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.