-1}{n}\sum_i \left[T_{n-1,i} - \frac{1}{n}\sum_j T_{n-1,j}\right]^2 \text{ (by (1.4))} \\
=& \frac{n-1}{n}\sum_i \left[T_{n-1,i} - T_n - \frac{1}{n}\sum_j (T_{n,j} - T_n)\right]^2 \\
=& \frac{n-1}{n}\sum_i \left[\nabla g(\overline{X}_n)^T(\overline{X}_{n-1,i} - \overline{X}_n) + R_{n,i} - \overline{R}_n\right]^2 \text{ (by (1.10) and (1.9))} \\
=& \frac{n-1}{n}\nabla g(\overline{X}_n)^T \cdot \sum_i (\overline{X}_{n-1,i} - \overline{X}_n)(\overline{X}_{n-1,i} - \overline{X}_n)^T \cdot \nabla g(\overline{X}_n) \\
&+ \frac{n-1}{n}\sum_i (R_{n,i} - \overline{R}_n)^2 \\
&+ 2\frac{n-1}{n}\sum_i (R_{n,i} - \overline{R}_n)(\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla g(\overline{X}_n) =: A_n + B_n + 2C_n \ .
\end{aligned}
$$

By the previous display, the result is proved if $A_n/\sigma_n^2 \xrightarrow{P} 1$, $B_n/\sigma_n^2 \xrightarrow{P} 0$, and $C_n/\sigma_n^2 \xrightarrow{P} 0$.

(I) We show that $A_n/\sigma_n^2 \xrightarrow{P} 1$.

By using the definition of A_n, the definition of σ_n^2, and (1.7) we obtain

$$
\begin{aligned}
\frac{A_n}{\sigma_n^2} =& \frac{(n-1)\cdot \nabla g(\overline{X}_n)^T \cdot \sum_i (\overline{X}_{n-1,i} - \overline{X}_n)(\overline{X}_{n-1,i} - \overline{X}_n))^T \cdot \nabla g(\overline{X}_n)}{\nabla g(\mu)^T \mathrm{Var}(X) \nabla g(\mu)} \\
=& \frac{\nabla g(\overline{X}_n)^T \cdot (n-1)^{-1}\sum_i (\overline{X}_n - X_i)(\overline{X}_n - X_i)^T \cdot \nabla g(\overline{X}_n)}{\nabla g(\mu)^T \mathrm{Var}(X) \nabla g(\mu)} \text{ (by (1.7))} \\
=& 1 + o_P(1)
\end{aligned}
$$

where the last line follows from $\nabla g(\overline{X}_n) = \nabla g(\mu) + o_P(1)$ and $(n-1)^{-1}\sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T = \mathrm{Var}(X) + o_P(1)$.

(II) We show that $B_n/\sigma_n^2 \xrightarrow{P} 0$.

By definition of B_n and σ_n^2 we get

$$
\begin{aligned}
\frac{B_n}{\sigma_n^2} =& \frac{n-1}{\sigma_n^2} \cdot \frac{1}{n}\sum_i (R_{n,i} - \overline{R}_n)^2 \\
\leq& \frac{n-1}{\sigma_n^2} \cdot \frac{1}{n}\sum_i R_{n,i}^2 \\
\leq& \frac{n-1}{\nabla g(\mu)^T \mathrm{Var}(X) \nabla g(\mu)} \sum_i R_{n,i}^2 \\
\leq& \frac{n-1}{\nabla g(\mu)^T \mathrm{Var}(X) \nabla g(\mu)} \sum_i ||\nabla g(\xi_{n,i}) - \nabla g(\overline{X}_n)||^2 \ ||\overline{X}_{n-1,i} - \overline{X}_n||^2
\end{aligned}
$$

where the last line follows from the definition of $R_{n,i}$ in (1.9) and the Cauchy-Schwartz inequality. From Equation (1.8), to prove the desired result it suffices to show that

$\max_{i=1,..,n} |||\nabla g(\xi_{n,i}) - \nabla g(\overline{X}_n)|| = o_P(1)$. To this end, notice that

$$\max_{i=1,..,n} ||\nabla g(\xi_{n,i}) - \nabla g(\overline{X}_n)||^2 \leq 4 \cdot \max_{i=1,..,n} ||\nabla g(\xi_{n,i}) - \nabla g(\mu)||^2 + 4 \cdot ||\nabla g(\overline{X}_n) - \nabla g(\mu)||^2 .$$

By the continuity of ∇g at μ, $\overline{X}_n = \mu + o_P(1)$, and the continuous mapping theorem, the last term on the RHS of the above display is $o_P(1)$. For the first term, using again the continuity of ∇g at μ, for an arbitrary $\epsilon > 0$ there exists a $\delta > 0$ such that if $||x - \mu|| < \delta$ then $||\nabla g(x) - \nabla g(\mu)|| < \epsilon$. Hence,

$$P\left(\max_{i=1,..,n} ||\nabla g(\xi_{n,i}) - \nabla g(\mu))|| \geq \epsilon\right) \leq P\left(\max_{i=1,..,n} ||\xi_{n,i} - \mu|| \geq \delta\right) . \tag{1.11}$$

Since $\xi_{n,i} \in [\overline{X}_{n-1,i}, \overline{X}_n]$ we have that $\max_{i=1,..,n} ||\xi_{n,i} - \mu|| \leq \max_{i=1,..,n} ||\overline{X}_{n-1,i} - \overline{X}_n|| + ||\overline{X}_n - \mu||$, with $||\overline{X}_n - \mu|| = O_P(n^{-1/2})$ (by the CLT) and $\max_{i=1,..,n} ||\overline{X}_{n-1,i} - \overline{X}_n||^2 \leq \sum_i ||\overline{X}_{n-1,i} - \overline{X}_n||^2 = O_P(n^{-1})$ (see Equation (1.8)). So,

$$\max_{i=1,...,n} \|\xi_{i,n} - \mu\| = O_P(n^{-1/2}) , \tag{1.12}$$

the RHS of (1.11) is $o(1)$, and we conclude.

(III) We prove that $C_n/\sigma_n^2 \xrightarrow{P} 0$.
From the definition of C_n and the Cauchy-Schwartz inequality

$$\begin{aligned}\left|\frac{C_n}{\sigma_n^2}\right| &= \frac{n-1}{n}\left|\sum_i \frac{(R_{n,i} - \overline{R}_n)}{\sigma_n}\frac{(\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla g(\overline{X}_n)}{\sigma_n}\right| \\ &\leq \left\{\sum_i \frac{(R_{n,i} - \overline{R}_n)^2}{\sigma_n^2}\right\}^{1/2} \cdot \left\{\sum_i \frac{\left[(\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla g(\overline{X}_n)\right]^2}{\sigma_n^2}\right\}^{1/2} \\ &\leq \left\{\frac{n}{n-1}\frac{B_n}{\sigma_n^2}\right\}^{1/2} \cdot \left\{\frac{n}{n-1}\frac{A_n}{\sigma_n^2}\right\}^{1/2} = o_P(1) ,\end{aligned}$$

where the last two equalities follow from the definitions of A_n and B_n, and the results of (I) and (II). □

1.3.2 Consistency of the Jackknife bias estimator

We start by defining the asymptotic bias of T_n. Consider the expansion in (1.5). Since $\xi_n \in [\overline{X}_n, \mu]$, by the LLN $\xi_n \xrightarrow{P} \mu$. So, under Assumption B we can apply the continuous mapping theorem and obtain that $\nabla^2 g(\xi_n) = \nabla^2 g(\mu) + o_P(1)$. Recall that by the CLT

$\overline{X}_n - \mu = O_P(n^{-1/2})$. Hence, the third term on the RHS of (1.5) is $o_P(n^{-1})$ and we have

$$\begin{aligned} T_n - \theta =& \nabla g(\mu)^T(\overline{X}_n - \mu) \\ &+ \frac{1}{2}(\overline{X}_n - \mu)^T \nabla^2 g(\mu)(\overline{X}_n - \mu) \\ &+ o_P\left(\frac{1}{n}\right) . \end{aligned} \tag{1.13}$$

Now, the first term on the RHS is $O_P(n^{-1/2})$ and has null expectation. The second term is of order $O_P(n^{-1})$ and its expectation is (see the comments below)

$$\begin{aligned} \mathbb{E}\left\{\frac{1}{2}(\overline{X}_n - \mu)^T \nabla^2 g(\mu)(\overline{X}_n - \mu)\right\} =& \frac{1}{2}\mathbb{E}\ \text{Trace}\left\{(\overline{X}_n - \mu)^T \nabla^2 g(\mu)(\overline{X}_n - \mu)\right\} \\ =& \frac{1}{2}\mathbb{E}\ \text{Trace}\left\{\nabla^2 g(\mu)(\overline{X}_n - \mu)(\overline{X}_n - \mu)^T\right\} \\ =& \frac{1}{2}\ \text{Trace}\left(\nabla^2 g(\mu)\ \mathbb{E}\left\{(\overline{X}_n - \mu)(\overline{X}_n - \mu)^T\right\}\right) \\ =& \frac{1}{2n}\text{Trace}\left(\nabla^2 g(\mu)\ \text{Var}(X)\right) . \end{aligned}$$

Indeed, for any conformable matrix B and vector c we have $c^T Bc = Trace(c^T Bc) = Trace(Bcc^T)$, which gives the first and second equality. The third equality follows by commuting expectation and $Trace$. Finally, the last equality is ensures by the iid assumption.

Now, the last term on the RHS of (1.13) can be ignored as it is $o_P(n^{-1})$ and hence of smaller order than the other two terms on the RHS. So, *heuristically*

$$\mathbb{E}\{T_n\} - \theta \approx \frac{a}{n} \text{ with } a := \frac{1}{2}\ \text{Trace}\left(\nabla^2 g(\mu)\ \text{Var}(X)\right) .$$

By the above display, we can define a/n as the *asymptotic bias* of T_n. At the end of this section we will show that $b_{jack} = a/n + o_P(n^{-1})$, so that the Jackknife bias estimator is consistent for the asymptotic bias of T_n. This rate implies that the bias-reduced Jackknife estimator T_{jack} has the same asymptotic distribution as T_n, i.e.

$$\begin{aligned} \sqrt{n}(T_{jack} - \theta) =& \sqrt{n}(T_n - \theta) + \sqrt{n} b_{jack} \\ =& \sqrt{n}(T_n - \theta) + \sqrt{n}\left(\frac{a}{n} + o_P(n^{-1})\right) \\ =& \sqrt{n}(T_n - \theta) + o_P(1) . \end{aligned}$$

In other words, the bias correction b_{jack} does not have an impact on the asymptotic behavior of T_n. Its advantage stands in the small-sample behavior of T_{jack}, in the sense that in finite sample T_{jack} will generally display a lower bias than T_n.[2]

[2]The *heuristic* argument for this is contained in Section 1.1.

Proposition 1.3.2. *Under Assumptions A and B*

$$n\ b_{jack} = a + o_P(1)$$

If moreover $\nabla^3 g$ is continuous in a neighborhood of μ and $\mathbb{E}||X||^4 < \infty$ then

$$n\ b_{jack} = a + O_P\left(n^{-1/2}\right)\ .$$

Proof. From the definition of b_{jack} in (1.1)

$$\begin{aligned} b_{jack} =& (n-1)\left(\frac{\sum_i T_{n-1,i}}{n} - T_n\right) \\ =& \frac{n-1}{n}\sum_i (T_{n-1,i} - T_n)\ . \end{aligned}$$

As obtained in the proof of Proposition 1.3.1 $\max_{i=1,..,n} ||\overline{X}_{n-1,i} - \mu|| \leq \max_{i=1,..,n} ||\overline{X}_{n-1,i} - \overline{X}_n|| + \|\overline{X}_n - \mu\| = o_P(1)$. So, by Assumption B and a second order Taylor expansion (see Proposition A.0.11), wpa1

$$\begin{aligned} T_{n-1,i} - T_n = g(\overline{X}_{n-1,i}) - g(\overline{X}_n) =& \nabla g(\overline{X}_n)^T(\overline{X}_{n-1,i} - \overline{X}_n) \\ &+ \frac{1}{2}(\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla^2 g(\xi_{n,i})(\overline{X}_{n-1,i} - \overline{X}_n)\ , \end{aligned}$$

where $\xi_{n,i} \in [\overline{X}_{n-1,i}, \overline{X}_n]$. From (1.7) $\sum_i(\overline{X}_{n-1,i} - \overline{X}_n) = 0$. So, from the previous two displays

$$\begin{aligned} n\ b_{jack} =& (n-1)\sum_i (T_{n-1,i} - T_n) \\ =& \frac{n-1}{2}\sum_i (\overline{X}_{n-1,i} - \overline{X}_n)^T \nabla^2 g(\xi_{n,i})(\overline{X}_{n-1,i} - \overline{X}_n) \\ =& \frac{1}{2(n-1)}\sum_i (X_i - \overline{X}_n)^T \nabla^2 g(\xi_{n,i})(X_i - \overline{X}_n) \text{ (from (1.7))} \\ =& \frac{1}{2(n-1)}\sum_i \text{Trace}\left\{\nabla^2 g(\xi_{n,i})(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right\} \\ =& \frac{1}{2(n-1)}\sum_i \text{Trace}\left\{\left[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)\right](X_i - \overline{X}_n)((X_i - \overline{X}_n)^T\right\} \\ &+ \frac{1}{2(n-1)}\sum_i \text{Trace}\left\{\nabla^2 g(\mu)(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T\right\} \\ =:& A_n + B_n \end{aligned} \tag{1.14}$$

where for the fourth equality we have used the property that $\text{Trace}(c^T Bc) = \text{Trace}(Bcc^T)$

for any conformable matrix B and vector c. Now,

$$\begin{aligned}
B_n =& \frac{1}{2} \text{ Trace} \left\{ \nabla^2 g(\mu) \frac{1}{n-1} \sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T \right\} \\
=& \frac{1}{2} \text{ Trace} \left\{ \nabla^2 g(\mu) \text{Var}(X) \right\} + o_P(1) \\
=& a + o_P(1) \ ,
\end{aligned}$$

where we have used $(n-1)^{-1} \sum_i (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T = \text{Var}(X) + o_P(1)$ and the definition of a. So, from the previous two displays to show the first part of the proposition it is sufficient to prove that $A_n = o_P(1)$. To this end, denote with $X_i^{(s)}$ and $\overline{X}_n^{(s)}$ the sth component of X_i and $\overline{X}_n$. Also, denote with $[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)]_{s,m}$ the element on the sth row and mth column of $[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)]$. Then, from the definition of A_n

$$\begin{aligned}
|A_n| \leq& \frac{2}{n-1} \sum_{i=1}^n \left| \text{Trace} \left\{ \left[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu) \right] (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T \right\} \right| \\
=& \frac{2}{n-1} \sum_{i=1}^n \left| \sum_{s,m} \left[\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu) \right]_{s,m} (X_i^{(m)} - \overline{X}_n^{(m)})(X_i^{(s)} - \overline{X}_n^{(s)}) \right| \\
\leq& \frac{1}{2(n-1)} \sum_{i=1}^n ||\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)|| \sum_{s,m} \left| (X_i^{(s)} - \overline{X}_n^{(s)})(X_i^{(m)} - \overline{X}_n^{(m)}) \right| \\
\leq& \max_{i=1,..,n} ||\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)|| \sum_{s,m} \frac{1}{2(n-1)} \sum_{i=1}^n \left| (X_i^{(s)} - \overline{X}_n^{(s)})(X_i^{(m)} - \overline{X}_n^{(m)}) \right|
\end{aligned} \tag{1.15}$$

where $||\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)||$ denotes the Forbenious norm of the matrix $\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)$. By proceeding similarly to Equation (1.11) and (1.12), we get that $\max_{i=1,..,n} ||\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)|| = o_P(1)$. Also, by Assumption A for each s, m the average on the RHS of the above display is $O_P(1)$. Thus, $|A_n| = o_P(1)$ and we have proved the first part of the proposition.

To prove the second part, notice that when $\mathbb{E}||X||^4 < \infty$ we can apply the CLT and have $(n-1)^{-1} \sum_i X_i X_i^T = \mathbb{E} X X^T + O_P(n^{-1/2})$, so that

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X}_n)(X_i-\overline{X}_n)^T =&\frac{1}{n-1}\sum_{i=1}^{n}X_iX_i^T-\frac{1}{n-1}\sum_{i=1}^{n}X_i\overline{X}_n^T\\ &-\overline{X}_n\frac{1}{n-1}\sum_{i=1}^{n}X_i^T+\frac{n}{n-1}\overline{X}_n\overline{X}_n^T\\ =&\mathbb{E}XX^T-\mu\mu^T+O_P\left(\frac{1}{\sqrt{n}}\right)\\ =&\mathrm{Var}(X)+O_P\left(\frac{1}{\sqrt{n}}\right)\ .\end{aligned}$$

Using the above display and the definition of B_n in (1.14) we find that

$$B_n=\frac{1}{2}\mathrm{Trace}\left\{\nabla^2 g(\mu)\mathrm{Var}(X)\right\}+O_P(n^{-1/2})\ .$$

Hence, from (1.14) it suffices to show that $|A_n|=O_P(n^{-1/2})$. To this end, from the Equation (1.15) and the arguments below it, we only have to prove that $\max_{i=1,..,n}||\nabla^2 g(\xi_{n,i})-\nabla^2 g(\mu)||=O_P(n^{-1/2})$. Consider a single component of $\nabla^2 g(\xi_{n,i})-\nabla^2 g(\mu)$ and drop its matrix index for notational simplicity. Since by assumption $\nabla^3 g$ is continuous in a neighborhood of μ, by a first order Taylor expansion there exists a C and a $\delta>0$ such that

$$\left|\nabla^2 g(x)-\nabla^2 g(\mu)\right|\leq C||x-\mu|| \text{ for all } x\in\mathrm{Ball}(\mu,\delta)\ .$$

Recall that from (1.12) $\max_{i=1,..,n}||\xi_{n,i}-\mu||=O_P(n^{-1/2})$, so the above display implies that wpa1

$$\max_{i=1,..,n}||\nabla^2 g(\xi_{n,i})-\nabla^2 g(\mu)||\leq C\max_{i=1,..,n}||\xi_{n,i}-\mu||=O_P(n^{-1/2})$$

which delivers the desired result.

□

1.4 Extensions

In some cases the jackknife does not provide a consistent estimator of the variance, i.e. $\upsilon_{jack}/\sigma_n^2$ does not converge in probability to 1. One of such cases is the estimation of the variances of the sample quantiles, e.g. the variance of the sample median. This lack of consistency is due to the lack of smoothness of the function g, see Shao and Tu (1995). This should not be surprising: the smoothness of g has been heavily used in the proof of Proposition 1.3.1. A valid alternative to the Jackknife in these situations is the *deleted d jackknife*. This can be seen as a generalization of the Jackknife studied in the previous pages, where instead of removing one observation we remove d observations and recompute

the statistic on the remaining sample. In particular, let $\mathbf{s}$ be a subset of size $n-d$ of the indices $\{1,\ldots,n\}$ and let $T_{n-d,\mathbf{s}}$ be the statistic computed with such observations. The *deleted-d jackknife variance estimator* is given by

$$v_{jack-d} := \frac{n-d}{d}\frac{1}{N}\sum_{\mathbf{s}\in\mathcal{S}}\left[T_{n-d,\mathbf{s}} - \frac{1}{N}\sum_{\mathbf{s}\in\mathcal{S}} T_{n-d,\mathbf{s}}\right]^2,$$

where $\mathcal{S}$ represents the collection of all subgroup of $\{1,\ldots,n\}$ of size $n-d$, where each group differs from the others by at least one element, and $N = \binom{n}{n-d}$ is the number of elements in $\mathcal{S}$. In combinatorial calculus $\mathcal{S}$ is also called collection of combinations of $\{1,\ldots,n\}$ of class $n-d$. Notice that when $d=1$, v_{jack-d} corresponds to the jackknife studied in the previous pages. The asymptotic properties of v_{jack-d} are obtained in Shao and Tu (1995) to whom we refer the reader. We will see in the following chapters that such a principle can be used also to estimate the distribution of a statistic, and it has a close connection with the *subsampling method*.

The jackknife principle can also be used to estimate the variance of a functional, where $T_n = g(F_n)$ and F_n is the empirical cdf. Notice that the empirical cdf is a function $x \mapsto (1/n)\sum_i 1(X_i \le x)$, so g maps such functions into $\mathbb{R}$. More details can be found in Shao and Tu (1995).

Notes

The first author to propose the jackknife method for bias estimation has been Quenouille (1949). Some time later, Tukey (1958) realized that the same principle could be used to estimate the variance of a statistic and popularized the Jackknife.

The presentation given in this chapter follows closely Chapter 2 of Shao and Tu (1995) who provide a very comprehensive and detailed theory of the Jackknife.

MacKinnon and White (1985) have been the first to propose the Jackknife for estimating the variance of the OLS estimator in the presence of heteroskedasticity. They provide a large simulation study showing that it outperforms the popular White-Huber estimator.

An idea similar to the Jackknife has been proposed by Angrist, Imbens, and Krueger (1999) for bias correcting the 2SLS estimator. They label their method as JIVE (Jackknife Instrumental Variable Estimator), although it is quite different from a Jackknife bias corrected estimator. Later, Hahn, Hausman, and Kuersteiner (2004) proposed a bias corrected jackknife estimator for the 2SLS, as that seen in Section 1.2. They provide evidence about the finite sample performances of such an estimator and show that in practice it outperforms several estimators for the linear model with endogenous regressors. Davidson and MacKinnon (2006) also show empirical evidence about several estimators for the IV model.

$\hat{t} = \sqrt{n}(\bar{X} - \mathbb{E}(X))$ By CLT, $\hat{t} \xrightarrow{P} N(0, Var(X))$

$\hat{t}^{PIV} = \dfrac{\sqrt{n}(\bar{X} - \mathbb{E}(X))}{\sqrt{\widehat{Var}(X)}}$ $\hat{t}^{PIV} \xrightarrow{P} N(0, 1)$

Idea behind Bootstrap:

Replace P_X with $\hat{P}_X$ and then

use $\hat{P}_X$ and random samples from $\hat{P}_X$ to compute estimations of statistic

In statistics:

$$\begin{tikzcd} P_X \arrow[r] \arrow[dr] & (X_1, \cdots, X_n) \arrow[d] \\ & T_n(X_1, \cdots, X_n, P_X) \end{tikzcd}$$

①

In Bootstrap:

$$\begin{tikzcd} \hat{P}_X \arrow[r] \arrow[dr] & (X_1^*, \cdots, X_n^*) \arrow[d] \\ & T_n(X_1^*, \cdots, X_n^*, \hat{P}_X) \end{tikzcd}$$

②

$X_1^*, \cdots, X_n^*$ is random;

$X_1, \cdots, X_n$ is fixed.

Principles of substitution: replace $\hat{P}_X$ by P_X

Principles of similarity: Build a T_n^* by mimicking the relationship ①

Resample with replacement from our sample $\{x_1, \cdots, x_n\}$.

$$X^* \sim \hat{P}_X$$

$$\hat{P}_X(X^* = X_j) = \frac{1}{n} \qquad \hat{P}_X(X^* \neq X_j, \forall j = 1, \cdots, n) = 0.$$

$$\hat{t} = \sqrt{n}(\bar{X} - \mathbb{E}(X)) \qquad \hat{t}^* = \sqrt{n}(\bar{X}^* - \hat{\mathbb{E}}(X^*))$$

In fact, $\hat{\mathbb{E}}(X^*) = \bar{X}$

$$\hat{F}_{\hat{t}^*}(z) = \hat{\mathbb{E}}\, \mathbb{1}\{\hat{t}^* \le z\}$$

$$= \hat{\mathbb{E}}\, \mathbb{1}\{\sqrt{n}(\bar{X}^* - \bar{X}) \le z\}$$

Monte-Carlo to compute $\hat{F}_{t^*}$:

1) Draw with replacement n observation from $\{X_1, \cdots X_n\}$ to get $\{X_i^*\}_{i=1}^n$

2) $\hat{t}_b^* = \sqrt{n}(\bar{X}^* - \bar{X})$

3) Repeat 1) & 2) B times $(B \to +\infty)$

$\{\hat{t}_b^*,\ b = 1, \cdots, B\}$ to approximate: Bootstrap distribution → compare it with asymptotique distribution of statistics. Usually be $\mathcal{N}(0, Var(X))$

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\hat{t}_b^* \le z\} \underset{LLN}{\simeq} \hat{\mathbb{E}}\, \mathbb{1}\{t^* \le z\} = \hat{F}_{t^*}(z)$$

$\hat{F}_{\hat{t}^*} \longrightarrow F_{\hat{t}}$ needs to prove.

$$P\Big(q_{\frac{\alpha}{2}} \le \overbrace{\sqrt{n}(\bar{X} - \mathbb{E}(X))}^{\hat{t}} \le q_{1-\frac{\alpha}{2}}\Big) = 1 - \alpha$$

↳ $\frac{\alpha}{2}$ quantile of $F_{\hat{t}}$

remplace the quantile in the CI.

$$\simeq P\Big(\hat{q}_{\frac{\alpha}{2}} \le \sqrt{n}(\bar{X} - \mathbb{E}(X)) \le \hat{q}_{1-\frac{\alpha}{2}}\Big)$$

BCI: $\left[\bar{X} - \frac{\hat{q}_{1-\frac{\alpha}{2}}}{\sqrt{n}},\ \bar{X} - \frac{\hat{q}_{\frac{\alpha}{2}}}{\sqrt{n}}\right]$

for $\mathbb{E}(X)$

Example: OLS model.

$$y_i = z_i^T \beta_0 + \varepsilon_i \qquad \mathbb{E}(\varepsilon_i x_i) = 0.$$

$$\beta_0 = (\mathbb{E}(zz^T))^{-1} \mathbb{E}(z \cdot y)$$

$$\hat{\beta} = (\overline{zz^T})^{-1} \overline{zy} \qquad \hat{t} = \sqrt{n}(\hat{\beta} - \beta_0)$$

Define $X_i = (y_i, z_i)$ $\qquad X_i^* = (y_i^*, z_i^*)$

$$\hat{\beta}^* = (\overline{z^* z^{*T}})^{-1} \overline{z^* y^*} \qquad \hat{t}^* = \sqrt{n}(\hat{\beta}^* - \hat{\beta})$$

Tricks:

Look at the random part in statistic ($\hat{\beta}$ here)

replace samples by its resampled observations.

In the expectation part in statistic (β_0 here),

replace P_x by $\hat{P}_x$

Then use $\hat{\mathbb{E}}(X^*) = \bar{X}$

Chapter 2

The General Theory of the Bootstrap

Goal: Test hypothesis or construct CIs about $\mathbb{E}(X)$ given $\{X_i\}_{i=1}^n$ $X_i \overset{iid}{\sim} P_X$

The bootstrap is a tool used to estimate the distribution of a statistic. In this chapter, we introduce the general theory of the bootstrap. First, we illustrate why and when the bootstrap yields refinements over the asymptotic approximations. Second, we prove the consistency of the bootstrap for constructing confidence intervals and tests of hypothesis. The results developed here will be used in the following chapters to show the validity of specific bootstrap schemes. The organization is as follows. Section 2.1 sets up the framework and illustrates the advantages of the bootstrap by mean of heuristic arguments. Section 2.2 provides sufficient conditions under which the bootstrap estimates consistently the distribution of a statistic. Section 2.3 shows how to build confidence intervals using the bootstrap method and proves the consistency of such confidence intervals under general sufficient conditions. Section 2.4 shows how the bootstrap can be used to construct tests of hypothesis and proves the consistency of such bootstrap tests. Finally, Section 2.5 provides the bootstrap delta method.

2.1 The framework and some intuition for the Bootstrap

The bootstrap is a method used to estimate the distribution of a statistic. It does so by generating artificial data from the sample data. The main employment of the bootstrap is in constructing confidence intervals and tests of hypothesis for an unknown parameter of interest. Its main advantage stands in providing refinements over the asymptotic approximation. The bootstrap can also be used to approximate the distribution of statistic when its asymptotic distribution depends on unknown parameters that cannot be accurately estimated.

In this section we first set up the framework, introduce some basic definitions, and define more precisely the bootstrap principle. Then, we show under which conditions the bootstrap provides good approximations of the distribution of a statistic.

The framework. We observe an iid sample $\{X_i\}_{i=1}^n$, with $X_i \sim P_X$ for $i = 1, .., n$. The observed sample is used to construct a statistic $T_n := T_n(X_1, .., X_n)$ that can be employed to obtain a confidence interval or a hypothesis test for an unknown population parameter. To this end, we are interested in the actual distribution of T_n

$$G_n(\tau, P_X) := \int 1\{T(x_1, .., x_n) \leq \tau\} \otimes_{i=1}^n P_X(dx_i) \ ,$$

where $1\{\cdot\}$ denotes the indicator function and $\otimes_{i=1}^n P_X$ is the joint probability of $\{X_i\}_{i=1}^n$ under the iid assumption. If the above cdf was known we could directly use it to construct confidence intervals or hypothesis tests about the parameter of interest. However, it is usually the case that such a distribution is unknown and must be estimated. One way to estimate it is to use the asymptotic approximation. It often holds that, as the sample size n grows, the statistic T_n converges in distribution to a specific random variable, so $T_n \rightsquigarrow T$ for some random variable T and

$$G_n(\tau, P_X) \to G_\infty(\tau, P_X) \text{ for each } \tau \in \mathbb{R} \ ,$$

where $G_\infty(\cdot, P_X)$ denotes the cdf of T. Notice that such cdf will in general depend on the probability P_X (below we provide a basic example). By the above display, $G_\infty(\cdot, P_X)$ can be used to approximate the *true finite sample distribution* of T_n. We refer to this approach as the *asymptotic approximation.*

Clearly, when $G_\infty(\cdot, P_X)$ depends on P_X the asymptotic approximation cannot be directly implemented, as it depends on unknown parameters defined at the population level. In such cases, we replace such unknown parameters by consistent estimates. This step is not needed when T_n is *asymptotically pivotal.* We say that T_n is *asymptotically pivotal* when $G_\infty(\cdot, P_X)$ does not depend on P_X. In this case, the asymptotic approximation can be directly implemented without estimating any additional parameter.

In order to estimate $G_n(\cdot, P_X)$, the bootstrap takes a different approach with respect to the asymptotic approximation. When an estimator $\widehat{P}_X$ of P_X is available, the bootstrap approximates $G_n(\cdot, P_X)$ by $G_n(\cdot, \widehat{P}_X)$, where

$$G_n(\tau, \widehat{P}_X) = \int 1\{T_n(x_1, .., x_n) \leq \tau\} \otimes_{i=1}^n \widehat{P}_X(x_i) \ .$$

So $G_n(\tau, \widehat{P}_X) = \widehat{\mathbb{E}}\{1\{T_n(X_1^*, .., X_n^*) \leq \tau\}\}$ with $\widehat{\mathbb{E}}$ representing the expectation where $X_1^*, .., X_n^*$ are independent and identically distributed according to the probability $\widehat{P}_X$. Such expectation considers only $\{X_i^*\}_{i=1}^n$ as random and as given the sample data. Notice that since $\widehat{P}_X$ depends on the data $\{X_i\}_{i=1}^n$, $G_n(\cdot, \widehat{P}_X)$ is a "random" cdf.

Example: Sample Mean. As an example, assume we are interested in testing a hypothesis about the expectation of X, i.e. $\mathbb{E}\{X\}$. Then, we can set $T_n := \sqrt{n}(\overline{X}_n - \mathbb{E}\{X\})$,

where $\overline{X}_n$ denotes the sample average of $\{X_i\}_{i=1}^n$. If X has finite second moments, the CLT ensures that $T_n \rightsquigarrow \mathcal{N}(0, Var(X))$. In this case, $G_\infty(\cdot, P_X)$ will be the cdf of $\mathcal{N}(0, \mathrm{Var}(X))$ that can be used to approximate the true cdf of T_n and hence to construct confidence intervals and hypothesis tests about $\mathbb{E}\{X\}$. Notice that the asymptotic distribution of T_n depends on P_X through $\mathrm{Var}(X) = \int (x - \mathbb{E}X)^2 \, P_X(dx)$, so such a statistic is not asymptotically pivotal. In practice, we first replace $\mathrm{Var}(X)$ by a consistent estimate and then use the cdf of $\mathcal{N}(0, \widehat{\mathrm{Var}}(X))$ to approximate $G_n(\cdot, P_X)$. Under a suitable normalization, T_n will become asymptotically pivotal. In particular, by redefining $T_n := \sqrt{n}(\overline{X}_n - \mathbb{E}\{X\})/\sqrt{\widehat{\mathrm{Var}}(X)}$, we will have that $T_n \rightsquigarrow \mathcal{N}(0,1)$ and $G_\infty(\cdot, P_X)$ will be the cdf of the standard normal and will not depend on P_X.
To construct the bootstrap approximation in this example, we can define $\widehat{P}_X$ to be the probability generated when we draw with replacement from the sample $\{X_i\}_{i=1}^n$. Then, in the non-asympotically pivotal case we have $T_n(X_1^*, .., X_n^*) = \sqrt{n}(\overline{X^*} - \widehat{\mathbb{E}}\{X^*\})$, where $\overline{X^*}$ represents the sample average of $\{X_i^*\}_{i=1}^n$, X_i^*'s are iid according to $\widehat{P}_X$, and $\widehat{\mathbb{E}}\{X^*\}$ is the expectation of X^* when X^* is distributed according to $\widehat{P}_X$. We will see in the next chapter that $\widehat{\mathbb{E}}\{X^*\} = \overline{X}$, so that the expectation in "the bootstrap world" equals the sample mean. For the asymptotically pivotal version of T_n we have $T_n(X_1^*, .., X_n^*) = \sqrt{n}(\overline{X^*} - \widehat{\mathbb{E}}\{X^*\})/\sqrt{\widehat{\mathrm{Var}^*}(X^*)}$, where $\widehat{\mathrm{Var}^*}(X^*)$ represents the sample variance of X^* in the "bootstrap world", i.e. $\widehat{\mathrm{Var}^*}(X^*) = \overline{(X^*)^2} - \overline{X^*}^2$.
In essence, within this example, the bootstrap replaces the true population with the sample data. If the sample is representative of the population (which happens when n is large), then we can expect $\widehat{P}_X$ to be close to P_X, so that $G_n(\cdot, \widehat{P}_X)$ will be close to $G_n(\cdot, P_X)$ and the bootstrap will yield a good approximation to the true distribution of the statistic T_n.

Bootstrap refinements. The utility of the bootstrap basically stands in two aspects. First, there are situations where the asymptotic distribution $G_\infty(\cdot, P_X)$ has an intricate expression or depends in a complicated way on unknown population parameters. Hence, it is not convenient to use the asymptotic approximation in practice, and the bootstrap provides an easy way to approximate $G_n(\cdot, P_X)$. Second, in some cases the bootstrap can yield refinements over the asymptotic approximation, in the sense that the approximation of $G_n(\cdot, P_X)$ provided by $G_n(\cdot, \widehat{P}_X)$ is more "precise" than the one provided by $G_\infty(\cdot, P_X)$. We will dig more into the latter aspect by using the *Edgeworth expansion* of $G_n(\cdot, P_X)$. The technicalities required to obtain such Edgeworth expansions are beyond the scope of these lecture notes, so we will only provide heuristic arguments that however will help our intuition.
Under some regularity conditions,

$$G_n(\tau, P_X) = G_\infty(\tau, P_X) + \frac{1}{n^{1/2}} g_1(\tau, P_X) + \frac{1}{n} g_2(\tau, P_X) + \frac{1}{n^{3/2}} g_3(\tau, P_X) + O\left(\frac{1}{n^2}\right) , \tag{2.1}$$

uniformly over τ, where $g_1(\cdot, P_X)$ and $g_3(\cdot, P_X)$ are even functions of their first argument (i.e. $g_1(\tau, P_X) = g_1(-\tau, P_X)$ and the same holds for $g_3(\cdot, P_X)$), while $g_2(\cdot, P_X)$ is an odd function of its first argument(i.e. $g_2(\tau, P_X) = -g_2(-\tau, P_X)$). Also, G_∞, g_1, g_2, g_3 are "continuous" functions of their second argument. Equation (2.1) is called *Edgeworth expansion* of $G_n(\cdot, P_X)$.

When $\widehat{P}_X$ is a "consistent" estimator of P_X, the previous expansion will also hold by replacing P_X with $\widehat{P}_X$, i.e.

$$G_n(\tau, \widehat{P}_X) = G_\infty(\tau, \widehat{P}_X) + \frac{1}{n^{1/2}} g_1(\tau, \widehat{P}_X) + \frac{1}{n} g_2(\tau, \widehat{P}_X) + \frac{1}{n^{3/2}} g_3(\tau, \widehat{P}_X) + O\left(\frac{1}{n^2}\right) , \tag{2.2}$$

almost surely and uniformly over τ.

To show the quality of the asymptotic approximation, from Equation (2.1) we obtain that

$$\sup_\tau |G_n(\tau, P_X) - G_\infty(\tau, P_X)| = O(n^{-1/2}) . \tag{2.3}$$

Thus, the error from the asymptotic approximation is of order $n^{-1/2}$.

To obtain the error from the bootstrap approximation, assume that the estimator $\widehat{P}_X$ is $n^{-1/2}$ consistent for P_X, i.e. $d(\widehat{P}_X, P_X) = O_P(n^{-1/2})$, where d represents a metric (i.e. a distance) over the space of probability measures. Such a rate will often hold for popular estimators $\widehat{P}_X$. Then, subtracting (2.2) from (2.1) gives

$$\begin{aligned}\sup_\tau \left|G_n(\tau, P_X) - G_n(\tau, \widehat{P}_X)\right| \leq & \sup_\tau \left|G_\infty(\tau, P_X) - G_\infty(\tau, \widehat{P}_X)\right| \\ & + \frac{1}{n^{1/2}} \sup_\tau \left|g_1(\tau, P_X) - g_1(\tau, \widehat{P}_X)\right| \\ & + O(n^{-1}) .\end{aligned}$$

The second term on the RHS is of order $O(n^{-1/2})$. Since $d(\widehat{P}_X, P_X) = O(n^{-1/2})$ and G_∞ is continuous in its second argument, we (heuristically) have that $\sup_\tau |G_\infty(\tau, \widehat{P}_X) - G_\infty(\tau, P_X)| = O(n^{-1/2})$. Gathering together these results leads to $\sup_\tau \left|G_n(\tau, P_X) - G_n(\tau, \widehat{P}_X)\right| = O(n^{-1/2})$. This rate implies that the approximation error of the bootstrap is $O(n^{-1/2})$ and hence of the same order as the asymptotic approximation. Accordingly, *nothing is lost by bootstrapping*, but so far nothing is gained neither.

The advantage of the bootstrap arises when the statistic T_n is asymptotically pivotal. In such a case, G_∞ does not depend on its second argument and $G_\infty(\tau, P_X) = G_\infty(\tau, \widehat{P}_X)$ for

Gain:
① when asymptotic distribution is complicated to compute.
② when T_n asymptotically pivotal, Bootstrap gains in terms of c.v.g rate.

The asymptotic distribution does not depend on distribution of data.

all $\tau \in \mathbb{R}$. Also, since g_1 is continuous in its second argument and $d(\widehat{P}_X, P_X) = O(n^{-1/2})$ we (heuristically) obtain that $\sup_\tau |g_1(\tau, \widehat{P}_X) - g_1(\tau, P_X)| = O(n^{-1/2})$. These rates and the previous display give

$$\sup_\tau \left| G_n(\tau, P_X) - G_n(\tau, \widehat{P}_X) \right| = O(n^{-1}) \ .$$

So, the approximation error from the bootstrap is of order n^{-1}. By comparing this rate with the error from the asymptotic approximation in (2.3), we find that the error from the bootstrap approximation is of smaller order than the error from the asymptotic approximation, as n increases. In other words, the *bootstrap provides refinements over the asymptotic approximation when the statistic is asymptotically pivotal.*

The bootstrap refinements are even more pronounced when T_n can take positive and negative values and we want to estimate the distribution of $|T_n|$ (i.e. we build confidence intervals and hypothesis tests based on $|T_n|$). To see this, assume again that T_n is asymptotically pivotal. From (2.1) we get

$$\begin{aligned} P(|T_n| \leq \tau) = G_n(\tau, P_X) - G_n(-\tau, P_X) =& G_\infty(\tau, P_X) - G_\infty(-\tau, P_X) \\ &+ \frac{1}{n^{1/2}} \left[g_1(\tau, P_X) - g_1(-\tau, P_X) \right] \\ &+ \frac{1}{n} \left[g_2(\tau, P_X) - g_2(-\tau, P_X) \right] \\ &+ O(n^{-3/2}) \ . \end{aligned}$$

The bootstrap estimate of $P(|T_n| \leq \tau)$ is $G_n(\tau, \widehat{P}_X) - G_n(-\tau, \widehat{P}_X)$. So, (2.2) leads to

$$\begin{aligned} G_n(\tau, \widehat{P}_X) - G_n(-\tau, \widehat{P}_X) =& G_\infty(\tau, \widehat{P}_X) - G_\infty(-\tau, \widehat{P}_X) \\ &+ \frac{1}{n^{1/2}} \left[g_1(\tau, \widehat{P}_X) - g_1(-\tau, \widehat{P}_X) \right] \\ &+ \frac{1}{n} \left[g_2(\tau, \widehat{P}_X) - g_2(-\tau, \widehat{P}_X) \right] \\ &+ O(n^{-3/2}) \ . \end{aligned}$$

Since g_1 is an even function of its first argument $g_1(\tau, P_X) - g_1(-\tau, P_X) = 0$ and $g_1(\tau, \widehat{P}_X) - g_1(-\tau, \widehat{P}_X) = 0$. Also, from $d(\widehat{P}_X, P_X) = O(n^{-1/2})$ and the continuity of g_2 in its second argument we (heuristically) obtain that $\sup_\tau |g_2(\tau, \widehat{P}_X) - g_2(\tau, P_X)| = O(n^{-1/2})$. Finally, notice that since T_n is asymtptotically pivotal $G_\infty(\cdot, P_X) = G_\infty(\cdot, \widehat{P}_X)$. By these features and the previous two displays we find that

$$P(|T_n| \leq \tau) - \left[G_n(\tau, \widehat{P}_X) - G_n(-\tau, \widehat{P}_X) \right] = O(n^{-3/2}) \ .$$

Hence, when we use the bootstrap to estimate the distribution of $|T_n|$ the approximation error is of order $n^{-3/2}$, and we obtain better refinements with respect to when we estimate

the distribution of T_n. This gain also holds for the asymptotic approximation: by the arguments above, the asymptotic approximation makes an error of order n^{-1} instead of $n^{-1/2}$. However, the quality of the refinements remains in favor of the bootstrap.

2.2 Consistency of the Bootstrap

In this section we show that under some sufficient conditions the bootstrap "consistently" estimates the distribution of a statistic. Denote with $\mathcal{P}$ a space of probability measures and with d a metric on it.

In partice, we do not use it.

Proposition 2.2.1. *Assume that*

$$G_n(\tau, P_X) = F_{T_n}(\tau) = P(T_n(X_1, \cdots, X_n, P_X) \leq \tau)$$

(i) $T_n \rightsquigarrow T$, *with* T *having cdf* $G_\infty(\cdot, P_X)$

(ii) $G_\infty(\cdot, P)$ *is continuous for all* $P \in \mathcal{P}$

(iii) for all $\tau \in \mathbb{R}$ *and any sequence* $(P_n)_n$ *in* $\mathcal{P}$ *such that* $d(P_n, P_X) \to 0$ *it holds that* $G_n(\tau, P_n) \to G_\infty(\tau, P_X)$

(iv) $d(\widehat{P}_X, P_X) = o_P(1)$.

Then,

$$\sup_\tau \left| G_n(\tau, \widehat{P}_X) - G_n(\tau, P_X) \right| = o_P(1) \ .$$

Before proceeding to the proof, let us comment on the conditions. The first two conditions require that the statistic must converge to a distribution that is well behaved in terms of continuity, so that $G_n(\tau, P_X) \to G_\infty(\tau, P_X)$ for each $\tau \in \mathbb{R}$. The third condition requires that also the finite sample distribution of T_n (i.e. $G_n(\cdot, P_X)$) must be "regular" with respect to the probability generating the data. In particular, condition (iii) requires that replacing P_X with a probability P_n that is close to it must not change the limit behavior of $G_n(\cdot, P_X)$ as n grows. This allows us to replace the true distribution of the data P_X with a consistent estimate $\widehat{P}_X$ so as to obtain an approximation of the true distribution of the statistic.

Proof. Recall that $G_n(\cdot, P_X)$ is the cdf of T_n. Combining (i), (ii), and Polya's Theorem (see Proposition A.0.12) ensures that $\sup_\tau |G_n(\tau, P_X) - G_\infty(\tau, P_X)| = o(1)$. Thus, it is sufficient to show that $\sup_\tau |G_n(\tau, \widehat{P}_X) - G_\infty(\tau, P_X)| = o_P(1)$. By contradiction assume that this is not the case. Then, there exists a $\delta > 0$ and an $\epsilon > 0$ such that

$$P\left(\sup_\tau \left| G_{n_1}(\tau, \widehat{P}_X^{(n_1)}) - G_\infty(\tau, P_X) \right| > \delta \right) \geq \epsilon$$

along a subsequence (n_1) of (n), where $\widehat{P}_X^{(n_1)}$ is an estimator based on the observations $\{X_i\}_{i=1}^{n_1}$. From condition (iv) and Proposition A.0.4, we can extract a subsequence (n_2)

of (n_1) such that $d(\widehat{P}_X^{(n_2)}, P_X) = o(1)$ almost surely. Hence, from (iii) it must hold that almost surely $|G_{n_2}(\tau, \widehat{P}_X^{(n_2)}) - G_\infty(\tau, P_X)| = o(1)$ for all $\tau \in \mathbb{R}$. Since by (ii) $G_\infty(\cdot, P_X)$ is continuous, using Proposition A.0.5 and Polya's theorem such pointwise convergence can be turned into a uniform convergence, so that $\sup_\tau |G_{n_2}(\tau, \widehat{P}_X^{(n_2)}) - G_\infty(\tau, P_X)| = o(1)$ almost surely. This implies that

$$P\left(\sup_\tau \left|G_{n_2}(\tau, \widehat{P}_X^{(n_2)}) - G_\infty(\tau, P_X)\right| > \delta\right) = o(1) \ .$$

Since (n_2) is a subsequence of (n_1) the previous two displays contradict each other. Hence, $\sup_\tau |G_n(\tau, \widehat{P}_X) - G_\infty(\tau, P_X)| = o_P(1)$ and we conclude.

□

2.3 Bootstrap Confidence Intervals

Having motivated the use of the bootstrap, in this section we present the construction of the bootstrap confidence intervals and the general conditions ensuring their consistency. We want to make inference on a population parameter θ_0 and have an estimator $\widehat{\theta}$. The statistic T_n takes the following form

$$T_n = \frac{\widehat{\theta}_n - \theta_0}{\widehat{\sigma}_n},$$

where $\widehat{\sigma}_n$ is a normalization factor depending on the data. Assume we have an estimator of the cdf of $(\widehat{\theta} - \theta_0)/\widehat{\sigma}_n$, say $\widehat{F}$, depending on the sample observations. For example, if we resample the data according to a probability $\widehat{P}$ constructed from the sample, and $(\widehat{\theta}^*, \widehat{\sigma}_n^*)$ are the versions of $(\widehat{\theta}, \widehat{\sigma}_n)$ computed with the data resampled from $\widehat{P}$, then $\widehat{F}$ can be set to the cdf of $(\widehat{\theta}^* - \widehat{\theta})/\widehat{\sigma}_n$. Given $\widehat{F}$ we can construct confidence intervals about θ_0 by using $(\widehat{\theta} - \theta_0)/\widehat{\sigma}_n$. To this end, let the β quantile of $\widehat{F}$ be

$$\widehat{F}^{-1}(\beta) := \inf\left\{q : \widehat{F}(q) \geq \beta\right\} \ .$$

The *bootstrap confidence interval* at the level $1 - \alpha - \beta$ is defined as

$$\left\{\theta \ : \ \widehat{\theta} - \widehat{\sigma}_n \widehat{\xi}_{1-\alpha} \leq \theta < \widehat{\theta} - \widehat{\sigma}_n \widehat{\xi}_\beta\right\} \tag{2.4}$$

where $\widehat{\xi}_\beta := \widehat{F}^{-1}(\beta)$. replace ξ_β by $\hat{\xi}_\beta$

Example. To better fix the ideas, consider the example of the sample mean introduced in the previous section. Then, $\widehat{\theta}_n = \overline{X}$, $\theta_0 = \mathbb{E}X$, and $\widehat{\sigma}_n = \sqrt{\mathrm{Var}(X)/n}$. In terms of the notation of the previous section, $T_n(X_1, .., X_n) = (\widehat{\theta}_n - \theta_0)/\widehat{\sigma}_n$. Let $\widehat{P}_X$ be the probability generated from drawing with replacement from the sample $\{X_i\}_{i=1}^n$. An estimator $\widehat{F}_n$ of

$\hat{P}_X$: boostrap distribution.

Bootstrap variance: $T_n^* = \dfrac{\bar{X}^* - \hat{\mathbb{E}}(X^*)}{\sqrt{\dfrac{\hat{V}^*(X^*)}{n}}}$ → need to compute expectation in Bootstrap world.

$\hat{V}^*(X^*) = \overline{(X^*)^2} - (\bar{X}^*)^2$

the distribution of $T_n(X_1, .., X_n)$ in this case is given by $\widehat{F}(\tau) := \widehat{\mathbb{E}}1\{T_n(X_1^*, .., X_n^*) \leq \tau\}$, where $\widehat{\mathbb{E}}$ is the expectation considering $\{X_i^*\}_{i=1}^n$ independent and identically distributed according to the probability $\widehat{P}_X$. The expectation $\widehat{\mathbb{E}}$ considers only $(X_1^*, \ldots, X_n^*)$ as random and the sample observations as fixed. Clearly, since $\widehat{P}_X$ depend on the data, such expectation will be stochastic.

$\hat{F}(\tau) = \hat{P}(T_n^* \leq \tau)$ $\quad X_i^* \overset{iid}{\sim} \hat{P}_X$

Constructing bootstrap confidence intervals. To construct a bootstrap confidence interval there are three elements that need to be chosen. First, we need to chose an estimator $\widehat{F}$ of the distribution of T_n. Second, we need to chose the normalization factor $\widehat{\sigma}_n$. Third, we need to chose how to compute the quantiles $\widehat{\xi}_\beta$ and $\widehat{\xi}_{1-\alpha}$. The decision of choosing $\widehat{F}$ will depend on the specific resampling scheme we want to implement, and this will often depend on the specific framework we have at hand. We will now briefly discuss the remaining two choices of $(\widehat{\xi}_{1-\alpha}, \widehat{\xi}_\beta)$ and $\widehat{\sigma}_n$.

There are several ways to choose $\widehat{\sigma}_n$ for constructing bootstrap confidence intervals. The main ones are: *the percentile t-method*, the *the percentile method*, and *Efron's percentile method.*

With the *percentile t-method*, $\widehat{\sigma}_n$ is chosen so as to make T_n is asymptotically pivotal (i.e. its asymptotic distribution does not depend on unknown parameters). For example, if $\sqrt{n}(\widehat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, V)$, then $\boxed{\widehat{\sigma}_n := \sqrt{\widehat{V}/n},}$ where $\widehat{V}$ is a consistent estimator of V, and $T_n = (\widehat{\theta} - \theta_0)/\widehat{\sigma}_n \rightsquigarrow \mathcal{N}(0, 1)$. So, the percentile t-method gives rise to a "studentized statistic" whose asymptotic standard deviation is 1. If the data is resampled according to an estimated probability $\widehat{P}$, and $(\widehat{\theta}^*, \widehat{\sigma}_n^*)$ are the versions of $(\widehat{\theta}, \widehat{\sigma}_n)$ computed according to the bootstrapped data (i.e. the data drawn from $\widehat{P}$), then the quantiles in (2.4) are computed according to

the best one

$$\widehat{\xi}_\gamma := \inf\left\{\xi \,:\, \widehat{P}\left(\frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\sigma}_n^*} \leq \xi\right) \geq \gamma\right\} \tag{2.5}$$

for $\gamma = 1 - \alpha, \beta$.

In the case of the *percentile method*, $\widehat{\sigma}_n$ is set to $1/\sqrt{n}$. The percentile method does not make the statistic asymptotically pivotal and, as seen in the previous section, the bootstrap in this case does not provide refinements over the asymptotic approximation. The main reason for choosing the percentile method over the percentile t-method arises when the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta_0)$ is difficult to estimate, so that it is more convenient to use a non-studentized version of T_n.

A drawback of both the percentile t-method and the percentile method is that their bootstrap confidence intervals are *not invariant to monotonic transformations*. Given a monotonic transformation ϕ (i.e. an increasing function), we could first obtain a bootstrap confidence interval for the transformed parameter $\phi(\theta_0)$ and then apply the inverse of ϕ to such confidence interval so as to get the interval for θ_0. In particular, consider the statistic

$T_n^\phi := (\phi(\widehat{\theta}) - \phi(\theta_0))/\widehat{\sigma}^\phi$, with $\widehat{\sigma}^\phi$ being a normalization factor depending on whether we are using the percentile method or the percentile t-method. The bootstrap confidence interval for $\phi(\theta_0)$ will be

$$\left\{\phi(\theta) \,:\, \phi(\widehat{\theta}) - \widehat{\sigma}^\phi \widehat{\xi}^\phi_{1-\alpha} \leq \phi(\theta) < \phi(\widehat{\theta}) - \widehat{\sigma}^\phi \widehat{\xi}^\phi_\beta\right\} ,$$

where $\widehat{\xi}^\phi_\beta$ is the β quantile of the cdf estimating the true cdf of T_n^ϕ. Since ϕ is a monotonic transformation, its inverse ϕ^{-1} can be applied to both sides of the above interval to get a confidence region for θ_0,

$$\left\{\theta \,:\, \phi^{-1}\left(\phi(\widehat{\theta}) - \widehat{\sigma}^\phi \widehat{\xi}^\phi_{1-\alpha}\right) \leq \theta < \phi^{-1}\left(\phi(\widehat{\theta}) - \widehat{\sigma}^\phi \widehat{\xi}^\phi_\beta\right)\right\} .$$

Notice that there is no guarantee that the above confidence region based on the transformation ϕ will coincide with the confidence interval (2.4) based on the originary statistic $(\widehat{\theta} - \theta_0)/\widehat{\sigma}_n$. In this sense, such intervals are not invariant to monotonic transformations.

Differently from the percentile method and the percentile t-method, the *Efron's percentile interval* is invariant to monotonic transformations. Similarly as above, assume that $\widehat{P}$ is a data-dependent probability used to resample our observations, and let $\widehat{\theta}^*$ be the bootstrap version of $\widehat{\theta}$. The confidence interval in the *Efron's percentile method* is defined as

$$\boxed{\left\{\theta \,:\, \widehat{\zeta}_\beta \leq \theta < \widehat{\zeta}_{1-\alpha}\right\}} \tag{2.6}$$

where $\widehat{\zeta}_\beta$ is the β quantile of the bootstrap distribution of $\widehat{\theta}^*$, i.e.

$$\widehat{\zeta}_\gamma := \inf\left\{\zeta \,:\, \widehat{P}\left(\widehat{\theta}^* \leq \zeta\right) \geq \gamma\right\} \tag{2.7}$$

for $\gamma = 1-\alpha, \beta$. Since $\phi(\widehat{\zeta}_\gamma)$ is the γ quantile of $\phi(\widehat{\theta}^*)$, the confidence interval based on Efron's percentile method is invariant to monotonic transformations.[1]

Notice that Efron's bootstrap confidence intervals are based on bootstrapping $\widehat{\theta}$, while the confidence intervals based on the percentile method and the percentile t-method are based on bootstrapping $(\widehat{\theta} - \theta_0)/\widehat{\sigma}_n$. Also, the confidence intervals obtained with Efron's percentile method coincide with the confidence intervals based on the percentile method if $\alpha = \beta$ and $\widehat{\theta}^* - \widehat{\theta}$ is symmetrically distributed. In particular, consider the confidence intervals obtained by setting $\beta = \alpha$. From (2.4) and (2.6) the confidence interval obtained with the percentile method coincides with the confidence interval obtained with Efron's method if and only if $\widehat{\zeta}_\alpha = \widehat{\theta} - \widehat{\xi}_{1-\alpha}/\sqrt{n}$ and $\widehat{\zeta}_{1-\alpha} = \widehat{\theta} - \widehat{\xi}_\alpha/\sqrt{n}$. Since (2.5) and (2.7) imply that $\widehat{\zeta}_\gamma = \widehat{\theta} + \widehat{\xi}_\gamma/\sqrt{n}$ for any $\gamma \in (0,1)$, the confidence intervals coincide if $\widehat{\xi}_\alpha = -\widehat{\xi}_{1-\alpha}$. This in turn happens if the distribution of $\widehat{\theta}^* - \widehat{\theta}$ is symmetric about zero.

[1] For any quantile $\widehat{\zeta}_\gamma$, we have that $\widehat{P}(\widehat{\theta}^* \leq \widehat{\zeta}_\gamma) \geq \gamma$.

Percentile method:

$$\left\{\theta : \hat{\theta} - \frac{\hat{\xi}_{1-\alpha}}{\sqrt{n}} \le \theta \le \hat{\theta} - \frac{\hat{\xi}_{\beta}}{\sqrt{n}}\right\}$$

Efron's method:

$$\left\{\theta : \hat{\zeta}_{\beta} \le \theta \le \hat{\zeta}_{1-\alpha}\right\}$$

where $\hat{\xi}_{\gamma}$ = γ quantile of the bootstrap distribution of $T_n^* = \dfrac{\hat{\theta}^* - \hat{\theta}}{\sqrt{\frac{1}{n}}}$

$\hat{\zeta}_{\gamma}$ = γ quantile of the bootstrap distribution of $\hat{\theta}^*$

Remark:

The CI with the Percentile method coincide with the CI of Efron's method if ① $\alpha = \beta$

② $\hat{\theta}^* - \hat{\theta}$ has a bootstrap distribution symmetric about 0.

(usually not satisfied)

Let us now briefly discuss how to choose the quantiles $\widehat{\xi}_\beta, \widehat{\xi}_{1-\alpha}$ and $\widehat{\zeta}_\beta, \widehat{\zeta}_{1-\alpha}$. In order to construct confidence intervals at the $1-\alpha-\beta$ level, we have different ways to choose such quantiles.

- To obtain *equal-tailed confidence intervals*, we can set $\alpha = \beta$ so as to get the level $1-2\alpha$. In this case, the confidence interval for the percentile-t method and the percentile method will be given by $[\widehat{\theta} - \widehat{\sigma}_n \widehat{\xi}_{1-\alpha}, \widehat{\theta} - \widehat{\sigma}_n \widehat{\xi}_\alpha]$. The confidence interval for Efron's percentile method will instead be $[\widehat{\zeta}_\alpha, \widehat{\zeta}_{1-\alpha}]$.

- To obtain a *symmetric confidence interval* for the percentile t-method and the percentile method, we first obtain the $1-\alpha$ quantile of $|\widehat{\theta}^* - \widehat{\theta}|/\widehat{\sigma}_n$, say $\widetilde{\xi}_{1-\alpha}$, and then set $[\widehat{\theta} - \widehat{\sigma}_n \widetilde{\xi}_{1-\alpha}, \widehat{\theta} + \widehat{\sigma}_n \widetilde{\xi}_{1-\alpha}]$ as a confidence interval at the $1-2\alpha$ level. Finally, the $1-2\alpha$ symmetric confidence interval for Efron's percentile method will be $[-\widetilde{\zeta}_{1-\alpha}, \widetilde{\zeta}_{1-\alpha}]$, where $\widetilde{\zeta}_{1-\alpha}$ is the $1-\alpha$ quantile of $\widehat{\theta}^*$.

Having explained how to construct bootstrap confidence intervals, we now show under which conditions they are valid, i.e. consistent. We say the the confidence interval is *asymptotically consistent* if

$$P\left(\widehat{\theta}_n - \widehat{\sigma}_n \widehat{F}_n^{-1}(1-\alpha) \le \theta_0 < \widehat{\theta}_n - \widehat{\sigma}_n \widehat{F}_n^{-1}(\beta)\right) = 1 - \alpha - \beta + o(1).$$

$\hat{F}^{-1}(\beta)$: β quantile of $\hat{F}$

The following result displays some sufficient conditions for the the consistency of the bootstrap confidence intervals. We will often use it in the next chapters to establish the validity of specific bootstrap schemes in the construction of confidence intervals.

$\hat{F}_n$ = boostrap cdf of $T_n^* = \dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}_n^*}$

Proposition 2.3.1. *Assume that*

(i) $(\widehat{\theta}_n - \theta_0)/\widehat{\sigma}_n \rightsquigarrow T$, *where* T *is a random variable with cdf* F

(ii) F *s strictly increasing and continuous*

difficult to prove.

(iii) $\sup_\tau |\widehat{F}_n(\tau) - F(\tau)| = o_P(1)$. normalité

$T_n = \dfrac{\hat{\theta} - \theta_0}{\hat{\sigma}_n}$

Then, the bootstrap confidence intervals for θ_0 *are asymptotically consistent.*

Bootstrap cdf of statistic

Proof. To show the desired result, we will first argue along subsequences. Consider an arbitrary subsequence (n_1) of (n). By condition (iii) and Proposition A.0.4 there exists a subsequence (n_2) of (n_1) such that $\sup_\tau |\widehat{F}_{n_2}(\tau) - F(\tau)| = o(1)$ almost surely. This, condition (ii), and Proposition A.0.13 imply that

$$\widehat{F}_{n_2}^{-1}(\gamma) = F^{-1}(\gamma) + o(1) \text{ for each } \gamma \in (0,1) \text{ almost surely} .$$

Since (n_2) is a subsequence of (n), by condition (i) we will also have

$$\frac{\widehat{\theta}_{n_2} - \theta_0}{\widehat{\sigma}_{n_2}} \rightsquigarrow T .$$

Now, fix $\gamma \in (0,1)$. The previous two displays together with Proposition A.0.7 and the Continuous Mapping Theorem (see Proposition A.0.8) ensure that

c.vg of cdf

$$\frac{(\widehat{\theta}_{n_2} - \theta_0)}{\widehat{\sigma}_{n_2}} - \widehat{F}_{n_2}^{-1}(\gamma) \rightsquigarrow T - F^{-1}(\gamma) \ .$$

Hence, the cdf of the random variable on the LHS will converge pointwise to the cdf of $T - F^{-1}(\gamma)$ *at each continuity point of the latter*, see Proposition A.0.5. At each $c \in \mathbb{R}$ the cdf of $T - F^{-1}(\gamma)$ is equal to $F(c + F^{-1}(\gamma))$.[2] Accordingly, the continuity of F ensures the continuity of the cdf of $T - F^{-1}(\gamma)$ over the entire real line and hence at 0. So, we find that

$= P(T - F^{-1}(\gamma) \leq 0) = P(T \leq F^{-1}(\gamma)) = F(F^{-1}(\gamma))$

$$P\left(\frac{(\widehat{\theta}_{n_2} - \theta_0)}{\widehat{\sigma}_{n_2}} - \widehat{F}_{n_2}^{-1}(\gamma) \leq 0\right) = F\left(F^{-1}(\gamma)\right) + o(1) = \gamma + o(1) \ .$$

$= \gamma$

Since the above display holds for any $\gamma \in (0,1)$, we also have

$$\begin{aligned} P\left(\widehat{\theta}_{n_2} - \widehat{\sigma}_{n_2}\widehat{F}_{n_2}^{-1}(1-\alpha) \leq \theta_0 < \widehat{\theta}_{n_2} - \widehat{\sigma}_{n_2}\widehat{F}_{n_2}^{-1}(\beta)\right) =& P\left(\widehat{F}_{n_2}^{-1}(\beta) < \frac{\widehat{\theta}_{n_2} - \theta_0}{\widehat{\sigma}_{n_2}} \leq \widehat{F}_{n_2}^{-1}(1-\alpha)\right) \\ =& P\left(\frac{\widehat{\theta}_{n_2} - \theta_0}{\widehat{\sigma}_{n_2}} \leq \widehat{F}_{n_2}^{-1}(1-\alpha)\right) \\ &- P\left(\frac{\widehat{\theta}_{n_2} - \theta_0}{\widehat{\sigma}_{n_2}} \leq \widehat{F}_{n_2}^{-1}(\beta)\right) \\ =& 1 - \alpha - \beta + o(1) \ . \end{aligned} \tag{2.8}$$

By recalling that (n_2) is a subsequence of (n_1) and (n_1) is an *arbitrary* subsequence of (n), Equation (2.8) must also hold along the entire sequence (n). This proves the desired result.[3]

□

2.4 Bootstrap hypothesis tests

The bootstrap method can also be used to construct tests of hypothesis for the parameter θ_0. We will consider two-sided tests, as the arguments for such a case extend straightforwardly to one-sided tests.

[2] To see this, let μ be the probability defining T. Then, the cdf of $T - F^{-1}(\gamma)$ evaluated at c is $\mu(T - F^{-1}(\gamma) \leq c) = \mu(T \leq F^{-1}(\gamma) + c) = F(F^{-1}(\gamma) + c)$, where in the last equality we have used the fact that F is the cdf of T.

[3] To see this, denote with S_{n_2} the probability on the LHS of (2.8). Assume that the result of the proposition does not hold. Then, there exists a subsequence (n_4) of (n) such that $|S_{n_4} - (1 - \alpha - \beta)| \geq \delta$ with $\delta > 0$. (2.8) ensures that from the subsequence (n_4) we can extract a further subsequence (n_5) such that $S_{n_5} = 1 - \alpha - \beta + o(1)$ so that we reach a contradiction.

Assume we are interested in testing

$$\mathcal{H}_0 : \theta_0 = \theta^{\mathcal{H}_0} \text{ versus } \mathcal{H}_1 : \theta_0 \neq \theta^{\mathcal{H}_0} \ ,$$

where $\theta^{\mathcal{H}_0}$ represents a specific value set by the researcher. A test for the above hypothesis can be built by starting from the two-sided confidence interval. In particular the decision rule can be defined as

- Do not reject $\mathcal{H}_0$ if

$$\widehat{\theta}_n - \widehat{\sigma}_n \widehat{F}_n^{-1}(1-\alpha) \leq \theta^{\mathcal{H}_0} < \widehat{\theta}_n - \widehat{\sigma}_n \widehat{F}_n^{-1}(\beta)$$

- Reject $\mathcal{H}_0$ otherwise

The following proposition shows the validity of the bootstrap two-sided test.

Proposition 2.4.1. *Under the same assumptions of Proposition 2.3.1 and $\widehat{\sigma}_n \xrightarrow{P} 0$ we have that*

(i) if $\mathcal{H}_0$ is true $P(\text{ accept } \mathcal{H}_0) = 1 - \alpha - \beta + o(1)$

(ii) if $\mathcal{H}_0$ does not hold, $P(\text{ reject } \mathcal{H}_0) = 1 + o(1)$

Proof. Assume first that $\mathcal{H}_0$ holds true. Then,

$$\begin{aligned} P(\text{ accept } \mathcal{H}_0) =& P\left(\widehat{\theta}_n - \widehat{\sigma}_n \widehat{F}_n^{-1}(1-\alpha) \leq \theta_0 < \widehat{\theta}_n - \widehat{\sigma}_n \widehat{F}_n^{-1}(\beta)\right) \\ =& 1 - \alpha - \beta + o(1) \end{aligned}$$

where the last equality follows from Proposition 2.3.1.
Assume now that $\mathcal{H}_1$ holds. Then,

$$P(\text{ reject } \mathcal{H}_0) = P\left(\theta^{\mathcal{H}_0} \geq \widehat{\theta}_n - \widehat{\sigma}_n \widehat{F}_n^{-1}(\beta)\right) + P\left(\theta^{\mathcal{H}_0} < \widehat{\theta}_n - \widehat{\sigma}_n \widehat{F}_n^{-1}(1-\alpha)\right) \ . \tag{2.9}$$

So, it suffices to sow that at least one of the two probabilities converges to 1. We distinguish two cases: $\theta_0 > \theta^{\mathcal{H}_0}$ and $\theta_0 < \theta^{\mathcal{H}_0}$.

Case $\theta_0 > \theta^{\mathcal{H}_0}$. Consider an arbitrary $\epsilon > 0$ and define $\Delta_n := F^{-1}(1-\alpha) - \widehat{F}_n^{-1}(1-\alpha)$. Then,

$$
\begin{aligned}
P\left(F^{-1}(1-\alpha)+\epsilon<\frac{\widehat{\theta}_n-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n}\right) \leq & P\left(\epsilon+\Delta_n+\widehat{F}_n^{-1}(1-\alpha)<\frac{\widehat{\theta}_n-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} \cap |\Delta_n|<\epsilon\right) \\
& + P(|\Delta_n| \geq \epsilon) \\
\leq & P\left(\widehat{F}_n^{-1}(1-\alpha)<\frac{\widehat{\theta}_n-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n}\right)+P(|\Delta_n| \geq \epsilon) \ .
\end{aligned}
\tag{2.10}
$$

Proposition A.0.13 and $\sup_\tau |\widehat{F}_n(\tau)-F(\tau)| = o_P(1)$ ensure that $P(|\Delta_n| \geq \epsilon) = o(1)$. Since the first probability on the RHS of the above display is equal to the second addendum on the RHS of (2.9), it is sufficient to prove that the LHS of the above display converges to 1. For any $M > [F^{-1}(1-\alpha)+\epsilon] \vee 0$ such a term is bounded from below by

$$
\begin{aligned}
P\left(\frac{\widehat{\theta}_n-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n}>M\right) \geq & P\left(\left|\frac{\widehat{\theta}_n-\theta_0}{\widehat{\sigma}_n}\right| \leq M \ \cap \frac{\theta_0-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n}>2M\right) \\
= & 1-P\left(\left|\frac{\widehat{\theta}_n-\theta_0}{\widehat{\sigma}_n}\right|>M \ \cup \ \frac{\theta_0-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} \leq 2M\right) \\
\geq & 1-P\left(\left|\frac{\widehat{\theta}_n-\theta_0}{\widehat{\sigma}_n}\right|>M\right)-P\left(\frac{\theta_0-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} \leq 2M\right) \ .
\end{aligned}
$$

Now, since $(\widehat{\theta}_n-\theta_0)/\widehat{\sigma}_n = O_P(1)$, the first probability on the RHS can be made arbitrarily small by choosing M large enough. Since $\theta_0-\theta^{\mathcal{H}_0} > 0$ and $\widehat{\sigma}_n = o_P(1)$, the second probability converges to zero. Hence, the result is proved for the case $\theta_0-\theta^{\mathcal{H}_0} > 0$.

Case $\theta_0 < \theta^{\mathcal{H}_0}$. Using the same reasoning as in (2.10), for any $\epsilon > 0$ we have

$$
P\left(\frac{\widehat{\theta}_n-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} \leq F^{-1}(\beta)-\epsilon\right) \leq P\left(\frac{\widehat{\theta}_n-\theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} \leq \widehat{F}_n^{-1}(\beta)\right)+P(|\Delta_n|>\epsilon) \ ,
$$

with $\Delta_n := F^{-1}(\beta)-\widehat{F}_n^{-1}(\beta)$. As noticed in the first part of the proof, $\Delta_n = o_P(1)$. Since the first probability on the RHS of the above display equals the first term on the RHS of (2.9), it is sufficient to prove that such a probability converges to 1. For

$M < [F^{-1}(\beta) - \epsilon] \wedge 0$, such probability is bounded from below by

$$
\begin{aligned}
P\left(\frac{\widehat{\theta}_n - \theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} \leq M\right) \geq & P\left(\left|\frac{\widehat{\theta}_n - \theta_0}{\widehat{\sigma}_n}\right| \leq |M| \ \cap \frac{\theta_0 - \theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} < 2M\right) \\
= & 1 - P\left(\left|\frac{\widehat{\theta}_n - \theta_0}{\widehat{\sigma}_n}\right| > |M| \ \cup \frac{\theta_0 - \theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} \geq 2M\right) \\
\geq & 1 - P\left(\left|\frac{\widehat{\theta}_n - \theta_0}{\widehat{\sigma}_n}\right| > |M|\right) - P\left(\frac{\theta_0 - \theta^{\mathcal{H}_0}}{\widehat{\sigma}_n} \geq 2M\right) .
\end{aligned}
$$

As already argued before, the first probability on the RHS can be made arbitrarily small by choosing $|M|$ large enough. Also, since $M < 0$, $\theta^{\mathcal{H}_0} - \theta_0 > 0$, and $\widehat{\sigma}_n = o_P(1)$ it holds that $P(\theta_0 - \theta^{\mathcal{H}_0} \geq 2M\widehat{\sigma}_n) = P((\theta^{\mathcal{H}_0} - \theta_0)/(2|M|) \leq \widehat{\sigma}_n) = o(1)$. This delivers the desired result.

$T_n = \frac{\exp(\bar{X}) - \exp(\mathbb{E}(X))}{\sqrt{\frac{1}{n}}}$

2.5 The Bootstrap Delta Method

In several occasions we will have to prove the validity of the bootstrap for transformations of a certain statistic. Assume to have an estimator $\widehat{\theta}_n$, its bootstrap counterpart $\widehat{\theta}^*$, and a transformation ϕ. Imagine we are interested in the validity of the bootstrap for $\sqrt{n}[\phi(\widehat{\theta}) - \phi(\theta_0)]$, i.e. we are interested in showing that $\sqrt{n}[\phi(\widehat{\theta}^*) - \phi(\widehat{\theta})]$ estimates consistently the distribution of $\sqrt{n}[\phi(\widehat{\theta}) - \phi(\theta_0)]$. The bootstrap delta method ensures that if the bootstrap is valid for $\widehat{\theta}$ (i.e. if $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ estimates consistently the distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$), then it will be valid also for $\phi(\widehat{\theta})$ (i.e. the bootstrap will estimate consistently the distribution of $\sqrt{n}(\phi(\widehat{\theta}) - \phi(\theta_0))$). Such a result is valid as long as the transformation ϕ is smooth.

Proposition 2.5.1. *Consider a parameter $\theta_0 \in \mathbb{R}^s$ and an estimator $\widehat{\theta}_n$ such that $\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow T$ for a certain random variable T with a continuous cdf.*
Let $\phi : \mathbb{R}^s \mapsto \mathbb{R}^k$ be a function that is continuously differentiable in a neighborhood of θ_0.
Let $\widehat{\theta}_n^*$ be the bootstrap counterpart of $\widehat{\theta}_n$.
If $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}) \rightsquigarrow T$ in probability, then

$$
\begin{gathered}
\sqrt{n}(\phi(\widehat{\theta}_n) - \phi(\theta_0)) \rightsquigarrow \phi_{\theta_0} T \\
\text{and } \sqrt{n}(\phi(\widehat{\theta}_n^*) - \phi(\widehat{\theta}_n)) \rightsquigarrow \phi_{\theta_0} T \text{ in probability}
\end{gathered}
$$

where ϕ_{θ_0} denotes the gradient of ϕ computed at θ_0.

To better grasp the proposition, recall that "$\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \rightsquigarrow T$ in probability" means that the bootstrap cdf of $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$, say $\widehat{F}_n$, converges in probability to the cdf of T,

Delta method:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} T$$

$\phi : \mathbb{R} \to \mathbb{R}$ is differentiable at θ_0 with ϕ' continue on θ_0

Then $\sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta_0)) \xrightarrow{d} \phi'(\theta_0) \cdot T$

say F_T. Hence, since the cdf of T is continuous, we have $\sup_x \left|\widehat{F}_n(x) - F_T(x)\right| = o_P(1)$.
Similarly, "$\sqrt{n}(\phi(\widehat{\theta}_n^*) - \phi(\widehat{\theta}_n)) \rightsquigarrow \phi_{\theta_0} T$ in probability" means that the bootstrap cdf of $\sqrt{n}(\phi(\widehat{\theta}_n^*) - \phi(\widehat{\theta}_n))$ converges uniformly in probability to the cdf of $\phi_{\theta_0} T$.
The proposition therefore implies that the bootstrap cdf of $\sqrt{n}(\phi(\widehat{\theta}_n^*) - \phi(\widehat{\theta}_n))$ estimates consistently the cdf of $\sqrt{n}(\phi(\widehat{\theta}_n) - \phi(\theta_0))$.

Proof. Let $\widehat{F}_n$ be the bootstrap cdf of $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$. Consider an arbitrary subsequence of (n). By the assumptions on $\widehat{\theta}_n$ and $\widehat{\theta}_n^*$ there exists a further subsequence (m) such that $\widehat{\theta}_m = \theta_0 + o(1)$ and $\sup_x \left|\widehat{F}_m - F_T(x)\right| = o(1)$, where the two equalities hold almost surely. So, they hold on a set A having probability one. Fix an event $\omega \in A$ and denote with $P_n^{*,\omega}$ the probability in the bootstrap world that corresponds to such fixed event. Assume that $\sqrt{m}\|\widehat{\theta}_m^* - \widehat{\theta}_m\| \leq M$ for some $M > 0$.
Since ϕ is continuously differentiable at θ_0, for any $\eta > 0$ arbitrarily small there exists a $\delta > 0$ such that $\|\theta - \theta_0\| < \delta$ implies $\|\phi_\theta - \phi_{\theta_0}\| < \eta/M$. So, fix $\eta > 0$ arbitrarily small and consider the corresponding δ.
Since $\widehat{\theta}_m = \theta_0 + o(1)$, for each m large enough we have $\|\widehat{\theta}_m - \theta_0\| < \delta/2$. Using this and $\sqrt{m}\|\widehat{\theta}_m^* - \widehat{\theta}_m\| \leq M$ leads to $\|\widehat{\theta}_m^* - \theta_0\| \leq \|\widehat{\theta}_m^* - \widehat{\theta}_m\| + \|\widehat{\theta}_m - \theta_0\| \leq M/\sqrt{m} + \delta/2 < \delta$ for any m large enough. Hence, by Taylor's Theorem (see Proposition A.0.11) we have

$$\sqrt{m}\left[\phi(\widehat{\theta}_m^*) - \phi(\widehat{\theta}_m)\right] = \phi_{\widetilde{\theta}^*}\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m) \text{ with } \widetilde{\theta}^* \in [\widehat{\theta}_m^*, \widehat{\theta}_m]\,.$$

Since $\widetilde{\theta}^* \in [\widehat{\theta}_m^*, \widehat{\theta}_m]$, with $\|\widehat{\theta}_m - \theta_0\| < \delta$ and $\|\widehat{\theta}_m^* - \theta_0\| < \delta$, we also have that $\|\widetilde{\theta}_m^* - \theta_0\| < \delta$. So, the definition of δ ensures that $\|\phi_{\widetilde{\theta}^*} - \phi_{\theta_0}\| < \eta$ and hence by the above display

$$\begin{aligned}\left\|\sqrt{m}\left[\phi(\widehat{\theta}_m^*) - \phi(\widehat{\theta}_m)\right] - \phi_{\theta_0}\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_{n_1})\right\| \leq & \|\phi_{\widetilde{\theta}_m^*} - \phi_{\theta_0}\| \cdot \sqrt{m}\|\widehat{\theta}_m^* - \widehat{\theta}_m\| \\ \leq & \eta\,.\end{aligned}$$

Since the inequality just obtained is implied by $\sqrt{m}\|\widehat{\theta}_m^* - \widehat{\theta}_m\| \leq M$, we obtain that

$$P_m^{*,\omega}\left(\left\|\sqrt{m}\left[\phi(\widehat{\theta}_m^*) - \phi(\widehat{\theta}_m)\right] - \phi_{\theta_0}\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m)\right\| > \eta\right) \leq P_m^{*,\omega}\left(\sqrt{m}\|\widehat{\theta}_m^* - \widehat{\theta}_m\| > M\right)$$

Recall that $\sup_x |\widehat{F}_m^\omega(x) - F_T(x)| = o(1)$. Convergence in distribution implies boundedness in probability, so the probability on the RHS of the previous display can be made arbitrarily small by choosing M large enough. Since η was arbitrarily chosen, we find that

$$P_m^{*,\omega}\left(\left\|\sqrt{m}\left[\phi(\widehat{\theta}_m^*) - \phi(\widehat{\theta}_m)\right] - \phi_{\theta_0}\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m)\right\| > \eta\right) \to 0$$

as $m \to 0$ for any $\eta > 0$. As already noticed,$\sup_x |\widehat{F}_m^\omega(x) - F_T(x)| = o(1)$, so by the

continuous mapping theorem for convergence in distribution, we get that

$$\phi_{\theta_0}\sqrt{m}(\widehat{\theta}^*_m - \widehat{\theta}_m) \rightsquigarrow \phi_0 T \text{ conditionally on } \omega$$

i.e. for the fixed ω considered, the cdf of $\phi_{\theta_0}\sqrt{m}(\widehat{\theta}^*_m - \widehat{\theta}_m)$ converges uniformly to the cdf of $\phi_{\theta_0}T$. Combining the two above displays and Slutzky's Theorem ensures that

$$\sqrt{m}\left[\phi(\widehat{\theta}^*_m) - \phi(\widehat{\theta}_m)\right] \rightsquigarrow \phi_0 T \text{ conditionally on } \omega$$

i.e. for the fixed ω considered the cdf of $\sqrt{m}\left[\phi(\widehat{\theta}^*_m) - \phi(\widehat{\theta}_m)\right]$ converges uniformly to the cdf of $\phi_{\theta_0}T$. Since ω is an arbitrary element of A and A has probability one, we find that

$$\sqrt{m}\left[\phi(\widehat{\theta}^*_m) - \phi(\widehat{\theta}_m)\right] \rightsquigarrow \phi_0 T \text{ almost surely}$$

i.e. the cdf of the LHS converges uniformly and almost surely to the cdf of the RHS. Since (m) was a subsequence of an arbitrary subsequence of (n) (see the beginning of this proof),the result of the proposition is proved. □

In many cases, statistics of the type $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ are not asymptotically pivotal, in the sense that their asymptotic distributions depend on unknown features of the DGP. This happens e.g. for the sample mean, where $\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, V)$, with the variance V that is unknown but estimable. As already seen at the beginning of this chapter, the bootstrap does not provide refinements over the asymptotic approximation when the statistic is not asymptotically pivotal. So, when we can choose, it is better to use the bootstrap for a statistic whose asymptotic distribution is parameter-free and totally known. In a great number of cases, we can make a statistic asymptotically pivotal by dividing it by the estimator of the asymptotic standard deviation. So, in a great number of cases we conduct inference by bootstrapping $\sqrt{n}(\widehat{\theta}_n - \theta_0)/\sqrt{\widehat{V}}$.
Now, imagine that we have obtained the validity of the bootstrap for $\sqrt{n}(\widehat{\theta}_n - \theta_0)$, but our interest is in proving the validity of the bootstrap for its studentized version $\sqrt{n}(\widehat{\theta}_n - \theta_0)/\sqrt{\widehat{V}}$. The following proposition allows to obtain this. We will denote by $\mathbb{M}^k$ the space of square matrices of dimension k.

Proposition 2.5.2. *Let $g : \mathbb{R}^s \mapsto \mathbb{M}^k$ be continuous at μ and assume that $\sqrt{n}(\widehat{\theta}^*_n - \widehat{\theta}_n) \rightsquigarrow T$ in probability. If $\|\overline{Z}_n - \mu\| = o_P(1)$ and $\|\overline{Z}^*_n - \overline{Z}_n\| = o_{P^*}(1)$ then*

$$g(\overline{Z^*}_n)\sqrt{n}(\widehat{\theta}^*_n - \widehat{\theta}_n) \rightsquigarrow g(\mu)\,T \text{ in probability}$$

In the above proposition $\|\overline{Z}_n - \overline{Z}_n\| = o_{P^*}(1)$ means that for any $\varepsilon > 0$ $P^*_n(\|\overline{Z^*}_n - \overline{Z}_n\| > \varepsilon)$ converges in probability to zero, where P^*_n denotes the probability in the bootstrap world.

$$\beta_0 = \mathbb{E}(XX^T)^{-1}\,\mathbb{E}(XY)$$

$$\hat{\beta} = (\overline{XX^T})^{-1}\,\overline{XY}$$

$$\hat{\beta}^* = (\overline{X^*X^{*T}})^{-1}\,\overline{X^*Y^*}$$

$$\frac{\sqrt{n}\,(\hat{\beta}-\beta_0)}{\sqrt{\hat{\Sigma}}} \quad \text{vs.} \quad \frac{\sqrt{n}\,(\hat{\beta}^*-\hat{\beta})}{\sqrt{\hat{\Sigma}^*}}$$

Proof. Since for any $\varepsilon > 0$ $P_n^*(\|\overline{Z^*}_n - \overline{Z}_n\| > \varepsilon) = o_P(1)$, it is possible to construct a sequence $(\varepsilon_n)_n$ such that $\varepsilon_n \to 0$ and $P_n^*(\|\overline{Z^*}_n - \overline{Z}_n\| > \varepsilon_n) = o_P(1)$.
Now, consider an arbitrary subsequence of (n). Then, it is possible to extract a further subsequence (m) such that almost surely: $P_m^*(\|\overline{Z}_m - \overline{Z}_m\| > \varepsilon_m) = o(1)$, $\|\overline{Z}_m - \mu\| = o(1)$, and $\sup_x |\widehat{F}_m(x) - F_T(x)| = o(1)$, where $\widehat{F}_m$ denotes the bootstrap cdf of $\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m)$. So, these convergences hold on a set A having probability one. Fix an event $\omega \in A$. Denote with $P_n^{*,\omega}$ the probability in the bootstrap world that corresponds to such a fixed event and with $\widehat{F}_n^\omega$ the cdf of $\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m)$ that corresponds to such an event. Then, for any $\varepsilon > 0$ we have

$$P_m^{*,\omega}\left(\|\overline{Z^*}_m - \mu\| > \varepsilon\right) = o(1)$$

where we have used the fact that $\|\overline{Z}_m^* - \mu\| \leq \|\overline{Z}_m^* - \overline{Z}_m\| + \|\overline{Z}_m - \mu\|$, with $\|\overline{Z}_m - \mu\| = o(1)$ and $P_m^{*,\omega}(\|\overline{Z}_m^* - \overline{Z}_m\| > \varepsilon) = o(1)$. By the above display, the continuity of g at μ, and the Continuous Mapping Theorem (see Proposition A.0.8) we get

$$P_m^{*,\omega}\left(\|g(\overline{Z}_m^*) - g(\mu)\| > \varepsilon\right) = o(1)$$

for any $\varepsilon > 0$. Using the above display, $\sup_x |\widehat{F}_m^\omega(x) - F_T(x)| = o(1)$, and Slutsky's theorem, we obtain that

$$g(\overline{Z}_m^*)\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m) \rightsquigarrow g(\mu)\,T \text{ conditionally on } \omega\,.$$

So, for the fixed ω considered, the cdf of $g(\overline{Z}_m^*)\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m)$ converges uniformly to the cdf of $g(\mu)\,T$. Since ω was an arbitrary element of A and A has probability one, we find that

$$g(\overline{Z}_m^*)\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m) \rightsquigarrow g(\mu)\,T \text{ almost surely }\ ,$$

i.e. the cdf of $g(\overline{Z}_m^*)\sqrt{m}(\widehat{\theta}_m^* - \widehat{\theta}_m)$ converges uniformly and almost surely to the cdf of $g(\mu)\,T$. Recalling that (m) was a subsequence of an arbitrary subsequence of (n) concludes the proof.

□

Notes

The first part of this chapter showing the bootstrap refinements by mean of the Edgeworth expansions is closely inspired to Davison and Hinkley (1997) and Horowitz (2001). The arguments provided here are heuristics, but a detailed theory for Edgeworth expansions can be found in Hall (2013). The parts of the chapter showing the consistency of the bootstrap for the construction of the confidence intervals and the bootstrap delta method follow Chapter 23 in van der Vaart (2000).
Beyond the already cited texts, an introducton to the theory of the bootstrap is contained in Shao and Tu (1995) and Efron and Tibshirani (1994).

Chapter 3

The Pairwise Bootstrap

The pairwise bootstrap is one of the most common bootstrap schemes. In this chapter we will show its validity for constructing confidence intervals and tests of hypothesis for unknown population parameters. We will first focus on the case where such parameters can be expressed as functions of expectations. This holds in many estimation frameworks, like OLS, IVs, etc. Then, we will consider population parameters having other forms, and in particular we will focus on the estimation of nonparametric densities, nonparametric mean regressions, and M estimation. The organization goes is as follows. Section 3.1 introduces some necessary technical concepts. Section 3.2 states and proves a bootstrap Central Limit Theorem for the pairwise resampling scheme. Section 3.3 shows the validity of the pairwise bootstrap for parameters that are functions of the population mean. Section 3.4 consider the extension to nonparametric density estimation and mean regressions. Section 3.5 considers the extension to a class of M estimators. Finally Section 3.7 provides a list of counterexamples where the pairwise bootstrap does not work.

3.1 Construction of a Probability in the Pairwise Bootstrap Scheme

Assume to have a sample $\{X_i\}_{i=1}^n$ of iid random variables taking values in $\mathbb{R}^M$. Essentially the pairwise bootstrap is based on the idea of replacing the population with the sample data $\{X_i\}_{i=1}^n$ and drawing *with replacement* observations from it. In this section, we first define rigorously the probability at the basis of the pairwise bootstrap and then show how to compute expectations based on such a probability.

$X_i \in \mathbb{R}^M$

Consider any subset $A \subset \mathbb{R}^M$ and define the indicator function as

$$\mathbb{I}(X_i \in A) = \begin{cases} 1 \text{ if } X_i \in A \\ 0 \text{ otherwise} \end{cases}$$

The probability at the basis of the pairwise bootstrap is defined as

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in A) \ , \quad A \subseteq \mathbb{R}^M \tag{3.1}$$

Notice that $\mathbb{P}_n$ is random probability, as it depends on the sample data $\{X_i\}_{i=1}^n$. In order to prove rigorously that it is a probability measure we need to verify two conditions. The first is that $\mathbb{P}_n(\mathbb{R}^M) = 1$. The second is that for any countable collection $(A_s)_s$ of pairwise disjoint sets (i.e. of sets such that $A_s \cap A_k = \emptyset$ for $k \neq s$) it holds that $\mathbb{P}_n(\cup_s A_s) = \sum_s \mathbb{P}_n(A_s)$. Now, the first of these two conditions is obviously satisfied. To verify the second condition, consider any countable collection $(A_s)_s$ of pairwise disjoint sets. Then, $\mathbb{I}(X_i \in \cup_s A_s)/n = \sum_s [\mathbb{I}(X_i \in A_s)/n]$. In fact, if $X_i \in \cup_s A_s$ we have $\mathbb{I}(X_i \in A_s) = 1$ for a *unique* s, as $(A_s)_s$ are pairwise disjoint. If instead $X_i \notin \cup_s A_s$, then $\mathbb{I}(X_i \in A_s) = 0$ for all s. Thus,

$$\mathbb{P}_n(\cup_s A_s) = \sum_{i=1}^{n} \frac{\mathbb{I}(X_i \in \cup_s A_s)}{n} = \sum_{i=1}^{n}\sum_{s} \frac{\mathbb{I}(X_i \in A_s)}{n} = \sum_{s}\sum_{i=1}^{n} \frac{\mathbb{I}(X_i \in A_s)}{n} = \sum_{s} \mathbb{P}_n(A_s) \ ,$$

where in the third equality we have exchanged the summations thanks to Tonelli-Fubini's theorem. Thus, $\mathbb{P}_n$ is a probability measure and as such it defines an expectation that we denote with $\mathbb{E}_n$.

If $X^* \sim \mathbb{P}_n$, for any function g the expectation $\mathbb{E}_n g(X^*)$ can be expressed as a sample mean, as the following proposition shows.

Proposition 3.1.1. *Let $\mathbb{E}_n$ be the expectation defined by the probability measure in (3.1) and let $X^* \sim \mathbb{P}_n$. Then, for any function $g : \mathbb{R}^M \mapsto \mathbb{R}$*

$$\mathbb{E}_n g(X^*) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \quad \longrightarrow \quad \hat{\mathbb{E}}(X^*) = \bar{X}$$

Proof. From the definition of $\mathbb{P}_n$ in (3.1), the set containing the sample observations $\{X_1, \ldots, X_n\}$ has $\mathbb{P}_n$ probability one, i.e. $\boxed{\mathbb{P}_n(\{X_1, \ldots, X_n\}) = 1}$. Hence, a random variable $X^* \sim \mathbb{P}_n$ can take only the values $X_1, \ldots, X_n$. Assume now that such a set has no tie (i.e. $X_1, \ldots, X_n$ are all mutually different). Then, using again (3.1) leads to $\mathbb{P}_n(\{X_i\}) = 1/n$ for each $i = 1, \ldots, n$. Accordingly,

$$\mathbb{E}_n g(X^*) = \sum_{i=1}^{n} g(X_i)\, \mathbb{P}_n(\{X_i\}) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \, .$$

The proof in the case where the set $\{X_1, \ldots, X_n\}$ has ties proceeds along the same lines. $\square$

3.2 The Pairwise Bootstrap CLT

In this section we prove a CLT that will be used to show the validity of the pairwise bootstrap for constructing confidence intervals and tests of hypothesis. The proof reported here is based on the Lindeberg-Feller CLT. Alternative proving strategies are based on the convergence of the characteristic function or use the Berry Hessen inequality.

$$\mathbb{E}_n[X_i^*] = \overline{X}$$

Proposition 3.2.1. *Assume that $\mathbb{E}||X||^2 < \infty$. Let $\{X_i^*\}_{i=1}^n$ be an iid sample from $\mathbb{P}_n$. Then, with probability one*

i.e.

$$\sqrt{n}(\overline{X^*} - \overline{X}) \rightsquigarrow \mathcal{N}(0, Var(X)) \ , \qquad P_n(\sqrt{n}(\overline{X}^* - \overline{X}) \leq t) \xrightarrow{\text{proba}} \Phi_{Var(X)}(t)$$

where

$$\overline{X^*} := \frac{1}{n}\sum_{i=1}^n X_i^* \ \textit{and} \ \overline{X} := \frac{1}{n}\sum_{i=1}^n X_i \ .$$

$\forall t \in \mathbb{R}$

(pointwise).

The above result has to be understood in the following sense. Let P_n^* be the probability measure that consider $\{X_i^*\}_{i=1}^n$ iid from $\mathbb{P}_n$. Clearly, since $\mathbb{P}_n$ is random, also P_n^* will be random. Then, the cdf of $\overline{X^*}$ given by $s \mapsto P_n^*\left(\sqrt{n}\left(\overline{X^*} - \overline{X}\right) \leq s\right)$ will also be random. The above theorem states that with probability one we have

$$\sup_s \left|P_n^*\left(\sqrt{n}\left(\overline{X^*} - \overline{X}\right) \leq s\right) - \Phi_V(s)\right| = o(1) \ ,$$

where Φ_V denotes the cdf of $\mathcal{N}(0, V = Var(X))$.

Proof. We will prove the conditions of the Lindeberg Feller CLT (Proposition A.0.9). To this end, denote with E_n^* the expectation that considers the bootstrap sample $\{X_i^*\}_{i=1}^n$ iid from $\mathbb{P}_n$. E_n^* is called *bootstrap expectation*. Recall that $\mathbb{E}_n$ is the expectation defined by the probability $\mathbb{P}_n$. Then,

$$E_n^* X_i^* = \mathbb{E}_n X_i^* = \overline{X} \ ,$$

where the second equality is obtained from Proposition 3.1.1. Accordingly, each term in the sum

$$\sqrt{n}(\overline{X^*} - \overline{X}) = \sum_{i=1}^n \frac{X_i^* - \overline{X}}{\sqrt{n}} =: \sum_{i=1}^n Z_i^* \tag{3.2}$$

has zero bootstrap expectation. To show the second condition of the Lindeberg Feller

CLT, i.e. the convergence of the covariances, notice that

$$\begin{aligned}\sum_{i=1}^{n} E_n^* Z_i^* Z_i^{*T} =& \sum_{i=1}^{n} \frac{E_n^* X_i^* X_i^{*T} - [E_n^* X_i^*]\, \overline{X}^T - \overline{X}\, \left[E_n^* X_i^{*T}\right] + \overline{X}\, \overline{X}^T}{n} \\ =& E_n^* X_i^* X_i^{*T} - [E_n^* X_i^*]\, \overline{X}^T - \overline{X}\, \left[E_n^* X_i^{*T}\right] + \overline{X}\, \overline{X}^T \\ =& \mathbb{E}_n X_i^* X_i^{*T} - [\mathbb{E}_n X_i^*]\, \overline{X}^T - \overline{X}\, \left[\mathbb{E}_n X_i^{*T}\right] + \overline{X}\, \overline{X}^T \\ =& \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T - \overline{X}\, \overline{X}^T \ ,\end{aligned}$$

where in the second equality we have used the iid structure of $\{X_i^*\}_{i=1}^n$ and in the fourth equality we have used Proposition 3.1.1. By the Strong Law of Large Numbers, the RHS of the above expression converges almost surely to $Var(X)$. Hence, we get

$$\sum_{i=1} E_n^* \{Z_i^* Z_i^{*T}\} \to Var(X) \ P\text{-almost surely} \ . \tag{3.3}$$

The above convergence result is an "almost sure" version of condition 1 of the Liendeberg Feller CLT. To obtain the second condition, consider a sequence $(\epsilon_n)_n$ of strictly positive numbers decreasing to 0 such that $\sqrt{n}\epsilon_n \to \infty$. We first use Proposition 3.1.1 to obtain

$$\begin{aligned}\sum_{i=1}^{n} E_n^* \left|\left|\frac{X_i^*}{\sqrt{n}}\right|\right|^2 \mathbb{I}\left(\left|\left|\frac{X_i^*}{\sqrt{n}}\right|\right| > \epsilon_n\right) =& \frac{1}{n} \sum_{i=1}^{n} E_n^* \left|\left|X_i^*\right|\right|^2 \mathbb{I}\left(\left|\left|X_i^*\right|\right| > \sqrt{n}\epsilon_n\right) \\ =& \mathbb{E}_n \left|\left|X_i^*\right|\right|^2 \mathbb{I}\left(\left|\left|X_i^*\right|\right| > \sqrt{n}\epsilon_n\right) \\ =& \frac{1}{n} \sum_{i=1}^{n} \left|\left|X_i\right|\right|^2 \mathbb{I}\left(\left|\left|X_i\right|\right| > \sqrt{n}\epsilon_n\right) \ ,\end{aligned} \tag{3.4}$$

where in the second equality we have used the iid structure of $\{X_i^*\}_{i=1}^n$ and in the last equality we have used Proposition 3.1.1. Since by assumption $\mathbb{E}||X||^2$ is finite, the Lebesgue Dominated Convergence Theorem (see Proposition A.0.10(ii)) ensures that $\mathbb{E}||X||^2\mathbb{I}(||X|| > N) \to 0$ for any $N \to \infty$. Hence, for any $\delta > 0$ there exists a N_δ large enough such that $\mathbb{E}||X||^2\mathbb{I}(||X|| > N_\delta) < \delta$. Now, consider a countable sequence $(\delta_s)_s$ decreasing to 0 and the associated sequence $(N_{\delta_s})_s$. Fix s. Since $\sqrt{n}\epsilon_n \to \infty$, for some n large enough and by the Strong Law of Large Numbers we have

$$\begin{aligned}0 \leq \frac{1}{n} \sum_{i=1}^{n} \left|\left|X_i\right|\right|^2 \mathbb{I}\left(\left|\left|X_i\right|\right| > \sqrt{n}\epsilon_n\right) \leq& \frac{1}{n} \sum_{i=1}^{n} \left|\left|X_i\right|\right|^2 \mathbb{I}\left(\left|\left|X_i\right|\right| > N_{\delta_s}\right) \\ =& \mathbb{E}||X||^2 \mathbb{I}\left(\left|\left|X\right|\right| > N_{\delta_s}\right) + o(1) \ P\text{-almost surely} \\ \leq& \delta_s + o(1) \ P\text{-almost surely} \ .\end{aligned}$$

Since $(\delta_s)_s$ is a discrete sequence tending to 0 and the countable intersection of events

with probability one still has probability one, the previous display implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} ||X_i||^2 \mathbb{I}\left(||X_i|| > \sqrt{n}\epsilon_n\right) = 0 \ P\text{- almost surely}$$

From the above display and (3.4) we find that

$$\lim_{n\to\infty} \sum_{i=1}^{n} E_n^* \left|\left|\frac{X_i^*}{\sqrt{n}}\right|\right|^2 \mathbb{I}\left(\left|\left|\frac{X_i^*}{\sqrt{n}}\right|\right| > \epsilon_n\right) = 0 \ P\text{- almost surely.}$$

Using the above display and Equation (3.3), we obtain that over an event with probability one the conditions of the Liendeberg Feller CLT holds for the sum in (3.2). This shows the desired result.

□

3.3 Applications to the Main Examples

In this section we apply the results on the bootstrap to two popular estimators, namely the OLS and the 2SLS estimators.

Bootstrapping OLS estimators. We observe $\{Y_i, Z_i\}_{i=1}^n$, with Z_i valued in $\mathbb{R}^k$ and Y_i real valued. We are interested in making inference on the coefficients of the linear projection of Y onto Z. So, given the regression model

$$Y_i = Z^T\beta_0 + \varepsilon \text{ with } \mathbb{E}\varepsilon Z = 0$$

we are interested in making inference on β_0. Assume that $\mathbb{E}ZZ^T$ is invertible (i.e. there is no collinearity between the explanatory variables Z). The OLS estimator is

$$\widehat{\beta} := \left(\overline{ZZ^T}\right)^{-1} \overline{ZY}$$

Also, notice that the coefficients β_0 are identified as

$$\beta_0 = \left(\mathbb{E}ZZ^T\right)^{-1} \mathbb{E}ZY$$

From basic statistical/econometric textbooks, we know that delta method. proposition 2.5.1.

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}(0, \Sigma) \text{ with } \Sigma = (\mathbb{E}ZZ^T)^{-1}(\mathbb{E}\varepsilon^2 ZZ^T)(\mathbb{E}ZZ^T)^{-1}. \tag{3.5}$$

"Dividing" by the asymptotic standard deviation yields an asymptotically pivotal statistic

$$\widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}(0, I)$$

where $\widehat{\Sigma}^{-1/2}$ is the unique matrix satisfying $\widehat{\Sigma}^{-1/2}\widehat{\Sigma}^{-1/2} = \widehat{\Sigma}^{-1}$ and

$$\widehat{\Sigma} = \left(\overline{ZZ^T}\right)^{-1}\left(\overline{ZZ^T\widehat{\varepsilon}^2}\right)\left(\overline{ZZ^T}\right)^{-1} \text{ with } \widehat{\varepsilon}_i = Y_i - Z_i^T\widehat{\beta}\,.$$

Let us focus first on bootstrapping $\sqrt{n}(\widehat{\beta}-\beta_0)$. So, let $\{Y_i^*, Z_i^*\}_{i=1}^n$ be iid extractions with replacement from the original sample $\{Y_i, Z_i\}_{i=1}^n$. The bootstrap version of the statistic writes as

$$\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$$

where $\widehat{\beta}^*$ has the same expression as $\widehat{\beta}$ but is based on the bootstrapped data $\{Y_i^*, X_i^*\}_{i=1}^n$. To show the validity of the bootstrap statistic notice first that

$$\beta_0 = \phi\left(\mathbb{E}X\right)\,,\ \widehat{\beta} = \phi\left(\overline{X}\right)\ \text{, and } \widehat{\beta}^* = \phi\left(\overline{X^*}\right)$$

where $X = (ZZ^T, ZY)$ and the function ϕ is differentiable at $\theta_0 := \mathbb{E}Z$. If X has finite second moments, then an application of the classical CLT yields $\sqrt{n}(\overline{X} - \mathbb{E}X) \rightsquigarrow \mathcal{N}(0, Var(X))$ while an application of the bootstrap CLT (i.e. Proposition 3.2.1) yields $\sqrt{n}(\overline{X^*} - \overline{X}) \rightsquigarrow \mathcal{N}(0, Var(X))$ in probability. Since ϕ is continuously differentiable at $\mathbb{E}X$, the *Bootstrap Delta Method* (i.e. Proposition 2.5.1) and the previous display yield

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow \phi_{\theta_0}T \text{ and } \boxed{\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \rightsquigarrow \phi_{\theta_0}T} \text{ in probability}$$

where $T = \mathcal{N}(0, Var(X))$. By Equation (3.5), the above display, and the principle of the uniqueness of the limit the distribution $\phi_{\theta_0}T$ must correspond to a $\mathcal{N}(0, \Sigma)$.
The above display shows the validity of the bootstrap for a statistic that is not asymptotically pivotal. We now prove the validity of the bootstrap for the asymptotically pivotal statistic $\widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta_0)$. Its bootstrap version is $\widehat{\Sigma}^{*\,-1/2}\sqrt{n}(\widehat{\beta^*} - \widehat{\beta})$. Notice that

$$\widehat{\Sigma}^{-1/2} = g\left(\overline{U}\right) \text{ and } (\widehat{\Sigma}^*)^{-1/2} = g\left(\overline{U^*}\right)$$

where $U := (ZZ^T, ZZ^TY^2, ZZ^T \otimes YZ, ZY, ZZ^T \otimes ZZ^T)$, since $\widehat{\varepsilon}_i^2 = Y_i^2 - 2Y_iZ_i^T + \widehat{\beta}^T Z_i Z_i^T \widehat{\beta}$. The function g is continuous at $\mathbb{E}U$. Since $\overline{U} = \mathbb{E}U + o_P(1)$, by the Continuous Mapping Theorem (Proposition A.0.8) $g(\overline{U}) = g(\mathbb{E}U) + o_P(1)$. So, using again the Continuous Mapping Theorem and Proposition A.0.7 we obtain that

$$g(\overline{U})\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \rightsquigarrow g(\mathbb{E}U)\,\phi_{\theta_0}T\,.$$

To obtain that the bootstrap is valid for the asymptotically pivotal statistics, it suffices to show that $\overline{U^*} = \overline{U} + o_{P^*}(1)$ and then apply Proposition 2.5.2. To this end, assume that U has finite second moments and fix an arbitrary $\eta > 0$. Assume without loss of generality that U is unidimensional. Then, since $E_n^*\overline{U^*} = \overline{U}$ (from Proposition 3.1.1), by

$$\beta^* = \overline{Z^* Z^{*T}} \; \overline{Z^* y^*}$$

$$\sqrt{n}(\hat{\beta}^* - \hat{\beta}) \rightsquigarrow N(0, \Sigma).$$ Bootstrap CLT + Delta method.

In terms of cdf:

$$\| \hat{F}_{T^*} - \Phi_\Sigma \|_\infty \xrightarrow{P} 0. \qquad (T^* = \sqrt{n}(\hat{\beta}^* - \hat{\beta}))$$

Thus, $\| \hat{F}_{T^*} - F_{T_n} \|_\infty \leq \| \hat{F}_{T^*} - \Phi_\Sigma \|_\infty + \| F_{T_n} - \Phi_\Sigma \|_\infty$

$(T_n = \sqrt{n}(\hat{\beta} - \beta_0))$ $\qquad \to 0$ by $\qquad \to 0$ by CLT

Confidence Interval of β_0:

$$\left[\hat{\beta} - \frac{\hat{\xi}_{1-\alpha}}{\sqrt{n}}, \; \hat{\beta} + \frac{\hat{\xi}_{1-\alpha}}{\sqrt{n}} \right]$$

$\hat{\xi}_{1-\alpha}$: $(1-\alpha)$ quantile of $|\hat{F}_{T_n^*}|$

Chebyshev's inequality we obtain

$$\begin{aligned}
P_n^*\left(|\overline{U^*}-\overline{U}|>\eta\right) \leq & \frac{1}{\eta^2} Var^*(\overline{U^*}) \\
= & \frac{1}{\eta^2\, n^2} Var^*(\sum_{i=1}^n U_i^*) \\
= & \frac{1}{\eta^2\, n} Var^*(U_i^*) \\
= & \frac{1}{\eta^2\, n}\left[E^*((U_i^*)^2)-(E_n^* U_i^*)^2\right] \\
= & \frac{1}{\eta^2\, n}\left[\overline{U^2}-(\overline{U}^2)\right] \\
= & O_P(n^{-1})\,.
\end{aligned}$$

So, the pairwise bootstrap is consistent for the statistic $(\widehat{\Sigma}^*)^{-1/2}\sqrt{n}(\widehat{\beta}^*-\widehat{\beta})$. By theorem 2.3.1, this implies that the bootstrap confidence intervals based on the pairwise scheme are consistent.

Bootstrapping 2SLS Estimators. Assume that the model is

$$Y = Z^T\beta_0 + \varepsilon \text{ with } \mathbb{E}W\varepsilon = 0$$

where Y is a scalar, Z, W are two vectors and ε is unobserved. The coefficient β_0 is identified as

$$\beta_0 = \left[(\mathbb{E}ZW^T)(\mathbb{E}WW^T)^{-1}(\mathbb{E}WZ^T)\right]^{-1}(\mathbb{E}ZW^T)(\mathbb{E}WW^T)^{-1}(\mathbb{E}WY)$$

and the 2SLS estimator is

$$\beta_0 = \left[(\overline{ZW^T})(\overline{WW^T})^{-1}(\overline{WZ^T})\right]^{-1}(\overline{ZW^T})(\overline{WW^T})^{-1}(\overline{WY})\,.$$

Since both β_0 and $\widehat{\beta}$ are smooth functions of sample means, the validity of the bootstrap for the 2SLS estimator is obtained in exactly the same way as for the OLS case.

3.4 Inference for a Non-Parametric Density (optional)

There are several situations where an estimator cannot be expressed as a transformation of a sample mean. In such cases, the consistency of the pairwise bootstrap cannot be proved as done in the previous section for OLS and 2SLS estimators. The nonparametric estimator of a density is one of such cases. In this section we show in details the validity of the pairwise bootstrap for a kernel estimator of a nonparametric density. A clear and comprehensive treatment of nonparametric estimation is provided in ???. We refer the

x is a fixed point

kernel density estimator: $\hat{f}(x) = \frac{1}{n \cdot h} \cdot \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$ $\{x_i\}_{i=1}^{n}$

(1) Draw with replacement "n" observations from $\{x_i\}_{i=1}^{n}$, noted as $\{x_i^*\}_{i=1}^{n}$

(2) We compute the statistic $\hat{A}^* = \frac{\sqrt{n \cdot h}\,(\hat{f}^*(x) - \hat{f}(x))}{\hat{s}^*(x)}$

(3) Repeat steps (1) & (2) B times $\{A_b^*,\ b = 1, \cdots, B\}$.

CI for true density $f(x)$:

$$\left[\hat{f}(x) - \frac{\hat{s}(x)}{\sqrt{n \cdot h}} \cdot \hat{q}_{\frac{\alpha}{2}},\ \hat{f}(x) + \frac{\hat{s}(x)}{\sqrt{n \cdot h}} \cdot \hat{q}_{1-\frac{\alpha}{2}}\right]$$

where $\hat{q}$ is the empirical quantile of $\hat{A}^*$

Attention:

it is confidence interval for a single point x.

reader to such book for a deeper understanding of these techniques.
Assume to observe an iid sample $\{ Z_i\}_{i=1}^n$ of continuously distributed random variables. We focus on the case of real-valued random variables for the ease of exposition. The multivariate case can be treated in exactly the same way, at the cost of a heavier notation. Let us denote with f the density of each observation Z_i. Assume we want to to estimate such density nonparametrically, i.e. without making any parametric assumption on it. A popular method employed is the kernel method. So, we define the estimator

$$\widehat{f}(z) := \frac{1}{n\,h} \sum_{i=1}^{n} K\left(\frac{Z_i - z}{h}\right)$$

where $K() : \mathbb{R} \mapsto \mathbb{R}$ is a continuous function that is symmetric about zero and satisfies $\int K(u)\,du = 1$, $\int u^j\,K(u)\,du = 0$ for $j = 1, .., r-1$, and $\int |u^r\,K(u)|du < \infty$. K is called kernel function of order r. the parameter h is shrinking to zero as the sample size increases, and is called *bandwidth*. If f is r times continuously differentiable, by classical nonparametric techniques we can show that

$$\sqrt{nh}(\widehat{f}(z) - f(z) - h^r B(z)) \rightsquigarrow \mathcal{N}\left(0, f(z) \int K(u)^2\,du\right)$$
$$\text{with } B(z) := \frac{1}{r!}\partial^r f(z) \int u^r K(u) du$$

$B(z)$ is a bias term. Assuming $nh^{2r+1} = o(1)$, such bias term can be ignored and the above convergence result can be employed to construct confidence intervals for $f(z)$.
As highlighted several times, our interest is always in bootstrapping statistics that are asymptotically pivotal. To pivotize the above statistic we can divide by the square root of

$$\boxed{\widehat{s}(z)^2 := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{Z_i - z}{h}\right)^2 - h\left[\widehat{f}(z)\right]^2}.$$

we can show that

$$\widehat{s}(z)^2 = f(z) \int K(u)^2\,du + o_P(1)$$

so that

$$\frac{\sqrt{nh}(\widehat{f}(z) - f(z))}{\widehat{s}(z)} \rightsquigarrow \mathcal{N}(0,1)\,.$$

In what follows we will prove the validity of the pairwise bootstrap for the above statistic. This will allow us to construct confidence intervals for $f(z)$.
The techniques we employ here can also be used to construct confidence intervals about a nonparametric regression, say $\mathbb{E}\{Y|Z = z\}$. The prove in such a case are obtained along the same lines as for the nonparametric density estimator.
Let $\{Z_i^*\}_{i=1}^n$ be an iid sample extracted with replacement from the original sample $\{Z_i\}_{i=1}^n$.

We define the bootstrap counterpart of the above statistic as

$$\sqrt{nh}\frac{(\widehat{f}^*(z)-\widehat{f}(z))}{\widehat{s}^*(z)} \text{ with } \widehat{f}^*(z) := \frac{1}{nh}\sum_{i=1}^n K\left(\frac{Z_i^*-z}{nh}\right)$$
$$\widehat{s}^*(z)^2 := \frac{1}{nh}\sum_{i=1}^n K\left(\frac{Z_i^*-z}{nh}\right)^2 - h\left[\widehat{f}^*(z)\right]^2 .$$

Proposition 3.4.1. *Assume that K is a kernel of order r symmetric about zero, f is r times continuously differentiable, $n\,h \to \infty$ and $n\,h^{2r+1} = o(1)$. Then,*

$$\sqrt{nh}\frac{(\widehat{f}^*(z)-\widehat{f}(z))}{\widehat{s}^*(z)} \rightsquigarrow \mathcal{N}(0,1) \text{ in probability}.$$

Proof. First notice that by Proposition 3.1.1 $E_n^*\widehat{f}^*(z) = \widehat{f}(z)$. So, using the expression of $\widehat{f}^*(z)$ we get

$$\begin{aligned}\sqrt{nh}(\widehat{f}^*(z)-\widehat{f}(z)) &= \sum_{i=1}^n \left[\frac{1}{\sqrt{nh}}K\left(\frac{Z_i^*-z}{nh}\right) - E_n^*\frac{1}{\sqrt{nh}}K\left(\frac{Z_i^*-z}{h}\right)\right] \\ &=: \sum_{i=1}^n \left[Y_{n,i}^* - E_n^* Y_{n,i}^*\right]\end{aligned}$$

where we have defined $Y_{n,i}^* := K((Z_i^*-z)/h)/\sqrt{nh}$. We now prove the the conditions of the Lindeberg-Feller CLT.
For the first condition, let $(\delta_n)_n$ be a sequence of numbers converging to zero such that $nh\delta_n^2 \to \infty$. Then,

$$\begin{aligned}\sum_{i=1}^n E_n^*|Y_{n,i}^*|^2\mathbb{I}(|Y_{n,i}^*|>\delta_n) &= nE_n^*|Y_{n,i}^*|^2\mathbb{I}(|Y_{n,i}^*|>\delta_n) \\ &= \frac{1}{h}E_n^* K\left(\frac{Z_i^*-z}{h}\right)^2 \mathbb{I}\left(K\left(\frac{Z_i^*-z}{nh}\right)^2 > nh\delta_n^2\right) \\ &= \frac{1}{nh}\sum_{i=1}^n K\left(\frac{Z_i-z}{h}\right)^2 \mathbb{I}\left(K\left(\frac{Z_i-z}{nh}\right)^2 > nh\delta_n^2\right) \\ &=: A_n \end{aligned} \tag{3.6}$$

where in the last equality we have used again Proposition 3.1.1. Notice first that (see below)

$$\begin{aligned}\mathbb{E}A_n &= \frac{1}{h}\int K\left(\frac{z'-z}{h}\right)^2 \mathbb{I}\left(K\left(\frac{z'-z}{h}\right)^2 > nh\delta_n^2\right) f(z')\,dz' \\ &= \int K(u)^2\mathbb{I}(K(u)^2 > nh\delta_n^2)\, f(z+uh)\,du = o(1)\,.\end{aligned}$$

For the first equality, we have used the change of variable $(z'-z)/h = u$, so that $dz' = h\,du$. To justify the second equality, notice that since $nh\delta_n^2 \to \infty$, for each fixed u the argument of the integral converges to 0. Such argument is upperbounded by a constant$\times K(u)^2$ that is integrable. So, by the Lebesgue Dominated Convergence Theorem (see Proposition A.0.10), we get that the integral is $o(1)$. Also (see the comments below)

$$\begin{aligned}
Var(A_n) =& \frac{1}{nh^2} Var\left\{K\left(\frac{Z_i - z}{h}\right)^2 \mathbb{I}\left(K\left(\frac{Z_i - z}{nh}\right)^2 > nh\delta_n^2\right)\right\} \\
\leq& \frac{1}{nh^2}\mathbb{E}\left\{K\left(\frac{Z_i - z}{h}\right)^4\right\} \\
=& \frac{1}{nh}\int K(u)^4 f(z+uh)\,du \\
=& O\left(\frac{1}{nh}\right).
\end{aligned}$$

The first equality is obtained by the iid assumption. The second equality is immediate. The third one is obtained by a change of variable, similarly to what done for $\mathbb{E}A_n$. Finally, the last equality follows from the fact that f is bounded and $\int K(u)^4 du < \infty$.
The above display implies that $A_n = \mathbb{E}A_n + o_P(1)$[1]. Since $\mathbb{E}A_n = o(1)$ we obtain that $A_n = o_P(1)$. So, by (3.6) we get that the first condition of the Lindeberg-Feller CLT holds, i.e.

$$\sum_{i=1}^{n} E_n^*|Y_{n,i}^*|^2 \mathbb{I}(|Y_{n,i}^*| > \delta_n) = o_P(1)\,. \tag{3.7}$$

For the second condition,

$$\begin{aligned}
\sum_{i=1}^{n} Var^*(Y_{n,i}^*) =& n\,Var^*\left\{\frac{1}{\sqrt{nh}}K\left(\frac{Z_i^* - z}{h}\right)\right\} \\
=& \frac{1}{h}Var^*\left\{K\left(\frac{Z_i^* - z}{h}\right)\right\} \\
=& \frac{1}{h}\left[E_n^* K\left(\frac{Z_i^* - z}{h}\right)^2 - \left\{E_n^* K\left(\frac{Z_i^* - z}{h}\right)\right\}^2\right] \\
=& \frac{1}{h}\left[\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h}\right)^2 - h^2\left\{\widehat{f}(z)\right\}^2\right] \\
=& \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h}\right)^2 - h\left\{\widehat{f}(z)\right\}^2.
\end{aligned} \tag{3.8}$$

To obtain the first three equalities we have used the iid structure, the definition of $Y_{n,i}^*$ and the property of the variance. For the fourth equality, we have used Proposition 3.1.1.

[1] Indeed by Chebyshev's inequality we have that $P(|A_n - \mathbb{E}A_n| > \varepsilon) \leq \varepsilon^{-2} Var(A_n)$, so $Var(A_n) = o(1)$ ensures that $A_n = \mathbb{E}A_n + o_P(1)$.

Next, we show that the RHS of (3.8) converges in probability to $f(Z)\int K(u)^2du$. The expectation of the first term is given by

$$\frac{1}{h}\int K\left(\frac{z'-z}{h}\right)^2 f(z')dz' = \int K(u)^2 f(z+uh)du \; = f(z)\int K(u)^2du + o(1)\,.$$

For the first equality, we have used the change of variable $(z'-z)/h = u$, so that $dz' = hdu$. To justify the second equality, notice that for a fixed u the argument of the above integral converges to $K(u)^2 f(z)$. Since the argument can be bounded by a constant$\times K(u)^2$ and $\int K(u)^2du < \infty$, the Lebesgue Dominated Convergence Theorem yields the second equality. By proceeding as for $Var(A_n)$, we can show that the variance of the first term on the RHS of (3.8) goes to 0. Hence, we get that

$$\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{Z_i - z}{h}\right)^2 = f(z)\int K(u)^2du + o_P(1)\,.$$

Using a similar reasoning, we obtain $\widehat{f}(z) = f(z) + o_P(1)$. So, gathering the results we have proved the second condition of the Lindeberg-Feller CLT, i.e.

$$\sum_{i=1}^{n} Var^*(Y^*_{n,i}) = f(z)\int K(u)^2du + o_P(1)\,. \tag{3.9}$$

We now argue along subsequences to show that $\sum_{i=1}^{n}(Y^*_{n,i} - E^*_n Y^*_{n,i}) \rightsquigarrow \mathcal{N}(0, f(z)\int K(u)^2du)$ in probability. In particular, consider an arbitrary subsequence of (n). Then, from (3.7) and (3.9) it is possible to extract a further subsequence (m) such that $\sum_{i=1}^{m} E^*_m|Y^*_{m,i}|^2\mathbb{I}(|Y^*_{m,i}| > \delta_n) = o(1)$ and $\sum_{i=1}^{m} Var^*(Y^*_{m,i}) = f(z)\int K(u)^2du + o(1)$ almost surely. Hence, applying the Lindeberg feller CLT we get that

$$\sum_{i=1}^{m}(Y^*_{m,i} - E^*_m Y^*_{m,i}) \rightsquigarrow \mathcal{N}(0, f(z)\int K(u)^2du) \text{ almost surely}$$

i.e. the cdf of the element on the LHS converges uniformly almost surely to the cdf of the normal on the RHS. Since (m) is a subseuence of an arbitrary subsequence of (n), we have obtained that

$$\sum_{i=1}^{n}(Y^*_{n,i} - E^*_n Y^*_{n,i}) \rightsquigarrow \mathcal{N}(0, f(z)\int K(u)^2du) \text{ in probability.}$$

From Exercise ?? we get that $\widehat{s}^*(z)^2 = f(z)\int K(u)^2du + o_{P^*}(1)$, i.e.

$$P\left(|\widehat{s}^*(z)^2 - f(z)\int K(u)^2du| > \varepsilon\right) = o_P(1)$$

for any $\varepsilon > 0$. So, using the previous two displays and Proposition **??** we get the final

result.

□

3.5 Pairwise Bootstrap for M Estimators

M-estimation is a particular type of estimation procedure covering a wide range of estimators in econometrics and statistics. Assume to observe an iid sample $\{X_i\}_{i=1}^n$ of vectors in $\mathbb{R}^M$. Let Θ be a compact subset of $\mathbb{R}^s$ and $m : \Theta \times \mathbb{R}^M \mapsto \mathbb{R}$. Let θ_0 be defined as

$$\theta_0 := \arg\min_{\theta\in\Theta} \mathbb{E}m(\theta, X)\,.$$

An M estimator is defined as the minimizer of the empirical counterpart of the above expectation:

$$\widehat{\theta} := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n m(\theta, X_i)\,. \tag{3.10}$$

The name "M estimator" comes from the fact that such an estimator is based on the *mean* of $m(\theta, X)$. Examples of M-estimators are **maximum-likelihood estimators** and **Non-linear least-squares**. Under some regularity conditions $\boxed{\sqrt{n}(\widehat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, \Sigma)\,.}$ The bootstrap theory so far developed in this chapter cannot be applied to such type of estimators, as they cannot be expressed as functions of the sample mean. However, under some regularity conditions the pairwise bootstrap will still be valid in this context. In particular, define

$$\boxed{\widehat{\theta}^* := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n m(\theta, X_i^*)\,,}$$

$\hat{\theta}$ is the true parameter in Bootstrap world.

where $\{X_i^*\}_{i=1}^n$ is the bootstrap sample generated with the pairwise scheme. Then, $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ will consistently estimate the distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$. Hence, confidence intervals and tests of hypothesis about θ_0 can be constructed by the pairwise bootstrap scheme.

$|P_n^*(\hat{t}^* \le \bar{t}) - P(\hat{t} \le \bar{t})| \xrightarrow{\text{proba}} 0 \quad \forall \bar{t} \in \mathbb{R}^M$

Let us now provide some intuition of why the pairwise bootstrap is valid in this context. If we assume that m is a smooth function with respect to θ, $\widehat{\theta}^*$ must satisfy the first-order conditions

$$\frac{1}{n}\sum_{i=1}^n \partial_\theta m(\widehat{\theta}^*, X_i^*) = 0\,,$$

so a mean-value expansion of the LHS around $\widehat{\theta}$ gives

$$\underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \partial_\theta m(\widehat{\theta}^*, X_i^*)}_{=0} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \partial_\theta m(\widehat{\theta}, X_i^*) + \partial_\theta\partial_{\theta^T} m(X_i^*, \widetilde{\theta}^*)\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$$

where $\widetilde{\theta}^* \in [\widehat{\theta}^*, \widehat{\theta}]$. So, the asymptotic behavior of $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ will be given by the asymptotic behavior of the first term on the RHS. Such a term *has zero expectation in the bootstrap world*: its bootstrap expectation coincides the the first-order conditions of the minimization problem in (3.10) that are equal to zero. Thus, we can verify that such a term satisfies the conditions of the Liendeberg-feller CLT and obtain that $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ converges to the same normal distribution as $\sqrt{n}(\widehat{\theta} - \theta_0)$. This ensures the consistency of the bootstrap.

Of course, for the bootstrap to provide refinements over the asymptotic approximation, we need to consider an asymptotically pivotal statistic. So, if we have a consistent estimator of Σ, say $\widehat{\Sigma}$, we can consider $\widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\theta} - \theta_0)$ and its bootstrap counterpart $\widehat{\Sigma}^{*-1/2}\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$.

For a rigorous proof of the validity of the pairwise bootstrap for M estimators, we would need tools typical of Empirical Process Theory, and in particular the concepts of Covering/Bracketing numbers and Glivenko-Cantelli results. References are given in the final notes to this chapter.

3.6 Pairwise Bootstrap for GMM

In the context of the Generalized method of Moments (GMM), the true parameter $\theta_0 \in \Theta \subset \mathbb{R}^s$ is the unique solution to

$$\mathbb{E}m(X, \theta) = 0 \qquad \theta_0 = \arg\min_{\theta \in \Theta} \| \mathbb{E}m(X,\theta) \|^2_{\Psi} \tag{3.11}$$

where here $m(X, \theta) \in \mathbb{R}^K$ and $K > s$. So the above system has more equations than unknown. The GMM estimator therefore minimizes the distance between the sample counterpart of the above moments and 0:

$$\widehat{\theta} := \arg\min_{\theta \in \Theta} \overline{m(X, \theta)}^T \widehat{\Psi}\, \overline{m(X, \theta)} := \| \overline{m(X,\theta)} \|^2_{\widehat{\Psi}} \qquad \widehat{\Psi} \to \Psi_0$$

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \| \overline{m(X,\theta)} \|^2_{\widehat{\Psi}}$$

where Ψ is a *positive definite* matrix that might depend on the sample data. Under some regularity conditions it holds that $\sqrt{n}(\widehat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, \Sigma)$.

The bootstrap counterpart of $\widehat{\theta}$ is

$$\widehat{\theta}^* := \arg\min_{\theta \in \Theta} \overline{m(X^*, \theta)}^T \widehat{\Psi}^* \overline{m(X^*, \theta)} := \| \overline{m(X^*, \theta)} \|^2_{\widehat{\Psi}^*} \tag{3.12}$$

where $\{X_i^*\}_{i=1}^n$ is an iid sample obtained by the pairwise resampling scheme. By using the same arguments as for M estimators, it is possible to show that the bootstrap cdf of $\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ estimates consistently the distribution of $\sqrt{n}(\widehat{\theta} - \theta_0)$. The same holds for the asymptotically pivotal versions. So, the bootstrap cdf of the statistic $\widehat{\Sigma}^{*-1/2}\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ estimates consistently the cdf of $\widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\theta} - \theta_0)$.

However, some authors have remarked that also for the asymptotically pivotal case the bootstrap does not provide refinements over the asymptotic approximation. The reason is that while in the real world the true value of the parameter θ_0 satisfies the populations

$\mathbb{E}_n \overline{m(X^*, \hat{\theta})} = \overline{m(X, \hat{\theta})}$

moment conditions (3.11), in the bootstrap world the true value of the parameter $\widehat{\theta}$ does not satisfies the bootstrap moment conditions $\overline{m(X^*, \theta)} = 0$. Thus, the bootstrap does not fully mimic what happens in the real world. In other words, since the bootstrap in this case is missing some information it cannot provide significant gains over the asymptotic approximation. The solution to this is to recenter the moment conditions in the bootstrap minimization program. So, consider the moment conditions

$$\overline{m(X, \theta) - \overline{m(X, \widehat{\theta})}} = 0$$

that are satisfied for $\theta = \widehat{\theta}$. Instead of (3.12) we minimize the distance between 0 and the (bootstrap) sample counterpart of the above moment conditions:

$$\widehat{\theta}^* := \arg\min_{\theta \in \Theta} \overline{m(X^*, \theta) - \overline{m(X, \widehat{\theta})}}^T \widehat{\Psi}^* \overline{m(X^*, \theta) - \overline{m(X, \widehat{\theta})}}. \tag{3.13}$$

Then the bootstrap cdf of $\widehat{\Sigma}^{*-1/2}\sqrt{n}(\widehat{\theta}^* - \widehat{\theta})$ will estimate consistently the true cdf of $\widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\theta} - \theta)$ and the bootstrap will provide refinements over the asymptotic approximation.

3.7 When the Pairwise Bootstrap is Inconsistent

In the previous pages of this chapter we have focused on cases when the pairwise bootstrap works. In this section, we show some examples where the pairwise bootstrap does *not* estimate consistently the distribution of a statistic.

Estimation of a max. Let $\{X_i\}_{i=1}^n$ be an iid sample, with $X_i \sim U[0, \theta]$. Our interest in on estimating θ. A natural estimator is $\widehat{\theta}_n := \max\{X_1, .., X_n\}$. Let us consider the statistic as $n(\widehat{\theta}_n - \theta)$. Notice that if $x \geq 0$ then $P(n(\widehat{\theta}_n - \theta) \leq x) = 1$, as $\widehat{\theta}_n - \theta$ will always be negative. If $x < 0$, then for each n large enough

$$\begin{aligned} P\left(n(\widehat{\theta}_n - \theta) \leq x\right) &= P\left(\widehat{\theta}_n \leq \theta + \frac{x}{n}\right) = P\left(X_i \leq \theta + \frac{x}{n} \text{ for all } i = 1, .., n\right) \\ &= \left[P\left(X_1 \leq \theta + \frac{x}{n}\right)\right]^n = \left[\left(\frac{\theta + x/n}{\theta}\right)\right]^n \end{aligned}$$

where the last equality follows from the fact that $X_1 \sim U[0, \theta]$, $x < 0$, and $\theta + x/n \geq 0$ for sufficiently large n. By the above display, for a fixed $x < 0$, when $n \to \infty$ we obtain that $P(n(\widehat{\theta}_n - \theta) \leq x) = [1 + (x/\theta)/n]^n \to \exp(x/\theta)$, by definition of Nepero's constant. So, we find that $n(\widehat{\theta}_n - \theta) \rightsquigarrow T$, where $F_T(x) = \exp(x/\theta)$ if $x \leq 0$ and $F_T(x) = 1$ if $x \geq 0$. Now, the pairwise bootstrap counterpart of $n(\widehat{\theta}_n - \theta)$ is $n(\widehat{\theta}_n^* - \widehat{\theta}_n)$, with $\widehat{\theta}_n^* = \max\{X_1^*, .., X_n^*\}$

and X_i^*'s being iid from the sample $\{X_1, .., X_n\}$. Then, $P_n^*(n(\widehat{\theta}_n^* - \widehat{\theta}_n) \leq 0) = 1$ and

$$\begin{aligned} P_n^*\left(n(\widehat{\theta}_n^* - \widehat{\theta}_n) < 0\right) =& P_n^*\left(\widehat{\theta}_n^* < \widehat{\theta}_n\right) = P_n^*\left(X_i^* < \widehat{\theta}_n \text{ for all } i = 1, .., n\right) \\ &= \left[\mathbb{E}_n^* \mathbb{I}(X_1^* < \widehat{\theta}_n)\right]^n = \left[\frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i < \widehat{\theta}_n)\right]^n \\ &= \left[\frac{n-1}{n}\right]^n = \left[1 - \frac{1}{n}\right]^n = \exp(-1) + o(1)\,. \end{aligned}$$

Now, let us proceed by contradiction and assume that the pairwise bootstrap estimates consistently the distribution of $n(\widehat{\theta}_n - \theta)$. This implies that the bootstrap cdf of $n(\widehat{\theta}^* - \widehat{\theta}_n)$ converges uniformly in probability to F_T. Then, for every $x' \in (-\theta, 0)$ we have

$$P_n^*\left(n(\widehat{\theta}_n^* - \widehat{\theta}_n) < 0\right) \geq P_n^*\left(n(\widehat{\theta}_n^* - \widehat{\theta}_n) \leq x'\right) = F_T(x') + o_P(1)\,.$$

Gathering together the previous two displays gives that $o_P(1) \geq F_T(x') - \exp(-1) = \exp(x'/\theta) - \exp(-1) > 0$ (since $x' > -\theta$). However, this is an absurd, so the contradiction assumption cannot hold and the pairwise bootstrap is inconsistent.

Notes

The proof of the bootstrap CLT provided in the first part of this chapter follows van der Vaart (2000). The proof of the bootstrap consistency for M Estimators requires tools typical of Empirical process Theory, and in particular bracketing and covering numbers. A clear presentation of these concepts can be found in van der Vaart (2000). Such a text, van der Vaart and Wellner (1996), and Kosorok (2007) contain proofs for the validity of M estimators when the function m defined in Section 3.5 is not necessarily smooth.

Chapter 4

The Residual, the Wild, and the Weighted Bootstrap

The pairwise bootstrap can be applied to a variety of situations, but it does not exploit all the information suggested by the model. This chapter introduces other resampling schemes such as the residual bootstrap and the wild bootstrap that resample the observations by using the structure suggested by the model. The main advantage is that these resampling schemes provide refinements with respect to the pairwise bootstrap when the model is correctly specified. The organization goes as follows. In Section 4.1 we introduce the residual bootstrap. Section 4.2 sets up the wild bootstrap.

4.1 The Residual Bootstrap

We want to make inference on the parameters of the linear model with homoscedastic errors

$$Y = Z^T\beta_0 + \varepsilon \text{ where } \mathbb{E}Z\varepsilon = 0 \text{ and } \mathbb{E}\{\varepsilon^2|Z\} = \sigma^2. \tag{4.1}$$

important for Bootstrap to work.

Assume that $\mathbb{E}ZZ^T$ is a full-rank matrix. We observe an iid sample $\{Y_i, Z_i\}_{i=1}^n$ and we estimate β_0 by OLS, $\widehat{\beta} = (\overline{Z\,Z^T})^{-1}(\overline{ZY})$.

The idea at the basis of the residual bootstrap is keep the explanatory variables Z_i fixed and resample the residuals $\widehat{\varepsilon}_i = Y_i - Z_i^T\widehat{\beta}$ so as to construct artificial observations Y_i^* (for $i = 1, \ldots, n$). Then, the bootstrap sample $\{Y_i^*, Z_i\}_{i=1}^n$ is used to construct a bootstrap version $\widehat{\beta}^*$ of $\widehat{\beta}$ or a bootstrap version of the statistic of interest.

Let us define $\overline{\widehat{\varepsilon}} := (1/n)\sum_{i=1}^n \widehat{\varepsilon}_i$. The practical steps to implement the residual bootstrap are the following:

1. Resample with replacement n observations from $\{\widehat{\varepsilon}_i : i = 1, \ldots, n\}$ to get $\{\varepsilon_i^*\}_{i=1}^n$ and set $r_i^* = \varepsilon_i^* - \overline{\widehat{\varepsilon}}$

2. Set $Y_i^* = Z_i^T\widehat{\beta} + r_i^*$ for $i = 1, \ldots, n$ and obtain the bootstrap sample $\{Y_i^*, Z_i\}_{i=1}^n$

Residual Bootstrap:

$$Y_i^* = Z_i^T \hat{\beta} + r_i^* , \quad r_i^* = \varepsilon_i^* - \bar{\hat{\varepsilon}} , \quad \hat{\varepsilon} = Y - Z^T \hat{\beta}$$

$$\varepsilon_i^* \overset{iid}{\sim} P_{\hat{\varepsilon}} \quad \text{where} \quad P_{\hat{\varepsilon}}(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(\hat{\varepsilon}_i \in A) \quad \forall A \subset \mathbb{R}.$$

Interpretation:

$$P_{\hat{\varepsilon}}(\varepsilon^* = \hat{\varepsilon}_j) = P_{\hat{\varepsilon}}(\{\hat{\varepsilon}_j\}) = \frac{1}{n}$$

$$P_{\hat{\varepsilon}}(\varepsilon^* \neq \hat{\varepsilon}_j, \forall j) = 0$$

$\hat{\varepsilon}^*$ is drawn from $\{\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n\}$

we get bootstrap samples: $\{Y_i^*, Z_i\}$

$$\hat{\beta}^* = (\overline{ZZ^T})^{-1} \overline{Y^* Z}$$

$$\hat{t} = \sqrt{n}(\hat{\beta} - \beta_0) \qquad \hat{t}^{AP} = \hat{\Sigma}^{-\frac{1}{2}} \sqrt{n}(\hat{\beta} - \beta_0)$$

$$\hat{\Sigma} = (\overline{ZZ^T})^{-1} \hat{\sigma}^2 , \quad \hat{\sigma}^2 = \overline{\hat{\varepsilon}^2}$$

$$\hat{t}^* = \sqrt{n}(\hat{\beta}^* - \hat{\beta}) \qquad \hat{t}^{AP*} = (\hat{\Sigma}^*)^{-\frac{1}{2}} \sqrt{n}(\hat{\beta}^* - \hat{\beta})$$

$$\hat{\Sigma}^* = (\overline{ZZ^T})^{-1} (\hat{\sigma}^*)^2 , \quad (\hat{\sigma}^*)^2 = \overline{(\hat{\varepsilon}^*)^2}$$

MC for Residual Bootstrap:

1) Draw with remplacement n observations from $\{\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n\}$, noted as $(\hat{\varepsilon}_i^*)$

2) Compute (Y_i^*)

3) Compute $\hat{t}^*$ from (Y_i^*, Z_i)

4) Repeat 1) → 3) B times, get $(\hat{t}_b^*)_{b=1}^{B}$. Compute its quantiles to get confidence interval of β_0.

3. Use the bootstrap sample of the previous step to compute the bootstrap counterpart of the T_n statistic, e.g. $T_n^* := \sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$ or $T_n^* := \widehat{\Sigma}^{*\,-1/2}\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$ (below we provide more details about the structure of T_n^*)

4. Repeat the above steps B times, with B large, so as to get the collection $\{T_{n,b}^* : b = 1, \ldots, B\}$ and use such a collection to estimate the true distribution of the statistic T_n.

Essentially, the residual bootstrap applies the same idea as the pairwise bootstrap, but only to the **centered residuals** of the linear model. It uses the information of the model, in the sense that it exploits its linearity and its separability from the errors. Originally, the residual bootstrap was developed for a linear model like (4.1) where the explanatory variables are fixed and non-random. However, such a resampling scheme remains still valid when the Z's are random.

From a mathematical point of view, the resampling scheme is defined as

$$Y_i^* = Z_i^T\widehat{\beta} + r_i^* \text{ with } r_i^* = \varepsilon_i^* - \overline{\widehat{\varepsilon}} \text{ and } \varepsilon_i^* \sim \text{iid}\widehat{P}_{\widehat{\varepsilon}} \tag{4.2}$$

where $\widehat{P}_{\widehat{\varepsilon}}$ is the probability

$$\widehat{P}_{\widehat{\varepsilon}}(A) := \frac{1}{n}\sum_{i=1}^{n} 1\{\widehat{\varepsilon}_i \in A\}$$

for any set $A \subset \mathbb{R}$. Notice that the probability $\widehat{P}_{\widehat{\varepsilon}}$ has the same structure as the probability $\mathbb{P}_n$ of the previous chapter.

An important aspect to notice is that the linear model in the bootstrap world, i.e. in (4.2), is characterized by *the bootstrap version* of the moment conditions characterizing the linear model in the population, i.e. in (4.1). To see this, denote with E_n^* the expectation that considers as random only the residuals ε_i^* for $i = 1, \ldots, n$, with $\varepsilon_i^* \sim \text{iid}P_{\widehat{\varepsilon}}$. Then, the bootstrap versions of the population moments $\mathbb{E}Z\varepsilon$ are

$$E_n^* Z_i r_i^* = Z_i E_n^* r_i^* = Z_i \left(E_n^*\varepsilon^* - \overline{\widehat{\varepsilon}}\right) = 0\,,$$

where the last step follows from Proposition 3.1.1 of the previous chapter and the definition of $\widehat{P}_{\widehat{\varepsilon}}$. So, $\widehat{\beta}$ is the unique vector satisfying the equations

$$E_n^*\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i\left(Y_i^* - Z_i^T\beta\right)\right\} = 0\,.$$

This is exactly what happens in the population, as β_0 is the unique vector satisfying

$$\mathbb{E}Z(Y - Z^T\beta) = 0\,.$$

Hence, as well as β_0 is the "true parameter" at the population level $\widehat{\beta}$ is the "true parameter" at the bootstrap level.

Having established the main features of the residual bootstrap scheme, we now show its validity for constructing hypothesis tests and confidence intervals about β_0. Given the bootstrap sample $\{Y_i^*, Z_i\}_{i=1}^n$ we compute the bootstrap counterpart of $T_n := \sqrt{n}(\widehat{\beta} - \beta_0)$ as

$$T_n^* := \sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \text{ where } \widehat{\beta}^* := (\overline{Z\,Z^T})^{-1}\overline{Z\,Y^*}\,.$$

Notice that we are replacing (i) the estimator $\widehat{\beta}$ with its bootstrap counterpart $\widehat{\beta}^*$ and (ii) β_0 (which is the true parameter in the population) with $\widehat{\beta}$ (which is the true parameter in the "bootstrap world").
To compute the bootstrap version of the asymptotically pivotal statistic, recall that we are assuming homoscedastic errrors, $\mathbb{E}\{\varepsilon^2|Z\} = \sigma^2$. So,

$$T_n = \widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta_0) \text{ with } \widehat{\Sigma} = \widehat{\sigma}(\overline{Z\,Z^T})^{-1} \text{ and } \widehat{\sigma} = \overline{(Y - Z^T\widehat{\beta})^2}$$

and its bootstrap version is

$$T_n^* := \widehat{\Sigma}^{*\,-1/2}\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \text{ with } \widehat{\Sigma} := \widehat{\sigma}^*(\overline{Z\,Z^T}) \text{ and } \widehat{\sigma}^* = \overline{(Y^* - Z^T\widehat{\beta}^*)^2}$$

Proposition 4.1.1. *Assume that the error terms are homoscedastic (i.e. $\mathbb{E}\{\varepsilon^2|Z\} = \sigma^2$) and Z have finite third order moment. Then, the residual bootstrap is consistent, i.e. the bootstrap distribution of*

$$\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \text{ and } \widehat{\Sigma}^{*\,-1/2}\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$$

Asymptotically pivotal

estimate consistently the distribution of

distribution

distribution

$$\sqrt{n}(\widehat{\beta} - \beta_0) \text{ and } \widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta_0)\,,$$

respectively.

Proof. We first consider the non-asymptotically pivotal statistic. Then,

$$\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) = \left(\overline{ZZ^T}\right)^{-1}\sqrt{n}\,\overline{Zr^*} \text{ with } \overline{ZZ^T} = \mathbb{E}ZZ^T + o(1) \text{ almost surely.} \tag{4.3}$$

We now prove that $\sqrt{n}\overline{Zr^*}$ satisfies the conditions of the Liendeberg-Feller CLT. The procedure is similar to the proof of Proposition 3.2.1. So, consider

$$\sqrt{n}\,\overline{Zr^*} = \sum_{i=1}^{n} \frac{Z_i r_i^*}{\sqrt{n}} \text{ with } r_i^* = \varepsilon_i^* - \overline{\widehat{\varepsilon}}. \tag{4.4}$$

For the convergence of the covariances, recall that E_n^* is the expectation considering as random only the residuals $\{\varepsilon_i^*\}_{i=1}^n$ and we have already obtained that $E_n^* r_i^* Z_i = 0$. So,

$$\sum_{i=1}^{n} E_n^* \left\{ \left(\frac{Z_i r_i^*}{\sqrt{n}}\right) \left(\frac{Z_i r_i^*}{\sqrt{n}}\right)^T \right\} = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T E_n^* r_i^{*2} = E_n^* r_i^{*2} \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T$$

where the second equality uses the fact that the residuals r_i^* for $i = 1, \ldots, n$ are iid. Since $E_n^* \varepsilon_i^* = \overline{\widehat{\varepsilon}}$ (from Proposition 3.1.1), by the formula of the variance we have (see the comments below)

$$\begin{aligned}
E_n^* r_i^{*2} &= E_n^* \left(\varepsilon_i^* - E_n^* \varepsilon_i^*\right)^2 = E_n^* \varepsilon_i^{*2} - \left(E_n^* \varepsilon_i^*\right)^2 = \overline{\widehat{\varepsilon}^2} - \left(\overline{\widehat{\varepsilon}}\right)^2 \\
&= \overline{\left[\varepsilon - Z^T(\widehat{\beta} - \beta_0)\right]^2} - \left(\overline{\varepsilon} - \overline{Z^T}(\widehat{\beta} - \beta_0)\right)^2 \\
&= \overline{\varepsilon^2} - 2\overline{\varepsilon Z^T}(\widehat{\beta} - \beta_0) + (\widehat{\beta} - \beta_0)^T \overline{ZZ^T}(\widehat{\beta} - \beta_0) + o(1) \text{ almost surely} \\
&= \overline{\varepsilon^2} + o(1) \text{ almost surely} \\
&= \sigma^2 + o(1) \text{ almost surely}
\end{aligned}$$

where the third equality uses Proposition 3.1.1, the fourth equality uses $\widehat{\varepsilon} = \varepsilon - Z^T(\widehat{\beta} - \beta_0)$, and the remaning equalities use $\widehat{\beta} - \beta_0 = o(1)$ almost surely and the strong law of large numbers. Gathering together the previous two displays gives

$$\sum_{i=1}^{n} E_n^* \left\{ \left(\frac{Z_i r_i^*}{\sqrt{n}}\right) \left(\frac{Z_i r_i^*}{\sqrt{n}}\right)^T \right\} = \sigma^2 \mathbb{E} Z Z^T + o(1) \text{ almost surely }. \tag{4.5}$$

So, the first condition of the Liendeberg- Feller CLT is satisfied. To prove the second condition, let $(\eta_n)_n$ be a sequence of positive numbers converging to zero such that $\sqrt{n}\eta_n \to \infty$.

Then,

$$\begin{aligned}\sum_{i=1}^{n} E_n^* \left\| \frac{Z_i r_i^*}{\sqrt{n}} \right\|^2 1\left\{ \left\| \frac{Z_i r_i^*}{\sqrt{n}} \right\| > \eta_n \right\} =& \frac{1}{n} \sum_{i=1}^{n} E_n^* \left\| Z_i r_i^* \right\|^2 1\left\{ \left\| \frac{Z_i r_i^*}{\sqrt{n}\eta_n} \right\| > 1 \right\} \\ \leq& \frac{1}{\sqrt{n}\eta_n} \frac{1}{n} \sum_{i=1}^{n} E_n^* \left\| Z_i r_i^* \right\|^3 \\ \leq& \frac{1}{\sqrt{n}\eta_n} \frac{1}{n} \sum_{i=1}^{n} \left\| Z_i \right\|^3 \; E_n^* |r_i^*|^3 \\ =& \frac{1}{\sqrt{n}\eta_n} \overline{\|Z\|^3} \, E_n^* |r_i^*|^3 \, .\end{aligned}$$

Since Z has finite third moments, to prove that the RHS of the above display is $o(1)$ almost surely, it is sufficient to show that $E_n^*|r_i^*|^3$ is upperbounded by a constant almost surely. Now, using the fact that for any two positive numbers a, b and integer p we have $(a+b)^p \leq 2^p a^p + 2^p b^p$, it holds that $E_n^*|r_i^*|^3 = E_n^*|\varepsilon^* - \overline{\widehat{\varepsilon}}|^3 \leq C\, E_n^*|\varepsilon^*|^3 + C\, |\overline{\widehat{\varepsilon}}|^3 = C\, \overline{|\widehat{\varepsilon}|^3} + C\, |\overline{\widehat{\varepsilon}}|^3 = C\, \overline{|\widehat{\varepsilon}|^3} + o(1)$ almost surely, where C is a fixed constant. Also, $\overline{|\widehat{\varepsilon}|^3} = \overline{|\varepsilon - Z^T(\widehat{\beta} - \beta_0)|^3} \leq C\, \overline{|\varepsilon|^3} + C\, \overline{|Z^T(\widehat{\beta} - \beta_0)|^3} = C\, \mathbb{E}|\varepsilon|^3 + \overline{\|Z\|^3}\, \|\widehat{\beta} - \beta_0\|^3 \leq 2\, C\, \mathbb{E}|\varepsilon|^3$ almost surely. Thus, $E_n^* r_i^*$ is upperbounded by a fixed constant almost surely, and by the previous display we obtain that the second condition of the Lindeberg-Feller CLT holds with probability one.

Since the two conditions of the Linbderg-Feller CLT holds almost surely for (4.4), we get that

$$\sqrt{n}\, \overline{Z\, r^*} \rightsquigarrow \mathcal{N}(0, \sigma^2 \mathbb{E} Z Z^T) \text{ with probability one ,}$$

where the asymptotic covariance is given by (4.5). By the above display, (4.3), and by combining Propositions A.0.7 and A.0.8 we find that

$$\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \rightsquigarrow \mathcal{N}(0, \sigma^2 \left(\mathbb{E} Z Z^T\right)^{-1}) \text{ almost surely .}$$

Since $\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow \mathcal{N}(0, \sigma^2 (\mathbb{E} Z Z^T)^{-1})$, we have proved the consistency of the residual bootstrap for the non-asymptotically pivotal statistic.

For the case of the asymptotically pivotal statistic, the result readily follows by applying the previous display and Proposition 2.5.2 . □

4.1.1 The residual bootstrap with heteroskedasticity

The above result shows the validity of the residual bootstrap in the presence of homoscedasticity. When instead the errors are heteroscedastic (i.e. $\mathbb{E}\{\varepsilon^2|Z\}$ is not constant), the residual bootstrap scheme needs to be slightly modified. We can write without los of generality $\mathbb{E}\{\varepsilon^2|Z\} = v(Z)\sigma^2$, where $\sigma^2 = \mathbb{E}\varepsilon^2$, and assume to have a consistent estimator $\widehat{v}$ of v in the sense that $\max_{i=1,\dots,n} |\widehat{v}(Z_i) - v(Z_i)| = o_P(1)$. Then, the residual

bootstrap scheme must be modified as follows:

$$Y_i^* = Z_i^T \widehat{\beta} + \widehat{v}(Z_i) r_i^* \text{ where } r_i^* = \varepsilon_i^* - \overline{\widehat{\varepsilon}} \text{ and } \varepsilon_i^* \sim \widehat{P}_{\widehat{\varepsilon}}.$$

The bootstrap version of the non-asymptotically pivotal statistic $(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}))$ can be computed as in the previous section. For the bootstrapped asymptotically pivotal statistic $\widehat{\Sigma}^* \sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$ we have to consider the bootstrap version of the *White-Huber* estimator of the asymptotic covariance.
By following similar steps as in the proof of Proposition 4.1.1, it is possible to show that the residual bootstrap remains valid in the presence of heteroskedasticity as long as $\max_{i=1,\dots,n} |\widehat{v}(Z_i) - v(Z_i)| = o_P(1)$. So, the validity of the residual bootstrap requires to estimate consistently the heteroskedasticity structure. This can be done easily if we assume that $\mathbb{E}\{\varepsilon^2|Z\}$ has a known parametric form (i.e. if we know the functional form of v). If we do not know the functional form of v we can employ nonparametric estimators to obtain a consistent estimate of v.

4.2 The Wild Bootstrap

The idea of the *wild bootstrap* is to resample the residuals of a regression model by using a random source *external* to the model. From a mathematical point of view, the resampling scheme is

$$Y_i^* = Z_i^T \widehat{\beta} + \xi_i \widehat{\varepsilon}_i \tag{4.6}$$

where $\{\xi_i\}_{i=1}^n$ are iid according to a **known** user-chosen probability P_ξ, with $\mathbb{E}\xi = 0$ and $\mathbb{E}\xi^2 = 1$.

The wild bootstrap scheme is implemented through the following steps

1. Resample n iid observations from P_ξ and get $\{\xi_i\}_{i=1}^n$
2. Construct the artificial observations $Y_i^* := Z_i^T \widehat{\beta} + \xi_i \widehat{\varepsilon}_i$ for $i = 1, \dots, n$
3. Use the bootstrap sample $\{Y_i^*, Z_i\}_{i=1}^n$ to compute the bootstrap counterpart of the T_n statistic, i.e. $T_n^* = \sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$ or $T_n^* = \widehat{\Sigma}^{*-1/2}\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$ (see below for the details)
4. Repeat the above steps B times so as to get the collection$\{T_{n,b}^* : b = 1, \dots, B\}$ and use such a collection to estimate the distribution of T_n.

In the bootstrap world, the only random elements are the weights $\{\xi_i\}_{i=1}^n$ and the sample data is given. The name "wild bootstrap" comes from the fact that the source of randomness ξ comes from *outside* the model: the bootstrap weights $\{\xi_i\}_{i=1}^n$ are resampled from

a probability P_ξ that is user-chosen and is not related to the model.

Similarly to the residual bootstrap, the linear model in (4.6) (i.e. in the bootstrap world) is characterized by the bootstrap counterpart of the moment conditions characterizing the linear model in (4.1) (i.e. in the population). To see this, let E^*_ξ be the expectation that considers as random only the bootstrap weights $\{\xi_i\}_{i=1}^n$ and as given the sample data. Then, $E^*_\xi Z_i\,(\xi_i\widehat{\varepsilon}_i) = 0$ so that $\widehat{\beta}$ is the unique vector satisfying

$$E^*_\xi \frac{1}{n}\sum_{i=1}^n Z_i(Y_i^* - Z_i^T\beta) = 0\,.$$

These equations are the bootstrap counterpart of the moment conditions

$$\mathbb{E}Z(Y - Z^T\beta) = 0$$

that are satisfied only for $\beta = \beta_0$. So, as well as β_0 is the true parameter at the population level, $\widehat{\beta}$ is the "true" parameter at the bootstrap level.

For each bootstrap sample $\{Y_i^*, Z_i\}_{i=1}^n$, the bootstrap counterpart of the statistic $T_n = \sqrt{n}(\widehat{\beta} - \beta_0)$ is computed as

$$T_n^* := \sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \text{ where } \widehat{\beta}^* := (\overline{ZZ^T})^{-1}\overline{ZY^*}$$

and $\widehat{\beta}$ is the true parameter in the bootstrap world. For the asymptotically pivotal statistic $T_n = \widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta_0)$, the bootstrap counterpart is

$$T_n^* = \widehat{\Sigma}^{*\,-1/2}\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$$

where

$$\widehat{\Sigma}^{*\,-1/2} = (\overline{ZZ^T})^{-1}\overline{(Y^* - Z^T\widehat{\beta}^*)^2 ZZ^T}\,(\overline{ZZ^T})^{-1}\,.$$

Differently from the residual bootstrap, the wild bootstrap remains valid under heteroskedasticity of the error terms without any further modification. So, for the wild bootstrap to be consistent we do not have to estimate the structure of the heteroskedasticity.

To prove the validity of the wild bootstrap we need the following central limit theorem known as *Conditional Multiplier Central Limit Theorem.*

Proposition 4.2.1. *Let $\{X_i\}_{i=1}^n$ be a sequence of iid random vectors in $\mathbb{R}^K$ with $\mathbb{E}\|X\|^2 < \infty$. If $\{\xi_i\}_{i=1}^n$ is a sequence of iid random variables independent from $\{X_i\}_{i=1}^n$ with $E^*_\xi\xi = 0$*

and $E^*_\xi \xi^2 = 1$*, then for* P*-almost all sequences* $\{X_i\}_{i=1}^n$ *we have*

$$\sqrt{n}\,\overline{\xi X} \rightsquigarrow \mathcal{N}(0, \mathbb{E}XX^T)$$

The result of the above proposition says that

$$\sup_{s\in\mathbb{R}^K} \left| P_\xi(\sqrt{n}\,\overline{\xi X} \le s) - \Phi_{\mathbb{E}XX^T}(s) \right| = o(1) \text{ almost surely}$$

where P_ξ considers as random only the bootstrap weights $\{\xi_i\}_{i=1}^n$ and as fixed the data $\{X_i\}_{i=1}^n$, and $\Phi_{\mathbb{E}XX^T}$ is the cdf of $\mathcal{N}(0, \mathbb{E}XX^T)$.

Proof. As usual, we need to verify the conditions of the Liendeberg-Feller CLT for the sum $\sqrt{n}\,\overline{\xi X}$, where only the bootstrap weights $\{\xi_i\}_{i=1}^n$ are random and the data $\{X_i\}_{i=1}^n$ is fixed. We consider

$$\sqrt{n}\,\overline{\xi X} = \sum_{i=1}^n \frac{\xi_i X_i}{\sqrt{n}}\,.$$

Since $E^*_\xi \xi = 0$, $E^*_\xi \xi^2 = 1$, and $\{\xi_i\}_{i=1}^n$ is independent from $\{X_i\}_{i=1}^n$, we have

$$\sum_{i=1}^n E^*_\xi \left\{ \left(\frac{\xi_i X_i}{\sqrt{n}}\right) \left(\frac{\xi_i X_i}{\sqrt{n}}\right)^T \right\} = \overline{XX^T} = \mathbb{E}XX^T + o(1) \text{ almost surely }\,.$$

This gives the first condition of the Liendeberg-Feller CLT. For the second condition of the Liendeberg-feller CLT, let $(\eta_n)_n$ and $(\tau_n)_n$ be two sequences of positive numbers decreasing to 0 such that $\eta_n/\tau_n \to \infty$ and $\tau_n\sqrt{n} \to \infty$. Then, using $1 = 1\{|\xi| \le \eta_n/\tau_n\} + 1\{|\xi_i| > \eta_n/\tau_n\}$ leads to

$$\begin{aligned}\sum_{i=1}^n E^*_\xi \left\| \frac{\xi_i X_i}{\sqrt{n}} \right\|^2 1\left\{ \left\| \frac{\xi_i X_i}{\sqrt{n}} \right\| > \eta_n \right\} \le & \overline{\|X\|^2}\; E^*_\xi |\xi|^2\, 1\left\{ |\xi| > \frac{\eta_n}{\tau_n} \right\} \\ & + \frac{1}{n}\sum_{i=1}^n \|X_i\|^2\, 1\{\|X_i\| > \tau_n\sqrt{n}\}\,.\end{aligned}$$

Since X has finite second moments $\overline{\|X\|^2} = \mathbb{E}\|X\|^2 + o(1)$ almost surely. Also, since ξ has finite second moments, by the Lebesgue Dominated Convergence Theorem (see Proposition A.0.10) the expectation in the first term on the RHS is $o(1)$. By arguments similar to the last part of the proof of Proposition 3.2.1, we get that the second term is $o(1)$ almost surely. So, the sum on the LHS of the previous display is $o(1)$ almost surely.
Finally, we obtain that the two conditions of the Liendberg-feller CLT hold with probability 1. This delivers the desired result.

□

We are now ready to prove the validity of the wild bootstrap for the linear model.

Proposition 4.2.2. *If* $\{\xi_i\}_{i=1}^n$ *is a sequence of iid bootstrap weights independent from the*

sample data $\{Y_i, Z_i\}_{i=1}^n$ *with* $E_\xi^* \xi = 0$ *and* $E_\xi^* \xi^2 = 1$ *then the wild bootstrap is consistent, i.e. the bootstrap distribution of*

$$\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \text{ and } \widehat{\Sigma}^{*\,-1/2}\sqrt{n}(\widehat{\beta}^* - \widehat{\beta})$$

estimates consistently the distribution of

$$\sqrt{n}(\widehat{\beta} - \beta_0) \text{ and } \widehat{\Sigma}^{-1/2}\sqrt{n}(\widehat{\beta} - \beta_0)$$

respectively.

Proof. We will first obtain an approximation on the *joint* probability space, i.e. by considering as random *both* the sample data and the bootstrap weights. To this end, let us denote with Q the probability that considers as random *both* the sample data and the bootstrap weights. Now,

$$\begin{aligned}\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) =& (\overline{ZZ^T})^{-1}\sqrt{n}\,\overline{Z\xi\widehat{\varepsilon}} \\ =& (\overline{ZZ^T})^{-1}\sqrt{n}\,\overline{Z\xi\varepsilon} - (\overline{ZZ^T})^{-1}\,\overline{\xi ZZ^T}\,\sqrt{n}\,(\widehat{\beta} - \beta_0)\,.\end{aligned}$$

Let us show that the second term on the RHS is negligible. Since both $\sqrt{n}(\widehat{\beta} - \beta_0)$ and $\overline{ZZ^T}$ are bounded in probability, it is sufficient to show that the $\overline{\xi ZZ^T}$is negligible. Since Z has finite second moments, an application of the Law of Large Numbers gives $\overline{\xi ZZ^T} = \mathbb{E}_Q \xi ZZ^T + o_Q(1)$, where we consider as random both $\{\xi_i\}_{i=1}^n$ and $\{Y_i, Z_i\}_{i=1}^n$. Since $\{\xi_i\}_{i=1}^n$ have zero expectations and are independent from $\{Y_i, Z_i\}_{i=1}^n$, $\mathbb{E}_Q \xi ZZ^T = 0$. For the first term on the RHS of the previous display, notice that $\sqrt{n}\,\overline{Z\xi\varepsilon}$ satisfies a CLT (when both the sample data and the bootstrap weights are random), so it is bounded in probability. Also, $\overline{ZZ^T} = \mathbb{E}ZZ^T + o_P(1)$. Hence, the first term on the RHS of the above display is equal to $\left[\mathbb{E}ZZ^T\right]^{-1}\sqrt{n}\,\overline{Z\xi} + o_Q(1)$. Gathering together these results leads to

$$\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) = \underbrace{\left[\mathbb{E}ZZ^T\right]^{-1}\sqrt{n}\,\overline{Z\xi\varepsilon}}_{=:G_n} + \Delta_n \text{ where } \Delta_n = o_Q(1)$$

over the joint probability space, i.e. by considering a random both the bootstrap weights and the sample data. In view of this approximation, for any $\delta > 0$

$$\begin{aligned}P_\xi\left(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \le s\right) =& P_\xi(G_n \le x - \Delta_n) \\ \le& P_\xi(G_n \le x + \delta) + P_\xi(|\Delta_n| > \delta) \\ =& \Phi_\Psi(s + \delta) + P_\xi(|\Delta_n| > \delta) + o(1) \text{ almost surely }\,,\end{aligned}$$

where the last line follows from Proposition 4.2.1 and Φ_Ψ is the cdf of the $\mathcal{N}(0, \Psi)$ with

$\Psi := \mathbb{E}(ZZ^T)^{-1}(\mathbb{E}\varepsilon^2 ZZ^T)(\mathbb{E}ZZ^T)^{-1}$. Now, for any $\eta > 0$

$$\begin{aligned} P\left(P_\xi(|\Delta_n| > \delta) > \eta\right) \leq & \eta^{-1}\mathbb{E}P_\xi(|\Delta_n| > \delta) = \eta^{-1}\mathbb{E}E^*_\xi 1\{|\Delta_n| > \delta\} \\ = & \eta^{-1}Q(|\Delta_n| > \eta) = o(1)\,. \end{aligned}$$

So,

$$P_\xi(|\Delta_n| > \delta) = o_P(1)\,.$$

Also, since $G_n = \sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) - \Delta_n$ we get

$$\begin{aligned} P_\xi\left(G_n \leq s - \delta\right) = & P_\xi\left(\sqrt{n}(\widehat{\beta}^* - \beta_0) \leq s - \delta + \Delta_n\right) \\ \leq & P_\xi\left(\sqrt{n}(\widehat{\beta}^* - \beta_0) \leq s\right) + P_\xi(|\Delta_n| > \delta) \end{aligned}$$

Using again Proposition 4.2.1 the LHS of the above display is equal to $\Phi_\Psi(s-\delta) + o(1)$ almost surely. By putting together the previous 4 displays we obtain

$$\Phi_\Psi(s-\delta) + o_P(1) \leq P_\xi\left(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \leq s\right) \leq \Phi_\Psi(s+\delta) + o_P(1)\,.$$

Since δ was arbitrarily chosen and Φ_Ψ is a continuous cdf, we can set δ so as to make $\Phi_\Psi(s-\delta)$ and $\Phi_\Psi(s+\delta)$ as close as possible to $\Phi_\Psi(s)$. Hence,

$$P_\xi\left(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \leq s\right) = \Phi_\Psi(s) + o_P(1)$$

for any fixed $s \in \mathbb{R}^K$. Since $\Phi_\Psi(\cdot)$ is a continuous cdf, the above pointwise convergence implies a uniform convergence. This delivers the desired result for the non-asymptotically pivotal statistic.

To prove the validity of the bootstrap for the asymptotically pivotal statistic, we need to combine this result with Proposition 2.5.2. □

Appendix A

Background Statistical Results

In this Appendix we provide the main theorems and results used.
The two following propositions are the Markov's inequality and the Chebyshev's inequality.

Proposition A.0.1. *For a random variable X and for any $a > 0$*

$$P\left(|X| \geq a\right) \leq \frac{\mathbb{E}|X|}{a}$$

Proposition A.0.2. *If $\mathbb{E}|X|^2$ is finite, for any $a > 0$*

(i) $P\left(|X| \geq a\right) \leq \frac{\mathbb{E}X^2}{a^2}$

(ii) $P\left(|X - \mathbb{E}X| \geq a\right) \leq \frac{Var(X)}{a^2}$

The following result is a direct consequence of the convergence in probability.

Proposition A.0.3. *If $Z_n \xrightarrow{P} 0$, there exists a sequence $(\eta_n)_n$ converging to zero such that $P(|X_n| > \eta_n) = o(1)$.*

Proposition A.0.4. *If $X_n \xrightarrow{P} 0$ then there exists a subsequence (n_1) of (n) such that $X_{n_1} \xrightarrow{a.s.} 0$.*

A sequence of random vectors $\{X_n\}_n$ is said to converge in distribution (or weakly) to a random vector X if for any continuous bounded functions f it holds that $\mathbb{E}f(X_n) \to \mathbb{E}X$. We write that $X_n \rightsquigarrow X$. Convergence in distribution is also called "weak convergence". The following proposition is characterization of the convergence in distribution in terms of convergence of cdf's.

Proposition A.0.5. *$X_n \rightsquigarrow X$ if and only if $P(X_n \leq x) \to P(X \leq x)$ for any continuity point of $P(X \leq \cdot)$.*

Proposition A.0.6. *If the sequence of cdf's $\{F_n\}_n$ converges pointwise to a cdf F (i.e. $F_n(x) \to F(x)$ for all x) and F is continuous, then*

$$\sup_x |F_n(x) - F(x)| = o(1)\,.$$

Proposition A.0.7. *Let $(X_n, Y_n)_n$ be a sequence of random vectors. If $X_n \rightsquigarrow X$ and $Y_n \overset{a.s.}{\to} c$ for a constant c, then $(X_n, Y_n) \rightsquigarrow (X, c)$*

Proposition A.0.8. *Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be continuous at every point of a set C such that $P(X \in C) = 1$.*

(i) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$

(ii) If $X_n \overset{P}{\to} X$, then $g(X_n) \overset{P}{\to} g(X)$

(iii) If $X_n \overset{a.s.}{\to} X$, then $g(X_n) \overset{a.s.}{\to} g(X)$

Proposition A.0.9. *For each n let $Y_{n,1}, \ldots, Y_{n,k_n}$ be a independent random vectors with finite variances such that*

$$\sum_{i=1}^{k_n} \mathbb{E}\|Y_{n,i}\|^2 \mathbb{I}(\|Y_{n,i}\| > \eta) \to 0 \text{ for all } \eta > 0$$

$$\sum_{i=1}^{k_n} Cov(Y_{n,i}) \to \Sigma\,.$$

Then, $\sum_{i=1}^{k_n}(Y_{n,i} - \mathbb{E}Y_{n,i}) \rightsquigarrow \mathcal{N}(0, \Sigma)$.

Proposition A.0.10. *Let μ be a measure defined o a measurable space $(A, \mathcal{A})$.*

(i) If $g_n : A \mapsto \mathbb{R}$ is a sequence of positive function increasing μ-almost everywhere towards g (i.e. $g_n(x) \uparrow g(x)$ for all $x \in B$ with $\mu(B^c) = 0$) then $\int g_n(x)\mu(dx) \uparrow \int g(x)\mu(dx)$.

(ii) If g_n is a sequence of functions converging pointwise μ-almost everywhere towards g (i.e. $g_n(x) \to g(x)$ for all $x \in B$ with $\mu(B^c) = 0$) and $|g_n| \leq f$ μ- amost everywhere with $\int f(x)\mu(dx) < \infty$, then $\int g_n(x)\mu(dx) \to \int g(x)\mu(dx)$ and the previous integral is finite.

Proposition A.0.11. *Assume that the function g defined on $\mathbb{R}^k$ has continuous partial derivatives of order q at each point of an open set $S \subset \mathbb{R}^k$. Let $x \in S$. For each point y, with $y \neq x$, such that the segment $L(x, y)$ joining x and y lies in S, there exists a point z in the interior of $L(x, y)$ such that*

$$\begin{aligned} g(y) =& g(x) + \sum_{m=1}^{q-1} \frac{1}{m!} \sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} \frac{\partial^m g(t_1, \ldots, t_k)}{\partial t_{i_1} \ldots \partial t_{i_m}}\Big|_{t=x} \cdot \prod_{j=1}^{m} (y_{i_j} - x_{i_j}) \\ &+ \frac{1}{q!} \sum_{i_1=1}^{k} \cdots \sum_{i_q=1}^{k} \frac{\partial^q g(t_1, \ldots, t_k)}{\partial t_{i_1} \ldots \partial t_{i_q}}\Big|_{t=z} \cdot \prod_{j=1}^{q} (y_{i_j} - x_{i_j})\,. \end{aligned}$$

Proposition A.0.12. *(Polya's Theorem) Let $(X_n)_n$ be a sequence of random vectors in $\mathbb{R}^s$ each with cdf F_n. If $X_n \rightsquigarrow X$ and X has a continuous cdf F, then*

$$\sup_x |F_n(x) - F(x)| = o(1)$$

Proposition A.0.13. *Let $(F_n)_n$ be a sequence of cdf's such that $\sup_\tau |F_n(\tau) - F(\tau)| = o(1)$ for a cdf F that is strictly increasing and continuous. Then, for each $t \in (0, 1)$*

$$F_n^{-1}(t) = F^{-1}(t) + o(1)$$

Bibliography

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67, publisher: Wiley.

Davidson, R. and J. G. MacKinnon (2006): "The Case against JIVE," *Journal of Applied Econometrics*, 21, 827–833, publisher: Wiley.

Davison, A. C. and D. V. Hinkley (1997): *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press.

Efron, B. and R. J. Tibshirani (1994): *An Introduction to the Bootstrap*, CRC Press.

Hahn, J., J. Hausman, and G. Kuersteiner (2004): "Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations," *The Econometrics Journal*, 7, 272–306, publisher: [Royal Economic Society, Wiley].

Hall, P. (2013): *The Bootstrap and Edgeworth Expansion*, Springer Science & Business Media.

Horowitz, J. L. (2001): "The Bootstrap," in *Handbook of Econometrics, Chapter 52*, Elsevier, Handbook of Econometrics, 3159–3228.

Kosorok, M. R. (2007): *Introduction to Empirical Processes and Semiparametric Inference*, Springer.

MacKinnon, J. G. and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305–325.

Quenouille, M. H. (1949): "Approximate Tests of Correlation in Time-Series," *Journal of the Royal Statistical Society. Series B (Methodological)*, 11, 68–84, publisher: [Royal Statistical Society, Wiley].

Shao, J. and D. Tu (1995): *The Jackknife and Bootstrap*, Springer Series in Statistics, New York: Springer-Verlag.

Tukey, J. (1958): "Bias and confidence in not quite large samples," *Annals of Mathematical Statistics*, 29.

van der Vaart, A. W. (2000): *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. W. and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, New York: Springer-Verlag.