Africa's Slave Trade: Implementing Machine Learning for Instrumental Variable Analysis

Christopher Colon, Anel Rodriguez, Charlie Stutts, Madison Stevens, and Davis Xu

ECMT 680: Financial Econometrics

Dr. M. Jahangir Alam

Texas A&M University

Spring 2024

Abstract

Our analysis is based on the study by Nathan Nunn (2008), who utilizes *Instrumental Variable* analysis. We replicate the methodology while incorporating new economic data from the World Bank and contribute to the results by applying a Machine Learning technique, *Double Lasso*, with the objective of demonstrating that the use of this machine learning methodology can confirm and improve Nunn's results. Both results, from our replication and from the application of the machine learning method, confirm the findings of the original paper, which concludes that the slave trades have an adverse effect on economic development in African countries. Finally, when we apply the Double Lasso, we can also see that the results become more accurate and precise than the original results.

I. Introduction

Our analysis is based on a replication of the study *The Long-Term Effects of Africa's Slave Trades* by Nathan Nunn (2008) and a contribution to the original method, applying a Machine Learning (ML) technique, Double Lasso. Through this, we intend to examine the application of the above-mentioned machine learning approach to the study, to confirm and improve its results and interpretation.

Nunn's study explores how slave trades (1400 – 1900) continue affecting the economic development of African countries. It uses slave data and 2000 economic data to analyze the possible existence of a causal relationship between the slave trades and the economic performance in African countries using an Instrumental Variable (IV) analysis as a methodology. The findings show a strong and significant negative relationship between the slaves exported and the economic development of those countries, demonstrating that the regions most affected by the slave trades have the poorest economic performance today.

Commented [SM1]: Might need to cut this out so the data & replication methodology section are not repetitive.

Commented [RA2R1]: I modified it! You can expand in the corresponding section

The method used in the original paper is a traditional econometric method to assess causality, which helps to reveal the strong effect of slavery on economic growth. However, in order to understand better the impact nuances and analyze further the true long-term impact of slave trades on the African economy, we consider that it is fundamental not only to replicate the paper but also explore beyond, using other economic data (i.e., real GDP data from other years), and apply a new technique to improve the precision of the results.

Given the above, the objectives of our paper can be described as follows:

1. Replicate the original results and incorporate new economic data to explore the effects in the subsequent years.
2. Apply a machine learning methodology, in this case Double Lasso, to confirm the original findings and get more robust and precise results.

Thus, our analysis is divided into two parts: the replication and the application of the machine learning method.

The first part of our study focuses on replicating the main results of the original paper using the same method, IV analysis with a Two-Stage Least-Squares (2SLS) regression. The study's replication is fundamental because it allows us to verify the results and detect errors, promote cumulative knowledge, and provide educational value. However, we go beyond a simple replication because we add new economic data from the World Development Indicators of the World Bank to the original dataset. Specifically, we add the logarithms of the real GDP for the period 2001 – 2022 and the logarithm of the average of the real GDP for the same period. These variables allowed us to analyze the impact of the slave trades on economic development of those years in the different African countries and have a more complete picture of those effects in a

longer period, i.e., 22-years period. The results of this further exploration confirm the original findings, showing a high negative relationship between the variables analyzed for all those years.

The second part of our analysis is the incorporation of the machine learning method. We employ Double Lasso Regression while still using the IV approach. Applying a ML technique such as Double Lasso Regression brings significant advantages like handling the data well and capturing the non-linear relationships. Double Lasso is a variable selection and regularization technique which consists in a two-step approach to enhance causal inference. In this study, this approach is useful for several reasons, such as the selection of the most important variables to explain the economic long-term effects of slave exports; the identification of causal effects of the slave trade on economic development in a context where there may be many potential confounding variables; and the reduction of overfitting which makes the model more interpretable and robust. The results applying this method confirm the findings of the replication, and we can see that the coefficients of our interest variable, for the years analyzed, tend to be even more negative and significant, which emphasize the harmful impact of slave exports on the economic performance of the countries studied today. Additionally, the confidence intervals of those coefficients are less wide than those we get in the replication; thus, the results of the Double Lasso prove to be more precise and accurate.

The remainder of the study is structured as follows. Section II presents the literature review. Section III describes the data and the methodology used in the replication exercise and in the application of the machine learning approach. The results comparison is shown in section IV. A discussion of the main findings of this analysis is presented in section V. Discussion and Analysis VI concludes.

II. Literature Review

The lasso regression method was introduced by Robert Tibshirani in 1996 to improve the prediction accuracy of regression models by decreasing their complexity. This was a response to OLS regressions struggling to properly handle many variables, resulting in high variance. Lasso regression shrinks the variance of coefficient estimates through the use of a penalty, with some coefficients going to zero and being removed from the model. Ultimately, the more refined variable selection lasso offers allows for the creation of less complex and more accurate models.

Lasso regression methods have also evolved to meet various research needs. An example is Zou's adaptive lasso method introduced in 2006. This method uses adaptive weights, that vary the level of the penalty applied to each coefficient. This method helps to address inconsistencies in variable selection that can lead to inaccurate results, an issue sometimes experienced with the traditional method. In addition, group lasso was introduced in Yuan and Li's 2006 paper "Model selection and estimation in regression with grouped variables." This modification of the traditional

lasso method finds natural groupings within variables and applies penalties to groups rather than individual variables. This lasso method allows for better prediction when group effects are prominent and better addresses related variables.

The integration of machine learning methods and traditional econometric models is an area of interest as economists are trying to determine if the strengths of both techniques can be combined for economic analysis. Lasso regression generally performs well and has advantages over other machine learning methods due to its ability to handle high dimensional data and can be extended to several types of models and regressions.

However, Lasso for IV analysis has not always proven to be the most effective pairing such as when studying labor economics. The ability of lasso to create artificial exclusion restrictions that result in misleading conclusions sometimes poses an issue for IV analysis. This was seen in a study titled "Machine Labor" conducted by Angrist and Frandsen that evaluated whether attending highly selective or private institutions leads to higher earnings compared to attending less competitive or public universities. The researchers found that while effective for selecting control variables, lasso was less effective when selecting an instrumental variable. It often selected weak instruments that did not lead to more precise results.

Our replication enhancement intends to identify if lasso regression is an effective machine learning method to pair with IV analysis when using vast historical data to examine present-day economic outcomes. The enhancement will concentrate on the use of lasso regression as a method for control variable selection rather than instrument selection.

III. Methodology

Data

The original study uses historical shipping records data sourced from the Trans-Atlantic Slave Trade Database, European Port Registers, and observer and government official reports as well as ethnicity data for the slaves transported for the years 1400-1900. These data sources are used to approximate the number of slaves exported from each African country, with shipping record data offering insight into the number of slaves exported and ethnicity data providing indicators of where slaves originated from. Geographic data indicating each African country's distance from the main trade routes and country-level GDP data for 2000 are also used to support the model. In addition, country-level supplemental data for independent variables is included in the data set such as temperature, humidity, population, and island dummies.

For replication and enhancement of the study with machine learning, we sourced the original data from the author Nathan Nunn's website and expanded the GDP data sourced from the World Bank to include the years 2001 to 2022.

Replication

Our replication utilizes the original study's method of IV analysis to determine the relationship between slave exports and the current economic outcomes of African countries. The model implemented uses slave exports ln_(export/area) as the independent variable and country-level GDP as the dependent variable to measure economic outcomes. To address potential issues of endogeneity the distance of each country from the Trans-Atlantic, Indian Ocean, Red Sea, and trans-Saharan slave trade routes is used as the instrumental variable. Our replication code reproduces the study's Model 1-3 results, displayed in the original paper's Table IV. Particularly, our analysis focuses on Model 3 which includes colonizer fixed effects and geography controls. When reproducing Model 3 results, we expanded the model to also include the years 2001-2022.

To replicate Model 3, we initially regressed slave exports on the distance from trade routes and the independent variables falling under colony and geography effects. Once the predicted values for slave exports were produced, they were regressed on the GDP from 2000 to identify the causal relationship. This is the result shown in Table IV, Column 3 of the original paper. After replicating the coefficient of interest for slave exports, the model was executed using 2001-2022 GDP data to increase the breadth of results for our analysis. The following sections IV and V detail the replication results and their implications when compared to the outcomes obtained with machine learning techniques.

Machine Learning Enhancement

Our study enhances Nunn's analysis by implementing the lasso variable selection method to uncover the true effects of historical slave exports on current GDP. This was accomplished by running the lasso twice; one between the endogenous variable ln_(export/area) and the list of potential instruments, and the other between the current GDP and the list of potential instruments. Any variable that was in either list was then incorporated into the regression model.

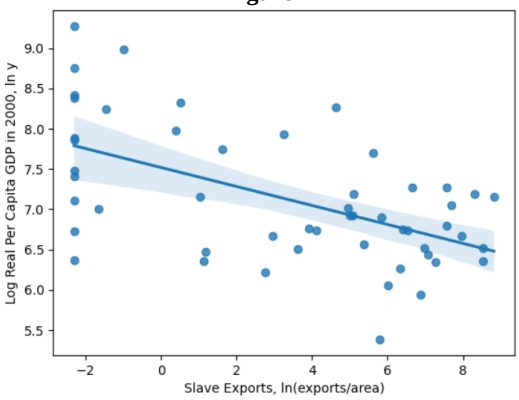
These steps were repeated for the GDP years between 2000-2022. The resulting coefficients are modeled in a graph, with the results from the replication as a comparison. Confidence intervals are included around these results to determine the significance of our machine learning implementations.

IV. Results and Comparison

Replication

Our replication of Nunn's work yielded the same coefficient as the original paper for the ln_(exports/area) variable we chose for our study. This gives confidence in our extended replication results and shows that our replication method was sound. The graph below shows the results of our replication of Nunn's work for the economic outcomes of the year 2000.

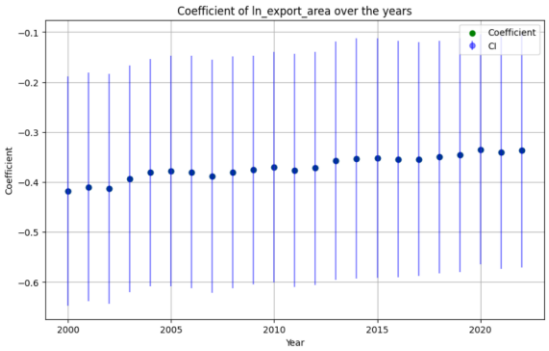
Figure 1



Nunn's analysis provided a coefficient of -.286 between current economic outcomes and the log of slave exports normalized by land area. Our paper produced the same -.286 coefficient between these variables. The confidence interval is shaded in a lighter blue.

When extending the years studied of the original paper, we had to normalize the coefficient in the original paper to the rest of our coefficients. This is because the GDP data in the original paper had a base year of 1990, while the GDP data for 2001-2022 used in our extension of the study had a base year of 2015. Once we normalized the coefficient for the year 2000 found in the original paper, we studied the coefficients for slave exports over the 2000 to 2022 time period, which is displayed in the following graph. We found a significant negative correlation between outcomes and ln(exports/area) for every year in the extended study, providing further confidence in the accuracy of the original paper's results.

Figure 2

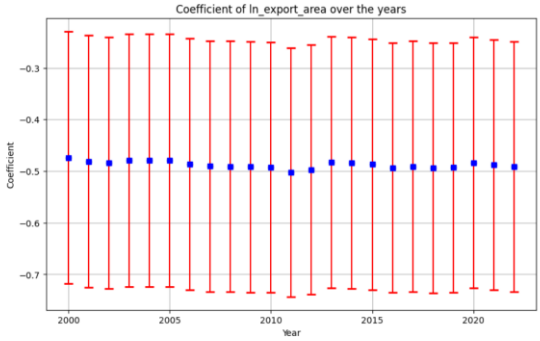


The average coefficient in the extended replication was -.365, with the most significant year being 2000 (-.418) and the least significant year being 2019 (-.346). Every year showed a significant negative correlation, as shown by the confidence intervals provided in the graph in blue.

Double Lasso

After our extended replication was done, we then implemented our double lasso method to enhance the results of the original paper and the extended replication. Our machine learning enhancement produced stronger correlations for every year in the study, showing that the double lasso method was effective in trying to enhance the findings of the original paper. The following graph shows the coefficients using the double lasso method.

Figure 3



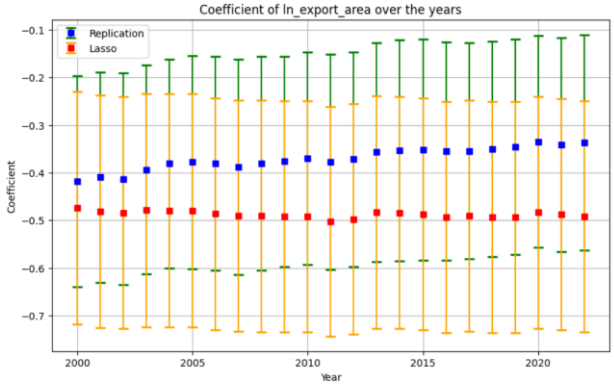
The average coefficient in the double lasso regression method was -.488, with the most significant year being 2000 (-.474) and the most significant year being 2011 (-.502). The negative

coefficients were significant for every year in the study (2000 to 2022), as shown by the confidence intervals provided in red.

Comparison of Results

As stated above, the double lasso regression method produced stronger negative coefficients between the current economic outcomes and the logs of export volumes normalized by area for every year included in our study. The average coefficient for the years in the double lasso method was .123 larger in absolute value than the extended replication of the original study. The graph below shows this difference in coefficient strength.

Figure 4



The difference in results between the double lasso regression method and the extended replication of the original study is not significant, as the confidence intervals for the double lasso (orange) overlap with the confidence interval for the replication (green). There is a consistent difference in coefficient size, but the difference is not large enough to be significant in a 95% confidence interval. That said, the double lasso method produced stronger correlations in every year studied.

Discussing Accuracy and Precision

The replication method had an average standard error of .1187, indicating that the results of the regression were significant enough to confidently say that the two variables had a negative correlation. There were no significant outliers in the years studied, giving confidence to the idea that the results were consistent. The double lasso method had an average standard error of .124, indicating again that the results of the regressions were significant enough to determine a negative correlation between the economic outcomes of countries and the log of the export variable

normalized by land area. Additionally, the confidence intervals for both the replication and the double lasso regression methods never crossed zero, indicating significant results.

The coefficient to standard error ratio, or t stat, was larger for the double lasso method, indicating that the double lasso method was more precise than the replication of the results in the original study. The t stats were large enough to deem the coefficients significant for both the replication and double lasso, indicating significant results.

V. Discussion and Analysis

Interpretation of Findings

The findings from the replication of Nunn's seminal 2008 study alongside the application of the Double Lasso method both reinforce and expand upon the original conclusions concerning the detrimental effects of slave trades on economic development in Africa. Our replication validates Nunn's methodology and findings, with the additional economic data from 2001-2022 offering new insights into the persistent economic challenges faced by affected regions.

This extended dataset highlights not only consistent underperformance relative to other regions but also the variability in recovery and growth trajectories among African nations. Such long-term perspectives are crucial for understanding the depth and endurance of the slave trade's impacts, suggesting that the consequences are not only deep-seated but have also evolved in complexity over the past two decades.

Broader Implications

The application of machine learning techniques, particularly the Double Lasso method, in this study illustrates a significant advancement in the robustness and precision of causal inference within economic research. This methodological enhancement facilitates a more nuanced analysis of complex datasets, often characterized by multicollinearity and high dimensionality, which traditional econometric techniques may not handle as effectively.

The implications for policymaking are profound; such techniques allow for more accurate identification of causal relationships, which is essential for formulating policies aimed at economic recovery and sustainable development. Furthermore, the insights gained from this approach can be applied to other regions globally that have experienced similar historical disruptions, providing a blueprint for how advanced analytical techniques can inform better policy responses.

Acknowledgment of Limitations

While the replication of historical economic studies offers valuable confirmations of past findings, such endeavors inherently encounter challenges related to data availability, quality, and the potential biases in historical records. These issues are compounded in studies dealing with periods as remote as the slave trades, where data was incomplete or unrecorded.

Additionally, the Double Lasso technique, despite its strengths, involves assumptions about variable selection that may not always be held, particularly in complex economic contexts where the relationships between variables are not fully understood. There is also the risk of overfitting, even though the method includes regularization parameters designed to mitigate this risk. These limitations must be carefully managed to ensure the validity and reliability of the conclusions drawn.

Comparison of Results

The comparative analysis between the results obtained from the traditional IV approach and those derived via the Double Lasso method did not show significant enhancements in the latter's capability to provide narrower confidence intervals but provided more pronounced negative coefficients. These improvements suggest a stronger and more definitive negative impact of the slave trade on economic development than previously estimated.

The Double Lasso's ability to refine these estimates speaks to its efficacy in handling complex, multi-variable environments typical of economic data, thereby contributing to a more detailed and nuanced understanding of the economic impacts.

Quality of Analysis

The findings not only corroborate the existing literature but also extend it by incorporating newer methodologies and data. Furthermore, the academic value of replicating seminal research and integrating innovative analytical techniques cannot be overstated. This process not only reinforces the reliability of historical economic analyses but also encourages the adoption of advanced methodologies in the exploration of economic phenomena, fostering a richer understanding and fostering continued academic inquiry.

VI. Conclusions

Combining data from the original IV estimates along with applying a Machine Learning (ML) technique, Double Lasso, over a larger period our analysis reaffirms the detrimental ongoing long-term effects of Africa's slave trades on economic development in African countries. Our

findings not only support the initial negative impact found in the paper but improve our understanding and ability to use Machine learning methods like Double Lasso to refine and enhance the precision of our data while accounting the uncertainty in causal estimates.

Future research could consist of machine learning tools used across immense data sets and exploring additional potentially casual economic indicators. By analyzing a variety of economic contexts, we further our knowledge and awareness of other long-term historical impacts. Overall, these studies help further refine our understanding of causal chains linking past events with outcomes, relationships that were previously hidden due to the complexities of historical data.

Reflecting on our paper's journey from replication to innovation, our combination of traditional econometric models with applied modern machine learning techniques allows us to improve the quality and depth of the initial analysis but also sets a precedent for future research endeavors. By the use of both replication and innovation, economic research can stride forward, grounded in historical insights yet elevated by technological advancement.

VII. References

Angrist, J. D., & Frandsen, B. (2022). Machine Labor. Journal of Labor Economics, 40(S1). https://doi.org/10.1086/717933

Lennon, C., Rubin, E., & Waddell, G. R. (2022). *Machine learning (too much) in 2SLS: Insights from a bias decomposition* (thesis).

Nunn, N. (2008). The Long-Term Effects of Africa's Slave Trades. The Quarterly Journal of Economics Vol. 123, No. 1 (Feb. 2008), 139-176.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x