MŰEGYETEM 1782

Budapest University of Technology and Economics

Master Thesis

An Essay on Linear Regression Bootstrapping

Author:
Viet Hung Pham

Supervisors:
Olivér Kiss
Dr Miklós Koren
Dr Edith Alice Kovács

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Mathematics

in the

Faculty of Natural Sciences
Institute of Mathematics

Budapest, 2022

BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS

Abstract

Faculty of Natural Sciences
Institute of Mathematics

Master of Science in Mathematics

An Essay on Linear Regression Bootstrapping

by Viet Hung Pham

Linear regression, especially the ordinary least squares (OLS), and bootstrapping are widely used techniques in computational statistics, econometrics and machine learning. Despite being an old method dating back to Carl Friedrich Gauss's time, linear regression has gained more popularity since the breakthrough article by White (1980) about relaxing the homoscedasticity assumption of the Gauss-Markov theorem. Bootstrapping is a resampling technique invented by Efron (1979). It is a foundation for many statistical techniques, including the bias-corrected and accelerated (BCa) bootstrap confidence interval and the linear regression bootstrapping. The wild bootstrap, the most used linear regression bootstrap method in econometrics, was invented by Wu (1986) and Liu (1988) and has been further generalized by MacKinnon, Nielsen, and Webb (2022) for clustered data.

The thesis consists of four parts. The first chapter discusses the theoretical backgrounds of resampling methods (jackknife and bootstrap), emphasizing the bootstrap principle and confidence intervals. Chapter 2 details the Ordinary Least Squares (OLS) properties using the 5+1 Gauss-Markov assumptions. The third chapter focuses on linear regression bootstrapping, which merges the previous two parts. Three regression bootstrap methods are discussed, with their advantages and disadvantages. It is shown that the wild bootstrap is the regression method that best mirrors the true linear model compared to the other two methods. Chapter 4 presents experiments using the wild bootstrap on Hungarian companies' balance sheet data from 2014. These simulations examine the bootstrap distribution of regression coefficients and how BCa-based confidence intervals behave. The results of the experiments show that the wild bootstrap with appropriate replication numbers and confidence interval types is an effective tool for analyzing the characteristics of regression coefficients.

Acknowledgements

First and foremost, I would like to express my greatest gratitude to Olivér Kiss and Dr Miklós Koren for their guidance and continuous help throughout the thesis. I want to thank Dr Edith Alice Kovács for her dedicated support. I also owe tremendous gratitude to Dr Tim Hesterberg, Dr Bradley Efron, Dr James G. MacKinnon and Dr László Mátyás for their invaluable insights and advice related to bootstrapping and heteroscedasticity. I am also thankful to Dr Ádam Szeidl, the co-head of CEU MicroData, for giving me a chance to work as a Research Assistant in his research group. I cannot also leave out my friend György Ruzicska, who recommended me to work for CEU MicroData. Finally, I would like to thank my family and friends for their constant support throughout my master's degree.

Contents

List of Figures

List of Abbreviations

OLS	**O**rdinary **L**east **S**quares
DGP	**D**ata **G**enerating **P**rocess
BLUE	**B**est **L**inear **U**nbiased **E**stimator
RV	**R**andom **V**ariable
CDF	**C**umulative **D**istribution **F**unction
eCDF	**e**mpirical **C**umulative **D**istribution **F**unction
PMF	**P**robability **M**ass **F**unction
PDF	**P**robability **D**ensity **F**unction
MSE	**M**ean **S**quared **E**rror
LLN	**L**aw of **L**arge **N**umbers
sLLN	**s**trong **L**aw of **L**arge **N**umbers
wLLN	**w**eak **L**aw of **L**arge **N**umbers
CLT	**C**entral **L**imit **T**heorem
MC	**M**onte **C**arlo
DKW	**D**voretzky–**K**iefer–**W**olfowitz
LOTUS	**L**aw **O**f **T**he **U**nconscious **S**tatistician
HCCME	**H**eteroscedasticity-**C**orrected **C**ovariance **M**atrices **E**stimator
HC	**H**eteroskedasticity-**C**onsistent
HCE	**H**eteroskedasticity-**C**onsistent Standard **E**rrors
SE	**S**tandard **E**rrors
i.i.d	**i**ndependent **i**dentically **d**istributed
CI	**C**onfidence **I**nterval
BC	**B**ias-**C**orrected
BCa	**B**ias-**C**orrected and **a**ccelerated
LM	**L**agrange **M**ultiplier
LOO	**L**eave-**O**ne-**O**ut
TSS	**T**otal **S**um of **S**quares
ESS	**E**xplained **S**um of **S**quares
RSS	**R**esidual **S**um of **S**quares

List of Symbols

$X_n \xrightarrow{P} X$	convergence in probability
$X_n \xrightarrow{d} X$	convergence in distribution
$X_n \xrightarrow{a.s.} X$	almost sure convergence
$\lim_{n\to\infty}$	limit
$\text{plim}_{n\to\infty}$	probability limit
$\sum_{i=1}^{n}$	summation from $i = 1$ to $i = n$
F	CDF (Cumulative Distribution Function)
$\hat{F}_n$	eCDF (empirical Cumulative Distribution Function)
F_{d_1, d_2}	F distribution with d_1 and d_2 degrees of freedom
LM	LM statistic
$\boldsymbol{R^2}$	R-squared
$\mathbb{R}$	real number
T	statistical functional
Cov	covariance
$\det$	determinant
$sd(X)$	$\sqrt{\mathbb{V}(X)}$ (standard deviation)
se	$\sqrt{\mathbb{V}(\hat{\theta}_n)}$ (standard error)
$\widehat{se}$	estimated standard error
$bias$	$\mathbb{E}(\hat{\theta}) - \theta$
$\widehat{bias}$	estimated bias
$jack$	jackknife
$boot$	bootstrap
$\hat{\theta}_n$	estimator of a parameter θ
$\mathbb{P}(X)$	probability
$\mathbb{E}(X)$	expectation
$\mathbb{V}(X)$	variance
$\mathcal{C}(\hat{\theta})$	confidence interval
Φ	CDF of a standard normal random variable
$z_{1-\alpha}$	$\Phi^{-1}(1 - \alpha)$
$\nabla_{\boldsymbol{\beta}} f$	gradient of f with respect to $\boldsymbol{\beta}$
$\boldsymbol{X}^T$	transpose of $\boldsymbol{X}$
$\boldsymbol{X}^{-1}$	inverse of $\boldsymbol{X}$
$\boldsymbol{X} > 0$	positive definite
$\boldsymbol{X} \geq 0$	positive semidefinite

$\mathrm{Im}(\boldsymbol{X})$	image of $\boldsymbol{X}$
$\mathrm{diag}\,(x_1,\ldots,x_n)$	$n \times n$ diagonal matrix with diagonal elements $x_1,\ldots,x_n$
$\boldsymbol{I_n}$	$n \times n$ identity matrix
χ^2_k	chi-square distribution with k degrees of freedom
$N(0,1)$	standard normal distribution
$N(\mu,\sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathrm{Bin}(n,p)$	binomial distribution with parameters n and p
$\#\{\cdot\}$	cardinality of a set
$\lVert\cdot\rVert$	norm
$\lVert\cdot\rVert_\infty$	sup-norm or max-norm
sup	supremum
$\odot$	Hadamard product
$\sim$	is distributed as
$\dot\sim$	is *approximately* distributed as
$\approxeq$	approximately equal
$\equiv$	identically equal

Dedicated to my late father…

Chapter 1

Resampling

1.1 Introduction

One of the main goals of statistical analysis is to "extract all the information from the data" (Rao, 1989, p. 91) to gauge (relative) accuracies of an estimator or make inferences about the population and its properties from the sample data at hand. The procedure can be done by *estimating a statistic*, a function of the data selected according to the question of interest. A closely-related concept is the *sampling distribution* which is the probability distribution of a given random-sample-based statistic. In conventional methods, prior knowledge or hypotheses are usually assumed for a statistic's sampling distribution before applying statistical procedures. Knowing the estimator's accuracy in estimation problems is vital, as any estimator could have an estimation error. Such accuracy measures are bias, standard error, and mean squared error. They can be deployed to select the appropriate from a set of estimators. They are also part of the characterization of the estimator's sampling distribution (Shao and Tu, 2012).

The **data generating process** (DGP) generates the data one is interested in, primarily unknown in the real world. The DGP determines the sampling distribution of a statistic and its properties for our observed data. The aim is to approximate them as accurately as possible for predictions or inferential tasks. For **parametric** problems, it is not necessary to approximate the accuracy measures; some theorems prove the superiority of a specific estimator for a given problem. For example, ordinary least squares (OLS) (see Chapter 2) is the best linear unbiased estimator (BLUE) by the Gauss-Markov theorem (see Theorem 2.3.2).

However, in real-life data/observations, **nonparametric** tasks occur; the parameters of the problems are unknown, unlike in the case of **parametric** ones. There are situations where the distribution of a nonparametric statistic can be greatly approximated by an appropriate parametric one, such as the number of earthquakes in a year can be modelled using Poisson distribution because of the *Poisson paradigm* (Blitzstein and Hwang, 2019).

In most cases, sampling distributions of statistics do not follow or depend on any well-known/named distributions. Relative accuracies of estimators need to be assessed using the observed data. **Jackknife** and **bootstrap** are viable tools for estimating accuracy measures, especially bias and standard errors, and constructing confidence intervals.

1.2 Conventional Methods

In conventional methods, gauging accuracy measures is done by using empirical counterparts of explicit theoretical formulas. These formulas are derived from a postulated model. For example, the variance for the sample mean has a closed-form theoretical formula. However, there is no such analytical solution determining – for example – the variance of the sample median.

The followings are the weaknesses and disadvantages of using conventional techniques (Shao and Tu, 2012):

- Assumptions of many unrealistic and highly restrictive conditions. For instance, equal variances in the assumption of OLS (see OLSA.3 assumption in Chapter 2).
- Asymptotic theories require large sample sizes to have proper accuracy measures.
- Postulated models are the foundation of some of the theoretical formulas or their approximations. The resulting accuracy estimator may have limitations whenever a model is marginally misspecified.
- It is sometimes difficult or impossible to derive a closed-form/analytical formula for a statistic. An excellent example of that is the variance of the sample median.
- There are cases when theoretical formulas are too complicated or impractical for estimating accuracy measures.

1.3 Empirical CDF and Statistical Functionals

1.3.1 Empirical CDF and its Properties

Definition 1.3.1 (CDF)**.** The **cumulative distribution function** (CDF) of a random variable (RV) X is the function F_X given by

$$F_X(x) = P(X \leq x). \tag{1.1}$$

Depending on the context, the notation can usually be simplified by removing the X subscript from F_X.

Definition 1.3.2 (Empirical CDF)**.** Let i.i.d. $X_1, \ldots, X_n$ be a random sample from a CDF F. The usually unknown F underlying distribution can be estimated by the **empirical CDF** (eCDF) $\hat{F}_n$ that puts mass $1/n$ at each data point X_i, that is,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \tag{1.2}$$

where I is an indicator RV such that

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise.} \end{cases}$$

The following three theorems introduce essential asymptotic properties of the **eCDF**, which serves as a foundation for modern computational statistics and econometrics. Van der Vaart (2000, chap. 19) and Wasserman (2006, sec. 2.1) present further details on the topic.

Theorem 1.3.1 (sLLN and asymptotic Normal CLT for eCDF)**.** *Let $X_1, \ldots, X_n \sim F$ and let $\hat{F}_n$ be the eCDF and the estimator for F. At any fixed value of x:*

- *$I(X_i \leq x)$ is a Bernoulli RV with parameter $p = F(x)$*
- *Therefore, $n\hat{F}_n(x)$ is a Binomial RV with mean $nF(x)$ and variance $nF(x)\,(1 - F(x))$. It implies that $\hat{F}_n(x)$ is an **unbiased** estimator for $F(x)$. In mathematical notations,*

  $$n\hat{F}_n(x) \sim Bin\Big(nF(x),\ nF(x)\big(1 - F(x)\big)\Big)$$

  which implies

  $$\mathbb{E}\Big(\hat{F}_n(x)\Big) = F(x) \quad \text{and} \quad \mathbb{V}\Big(\hat{F}_n(x)\Big) = \frac{F(x)\big(1 - F(x)\big)}{n}\,.$$

*The first equality from the left represents the property of unbiased eCDF $\hat{F}_n$. Hence, by the **strong Law of Large Numbers** (sLLN), $\hat{F}_n$ is a (strongly) consistent estimator of F for every value x,*

$$\hat{F}_n(x) \xrightarrow{a.s.} F(x) \qquad \forall x. \tag{1.3}$$

*By the **Central Limit Theorem** (CLT), $\hat{F}_n(x)$ estimator is asymptotically normal,*

$$\sqrt{n}\,\Big(\hat{F}_n(x) - F(x)\Big) \xrightarrow{d} N\Big(0,\ F(x)\big(1 - F(x)\big)\Big)\,. \tag{1.4}$$

There is a stronger result called the *Glivenko-Cantelli* theorem, which extends the sLLN for eCDF and states uniform convergence (Van der Vaart, 2000).

Theorem 1.3.2 (Glivenko-Cantelli)**.** *Let i.i.d. $X_1, \ldots, X_n \sim F$. Then*

$$\|\hat{F}_n - F\|_\infty = \sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0 \tag{1.5}$$

where the sup-norm expression is known as the Kolmogorov-Smirnov statistic *(Van der Vaart, 2000). It is used for testing the goodness-of-fit between the eCDF $\hat{F}_n(x)$ and the assumed true CDF F.*

The **DKW inequality**, named after Aryeh Dvoretzky, Jack Kiefer and Jacob Wolfowitz, describes how close an eCDF will be to the true CDF from which the empirical samples are drawn (Dvoretzky, Kiefer, and Wolfowitz, 1956). That is to say, DKW inequality provides a bound on the tail probabilities of the sup-norm expression in (1.5).

Theorem 1.3.3 (DKW inequality)**.** *Let i.i.d.* $X_1, \dots, X_n \sim F$*. Then,*

$$\mathbb{P}\left(\sup_x |\hat{F}_n(x) - F(x)| \right) \leq 2e^{-2n\epsilon^2} \,. \tag{1.6}$$

1.3.2 Statistical Functionals

Definition 1.3.3 (Functional, Statistical Functional)**.** A **functional** is a mapping T from a space $\mathcal{F}$ (collection of functions) into the field of real numbers, that is, $T : \mathcal{F} \mapsto \mathbb{R}$. A **statistical functional** is a mapping T from *CDF* F to $\mathbb{R}$.

Definition 1.3.4 (Plug-in estimator)**.** The **plug-in estimator** of $\theta = T(F)$ is given by

$$\hat{\theta}_n = T(\hat{F}_n). \tag{1.7}$$

The known eCDF $\hat{F}_n$ is used in the statistical functional instead of the unknown CDF F .

Definition 1.3.5 (Linear Functional)**.** T is a **linear functional** if for some function g(x)

$$T(F) = \int g(x)\, dF(x) \,. \tag{1.8}$$

Because the eCDF $\hat{F}_n$ is discrete in which each X_i is weighted uniformly $(1/n)$, then

Theorem 1.3.4. *The plug-in estimator for the linear functional in (1.8) is*

$$\hat{\theta}_n = T(\hat{F}_n) = \int g(x)\, d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) + c \tag{1.9}$$

where c is a constant.

Examples of plug-in linear functionals are the sample mean and (biased) sample variance. A notable example of plug-in nonlinear functionals is the sample quantile. The reason for the usage of statistical functionals is that they rigorously express population parameters of interest. Note that a random sample i.i.d. $X_1, \dots, X_n$ is observed from an unknown DGP characterized by the underlying CDF F. As CDF F is also unknown, using the eCDF $\hat{F}_n$ is the best option to describe the underlying F using our random sample. Because the plug-in statistical functionals map $\hat{F}_n$ into $\mathbb{R}$, they can be viewed as quantities describing the features of the population through a random sample. For example, the sample mean, variance, and quantiles of $\hat{F}_n$ are quantities depicting the population through a random sample. However, the quality of our random sample matters a lot. With a representative population sample, plug-in statistical functionals can accurately estimate population parameters. The jackknife and bootstrap are resampling methods that harness the power of *plug-in statistical functionals* to approximate the bias and standard error of a parameter and construct a confidence interval (CI) using the observed random sample.

For a more detailed presentation on the statistical functionals, see Wasserman (2006, sec. 2.2), Wasserman (2004, sec. 7.2) and Chen (2020, sec. 10.2, 10.4).

1.4 Jackknife

A nonparametric resampling technique was developed by Quenoille (1949) to eliminate bias in an estimate. It was popularized by Tukey (1958), who coined Quenoille's method as the **jackknife**. The concept was further developed by Efron and Tibshirani (1994) and Shao and Tu (2012). The reason for this naming is to highlight the all-purpose nature of this tool to solve statistical problems. Even though this thesis's primary focus is bootstrapping, the jackknife method is used in the construction of *bias-corrected and accelerated* (BCa) CI. In addition, it was shown by MacKinnon and White (1985) that the HC3 covariance matrix estimator is numerically the same as the jackknife estimate of variance for the OLS estimator $\hat{\boldsymbol{\beta}}$.

The jackknife is mainly used for computing bias, standard error of a parameter estimate, and constructing confidence intervals. It is computationally less expensive than the bootstrap, but the latter has more advantageous properties, for example, in the case of nonlinear and non-smooth statistics. In those cases, the jackknife fails (Efron and Tibshirani, 1994). As for linear statistics (linear functional) in the form of (1.9), the jackknife is a *linear approximation* to the bootstrap.

There are two types of the jackknife method; the classic *leave-one-out* and the *delete-d* jackknife. Regarding methodology, the only major difference is that the *delete-d* jackknife leaves out $d > \sqrt{n}$ observations instead of leaving out one observation before applying the statistics of interest to the remaining ones. As a result, the inconsistency of the *leave-one-out* jackknife can be fixed for non-smooth statistics, such as quantiles, at the cost of more computational resources. For a more detailed explanation of the delete-d jackknife, see Efron and Tibshirani (1994, sec. 11.7) and Shao and Tu (2012, sec. 2.3).

Leave-One-Out Jackknife

Usually, the *leave-one-out* version is meant when referring to the **jackknife**. Some of the necessary technical aspects of this statistical method based on Efron and Tibshirani (1994, chap. 11) are described below.

First, denote S_n a set of observed values of random variables or random vectors. If the i^{th} value/vector is left out from S_n, let $S^*_{(-i)}$ be that set after the removal. Let us have a sample $S_n = (X_1, X_2, \dots, X_n)$, where i.i.d. $X_1, \dots, X_n \sim F$, and an estimator $\hat{\theta} = T(S_n)$. Removing the i^{th} observation at a time is called the **leave-one-out** principle.

Definition 1.4.1 (Jackknife Sample)**.** The i^{th} **jackknife sample** is

$$S^*_{(-i)} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \qquad \forall i = 1, 2, \ldots, n. \tag{1.10}$$

Definition 1.4.2 (Jackknife Replication)**.** The i^{th} **jackknife replication** of $\hat{\theta}$ is

$$\hat{\theta}^*_{(-i)} = T(S^*_{(-i)})\,. \tag{1.11}$$

The algorithm of the leave-one-out jackknife is given in Figure 1.1.

```
Algorithm 1: The (Leave-One-Out) Jackknife
Input  : $S_n = (X_1, X_2, \ldots, X_n)$ – observed samples of size $n$
Output: $J_{\theta_{jack}} = [\hat{\theta}^*_{(-1)}, \ldots, \hat{\theta}^*_{(-n)}]$ – jackknifed parameters
Init: $N$ := sample size
      $J_{\theta_{jack}}$ := [ ] – an empty array of parameter estimates
      $T$ – A function of a statistic
for $i$ := 1 to $N$ do
    $S^*_{(-i)}$ := remove $i^{\text{th}}$ element from $S_n$
    $\hat{\theta}^*_{(-i)} := T\big(S^*_{(-i)}\big)$
    $J_{\theta_{jack}}[i] := \hat{\theta}^*_{(-i)}$
end
```

FIGURE 1.1: The (Leave-One-Out) Jackknife Algorithm

Before defining the jackknife estimates of bias and standard error, let

$$\hat{\theta}^*_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}^*_{(-i)} \tag{1.12}$$

be the *mean of the jackknife replications.*

Definition 1.4.3 (Jackknife Estimate of Bias)**.** The **jackknife estimate of bias** is provided by

$$\widehat{bias}_{jack} = (n-1)\left(\hat{\theta}^*_{(\cdot)} - \hat{\theta}\right). \tag{1.13}$$

Definition 1.4.4 (Jackknife Estimate of Standard Error)**.** The **jackknife estimate of the standard error** is

$$\widehat{se}_{jack} = \left[\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}^*_{(-i)} - \hat{\theta}^*_{(\cdot)}\right)^2\right]^{1/2}. \tag{1.14}$$

The question arises over why the $(n-1)$ factor appears in both formulas instead of 1. The short answer is that the $(n-1)$ "inflation factor" is necessary because the jackknife deviations of bias and standard errors in (1.13) and (1.14), respectively, that is,

$$\left(\hat{\theta}^*_{(\cdot)} - \hat{\theta}\right) \quad \text{and} \quad \left(\hat{\theta}^*_{(-i)} - \hat{\theta}^*_{(\cdot)}\right)^2$$

are usually much smaller than the bootstrap deviations of those accuracy measures. The reason for this is that a typical jackknife sample exhibits a much closer resemblance to the original sample than a typical bootstrap sample.

Shao and Tu (2012, chap. 2) provide further details on the properties of jackknife.

1.5 Bootstrap

Another major resampling method, the bootstrap, was devised by Efron (1979). His work was inspired by preceding results on the jackknife. The term bootstrap was explained in Efron and Tibshirani (1994, p. 5):

> "The use of the term bootstrap derives from the phrase to pull oneself up by one's bootstrap, widely thought to be based on one of the eighteenth century Adventures of Baron Munchausen, by Rudolph Erich Raspe. (The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.)"

It has become one of the most significant modern computational statistical results (Efron and Hastie, 2016; Hansen, 2022a). Since then, several branches of bootstrapping have been developed, such as linear regression bootstrapping, time series bootstrapping, bootstrap aggregating and constructions of bootstrap confidence intervals. In this section, the bootstrap principle, bootstrap estimates of bias and standard error, and bootstrap confidence intervals are detailed. Before discussing those topics, the idea of simulation is presented through an example of approximating the variance (see Wasserman, 2006, sec. 8.1).

1.5.1 Monte Carlo Simulation

Draw i.i.d. $X_1, \dots, X_n \sim F$ with finite mean μ and finite variance σ^2. By the wLLN,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \int x \, dF(x) = \mathbb{E}(X) = \mu \tag{1.15}$$

as $n \to \infty$. The LLN states that as n grows, the sample mean $\bar{X}_n$ converges to the true mean $\mathbb{E}(X) = \mu$. In other words, if drawing a large sample from the unknown CDF F, the sample mean $\bar{X}_n$ is a consistent estimator of the population mean $\mathbb{E}(X) = \mu$. The procedure described above is called the **Monte Carlo simulation** or the Monte Carlo method. In a Monte Carlo simulation, an arbitrarily large number of n can be determined, which results in an insignificant difference between the sample mean $\bar{X}_n$ and the true mean $\mathbb{E}(X) = \mu$. The *Law Of The Unconscious Statistician* (LOTUS) can be used for generalizing (1.15). Let X be a random variable with CDF F, and let g be any function with finite mean, then

$$\widehat{\eta}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{P} \int g(x) \, dF(x) = \mathbb{E}\big(g(X)\big) = \eta \tag{1.16}$$

as $n \to \infty$. $\widehat{\eta}_n$ is called the Monte Carlo estimator of the expectation η.

Using the idea above with the LOTUS, it can be shown that the sample variance can be used to estimate $\mathbb{V}(X)$ through a Monte Carlo simulation,

$$\begin{aligned}\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2 \\ &\xrightarrow{P} \int x^2\,dF(x) - \left(\int x\,dF(x)\right)^2 = \mathbb{V}(X)\,.\end{aligned} \tag{1.17}$$

The $\widehat{\sigma}_n^2$ is called the Monte Carlo estimator of the variance of X.

1.5.2 Bootstrap Principle

The goal of bootstrapping, as shown in Figure 1.2, is to estimate the true sampling distribution of some quantity T (such as mean and regression coefficients) with its relative accuracies, such as bias or standard error. The true sampling distribution is calculated by taking new samples (Sample World) from the true population, computing T-s and gathering all these T values to describe the true sampling distribution. Nonetheless, it is a really expensive approach. An alternative is to use a single sample to generate new samples called **bootstrap samples** by a sampling approach called

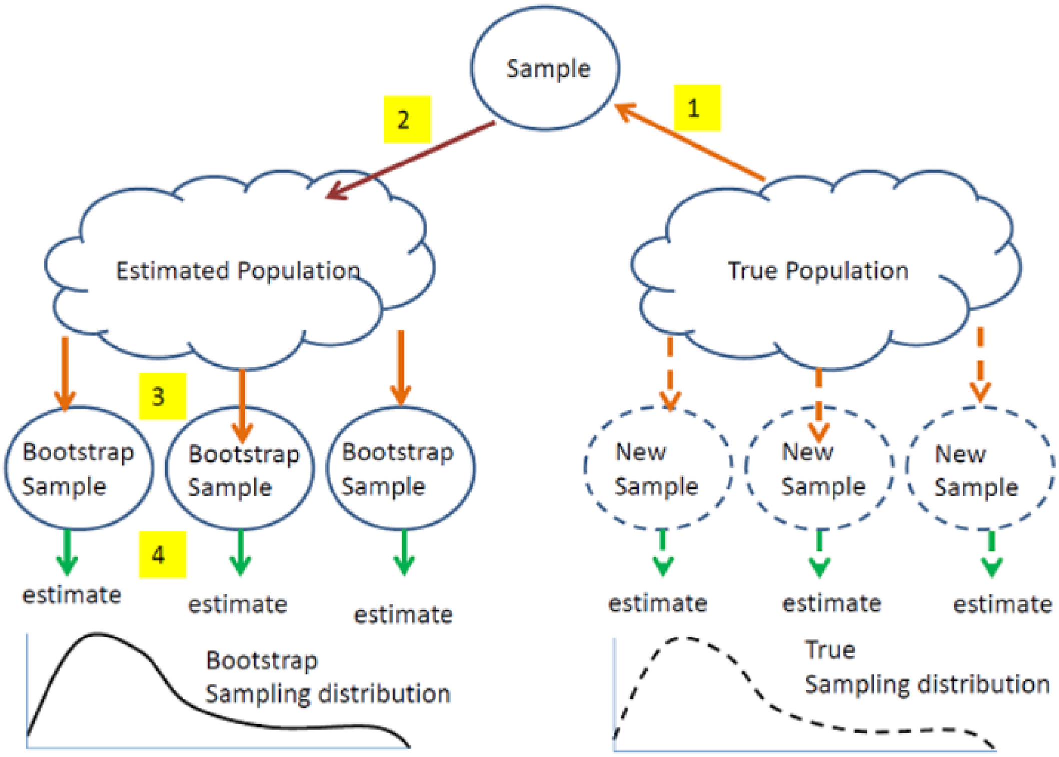


FIGURE 1.2: The Goal of Bootstrapping (Altman, 2018)

sampling with replacement. By utilising these bootstrap samples, the *bootstrap sampling distribution* of an estimator and its accuracy measures can be determined.

Let i.i.d. $X_1, \ldots, X_n \sim F$, the original, random set of observations. The CDF F of the data is unknown. The question is how to simulate from the CDF F when the observation data are from eCDF $\hat{F}_n$ (see Definition 1.3.2). The short answer is to **sample with replacement** from these n i.i.d. $X_1, \ldots, X_n$ points. It would lead to a set of new observations $X_1^{*(1)}, \ldots, X_n^{*(1)}$, called a **bootstrap sample**. If repeating this *sampling with replacement* procedure from the original set of observations, another set of bootstrap sample $X_1^{*(2)}, \ldots, X_n^{*(2)}$ is yielded. Repeating this sampling procedure for B rounds results in the following B bootstrap samples (Chen, 2019):

$$\begin{array}{ccc} X_1^{*(1)} & \cdots & X_n^{*(1)} \\ X_1^{*(2)} & \cdots & X_n^{*(2)} \\ \vdots & \vdots & \vdots \\ X_1^{*(B)} & \cdots & X_n^{*(B)} \,. \end{array} \tag{1.18}$$

All these bootstrap samples have their own eCDFs, which are $\hat{F}_n^{*(b)}$ $(b = 1, \ldots, B)$. Interestingly, it can be shown that the probability that an individual observation will occur at least once in the bootstrap sample is

$$\mathbb{P}(\text{an observation in a bootstrap sample}) = 1 - \left(1 - \frac{1}{n}\right)^n \tag{1.19}$$

where

$$\lim_{n\to\infty} 1 - \left(1 - \frac{1}{n}\right)^n = 1 - e^{-1} \approx 0.632\,. \tag{1.20}$$

The approximation 0.632 is outstanding even for small n (Hansen, 2022a). An elegant proof of this task can be carried out with the isomorphic problem called *de Montmort's matching problem* (Blitzstein and Hwang, 2019, Example 1.6.4).

The whole *sampling with replacement* approach is called the **nonparametric bootstrapping** or **bootstrap principle**. The implicit assumption of nonparametric bootstrapping is that the observed random sample $X_1, \ldots, X_n$ is "somewhat" representative of the population of the whole.

The following diagrams sum up the discussion above (Wasserman, 2004):

$$\begin{array}{rlllll} \text{Population World} & F & \Longrightarrow & X_1, \ldots, X_{n+} & \Longrightarrow & \theta \ = T(F) \\ \text{Sample World} & \hat{F}_n & \Longrightarrow & X_1, \ldots, X_n & \Longrightarrow & \hat{\theta} \ = T(\hat{F}_n) \\ \text{Bootstrap World} & \hat{F}_n^* & \Longrightarrow & X_1^*, \ldots, X_n^* & \Longrightarrow & \hat{\theta}^* = T(\hat{F}_n^*) \end{array} \tag{1.21}$$

where $T(\hat{F}_n)$ and $T(\hat{F}_n^*)$ use the plug-in estimator principle (see Definition 1.3.4). The difference between the Population World and Sample World is that the former has all the observation points (denoted as $n+$), and the latter has a random sample of n observations from the population $n+$. Having collected all $\hat{\theta}^* = T(\hat{F}_n^*)$ (here, the bootstrap sample indicator index is removed for the sake of readability), the bootstrap sampling distribution of the estimator in question can be drawn. Therefore, all accuracy measures of interest, such as bias and standard errors, can be computed (see Section 1.5.3). Furthermore, the bootstrap confidence interval of the estimator can also be determined (see Section 1.5.4). The algorithmic details of the bootstrapping discussed above are given in Figure 1.3.

```
Algorithm 2: The Bootstrap
  Input  : $\hat{F}_n = (X_1, X_2, \ldots, X_n)$ – observed sample of size $n$
  Output: $L_{\theta_{boot}} = [\hat{\theta}^*_1, \ldots, \hat{\theta}^*_n]$ – bootstrapped parameters
  Init: $B$ – # of bootstrap repetitions
        $b := 1$
        $L_{\theta_{boot}} := [\,]$ – an empty array of parameter estimates
        $T$ – A function of a statistic
  while $b < B + 1$ do
      $\hat{F}^{*(i)}_n$ := a sample size $n$ from $S_n$ by sampling with replacement
      $\hat{\theta}^*_b := T\left(\hat{F}^{*(b)}_n\right)$
      $L_{\theta_{boot}}[b] := \hat{\theta}^*_b$
      $b{+}{+}$
  end
```

FIGURE 1.3: The Bootstrap Algorithm

1.5.3 Bootstrap Estimates of Bias and Standard Error

For the majority of statistics, the bootstrap estimates of bias and standard error can be approximated by the Monte Carlo simulation defined in Section 1.5.1 and the bootstrap algorithm in Figure 1.3. The starting point of this section is the Sample World (1.21) with eCDF $\hat{F}_n$ and its plug-in estimator $\hat{\theta} = T(\hat{F}_n)$. This subsection is based on Efron and Tibshirani (1994, chap. 6, 10).

Definition 1.5.1 (Bootstrap Replication)**.** The b^{th} **bootstrap replication** of $\hat{\theta}$ is

$$\hat{\theta}^*_b := T\left(\hat{F}^{*(b)}_n\right). \tag{1.22}$$

Before turning to the bootstrap estimates of bias and standard error, generate independent bootstrap samples $\hat{F}^{*(b)}_n$, where $b = 1, \ldots, B$, and evaluate bootstrap replications $\hat{\theta}^*_b := T\left(\hat{F}^{*(b)}_n\right)$. Let

$$\hat{\theta}^*_{(\cdot)} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*_b = \frac{1}{B}\sum_{b=1}^{B} T\left(\hat{F}^{*(b)}_n\right) \tag{1.23}$$

be the *mean of the bootstrap replications.*

Definition 1.5.2 (Bootstrap Estimate of Bias)**.** The **bootstrap estimate of bias** is given by

$$\widehat{bias}_{boot} = \hat{\theta}^*_{(\cdot)} - T(\hat{F}_n)\,. \tag{1.24}$$

Definition 1.5.3 (Bootstrap Estimate of Standard Error)**.** The **bootstrap estimate of the standard error** is

$$\widehat{se}_{boot} = \left[\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \hat{\theta}^*_{(\cdot)}\right)^2\right]^{1/2}. \tag{1.25}$$

The bootstrap relies on two approximations (Caron, 2019):

1. Approximate $\mathbb{V}_F(\hat{\theta}_n)$ by $\mathbb{V}_{\hat{F}_n}(\hat{\theta}^*)$ using $\hat{F}_n$
2. Approximate $\mathbb{V}_{\hat{F}_n}(\hat{\theta}^*)$ by Monte Carlo (MC) simulation to get $\widehat{se}^2_{boot}$

To sum up:

$$\mathbb{V}_F(\hat{\theta}_n) \overset{\text{eCDF}}{\approxeq} V_{\hat{F}_n}(\hat{\theta}^*) \overset{\text{MC}}{\approxeq} \widehat{se}^2_{boot}\,. \tag{1.26}$$

As mentioned in Section 1.4, even though the bootstrap is more computationally expensive than the jackknife, the bootstrap can be used to determine the standard error of nonlinear and non-smooth statistics. A prime example of nonlinear statistics where the difference can be seen is the quantile functions, such as the median.

Shao and Tu (2012, chap. 3) present further details on the properties of bootstrap.

1.5.4 Bootstrap Confidence Interval

Confidence intervals are one of the most applied statistical tools for data analysis. The advantage of using confidence intervals is their intuitive capability to merge point estimation and hypothesis testing into a single inferential statement (DiCiccio and Efron, 1996). With the advancement in the bootstrap confidence interval in the 80s and 90s, a more accurate confidence interval can be built without using strong assumptions such as a normality assumption on the sample distribution of an estimator. Bootstrap confidence intervals offer an order-of-magnitude improvement – from first-order to second-order accuracy – upon the accuracy of the standard intervals such as $\hat{\theta} \pm z_{\alpha/2}\,\hat{\sigma}$ for nominal $100\,(1-\alpha)\%$ two-sided coverage (Efron and Narasimhan, 2020). z_α is the 100αth percentile point of a standard normal distribution. Examples of bootstrap confidence intervals that are discussed in this section are percentile, bias-corrected (BC), bias-corrected accelerated (BCa), and bootstrap-t. Before detailing these methods, define some fundamental concepts.

Foundational Concepts in Bootstrap Confidence Intervals

First, define the concept of an accurate confidence interval (Hesterberg, 2015, chap. 5), (Efron and Tibshirani, 1994, chap. 14).

Definition 1.5.4 (Accurate confidence intervals)**.** An **accurate** $100\,(1-\alpha)\%$ **confidence interval** misses $100\,(\alpha/2)\%$ of the time on each side. The central $1-\alpha$ confidence interval $(\hat{\theta}_{\text{low}}, \hat{\theta}_{\text{up}})$ is meant to have a probability $\alpha/2$ of not covering the population parameter θ from below and above:

$$\mathbb{P}(\theta < \hat{\theta}_{\text{low}}) = \frac{\alpha}{2} \qquad \text{and} \qquad \mathbb{P}(\theta > \hat{\theta}_{\text{up}}) = \frac{\alpha}{2}\,. \tag{1.27}$$

An interval that **undercovers** on one side and **overcovers** on the other is *biased.*

According to the definition, the *accurate* 95% *confidence interval* is when it misses 2.5% of the time on each side. To demonstrate the definition above with a simple example, consider a confidence interval whose overall coverage of 95% is correct but misses 4% on one side and 1% on the other. In this case, on one side, it *overcovers* and *undercovers* on the other side. It is **not** an accurate confidence interval even though the overall coverage is all right. The confidence interval, in this case, is called biased.

Another aspect of a bootstrap confidence interval is how fast the constructed confidence interval converges to the accurate confidence interval. The convergence can be described with first- and second-order accuracy (Hesterberg, 2015, chap. 5), (Efron and Tibshirani, 1994, chap. 14).

Definition 1.5.5 (First and Second Accurate Inferences)**.** A confidence interval is **first-order** or **second-order accurate** if the one-sided non-coverage probabilities differ from the nominal values by $O\big(n^{-1/2}\big)$ or $O\big(n^{-1}\big)$, respectively. That is, the errors in matching (1.27) go to zero at the rate of $\boldsymbol{1/\sqrt{n}}$ for a first-order accurate or $\boldsymbol{1/n}$ for a second-order accurate confidence interval as sample size n grows,

$$\mathbb{P}(\theta < \hat{\theta}_{\text{low}}) = \frac{\alpha}{2} + \frac{c_{\text{low}}}{\sqrt{n}} \qquad \text{and} \qquad \mathbb{P}(\theta > \hat{\theta}_{\text{up}}) = \frac{\alpha}{2} + \frac{c_{\text{up}}}{\sqrt{n}} \tag{1.28}$$

$$\mathbb{P}(\theta < \hat{\theta}_{\text{low}}) = \frac{\alpha}{2} + \frac{c_{\text{low}}}{n} \qquad \text{and} \qquad \mathbb{P}(\theta > \hat{\theta}_{\text{up}}) = \frac{\alpha}{2} + \frac{c_{\text{up}}}{n} \tag{1.29}$$

respectively, for two constants c_{low} and c_{up}. A procedure needs to handle *bias*, *skewness*, and *transformations* to be *proper* second-order accurate.

When examining a bootstrap confidence interval, a crucial concept is the *transformation-invariant* or *transformation-respecting* property. After obtaining a confidence interval from a bootstrap sampling distribution, applying a monotonic function on this original bootstrap sampling distribution would result in a new bootstrap distribution with a new confidence interval preserving the same asymmetry and bias as the original one (Hesterberg, 2015).

Definition 1.5.6 (Transformation Invariance)**.** Suppose $\hat{\theta}$ is observed from a family of densities $f_\theta(\hat{\theta})$, and a $1-\alpha$ confidence interval for $\mathcal{C}(\hat{\theta})$ is built for the true parameter θ. Let ϕ be a monotonic increasing function of θ

$$\phi = m(\theta) \tag{1.30}$$

and therefore, $\hat{\phi} = m(\hat{\theta})$ is the point estimate. Then $\mathcal{C}(\hat{\theta})$ maps into $\mathcal{C}^{\phi}(\hat{\phi})$ pointwise and the resulting $1-\alpha$ confidence interval for ϕ is

$$\mathcal{C}^{\phi}(\hat{\phi}) = \left\{\phi = m(\theta) \mid \theta \in \mathcal{C}(\hat{\theta})\right\}. \tag{1.31}$$

The definition states that if the event $\{\theta \in \mathcal{C}(\hat{\theta})\}$ has an occurrence probability of α, then the event $\{\phi \in \mathcal{C}^{\phi}(\hat{\phi})\}$ must have the same probability (Efron and Hastie, 2016). It has been proven that the percentile and BC methods are transformation-respecting but not second-order accurate and can partially control skewness. Bootstrap-t is second-order accurate but not transformation-respecting. The standard technique is neither. However, BCa can handle all these aspects to be a proper second-order accurate confidence interval. It is recommended to use BCa intervals for a general use case, especially for nonparametric problems (Efron and Tibshirani, 1994).

The upcoming sections describing the different methods of constructing bootstrap confidence intervals follow Efron and Hastie (2016) and Efron and Tibshirani (1994). The initial step of constructing bootstrap confidence intervals is to generate bootstrap replications

$$\hat{\theta}^*_b \quad \text{where} \;\; b = 1, \ldots, n \tag{1.32}$$

using Algorithm 2.

Percentile Confidence Intervals

Definition 1.5.7 (Bootstrap CDF)**.** The **bootstrap CDF** $\hat{G}(x)$ is the proportion of bootstrap samples less than x

$$\hat{G}(x) = \frac{\#\{\hat{\theta}^*_b \leq x\}}{B}. \tag{1.33}$$

Definition 1.5.8. The $\alpha^{\mathbf{th}}$ **percentile point** $\hat{\theta}^{*(\alpha)}$ of the bootstrap distribution is provided by the inverse function of $\hat{G}$,

$$\hat{\theta}^{*(\alpha)} = \hat{G}^{-1}(\alpha). \tag{1.34}$$

Using the formulas above, the *percentile confidence interval* (or percentile interval) can be derived.

Theorem 1.5.1. *The* $100\,(1-\alpha)\%$ ***percentile confidence interval*** *is given by*

$$\left[\hat{\theta}^{*(\alpha/2)},\, \hat{\theta}^{*(1-\alpha/2)}\right] = \left[\hat{G}^{-1}(\alpha/2),\, \hat{G}^{-1}(1-\alpha/2)\right]. \tag{1.35}$$

A variation of the percentile interval is called the *reverse percentile interval* or *"basic bootstrap confidence limits"* (Davison and Hinkley, 1997, p. 29). The starting point for the reverse percentile is that the behaviour of $\theta - \hat{\theta}$ can be approximated by the behaviour of $\hat{\theta} - \hat{\theta}^*$. It leads to the following formula:

Theorem 1.5.2. *The* $100\,(1-\alpha)\%$ ***reverse percentile confidence interval*** *is provided by*

$$\left[2\hat{\theta}-\hat{\theta}^{*(1-\alpha/2)},\,2\hat{\theta}-\hat{\theta}^{*(\alpha/2)}\right]=\left[2\hat{\theta}-\hat{G}^{-1}(1-\alpha/2),\,2\hat{\theta}-\hat{G}^{-1}(\alpha/2)\right] \tag{1.36}$$

where $\hat{\theta}=T(\hat{F}_n)$ *from the Sample World (see 1.21).*

"This is the mirror image of the bootstrap percentile interval; it reaches as far as above the $\hat{\theta}$ as the percentile interval reaches below" (Hesterberg, 2015, p. 55).

Even though both methods are intuitive for constructing a confidence interval, it turns out that they do not perform well on small-sample problems because of the narrowness bias stemming from the narrow bootstrap sampling distribution (Hesterberg, 2011). The reverse percentile method also carries the disadvantages of being asymmetrical in the wrong direction for skewed data and nonlinear transformations. However, the percentile interval technique is a reasonable choice in the absence of information about asymmetrical bootstrap sampling distributions (Hesterberg, 2015). Fortunately, there is a method called **bootknife resampling** to circumvent the narrow bootstrap distribution and narrowness bias. For drawing a bootstrap sample, leave out one observation randomly and then *sample with replacement* a sample size of n from the remaining $n-1$ observations. (Hesterberg, 2004). As a result, bootknife resampling may give an extra amount of variability for the parameter of interest in case of small sample sizes.

Bias-Corrected and accelerated (BCa) Confidence Interval

The construction of a BCa confidence interval is introduced by Efron (1987). The BCa interval depends on two constants, **bias-correction** and **acceleration** factors denoted by $\hat{z}_0$ and $\hat{a}$, respectively. The endpoints of a BCa confidence interval are provided by the percentiles of the bootstrap sampling distribution. However, they are not necessarily the same as the percentile confidence interval (see 1.35).

Definition 1.5.9. The **bias-correction** factor, $\hat{z}_0$ is the proportion of bootstrap replications less than the original estimate $\hat{\theta}$

$$\hat{z}_0=\Phi^{-1}\left(\frac{\#\{\hat{\theta}_b^*<\hat{\theta}\}}{B}\right) \tag{1.37}$$

where Φ^{-1} is the inverse function of the standard normal CDF.

The aim of $\hat{z}_0$ is to measure the difference between the median of $\hat{\theta}^*$ and $\hat{\theta}$ in normal units (Efron and Tibshirani, 1994).

Definition 1.5.10. The **acceleration** factor, $\hat{a}$, is

$$\hat{a} = \frac{1}{6} \frac{\sum_{i=1}^{n} \left(\hat{\theta}^*_{(\cdot)} - \hat{\theta}^*_{(-i)} \right)^3}{\left[\sum_{i=1}^{n} \left(\hat{\theta}^*_{(\cdot)} - \hat{\theta}^*_{(-i)} \right)^2 \right]^{3/2}} \tag{1.38}$$

where $\hat{\theta}^*_{(-i)}$ and $\hat{\theta}^*_{(\cdot)}$ are defined by (1.11) and (1.12), respectively, using the *jackknife principle.*

The acceleration factor $\hat{a}$ depicts the normalized rate of change of the standard error of $\hat{\theta}$ with respect to the population parameter θ. This acceleration value is meant to correct the possibly unrealistic assumption of standard normal approximation $\hat{\theta} \sim N(\theta, \sigma_{\hat{\theta}})$ that $\sigma_{\hat{\theta}}$ is the same for all θ (Efron and Tibshirani, 1994). Roughly speaking, the acceleration constant $\hat{a}$ corrects for skewness in the bootstrap sampling distribution using the *jackknife estimation.*

Using the components defined above, the BCa confidence interval can be constructed.

Theorem 1.5.3. *The* $100\,(1-\alpha)\%$ ***BCa confidence interval*** *is given by*

$$\left[\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)} \right] = \left[\hat{G}^{-1}(\alpha_1), \hat{G}^{-1}(\alpha_2) \right] \tag{1.39}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}\big(\hat{z}_0 + z_{\alpha/2}\big)} \right)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}\big(\hat{z}_0 + z_{1-\alpha/2}\big)} \right), \tag{1.40}$$

and z_α *is the* 100α*th percentile point of the standard normal distribution.*

It has been proven that the BCa confidence interval has all the desirable properties. It is second-order accurate, transformation-invariant, and can control for skewness. It can be deduced that standard, percentile and BC (bias-corrected) confidence intervals are special cases of the BCa confidence interval. By setting *acceleration factor* $\hat{a}$ *to* 0, BC, and by further setting *bias-correction factor* $\hat{z}_0$ *to* 0, percentile confidence intervals are obtained from the BCa. If the bootstrap CDF $\hat{G}$ is normal, the percentile method reduces to the standard interval.

Methods for Speeding Up BCa Methods

Generally, it is recommended to use the BCa to construct a bootstrap confidence interval for a nonparametric task because it contains all the desirable attributes discussed above. Nonetheless, a BCa can be slower for large sample sizes due to the evaluation of the acceleration factor $\hat{a}$. There are two methods to overcome this issue which can speed up the construction of a BCa confidence interval. These techniques are demonstrated through an example.

Example 1.5.1. Suppose there are $10,000$ balls with some weights. It is possible to have balls with the same weight. The task is to determine the 95% confidence interval for the median ball weight. Assume that the DGP is unknown for the weights of the balls, therefore, this is a nonparametric problem. After applying Algorithm 2, a bootstrap sampling distribution is attained for the median of ball weights. Using the original BCa methods, the **acceleration constant** $\hat{a}$ is achieved by performing $10,000$ jackknife sampling and evaluation rounds in line with Algorithm 1. Sampling and evaluating $10,000$ times is a resource-intensive task. The two methods to make BCa intervals more computationally efficient are

- **Inner Group Jackknife**: Efron and Narasimhan (2020) proposed a **grouping** technique for this "inner jackknife" for calculating the acceleration factor $\hat{a}$ in BCa. Use 50 groups of 200 balls each in which the jackknife sample size would be reduced to 50 from $10,000$. Calculating sample medians for these 50 groups and then applying the jackknife algorithm to calculate $\hat{a}$ (see Algorithm 1).
- **Inner Random Subset Jackknife**: Choose a certain proportion of $10,000$ balls randomly to apply the jackknife algorithm. For example, choosing 10% as a proportion parameter is equivalent to drawing 1000 balls out $10,000$ **randomly** and then compute acceleration value $\hat{a}$.

Conjecture 1.5.4. *The inner group jackknife and inner random subset jackknife methods are asymptotically second-order accurate.*

Tim Hesterberg and I conjecture that, as long as the number of groups (for inner group jackknife) and the number of subsets (for inner random subset jackknife) used in calculating $\hat{a}$ go to infinity as the sample size n goes to infinity, the resulting BCa intervals are still going to be second-order accurate. For the inner random subset jackknife, estimating $\hat{a}$ with some random error would not significantly affect limits. The reason why both methods may work is because $\hat{a}$ constant does not need to be very accurate for large sample sizes.

Bootstrap-t Interval

The backbone of the bootstrap-t method comes from the t-statistic devised by William Gosset. Whenever the sampling distribution of a sample statistic $\hat{\theta}$ is a normal distribution with *unknown* variance, the standardized random variable t can be computed,

$$t = \frac{\hat{\theta} - \theta}{\widehat{se}} \sim t_{n-1} \tag{1.41}$$

where $\widehat{se}^2$ is an unbiased estimate of the population σ^2. The t random variable is distributed as *Student's* t with $n-1$ degrees of freedom. To this end, the following confidence interval can be constructed

$$\hat{\theta} \pm t_{n-1,\,\alpha/2} \cdot \widehat{se}\,. \tag{1.42}$$

The Student's t distribution does not consider the population's skewness or other errors. However, the bootstrap-t interval does adjust for these aspects, and unlike Student's t, it does not assume normality for the parameter of interest.

After generating bootstrap replications according to (1.32), the corresponding standard error $\widehat{se}^*_b$ can be calculated. For each $(\hat{\theta}^*_b,\, \widehat{se}^*_b)$ pair, the following statistic can be computed

$$t^*_b = \frac{\hat{\theta}^*_b - \hat{\theta}}{\widehat{se}^*_b} \tag{1.43}$$

where $\hat{\theta}$ plays the role of θ in the Bootstrap World (see 1.21). Collect these t^*_b $(b = 1, \ldots, B)$ into an array called $\tilde{t}$. The bootstrap-t confidence interval can be achieved after knowing the $(\alpha/2)$th percentile in $\tilde{t}$.

Theorem 1.5.5. *The* $100\,(1-\alpha)\%$ ***bootstrap-t confidence interval*** *is provided by*

$$\hat{\theta} \pm \tilde{t}_{\alpha/2} \cdot \widehat{se}. \tag{1.44}$$

It has been shown that the bootstrap-t (or studentized bootstrap) confidence interval is second-order accurate and can correct for skewness. However, the most significant disadvantage of this method is that it is *not* transformation-invariant (Efron and Tibshirani, 1994).

Chapter 2

Linear Regression

This chapter focuses on the multivariate linear regression model with the ordinary least squares (OLS) estimation method. It is one of the most widely used techniques in econometrics because of its explanatory and multi-purpose capabilities. Firstly, the general linear regression model specification is described. In order to derive the most important finite-sample properties of the OLS estimator, the Gauss-Markov theorem is stated with its assumptions. Then, the properties of the OLS estimator are detailed. Both normality and asymptotic properties of variance and confidence interval of the OLS estimator are described. The final topic of interest of this chapter is heteroscedasticity. Heteroscedastic error means that the error terms' standard deviations are not constant. The challenge is detecting and, if detected, correcting them to make them consistent when computing the variance of the OLS estimator. This chapter is mainly based on Buteikis (2020, chap. 4) and Hansen (2022a, chap. 4).

2.1 Model Specification

The multivariable regression model with n observations and d independent variables, where $(d+1) < n$, is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_d X_{id} + \epsilon_i \tag{2.1}$$

where $i = 1, \ldots, n$.
In matrix formula:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1d} \\ 1 & X_{21} & \cdots & X_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nd} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} . \tag{2.2}$$

In a more dense form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \, . \tag{2.3}$$

The notation is as follows:

- $\boldsymbol{Y}$ – An $n \times 1$ column-vector where $\boldsymbol{Y}_i$ is the i^{th} element of $\boldsymbol{Y}$.

- $\boldsymbol{X}$ – An $n \times (d+1)$ matrix (including the constant $\mathbf{1}$). Also called a *design matrix*, *regressor matrix* or *covariate matrix*. It contains all the explanatory variables.
- $\boldsymbol{X}_{\cdot j}$ – An $n \times 1$ j^{th} regressor column-vector of $\boldsymbol{X}$ where $j = 0, \dots d$. $\boldsymbol{X}_{\cdot 0}$ is the constant $\mathbf{1}$.
- $\boldsymbol{X}_{i\cdot}$ – A $(d+1) \times 1$ i^{th} observation column-vector of $\boldsymbol{X}$ where $i = 1, \dots n$.
- $\boldsymbol{X}_{ij}$ – The ij^{th} element of $\boldsymbol{X}$.
- $\boldsymbol{\beta}$ – A $(d+1) \times 1$ parameter/coefficient column-vector where β_0 is the coefficient of the constant term, and β_i is the i^{th} coefficient term.
- $\boldsymbol{\epsilon}$ – An $n \times 1$ error column-vector where ϵ_i is the i^{th} error term.

2.2 Model Assumptions

There are five required and one (partially) optional assumption. They are called Gauss-Markov assumptions.

OLSA.1: Observations are Mutually Independent

The random variables $(\boldsymbol{Y}_i, \boldsymbol{X}_{i\cdot})$ for $i = 1, \dots, n$ are independent and identically distributed.

To paraphrase, whenever taking any two observations $j \neq k$ in a sample, the $(\boldsymbol{Y}_j, \boldsymbol{X}_{j\cdot})$ are independent of the $(\boldsymbol{Y}_k, \boldsymbol{X}_{k\cdot})$ despite having the same distribution. This assumption might be violated if there are some dependence/connections between a group of observations in the sample, such as being classmates at a school (in a setting where observations are students). This notion is called **clustered dependence** which assumes that the observations are grouped into clusters. Hansen (2022a, sec. 4.21) provides more details on the topic.

OLSA.2: Linear Model DGP

The underlying *Data Generating Process* (DGP) for the population is **linear** and exhibits **strict exogeneity**,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.4}$$

$$\mathbb{E}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \mathbf{0}, \tag{2.5}$$

respectively.

Regardless of how dependent and independent variables are transformed, the model must be linear in parameters. The implications of this assumption are

$$\mathbb{E}(\boldsymbol{\epsilon}) = \mathbb{E}\big(\mathbb{E}(\boldsymbol{\epsilon}|\boldsymbol{X})\big) = \mathbf{0} \tag{2.6}$$

$$Cov\,(\boldsymbol{\epsilon}, \boldsymbol{X}) = \mathbf{0} \tag{2.7}$$

$$\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}\,. \tag{2.8}$$

OLSA.3: Conditional Homoscedasticity

The conditional variances of every error term are the same,

$$\mathbb{V}\,(\epsilon_i\,|\,\boldsymbol{X}) = \mathbb{E}\big(\epsilon_i^2\,|\,\boldsymbol{X}\big) = \sigma^2(\boldsymbol{X}) = \sigma^2 \quad i = 1, \ldots, n. \tag{2.9}$$

The general error covariance matrix conditional on $\boldsymbol{X}$ is given by

$$\mathbb{V}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \begin{bmatrix} \mathbb{V}(\epsilon_1\,|\,\boldsymbol{X}) & Cov\,(\epsilon_1,\epsilon_2\,|\,\boldsymbol{X}) & \cdots & Cov\,(\epsilon_1,\epsilon_n\,|\,\boldsymbol{X}) \\ Cov\,(\epsilon_2,\epsilon_1\,|\,\boldsymbol{X}) & \mathbb{V}(\epsilon_2\,|\,\boldsymbol{X}) & \cdots & Cov\,(\epsilon_2,\epsilon_n\,|\,\boldsymbol{X}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\,(\epsilon_n,\epsilon_1\,|\,\boldsymbol{X}) & Cov\,(\epsilon_n,\epsilon_2\,|\,\boldsymbol{X}) & \cdots & \mathbb{V}(\epsilon_n\,|\,\boldsymbol{X}) \end{bmatrix}. \tag{2.10}$$

(2.9) is formed with the implicit assumption of OLSA.2, and therefore, combining with the general form of the error covariance matrix results in the following error covariance matrix given $\boldsymbol{X}$

$$\begin{bmatrix} \sigma^2 & Cov\,(\epsilon_1,\epsilon_2\,|\,\boldsymbol{X}) & \cdots & Cov\,(\epsilon_1,\epsilon_n\,|\,\boldsymbol{X}) \\ Cov\,(\epsilon_2,\epsilon_1\,|\,\boldsymbol{X}) & \sigma^2 & \cdots & Cov\,(\epsilon_2,\epsilon_n\,|\,\boldsymbol{X}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\,(\epsilon_n,\epsilon_1\,|\,\boldsymbol{X}) & Cov\,(\epsilon_n,\epsilon_2\,|\,\boldsymbol{X}) & \cdots & \sigma^2 \end{bmatrix}. \tag{2.11}$$

It turns out that the error covariance matrix needs to be diagonal, leading to another assumption.

OLSA.4: Conditionally Uncorrelated Error Terms

The covariance between different error term pairs, given $\boldsymbol{X}$, is **zero**, that is, all pairwise error terms are uncorrelated,

$$Cov\,\big(\epsilon_i, \epsilon_j|\boldsymbol{X}\big) = 0 \qquad \forall i \neq j. \tag{2.12}$$

This condition states that any error pairs given $\boldsymbol{X}$ are uncorrelated. The implication of this condition is that no cross-correlation exists between error terms conditionally. Define the general error covariance matrix in (2.10) to be

$$\boldsymbol{\Sigma} := \mathbb{V}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \mathbb{E}\big(\,\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\,\big) \tag{2.13}$$

where the i^{th} diagonal element of $\boldsymbol{\Sigma}$ from the OLSA.2 assumption is

$$\mathbb{V}\left(\epsilon_i \,|\, \boldsymbol{X}\right) = \mathbb{E}\big(\epsilon_i^2 \,|\, \boldsymbol{X}\big) = \mathbb{E}\big(\epsilon_i^2 \,|\, \boldsymbol{X}_{i,\cdot}\big) = \sigma_i^2 \,. \tag{2.14}$$

The off-diagonal element ij^{th} of $\boldsymbol{\Sigma}$ is

$$Cov\left(\epsilon_i, \epsilon_j | \boldsymbol{X}\right) = \mathbb{E}\left(\epsilon_i \epsilon_j \,|\, \boldsymbol{X}\right) = \mathbb{E}\left(\epsilon_i \,|\, \boldsymbol{X}\right) \, \mathbb{E}\left(\epsilon_j \,|\, \boldsymbol{X}\right) = 0 \tag{2.15}$$

from the OLSA.4 assumption. As a result, $\boldsymbol{\Sigma}$ is a diagonal matrix with i^{th} diagonal element σ_i^2,

$$\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} . \tag{2.16}$$

In the special case of conditional homoscedasticity (OLSA.3), where $\mathbb{E}\big(\epsilon_i^2 \,|\, \boldsymbol{X}\big) = \sigma_i^2(\boldsymbol{X}) = \sigma^2$, $\boldsymbol{\Sigma}$ becomes

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} =: \sigma^2 \, \boldsymbol{I_n} \tag{2.17}$$

where $\boldsymbol{I_n}$ is the $n \times n$ identity matrix.

OLSA.5: No Exact Collinearity Between Regressors

Regressor vectors $\boldsymbol{X}_{\cdot j}$ $(j = 0, \ldots, d)$ are **linearly independent**. That is,

$$\beta_0 \mathbf{1} + \beta_1 \boldsymbol{X}_{\cdot 1} + \ldots + \beta_d \boldsymbol{X}_{\cdot d} = 0 \tag{2.18}$$

if and only if

$$\beta_0 = \ldots = \beta_d = 0 \,. \tag{2.19}$$

In other words, $\boldsymbol{X}$ is a *full rank* matrix where $\operatorname{rank}(\boldsymbol{X}) = d + 1$. However, the most practical way of describing the assumption is

$$\det\Big(\boldsymbol{X}^T \boldsymbol{X}\Big) \neq 0, \tag{2.20}$$

that is, $(\boldsymbol{X}^T \boldsymbol{X})$ is a *nonsingular* matrix.

OLSA.6: Conditionally Normally Distributed Error

$$\boldsymbol{\epsilon} | \boldsymbol{X} \sim N\Big(\mathbf{0}, \sigma^2 \boldsymbol{I_n}\Big) \,. \tag{2.21}$$

The normality assumption is required for small sample cases. However, in large sample cases due to LLN, the violation of this condition has marginal consequences.
It also implies that

$$\boldsymbol{Y}|\boldsymbol{X} \sim N\Big(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I_n}\Big)\,. \tag{2.22}$$

$\boldsymbol{Y}$ is normally distributed given the $\boldsymbol{X}$.

2.3 OLS Estimation of Regression Parameters

Theorem 2.3.1. *The **OLS estimator**, given* OLSA.1 – OLSA.5 *hold true, is*

$$\hat{\boldsymbol{\beta}} = \Big(\boldsymbol{X}^T\boldsymbol{X}\Big)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,. \tag{2.23}$$

Proof. Let $\boldsymbol{X}$ be an $n \times (d+1)$ matrix with rank $d+1$ where $d+1 < n$. From OLSA.2 assuming $\boldsymbol{\epsilon} \neq \boldsymbol{0}$, let $\boldsymbol{Y}$ be an $n \times 1$ vector such that $\boldsymbol{Y} \notin \mathrm{Im}(\boldsymbol{X})$, that is, $\boldsymbol{Y} = \boldsymbol{X\beta}$ has no $\boldsymbol{\beta}$ solution. The aim is to find a $\boldsymbol{\beta}$ such that $\boldsymbol{X\beta}$ is as close as possible to $\boldsymbol{Y}$ as measured by the square of the Euclidean norm $\|\boldsymbol{X\beta} - \boldsymbol{Y}\|_2^2$,

$$\begin{aligned}\|\boldsymbol{X\beta} - \boldsymbol{Y}\|_2^2 &= (\boldsymbol{X\beta} - \boldsymbol{Y})^T(\boldsymbol{X\beta} - \boldsymbol{Y})\\ &= \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta} - 2\boldsymbol{Y}^T\boldsymbol{X\beta} + \boldsymbol{Y}^T\boldsymbol{Y}\,.\end{aligned}$$

Take the gradient with respect to $\boldsymbol{\beta}$ and set the gradient to $\boldsymbol{0}$,

$$\begin{aligned}\boldsymbol{0} &= \nabla_{\boldsymbol{\beta}}\Big(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta} - 2\boldsymbol{Y}^T\boldsymbol{X\beta} + \boldsymbol{Y}^T\boldsymbol{Y}\Big)\\ &= 2\boldsymbol{X}^T\boldsymbol{X\beta} - 2\boldsymbol{X}^T\boldsymbol{Y}\,.\end{aligned}$$

Rearranging the last equation while dividing by two yields

$$\boldsymbol{X}^T\boldsymbol{X\beta} = \boldsymbol{X}^T\boldsymbol{Y}\,.$$

Left multiplication of both sides by $\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}$ provides the OLS estimator of interest

$$\hat{\boldsymbol{\beta}} = \Big(\boldsymbol{X}^T\boldsymbol{X}\Big)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,.$$

□

Remark.

- *Projection*
  If the OLS estimator is multiplied from left (2.23) by $\boldsymbol{X}$, $n \times 1$ *projection vector* $\boldsymbol{p}$ is received,

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{p} = \boldsymbol{X}\Big(\boldsymbol{X}^T\boldsymbol{X}\Big)^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \tag{2.24}$$

which projects $\boldsymbol{Y}$ onto the subspace of $\boldsymbol{p}$, happening to be the column space of $\boldsymbol{X}$. The $n \times n$ projection matrix $\boldsymbol{P}$ is

$$\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T. \tag{2.25}$$

The significance of the projection is that it gives a geometric meaning to the OLS estimation by projecting the $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$, $\mathrm{Im}(\boldsymbol{X})$. It uses the fact that the residual $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is *perpendicular/orthogonal* to the subspace when the projection takes place. Using the orthogonality yields the following **normal equation**

$$\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0. \tag{2.26}$$

Rearranging for the $\hat{\boldsymbol{\beta}}$ yields the OLS estimator. Strang (2016, sec. 4.2, 4.3) provides more detailed information about the geometric interpretation of the OLS estimation.

- *Relaxing OLSA.5 Assumption*
  If linear independence between regressor vectors $\boldsymbol{X}_{\cdot j}$ $(j = 0, \ldots, d)$ is not required, the matrix inversion does not work in the OLS estimation formula in (2.23) because $\boldsymbol{X}^T\boldsymbol{X}$ becomes a singular matrix; hence its determinant is equal to 0. One of the ways to circumvent the problem is to use the **Moore-Penrose inverse** or **pseudoinverse**, which relies on the *singular value decomposition*. Even though the $\hat{\boldsymbol{\beta}}$ can be calculated with the pseudoinverse's help, the residual terms become considerable in their magnitude.

Corollary 2.3.1.1. *Using the OLS estimator from (2.23), the* ***OLS prediction*** *is*

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{2.27}$$

where $\hat{\boldsymbol{Y}}$ *is the* ***fitted value****.*

The multivariable regression model defined in (2.1) is usually unknown. That is, the true systematic component $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$ (2.8) and the true random error $\boldsymbol{\epsilon}$ cannot be observed. However, the OLS can be used to estimate unknown coefficients $\boldsymbol{\beta}$ and target variable $\boldsymbol{Y}$. So the regression can be written as

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \hat{\boldsymbol{\epsilon}} \tag{2.28}$$

where $\hat{\boldsymbol{\epsilon}}$ is the estimated residual. Equivalently,

$$\boldsymbol{Y}_i = \hat{\boldsymbol{Y}}_i + \hat{\epsilon}_i \qquad i = 1, \ldots, n. \tag{2.29}$$

Hence, the OLS can be considered as a decomposition of each $\boldsymbol{Y}_i$ into two parts; a fitted value $\hat{\boldsymbol{Y}}_i$ and an estimated residual term $\hat{\epsilon}_i$.

Theorem 2.3.2 (Gauss-Markov). *If the conditions* OLSA.1 – OLSA.5 *hold true, the OLS estimator* $\hat{\boldsymbol{\beta}}$ *is the* ***Best Linear Unbiased Estimator*** *(BLUE). It is also consistent with the true parameter values of the multivariate linear regression model.*

2.4 Variance of the OLS Estimator

Consider the general error covariance formula, $\boldsymbol{\Sigma}$, defined in (2.10). For any $n \times (d+1)$ matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{X})$,

$$\mathbb{V}\big(\boldsymbol{A}^T\boldsymbol{Y} \,|\, \boldsymbol{X}\big) = \mathbb{V}\big(\boldsymbol{A}^T\boldsymbol{\epsilon} \,|\, \boldsymbol{X}\big) = \boldsymbol{A}^T\boldsymbol{\Sigma}\boldsymbol{A}. \tag{2.30}$$

Plugging $\boldsymbol{X}\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}$ into $\boldsymbol{A}$ leads to

$$\mathbb{V}_{\hat{\boldsymbol{\beta}}} := \mathbb{V}\left(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{X}\right) = \big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X}\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1} \tag{2.31}$$

where $\hat{\boldsymbol{\beta}} = \boldsymbol{A}^T\boldsymbol{Y}$. $\mathbb{V}_{\hat{\boldsymbol{\beta}}}$ is the (general) variance of the OLS estimator. The formula for the $\mathbb{V}_{\hat{\boldsymbol{\beta}}}$ in (2.31) is also known as the **sandwich estimator**. The name came from the idea that $\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X}$ is the *filling* between two matrices $\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}$. It is also worth noting that the filling

$$\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X} = \sum_{i=1}^{n}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\sigma_i^2 \tag{2.32}$$

is a weighted version of $\big(\boldsymbol{X}^T\boldsymbol{X}\big)$. In the case of conditional homoscedasticity (OLSA.3), the covariance matrix $\mathbb{V}_{\hat{\boldsymbol{\beta}}}$ reduces to

$$\mathbb{V}_{\hat{\boldsymbol{\beta}}} = \sigma^2\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1} \tag{2.33}$$

because $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I_n}$, and hence $\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{X}\sigma^2$. To sum up the derivation above, the following theorem can be stated.

Theorem 2.4.1 (Variance of the OLS estimator). *Whenever a linear model DGP (OLSA.2) with i.i.d. observations (OLSA.1) holds true*

$$\mathbb{V}_{\hat{\boldsymbol{\beta}}} := \mathbb{V}\left(\hat{\boldsymbol{\beta}} \,|\, \boldsymbol{X}\right) = \big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X}\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1} \tag{2.34}$$

where $\boldsymbol{\Sigma}$ *is defined in (2.10). Furthermore, if the error is homoscedastic (OLSA.3), the variance of the OLS estimator simplifies to*

$$\mathbb{V}_{\hat{\boldsymbol{\beta}}} = \sigma^2\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}. \tag{2.35}$$

Based on the above mentioned properties of the variance of the OLS estimator, the *modern* version of the Gauss-Markov theorem can be stated (Hansen, 2022b).

Theorem 2.4.2 (Modern Gauss-Markov). *If* $\hat{\boldsymbol{\beta}}$ *is an unbiased estimator of* $\boldsymbol{\beta}$ *then*

$$\sigma^2\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1} \leq \mathbb{V}_{\hat{\boldsymbol{\beta}}}. \tag{2.36}$$

In other words, the general sandwich estimator defined in (2.31) reaches its minimum value at its homoscedastic value as long as the OLS estimator is unbiased.

Generally, the $\mathbf{\Sigma}$, and in the case of the homoscedastic error σ^2, are unknown. These quantities are necessary to gauge with estimators that have the potential to be unbiased and consistent. $\hat{\sigma}^2$ can be used for estimating σ^2 in the following way.
The **estimated residual** can be obtained from (2.29), the difference between the actual and the fitted values

$$\hat{\boldsymbol{\epsilon}}^T = [\hat{\epsilon}_1, \dots, \hat{\epsilon}_n]^T = \left[\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right]^T . \tag{2.37}$$

Using this information, $\hat{\sigma}^2$ can be constructed.

Theorem 2.4.3. *The estimator of σ^2 is*

$$\hat{\sigma}^2 = \frac{1}{n-(d+1)} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}{n-(d+1)} , \tag{2.38}$$

where $d+1$ is the number of the parameters.

It turns out that $\hat{\sigma}^2$ is an unbiased and *consistent* estimator of σ^2 (see Greene, 2017a, sec. 4.4.2). Usually in statistical packages, such as statsmodels (Seabold and Perktold, 2010), the default covariance matrix for $\hat{\boldsymbol{\beta}}$ is $\hat{\sigma}^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$. Nonetheless, it is not the most representative error covariance matrix (see Section 2.6) on real-life data.

2.4.1 Properties of the OLS Variance Estimator

Theorem 2.4.4. *If, additionally to* OLSA.1 - OLSA.5*, the assumption* OLSA.6 *also holds, the conditional distribution of the OLS estimator is normal, that is,*

$$\hat{\boldsymbol{\beta}} \,|\boldsymbol{X} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right) . \tag{2.39}$$

Element-wise, it would be

$$\hat{\beta}_j \,|\boldsymbol{X} \sim N\left(\beta_j, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}_{jj}\right) \tag{2.40}$$

where

$$\mathbb{V}_{\hat{\beta}_j} := \mathbb{V}(\hat{\beta}_j \,|\, \boldsymbol{X}) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}_{jj} . \tag{2.41}$$

It can be demonstrated that OLSA.6 assumption is not necessary when the *asymptotic distribution* of $\hat{\beta}$ with i.i.d. observations is considered. Under suitable assumptions (see Greene, 2017a, sec. 4.4.3), it can be shown using the OLSA.1 assumption that

$$\hat{\boldsymbol{\beta}} \,|\boldsymbol{X} \sim N\left(\boldsymbol{\beta}, \frac{\sigma^2}{n} \boldsymbol{Q}^{-1}\right) \tag{2.42}$$

where

$$\operatorname*{plim}_{n\to\infty} \frac{\boldsymbol{X}^T \boldsymbol{X}}{n} = \boldsymbol{Q} \tag{2.43}$$

is a positive definite matrix. Based on the i.i.d. $(\boldsymbol{Y}_i, \boldsymbol{X}_{i\cdot})$, the alternative way to obtain the result above is to assume

$$\mathbb{E}(\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T) = \boldsymbol{Q}\,. \tag{2.44}$$

By the *LLN*, $\operatorname{plim}_{n\to\infty}(1/n)\sum_i \boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T = \boldsymbol{Q}$ and using the *Theorem D.14* from Greene (2017b), $\operatorname{plim}_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1} = \boldsymbol{Q}^{-1}$. If observations are not identically distributed, for instance, if $\mathbb{E}(\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T) = \boldsymbol{Q}_i$, the **Lindeberg-Feller CLT** (Theorem D.19A) can be applied under suitable, more general assumptions (Greene, 2017a). Essentially, it yields the same result. When working with real-life data, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is used to estimate $(1/n)\boldsymbol{Q}^{-1}$, and σ^2 is approximated by $\hat{\sigma}^2$ from (2.38).

If the assumption OLSA.6 holds, $\boldsymbol{\epsilon}$ is conditionally normally distributed, and therefore, $\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}$ is normally distributed for the asymptotic case. However, even if OLSA.6 is not assumed, the *CLT* applies. Therefore, it leads to the asymptotic normality of the OLS estimator *if the feature variables are well-behaved and observations are independent* (Greene, 2017a).

2.5 The Confidence Interval of the OLS Estimator

2.5.1 Normality Assumption with Known Variance

If error terms are (conditionally) normally distributed as the assumption OLSA.6 states, it is known that $\hat{\beta}_j$ is (conditionally) normally distributed (see 2.40). Standardizing it and using the fact that $\mathbb{E}(\hat{\beta}_j) = \beta_j$ result in the following formula:

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathbb{V}_{\hat{\beta}_j}}} \sim N(0,1) \qquad j = 0,\ldots,d \tag{2.45}$$

where $\mathbb{V}_{\hat{\beta}_j}$ is defined in (2.41). The 95% confidence level is

$$\mathbb{P}(-1.96 \le Z_j \le 1.96) = 0.95\,. \tag{2.46}$$

Substituting and rearranging terms yield

$$\mathbb{P}\left(\hat{\beta}_j - 1.96\sqrt{\mathbb{V}_{\hat{\beta}_j}} \le \beta_j \le \hat{\beta}_j + 1.96\sqrt{\mathbb{V}_{\hat{\beta}_j}}\right) = 0.95 \tag{2.47}$$

or equivalently, the interval estimator is

$$\hat{\beta}_j \pm 1.96\sqrt{\mathbb{V}_{\hat{\beta}_j}}\,. \tag{2.48}$$

2.5.2 Normality Assumption with Unknown Variance

The initial assumptions are the same as in the previous subsection; disturbances are conditionally normally distributed. However, the variance is unknown. In this case, the unknown σ^2 can be estimated by the sample variance $\hat{\sigma}^2$ as shown in (2.38). Nevertheless, the ratio now has a t-distribution with $\big(n - (d+1)\big)$ instead of the standard normal distribution:

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\mathbb{V}}_{\hat{\beta}_j}}} \sim t_{n-(d+1)} \qquad j = 0, \ldots, d \tag{2.49}$$

where

$$\hat{\mathbb{V}}_{\hat{\beta}_j} = \hat{\sigma}^2 \big(\boldsymbol{X}^T \boldsymbol{X}\big)^{-1}_{jj} \,. \tag{2.50}$$

Note that (2.41) and (2.50) are almost the same, except that the latter uses the sample variance instead of the population variance.

So the $100\,(1-\alpha)\%$ confidence interval is

$$\mathbb{P}\left(\hat{\beta}_j - t_{n-(d+1),\,\alpha/2}\sqrt{\hat{\mathbb{V}}_{\hat{\beta}_j}} \le \beta_j \le \hat{\beta}_j + t_{n-(d+1),\,\alpha/2}\sqrt{\hat{\mathbb{V}}_{\hat{\beta}_j}}\right) \\ = 1 - \alpha \tag{2.51}$$

or equivalently, the interval estimator is

$$\hat{\beta}_j \pm t_{n-(d+1),\,\alpha/2}\sqrt{\hat{\mathbb{V}}_{\hat{\beta}_j}} \,. \tag{2.52}$$

2.5.3 Relaxing the Normality Assumption

Two methods that can be used to construct a confidence interval for the OLS estimator when OLSA.6 (normality assumption) is not assumed; the asymptotic (large sample) and bootstrap regression techniques. The latter is discussed in Chapter 3.

Asymptotic Methods

If error terms are not normally distributed, (2.42) can be utilized to obtain the following limiting distribution of the statistic

$$Z_j = \frac{\sqrt{n}\,(\hat{\beta}_j - \beta_j)}{\sqrt{\sigma^2\,\boldsymbol{Q}_{jj}}} \sim N(0,1) \tag{2.53}$$

where $\boldsymbol{Q}_{jj}$ is the j^{th} diagonal element of $\boldsymbol{Q}$ defined in (2.43). σ^2 can be estimated by the consistent $\hat{\sigma}^2$ based on *Theorem D.16* from Greene (2017b), and (2.38). It eventually results in a statistic with the same limiting distribution. Combining it with the approximation of $\boldsymbol{Q}$ with $(\boldsymbol{X}^T\boldsymbol{X}/n)$ leads to (2.49). Under the assumption of asymptotic properties, the *t-statistic* becomes standard normal. That is, the *t-statistic* is asymptotically standard normal **without error terms being normally distributed**. To paraphrase, the $100\,(1-\alpha)\%$ CI can be calculated with (2.51) but by substituting critical t-values for z-values (standard normal values) as the degree of freedom becomes "large enough".

Other Methods

The above-mentioned asymptotic methods can be generalized also for cases when OLSA.3 does not hold, that is, conditional homoscedasticity is not present. In other words, when heteroscedasticity occurs. In Section 2.6, the homoscedastic error covariance matrix is corrected for heteroscedasticity using the Heteroscedasticity-Corrected Covariance Matrices Estimator (HCCME). Bootstrapping linear regression, which is discussed in Chapter 3, provides another option for calculating variance and, therefore, a confidence interval for the OLS estimator. In practice, it is unnecessary to bootstrap if having a relatively large sample size with conditionally uncorrelated error terms (OLSA.4). HCCME (especially the HC3) is robust enough to estimate the covariance matrix and hence construct the appropriate CI based on (2.51) with minor modifications detailed in the last part of the above-mentioned asymptotic methods. However, for a smaller sample size, it could make sense to use bootstrapping even with the presence of OLSA.4 for calculating the variance for the OLS estimator. Furthermore, when OLSA.1, OLSA.3 and OLSA.4 do not hold, which are the characteristics of clustered data, cluster bootstrapping (especially wild cluster bootstrap) is usually the way to go. It is beyond the scope of the thesis, so it is not discussed. Cameron, Gelbach, and Miller (2008), MacKinnon and Webb (2017) and Roodman et al. (2019) provide a more detailed discussion of the cluster bootstrapping method.

2.6 Heteroscedasticity

Suppose that all the model assumptions hold but the conditional homoscedasticity (OLSA.3). In this case, conditional **heteroscedasticity** is present when doing regression. Therefore, (2.17) is not valid anymore,

$$\mathbb{V}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}) = \boldsymbol{\Sigma} \neq \sigma^2\, \boldsymbol{I_n}. \tag{2.54}$$

Instead, the more general (2.16) prevails. It is worth noting that OLS estimator $\hat{\boldsymbol{\beta}}$ remains unbiased even with a biased and inconsistent OLS variance estimator. The question is how to detect heteroscedasticity and correct the covariance matrix so that the OLS variance estimator is unbiased and consistent.

2.6.1 Testing for Heteroscedasticity

Before determining the formula for HCCME, heteroscedasticity detection should be done. There are several methods for testing for heteroscedasticity, but the preferred technique is to use the **White test**. The White test can be used with two statistics to determine heteroscedasticity's presence; **F** or **LM** (Lagrange Multiplier) statistic. However, both methods use **R-squared** as their underlying measure. A more in-depth discussion on **R-squared** can be found in Appendix B.

White Test

Suppose the estimated residual has already been obtained from the OLS (see 2.37). The White test is the generalization of another heteroscedasticity testing technique called the *Breusch-Pagan test*. The Breusch-Pagan test assumes that heteroscedasticity may be a linear function of all the independent variables in the model. However, the White test allows having nonlinear explanatory variables on the estimated squared residual of the OLS serving now as the response variable in the following model (Pedace, 2013):

$$\begin{aligned}\hat{\boldsymbol{\epsilon}} \odot \hat{\boldsymbol{\epsilon}} = \alpha_0 + \alpha_1 \boldsymbol{X}_{\cdot 1} + \ldots + \alpha_d \boldsymbol{X}_{\cdot d} + \alpha_{d+1} \boldsymbol{X}_{\cdot 1} \odot \boldsymbol{X}_{\cdot 1} \\ + \ldots + \alpha_{2d} \boldsymbol{X}_{\cdot d} \odot \boldsymbol{X}_{\cdot d} + \alpha_{2d+1} \left(\boldsymbol{X}_{\cdot 1} \odot \boldsymbol{X}_{\cdot 2}\right) + \ldots + u_i\end{aligned} \tag{2.55}$$

where the u_i is the error of this regression containing nonlinear regressors, namely cross-products and squares $(\boldsymbol{X}_{\cdot i} \odot \boldsymbol{X}_{\cdot j})$ for $i \neq j$. The number of α_i-s involved in the model is $d + d + \big(d \cdot (d-1)\big)/2$. For example, with $d = 5$ independent variables, it would generate 20 regressors. It can be seen that for a not-large d, it would result in many α_i. Instead, a simpler model can be used on $\hat{\boldsymbol{\epsilon}}^2$ for the White test.

$$\hat{\boldsymbol{\epsilon}} \odot \hat{\boldsymbol{\epsilon}} = \delta_0 + \delta_1 \hat{\boldsymbol{Y}} + \delta_2 \hat{\boldsymbol{Y}} \odot \hat{\boldsymbol{Y}} + u_i \tag{2.56}$$

where $\hat{\boldsymbol{Y}}$ is the fitted value from the OLS,

$$\hat{\boldsymbol{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{X}_{\cdot 1} + \ldots + \hat{\beta}_d \boldsymbol{X}_{\cdot d} \,. \tag{2.57}$$

Before detailing the proper algorithm for the White test, define **LM** and **F** statistics for heteroscedasticity.

Remark. $\odot$ in (2.55) and (2.56) is called **element-wise multiplication** or **Hadamard-product**. From now on, element-wise multiplication between two identical arrays or matrices is denoted by squaring them. For instance, $\hat{\boldsymbol{\epsilon}}^2 := \hat{\boldsymbol{\epsilon}} \odot \hat{\boldsymbol{\epsilon}}$.

Definition 2.6.1 (LM statistic)**.** The **LM** statistic for heteroscedasticity is the sample size times the R-squared of $\hat{\boldsymbol{\epsilon}}^2$

$$\text{LM} = n \boldsymbol{R}^{\mathbf{2}}_{\hat{\boldsymbol{\epsilon}}^2} \;\sim\; \chi^2_d \,. \tag{2.58}$$

A regression model with d regressors and $\hat{\boldsymbol{\epsilon}}^2$ as a dependent variable yield an LM statistic distributed as chi-squared distribution with d degrees of freedom.

Definition 2.6.2 (R-squared form of the F statistic)**.** The **R-squared form of the F statistic** for heteroscedasticity can be defined as

$$F = \frac{\boldsymbol{R}^{\mathbf{2}}_{\hat{\boldsymbol{\epsilon}}^2} \,/\, d}{\big(1 - \boldsymbol{R}^{\mathbf{2}}_{\hat{\boldsymbol{\epsilon}}^2}\big) \,/\, \big(n - (d+1)\big)} \;\sim\; F_{d,\, n-d-1} \tag{2.59}$$

where d is the number of regressors, and the regression model is the same for the LM statistic. The F statistic is distributed as F-distribution with d and $n-d-1$ degrees of freedom.

White Test for Heteroscedasticity

1. Estimate the model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ using OLS.
2. Obtain the predicted $\hat{\boldsymbol{Y}}$ and the squared residual $\hat{\boldsymbol{\epsilon}}^2$.
3. Run the OLS in (2.56). Retrieve the R-squared value $\boldsymbol{R}^{\mathbf{2}}_{\hat{\boldsymbol{\epsilon}}^2}$ from this regression.
4. For the following hypothesis test,
$$\begin{cases} H_0: & \delta_1 = \delta_2 = 0 \\ H_1: & \exists i \in \{1,2\} : \delta_i \neq 0\,, \end{cases}$$
apply either the $F \sim F_{2,n-3}$ or LM $\sim \chi^2_2$.
5. Calculate the p-value.

2.6.2 Heteroskedasticity-Consistent Covariance Matrix Estimators

If the residual terms turn out to be heteroskedastic after the White test, the OLS estimate can be used with **White-correction** on $\mathbb{V}_{\hat{\beta}_j}$ from (2.33) (Buteikis, 2020). The general error covariance matrix is

$$\mathbb{V}_{\hat{\boldsymbol{\beta}}} := \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \tag{2.60}$$

where $\boldsymbol{\Sigma}$ is defined in (2.16). The equivalent formula for the unknown $\boldsymbol{\Sigma}$ can be obtained by

$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_1^2, \dots, \sigma_n^2\right) = \mathbb{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \mid \boldsymbol{X}\right) = \mathbb{E}\left(\tilde{\boldsymbol{\Sigma}} \mid \boldsymbol{X}\right) \tag{2.61}$$

where $\tilde{\boldsymbol{\Sigma}} = \operatorname{diag}\left(\epsilon_1^2, \dots, \epsilon_n^2\right)$. Therefore, $\tilde{\boldsymbol{\Sigma}}$ is a conditionally unbiased estimator for $\boldsymbol{\Sigma}$. Assume that the true error terms are known, so the following unbiased estimator for $\mathbb{V}_{\hat{\boldsymbol{\beta}}}$ could be determined (Hansen, 2022a)

$$\begin{aligned} \mathbb{V}^{\text{ideal}}_{\hat{\boldsymbol{\beta}}} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{\Sigma}}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\operatorname{diag}\left(\epsilon_1^2, \dots, \epsilon_n^2\right)\,\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\epsilon_i^2\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}. \end{aligned} \tag{2.62}$$

As the true error terms are not observable, $\mathbb{V}^{\text{ideal}}_{\hat{\boldsymbol{\beta}}}$ is not a viable estimator. However, replacing ϵ_i^2 with other appropriate observable values results in one of the

Heteroskedasticity-Consistent Covariance Matrix Estimators (HCCME)

$$\mathbb{V}_{\hat{\boldsymbol{\beta}}}^{\text{HCCME}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \tag{2.63}$$

where $\hat{\boldsymbol{\Sigma}}$ is estimated by a function of $\hat{\epsilon}_i^2$. All known forms of HCCME are HC0, HC1, HC2, HC3, HC4, HC4m and HC5.

Theorem 2.6.1 (Eicker, 1963; White, 1980)**.** *The **HC0 error covariance matrix** can be gained from (2.62) by substituting the true ϵ_i^2 for the estimated $\hat{\epsilon}_i^2$*

$$\begin{aligned}\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{HC0} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\operatorname{diag}\left(\hat{\epsilon}_i^2,\ldots,\epsilon_n^2\right)\,\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\hat{\epsilon}_i^2\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}.\end{aligned} \tag{2.64}$$

HC0 is originally developed by Eicker (1963). Sometimes HC0 is also called the Eicker-White covariance matrix estimator. However, one of his revolutionary discoveries in White (1980) was

$$\hat{\mathbb{V}}_n \equiv \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\hat{\epsilon}_i^2 \xrightarrow{a.s.} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\epsilon_i^2\right). \tag{2.65}$$

The expression on the rightmost side of (2.65) is an average of n matrix expectations, and each one is unknown and impossible to estimate consistently. Until 1980, even though Eicker (1963), Hinkley (1977) and S. D. Horn, R. A. Horn, and Duncan (1975) were the forerunners of HCO, HC1 and HC2, respectively, it was widely believed by econometricians that it was necessary to estimate each matrix expectation separately to estimate an average of expectations consistently (MacKinnon, 2013). The primary finding of White (1980) was to prove that it is not necessary at all.
The result shown in (2.65) makes it possible to get the asymptotic covariance matrix estimator

$$(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1}\,\hat{\mathbb{V}}_n\,(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1}. \tag{2.66}$$

Moreover, White (1980) showed that the covariance matrix estimator in (2.66) consistently estimates the asymptotic covariance matrix of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$. This result by White (1980) has turned out to be one of the most cited econometrics articles of all time. The covariance matrix estimator (2.64) is the finite-sample version of the (2.66) in which the factors of n have been removed (MacKinnon, 2013). This HCCME became known as HC0, coined by White. HC1 and HC2 were named by MacKinnon and White (1985), although the formulas for them had been invented in the 70s. Regarding HC0, it turns out that $\hat{\epsilon}_i^2$ is biased towards zero (Hansen, 2022a). Fortunately, scaling $\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{\text{HC0}}$ by $n\,/\,\big(n-(d+1)\big)$ reduces the bias.

Theorem 2.6.2 (Hinkley, 1977; MacKinnon and White, 1985)**.** *The* ***HC1 error covariance matrix*** *is given by*

$$
\begin{aligned}
\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{\mathit{HC1}} &= \frac{n}{n-(d+1)} \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \operatorname{diag}\left(\hat{\epsilon}_i^2, \ldots, \epsilon_n^2\right) \, \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \frac{n}{n-(d+1)} \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T \hat{\epsilon}_i^2\right) \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}.
\end{aligned}
\tag{2.67}
$$

Other variants of HC (heteroskedasticity-consistent) estimators are based on the so-called **hat matrix**, $\boldsymbol{H}$.

Definition 2.6.3. The hat matrix, $\boldsymbol{H}$, is defined as

$$
\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T. \tag{2.68}
$$

Generally speaking, the i^{th} diagonal element of $\boldsymbol{H}$, h_{ii} quantifies how far a specific $\boldsymbol{X}_{i,\cdot}$ is from the rest of the values in $\boldsymbol{X}$. The farther it is, the higher the leverage h_{ii} value is. The value of the h_{ii} ranges from 0 to 1. A more detailed discussion on the hat matrix is provided in Appendix C. The rest of the HCCME formulas use the diagonal of the hat matrix, $\boldsymbol{H}$, to scale the observed residual terms, $\hat{\epsilon}_i$, in the sandwich estimator defined in (2.63).

Theorem 2.6.3 (S. D. Horn, R. A. Horn, and Duncan, 1975; MacKinnon and White, 1985)**.** *The* ***HC2 error covariance matrix*** *is provided by*

$$
\begin{aligned}
\mathbb{V}_{\hat{\boldsymbol{\beta}}}^{\mathit{HC2}} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \operatorname{diag}\left(\frac{\hat{\epsilon}_i^2}{1-h_{11}}, \ldots, \frac{\hat{\epsilon}_n^2}{1-h_{nn}}\right) \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \left(\sum_{i=1}^{n} (1-h_{ii})^{-1} \boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T \hat{\epsilon}_i^2\right) \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}.
\end{aligned}
\tag{2.69}
$$

Theorem 2.6.4 (MacKinnon and White, 1985)**.** *The* ***HC3 error covariance matrix*** *is obtained by*

$$
\begin{aligned}
\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{\mathit{HC3}} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \operatorname{diag}\left(\frac{\hat{\epsilon}_i^2}{(1-h_{11})^2}, \ldots, \frac{\hat{\epsilon}_n^2}{(1-h_{nn})^2}\right) \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \left(\sum_{i=1}^{n} (1-h_{ii})^{-2} \boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T \hat{\epsilon}_i^2\right) \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}.
\end{aligned}
\tag{2.70}
$$

HC3 was originally motivated as a simplified version of the jackknife covariance estimator. The formula was derived by James G. MacKinnon and Halbert White in 1985 based on the jackknife principle. They showed that the HC3 is numerically the same as the jackknife variance estimate for the OLS estimator. However, it is more computationally efficient than the jackknife variant. For more details, see Appendix D. HC2 is unbiased (under homoscedasticity), whereas HC3 is robust for any $\boldsymbol{X}$ (Hansen, 2022a). HC3 does not perform poorly under homoscedasticity and outperforms all the HCE-s in the presence of heteroscedasticity (Long and Ervin, 2000). They also suggest using HC3 all of the time. Furthermore, Davidson and MacKinnon (1993) strongly recommend using either HC2 or HC3 in favour of HC0 when computing the variance of the OLS estimator. In addition, the relationship between HC0, HC2 and HC3 is shown by the following theorem.

Corollary 2.6.4.1.

$$\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{HC0} < \hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{HC2} < \hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{HC3}. \tag{2.71}$$

The inequalities arise from the fact that $\left(1 - h_{ii}\right)^{-2} > \left(1 - h_{ii}\right)^{-1} > 1$.

Remark. The matrix relation here means that if a "smaller" error covariance matrix is subtracted from a "larger" one, the operation results in a positive definite matrix.

There are certain forms of heteroscedasticity when HC3 may fail. For instance, when the regressors are from heavy-tailed distributions and the errors are from light-tailed distributions (Buteikis, 2020). HC3 underperforms when there are high leverage points and non-normal errors (Cribari-Neto, 2004). Cribari-Neto proposed a new HCE variant, HC4, which outperforms HC3 in these cases.

Theorem 2.6.5 (Cribari-Neto, 2004)**.** *The* ***HC4 error covariance estimator*** *is provided by*

$$\begin{aligned}\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{HC4} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T \operatorname{diag}\left(\frac{\hat{\epsilon}_i^2}{(1-h_{11})^{\delta_1}}, \dots, \frac{\hat{\epsilon}_n^2}{(1-h_{nn})^{\delta_n}}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-\delta_i}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\hat{\epsilon}_i^2\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\end{aligned} \tag{2.72}$$

where

$$\delta_i = min\left\{4, \frac{nh_{ii}}{d+1}\right\}. \tag{2.73}$$

Cribari-Neto and Bernardino (2011) found that HC4 performance deteriorates when increasing the number of regressors and the maximal leverage point, h_{ii}, is extreme. They suggested using HC4m, which also turns out to be more reliable than HC4 for inference-based testing under *both* normal and nonnormal random errors.

Theorem 2.6.6 (Cribari-Neto and Bernardino, 2011)**.** *The* ***HC4m error covariance estimator*** *is provided by*

$$\begin{aligned}\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{HC4m} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T \operatorname{diag}\left(\frac{\hat{\epsilon}_i^2}{(1-h_{11})^{\delta_1}}, \dots, \frac{\hat{\epsilon}_n^2}{(1-h_{nn})^{\delta_n}}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-\delta_i}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\hat{\epsilon}_i^2\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\end{aligned} \tag{2.74}$$

where

$$\delta_i = min\left\{\gamma_1, \frac{nh_{ii}}{d+1}\right\} + min\left\{\gamma_2, \frac{nh_{ii}}{d+1}\right\}. \tag{2.75}$$

The suggested values for γ parameters for the best approximations are $\gamma_1 = 1, \gamma_2 = 1.5$.

Cribari-Neto, Tatiene, and Vasconcellos (2007) proposed another HCCME called HC5. It can be viewed as the generalization of HC4.

Theorem 2.6.7 (Cribari-Neto, Tatiene, and Vasconcellos, 2007)**.** *The **HC5 error covariance estimator** is provided by*

$$\begin{aligned}\hat{\mathbb{V}}_{\hat{\boldsymbol{\beta}}}^{HC5} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T \operatorname{diag}\left(\frac{\hat{\epsilon}_i^2}{(1-h_{11})^{\delta_1}}, \dots, \frac{\hat{\epsilon}_n^2}{(1-h_{nn})^{\delta_n}}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}(1-h_{ii})^{-\delta_i}\,\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\hat{\epsilon}_i^2\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\end{aligned} \tag{2.76}$$

where

$$\delta_i = min\left\{\frac{nh_{ii}}{d+1}, max\left\{4, \frac{nkh_{max}}{d+1}\right\}\right\}. \tag{2.77}$$

with $h_{max} = max\left\{h_{11}, \dots, h_{nn}\right\}$ is the maximal leverage and $0 \leq k \leq 1$ pre-defined constant. The suggested value for the constant parameter k is 0.7.

It can be seen that if $k = 0$, HC5 reduces to HC4. According to the authors, HC5 is the first HCCME in which all squared residuals are discounted more heavily as the maximal leverage increases.

Using any variants of HCCME means trading off efficiency for consistency (Buteikis, 2020). Any specifications of HCCME result in greater variability than the model-based homoscedastic one due to the Modern Gauss-Markov Theorem (Theorem 2.4.2). Nonetheless, they require a minimal set of assumptions on random error terms compared to conditional homoscedasticity assumption OLSA.3.

Chapter 3

Linear Regression Bootstrapping

Consider the following linear model satisfying OLSA.1 - OLSA.5 except for the OLSA.3 (conditional homoscedasticity)

$$\boldsymbol{Y}_i = \boldsymbol{X}_{i\cdot}^T \boldsymbol{\beta} + \epsilon_i, \quad \mathbb{E}(\epsilon_i | \boldsymbol{X}_{i\cdot}) = 0, \quad \mathbb{E}(\epsilon_i \epsilon_j | \boldsymbol{X}) = 0 \ \forall i \neq j. \tag{3.1}$$

There are three methods to apply bootstrap DGP for the model (3.1): residual, pairs, and wild bootstrap. The residual bootstrap requires an even stronger assumption than any of the Gauss-Markov assumptions. All of these methods can be applied to nonlinear regression models with minor changes (MacKinnon, 2006).

3.1 Residual Bootstrap

If error terms are i.i.d. with common variance σ^2, a very accurate inference can be achieved using the residual bootstrap (MacKinnon, 2006). The i.i.d. assumption is even stronger than any of the Gauss-Markov assumptions and implies homoscedasticity. To use the residual bootstrap, first, compute the OLS estimator $\hat{\boldsymbol{\beta}}$ and residual $\hat{\boldsymbol{\epsilon}}$. Then using the residual $\hat{\boldsymbol{\epsilon}}$, the residual bootstrap DGP can be formulated as follows

$$\boldsymbol{Y}_i^* = \boldsymbol{X}_{i\cdot}^T \hat{\boldsymbol{\beta}} + f(\hat{\epsilon}_i^*) \quad f(\hat{\epsilon}_i^*) \sim \text{eCDF}(\hat{\epsilon}_i) \tag{3.2}$$

where $f(\hat{\epsilon}_i)$ is a rescaled version of the i^{th} residual term, $\hat{\epsilon}_i$ $(i = 1, \ldots, n)$. The bootstrap errors $\hat{\epsilon}_i^*$ are drawn from the empirical CDF (eCDF) of $\hat{\epsilon}_i$. That is, each $\hat{\epsilon}_i^*$ is assigned a probability $1/n$ by the eCDF in the following way

$$\mathbb{P}\big(\hat{\epsilon}_i^* = \hat{\epsilon}_i\big) = \frac{1}{n} \qquad \forall i = 1, \ldots, n. \tag{3.3}$$

Usual choices for rescaling recommended by MacKinnon (2012) are

$$\text{t1}: \quad f(\hat{\epsilon}_i) = \left(\frac{n}{n-(d+1)}\right)^{1/2} \hat{\epsilon}_i \tag{3.4}$$

$$\text{t2}: \quad f(\hat{\epsilon}_i) = \frac{\hat{\epsilon}_i}{(1-h_{ii})^{1/2}} \tag{3.5}$$

$$\text{t3}: \quad f(\hat{\epsilon}_i) = \frac{\hat{\epsilon}_i}{1-h_{ii}} \tag{3.6}$$

where h_{ii} is the leverage defined in (C.4). The t1, t2, and t3 transformations are inspired by the HC1, HC2, and HC3 covariance matrices. Rescaling residuals is desirable to have the correct variance if the quantity to be bootstrapped is not invariant to the variance of the error terms (MacKinnon, 2006). It is worth noting that the design matrix $\boldsymbol{X}$ is kept fixed at their (original) sample value.

After resampling the residuals terms provided in (3.2), a bootstrap sample is made up of

$$[\boldsymbol{Y}^* \ \boldsymbol{X}] = \begin{bmatrix} \boldsymbol{Y}_1^* & \boldsymbol{X}_{1\cdot} \\ \boldsymbol{Y}_2^* & \boldsymbol{X}_{2\cdot} \\ \vdots & \vdots \\ \boldsymbol{Y}_n^* & \boldsymbol{X}_{n\cdot} \end{bmatrix}. \tag{3.7}$$

Given this bootstrap sample, the regression estimator is

$$\hat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*. \tag{3.8}$$

This procedure is repeated B times according to the *bootstrap principle*. As a result, bootstrap sampling distributions can be obtained for each $\hat{\boldsymbol{\beta}}_i$.

In practice, the residual bootstrap is rarely used because of the very strong i.i.d. error term assumption, which means that the residual bootstrap is only helpful if the error terms exhibit homoscedasticity.

3.2 Pairs Bootstrap

The pairs bootstrap method for regression was proposed by Freedman (1981) to accommodate heteroscedasticity. Unlike the residual bootstrap, which resamples residual terms, the pairs bootstrap resamples from the matrix $[\boldsymbol{Y} \ \boldsymbol{X}]$. Technically, the resampling procedure is carried out on the joint eCDF of $[\boldsymbol{Y}_i \ \boldsymbol{X}_{i\cdot}]$ $(i = 1, \ldots, n)$. A typical bootstrap sample under the pairs bootstrap DGP is

$$[\boldsymbol{Y}^* \ \boldsymbol{X}^*] = \begin{bmatrix} \boldsymbol{Y}_1^* & \boldsymbol{X}_{1\cdot}^* \\ \boldsymbol{Y}_2^* & \boldsymbol{X}_{2\cdot}^* \\ \vdots & \vdots \\ \boldsymbol{Y}_n^* & \boldsymbol{X}_{n\cdot}^* \end{bmatrix} \tag{3.9}$$

where $[\boldsymbol{Y}_i^* \ \boldsymbol{X}_{i\cdot}^*] \sim \text{eCDF}\big([\boldsymbol{Y}_i \ \boldsymbol{X}_{i\cdot}]\big)$ in which for each k

$$\mathbb{P}\big(\boldsymbol{Y}_k^* = \boldsymbol{Y}_i, \ \boldsymbol{X}_{k\cdot}^* = \boldsymbol{X}_{i\cdot}\big) = \frac{1}{n} \qquad \forall i = 1, \ldots, n. \tag{3.10}$$

Given a bootstrap sample, the OLS estimator can be attained

$$\hat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^{*T}\boldsymbol{X}^*\right)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}^*. \tag{3.11}$$

The process is repeated B times for the B bootstrap samples to get the bootstrap sampling distributions for each OLS estimate term.

Although the pairs bootstrap in this form can handle heteroscedasticity, it has major drawbacks. One of the disadvantages of using the pairs bootstrap is that it is viable for $\boldsymbol{X}^{*T}\boldsymbol{X}^*$ to be singular in a bootstrap sample. It means that the bootstrap OLS estimator $\hat{\boldsymbol{\beta}}^*$ is not defined. For instance, the probability that a bootstrap sample is made up of one observation n times is $n^{-(n-1)}$, which is a small probability but still has a chance to occur (Hansen, 2022a). There are several possible ways to circumvent this problem. The first way is to calculate the pseudoinverse of $\boldsymbol{X}^{*T}\boldsymbol{X}^*$. Better solutions are based on the following ratio

$$\lambda^* = \frac{\lambda_{\min}\left(\boldsymbol{X}^{*T}\boldsymbol{X}^*\right)}{\lambda_{\min}\left(\boldsymbol{X}^T\boldsymbol{X}\right)} \tag{3.12}$$

which expresses the ratio of the smallest eigenvalue of the bootstrap design matrix to that of the data design matrix. Shao and Tu (2012) recommend using a tolerance value τ set to $1/2$ for λ^*. If $\lambda^* < \tau$ in a bootstrap sample, the corresponding bootstrap OLS estimator can be treated as missing. The alternative is to define the following trimming rule using the τ tolerance to ensure the bootstrap OLS estimator will be well-behaved (Hansen, 2022a),

$$\hat{\boldsymbol{\beta}}^* = \begin{cases} \hat{\boldsymbol{\beta}}^* & \text{if } \lambda^* \geq \tau \\ \hat{\boldsymbol{\beta}} & \text{if } \lambda^* < \tau. \end{cases} \tag{3.13}$$

The other deficiency is that, unlike the residual and wild bootstrap, the pairs bootstrap does not condition on $\boldsymbol{X}$. The regressors in the pairs bootstrap are not exogenous because of the nature of drawing from the data matrix itself. Once the $\boldsymbol{X}_{i\cdot}^*$ is quantified, so is $\hat{\epsilon}_i^*$. In other words, $\mathbb{E}(\hat{\epsilon}_i^*|\boldsymbol{X}_{i\cdot}^*) \neq 0$. For example, $\mathbb{E}(\hat{\epsilon}_i^*|\boldsymbol{X}_{i\cdot}^* = \boldsymbol{X}_{i\cdot}) = \sum_{i=1}^{n} \hat{\epsilon}_i \, \mathbb{P}(\hat{\epsilon}_i|\boldsymbol{X}_{i\cdot}^* = \boldsymbol{X}_{i\cdot}) = \hat{\epsilon}_i \neq 0$ (Flachaire, 2005). As a result, the bootstrap principle's assumption is not met because the pairs bootstrap DGP is not close to the true linear model DGP (OLSA.2) where there is a strict exogeneity condition, $\mathbb{E}(\boldsymbol{\epsilon}|\boldsymbol{X}) = \boldsymbol{0}$. These theoretical considerations were reinforced by Monte Carlo simulations showing the inaccuracy of pairs bootstrap methods (Horowitz, 2001). Nonetheless, a modified version of the pairs bootstrap proposed by Flachaire (1999) addresses the issues of exogeneity and some asymptotic refinements related to test statistics detailed in Beran (1988) and Davidson and MacKinnon (1999). The improved version pairs bootstrap DGP is

$$\boldsymbol{Y}_i^* = \boldsymbol{X}_{i\cdot}^{*T}\hat{\boldsymbol{\beta}} + f(\hat{\epsilon}_i^*) \tag{3.14}$$

in which the bootstrap sample $[\boldsymbol{Y}^* \; \boldsymbol{X}^*]$ is generated by resampling first from $[\boldsymbol{X} \; f(\hat{\boldsymbol{\epsilon}})]$, where $f(\hat{\epsilon}_i)$ is a t2 transformation from (3.5). The target $\boldsymbol{Y}^*$ is computed in (3.14).

3.3 Wild Bootstrap

The wild bootstrap was proposed by Wu (1986) and Liu (1988) and further developed by Mammen (1993). It is considered to be the intermediate way between residual and pairs bootstrap because it amalgamates the exogeneity assumption from the residual bootstrap ($\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$) and the ability to control for the heteroscedasticity from the pairs bootstrap (MacKinnon, 2012). Therefore, it is usually the default choice for bootstrap regression tasks with heteroscedastic errors, especially on clustered data.

In order to use the wild bootstrap, first, OLS estimator $\hat{\boldsymbol{\beta}}$ and residual $\hat{\boldsymbol{\epsilon}}$ should be computed. Then using the residual $\hat{\boldsymbol{\epsilon}}$, the wild bootstrap DGP can be defined as

$$\boldsymbol{Y}_i^* = \boldsymbol{X}_{i\cdot}^T \hat{\boldsymbol{\beta}} + \xi_i^* f(\hat{\epsilon}_i) \tag{3.15}$$

where ξ_i^* $(i = 1, \ldots, n)$ is a random variable with usually mean 0, and $f(\hat{\epsilon}_i)$ is one of the rescaled versions of $\hat{\epsilon}_i$ defined in (3.4), (3.5), and (3.6). Hence, usually $E\big(\xi_i^* f(\hat{\epsilon}_i)\big) = 0$, although the expectation of $f(\hat{\epsilon}_i)$ may not. After $\xi_i^* f(\hat{\epsilon}_i)$ residual term becomes known, the process of getting a bootstrap sample for the wild bootstrap is the same as described at the residual bootstrap at this point (see 3.7). After computing the bootstrap OLS estimator (similar to 3.8) from a bootstrap sample, a bootstrap sampling distribution can be computed due to the bootstrap principle.

In an ideal world, rescaled residual terms should have the same moments as bootstrap residual terms (MacKinnon, 2012). It means that the following conditions on ξ_i^* should be met

$$\mathbb{E}\big(\xi_i^*\big) = 0, \quad \mathbb{E}\big(\xi_i^{*2}\big) = 1, \quad \mathbb{E}\big(\xi_i^{*3}\big) = 1, \quad \mathbb{E}\big(\xi_i^{*4}\big) = 1. \tag{3.16}$$

However, it has been shown that this cannot be achieved. For a more detailed discussion on the reason, see MacKinnon (2012, p. 4).

There are various choices for ξ_i^*. One of the most popular is **Mammen's two-point distributions** (Mammen, 1993):

$$\xi_i^* = \begin{cases} -\dfrac{\sqrt{5}-1}{2} & \text{with probability } \dfrac{\sqrt{5}+1}{2\sqrt{5}} \approx 0.7236 \\[2ex] \dfrac{\sqrt{5}+1}{2} & \text{with probability } \dfrac{\sqrt{5}-1}{2\sqrt{5}} \approx 0.2764. \end{cases} \tag{3.17}$$

These seemingly strange values with their corresponding probabilities are the solution for the following system of equations

$$\begin{aligned} p_1\xi_1 + (1-p_1)\,\xi_2 &= 0 \\ p_1\xi_1^2 + (1-p_1)\,\xi_2^2 &= 1 \\ p_1\xi_1^3 + (1-p_1)\,\xi_2^3 &= 1. \end{aligned} \tag{3.18}$$

Even though $E\big(\xi_i^* f(\hat{\epsilon}_i)\big) \neq 0$, the moments for the Mammen distribution are

$$\mathbb{E}\big(\xi_i^*\big) = 0, \quad \mathbb{E}\big(\xi_i^{*2}\big) = 1, \quad \mathbb{E}\big(\xi_i^{*3}\big) = 1, \quad \mathbb{E}\big(\xi_i^{*4}\big) = 2. \tag{3.19}$$

It gets the first three moments right; however, the fourth one is 2 instead of 1. No distribution has the correct third but a fourth moment smaller than 2 (MacKinnon, 2012). Another commonly used distribution is the **Rademacher distribution** proposed by Davidson and Flachaire (2008)

$$\xi_i^* = \begin{cases} -1 & \text{with probability } \ 1/2 \\ 1 & \text{with probability } \ 1/2 \end{cases} \tag{3.20}$$

for which the moments are

$$\mathbb{E}\big(\xi_i^*\big) = 0, \quad \mathbb{E}\big(\xi_i^{*2}\big) = 1, \quad \mathbb{E}\big(\xi_i^{*3}\big) = 0, \quad \mathbb{E}\big(\xi_i^{*4}\big) = 1. \tag{3.21}$$

It has the right fourth moment but the wrong third one. It can be seen from the distribution that it introduces symmetry into residuals. It is shown by Davidson and Flachaire (2008) that even if the error terms are asymmetric, the Rademacher distribution outperforms the Mammen distribution because getting the fourth moment right is more important than getting the third one wrong.
The standard normal is a natural but inferior option for ξ_i^* compared to both Mammen and Rademacher. The standard normal, which has mean 0 and variance 1, has the following moments

$$\mathbb{E}\big(\xi_i^*\big) = 0, \quad \mathbb{E}\big(\xi_i^{*2}\big) = 1, \quad \mathbb{E}\big(\xi_i^{*3}\big) = 0, \quad \mathbb{E}\big(\xi_i^{*4}\big) = 3. \tag{3.22}$$

That is, the standard normal combines the worst of both worlds of Rademacher's and Mammen's because it has the same undesirable third moment as the Rademacher distribution, and it has an even worse fourth moment than the Mammen distribution.

More refined versions of the Rademacher distribution were proposed by Webb (2013) with his 4- and 6-point distributions. The Webb 4-point distribution is defined as

$$\xi_i^* = -\sqrt{\frac{3}{2}};\ -\sqrt{\frac{1}{2}};\ \sqrt{\frac{1}{2}};\ \sqrt{\frac{3}{2}} \quad \text{with probability } 1/4 \text{ each} \tag{3.23}$$

for which

$$\mathbb{E}\big(\xi_i^*\big) = 0, \quad \mathbb{E}\big(\xi_i^{*2}\big) = 1, \quad \mathbb{E}\big(\xi_i^{*3}\big) = 0, \quad \mathbb{E}\big(\xi_i^{*4}\big) = 5/4. \tag{3.24}$$

The Webb 6-point distribution is defined as

$$\xi_i^* = -\sqrt{\frac{3}{2}};\ -\sqrt{\frac{2}{2}};\ -\sqrt{\frac{1}{2}};\ \sqrt{\frac{1}{2}};\ \sqrt{\frac{2}{2}};\ \sqrt{\frac{3}{2}} \quad \text{with probability } 1/6 \text{ each} \tag{3.25}$$

for which

$$\mathbb{E}\big(\xi_i^*\big) = 0, \quad \mathbb{E}\big(\xi_i^{*2}\big) = 1, \quad \mathbb{E}\big(\xi_i^{*3}\big) = 0, \quad \mathbb{E}\big(\xi_i^{*4}\big) = 7/6. \tag{3.26}$$

Both the 4- and 6-point distributions match the Rademacher distribution in terms of the first three moments. However, their fourth moments are smaller. Based on Monte Carlo simulations, Webb (2013) showed that his proposed 4- and 6-point distributions have the edge over the Rademacher distribution when applying to clustered data (when OLSA.1, OLSA.3 and OLSA.4 are not assumed) when the number of clusters is small. However, the clustered bootstrap is beyond the scope of this thesis.

The consensus is that wild bootstrap surpasses the pairs bootstrap based on evidence from Monte Carlo simulations; see Flachaire (2005), MacKinnon (2006), and Davidson and Flachaire (2008).

3.4 Bootstrap Covariance Matrices

The bootstrap standard errors or, more generally, the bootstrap covariance matrices of the OLS estimator $\hat{\boldsymbol{\beta}}$ measure the variation of each estimate $\hat{\beta}_i$ across the B bootstrap samples. No matter what the bootstrap DGP is, the bootstrap covariance matrix is analogous to the formula in (1.25)

$$\widehat{\mathbb{V}}^*(\hat{\boldsymbol{\beta}}) = \frac{1}{B-1}\sum_{j=1}^{B}\left(\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*\right)\left(\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*\right)^T \tag{3.27}$$

where $\bar{\boldsymbol{\beta}}^*$ is the average of the $\hat{\boldsymbol{\beta}}_j^*$ over B bootstrap samples. $\hat{\boldsymbol{\beta}}_j^*$ is the estimate for the $j^{\text{ th}}$ bootstrap sample.

For the residual bootstrap, if n and B become large, the bootstrap covariance matrix for the OLS estimator in (3.27) tends to be

$$\hat{\sigma}^{*2}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}, \tag{3.28}$$

where $\hat{\sigma}^{*2}$ is the average variance of the bootstrap error terms (MacKinnon, 2006). This matrix tends to be the same as the homoscedastic covariance matrix $\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ of $\hat{\boldsymbol{\beta}}$.

For the wild bootstrap, starting from the formula in (3.27), $\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*$ can be ignored if B is large enough because

$$\begin{aligned}\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^* &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\epsilon}}_j^*\right) - \bar{\boldsymbol{\beta}}^* && (3.29)\\ &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\epsilon}}_j^* + \left(\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^*\right) && (3.30)\end{aligned}$$

where $\mathbb{E}(\hat{\boldsymbol{\beta}}_j^*) = \hat{\boldsymbol{\beta}}$ if the OLS estimator is unbiased. $\hat{\boldsymbol{\epsilon}}_j^*$ is the residual estimate for the $j^{\,\text{th}}$ bootstrap sample. Therefore,

$$\widehat{\mathbb{V}}^*_{\text{wild}}(\hat{\boldsymbol{\beta}}) \approx \frac{1}{B-1}\sum_{j=1}^{B}\Big(\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\epsilon}}_j^*\Big)\Big(\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\epsilon}}_j^*\Big)^T \tag{3.31}$$

$$= \frac{1}{B-1}\sum_{j=1}^{B}\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\epsilon}}_j^*\,\hat{\boldsymbol{\epsilon}}_j^{*T}\boldsymbol{X}\big(\boldsymbol{X}^T\boldsymbol{X}\big)^{-1} \tag{3.32}$$

which resembles the HCCME in (2.63). Instead of using a diagonal matrix $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\epsilon}}_j^*\,\hat{\boldsymbol{\epsilon}}_j^{*T}$ is deployed. Because of the second moment, $\mathbb{E}\big(\xi_i^{*2}\big) = 1$, the diagonal element of $\hat{\boldsymbol{\epsilon}}_j^*\,\hat{\boldsymbol{\epsilon}}_j^{*T}$ have an expectation of $f^2(\hat{\epsilon}_i)$ $(i = 1, \ldots, n)$. Owing to $\mathbb{E}\big(\xi_i^*\xi_j^*\big) = 0$, off-diagonal elements have an expectation of zero. As a consequence, the wild bootstrap covariance estimator in (3.32) converges to the HCCME in (2.63) as B becomes large. Applying the rescaling transformation t1, t2, t3 in (3.4), (3.5), and (3.6), respectively, the bootstrap covariance matrix in (3.27) (including the wild bootstrap's one) converges to HC1, HC2, HC3 as $B \to \infty$. Flachaire (2002) reached the same conclusions as the results described above by MacKinnon (2012). A similar argument can be applied to the pairs bootstrap.

In summary, it is not worth using bootstrap regression methods (pairs, residual, wild) to estimate either homoscedasticity or heteroscedasticity covariance matrices because they are more computationally expensive than the conventional HCCMEs.

Chapter 4

Experiments

In this chapter, the primary focus is to discuss how bootstrap distributions and confidence intervals of regression coefficients change in wild bootstrap simulations when increasing the sample size and bootstrap replication numbers. The experiment relies on CEU MicroData's Hungarian balance sheet data in the **commerce sector** from 2014. The core of the experiments uses a Python package called ols_bootstrap that is to be released along with this thesis. A basic tutorial about the package can be found in the corresponding repository[1] along with the simulation itself[2].

4.1 The Model

The linear model is

$$ln\,(w_i/L_i) = \beta_0 + \beta_1\,ln(L_i) + \beta_2 D_i + \epsilon_i \qquad \forall i = 1, \ldots, n \tag{4.1}$$

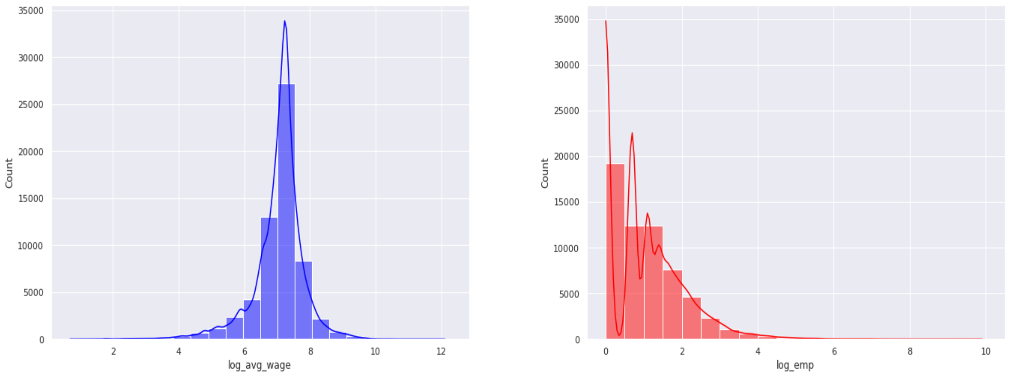
where w_i is the total wage bill, L_i is the number of people employed by firm i, and D_i is a dummy variable indicating whether firm i is a net exporter. $ln\,(w_i/L_i)$ on the left-hand side is the log average wage of the firm i. All Gauss-Markov assumptions but the OLSA.3 (conditional homoscedasticity) are hypothesized. The aim is to use the wild bootstrap defined in (3.15) to examine the behaviour of bootstrap distributions and confidence intervals of the coefficients. Note that experiments are not aimed at examining causal relationships.

4.2 The Dataset

In 2014, there were 99,643 firms in the commerce sector. However, the observations must be reduced because there were firms with 0 or missing employment and wage bills. The final dataset consists of 60,992 observed firms, from which 10,743 were net exporters, and the remaining 50,249 were not. If a firm's export value had been greater than its import, it would have been deemed a net exporter with a label 1. Otherwise, the firm was considered a net importer with a label 0. As mentioned above, D_i indicates the net exporting property of a firm. From now on, it is referred

[1] https://github.com/pvh95/ols_bootstrap/blob/masterthesis2022/demonstration.ipynb

[2] https://github.com/pvh95/ols_bootstrap/blob/masterthesis2022/bca_application.ipynb



(A) Log Average Wage

(B) Log Employment

FIGURE 4.1: The Distribution of Log Average Wage and Employment

to as the is_exporter independent variable. This reduced dataset is treated as the population data, assuming no shadow companies exist.

It is worth mentioning the other independent variable, the **log_employment** ($ln(L_i)$), and the target variable, **log_average_wage** ($ln\left(w_i/L_i\right)$), in brief details. Their distribution plots can be found in Figure 4.1. The log average wage has a normal-like distribution peaking at around the value of 7. Various normality tests and the Q-Q Plot in Figure 4.2 show that the log average wage does not exhibit a normal distribution. As for log employment, it has a heavy-tailed distribution because most enterprises are small and medium-sized (SMEs).

4.3 Metrics about the Population Data

The population OLS estimates of the model described in (4.1) are

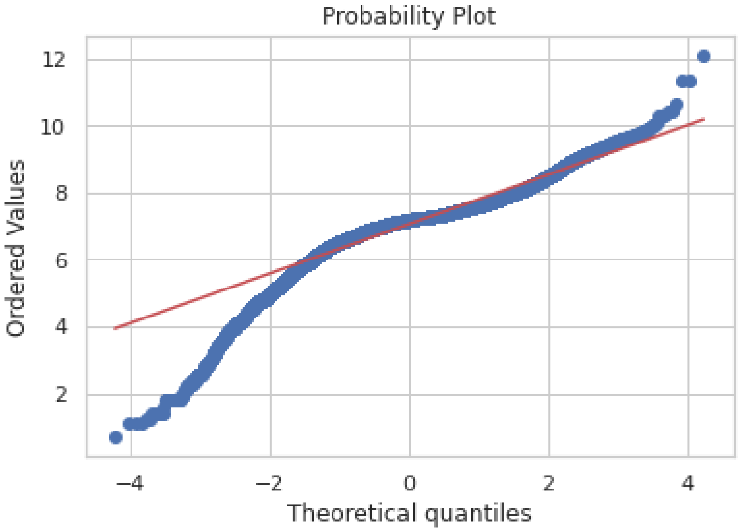


FIGURE 4.2: Q-Q Plot of Log Average Wage

- $\hat{\beta}_0 = 6.7833$ for the constant
- $\hat{\beta}_1 = 0.2105$ for the log_employment ($ln(L_i)$)
- $\hat{\beta}_2 = 0.3226$ for the is_exporter (D_i)

The correlation matrix heatmap in Figure 4.3 indicates a mild correlation between independent variables and the target. The danger of multicollinearity between explanatory variables is not present, as they have a correlation of 0.29. Another interesting statistic is the R-squared value of 0.123. It is not deemed as a high value but not negligible.

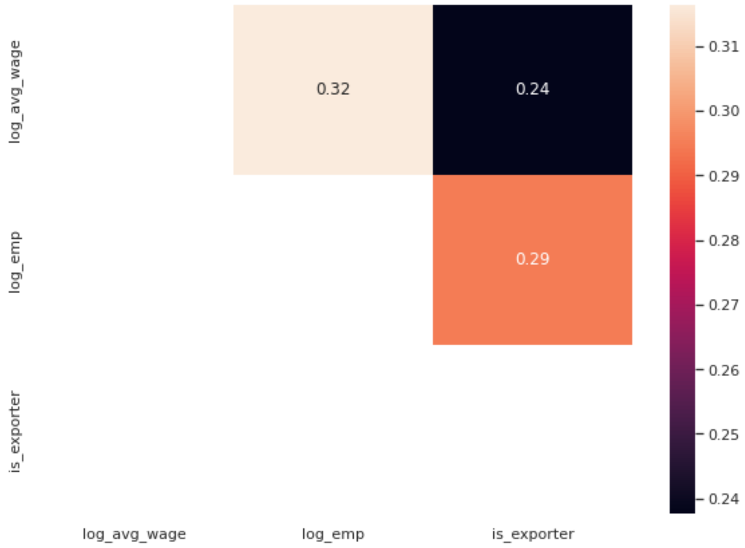


FIGURE 4.3: Heatmap of the Correlation Matrix

4.4 Confidence Interval Test

In the confidence interval test, the linear model in (4.1) is estimated by the wild bootstrap DGP in (3.15) with the Rademacher distribution (3.20), 10,000 bootstrap replications, t3 transformation (3.6) on the residual terms, and 95% confidence interval. The confidence interval test checks how many times the confidence intervals of bootstrap regression estimates contain the population coefficient values out of 1000 trials when setting 95% as the confidence level. Four types of BCA-based confidence intervals are used in this experiment; BCa (1.39), BC ((1.39) but with $\hat{a} := 0$), percentile (1.35), and reverse percentile (1.36). The test is carried out on three different sample sizes: 100, 600, and 1100. Note that this test does not take into account the number of occasions a confidence interval type misses on each side.

The result of the confidence interval test for $\hat{\beta}_1$ (log_employment) and $\hat{\beta}_2$ (is_exporter) from 1000 trials can be found in Table 4.1, 4.2, respectively. 0 means the number of confidence intervals out of 1000 that do not trap the population value. 1 analogously means the number of confidence intervals that include the population value. For the log_employment, the percentile seems to be the confidence interval type closest to the 95% confidence level for all three samples sizes. Regarding the is_exporter, the reverse percentile one is the closest to the predefined 95% confidence level.

sample size	**100**	**600**	**1100**
BCa	0: **60** 1: **940**	0: **46** 1: **954**	0: **56** 1: **944**
BC	0: **59** 1: **941**	0: **44** 1: **956**	0: **55** 1: **945**
Percentile	0: **58** 1: **942**	0: **44** 1: **956**	0: **55** 1: **945**
Reverse Percentile	0: **58** 1: **942**	0: **45** 1: **955**	0: **56** 1: **944**

TABLE 4.1: The 95% Confidence Interval Test for $\hat{\beta}_1$

sample size	**100**	**600**	**1100**
BCa	0: **67** 1: **933**	0: **58** 1: **942**	0: **46** 1: **954**
BC	0: **59** 1: **941**	0: **51** 1: **949**	0: **42** 1: **958**
Percentile	0: **60** 1: **940**	0: **51** 1: **949**	0: **46** 1: **954**
Reverse Percentile	0: **57** 1: **943**	0: **50** 1: **950**	0: **47** 1: **953**

TABLE 4.2: The 95% Confidence Interval Test for $\hat{\beta}_2$

4.5 Bootstrap Distribution Simulation

In this experiment, the linear model in (4.1) is estimated by the wild bootstrap DGP in (3.15) with the Webb 6-point distribution (3.25), t3 transformation (3.6) on the residual terms and 95% confidence interval. The sample sizes used in this experiment are 50, 200, 1000, and 10,000. The number of bootstrap replications is 10, 100, 1000, and 10,000. The aim is to examine how increasing the sample size, and the bootstrap replication number affects the bootstrap sampling distributions of regression coefficients. Figure E.1 and E.2 are the result of the simulations on the bootstrap distribution of the $\hat{\beta}_1$ (log employment). Three lines are drawn: the 2.5th and 97.5th percentile and the mean value. The population value of the β_1 is 0.2105. A vertical movement on one of the columns depicts fixing a bootstrap replication number while increasing the sample size. A horizontal movement on one of the rows represents fixing a sample size while increasing the bootstrap replication number. The effect of increasing the sample size is well-known and demonstrated in the figures. However, increasing the bootstrap replication number at a fixed sample size results in a smoother, more normal-like distribution for a bootstrap distribution. It can be witnessed even for a sample size of 10,000. The effect of increasing both factors is remarkable as it results in a narrower and more accurate confidence interval with a smoother and more normal-like distribution. Similar arguments can be applied to the simulations on the bootstrap distributions of the $\hat{\beta}_2$ (is_exporter), whose figures can be seen in the corresponding Jupyter Notebook[3].

4.6 Concluding Remarks

The thesis provided an overview of the bootstrap regression methods' theoretical background and showcased their practical effectiveness through various simulation tests. Chapter 1 presented the jackknife and bootstrap methods with a focus on building accurate confidence intervals. Chapter 2 introduced the OLS, one of the most frequently used statistical methods, followed by the examination of its properties and the potential to relax some of the Gauss-Markov assumptions. Chapter 3 discussed the fusion of the preceding chapters, the linear regression bootstrap. Three methods were detailed, but it was shown that the wild bootstrap is the one that is the closest to the true linear model. The last chapter demonstrated the power of the wild bootstrap in confidence interval tests and bootstrap distribution simulations with varying sample sizes and bootstrap replication numbers.

A future direction of the ols_bootstrap package is to develop more general models to handle clustered data. These cutting-edge models are called the *wild cluster bootstrap*, detailed in Roodman et al. (2019) and MacKinnon, Nielsen, and Webb (2022). Some of these models have already been implemented in a newly released Python package (wildboottest), which is envisioned to be merged with ols_bootstrap.

[3] https://github.com/pvh95/ols_bootstrap/blob/masterthesis2022/bca_application.ipynb

Appendix A

Background

A.1 Expectation

Let $X \in \mathbb{R}$ be a RV either discrete or continuous with CDF F. If X is discrete, let $A \subset \mathbb{R}$ be the finite (or countably infinite) sample space, and f be its probability mass function (PMF). If X is continuous, let f be its probability density function (PDF).

Definition A.1.1. The **expected value** (average value) of X is defined as

$$\mu := \mathbb{E}(X) = \int_{\mathbb{R}} x\, dF(x) = \begin{cases} \sum_{x \in A} x f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x)\, dx & \text{if } X \text{ is continuous} \end{cases} \tag{A.1}$$

assuming that the sum or integral is well-defined. From now on, $\int x\, dF(x) := \int_{\mathbb{R}} x\, dF(x)$.

Remark. The definition can easily be extended to $\mathbb{R}^d$ with a minor modification.

The unifying notation $\int x\, dF(x)$ is used to define expectations for both discrete and continuous RV conveniently. It has a special meaning in the measure theory.

A closely related concept built on the expectation is called the *Law of Unconscious Statistician* (**LOTUS**) or the *Rule of the Lazy Statistician.*

Theorem A.1.1 (LOTUS)**.** *Let g be a function, then*

$$E\big(g\,(X)\big) = \int g(x)\, dF(x) = \begin{cases} \sum_{x \in A} g(x) f(x) & \textit{if } X \textit{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f(x)\, dx & \textit{if } X \textit{ is continuous} \end{cases} \tag{A.2}$$

assuming that the sum or integral is well-defined.

An important application of LOTUS is for finding the *variance* for a random variable X.

Definition A.1.2. Let X be a RV with mean μ. The **variance** of X is

$$\mathbb{V}(X) = \mathbb{E}(X - \mu)^2 = \int (x - \mu)^2\, dF(x) \tag{A.3}$$

assuming the expectation exists. The standard deviation is $sd\,(X) = \sqrt{\mathbb{V}(X)}$.

A related concept to the expectation is a point estimator called *bias.*

Definition A.1.3. Let i.i.d. $X_1, \ldots, X_n \sim F$. A point estimator $\hat{\theta}_n$ of a true parameter θ is some function T of $X_1, \ldots, X_n$

$$\hat{\theta}_n = T\left(X_1, \ldots, X_n\right).$$

The **bias** of an estimator is

$$\operatorname{bias}\big(\hat{\theta}_n\big) = \mathbb{E}\big(\hat{\theta}_n\big) - \theta. \tag{A.4}$$

An estimator $\hat{\theta}_n$ is said to be **unbiased** if

$$\mathbb{E}\big(\hat{\theta}_n\big) = \theta\,. \tag{A.5}$$

A.2 Law of Large Numbers

Suppose that i.i.d $X_1, \ldots, X_n$ RVs are given from some distribution with finite mean μ and finite variance σ^2. The sample mean, which is also a RV, is defined as

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}\,. \tag{A.6}$$

The mean and variance of the sample mean are

$$\mathbb{E}\left(\bar{X}_n\right) = \mu \tag{A.7}$$

$$\mathbb{V}\left(\bar{X}_n\right) = \frac{\sigma^2}{n}\,. \tag{A.8}$$

There are some interesting properties to reveal when examining the *convergence* of $\bar{X}_n$. However, there are several definitions for the convergence of a RV. The first one is convergence probability which is probably the most used convergence.

Definition A.2.1. A sequence of a RVs X_n **converges in probability** to a RV X, if $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\, X_n - X \,| \leq \epsilon\right) = 1\,. \tag{A.9}$$

X is called the **probability limit** (or **plim**) of X_n. There are two ways of denoting convergence in probability

$$X_n \xrightarrow{P} X \quad \text{or} \quad \operatorname*{plim}_{n\to\infty} X_n = X\,. \tag{A.10}$$

With the knowledge of the sample mean and convergence in probability, the **weak Law of Large Numbers** (**wLLN**) can be stated

Theorem A.2.1 (wLLN)**.** *For all $\epsilon > 0$ as $n \to \infty$*

$$\bar{X}_n \xrightarrow{P} \mu \quad \textit{or} \quad \operatorname*{plim}_{n\to\infty} \bar{X}_n = \mu\,. \tag{A.11}$$

To paraphrase, wLLN states that the sample mean converges in probability to the population mean.

An estimator is consistent if it converges in probability to the population value.

Definition A.2.2. An estimator $\hat{\theta}_n$ of a parameter θ is consistent if $\operatorname{plim}_{n\to\infty} \hat{\theta}_n = \theta$.

It says that if a sample size n is sufficiently large for any given data distribution, the point estimator $\hat{\theta}_n$ will be arbitrarily close to the population θ with high probability.

The next convergence is called **almost sure convergence**, which is considered to be more powerful than convergence in probability.

Definition A.2.3. A sequence of a RVs X_n **converges almost surely** to a RV X if

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1\,. \tag{A.12}$$

It is denoted as

$$X_n \xrightarrow{a.s.} X\,. \tag{A.13}$$

The **almost sure** implies that a RV occurs probability equal to one. The **strong Law of Large Numbers** (**sLLN**) is built upon the concept of almost sure convergence.

Theorem A.2.2 (sLLN). *As $n \to \infty$*

$$\bar{X}_n \xrightarrow{a.s.} \mu\,. \tag{A.14}$$

The third kind of convergence is the "weakest" of all three. It is called **convergence in distribution**.

Definition A.2.4. A sequence of RVs X_n **converges in distribution** to a RV X, if

$$\lim_{n\to\infty} \mathbb{P}\left(X_n \leq x\right) = \mathbb{P}\left(X_n \leq X\right) \tag{A.15}$$

for all x at which the $F(x) = \mathbb{P}(X_n \leq X)$ is continuous. It is denoted as

$$X_n \xrightarrow{d} X\,. \tag{A.16}$$

There is a connection between the three types of convergence

$$X_n \xrightarrow{a.s.} X \quad\Longrightarrow\quad X_n \xrightarrow{P} X \quad\Longrightarrow\quad X_n \xrightarrow{d} X\,. \tag{A.17}$$

LLNs can also be stated in terms of some function of $\bar{X}_n$ due to the following theorem called the **continuous mapping theorem**.

Theorem A.2.3 (Continuous Mapping Theorem). *Let g be a continuous function. Then*

$$\begin{aligned}
X_n \xrightarrow{a.s.} X \quad&\Longrightarrow\quad g(X_n) \xrightarrow{a.s.} g(X)\\
X_n \xrightarrow{P} X \quad&\Longrightarrow\quad g(X_n) \xrightarrow{P} g(X)\\
X_n \xrightarrow{d} X \quad&\Longrightarrow\quad g(X_n) \xrightarrow{d} g(X)\,.
\end{aligned} \tag{A.18}$$

Some properties of convergent sequences of real numbers can be extended to sequences of RVs thanks to **Slutsky's theorem**.

Theorem A.2.4 (Slutsky). *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{d} c$ *(c is constant), then*

$$X_n + Y_n \xrightarrow{d} X + c \tag{A.19}$$

$$X_n Y_n \xrightarrow{d} Xc\,. \tag{A.20}$$

A.3 Central Limit Theorem

Lindeberg-Lévy CLT states that for large n, the distribution of $\bar{X}_n$ after standardization approaches a standard Normal distribution.

Theorem A.3.1 (Lindeberg-Lévy CLT). *As* $n \to \infty$

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1). \tag{A.21}$$

An alternative way of expressing the CLT is

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right)\ . \tag{A.22}$$

CLT can also be used to approximate the distribution of the unnormalized $\bar{X}_n$ for large enough sample sizes

$$\bar{X}_n \mathrel{\dot\sim} N\left(\mu, \frac{\sigma^2}{n}\right). \tag{A.23}$$

Appendix B

Goodness-of-Fit

B.1 R-squared

The sample average of the fitted values $\hat{\boldsymbol{Y}}_i$ is the same as the sample average of the observed target $\hat{\boldsymbol{Y}}$ due to OLSA.2. Therefore, the **sample mean** of the observed target can be defined as

$$\bar{\boldsymbol{Y}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i \,. \tag{B.1}$$

This **sample mean** can be used to define the following measures.

Definition B.1.1 (TSS)**.** The **total sum of squares** (**TSS**) is a measure of the *total* sample variation in $\boldsymbol{Y}_i$ around the sample mean defined as

$$\text{TSS} = \sum_{i=1}^{n} \left(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}} \right) . \tag{B.2}$$

Definition B.1.2 (ESS)**.** The **explained sum of squares** (**ESS**) is a measure of the sample variation in $\hat{\boldsymbol{Y}}_i$ around the sample mean, whihc is **explained by the OLS** defined as

$$\text{ESS} = \sum_{i=1}^{n} \left(\hat{\boldsymbol{Y}}_i - \bar{\boldsymbol{Y}} \right) . \tag{B.3}$$

Definition B.1.3 (RSS)**.** The **residual sum of squares** (**RSS**) is a measure of the sample variation in $\hat{\epsilon}_i$ around the sample mean, which is **not explained by the OLS** defined as

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 \,. \tag{B.4}$$

The following relationship can be drawn from the measures defined above.

Theorem B.1.1. *The total variation in* $\boldsymbol{Y}$ *can be decomposed into the explained sum of squares and the residual sum of squares*

$$TSS = ESS + RSS \,. \tag{B.5}$$

It is illuminating to have a numerical summary of how well the OLS regression fits the data. **R-squared** or $\boldsymbol{R^2}$, the **coefficient of determination**, provides this measure.

Definition B.1.4. Assuming that TSS is not zero, the ratio of the explained variation to the total variation is defined as

$$\boldsymbol{R^2} = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}\,. \tag{B.6}$$

$\boldsymbol{R^2}$ in a regression model explains the variation in the response explained by independent variables The value of $\boldsymbol{R^2}$ is between 0 and 1. The higher the value of $\boldsymbol{R^2}$, the greater the OLS fits the data. It is important to note that $\boldsymbol{R^2}$ neither gives information on whether a model is good or bad nor determines whether the data and predictions are biased. In other words, high $\boldsymbol{R^2}$ can be a poor model, and low $\boldsymbol{R^2}$ can be a great model.

B.2 Adjusted R-squared

This modified measure adjusts for the number of regressors in a linear model relative to the number of data points. For a multivariable regression model, $\boldsymbol{R^2}$ automatically increases as an additional explanatory variable is provided, whether this new regressor variable has a "meaningful added value" or not. In extreme cases, if a model contains $n-1$ independent variable, $\boldsymbol{R^2} = 1$. For such cases, it gives rise to the usage of $\boldsymbol{R^2}_{\text{adj}}$.

Definition B.2.1 (Adjusted R-squared)**.** The **adjusted R-squared** is defined as

$$\boldsymbol{R^2}_{\text{adj}} = 1 - \frac{\text{RSS}\,/\,(n-d-1)}{\text{TSS}\,/\,(n-1)}\,. \tag{B.7}$$

$\boldsymbol{R^2}_{\text{adj}} \leq \boldsymbol{R^2}$ and it is also an unbiased estimator of the population $\boldsymbol{R^2}$.

Appendix C

The Hat Matrix

This appendix summarizes Pardoe, Simon, and Young (2018, Lec 9). As discussed in Section 2.6.2, the hat matrix plays a vital role in HCEs (heteroskedasticity-consistent standard errors) from HC2 onwards by scaling the estimated residual terms in the sandwich estimator. The question is where this matrix arises from and why it is called the "hat matrix". The formulation of linear regression can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,.$$

The OLS prediction to the linear model is

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

where the OLS estimator $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,.$$

Putting together the two equations above provides another way of expressing the predicted response

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,.$$

The fitted value is gained by left multiplying the $n \times 1$ vector, $\boldsymbol{Y}$, by a $n \times n$ matrix $\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$. This matrix is called the **hat matrix**

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,. \tag{C.1}$$

Therefore,

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}\,. \tag{C.2}$$

The name "hat matrix" comes from putting "ˆ" on the observed dependent variable $\boldsymbol{Y}$ to get the fitted value $\hat{\boldsymbol{Y}}$. In geometry, the hat matrix is called the **projection matrix**, shown in (2.25). The equivalent way of writing (C.2) for an individual fitted value, $\hat{\boldsymbol{Y}}_i \quad (i = 1, \ldots, n)$

$$\hat{\boldsymbol{Y}}_i = h_{i1}\boldsymbol{Y}_1 + h_{i2}\boldsymbol{Y}_2 + \ldots + h_{ii}\boldsymbol{Y}_i + \ldots + h_{in}\boldsymbol{Y}_n \tag{C.3}$$

where the **leverage**,

$$h_{ii} = [\boldsymbol{H}]_{ii} = \boldsymbol{X}_{i\cdot}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}_{i\cdot} \tag{C.4}$$

is the i^{th} diagonal element of the hat matrix. h_{ii} describes the influence $\boldsymbol{Y}_i$ has on $\hat{\boldsymbol{Y}}_i$ (Pardoe, Simon, and Young, 2018). In other words, when h_{ii} is small $\boldsymbol{Y}_i$ contributes less to the corresponding $\hat{\boldsymbol{Y}}_i$ than when h_{ii} is large. Another interpretation of leverages is h_{ii} measures the "weighted" distance between $\boldsymbol{X}_i$ and the mean of all the $\boldsymbol{X}_i$'s. Therefore, it is capable of detecting extreme data points when h_{ii} is "large". What does "large" mean for identifying extreme values in $\boldsymbol{X}$?
First, state the properties of the leverage h_{ii}

1. $0 \leq h_{ii} \leq 1$.
2. $\sum_{i=1}^{n} = d + 1$, the sum of h_{ii} equals to the number of independent variables.

The mean leverage value is defined as

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{d+1}{n} \, . \tag{C.5}$$

A usual way of flagging the outlier is when h_{ii} is *more than three times greater* than the mean leverage value, $\bar{h}$,

$$h_{ii} > 3\bar{h} = 3\left(\frac{d+1}{n}\right) . \tag{C.6}$$

However, it is not a universally agreed rule as some may use $2(d+1)/n$ as a cutoff value.

Even though a data point has high leverage, it does not necessarily mean it is also an **influential observation**. In a nutshell, an observation is influential if leaving out this observation from the sample results in a substantial alteration in parameter estimates of interest. There are several existing diagnostic methods to flag an influential point such as Cook Distance, Welsch Distance, DFFITS, DFBETAS and COVRATIO. Montgomery, Peck, and Vining (2021) provide a more detailed discussion of influential points. However, Hansen argues that it is not recommended to use those measures listed above for application as they are not based on statistical theory. He suggests using the LOO Estimator (see D.1) to measure the coefficient change.

Appendix D

The Jackknife and HC3

A brief description of the numerical equivalency between the jackknife error covariance matrix and HC3 is detailed based on Hansen (2022a).

Theorem D.0.1. *The* ***leave-one-out*** *(LOO) estimator and prediction error are provided by*

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}} - \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}_{i\cdot}\,\tilde{\epsilon}_i \tag{D.1}$$

$$\tilde{\epsilon}_i = \left(1 - h_{ii}\right)^{-1}\hat{\epsilon}_i\,, \tag{D.2}$$

respectively. h_{ii} *is the leverage value (see (C.4)), and* $\hat{\epsilon}_i$ *is the OLS prediction error.*

Denote the sample mean of the LOO estimators as

$$\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} - \tilde{\mu} \tag{D.3}$$

where

$$\tilde{\mu} = \frac{\sum_{i=1}^{n}\boldsymbol{X}_{i\cdot}\,\tilde{\epsilon}_i}{n}\,. \tag{D.4}$$

Take the difference between the LOO estimator $\hat{\boldsymbol{\beta}}_{(-i)}$ and the sample mean of the LOO estimator $\bar{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}}_{(-i)} - \bar{\boldsymbol{\beta}} = -\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}_{i\cdot}\,\tilde{\epsilon}_i - \tilde{\mu}\right). \tag{D.5}$$

The **jackknife error covariance matrix** of the OLS estimator $\hat{\boldsymbol{\beta}}$ is

$$\hat{\mathbb{V}}^{\text{jack}}_{\hat{\boldsymbol{\beta}}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\beta}}_{(-i)} - \bar{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_{(-i)} - \bar{\boldsymbol{\beta}}\right)^T \tag{D.6}$$

$$= \frac{n-1}{n}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_{i\cdot}\boldsymbol{X}_{i\cdot}^T\tilde{\epsilon}_i^2 - n\tilde{\mu}\tilde{\mu}^T\right)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \tag{D.7}$$

$$= \frac{n-1}{n}\hat{\mathbb{V}}^{\text{HC3}}_{\hat{\boldsymbol{\beta}}} - (n-1)\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\tilde{\mu}\tilde{\mu}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \tag{D.8}$$

where $\hat{\mathbb{V}}^{\text{HC3}}_{\hat{\boldsymbol{\beta}}}$ is the HC3 estimator given by (2.70). The second term in (D.8) is usually small because $\tilde{\mu}$ is usually small in magnitude. Therefore, $\hat{\mathbb{V}}^{\text{jack}}_{\hat{\boldsymbol{\beta}}} \approxeq \hat{\mathbb{V}}^{\text{HC3}}_{\hat{\boldsymbol{\beta}}}$.

Appendix E

Bootstrap Distribution of the Log Employment

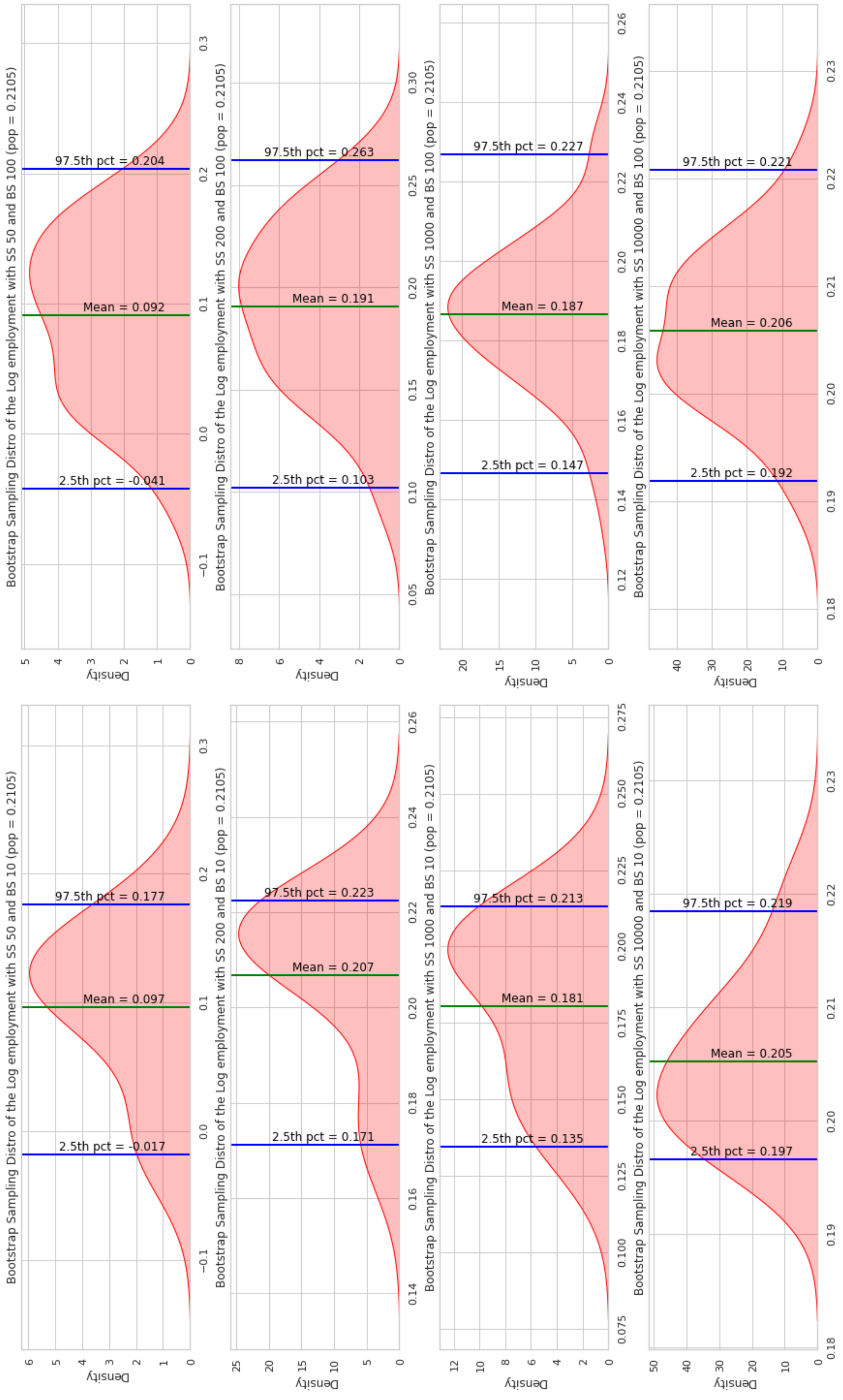


Figure E.1: Bootstrap distributions of Log Employment (Part 1)

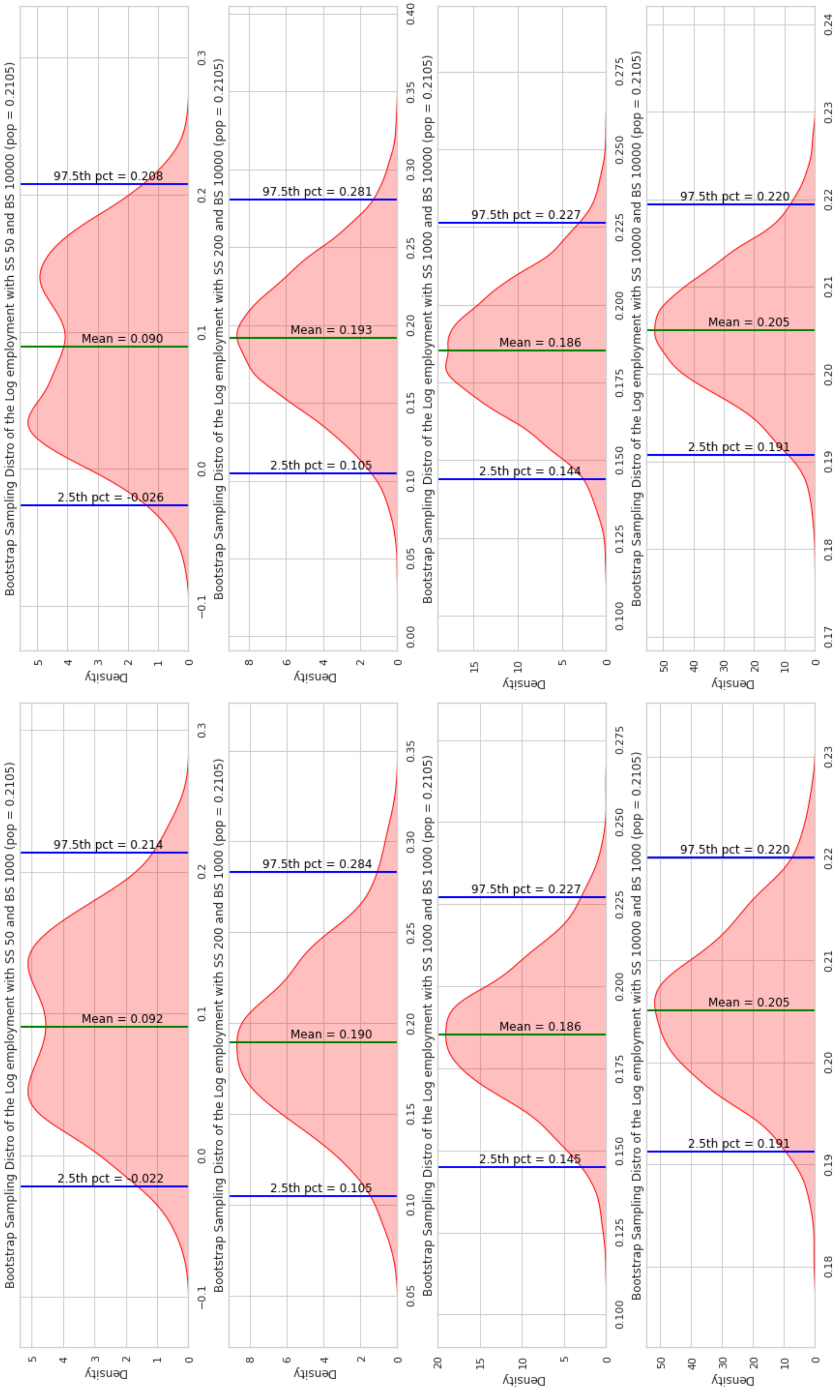


Figure E.2: Bootstrap distributions of Log Employment (Part 2)

Bibliography

White, Halbert (1980). “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity”. In: *Econometrica: Journal of the Econometric Society*, pp. 817–838.

Efron, Bradley (1979). “Bootstrap Methods: Another Look at the Jackknife: Annual Statistics”. In.

Wu, Chien-Fu Jeff (1986). “Jackknife, Bootstrap and other Resampling Methods in Regression Analysis”. In: *The Annals of Statistics* 14.4, pp. 1261–1295.

Liu, Regina Y (1988). “Bootstrap Procedures Under some Non-I.I.D. Models”. In: *The Annals of Statistics* 16.4, pp. 1696–1708.

MacKinnon, James G, Morten Ørregaard Nielsen, and Matthew D Webb (2022). *Fast and Reliable Jackknife and Bootstrap Methods for Cluster-robust Inference*. Queen’s Economics Department working paper. Department of Economics, Queen’s University. URL: https://books.google.hu/books?id=F0EVzwEACAAJ.

Rao, C Radhakrishna (1989). “Statistics and Truth”. In: *Putting Chance to Work*.

Shao, Jun and Dongsheng Tu (2012). *The Jackknife and Bootstrap*. Springer Science & Business Media.

Blitzstein, John K and Jessica Hwang (2019). *Introduction to Probability*. 2nd Edition. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. ISBN: 9780429766749. URL: http://probabilitybook.net.

Van der Vaart, Aad W (2000). *Asymptotic Statistics*. Vol. 3. Cambridge University Press.

Wasserman, Larry (2006). *All of Nonparametric Statistics*. Springer Science & Business Media.

Dvoretzky, Aryeh, Jack Kiefer, and Jacob Wolfowitz (1956). “Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator”. In: *The Annals of Mathematical Statistics*, pp. 642–669.

Wasserman, Larry (2004). *All of Statistics: A Concise Course in Statistical Inference*. Vol. 26. Springer.

Chen, Yen-Chi (2020). *Lecture 10: Statistical Functionals and the Bootstrap*. URL: https://faculty.washington.edu/yenchic/20A_stat512/Lec10_functional.pdf.

Tukey, John W (1958). “Bias and Confidence in not-quite Large Samples (Abstract)”. In: *The Annals of Mathematical Statistics* 29, p. 614.

Efron, Bradley and Robert J Tibshirani (1994). *An Introduction to the Bootstrap*. CRC press.

MacKinnon, James G and Halbert White (1985). “Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties”. In: *Journal of Econometrics* 29.3, pp. 305–325.

Efron, Bradley and Trevor Hastie (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science.* Cambridge University Press.

Hansen, Bruce E (2022a). *Econometrics.* Princeton University Press.

Altman, Naomi (2018). *STAT555 – Statistical Analysis of Genomics Data.* URL: https://online.stat.psu.edu/stat555/node/119/.

Chen, Yen-Chi (2019). *Lecture 10: The Bootstrap.* URL: http://faculty.washington.edu/yenchic/19A_stat535/Lec10_bootstrap.pdf.

Caron, Francois (2019). *SB1.2/SM2 Computational Statistics Lecture notes: The Bootstrap.* URL: https://www.stats.ox.ac.uk/~caron/teaching/sb1b/lecturebootstrap.pdf.

DiCiccio, Thomas J and Bradley Efron (1996). “Bootstrap Confidence Intervals”. In: *Statistical Science* 11.3, pp. 189–228.

Efron, Bradley and Balasubramanian Narasimhan (2020). “The Automatic Construction of Bootstrap Confidence Intervals”. In: *Journal of Computational and Graphical Statistics* 29.3, pp. 608–619.

Hesterberg, Tim C (2015). “What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum”. In: *The American Statistician* 69.4, pp. 371–386.

Davison, Anthony C and David V Hinkley (1997). *Bootstrap Methods and Their Application.* 1. Cambridge university press.

Hesterberg, Tim C (2011). “Bootstrap”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 3.6, pp. 497–526.

— (2004). “Unbiasing the Bootstrap-Bootknife-Sampling vs. Smoothing”. In: *American Statistical Association.* URL: https://drive.google.com/file/d/1eUo2nDIrd8J_yuh_uoZBaZ-2XCl_5pT7/edit.

Efron, Bradley (1987). “Better Bootstrap Confidence Intervals”. In: *Journal of the American Statistical Association* 82.397, pp. 171–185.

Buteikis, Andrius (2020). *Practical Econometrics and Data Science.* URL: http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/.

Strang, Gilbert (2016). *Introduction to Linear Algebra.* Wellesley-Cambridge Press.

Hansen, Bruce E (2022b). “A Modern Gauss-Markov Theorem”. In: *Econometrica* 90.3, pp. 1283–1294.

Greene, William H (2017a). *Econometric Analysis.* 8th Edition. Pearson.

Seabold, Skipper and Josef Perktold (2010). “Statsmodels: Econometric and Statistical Modeling with Python”. In: *9th Python in Science Conference.*

Greene, William H (2017b). *Econometric Analysis – Mathematical and Statistical Appendices A-E.* Pearson. URL: https://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller (2008). "Bootstrap-based Improvements for Inference with Clustered Errors". In: *The Review of Economics and Statistics* 90.3, pp. 414–427.

MacKinnon, James G and Matthew D Webb (2017). "Wild Bootstrap Inference for Wildly Different Cluster Sizes". In: *Journal of Applied Econometrics* 32.2, pp. 233–254.

Roodman, David et al. (2019). "Fast and Wild: Bootstrap Inference in Stata Using boottest". In: *The Stata Journal* 19.1, pp. 4–60.

Pedace, Roberto (2013). *Econometrics For Dummies.* John Wiley & Sons.

Eicker, Friedhelm (1963). "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions". In: *Annals of Mathematical Statistics* 34, pp. 447–456.

Hinkley, David V (1977). "Jackknifing in Unbalanced Situations". In: *Technometrics* 19.3, pp. 285–292.

Horn, Susan D, Roger A Horn, and David B Duncan (1975). "Estimating Heteroscedastic Variances in Linear Models". In: *Journal of the American Statistical Association* 70.350, pp. 380–385. DOI: 10.1080/01621459.1975.10479877. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1975.10479877.

MacKinnon, James G (2013). "Thirty Years of Heteroskedasticity-Robust Inference". In: *Recent Advances and Future Firections in Causality, Prediction, and Specification Analysis.* Springer, pp. 437–461.

Long, J. Scott and Laurie H. Ervin (2000). "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model". In: *The American Statistician* 54.3, pp. 217–224. DOI: 10.1080/00031305.2000.10474549. URL: https://www.tandfonline.com/doi/abs/10.1080/00031305.2000.10474549.

Davidson, Russell and James G MacKinnon (1993). *Estimation and Inference in Econometrics.* Vol. 63.

Cribari-Neto, Francisco (2004). "Asymptotic Inference Under Heteroskedasticity of Unknown Form". In: *Computational Statistics & Data Analysis* 45.2, pp. 215–233.

Cribari-Neto, Francisco and Wilton Bernardino (2011). "A New Heteroskedasticity-Consistent Covariance Matrix Estimator for the Linear Regression Model". In: *AStA Advances in Statistical Analysis* 95.2, pp. 129–146.

Cribari-Neto, Francisco, Souza Tatiene, and Klaus LP Vasconcellos (2007). "Inference under Heteroskedasticity and Leveraged Data". In: *Communications in Statistics—Theory and Methods* 36.10, pp. 1877–1888.

MacKinnon, James G (2006). "Bootstrap Methods in Econometrics". In: *Economic Record* 82, S2–S18.

— (2012). "Inference Based on the Wild Bootstrap". In.

Freedman, David A (1981). "Bootstrapping Regression Models". In: *The Annals of Statistics* 9.6, pp. 1218–1228.

Flachaire, Emmanuel (2005). “Bootstrapping Heteroskedastic Regression Models: Wild Bootstrap vs. Pairs Bootstrap”. In: *Computational Statistics & Data Analysis* 49.2, pp. 361–376.

Horowitz, Joel L (2001). “The Bootstrap”. In: *Handbook of Econometrics*. Vol. 5. Elsevier, pp. 3159–3228.

Flachaire, Emmanuel (1999). “A Better Way to Bootstrap Pairs”. In: *Economics Letters* 64.3, pp. 257–262.

Beran, Rudolf (1988). “Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements”. In: *Journal of the American Statistical Association* 83.403, pp. 687–697.

Davidson, Russell and James G MacKinnon (1999). “The Size Distortion of Bootstrap Tests”. In: *Econometric theory* 15.3, pp. 361–376.

Mammen, Enno (1993). “Bootstrap and Wild Bootstrap for High Dimensional Linear Models”. In: *The Annals of Statistics* 21.1, pp. 255–285.

Davidson, Russell and Emmanuel Flachaire (2008). “The Wild Bootstrap, Tamed at Last”. In: *Journal of Econometrics* 146.1, pp. 162–169.

Webb, Matthew D (2013). *Reworking Wild Bootstrap Based Inference for Clustered Errors*. Tech. rep. Queen’s Economics Department Working Paper.

Flachaire, Emmanuel (2002). “Bootstrapping Heteroskedasticity Consistent Covariance Matrix Estimator”. In: *Computational Statistics* 17.4, pp. 501–506.

Pardoe, Ian, Laura Simon, and Derek Young (2018). *STAT462 – Applied Regression Analysis*. URL: https://online.stat.psu.edu/stat462/node/77/.

Montgomery, Douglas C, Elizabeth A Peck, and G Geoffrey Vining (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.