The Effect of Plough Agriculture on Gender Roles: A Machine Learning Approach*

Anna Baiardi† and Andrea A. Naghi‡

May 2023

Abstract

This paper undertakes a replication in a wide sense of Alesina et al. (2013), which examines the relationship between historical plough agriculture and current gender roles. We revisit the main research question with recently developed causal machine learning methods, which allow to model the relationship of covariates with the treatment and the outcomes in a more flexible way, while also including interactions and nonlinearities that were not considered in the original analysis. Our results suggest an even larger negative effect of the historical plough adoption on female labor force participation than what the original analysis found. The paper highlights the benefits of using causal machine learning methods in applied empirical economics.

Keywords: machine learning, causal inference, double machine learning, average treatment effects

J.E.L. Classification: C01, C21, D04

*Baiardi acknowledges support from EU Horizon 2020, Marie Skłodowska-Curie individual grant (No. 840319). Naghi acknowledges support from EU Horizon 2020, Marie Skłodowska-Curie individual grant (No. 797286). Financial support from the United Nations Sustainable Development Funds is also gratefully acknowledged.

†Department of Economics, Erasmus University and Tinbergen Institute. Email: baiardi@ese.eur.nl.

‡Department of Econometrics, Erasmus University and Tinbergen Institute. Email: naghi@ese.eur.nl.

1 Introduction

Beliefs about the appropriate role of women in society are very different across regions. These disparities can be observed by analyzing, for instance, differences in labor force participation for women across societies. Choices about female labor supply have been shown to be partially explained by culture and norms (Fernandez, 2007; Fernández and Fogli, 2009). Thus, investigating the origin of gender norms is very important to understand the reasons behind these differences and assess whether appropriate policies can be implemented to reduce them (Hansen et al., 2015).

The paper by Alesina et al. (2013) is a seminal contribution addressing the question of the historical origins of gender roles. The authors test the hypothesis, originally developed by Boserup (1970), that today's gender norms have their roots in the agricultural practices that prevailed in pre-industrial times. The hypothesis compares the roles of shifting and plough cultivation. Since operating the plough requires considerable physical strength, men have an advantage in plough cultivation compared to women; in contrast, women could more easily participate in shifting cultivation, in which the use of the hoe and the digging stick are prevalent, and there is higher need for weeding, which was traditionally performed by women and children. Thus, where plough agriculture was prevalent, gender division of labor was more common. This division of labor would persist over time until the present day.

In this paper, we perform a replication in a wide sense by revisiting the main research question in Alesina et al. (2013) with new causal inference tools, namely causal machine learning (CML) methods. To this end, we connect the econometric theory on CML with empirical economics, serving as an illustration for applied researchers on the gains of implementing these newly available methods in observational studies. In our replication study, we employ the double/debiased machine learning method (DML) introduced in Chernozhukov et al. (2017, 2018), which provides consistent estimation and valid inference on the average treatment effect (ATE), in settings where high-dimensional nuisance parameters are estimated with machine learning methods. In our analysis, we combine the DML framework with the following machine learning methods: lasso, trees, neural net, random forest,

boosting and two hybrid methods: ensemble and best. Empirical economics studies that started to employ this method include Knaus (2018), Dube et al. (2020), Xu et al. (2021) and Baiardi and Naghi (2022), among others.

We revisit the question in Alesina et al. (2013) with CML tools because the functional form of the relationship of the confounders with plough use and gender norms is unknown, and guidance from economic theory or previous literature regarding the nuisance functions is limited.[1] CML methods are designed to flexibly estimate and capture complex interactions between the outcome, treatment and confounders, which is important when drawing causal conclusions based on the assumption of unconfoundedness. In addition, the original analysis undertaken with traditional causal inference tools (ordinary least squares and instrumental variables) cannot include all raw covariates, interactions and nonlinearities at once, because the number of confounders would be too large relative to the sample size.[2] In contrast, CML methods can handle a large number of covariates relative to the sample size, for example, by using regularized regressions, and thus can control at once for all potentially relevant linear and nonlinear confounders.

Our DML estimation results for the effect of plough cultivation on gender roles show a negative and significant effect, confirming the main findings in Alesina et al. (2013). In fact, the estimates suggest an even larger effect of the plough adoption, compared to the original findings. We attribute these differences to causal machine learning methods being able to capture more flexibly the effect of a large number of covariates.

In what follows, section 2 familiarizes the reader with the revisited paper and presents a description of the original analysis. Section 3 briefly describes the recently developed DML method that we implement, and presents the results of the replication in a wider sense. Section 4 summarizes some lessons learned from revisiting the paper with CML methods. The Online Supplementary Material includes details on the implementation of the CML methods used, as well as sensitivity analysis results.

[1] Other topics in applied economics, such as estimating the wage equation, or the gravity model, have been analyzed much more extensively, and researchers interested in these topics can benefit from more guidance from previous research regarding the functional form of the nuisance functions.

[2] For example, note that only a limited number of pre-specified nonlinear terms are included in Alesina et al. (2013).

2 Description of the Original Analysis

Alesina et al. (2013) consider several empirical strategies and data sets, and present results using country-level and individual-level regressions. Then, to tackle possible endogeneity issues, the paper follows two approaches: first, several potential confounders are included in the regressions; second, an instrumental variable (IV) strategy is used. We revisit the main question addressed in the original paper, focusing on the country-level results, as the majority of the regressions reported in the original paper are based on this data.

The baseline ordinary least squares (OLS) country-level results in the original analysis (reported in Table 4 of Alesina et al., 2013) are obtained by estimating the following regression:

$$y_c = \alpha + \beta plough\, use_c + \boldsymbol{X_c^H \Gamma} + \boldsymbol{X_c^C \Pi} + \epsilon_c,$$

where c stands for country. In the paper, three outcome variables are examined as measures of gender roles: female labour force participation, attitudes about women's work, and attitudes about women as leaders. The first outcome variable is an indicator variable that equals one if the individual is in the labor force in 2000; the second is the share of firms with a woman among its principal owners in the period 2003-2010; finally, the third is the proportion of seats held by women in the national parliament in 2000. The treatment variable, $plough\, use_c$, is calculated as the estimated proportion of individuals living in a country with ancestors that used the plough in pre-industrial agriculture. The vector X_c^H includes historical ethnographic variables at the country level. These controls capture the historical characteristics of ethnicities living in a country, and they are meant to account for differences between ethnicities that historically adopted the plough and those that did not. They include: ancestral suitability for agriculture, fraction of ancestral land that was tropical or subtropical, ancestral domestication of large animals, ancestral settlement patterns, and ancestral political complexity. The vector X_c^C denotes contemporary country-level controls: natural log of real per capita GDP, and its square. These are included as the level of economic development is believed to have an impact on female labour force participation, and the square of per capita GDP is intended to capture the observed U-shaped relation between the

two variables. Continent fixed effects are also added in some specifications.

As mentioned by Alesina et al. (2013), concerns about potential endogeneity in the baseline regressions arise. It is possible that plough agriculture may have been more common in countries that had less equal gender-role attitudes. This would cause the OLS estimates to be biased away from zero. Moreover, plough agriculture may have been more likely in areas where economic development was historically higher. If historical and contemporary economic development are correlated, and more economically advanced countries tend to have higher female labour force participation and more equal gender roles, OLS estimates may be biased towards zero. To tackle these issues, the following two solutions are offered in the paper.

First, motivated by the thought that the potential bias may be partly due to observable characteristics, a number of additional controls are added to the regressions. These include two groups of variables that may correlate with both plough agriculture in the past and current gender roles: historical characteristics of the ancestors of the current population living in a country, and current economic, social and cultural characteristics of countries. We list these variables in Section A of the Online Supplementary Material. Alesina et al. (2013) provide the rationale for including each of these controls, and details on how the variables are constructed.

Second, the authors use an instrumental variable approach, which exploits the fact that plough adoption is correlated with the suitability of the land for cereal crops that would benefit, and crops that would not benefit, from the plough. To this end, two instruments for plough adoption are constructed, based on the analysis by Pryor (1985). The first is the suitability for 'plough-positive' (i.e. which benefit most from the plough) cereal crops, and the second is the suitability for 'plough-negative' (i.e. which benefit least from the plough) cereal crops.[3] Alesina et al. (2013) show that the suitability for plough-positive crops, but not for plough-negative crops, is positively correlated with the use of the plough. On the validity of the exclusion restriction, the authors explain that the underlying assumption is that the suitability for plough-positive and plough-negative crops only affects current gender norms through the historical adoption of the plough; thus, the main

[3]See Alesina et al. (2013) for details on the data used and how the instruments are constructed.

concern with the instrumental variable strategy is the possibility that suitable areas for different crops could be correlated with geographic characteristics that have an effect on gender norms through other channels. To address this, Alesina et al. (2013) include in a robustness check a number of geo-climatic characteristics (listed in the Online Supplementary Material).

3 Causal Machine Learning Analysis

We implement the DML method developed in Chernozhukov et al. (2017, 2018). Following the notation in Chernozhukov et al. (2017), we consider the partially linear regression model:

$$Y = D\theta_0 + g_0(X) + U,$$
$$D = m_0(X) + V,$$

where Y is the outcome, D is the treatment variable, X are the set of covariates, and U and V the error terms. The main equation of interest is the first equation, where θ_0 is the average treatment effect (ATE). The second equation links the treatment to the covariates and keeps track of confounding effects. The functions $g_0(X)$ and $m_0(X)$ can be highly nonlinear and are estimated with a variety of ML methods.

Employing ML methods in this setting introduces bias due to regularization.[4] This is because by regularization, the less important coefficients are shrunk to zero which introduces a bias that transfers to the target parameter, similar to the omitted variable bias. Bias due to regularization is controlled by solving two prediction problems (hence the name 'double/debiased' machine learning). More precisely, in the first step, a ML method is used to fit m_0 in the second equation, partialling out the effect of X from D and obtaining residuals $\hat{V}$. Then, a ML method is used again to fit g_0 in the first equation, partialling out the effect of X from Y and obtaining residuals $\hat{W}$. Finally, we run the residuals-on-residuals regression, $\hat{W}$ on $\hat{V}$, to obtain an estimate of the low dimensional parameter θ_0. This is similar to a Frisch-Waugh-Lovell – style approach to estimate the target parameter.

[4]This approach also introduces bias due to overfitting which is then mitigated with sample splitting, see Chernozhukov et al. (2017).

Chernozhukov et al. (2018) extend the partially linear regression model to a partially linear IV model which allows for endogenous treatment. In this paper, we refer to this model as DML - IV. The instrument can be scalar or vector. Chernozhukov et al. (2018) also consider estimation of average treatment effects when the treatment effect is fully heterogeneous (the interactive model and the interactive IV model). We do not consider these extensions here as they require a binary treatment and our treatment variable is continuous. Our DML estimates are obtained with the Robinson-style 'partialling-out' score function (see Chernozhukov et al., 2018), but we also perform sensitivity checks with the alternative score function in the Online Supplement.

3.1 Results

In our analysis, we re-examine both the country-level OLS and IV regressions. For the OLS analysis, we begin by estimating a DML partially linear model that only includes the baseline set of controls as raw covariates. We then revisit the robustness analysis of this specification, by including as raw covariates the largest set of controls used in the robustness checks (this corresponds to Table 7, column 8 of the original paper), to which we also add the continent fixed effects.[5] This amounts to a total of 36 raw covariates. For the IV analysis, in addition to the baseline controls and in line with the original paper, we consider the geo-climatic characteristics the authors use in their IV robustness checks (Table A14 of the Online Appendix of the original paper). In the original paper, the geo-climatic characteristics are added linearly, in quadratic forms, and as linear interactions.[6] To these variables, we again add the continent fixed effects.

Table 1 reports the results of the DML partially linear model that replicates the baseline regression. In accordance with the original paper, the treatment effect estimates are negative and statistically significant. Both the coefficients and standard errors are close to those reported in Table 4 of Alesina et al. (2013)

[5]When revisiting the robustness analysis with DML, we include continent fixed effects, even though the original paper did not include them in their most complete robustness checks. As causal ML methods can handle a large number of covariates, we include all the covariates which were considered in the original paper, to ensure that all potential confounders are taken into account.

[6]In the original analysis, quadratic terms and linear interactions are not included in the same regressions.

Table 1: Country-level estimates, partially linear model

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) OLS
				Panel A: Female labour force participation 2000				
Plough use	-10.434	-9.287	-11.948	-11.243	-11.271	-11.501	-11.388	-12.401
	(3.195)	(2.817)	(3.129)	(3.181)	(3.227)	(3.182)	(3.18)	(2.964)
Observations	165	165	165	165	165	165	165	165
				Panel B: Share of firms with female ownership				
Plough use	-13.168	-12.316	-13.114	-12.769	-14.556	-13.287	-13.54	-15.241
	(3.954)	(3.772)	(3.945)	(4.488)	(3.828)	(4.128)	(4.301)	(4.06)
Observations	123	123	123	123	123	123	123	123
				Panel C: Share of political positions held by women 2000				
Plough use	-5.196	-3.617	-4.29	-5.41	-4.48	-4.817	-5.029	-4.821
	(1.946)	(1.707)	(1.745)	(1.777)	(1.853)	(1.804)	(1.825)	(1.782)
Observations	144	144	144	144	144	144	144	144
Raw covariates	7	7	7	7	7	7	7	7

Notes: Analysis of Table 4 (columns 1, 3, 5) of Alesina et al. (2013) using DML. Column 8 reports the original paper results. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Robust standard errors are reported in column 8. The number of covariates does not include the treatment variable.

(reproduced here for convenience in column 8 of Table 1), and reassuringly, fairly stable across the ML methods. This indicates that the estimates are robust to using a number of different ML methods to estimate the nuisance functions.

We find however very different results when carrying out the robustness analysis of this baseline specification with the DML method. Panel A of Table 2 reports the results. While the effect is still negative, albeit much smaller in absolute value, statistically significance is now lost. Interestingly, when Alesina et al. (2013) include all covariates at once (the estimate is reproduced in the last column of our Table 2), the treatment effect becomes smaller in absolute value, compared to when groups of covariates are added separately (see their Table 7, columns 1 to 7), or compared to the baseline specification (reproduced in column 8 of our Table 1). With DML, the treatment effect of interest does not only become smaller, but also statistically insignificant.[7]

Our findings up to this point would lead us to (mistakenly) conclude that the negative effect of plough adoption on attitudes towards gender roles may not be as large as suggested by the original analysis, and that the effect is not statistically significant. However, our estimates from the DML partially linear model may still be subject to endogeneity. While flexibly controlling for a large number of covariates can account for the confounding effect of observed characteristics, the remaining concern is that plough adoption may be correlated with unobserved characteristics that also affect the outcome. The instrumental variable approach suggested by Alesina et al. (2013) can alleviate this potential issue. We consider the same instruments as in the paper (described above) and we turn to re-evaluating the results by estimating a DML - IV model. Panel B of Table 2 reports the results. As in the original analysis, the estimated coefficients have a negative sign, and they are now statistically significant at the 10% level for most of the ML methods, with the exception of neural networks and ensemble. It is interesting to note that the magnitude of the coefficients is larger than in the DML partially linear model (both baseline and extended), supporting the hypothesis that the OLS estimates are biased towards zero. This is consistent with

[7]To assess the robustness of the DML estimates to using other causal ML methods, we also perform the analysis with the causal forest (Wager and Athey, 2018; Athey et al., 2019). The implementation details are described in the Online Supplementary Material. Table B.1 in the Online Supplement reports the results of the main OLS robustness check of Alesina et al. (2013). The results show that the estimates are very similar to those obtained with DML.

Table 2: Country-level estimates with full set of controls

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) OLS
	Panel A: DML, Partially linear Model. Outcome: Female labour force participation							
Plough use	-5.922	-6.744	-6.445	-5.812	-5.682	-5.925	-5.224	-9.234
	(5.636)	(5.063)	(4.834)	(4.742)	(6.023)	(4.998)	(4.91)	(4.301)
Observations	142	142	142	142	142	142	142	142
Raw covariates	36	36	36	36	36	36	36	30
	Lasso	Reg. Tree	Boosting	Forest	Neural Net.	Ensemble	Best	2SLS
	Panel B: DML-IV. Outcome: Female labour force participation							
Plough use	-38.345	-36.85	-39.429	-36.961	-20.725	-33.645	-38.712	-28.516
	(19.936)	(18.996)	(16.23)	(14.247)	(29.797)	(21.486)	(23.011)	(7.559)
Observations	160	160	160	160	160	160	160	160
Raw covariates	17	17	17	17	17	17	17	11

Notes: Analysis of the main robustness checks of Alesina et al. (2013) using DML. Column 8 reports the results of the most complete robustness checks for the OLS and IV specifications in the original paper. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Robust standard errors are reported in column 8. The number of covariates does not include the treatment variable.

the original paper, which also finds that the IV coefficients are larger than the OLS estimates. It is worth to further notice that compared to the IV results of the original paper, our DML - IV findings suggest an even larger effect of the plough adoption on female labour force participation. We attribute this to causal machine learning methods being able to control for a large number of covariates in a more flexible way.[8] Overall, when looking at both the robustness analysis and the IV analysis and comparing them to the baseline results, we notice that our estimates move in the same direction as the original paper estimates, but our estimates move even more, supporting the idea that DML controls more flexibly for relevant covariates.

To investigate whether the instruments still have predictive power when flexibly controlling for the confounders, we estimate the first stage of Table 2 Panel B, using the partially linear model with treatment as outcome and the instruments as treatments. The results are reported in Table 3.[9] As in the original 2SLS analysis, we notice a strong correlation with one of the two instruments (the land suitability for plough-positive crops), but no significant correlation with the other (plough-negative crops). Since plough-negative crops does not predict plough use, we undertake a robustness check by estimating the model considering only plough-positive crops as instrument for plough use; the results, shown in Table B.2 of the Online Supplementary Material, are overall consistent with those reported in Table 2, Panel B.[10]

Our DML estimates are obtained by tuning the parameters of the ML methods

[8] As explained above, our DML specification differ from the original paper's robustness analysis because it considers nonlinearities and it includes continent fixed effects. Therefore, the differences between the DML and the original estimates could, in principle, be driven by the continent fixed effects, and not by the nonlinearities. The original paper shows that adding the continent fixed effects to the baseline specification leads to very small changes in the OLS estimates (see Table 4 in the original paper), while it results in larger changes in the IV case (see Table 8 in the original paper). However, even in the IV case, including the continent fixed effects only increases the absolute size of the plough coefficient by 3-4 percentage points, while the DML coefficients exceed the OLS and 2SLS estimates by more than double that amount (with the exception of the neural network and ensemble estimates). Thus, we conclude that allowing for a more flexible nuisance function is likely to be driving at least part of the differences between the DML and the 2SLS (and OLS) estimates.

[9] Column 8 of Table 3 reports the first-stage estimated by OLS, including the full set of covariates at once.

[10] The F test of the instrument is within the range of 8.30 and 19.90, depending on the ML method used, with an average of 11.95. This is similar to the value of the F test for the instrument in the 2SLS estimation, which is 11.15.

Table 3: First stage estimates

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) OLS
Plough-positive crops	0.521	0.364	0.531	0.616	0.341	0.460	0.383	0.678
	(0.160)	(0.136)	(0.157)	(0.160)	(0.154)	(0.158)	(0.156)	(0.204)
Plough-negative crops	0.174	-0.071	0.134	0.080	0.193	0.208	0.178	0.232
	(0.160)	(0.151)	(0.150)	(0.176)	(0.134)	(0.147)	(0.151)	(0.204)
Observations	160	160	160	160	160	160	160	160

Notes: This table reports the results of the first-stage analysis of Table 2 Panel B. For the DML analysis, the standard errors are adjusted for variability across splits using the median method. Robust standard errors are reported in column 8.

via cross-validation, whenever possible/theoretically justifiable. For the remaining parameters that are not data-driven (such as the number of trees, or the leaf node size in the random forest), or for the algorithms for which adaptive tuning parameters are not yet available (e.g. neural network), we use default values. We detail the default parameter values in the implementation details section of the Online Supplementary Material. For robustness, we perform extensive sensitivity analysis on the values of these tuning parameters choices. We also implement several activation functions in the neural network. The results, reported in the online supplementary material, are in line with those reported in the main text. Interestingly, once we increase the number of hidden layers to 4, with the number of neurons set to 2, the DML-IV estimates for the neural network (Table 2 Panel B) become significant for both activation functions, further supporting the results of the other ML methods. Additionally, we perform robustness checks of our DML results using an alternative score function, increasing the number of folds, and trimming the propensity scores at 0.01 and 0.99. These results are reported in the online supplementary material and confirm our main findings.

4 Discussion and Takeaways

This study revisited the main results of the paper by Alesina et al. (2013) – obtained originally with traditional causal inference tools (OLS, IV) – with recently developed CML methods. Our results using the new causal tools show a negative and significant causal effect of the historical plough use on female labor

force participation, supporting the findings from Alesina et al. (2013). Although the main conclusion is the same when using CML methods, there are a couple of lessons we learned while performing this analysis, which could be of interest for applied economics researchers.

First, this empirical paper is a good illustration on how causal machine learning methods can serve as useful tools for the empirical researcher to perform supplementary analyses. In order to support the credibility of the empirical evidence, researchers typically report a number of different model specifications and evaluate the sensitivity of estimates to these alternatives – similar to the above-mentioned robustness checks performed in the original paper. The usual approach to evaluating the variability of estimates to different model specifications can be somewhat ad-hoc and not a systematic way of implementing sensitivity analysis. In addition, relevant covariates or interactions of these covariates which are not considered important a priori by the researcher might be missed. Instead, causal machine learning methods use systematic algorithms that compare a wide range of model specifications for the nuisance functions and choose the one that best fits the data. This makes them more robust methods for sensitivity analyses than the current practice in the literature. Indeed, the example discussed here shows that the robustness analysis performed with DML can suggest different conclusions compared to the original paper's robustness checks. Thus, we view causal machine learning methods as promising tools for sensitivity analysis in empirical work.

Second, this revisited empirical study illustrates the gains from combining modern ML tools with quasi-experimental methods such as instrumental variables. While causal ML methods can make the unconfoundedness assumption more plausible by flexibly controlling for observed confounders, they cannot account for unobserved confounders. In such settings, the researcher could combine causal ML methods with quasi-experimental methods such as IV, which potentially overcomes biases caused by unobserved factors. Integrating the two methods could provide powerful tools for the researcher's toolkit.

References

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the origins of gender roles: Women and the plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2021). DoubleML – An object-oriented implementation of double machine learning in R. arXiv:2103.09603 [stat.ML].

Baiardi, A. and Naghi, A. A. (2022). The value added of machine learning to causal inference: Evidence from revisited studies. Working paper, Erasmus University Rotterdam.

Boserup, E. (1970). *Woman's Role in Economic Development*. London: George Allen and Unwin Ltd.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Dube, A., Jacobs, J., Naidu, S., and Suri, S. (2020). Monopsony in online labor markets. *American Economic Review: Insights*, 2(1):33–46.

Fernandez, R. (2007). Women, work, and culture. *Journal of the European Economic Association*, 5(2-3):305–332.

Fernández, R. and Fogli, A. (2009). Culture: An empirical investigation of beliefs, work, and fertility. *American economic journal: Macroeconomics*, 1(1):146–77.

Hansen, C. W., Jensen, P. S., and Skovsgaard, C. V. (2015). Modern gender roles and agricultural history: the neolithic inheritance. *Journal of Economic Growth*, 20(4):365–404.

Knaus, M. C. (2018). A double machine learning approach to estimate the effects of musical practice on student's skills. *arXiv preprint arXiv:1805.10300*.

Pryor, F. L. (1985). The invention of the plow. *Comparative Studies in Society and history*, 27(4):727–743.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Xu, Y., Ghose, A., and Xiao, B. (2021). Mobile payment adoption: An empirical investigation on alipay. *Available at SSRN 3270523*.

Online Supplementary Material

A Implementation Details

Double Machine Learning. We obtain the DML estimation results with 100 splits and 2-fold cross-fitting. The reported estimates are median estimates across the splits. The standard errors are adjusted using the median method due to the variability resulting across the sample splits.

We report lasso estimates based on ℓ_1-penalized regressions, where we obtain the penalty parameter via 10-fold cross-validation. For the lasso, we include as controls the set of all raw covariates, the squared terms included in the original analysis, and all first order interactions. For the other ML methods, the controls are the set of raw covariates. The penalty parameter of the regression tree is obtained with 10-fold cross-validation. The reported random forest results are obtained with 1000 trees and the boosting results with 1000 boosted regression trees. For the boosting, the minimum number of observations in the terminal nodes is set to the default value of 1 and the fraction of randomly selected training observations is set to 0.5. For the neural networks, we used simple settings of 2 neurons, 1 hidden layer, and a decay parameter of 0.01. The baseline activation function is set to the linear function. For sensitivity analysis we change the activation function from a linear to the nonlinear softplus function. This is a smooth approximation of the Rectified Linear Unit (ReLU) function, commonly used in the literature (see Table B.10 below on the DML-IV results). Furthermore, we perform extensive robustness analyses on all the tuning parameters values of the ML methods that are not obtained via cross-validation, such as the number of trees for the random forest and boosting, the minimum number of observations in the end nods, the number of neurons, number of hidden layers. We report some of these results in Tables B.3 - B.10 below, for the DML-IV estimates. The results are consistent with those reported in the main text. Notice also that when the number of neurons is 2 and the hidden layers are increased to 4, the neural network estimates become significant for both the linear and the *SmoothReLU* activation functions, which corroborates the results found with the other ML methods.

The two hybrid methods used in our analysis are Ensemble and Best. *Ensemble* is a weighted average from lasso, boosting, random forest and neural net. The

weights minimize the average means squared out-of-sample prediction error. *Best* selects the best method among all the methods used, to estimate the nuisance functions, in term of the average out-of-sample prediction error.

We perform sensitivity checks varying the number of splits and folds and the results are very similar to those reported in the paper. We include in Table B.11 the results using 100 splits and 5 folds, and the results are overall consistent with those using 2 folds. We also implement sensitivity analysis to using different Neyman orthogonal score functions. Our main DML results are obtained using the partialling out score. We explore robustness when using an alternative score function (IV-type), using the DoubleML R package (Bach et al., 2021). The DoubleML package currently does not allow the implementation of the IV-type score for IV models with multiple instruments; thus, we perform this check for the results in Table 1 and Panel A of Table 2. The estimates are reported in Table B.12 and show that the results are not sensitive to the type of score function used. Finally, results with trimmed propensity scores on Table 2 Panel B are reported in Table B.13 and the results are consistent with our main findings.

Causal Forest. For the causal forest estimates, the values of the tuning parameters are optimised via cross-validation, with the exception of the number of trees, which is set to 2000. We also perform sensitivity checks with 500 and 1000 trees and the estimates are consistent with those reported in the main text.

The causal forest estimates are implemented with *orthogonalization*, as suggested in Section 6.1.1 of Athey et al. (2019). This is particularly useful when applying the method on observational studies. More precisely, we estimate the marginal outcomes and the propensity score by training separate regression forests. We then obtain the residual treatment and the residual outcome on which we finally train the causal forest.

Details on the Control Variables. When replicating the OLS robustness analysis reported in Table 2 Panel A, we include the baseline control variables, country fixed effects and the additional control variables. The additional covariates include historical and contemporaneous variables. The additional historical controls are: the intensity of agriculture; the proportion of subsistence provided by hunting and by the herding of large animals; the fraction of countries' ances-

tors without land inheritance rules, with patrilocal post-marital residence rules, and with matrilocal post-marital rules; the fraction of countries' ancestors with a nuclear and an extended family structure; the average year the ethnicities were sampled in the *Ethnographic Atlas*. The additional contemporary controls are: years of civil and interstate conflicts (1816-2007); terrain ruggedness; whether a country was under a communist regime after WWII; the fraction of a country's population with European descent; oil production per capita; agricultural, manufacturing and services shares of GDP; and the fraction of a country's population who is Catholic, Protestant, other Christian, Muslim, and Hindu. When replicating the IV robustness analysis reported in Table 2 Panel B, we include the baseline controls, country fixed effects, and the additional geo-climatic characteristics added in Table A14 of the Online Supplement of the original paper. The geo-climatic characteristics are: terrain slope, soil depth, average temperature, average precipitation. For the IV analysis with only one instrument (plough-positive crops) reported in Table B.2, we also include plough-negative crops as a control variable.

B Additional Tables

Table B.1: Analysis with Causal Forest

	Female labor force participation
Traditional plough use	-5.996
	(4.071)
Observations	142

Notes: Analysis of the main OLS robustness check of Alesina et al. (2013) using the causal forest. Standard errors are reported in parentheses.

Table B.2: DML-IV estimates with one instrument

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) 2SLS
				Panel A: DML-IV estimates				
Plough use	-30.456	-28.201	-26.152	-15.773	-24.178	-24.730	-26.530	-25.736
	(13.207)	(14.245)	(11.721)	(10.841)	(17.485)	(13.659)	(15.475)	(13.026)
				Panel B: First stage				
Plough use	0.584	0.457	0.626	0.678	0.490	0.585	0.510	.678
	(0.169)	(0.144)	(0.167)	(0.152)	(0.166)	(0.179)	(0.177)	(.203)
Observations	160	160	160	160	160	160	160	160

Notes: Analysis of the main IV robustness check of Alesina et al. (2013) using DML, using only one instrument: plouugh-positive crops. Column 8 reports the 2SLS estimates for the second stage (Panel A) and first stage (Panel B). Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Robust standard errors are reported in column 8.

Table B.3: DML Sensitivity Analysis on *Boosting*, Varying the Number of Trees

	Number of Trees: 500	Number of Trees: 2000
Plough use	-39.476	-43.424
	(15.062)	(16.559)
Observations	160	160

Notes: Analysis of Alesina et al. (2013) using DML-IV. Standard errors adjusted for variability across splits using the median method are reported in parentheses.

Table B.4: DML Sensitivity Analysis on *Boosting*, Varying the Minimum Number of Observations in the End Nodes

	Min. No. of Observations: 3	Min. No. of Observations: 5
Plough use	-43.553	-35.904
	(14.769)	(13.871)
Observations	160	160

Notes: Analysis of Alesina et al. (2013) using DML-IV. Standard errors adjusted for variability across splits using the median method are reported in parentheses.

Table B.5: DML Sensitivity Analysis on *Boosting*, Varying both the Number of Trees and the Minimum Number of Observations in the End Nodes

	No. of Trees: 500 Min. No. of Observations: 3	No. of Trees 500 Min. No. of Observations: 5
Plough use	-39.448	-45.162
	(14.475)	(15.156)
Observations	160	160
	No. of Trees: 2000 Min. No. of Observations: 3	No. of Trees 2000 Min. No. of Observations: 5
Plough use	-40.960	-39.524
	(15.002)	(16.027)
Observations	160	160

Notes: Analysis of Alesina et al. (2013) using DML-IV. Standard errors adjusted for variability across splits using the median method are reported in parentheses.

Table B.6: DML Sensitivity Analysis on *Random Forest*, Varying the Number of Trees

	Number of Trees: 500	Number of Trees: 2000
Plough use	-30.185	-30.226
	(15.138)	(14.409)
Observations	160	160

Notes: Analysis of Alesina et al. (2013) using DML-IV. Standard errors adjusted for variability across splits using the median method are reported in parentheses.

Table B.7: DML Sensitivity Analysis on the *Neural Net*, Varying the Number of Neurons

	Number of Neurons: 3	Number of Neurons: 5
Plough use	-15.592	-15.153
	(22.778)	(24.528)
Observations	160	160

Notes: Analysis of Alesina et al. (2013) using DML-IV. Standard errors adjusted for variability across splits using the median method are reported in parentheses.

Table B.8: DML Sensitivity Analysis on the *Neural Net*, Varying the Decay Parameter

	Decay Parameter: 0.02	Decay Parameter: 0.05
Plough use	-27.399	-26.729
	(19.065)	(17.550)
Observations	160	160

Notes: Analysis of Alesina et al. (2013) using DML-IV. Standard errors adjusted for variability across splits using the median method are reported in parentheses.

Table B.9: DML Sensitivity Analysis on the *Neural Net*, Linear Activation Function, Varying the Number of Hidden Layers and the Number of Neurons per Layer

	Hidden Layers: 2 Neurons per layer: 2	Hidden Layers: 3 Neurons per layer: 2	Hidden Layers: 4 Neurons per layer: 2
Plough use	-26.268	-26.863	-13.621
	(25.344)	(30.356)	(3.250)
Observations	160	160	160
	Hidden Layers: 2 Neurons per layer: 4	Hidden Layers: 3 Neurons per layer: 4	Hidden Layers: 4 Neurons per layer: 4
Plough use	-35.425	-27.363	-27.044
	(22.606)	(26.188)	(29.753)
Observations	160	160	160

Notes: Analysis of Alesina et al. (2013) using DML-IV. Standard errors adjusted for variability across splits using the median method are reported in parentheses.

Table B.10: DML Sensitivity Analysis on the *Neural Net*, *SmoothReLU* Activation Function, Varying the Number of Hidden Layers and the Number of Neurons per Layer

	Hidden Layers: 2 Neurons per layer: 2	Hidden Layers: 3 Neurons per layer: 2	Hidden Layers: 4 Neurons per layer: 2
Plough use	-33.953	-37.394	-39.094
	(18.997)	(17.810)	(18.104)
Observations	160	160	160
	Hidden Layers: 2 Neurons per layer: 4	Hidden Layers: 3 Neurons per layer: 4	Hidden Layers: 4 Neurons per layer: 4
Plough use	-30.346	-39.505	-40.553
	(27.008)	(22.416)	(26.584)
Observations	160	160	160

Notes: Analysis of Alesina et al. (2013) using DML-IV. Standard errors adjusted for variability across splits using the median method are reported in parentheses.

Table B.11: DML sensitivity analysis with 5 folds

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) OLS
	Panel A: DML, Partially linear model with full set of controls. Outcome: Female labour force							
Plough use	-4.408	-2.501	-5.591	-4.389	-4.262	-4.465	-4.204	-9.234
	(4.843)	(4.251)	(4.029)	(4.167)	(4.455)	(4.219)	(4.156)	(4.301)
Observations	142	142	142	142	142	142	142	142
	Lasso	Reg. Tree	Boosting	Forest	Neural Net.	Ensemble	Best	2SLS
	Panel B: DML-IV. Outcome: Female labour force							
Plough use	-48.897	-47.442	-47.816	-43.191	-20.339	-40.641	-44.467	-28.516
	(21.390)	(26.887)	(17.568)	(17.980)	(29.529)	(22.349)	(29.470)	(7.559)
Observations	160	160	160	160	160	160	160	160

Notes: Analysis of the main robustness checks of Alesina et al. (2013) using DML. The DML method is implemented using 100 splits and 5 folds. Column 8 reports the results of the most complete robustness checks for the OLS and IV specifications in the original paper. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Robust standard errors are reported in column 8.

Table B.12: DML Sensitivity analysis using alternative score function, partially linear model

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) OLS
	Panel A: Baseline controls, female labour force participation 2000					
Plough use	-12.53	-11.727	-11.999	-10.833	-13.02	-12.401
	(2.73)	(3.711)	(3.768)	(2.772)	(5.01)	(2.964)
	Panel B: Baseline controls, share of firms with female ownership					
Plough use	-11.083	-13.135	-13.552	-12.575	-12.311	-15.241
	(3.171)	(4.894)	(5.525)	(3.797)	(5.655)	(4.06)
	Panel C: Baseline controls, share of political positions held by women 2000					
Plough use	-2.130	-5.418	-5.268	-5.421	-6.117	-4.821
	(1.592)	(2.306)	(2.042)	(1.550)	(2.810)	(1.782)
	Panel D: Full set of controls, female labour force participation 2000					
Plough use	-11.113	-7.296	-6.551	-6.455	-7.520	-9.234
	(3.423)	(6.108)	(4.954)	(3.478)	(8.563)	(4.301)

Notes: Analysis of Table 4 (columns 1, 3, 5) and of the main OLS robustness check of Alesina et al. (2013) using DML, implementing the IV-type score function. Column 6 reports the original paper results. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Robust standard errors are reported in column 6.

Table B.13: DML Sensitivity analysis trimming extreme values of propensity score, DML-IV

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) OLS
Plough use	-38.364	-24.981	-37.822	-34.432	-25.245	-28.516
	(16.848)	(15.257)	(14.860)	(13.624)	(21.830)	(7.559)
Observations	160	160	160	160	160	160

Notes: Analysis of the main IV robustness check of Alesina et al. (2013) using DML. The propensity scores are trimmed at 0.01 and 0.99. Column 6 reports the original paper results. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Robust standard errors are reported in column 6.