Bootstrap Methods - Programming Assignment

Vien Dinh*, Derek Dirks† and Jorrim Prins‡

January 10, 2020

1 Introduction

An important econometric problem is the difference between asymptotic and finite sample behavior. Statistical inference methods that rely on asymptotics conclude convergence to specific distributions and these are used to calculate estimators and statistics. Small datasets are a reoccurring problem as asymptotic assumptions will lead to invalid outcomes, an alternative to the asymptotic approach has to be found.

An important method for finite sample data is the bootstrap algorithm, which measures the properties of an estimator or statistic by observing the available data. Monte Carlo simulation determines the distribution of such an estimator or statistic, by resampling the available data in every simulation.

2 Theory

The bootstrap algorithm can be executed by nonparametric, parametric or residual resampling. Nonparametric bootstrap treats the original sample as the population and generates B bootstrap samples from this new 'population', using sampling with replacement. Therefore, each bootstrap sample contains some original data points that appear multiple times, while other points may not appear at all. Parametric bootstrap relies on the conditional distribution of the data and generates new data from random draws of distribution. Residual resampling generates new data by resampling the observed residuals and can be seen as an intermediate between the previous two. A relevant statistic is calculated from the resampled dataset and this is repeated B times. The B bootstrap replications can be combined to find a bootstrapped version of the statistic. In this report the nonparametric bootstrap method is chosen as it is easier to compute than the other three.

3 Data & Methodology

This report uses data that is similar to the *cholesterol* dataset used by Efron and Hastie (2016). The dataset consists of information on 164 men that were administered a drug called cholestyramine over a timespan of seven years on average. Two variables were measured, the men's *decrease of cholesterol levels* over the total period and the proportion of the intended dose of cholestyramine that was actually taken (called *compliance*).

*Studentnummer: 11002115
†Studentnummer: 11029633
‡Studentnummer: 11038934

Nonparametric bootstrap is used to evaluate standard errors in a polynomial regression model with unknown degree, that looks as follows:

$$y = X_m\beta_m + \varepsilon \qquad \varepsilon \sim \left(0, \sigma^2 I\right)$$

The dimension of the regressor matrix is denoted by m, whereas the polynomial degree is m-1 as the model contains a constant. Predictions ($\hat{\theta}_x = x'\hat{\beta}_m$ with $\hat{\beta}_m = (X_m' X_m)^{-1} X_m' y$) are made by a regular OLS estimation and these OLS predictions are used as the statistic of importance. The optimal degree for the polynomial is determined by the C_p criterion specified below. For the *fixed* method, the criterion is evaluated after regression on the full dataset and bootstrap follows for estimations, the *adapted* method evaluates the criterion and estimates the statistic at every bootstrap replication.

$$C_p(m) = \|y - X_m\beta_m\|^2 + 2\sigma^2 m$$

The standard error of the predictions is finally evaluated by *bagging* the adaptive approach, reducing the variability in the model selection procedure and finding a *smoothed* estimator.

Differences between the estimators are evaluated by comparing the models for specific values of c, where c is used to form a specific polynomial x for prediction.

4 Results

The *fixed* method needs an optimal degree of the polynomial, calculating the C_p criterion with OLS results of the full dataset show the following values for the criterion:

Degree	C_p
0	71887
1	1131
2	1411
3	667
4	1591
5	1811
6	2758

Figure 1: Criterion values for OLS regression

The table shows a preference for models with a polynomial degree of 3 ($m = 4$). The *adaptive* approach decides the polynomial degree per bootstrap replication and the 4000 bootstrap replications used have the following proportion per degree (a degree of 0 is never chosen in this approach, as it is far from optimal):

	m=1	2	3	4	5	6
Proportion	0.19	0.13	0.35	0.08	0.20	0.05

Figure 2: Proportion of replication per polynomial degree (adaptive approach)

Model selection procedures like the adaptation method are only possible with a large amount of available data. With the relatively small available dataset (n = 164), the bootstrap algorithm is an effective method as new samples are generated. The optimal degree chosen by the *fixed* method is still chosen in 35% of the replications, but other degrees are definitely not negligible.

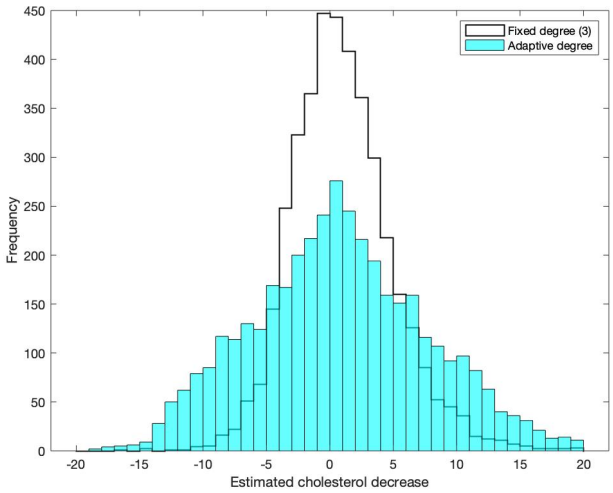

Figure 3: Histogram of bootstrap replications for $c = -2$

The distribution of 4000 bootstrap replications shows a high peak with low variance (estimated standard error of 4.11) for the *fixed* method, whereas the *adaptive* approach has a broader distribution (estimated standard error of 8.06). This shows that ignoring model selection in every replication results in confidence intervals that are smaller than they should be, they would be almost twice as big.

Figure 2 shows the proportion of replications estimated by degree, which is considerably spread out. This could lead to discontinuous estimators, Efron and Hastie (2016) show that this is actually the case in this example, and can be solved by *bagging*.

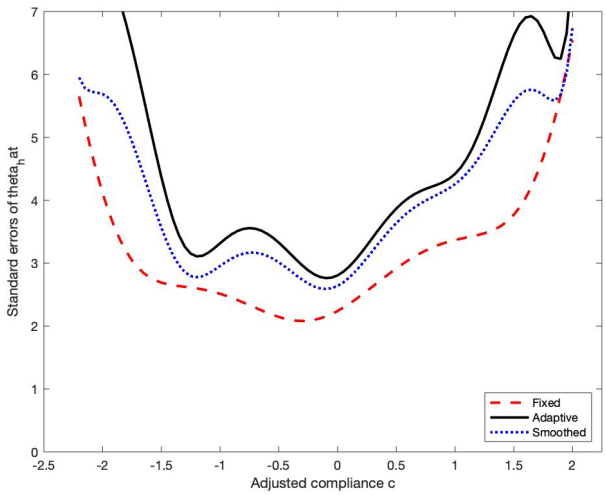

Figure 4: Bootstrap standard-error estimates of $\hat{\theta}_c$

The figure shows estimated standard errors of the *smoothed* estimator exactly between the other two estimators as expected. Predictions are made for c-values between -2.2 and 2, this includes most of the observed values in the sample. Average bootstrap standard error estimations are 3.16, 4.75 and 4.03 for the *fixed, adaptive and smoothed* approach respectively with and adaptive/fixed ratio of 1.50 and adaptive/smoothed ratio of 1.18.

5 Conclusion

Looking at a limited dataset of 164 observations regarding the decrease of men's cholesterol levels, a nonparametric bootstrap model is approximated in three different ways. A *fixed*, *adapted* and *smooth* method evaluate the criterion function and the standard errors are used to assess the performance of the different methods.

Comparing the fixed and adapted methods, the fixed method estimates standard errors that are lower for every compliance value, judging from figure 4. However, ignoring the model selection used in the adaptive method, confidence intervals and tests determined by the fixed method will be too short and thus overreject.

The smooth method is derived from the adaptive method as it uses a weighted estimation of the adaptive method. Hence, the smooth method is considered the preferred of the three.

However, these results should be viewed carefully. As the bootstrap method still relies on asymptotic theory, results using a relatively small dataset are likely to be invalid. One way to correct for this is to simply increase the sample size, while another would be to implement a nested bootstrap, which would lead to a asymptotic refinement since the statistic is asymptotically non-pivotal.

6 Appendix

[5] Model selection is important as statistical inference from the dataset will only be correct if valid assumptions are made and the model is correctly specified. From theory, the obvious assumption is an effect of the adjusted compliance on cholesterol level decrease and not the other way around, therefore use the latter as the dependent variable in a regression model. The perceived model is a polynomial one with m-1 degrees (which totals m regressors as the model includes a constant) where the ultimate number of polynomial features is to be determined by model selection.

Model selection procedures like the adaptation and smoothed method are only possible with a large amount of available data. With the relatively small available dataset (n = 164), the bootstrap algorithm can be an effective solution as new samples are generated from existing data.

Optimal models for large datasets are usually determined by splitting the data into a training and a test set and using these to achieve the best test fit, but the available dataset is not large enough to do this. Cross-validation could be a good alternative, as it splits the data into training and test sets multiple times, resampling the sets.

[6] As statistical inference methods rely on asymptotic theory, this report will likely generate invalid results as the data is limited to 164 observations. Even though adaptive and smoothed bootstrap methods in this report might perform better than usual asymptotic results, the results should be considered carefully. One way to validate the results is by simply having more data available. If another sample of the original data would be available, asymptotic theory behind the bootstrap is closer to true and the results become better. Estimates of the current data could be easily compared and validated with the new data.

A way to improve estimates is to implement the nested bootstrap. This method is especially useful when the statistic is asymptotically non-pivotal (which is the case, θ depends on unknown parameter σ via ϵ), since nested estimates of σ convert the statistic into an asymptotically pivotal one. Bootstrapping the existing bootstrap sample and subsequently applying the percentile-t method could lead to asymptotic refinement.