ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

BACHELOR THESIS ECONOMETRICS AND OPERATIONAL RESEARCH

Application of Double/Debiased Machine Learning

Author:
Jan Jakob HEIJENGA

Student ID number:
432951

Supervisor:
Dr. Andrea A. NAGHI

Second assessor:
Dr. Maria Grith

ABSTRACT

Available data grows fast, for example characteristics of individuals. When estimating the average treatment effect by controlling for characteristics using Ordinary Least Squares, this method will become inconsistent. The new method Double/Debiased Machine Learning delivers unbiased estimators and is able to construct confidence intervals that is root-N consistent, the sample size is N, and approximately normally distributed. This method is applied to the data sets of the two papers “The Long-Term Effects of Africa’s Slave Trades” and “The Slave Trade and the Origins of Mistrust in Africa”. These papers have drawn conclusions about the effect of Africa’s Slave Trade using OLS and Instrumental Variables, based on a large data set. The first paper uses a specific measure for the slave trade in every table. This thesis shows that it is better to use another more important measure. Moreover, this thesis shows more reliable estimations and confidence intervals for the treatment variable. Besides applying DML, the most important results of these two papers are replicate using OLS. For almost every result the estimations are the same, however, when estimating for channels of causality, the estimations differ.

July 17, 2018

aan Connie

1 INTRODUCTION

Over the past years the technology has increased and therefore also data sets have grown. For example, social media platforms, such as Netflix, Facebook and Google, collect a lot of data from their users. These data sets contain a lot of information about every single user. This information varies from their rating from movies to their personal shopping behaviour. All these variables are called covariates. However, there is a problem with estimating using these growing data sets. When you want to estimate the effect of a parameter of interest on the dependent variable, the often used method Ordinary Least Squares (OLS) will become inconsistent when the data contains more than three covariates Athey and Imbens (2015).

Thus, we want to estimate and construct confidence intervals for a parameter of interest when having a high-dimensional set of covariates. This set of covariates can be estimated with new statistical methods, called Machine Learning (ML) Methods. ML methods are increasingly being applied in the Economics literature. These methods are more about algorithms than about asymptotic statistical properties, such as given in the Econometrics courses.

These methods have the ability to learn with the data itself instead of explicitly being programmed. They construct algorithms that can learn from and predict the data. However, estimating with these methods causes two problems, namely overfitting and regularization bias. These problems are solved in Chernozhukov et al. (2017), where they show, when estimating the parameter of interest, the regularization bias and overfitting can be vanished by using two adjustments. First by making use of Neyman-orthogonal moments/scores, which makes the covariates more robust when estimating the parameter of interest, and second by making use of cross-fitting, this is a form of sample-splitting.Combining these methods is called Double or Debiased Machine Learning (DML). In their paper, they stated that DML delivers unbiased estimators that is root-N consistent and approximately normally distributed, this is necessary to construct valid confidence statements. However, the statistical theory of DML is fundamental and because it it relies on mild theoretical requirements, a lot of Machine Learning methods could be used when estimating the covariates' parameters.

Gaining knowledge about this method is of high importance. There has been done a lot of research on Machine Learning methods, however the DML paper dates back to 2017. Moreover, these methods are not covered in the Econometric courses. When someone is estimating using OLS, consisting multiple covariates, the estimation results will be unreliable, affecting the interpretation of empirical results. Because of the growing data sets, it is important that these consequences should concern everybody who is working with large data sets.

This topic is also relevant from a scientific point of view. The lesser is known about this topic, the more my research will add to current knowledge. Because the existing paper dates back to 2017, there is still a lot of progress to be made on this topic and it is relevant to delve into.

In Chernozhukov et al. (2017) they show DML methods in three empirical examples, first, to measure the effect of unemployment insurance bonus on unemployment duration, second, to measure the effect of 401(k) eligibility and participation on net financial assets and third to measure the effect of institutions on economic growth. In this thesis the DML methods are applied to data sets of two specific papers. The first is ”The Slave Trade and the Origins of Mistrust in Africa”, Nunn and Wantchekon (2011). This paper shows that current differences of Africans in trust levels has origins in the slave trades. The second paper is ”The Long-Term Effects of Africa's Slave Trades”, Nunn (2008). They show that a part of the current underdevelopment of Africans is related to the slave trade. Both of these papers use OLS and Instrumental Variables (IV) to draw their conclusions. However, the used data sets contain a lot of covariates, that is why it is interesting to apply DML the data sets of these papers. The target is to estimate and construct confidence intervals using DML with these papers, what will lead to more valid confidence statements.

In Section 2 is the methodology explained. This consist of the different Machine Learning methods and the theorem of DML with the explanation to overcome the regularization and overfitting bias. Section 3 shows which data is used in this thesis and section 4 shows a part of all the replications of the main tables from these two papers. The tables from Nunn and Wantchekon (2011) are 1, 5 and 10. These tables report OLS estimates of the relationship between measures of slave export and trust in neighbours, IV estimates of the effect of slave trade on the trust and OLS estimates of identifying channels of causality. The remaining tables 2, 3, 6, 9 are shown in the Appendix, these tables contains besides the original regressors, also additional controls. From Nunn (2008) table 3 is replicated, this table shows OLS estimates of the relationship between slave exports and income. All these tables have a lot of covariates in their regressions, which makes it interesting to replicate.

Besides replicating the important results, the DML method is also applied to the tables from section 4, these are shown in section 5. There are some very interesting results, which leads in some cases to other conclusions.

1.1 LITERATURE

The Long-Term Effects of Africa's Slave Trades
Nunn (2008) does in his paper research to the question: ”Can part of Africa's current underdevelopment be explained by its slave trades?” To find this out, he makes, using historical data, estimations of the slave export from each African country. His main finding is that he finds that the number of slaves negatively correlated is with the current economic performance. To make this statement more reliable, he examines if this statement is causal using Instrumental Variables. As a result, all the outcomes from the different regressions suggest a negative correlation between the slave export and the current economic performance.

The Slave Trade and the Origins of Mistrust in Africa
Nunn and Wantchekon (2011) show in their paper that the different levels in trust of African's are effects from the slave trade. In detail, they show that individuals whose antecedents were suffered a lot during the slave trade are in modern life less trusting. To make this statement more reliable, they also show that the relationship is causal. Besides, they show that the slave trades' impact is most influenced through internal factors, like values, norms and beliefs.

Double/Debiased Machine Learning for Treatment and Structural Parameters
Chernozhukov et al. (2017) is about estimating with a lot of covariates, in other words to make valid conclusions about a parameter of interest when having high-dimensional nuisance parameters. They consider Machine Learning methods to estimate these nuisance parameters. Although, these Machine Learning methods perform well in practice, they cause a huge bias in estimates of the parameter of interest when estimating the nuisance parameters using ML methods. These are regularization bias and overfitting, in their paper they shows that these biases can be removed by making two adjustments. First, the use of Neyman-orthogonal moments, this makes the nuisance parameters more robust when estimating the parameter of interest. And second, use cross-fitting, this is a method that splits the data into multiple parts, what results in a vanishing of the bias due to overfitting. These methods together are called double or debiased ML (DML). Moreover, they show that one can make valid confidence statements, because the parameter of interest is approximately normally distributed.

High-Dimensional Methods and Inference on Structural and Treatment Effects
Belloni et al. (2014) is a precursor of Chernozhukov et al. (2017). This paper already stated that the data sets become bigger, high-dimensional data. However, normal estimations methods cannot be applied anymore, their goal is to show how methods can be adjusted in that way, so that one can make high-quality inference about the parameters of the model. They also stated that regularization, thus the reduction of the dimension, is necessary to provide meaningful conclusions. However, this paper give just an overview about some methods that can be used and underlying theory, they also stated that further research is necessary to make it more applicable in economic relevant settings.

The elements of Statistical Learning
Tibshirani, Friedman, and Hastie (2009) give an overview about the statistical methods, which are available for construction prediction models when having high-dimensional data sets. They also stated that the available data grows, resulting in new statistical tools. They describe the important new ideas in Machine Learning, Bioinformatics and Data Mining, they also use a consistent terminology in their explanations. This book has a huge coverage, with a lot of Machine Learning methods, also provides this book a good basis for the used Machine Learning methods in this thesis.

2 Methodology

Machine Learning can be divided into two types of learning, namely Supervised Learning and Unsupervised Learning. Classification is an example of supervised learning. This is pure prediction, which classifies each observation into a category. The goal is to find a function based on the observations, a new observation can be assigned to one of the categories. Another example is regression, where you need to simplify a good fit for the conditional expectation. Unsupervised learning is used when the data has not been categorized or classified. In Chernozhukov et al. (2017) they only make use of supervised learning and because these methods are applied, this thesis is not going to delve into the unsupervised learning.

Causality is de-emphasized, because the ML methods are much more about prediction and fit. The first key component of Machine Learning methods is out-of-sample cross-validation. This means that the methods are validated by criticizing their properties out of sample. However, general causal problems do not have unbiased estimators of the true causal effects, how to get unbiased estimators will become clear later.

Regularization is the second key component. This means that there is some penalty to avoid over-fitting, rather than choosing the best fit. But there are two issues with regularization: choosing the amount of regularization and choosing the form of the regularization. To overcome the first issue, the ML literature has emphasized out-of-sample cross validation methods, see the first key component above, for choosing the amount of regularization, in other words for choosing the value of the penalty.

The last key component of ML methods is scalability. This ensures that methods can handle the large amount of data. This is important because the number of observations may be in billions and the number of covariates may run in millions. One of the ways to do this is using parallel computation. In short, large problems are divided into smaller ones, and these can be solved at the same time Athey and Imbens (2015).

The Machine Learning methods are good to reduce variance, using regularization, besides reducing the variance, they also perform well by making a trade off between overfitting bias and regularization bias. On the other hand, both overfitting and regularization bias in estimation the covariates cause a huge bias in estimating the parameter of interest which is obtained by estimate the parameter of interest using ML methods and estimate the nuisance parameters using ML methods, combining these two estimates results in an inconsistent estimator.

2.1 Double/Debiased Machine Learning

Following Chernozhukov et al. (2017), below is their used specification of the partially linear regression model:

$$Y = F\theta_0 + m_0(X) + U, \quad \mathrm{E}[U|X,F] = 0, \tag{1a}$$

$$F = g_0(X) + V, \quad \mathrm{E}[V|X] = 0, \tag{1b}$$

where the outcome variable is Y, F is the parameter of interest, vector

$$X = (X_1, ..., X_p)$$

consists of control variables, and V and U are the error terms. The relevant coefficient which is interesting to derive is θ_0. The second equation shows that the variable of interest depends on control variables. However, naively estimating (1a) using ML methods results in a huge bias in the estimators of θ_0, caused by regularization bias and overfitting.

Additional to this partially linear regression model, Chernozhukov et al. (2017) provide a model that allows for instrumental variable identification:

$$Y = F\theta_0 + m_0(X) + U, \quad \mathrm{E}[U|X, Z] = 0, \tag{2a}$$

$$Z = g(X) + V, \quad \mathrm{E}[V|X] = 0, \tag{2b}$$

where the instrumental variable is Z. If Z is equal to F, this equation is identical to (1a).

Overcome Regularization Bias

Suppose an estimation of θ_0 by making use of a sample split. This sample split consists of two parts, a main part with l observations, these are indexed by $j \in J$, and an auxiliary part with $N - l$, indexed by $j \in J^c$. Given that $\hat{m}_0$ is estimated using ML methods and the auxiliary sample, this results in the following estimate:

$$\hat{\theta}_0 = (\frac{1}{n}\sum_{i \in I} F_i^2)^{-1}\frac{1}{n}\sum_{i \in I} F_i(Y_i - \hat{m}_0(X_i)) \tag{3}$$

Though, this estimator has a rate of convergence slower than $1/\sqrt{n}$:

$$|\sqrt{n}(\hat{\theta}_0 - \theta_0| \rightarrow_p \infty \tag{4}$$

This is caused by the bias in learning m_0. To overcome this problem, Chernozhukov et al. (2017) recommended the use of orthogonalization. First the effect of X from F is partialling out. The resulting variable is $\hat{Q} = F - \hat{g}_0(X)$. The last term is an Machine Learning estimator of $g_0(X)$ and is obtained using the auxiliary sample. As well as the naive approach to estimate θ_0, m_0 is also estimated using the auxiliary sample. Using the main sample of observations, this results in the following estimator for θ_0:

$$\check{\theta}_0 = (\frac{1}{n}\sum_{i \in I} \hat{Q}_i F_i)^{-1}\frac{1}{n}\sum_{i \in I} \hat{Q}_i(Y_i - \hat{m}_0(X_i)) \tag{5}$$

The difference between the orthogonal and non-orthogonal ML estimator is shown below.

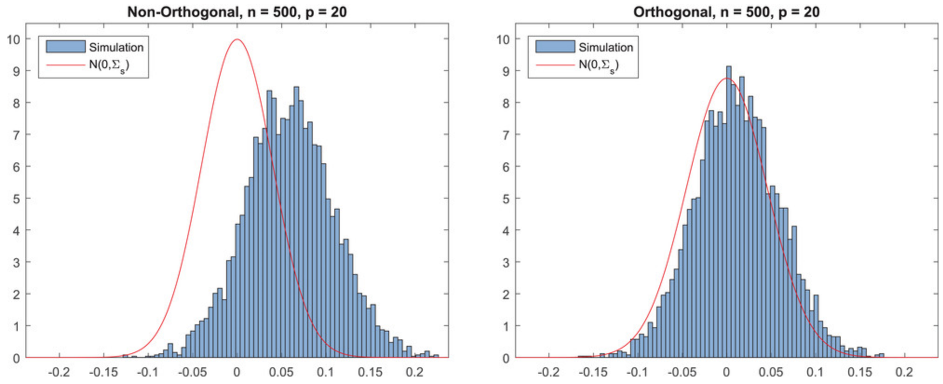

Figure 1: **Left Figure:** Simulation experiment of a non-orthogonal Machine Learning estimator, $\hat{\theta}_0$ from equation (3). This parameter is from the partial linear model in a simulation where the nuisance parameter m_0 is estimated using Random Forest. Because in this experiment there is used a small amount of variables, random forest is more favourable to use. As you can see, the blue histogram is shifted to the right, comparing to the normal approximation (red line). Thus it is clearly that there is a bias present. **Right Figure:** This is a simulation experiment of the orthogonal DML estimator, $\check{\theta}_0$ from equation (5). Here is also used Random Forest to estimate the nuisance parameters. As is clearly to see, the blue histogram is shifted to the left in comparing to the left figure and looks much more like the approximation of the normal distribution (red line). This figure is from Chernozhukov et al. (2017).

Clearly, the histogram is shifted to the right, this is the bias resulting from regularization. The right figure shows that the bias is almost completely removed by orthogonalization. A generalization of the orthogonalization principle is given by the Neyman orthogonality and moment conditions. The estimator $\check{\theta}_0$ can be viewed as a solution to

$$\partial_\eta \mathrm{E}\psi(W;\theta_0,\eta_0)[\eta-\eta_0]=0 \tag{6}$$

This equation is called the Neyman orthogonality and ψ is the Neyman orthogonal moment function. This orthogonality condition means that the used moment conditions to identify the parameter of interest are locally not sensitive to the value of the covariate, because of this one can plug in noisy estimations of these covariates without violating the moment conditions Chernozhukov et al. (2017).

To show that $\check{\theta}_0$ has good properties, the scaled estimation error of this estimator can be divided into three components, as shown in their paper:

$$\sqrt{n}(\check{\theta}_0-\theta_0)=a^*+b^*+c^*,$$

see the Appendix for the derivations, however, this is to show that it concentrates in a $N^{-1/2}$-neighbourhood of θ_0 and is approximately unbiased and normally distributed. The first term will satisfy $a^* \rightsquigarrow N(0,\Sigma)$ under some conditions. Moreover, the next term b^* takes the impact of regularization bias when estimating the two functions of X, this term can vanish when having a lot of data generating processes. At last, under weak conditions and making use of sample-splitting $c^*=o_p(1)$. This shows that $\check{\theta}_0$ has good properties.

Overcome Overfitting Bias

Besides the regularization bias, there is also a bias induced by overfitting. This bias can be removed by using sample splitting. This ensures that the remainder term, c^*, vanishes in probability. Chernozhukov et al. (2017) show that sample splitting allows to manage with this term. - Although, when estimating the parameter of interest using sample splitting, only the main sample is used. This could result in an extraordinary loss of efficiency, because only a part of the available data is used. They show that this loss can be removed by flipping the role of the main- and auxiliary sample, this results in a second estimator of the treatment variable of interest. It may regain full efficiency, when averaging these two resulting estimators. They call this procedure where they flip the roles of the auxiliary- and main sample to obtain more estimations and thereafter take the average of the results *cross-fitting*. This example is based on two splits, in general you can use a K-fold version of this process. How the bias results from overfitting vanish is shown below.

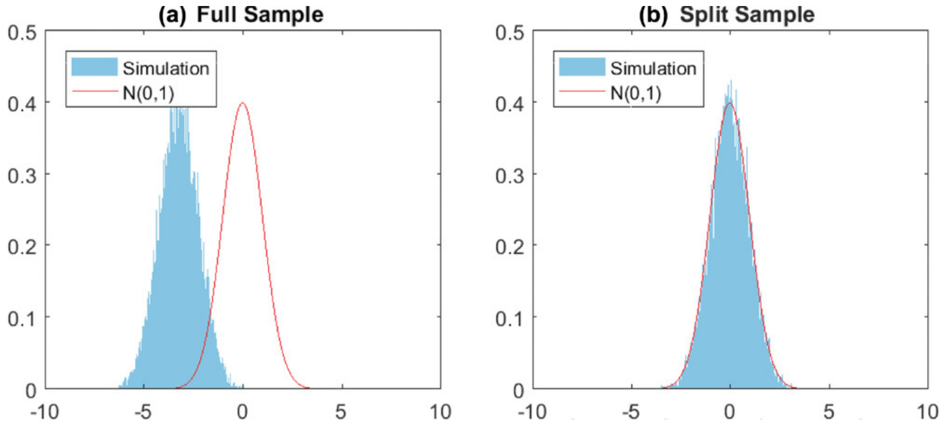

Figure 2: This figure shows the bias that is caused by overfitting. **Left Figure:** in this simulation experiment are the estimations of the nuisance parameters made with overfitting using the whole sample. The blue histogram shows the distribution of $\check{\theta}_0$ from equation (1a), and it is clearly to see that it is shifted to much to the left. This is the bias resulting from not making use of cross-fitting. **Right Figure:** in this simulation experiment there has been make use of sample-splitting in estimating $\check{\theta}_0$ from equation (1a), and as clearly to see the bias induced by overfitting is completely removed by making use of sample-splitting. This figure is from Chernozhukov et al. (2017).

The blue histogram in both figures shows a distribution for the studentized parameter of interest resulting from simulating using the full sample and a simple estimator, what makes sure that c^* explodes if sample-splitting is not used, for the covariates function. As you can see, the histogram is shifted to the left in comparison with the standard normal red line, this shift is the bias that results due to overfitting. At the other hand shows the Right Figure that the bias is completely removed by making use of sample splitting. Without sample splitting, it might lead to heavy biases in the estimators from the parameter of interest.

To apply the DML method, Chernozhukov et al. (2017) provide a code in the software R, which they use in their three empirical examples. In their first example they use the earlier described partially linear regression model, (1a) and in the third example they make use the partially linear regression model that allows for instrumental variable identification, (2a), these are the two formulas that will be used in this thesis. However, this code is fully constructed for their examples, thus this code has been changed in the way that it fits for the data sets of the other papers. In these papers they are very clear about which data they delete or which variables they split into dummy variables for example. This makes it easier to get the right data set for each regression.

2.2 Machine Learning Methods

In Chernozhukov et al. (2017) they applied several Machine Learning methods in their empirical examples, such as Lasso, Regression Trees, Boosting, Random Forest and a hybrid method Best. In this thesis these methods are used as well. To make these methods more clear, below is given a short description about these methods. Athey and Imbens (2015)

LASSO
Least Absolute Selection and Shrinkage Operator Tibshirani (1996) is a Machine Learning method that can be applied when having a linear regression with many regressors. It minimizes the sum of squared residuals with an additional term:

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \|\beta\|, \tag{7}$$

where $\|\beta\| = (\sum_{k=1}^{K} |\beta_k|)$. Rewrite (7) as follows to choose the penalty parameter:

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 \tag{8a}$$

$$s.t. \sum_{k=1}^{K} |\beta_k| \leq t \cdot \sum_{k=1}^{K} |\hat{\beta}_k^{ols}| \tag{8b}$$

where t is a scalar between zero and one. When t equals zero, it is easy to see that all estimates shrinks to zero. When t is equal to one, the estimates are not shrinking and it is just OLS. The penalty parameter λ in (7) or t in (8a) are chosen through cross-validation. In Econometrics LASSO is the most popular Machine Learning method. It has the advantage that is sets the coefficients that are not relevant equal to zero. This makes it easier to give an interpretation about the coefficients.

Regression Trees
This is a nonparametric regression and it is first introduced by Breiman et al. (1984). This is a method where the covariate space is divided into subspaces. The regression function is estimated as the average from the subspaces' outcomes. However, the sum of squared deviations will be minimized and is defined by:

$$Q(g) = \sum_{i=1}^{N} (Y_i - g(X_i))^2, \tag{9}$$

The start estimate of $g(X_i)$ is $\overline{Y}$. The data will be split depending on whether $X_{i,l} \leq t$ or $X_{i,l} > t$, for covariate l and threshold t. This split results into two averages, the average over the values where $X_{i,l} \leq t$, $\overline{Y}_{\text{left}}$ and the average over the the values that satisfied $X_{i,l} > t$, $\overline{Y}_{\text{right}}$. The estimator is defined as follows:

$$g_{l,t}(x) = \begin{cases} \overline{Y}_{\text{left}} & \text{if } x_l \leq t \\ \overline{Y}_{\text{right}} & \text{if } x_l > t \end{cases} \tag{10}$$

The optimal covariate and threshold results from:

$$(l^*, t^*) = \arg\min_{l,t} Q(g_{l,t}(x)) \tag{11}$$

The partition is constructed by whether $X_{i,l^*} \leq t^*$ or $X_{i,l^*} > t^*$. Repeat this procedure until equation (9) with a penalty for the number of splits is no longer minimized. The penalty term is selected in about the same way as the LASSO penalty term. The sample is also divided into B cross-validation samples. If the tree grows, use the full sample excluding the b-th cross validation sample, $g(b, \lambda)$, to get the squared errors over the cross-validation sample for each λ:

$$Q(\lambda) = \sum_{b=1}^{B} \sum_{i:I_i=b} (Y_i - g(b, \lambda))^2 \tag{12}$$

The penalty term is λ and minimizes (12). However, sometimes the growing tree is stopped too early, the way to fix this is by pruning the tree. A solution can be to grow a big tree, this can be done in several ways. For example, by making use of a small value for the penalty term. Another way is to grow the tree till the leaves have a pre-arranged small amount of observations. Thereafter, prune leaves until the objective function is no longer sufficiently improved.

Boosting
Weak learners are estimations of regression functions in a simple, and easy to compute, way. Examples are trees, kernels and neural networks. Boosting uses these weak learners over and over again to get a nice predictor for both regression and regression problems.

Random Forests
When applying Random Forests, draw several bootstrap samples from the data and start, given a bootstrap sample, with a single leaf tree. Then, select randomly L covariates out of the K covariates. Among this subset consisting of L covariates, select the optimal covariate and threshold. Set a certain minimum of units, if some leaves have more than this amount of units, start the procedure over. Otherwise average the trees over the bootstrap samples.

Best
The method Best selects the best performing Machine Learning methods when estimating nuisance functions. This method is based on the average out-of-sample prediction performance for the dependent variable. This prediction performance is obtained by estimating each nuisance function with each of foregoing described methods. Interesting to notice is that when one method performs it best in estimating each nuisance function, then the estimates of Best would be the same as the estimates of that method.

2.3 The slave Trade and the Origins of Mistrust in Africa

Following Nunn and Wantchekon (2011), below is their used baseline estimation equation:

$$trust_{j,f,e,d} = \alpha_d + \beta slave\ exports_f + \mathbf{X}'_{j,f,e,d}\mathbf{\Phi} + \mathbf{X}'_{e,d}\mathbf{\Gamma} + \mathbf{X}'_f\mathbf{\Omega} + \varepsilon_{j,f,e,d}, \tag{13}$$

in this equation d means from which country the respondent is coming from, e the district from the respondent, f the ethnic group and j the individual and $trust$ is either Trust in Neighbours or Trust in relatives or Trust in local council or Intergroup trust or Intragroup trust. The constant α_d are representing the country fixed effects, *slave exports*$_f$ is the measure for the slave export. Different kinds of control variables are used in the tables, $\mathbf{X}'_{j,f,e,d}$ represents the individual controls, $\mathbf{X}'_e$ represents the colonial population density and $\mathbf{X}'_f$ ethnicity-level colonial variables.

Thus, when comparing this equation with equation (1a): *trust* is equal to Y, β equals θ_0, $\alpha_c + \mathbf{X}'_{j,e,d,c}\mathbf{\Gamma} + \mathbf{X}'_{d,c}\mathbf{\Omega} + \mathbf{X}'_e\mathbf{\Phi}$ is the same as $g_0(X)$ and at last ε equals U.

In this thesis there are seven tables that are replicate from above paper, namely, table 1, 2, 3, 5, 6, 9 and 10. Table 1 shows the OLS estimates of the relationship between different measures of slave export and trust in neighbours, table 2 reports the OLS estimations of the relationship between a measure of slave export on multiple types of trust and table 3 shows the OLS estimates of also this relationship, though, in these regressions are used additional controls. Because this thesis replicate the important results from the paper, OLS must also be applied on equation (13) with the same controls variables as the original tables. Table 5 shows the IV estimations of the relationship between the slave trade and trust, table 6 shows the IV estimates of also this relationship, but there are used additional controls. The used instrument is defined by the distance of the ethnic group of an individuals from the coast during the slave period. At last, tables 9 and 10 are replicate, these tables report the OLS estimates for identifying channels of causality. All these OLS and IV estimates can be done using the software STATA. Within this program, you can give the in- and dependent variable(s) and all the covariates, the program returns afterwards all the estimates. For almost all tables the same results are obtained. However, when replicating table 10, the estimations of the external channel differ a bit, see section 4

2.4 The Long-Term Effects of Africa's Slave Trades

Nunn (2008) uses the following baseline estimating equation:

$$ln\ z_j = \alpha + \beta\ ln(export_j/area_j) + \mathbf{D}'_j\gamma + \mathbf{E}'_j\delta + \varepsilon_j, \tag{14}$$

in this equation is the outcome variable the natural log of real per capita GDP in country j in 2000, the parameter of interest is $ln(exports_j/area_j)$ and is a measure for the slave export. The first set of controls is $\mathbf{D}_i$ and represents who the colonizer was in each country. The other set of covariates is $\mathbf{D}_j$ and represent the differences in the climate and geography of countries. The parameter of interest is β. It denotes the estimated effect of slave exports on income. Thus, when comparing this with equation (1a), $ln\ z$ is the same as Y, β equals θ_0 and $\alpha + \mathbf{D}'_j + \mathbf{E}'_j$ is the same as $g_0(X)$, and at last ε_j equals U.

From this paper, table 3 has been replicated. This table shows the OLS estimations of the effect of slave exports on income. When replicating this table, just like the other paper, the software STATA is used when making the estimations.

3 Data

The Slave Trade and the Origins of Mistrust in Africa
To estimate the effect of the slave trade on the trust of descendants, has been used of the data set of Nunn and Wantchekon (2011) [1]. This data comes from the 2005 Afrobarometer surveys, the questions in these surveys are shown in the local languages from and are given in each country to a random sample of either 1,200 or 2,400 individuals. These surveys are given in 17 sub-Saharan countries. In total, 21,822 individuals completed the surveys, from this sample, 120 of the respondents are

[1] Source: `https://www.aeaweb.org/article?id=10.1256/aer.101.7.3121`

removed, these respondents have given wrong answers, remaining with 21,702 potential observations. The data set contains each dependent variable, the measures for trust, and all covariates that is used in their paper. This data set is used to replicate the results from their paper and to apply DML on it.

The long-term effects of Africa's slave trades
The authors' website provided the whole data set from this paper[2]. This data set contains some characteristics and control variables from 52 countries. These variables are used in the OLS and DML regressions. For each country, it contains the real per capita GDP, this data is used to calculate some variables, such as $ln(export/area)$. Moreover, it contains the colonizer fixed effects, this are dummy variables which indicates the colonizer's identity at the time of independence. Moreover, there are is a set of control variables, like average minimum temperature and a north Africa indicator, to capture the differences in climate and geography of the countries.

4 Replication Results

The most important results of Nunn and Wantchekon (2011) and Nunn (2008) are replicate. After this, DML is applied on some of these tables. These are table 1, 5 and 10 from the first paper and table 3 from latter paper. That is why this section only shows these replication tables. The tables 2, 3, 6 and 9 from Nunn and Wantchekon (2011), are shown in the Appendix. In all tables: *** means a significance level of 10 percent, ** means a significance level of 5 percent and * means a significance level of 1 percent.

4.1 The Slave Trade and the Origins of Mistrust in Africa

Table 1
This table shows OLS estimations of the relationship between different measures for slave export and trust in neighbours. The outcome variable is trust in neighbours and the different measures of the slave trade are shown in the columns. In each regression they use country fixed effects and individuals-, districts controls. The first column has the disadvantage that it does not account for any differences in the magnitude of ethnic groups. The next column normalizes the slave export measure by the inhabitation of each ethnic group after the slave trade period. The next column uses a measure that normalizes the slave exports by the area of land where an ethnic group lives.The remaining columns report the natural logarithm of the measures of the previous columns plus one.

As you can see, these results are the same as their findings, resulting in the same conclusions. All the estimated coefficients are negative and highly significant, which is consistent with their hypothesis that the slave trade is negative related to the trust of descendants. For the rest of their paper they use for the measure of slave export the specification from column 6, $ln(1 + export/area)$, this measure is precise and available for every ethnic group in their data set.

Table 1: Replication Table 1

Dependent variable: Trust of neighbours	Slave export (thousands)	Export/ historical pop	Exports/ area	ln(1+export)	ln(1+export/historical pop)	ln(1+export/area)
OLS estimation	$-0{\cdot}00068^{***}$	$-0{\cdot}531^{***}$	$-0{\cdot}019^{***}$	$-0{\cdot}037^{***}$	$-0{\cdot}743^{***}$	$-0{\cdot}159^{***}$
SE adj clust ethn	0·00014	0·147	0·005	0·014	0·187	0·034
SE adj two-way clust	0·00015	0·147	0·005	0·014	0·187	0·034
SE Conley (1999)	0·00013	0·165	0·005	0·015	0·212	0·034
Used controls:						
Individual	Yes	Yes	Yes	Yes	Yes	Yes
District	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Specifications:						
Observations	20,027	20,027	20,027	20,027	17,644	20,027
Districts	1,257	1,214	1,257	1,257	1,214	1,257
Ethnicities	185	157	185	185	157	185
R^2	0.16	0.15	0.16	0.15	0.15	0.16

This table shows OLS estimations. Below each estimated coefficient are three different standard errors displayed. Below the OLS estimates are shown the used controls and specifications of the regressions.

Table 5

This table reports the IV estimates of the relationship between the slave trade and trust. Using IV requires an instrumental variable that is correlated with the variable slave export and uncorrelated with any ethnic group's characteristics that may affect one of the five measures of trust. This paper uses the instrument 'historical distance of ethnic group from coast', because this

[2] Source: `https://scholar.harvard.edu/nunn/pages/data-0`

instrument captures an ethnic group's exposure to the external demand for slaves, since slaves were purchased at the coast before being shipped overseas. Moreover, it is likely that this distance is not correlated with any other thing that affects any of the five measures of trust. These estimates are controlled for the ethnicity-level colonial-, district- and individual- controls, country fixed effects and density of the population, besides the columns represent the five measures of trust.

Just like the estimates in table 5, the same estimations are produced. However, there are two other standard errors, for the dependent variables Trust of Neighbours and Intergroup Trust (both in bold). This does not make huge changes in the conclusions, the differences are only 0.001. All the first stage estimates are negative, implicating that the historical distance of ethnic group from coast is negatively related with slave exports, this is consistent with the historical record, the ethnic groups that have lived further away from the coast exported fewer slaves. The second stage estimates are all negative and highly significant, this means as well as the latter table that the slave exports is negatively correlated with the trust of descendants.

Table 2: Replication Table 5

	Trust of neighbours	Trust of relatives	Trust of local council	Intergroup Trust	Intragroup Trust
Second stage: Dependent variable is one of above trust					
ln(1 + export/area)	$-0{\cdot}245^{***}$	$-0{\cdot}190^{***}$	$-0{\cdot}221^{***}$	$-0{\cdot}174^{***}$	$-0{\cdot}251^{***}$
SE adj two-way clust	**0.701**	$0{\cdot}067$	$0{\cdot}060$	**0.081**	$0{\cdot}088$
Hausman test (p-value)	$0{\cdot}53$	$0{\cdot}88$	$0{\cdot}09$	$0{\cdot}41$	$0{\cdot}44$
R^2	$0{\cdot}15$	$0{\cdot}13$	$0{\cdot}20$	$0{\cdot}12$	$0{\cdot}15$
First stage: Dependent variable is ln(1 + export/area)					
Instrumental Variable	$-0{\cdot}0014^{***}$	$-0{\cdot}0014^{***}$	$-0{\cdot}0014^{***}$	$-0{\cdot}00014^{***}$	$-0{\cdot}00014^{***}$
SE adj two-way clust	$0{\cdot}0003$	$0{\cdot}0003$	$0{\cdot}0003$	$0{\cdot}0003$	$0{\cdot}0003$
Used Controls:					
District	Yes	Yes	Yes	Yes	Yes
Individual	Yes	Yes	Yes	Yes	Yes
Ethnicity-level colonial	Yes	Yes	Yes	Yes	Yes
Colonial population density	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Specifications:					
Clusters	147 / 1,187	147 / 1,187	146 / 1,194	147 / 1,184	147 / 1,186
Observations	16,679	16,709	15,905	16,473	16,636
F-stat of excl. instrument	26.8	26.9	27.4	27.0	27.1
R^2	0.81	0.81	0.81	0.81	0.81

This table shows Instrumental Variable estimates. Below each OLS estimate is shown the standard error. The used instrumental variable is Historical distance of ethnic group from coast. Moreover, are shown the used controls and specifications of the regressions.

Table 10

In this table there will be tests for channels of causality. The previous tables shows a negative relationship between the slave trade and the trust of descendants. In their paper, they give a possible second explanation why these two factors are correlated: "because it resulted in a deterioration of preexisting states, institutions, and legal structures." To estimate the effect of those two factors, there is a new slave export measure that is location-based. This measures the slave export from the area where the individual is living. Therefore, including these measures in the estimation equation, you can make more easily a distinguish between them.

When replicating the results from Nunn and Wantchekon (2011), there was happening something weird. The estimates of the baseline measure, ethnicity-based slave export measure, and its standard errors corresponded. Although, the estimates of the location-based slave export measure differs in almost each case, it differs in most cases not a lot, about 10 percent, it is strange that the estimates are not the same. Despite this difference, the OLS estimates remain significant. This leads to the same conclusions as in their paper. Because the negative estimates of the slave export measure that is location-based, it implies that location-based slave export affects the trust of descendants. However, in almost every case the ethnicity-based measure is twice as big as the location-based measure, suggesting that the primal measure is more important.

Table 3: Replication Table 10

	Trust of Neighbours	Trust of relatives	Trust of local council	Intergroup	Intragroup
Internal channel	$-0{\cdot}182^{***}$	$-0{\cdot}155^{***}$	$-0{\cdot}100^{***}$	$-0{\cdot}090^{***}$	$-0{\cdot}169^{***}$
SE adj two-way clust	$0{\cdot}029$	$0{\cdot}029$	$0{\cdot}023$	$0{\cdot}030$	$0{\cdot}033$
External channel	**-0.041****	**-0.058*****	**-0.068*****	$-0{\cdot}047^{**}$	**-0.039***
SE	**0.019**	**0.016**	**0.017**	**0.024**	**0.022**
Used controls:					
Baseline controls	Yes	Yes	Yes	Yes	Yes
Ethnicity-level colonial controls	Yes	Yes	Yes	Yes	Yes
Colonial population density	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Specifications:					
Clusters	146/269	146/269	145/272	146/269	146/269
Observations	15,972	15,999	15,221	15,773	15,931
R^2	0.16	0.13	0.20	0.12	0.16

This table shows OLS estimates. Below each OLS estimate is shown the standard error. The used Internal channel is a slave export measure based on ethnicity and the used External Channel is a slave export measure based on location. Moreover, are shown the used controls and specifications of the regressions.

4.2 The Long-Term Effects of Africa's Slave Trades

Table 3

This table shows the relationship between the slave export and the current economic performance. Just like as Nunn and Wantchekon (2011), the total number of slaves is normalized by land area. Every regression contains fixed effects of the colonizer, moreover the used control variables vary over the regressions as you can see. The same estimations are produced, resulting in the same conclusions. All the estimates of the parameter of interest, $ln(export/area)$, are negative and highly significant, suggesting a negative relationship between slave export and the individuals' income. This is also economically meaningful, when for example calculating the standardized beta coefficients of the estimates. If the standard deviation in $ln(export/area)$ with one increases, the standard deviation in $log\ income$ decreases with between 0.36 to 0.62.

Table 4: Replication Table 3

Outcome variable: log real per copita			Regressions:			
GDP in 2000	(1)	(2)	(3)	(4)	(5)	(6)
ln(export/area)	−0·112***	−0·108***	−0·076***	−0·085**	−0·128***	−0·103***
SE	0·024	0·037	0·029	0·035	0·034	0·034
Used controls:						
ln(coastline/area)		0·092**	0·085**	0·095**	0·083**	0·082**
SE		0·042	0·039	0·042	0·037	0·040
Distance from equator		−0·005	0·016	0·019	0·006	0·023
SE		0·020	0·017	0·018	0·017	0·017
Avg min temperature		−0·039	−0·019	−0·005	−0·037	−0·015
SE		0·028	0·028	0·027	0·025	0·026
Lowest monthly rainfall		0·008	−0·001	0·0001	−0·002	−0·001
SE		0·008	0·007	0·007	0·008	0·006
Longitude		−0·007	0·001	−0·004	−0·009	−0·004
SE		0·006	0·005	0·006	0·006	0·005
Avg max humidity		0·008	0·009	0·009	0·013	0·015
SE		0·012	0·012	0·012	0·010	0·011
Island indicator				−0·398		−0·150
SE				0·529		0·516
French legal origin				0·755	−0·141	0·643
SE				0·503	0·734	0·470
Percent Islamic				−0·008***	−0·003	−0·006*
SE				0·003	0·003	0·003
North Africa indicator				0·382		−0·304
SE				0·484		0·517
ln(gold prod/pop)					0·014	0·011
SE					0·015	0·017
ln(diamond prod/pop)					−0·048	−0·039
SE					0·041	0·043
ln(oil prod/pop)					0·088***	0·078***
SE					0·025	0·027
Colonizer fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Specifications:						
R^2	0.51	0.63	0.60	0.71	0.80	0.77
Number obs.	52	42	52	52	42	52

Notes: This table shows OLS estimates. The standard errors are shown below each OLS coefficient. Moreover, are shown the used controls and specifications of the regressions.

5 DML Results

The results in the DML tables 5, 6 and 7 are based on four methods for estimating the nuisance functions of the partially linear model, see equation (1a). There are three tree-based methods used: Trees, Forest and Boosting. In table 8, additional the ℓ_1-penalization based method Lasso is used. As in Chernozhukov et al. (2017) almost the same parameters and techniques are used, thus for the method Trees, to estimate each nuisance function there has been make use of a single Classification And Regression Tree (CART). The penalty parameter is using 10-fold cross-validation. Besides, the estimates obtained with Forest, are based on estimating each nuisance function using a random forest. This random forest has a minimum of 1000 trees. The method Boosting has been make use of boosted regression trees. The regularization parameters are, just like by Trees, obtained using 10-fold cross-validation. Bellonii et al. (2012) developed a selection rule that is used to obtain the penalty parameter of the method Lasso. The method Best selects the best performing Machine Learning methods for estimating nuisance functions. This selection is based on the average out-of-sample prediction performance for the dependent variable. This prediction performance is obtained by estimating each nuisance function with each of foregoing described methods. All the estimates in the tables are obtained using 2-fold cross-fitting and using 2 sample splits, with exception of table 8. In this table all the estimates are obtained with 100 sample splits. The first row of each table represents my replication of the parameter of interest, because that makes it easier to compare the DML results with the original OLS and IV estimates. In all tables: *** means a significance level of 10 percent, ** means a significance level of 5 percent and * means a significance level of 1 percent.

5.1 Main Results The Slave Trade and the Origins of Mistrust in Africa

Table 1

This table shows in the first row the OLS estimates of the relationship between trust in neighbours and different measures for slave trade. As you can see, the same estimations are produced as their paper. The estimates in each column represents the impact of slave trade. To define this, the table uses trust of neighbours. Thus the estimate in column 1 shows the effect from the total number of slaves on the trust of neighbours. All the estimates are negative and highly significant, which is consistent with their hypothesis that the slave trade adversely affected individuals' trust of those around them.

Table 5: DML Table 1

Outcome variable:	Slave export	Export/	historical	ln(1+export/	ln(1+export/historical pop)	ln(1+export/area
Trust of neighbours	(thousands)	historical pop	area			
Ordinary Least Squares	−0·00068***	−0·531***	−0·019***	−0·037***	−0·743***	−0·159***
	(0·00014)	(0·147)	(0·005)	(0·014)	(0·187)	(0·034)
Trees	−0·00045***	−0·332***	−0·014***	−0·030***	−0·533***	−0·118***
	(0·00004)	(0·040)	(0·001)	(0·005)	(0·062)	(0·010)
Forest	−0·00030***	−0·279***	−0·010***	−0·019***	−0·455***	−0·092***
	(0·00006)	(0·061)	(0·002)	(0·006)	(0·086)	(0·014)
Boosting	−0·00055***	−0·461***	−0·015***	−0·035***	−0·652***	−0·131***
	(0·00004)	(0·047)	(0·001)	(0·004)	(0·057)	(0·009)
Best	−0·00044***	−0·362***	−0·013***	−0·022***	−0·455***	−0·092***
	(0·00006)	(0·066)	(0·002)	(0·006)	(0·086)	(0·014)
Trees†	−0·00045***	−0·305***	−0·015***	−0·023	−0·455***	−0·100***
	(0·00004)	(0·043)	(0·001)	(0·046)	(0·057)	(0·011)

This table shows DML method estimates of the parameter of interest and at the first row the OLS estimates. Below each coefficient the standard error is reported. There are two splits used. The rows denote which method is used to estimate nuisance functions. The columns denotes the variable of interest. Best is based on the methods Trees, Forest and Boosting. Trees† used 100 splits instead of 2.

About all of the estimating results are negative and highly significant. This is consistent with the findings of Nunn and Wantchekon (2011). Comparing the method Best with Trees†, there is hardly no difference in the estimates, although the standard errors differ few times.

It is interesting to notice that even though the results are still negative and highly significant, the estimates differ relatively much between DML and OLS. For example the first column, the parameter of interest is Slave Exports. When applying DML the result varies between -0.00030 and -0.00055, while applying OLS the estimate is -0.00068, this is a percentage difference of more than 20%. Moreover, there is even a greater difference in the standard errors. The standard errors of DML are about 0.00005, while the standard error of OLS is equal to 0.00014, this is a difference of a factor 3. Leading to less extreme conclusions, because the DML estimates are all closer to zero than OLS. For example the effect of $ln(1 + export/area)$ on trust of neighbours: they found that when $ln(1 + export/area)$ increases one, the trust of neighbours decreases with about 0.159, however DML shows that the decrease will be about 0.100. The confidence interval will also be more precise: the 95% confidence interval of OLS is about $(-0.227, -0.091)$, while the confidence interval of DML is about $(-0.122, -0.079)$

Table 5
This table shows the estimates using IV and using DML with instrumental variables. The columns represent the dependent variable, a measure for individual trust. The used instrument is historical distance of ethnic group from coast. This instrument is correlated with slave export and uncorrelated with any ethnic group's characteristics that may affect one of the five measures of trust. These estimates represents the estimations of the parameter of interest $ln(1 + export/area)$.

Table 6: DML Table 5

Second stage: Dependent variable is one of the following trust	Trust of relatives	Trust of neighbours	Trust of local council	Intragroup trust	Intergroup trust
Instrumental variables	−0·190***	−0·245***	−0·221***	−0·251***	−0·174***
	(0·067)	(**0.071**)	(0·060)	(0·088)	(**0.081**)
Trees	−0·957***	0·075	−0·570	−0·233	−0·550
	(0·249)	(0·197)	(0·416)	(0·247)	(0·368)
Forest	−1·308	−0·459	−1·151	−1·425	−0·444
	(1·494)	(0·753)	(1·411)	(1·516)	(0·707)
Boosting	−0·417***	−0·483***	−0·457***	−0·491***	−0·404***
	(0·111)	(0·114)	(0·125)	(0·113)	(0·116)
Best	−2·056	−0·583	−2·057	−2·225	−0·856
	(1·492)	(0·752)	(1·402)	(1·518)	(0·706)
Trees†	−1·076***	−0·043	−0·155	−0·847***	−0·333
	(0·238)	(0·298)	(0·259)	(0·281)	(0·290)

This table shows DML method estimates of the parameter of interest and at the first row the IV estimates. Below each coefficient the standard error is reported. There are two splits used. The rows denote which method is used to estimate nuisance functions. The variable of interest is ln(1 + exports/area), and the used instrument is the historical distance of ethnic group from coast. Best is based on the methods Trees, Forest and Boosting. Trees† used 100 splits instead of 2.

First of interesting is to notice that the used DML methods vary a lot. Comparing Trees, Forest and Boosting: the range of trust of relatives is between -1.308 and -0.417, more remarkable is the range of trust of neighbours, which varies between -0.483 and even a positive estimation 0.075. The interpretation of this positive estimate is that when slave exports increases, the trust of neighbours also increases, and that is intuitively unreliable. This may be caused by by using a sample split of only 2. The method Best shows highly negatively estimates, though they have a relatively large standard error. It is more confident to compare Trees† with the replication results, because these estimations are based on 100 splits. Although these estimates are all negative, only two are significant. The 95% confidence interval for trust of relatives using the method IV is $(-0.324, -0.056)$, while the confidence interval using Trees† is equal to $(-1.552, -0.6)$. This makes their hypothesis, that slave exports are negatively correlated with trust of descendants, more reliable.

Table 10
The previous tables show that the slave trade has a negative effect on the trust of descendants. In the paper they have a possible second explanation why these two factors are correlated. In this table there will be test for channels of causality. In their words, there will be a direct estimate how much of the slave trade's effect on trust works through an individual's external environment-such as the rule of law and the trustworthiness of others- versus through individual's internal norms of mistrust. To estimate the effect of those two factors, there is a new slave export measure introduced. This measure is location-based, thus it measures the total number of slaves exported from and individual's residence. Therefore, including these measures in the estimation equation, you can easily make a distinction between them.

Within each used method, the first two rows indicate the estimates of internal channel, the slave export measure that is based on ethnicity, and below the standard errors. The second two rows represent the external channel, the slave export measure that is based on location, and below the standard errors.

Table 7: DML Table 10

	Trust of relatives	Trust of neighbours	Trust of local council	Intragroup trust	Intergroup trust
Ordinary Least Squares	−0·155	−0·182	−0·100	−0·169	−0·090
	(0.029)	(0·029)	(0·023)	(0·033)	(0·030)
	-0.058	**-0.041**	**-0.068**	**-0.039**	−0·047
	(0.016)	**(0.019)**	**(0.017)**	**(0.022)**	**(0.024)**
Trees	0·187	0·118	0·055	0·015	0·264
	(0.029)	(0·030)	(0·028)	(0·031)	(0·030)
	−0·042***	−0·030**	−0·044***	−0·045***	−0·028**
	(0·014)	(0·014)	(0·015)	(0·013)	(0·014)
Forest	−0·135	−0·133	0·197	−0·093	0·079
	(0.107)	(0·147)	(0·085)	(0·164)	(0·128)
	−0·013	−0·003	−0·026	−0·017	−0·027
	(0·018)	(0·017)	(0·018)	(0·017)	(0·016)
Boosting	−0·098***	−0·114***	−0·060*	−0·085***	0·021
	(0.030)	(0·031)	(0·032)	(0·030)	(0·030)
	−0·036**	−0·027*	−0·041***	−0·035**	−0·047***
	(0·015)	(0·014)	(0·015)	(0·014)	(0·014)
Best	−0·147	−0·133	0·106	−0·108	0·079
	(0.103)	(0·147)	(0·083)	(0·172)	(0·128)
	−0·019	−0·010	−0·029*	−0·017	−0·027
	(0·018)	(0·017)	(0·018)	(0·017)	(0·016)
Trees†	0·173	0·190	0·097	−0·009	0·100
	(0.028)	(0·032)	(0·033)	(0·028)	(0·029)
	−0·050***	−0·022	−0·045***	−0·044***	−0·036***
	(0·014)	(0·014)	(0·015)	(0·013)	(0·013)

This table shows DML method estimates of the parameter of interest and at the first row the OLS estimates. Below each coefficient the standard error is reported. There are two splits used. The rows denote which method is used to estimate nuisance functions. The variables of interest are the slave export measure that is ethnicity based and the slave export measure that is location based. Best is based on the methods Trees, Forest and Boosting. Trees† used 100 splits instead of 2.

Remarkable are the estimates of the method Trees. Each estimation of the baseline measure is positive, suggesting that slave exports positively affects the trust of descendants. The other two 2-splits methods look more reliable. The estimations of the method Best are more reliable, however they still have a relatively large standard error. The baseline estimates of Best are in most cases negative and insignificant. At the other hand the estimates of the location-based measure are all negative and insignificant. More interesting are the estimation results of Trees†. The estimations of the baseline measure are, with one exception, all positive and significant, suggesting that slave exports has affected positively trust through internal norms of trust. Moreover, the estimations of the location-based measure are all negative and, with one exception, all significant, suggesting that the slave exports has a negative relationship with trust through an external environment. The previous tables shows that slave export is negatively correlated with trust of descendants. But, because the baseline measure estimations are positively insignificant and the location-based measure estimations are negatively significant, suggesting that although baseline is positive, the external channel is more important, what differs from their conclusion.

This is very remarkable, it implies that the location-based slave export measure is more important. Although in their whole paper they used the ethnicity-based slave export measure as baseline. Concluding, these results suggest that in each case the external channel had to be used instead of the internal channel.

5.2 Replication Main Results The long-term effects of Africa's slave trades

Table 3
This table shows the relationship between the slave exports and the current economic performance. Just like in Nunn and Wantchekon (2011), is the slave export normalized by land area. Each column has other regressors, see previous section for the used regressors. The first row shows my replication of this paper. As you can see, the same estimations are produced. These estimates are negative and highly significant, suggesting a negative relationship between the current economic performance and slave export.

Table 8: DML Table 3

Outcome variable: log real per copita	Regressions:					
GDP in 2000	(1)	(2)	(3)	(4)	(5)	(6)
Ordinary Least Squares	−0·112***	−0·076***	−0·108***	−0·085**	−0·103***	−0·128***
	(0·024)	(0·029)	(0·037)	(0·035)	(0·034)	(0·034)
RLasso	−0·118***	−0·118***	−0·104***	−0·104***	−0·119***	−0·107***
	(0·025)	(0·026)	(0·029)	(0·027)	(0·025)	(0·029)
Trees	−0·118***	−0·095***	−0·104***	−0·084***	−0·133***	−0·107***
	(0·025)	(0·026)	(0·029)	(0·029)	(0·025)	(0·029)
Forest	−0·123***	−0·090***	−0·112***	−0·083**	−0·112***	−0·106***
	(0·023)	(0·034)	(0·036)	(0·033)	(0·028)	(0·033)
Best	−0·123***	−0·090***	−0·112***	−0·083**	−0·112***	−0·106***
	(0·024)	(0·034)	(0·036)	(0·033)	(0·028)	(0·033)

The table shows DML method estimates of the parameter of interest and at the first row the OLS estimates. Below each coefficient the standard error is reported. There are two splits used. The rows denote which method is used to estimate nuisance functions. The columns denotes the variable of interest. Best is based on the methods RLasso, Trees and Forest.

All these DML methods are based on 100 splits. Comparing the first three used DML methods, the estimations are approximately comparable. They differ at most about -0.02. In column (1) the estimations have a range between -0.123 and -0.118. All these estimates are negative and significant, this suggests that slave exports is negatively correlated with the current economic performance. The estimates of Best are also all negative and significant, consisting with the findings in Nunn (2008). Almost all the estimations are more negative than the replication results. They differ between 2 and at most 18 percent. Besides, in almost each case the magnitude of the standard errors decreases. For example in column (1), Best has a 95% confidence interval of $(-0.171, -0.075)$, while OLS has a confidence interval of $(-0.16, -0.064)$, which is less negative. Thus, these DML estimates are consistent with the findings in the Long-Term Effects of Africa's Slave Trades, however the DML estimates are more reliable.

6 Discussion

There have been made certain choices when conducting my research about applying DML to the data sets of two papers. For example, the Machine Learning method Neural Networks did not work, caused by the error ”0 (non-NA) cases”. This can be solved by removing dummy variables. Their was the choice by either applying one additional ML method that might improve Best or not applying Neural Networks. Because the important results from the two papers are also replicate and the comparison between the DML results and the replication results have to be done fairly, the second option is chosen. Moreover, this method can simulate strong non-linear relationships. However, it takes a lot of time and knowledge to tune the network and a lot of computation time to learn the network, this can take hours to months. The result is a true simulation, but there is no information about the confidence from the resulting estimator. In the context of this thesis, the use of the other Machine Learning methods is more relevant.

In Nunn (2008) they are applying Instrumental Variables with four instruments. However, the provided code in the software R is only able to use a single instrument. It was still possible to apply DML with a single instrument a time, but this makes it impossible to compare the DML results with the replication results in a good way.

Furthermore, it may be of interest to apply DML using 100 splits instead of 2 splits. By making use of 100 splits the bias may decrease more due to overfitting, then by making use of 2 splits. However, 100 splits are associated with much longer computation time. Moreover, the ML method Trees is a relative fast method, that is why every DML table has also an additional second Trees method with 100 splits. In this thesis, there is shown in general that it is more confident to use DML when estimating with multiple covariates instead of OLS.

7 Conclusion

Replicating the most important results of the two papers did not deliver surprising conclusions: the same estimations are reproduced, with few exceptions. However, these exceptions did not alter their conclusions. The aforementioned DML results of this thesis imply more reliable estimations and confidence intervals. Though the conclusions are not altered by these estimations, they do improve the reliability of the estimations. Besides constructing these estimations and confidence intervals, There is also show that Nunn and Wantchekon (2011) have used the wrong slave-export variable. The total number of slaves exported from and individual's residence is more important than the variable they primary used. The main conclusion on slave exports is that their results may not be reliable.

References

S. Athey and G. Imbens (eds.). *An introduction to Supervised and Unsupervised Learning*, July 2015. NBER Lectures. URL http://www.nber.org/econometric_minicourse_2015/.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, May 2014. doi: 10.1257/jep.28.2.29. URL http://www.aeaweb.org/articles?id=10.1257/jep.28.2.29.

A. Bellonii, D. Chen, V. Chernozhukovv, and C. s. Sparse models and methods for optimal instrument with application to eminent domain. *Econometric*, 80(6):2369–2429, 2012. doi: doi:10.3982/ECTA9626. URL https://doi.org/10.3982/ECTA9626.

Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor and Francis, 1984. ISBN 978-0412048418.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economics Review*, 107(5):261–65, May 2017. doi: 10.1257/aer.p20171038. URL http://www.aeaweb.org/articles?id=10.1257/aer.p20171038.

T. G. Conley. GMM estimation with cross sectional dependence. *Journal of Econometric*, 92(1):1–45, September 1999. URL http://www.sciencedirect.com/science/article/pii/S0304407698000840.

Nathan Nunn. The long-term effects of africa's slave trades. *Quarterly Journal of Economic*, 123(1):139–176, 2008. doi: 10.3386/w13367. URL https://scholar.harvard.edu/nunn/publications/long-term-effects-africas-slave-trades.

Nathan Nunn and Leonard Wantchekon. The slave trade and the origins of mistrust in africa. *American Economics Review*, 101(7):3221–52, December 2011. doi: 10.1257/aer.101.7.3221. URL http://www.aeaweb.org/articles?id=10.1257/aer.101.7.3221.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

Robert Tibshirani, Jerome Friedman, and Trevor Hastie. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer, 2009. ISBN 978-0-387-84857-0.

APPENDIX

A REPLICATION RESULTS

Below are the remainder replication tables reported. As well as the previous replication tables, the same estimates are produced with an exception in Table 10. This table has two other standard errors, although the difference between them is only 0.001, which does not make any difference in their conclusions

Slave Trade - Table 2

Table 9: Replication Table 2

	Trust of Neighbours	Trust of relatives	Trust of local council	Intergroup	Intragroup
ln(1 + export/area)	-0.159	-0.133	-0.111	-0.097	-0.144
	(0.034)	(0.037)	(0.021)	(0.028)	(0.032)
Used controls:					
District	Yes	Yes	Yes	Yes	Yes
Individual	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Specifications:					
Ethnicity clusters	185	185	185	185	185
Observations	20,027	20,062	19,733	19,765	19,952
District clusters	1,257	1,257	1,283	1,255	1,257
R^2	0.16	0.13	0.20	0.11	0.14

This table shows OLS estimations. The description for the used controls is given in 9. This table shows, in addition to table 1, the relationship between $ln(1 + export/area)$ and different measures for trust.
This table shows OLS estimations. Below each estimated coefficient are three different standard errors displayed. Below the OLS estimates are shown the used controls and specifications of the regressions.

Slave Trade - Table 3

Table 10: Replication Table 3

	Trust of Neighbours	Trust of relatives	Trust of local council	Intergroup	Intragroup
ln(1 + export/area)	-0.202	-0.178	-0.129	-0.115	-0.188
	(0.031)	(0.032)	**(0.021)**	(0.030)	**(0.032)**
Used Controls:					
District	Yes	Yes	Yes	Yes	Yes
Individual	Yes	Yes	Yes	Yes	Yes
Ethnicity-level colonial	Yes	Yes	Yes	Yes	Yes
Colonial population density	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Specifications:					
Ethnicity clusters	147	147	146	147	147
Observations	16,679	16,709	15,905	16,473	16,636
District clusters	1,187	1,187	1,194	1,184	1,186
R^2	0.16	0.13	0.21	0.12	0.16

This table shows OLS estimations. The description of the used controls is given in table 2. This table shows, in addition to table 9, the relationship between $ln(1 + export/area)$ and measures of trust also additional controls.

Slave Trade - Table 6

Table 11: Replication Table 6

	Trust of neighbours	Trust of relatives	Trust of local council	Intergroup Trust	Intragroup Trust
Second stage: Dependent variable is one of above trust					
ln(export/area + 1)	-0.271	-0.172	-0.262	-0.189	-0.254
	(0.088)	(0.076)	(0.075)	(0.103)	(0.109)
Hausman test (p-value)	0.42	0.98	0.05	0.44	0.53
R^2	0.16	0.13	0.20	0.12	0.15
First stage: Dependent variable is ln(export/area + 1)					
Instrumental Variable	-0.0015	-0.0015	-0.0015	-0.0015	-0.0015
	(0.0003)	(0.0003)	(0.0003)	(0.003)	(0.0003)
Used Controls:					
District	Yes	Yes	Yes	Yes	Yes
Individual	Yes	Yes	Yes	Yes	Yes
Colonial population density	Yes	Yes	Yes	Yes	Yes
Distances to Saharan city, route	Yes	Yes	Yes	Yes	Yes
Ethnicity-level colonial	Yes	Yes	Yes	Yes	Yes
Reliance on fishing	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Specifications					
Number of clusters	147/1,187	147/1,187	146/1,194	147/1,184	147/1,186
Number of observations	16,679	16,709	15,905	16,473	16,636
F-stat of excl. instrument	21.6	21.7	22.2	21.6	21.8
R^2	0.81	0.81	0.81	0.81	0.81

This table shows IV estimates. When estimating the relationship between $ln(export/area + 1)$ and measures for trust, this table has additional controls comparing to table 2. And the used Instrumental Variable is Historical distance of ethnic group from coast.

Slave Trade - Table 9

Table 12: Replication Table 9

	Intergroup trust				
	Within	Within	Within		
	district	town	province	Trust of local council	
Baseline measure	-0.120	-0.102	-0.098	-0.070	-0.072
	(0.027)	(0.028)	(0.029)	(0.019)	(0.019)
Alternative measure	-0.063	-0.091	-0.037		
	(0.030)	(0.029)	(0.035)		
Used Controls:					
Baseline controls	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Council fixed effects	No	No	No	Yes	Yes
Colonial population density	Yes	Yes	Yes	Yes	Yes
Public goods fixed effects	No	No	No	Yes	No
Ethnicity-level colonial	Yes	Yes	Yes	Yes	Yes
Specifications:					
Clusters	147/737	147/725	147/1,127	145/1,130	146/1,172
Observations	12,513	9,673	15,999	12,203	12,827
R^2	0.12	0.12	0.12	0.37	0.37

The table shows OLS estimates. Below table 10 is a description about the used controls.

B Deriving properties of $\check{\theta}_0$

As shown in section 2.1 the scaled estimation error of $\check{\theta}_0$ can be divided into three parts. In Chernozhukov et al. (2017), they show the derivation of the properties of these estimator:

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = d^* + e^* + f^*$$

The first term, d^* satisfies:

$$d^* = (EW^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i \in I} W_i U_i \rightsquigarrow N(0, \Sigma)$$

under weak conditions. The second term, e^*, has the regularization bias when estimating g_0 and m_0. In formula form:

$$e^* = (EW^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i \in I}(\hat{g}_0(X_i) - g_0(X_i))(\hat{m}_0(X_i) - m_0(X_i)),$$

This term is only dependently on the product of the remainder terms in $\hat{m}_0$ and $\hat{g}_0$. Thus when having a large range of data-generating processes, it will vanish. And at last under weak conditions will f^* satisfy: $f^* = o_p(x)$