New Tools for Predicting Economic Growth Using Machine Learning: A Guide for Theory and Policy

Conference Paper · November 2015

CITATIONS
4

READS
4,344

3 authors:

James Thomas Bang
Saint Ambrose University
59 PUBLICATIONS **244** CITATIONS
SEE PROFILE

Tinni Sen
Virginia Military Institute
18 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Atin Basuchoudhary
Virginia Military Institute
76 PUBLICATIONS **242** CITATIONS
SEE PROFILE

New Tools for Predicting Economic Growth Using Machine Learning: A Guide for Theory and Policy

James T. Bang
St. Ambrose University

Atin Basuchoudhary
Virginia Military Institute

Tinni Sen
Virginia Military Institute

1. Introduction

In this paper, we develop a machine-learning framework for predicting growth. Apart from creating this framework, we also envisage this book as a sort of primer for using Machine Learning to answer economic questions. While Machine Learning itself is not a new idea, advances in computing technology combined with a dawning realization of its applicability to economic questions makes it a new tool for economists (Varian, 2014).

We identify the research questions and issues in section 2 and state our proposed methodologies in section 3. Section 4 describes our data as well as some of the issues in our dataset. Sections 5 and 6 present our results and some concluding thoughts about policy implications and avenues for future research.

2. Identification of Research and Issues

Xavier Sala-i-Martin summarizes an extensive literature on economic growth in a series of papers culminating in a short paper that finds a robust correlation between economic

growth and some variables (Sala-i-Martin, 1997). Sala-i-Martin divides these "universal" correlates into 9 categories. These categories are as follows:

(1) **Geography.** For example, *absolute latitude* (distance from the equator) are negatively correlated with growth and certain regions such as sub-Saharan Africa and Latin America under-perform, on average.

(2) **Political institutions.** Measures of institutional quality like strong *Rule of Law*, *Political Rights*, and *Civil Liberties* improve growth, while instability measures like *Number of Revolutions and Military Coups* and *War* impede growth.

(3) **Religion.** Predominantly *Confucist/Buddhist* and *Muslim* countries grow faster, while predominantly *Protestant* and *Catholic* grow more slowly.

(4) **Market distortions and market performance.** For example *Real Exchange Rate Distortions* and *Standard Deviation of the Black Market Premium* correlate negatively with growth.

(5) **Investment and its composition.** *Equipment Investment* and *Non-Equipment Investment* are both positively correlated with growth.

(6) **Dependence on primary products.** *Fraction of Primary Products in Total Exports* are negatively correlated with growth while the *Fraction of gross domestic product* (GDP) *in Mining* is positively correlated with growth.

(7) **Trade.** A country's *Openness to Trade* increases growth.

(8) **Market orientation.** A country's *Degree of Capitalism* increases growth.

(9) **Colonial History.** *Former Spanish Colony* grow more slowly.

Sala-i-Martin's findings are standard in the growth literature. His econometric techniques cull the immense proliferation of explanatory variables into a tractable and parsimonious list. However, there are several problems with his approach.

First, many of the findings say nothing about why these variables matter, or which matter more than others. In fact, we note that if a country's GDP has a large *Fraction of Primary Products in Total Exports* it is likely to be a growth laggard, though if it has a high *Fraction of GDP* in Mining, it is in the high growth category. This sort of contradiction suggests that maybe the Sala-i-Martin list is not parsimonious enough. It is certainly not completely amenable to good theoretical explanations.

Second, econometric techniques focusing on growth and identifying the correlates of growth are not useful for *predicting* economic growth. Therefore, it is not possible to know whether certain theoretically ordained variables for explaining growth matter more than others. Machine Learning techniques, however, can create parsimonious lists of variables by focusing on the *predictive* power of variables. This approach can therefore distinguish between different theories of growth, as well as create better models of growth by whittling down the list of variables to those that are the best predictors of growth. Furthermore, Machine Learning has the advantage of not requiring any major assumptions about a variable's underlying distribution, or indeed any prior assumptions about theoretical links.

The predictive power of Machine Learning techniques has practical benefits as well. The policy maker, for instance, mainly needs to know the effect of a current change in policy on a future (out-of-sample) target. From the policy maker's perspective, a parsimonious list from the usual econometric techniques itself does not provide good policy levers for

increasing economic growth because existing studies that use these techniques neglect the issues of cross-validation and out-of-sample predictability. Thus, the policy maker has no idea whether moving a country towards a more capitalist direction, for example, will lead to higher growth because the econometric techniques used in finding these sorts of correlations have not usually been validated or tested for out-of-sample predictive ability. Machine Learning approaches address this problem by emphasizing both cross-validation as well as out-of-sample prediction scores for different specifications of the model. Moreover, policy makers can also get a sense of the relative importance of variables in predicting growth. Thus, they can prioritize policy levers according to which ones may have the greatest impact on economic growth.

Of course, both the policy maker and the academic remain concerned about causal links. However, the results of standard econometric techniques designed for establishing causal links, do not satisfy these concerns. This is because the growth literature's focus on growth accounting, and therefore on the correlates of growth, essentially end up generating long lists of possible correlates of growth. Such lists hamper standard econometric techniques since they are plagued by endogeneity problems. Of course, finding parsimonious lists of correlates of growth a la Sala-i-Martin certainly help in specifying econometric models to explore causal links. However, econometric techniques that attempt to solve the problem of endogeneity by using instrumental variables are problematic because the consistency of the estimated marginal effects depends critically on the strength and validity of the instruments. In fact, some of these instruments may actually lead to biased estimates (Bazzi & Clemens, 2013). Further, it is even possible for more sophisticated parametric methods to do *worse* than the traditional ordinary least squares

method. In contrast, Machine Learning techniques can help in this process by winnowing down the list of possible correlates by focusing on how well these variables predict out-of-sample. Thus, variables that appear to be robust correlates of growth but do not predict well out-of-sample, cannot really be causal variables, and can be eliminated. In this sense, Machine Learning can be helpful in exploring causal links to growth (Athey & Imbens, 2015).

Another problem in the growth literature is the paucity and unreliability of data for precisely the countries for which growth issues matter most. Standard statistical analyses do not perform well when there is missing data. Machine Learning can address this problem in a scientifically verifiable way by finding "surrogate" variables that can proxy those with missing data. These proxies are chosen by the Machine Learning techniques by their predictive abilities, and to that extent, provide a hard test for the usefulness of a particular proxy variable. Lastly, to the extent that econometric techniques focus on point parametric estimates, they ignore potential non-linear ways in which explanatory variables can affect economic growth.[1] Machine Learning techniques, on the other hand, can provide easily understandable graphs by capturing non-linear "marginal" effects of a particular variable on growth. Thus, we are able to say with some predictive accuracy whether over a certain range, a particular variable has a greater or lesser, negative or positive impact on growth, as well as identify other ranges where the same variable does not affect growth.

We plan to develop a model that will provide a framework for understanding the complex non-linear patterns that link formal political institutions, informal political

[1] Some regression techniques can capture these non-linearities to some extent. But, in these techniques, either the non-linearity has to be imposed on the model or the modeler has to use fairly complicated spline regression techniques. In any case, neither technique is usually scored by their ability to predict out-of-sample.

institutions, resource availability, and individual behavior to economic growth. Our empirical strategy will atheoretically incorporate the patterns that link underlying institutional, economic, political, social, and geographic factors to predict the rate of economic growth. Then, we will take those factors that our empirical model identifies as important, and suggest a roadmap to build a theoretical framework that explains how these fit into the story of growth.

3. Empirical Methods

We will build an empirical model using ML techniques (regression tree, artificial neural network, bootstrap aggregating, boosting, and random forest predictors) to provide an objective approach to finding *linear and non-linear* patterns on *how* publicly-available economic, geographic, and institutional variables *predict* growth (Hand, Mannila, & Smyth, 2001). First, we will identify the Machine Learning approach that best predicts growth. Then, using the best technique, we will identify the variables that form the pattern that best predicts growth. This is important because, at least theoretically, there may be reason to believe that many of the correlates of growth have complex non-linear impacts of growth because strategic complementarities between variables can lead to multiple possible growth equilibria. For example, openness to trade may improve growth *on average*, but only if a country is not too dependent on natural resources or other primary commodities that may lead to conflict among competing factions over the rents generated by the resource sector.

Machine Learning techniques identify tipping points in the range of a particular variable that may place a country in a lower or higher growth category. Moreover, Machine Learning can generate partial dependence plots. These graphs can illustrate how variables

identified as good predictors of growth relates (perhaps non-linearly) to growth. Further, by identifying the variables that have the *most* predictive power we could help develop a framework to distinguish between competing theoretical explanations of growth. Suppose, for instance, political economy models may suggest that income inequality is important in explaining growth, but neoclassical models may predict that education matters more. If ML methodologies rank income inequality as a better predictor of growth than population density, we can assume that the political economy model may itself be a better explanation of growth than the neoclassical model, or vice versa. This would then suggest greater econometric scrutiny of Theory A in teasing out causal patterns. Moreover, this Machine Learning approach can help eliminate correlates of conflict that do not predict economic growth well. Presumably, correlates that do not predict well cannot really be considered as variables that cause growth.

Our Machine Learning approach, and the econometric tests arising out of this approach, will help us better understand causal patterns explaining growth. Moreover, we offer a better understanding of how growth can be *predicted*, which will be of particular help to policy makers as they design policies of economic growth.

3.1 Classical and Other Regression Analysis

In general, using given data from a learning sample, $\mathcal{L} = \{(y_1, \mathbf{x}_1), \ldots (y_N, \mathbf{x}_N)\}$, any prediction function, $d(\mathbf{x}_i)$, maps the vector of input variables, $\mathbf{x}$, into the output variable, y. An effective prediction algorithm seeks to define parameters that minimize some error function over the predictions. Common error functions that many predictors use include the mean (or sum) of the absolute deviations of the observed values from the predicted values or the mean (or sum) of the squared deviations. In linear regression models $d(\mathbf{x}_i)$ is

simply a linear function of the inputs and their respective slope coefficients, plus a constant, $d(\mathbf{x}_i) = \mathbf{x}_i\beta$. For linear models, we can express the minimization condition as:

$$R_{MAD}(d) = \frac{1}{n}\sum_{i=1}^{N}|y_i - d(\mathbf{x}_i)|,$$

and:

$$R_{OLS}(d) = \frac{1}{n}\sum_{i=1}^{N}(y_i - d(\boldsymbol{x}_i))^2,$$

where $d(\mathbf{x}_i) = \mathbf{x}_i\beta$ is a linear function of the inputs.

Under certain conditions, the predictor ($\mathbf{x}_i\beta$) that minimizes the mean absolute deviations function can be shown to be a good estimate for the conditional *median* (or other conditional quantiles if the absolute deviation function is "tilted" as in quantile regression). Correspondingly, the predictor that minimizes the least squares function can be shown to be a good estimate for the conditional *mean*, or expected value of y, $E(y|\mathbf{x})$.

Although linear regression can sometimes yield good predictors, it is important to realize that the main objective of linear models is to estimate *causal* effects for one or more hypothesized determinants of y holding all of the other hypothesized determinants in a model constant. More sophisticated methods of estimating linear regression models (such as ones that use instrumental variables or other two-step methods) focus on purging the marginal causal effects of bias that might result from endogeneity, selection bias, or misspecifications of the functional form of the target variable address the problem of bias in the estimated marginal effects to the detriment of the model's overall predictive accuracy.[2]

[2] For example, it is fairly well-known that a two-stage least squares estimator can sometimes yield a *negative* value for the regression R^2. This implies that the sum of squared errors of the model exceed the total sum of squares of the target variable.

3.2 Regression Tree Predictors

Classification and regression trees (CART)[3] diagnose and predict outcomes by finding binary splits in the input variables to optimally divide the sample into subsamples with successively higher levels of purity in the outcome variable, *y*. So, unlike linear models, where the parameters are linear coefficients on each predictor, the parameters of the tree models are "if-then" statements that split the dataset according to the observed values of the inputs.

More specifically, a tree, *T*, has four main parts:

1. Binary splits to splits in the inputs that divide the subsample at each node, *t*;
2. Criteria for splitting each node into additional "child" nodes, or including it in the set of terminal nodes, T^*;
3. A decision rule, $d(\mathbf{x})$, for assigning a predicted value to each terminal node;
4. An estimate of the predictive quality of the decision rule, *d*.

The first step is achieved at each node by minimizing a measure of impurity. The most common measure of node impurity, and the one we use for our tree algorithms, is the mean square error, denoted $\hat{R}(d)=\frac{1}{n}\sum_{i=1}^{N}(y_i - d(\mathbf{x}_i))^2$. Intuitively, this method searches for the "best" cutoff in each of the inputs to minimize errors, then selecting which of the inputs yields the greatest improvement in node impurity using its optimal splitting point.

Then, a node is declared to be terminal in the second step if one of the following conditions is met: (1) that the best split determined by the application of step fails to improve the node impurity by more than a predetermined minimum improvement

[3] We provide only a brief summary of tree construction as it pertains to our objectives. For a full description of the CART algorithm, see Breiman, et al. (1984).

criterion; or (2) the split creates a "child" node that contains fewer observations than the minimum allowed.[4] At each terminal node, the decision rule typically assigns observations with a predicted outcome based on the outcome that is most frequent (more than one half of the observations in that node for binary outcomes, for example).[5]

The predictive quality of the rule is also evaluated using the *mean square error*, $\hat{R}(d)=\frac{1}{n}\sum_{i=1}^{N}(y_i - d(\boldsymbol{x}_i))^2$. This misclassification rate is often cross-validated by splitting the sample several times and re-estimating the misclassification rate each time to get an average misclassification of all of the cross-validated trees.

3.2.1 Boosting Algorithms

Combining ensembles of trees can often improve the predictive accuracy of a CART classifier. The bootstrap aggregating (bagging) predictor, boosting (adaptive boosting and other generalizations of boosting) algorithm, and random forest predictors all predict outcomes using ensembles of classification trees. The basic idea of these predictors is to improve the predictive strength of a "weak learner" by iterating the tree algorithm many times by either modifying the distribution (boosting) or randomly resampling the distribution (bagging). Then either classify the outcomes according to the outcome of the "strongest" learner once the algorithm achieves the desired error rate (boosting), or according to the outcome of a vote by the many trees (bagging).

[4] Note that there is a tradeoff here: setting lower values for the minimum acceptable margin of improvement or the minimum number of observations in a child node will lead to a more accurate predictor (at least within the sample the model uses to learn). However, improving the accuracy of the predictor within the sample will also lead to a more complex (and therefore less easily-interpreted) tree, and may lead to over-fitting in the sense that the model will perform more poorly out-of-sample.

[5] It is possible, however to consider decision rules that assign one class of a binary outcome anytime the proportion of observations exceeds one-third of the total observations in that node, especially if one type of misclassification error is more costly than the other.

Boosting is a way proposed by Freund and Schapire (1996) to augment the strength of a "weak learner" (a learning algorithm that predicts poorly) and making it a "strong learner." More specifically, for a given distribution $\mathcal{D}$ of importance values assigned to each observation in $\mathcal{L}$, and for a given desired error $\tilde{R}$, and failure probability, ϕ, a *strong learner* is an algorithm that has a sufficiently high probability (at least $1 - \phi$) of achieving an error rate no higher than $\tilde{R}$. A weak learner has a lower probability (less than $1 - \phi$) of achieving the desired error rate. Adaboost creates a set of M classifiers, $F = (f_1, \ldots, f_M)$that progressively re-weight the importance of each observation based on whether the previous classifier predicted it correctly or incorrectly. Modifications of the boosting algorithm for classification have been developed for regression trees by Freund and Schapire (1997) and Friedman(2001).

Starting with a $\mathcal{D}_1 = (1/N, \ldots, 1/N)$, suppose that our initial classifier, $f_1 = T$ (the single-tree CART predictor), is a "weak learner" in that the misclassification rate, $\hat{R}(d)$ is greater than the desired maximum desired misclassification rate, $\tilde{R}$. Next, for all observations in the learning sample, recalculate the distribution weights for the observations as:

$$\mathcal{D}_2 = \frac{\mathcal{D}_1(i)}{Z_2} \times \begin{cases} \frac{\hat{R}_1(d)}{1-\hat{R}_1(d)} & if\ d_1(\mathbf{x}_i) = y_i \\ 1 & otherwise \end{cases},$$

where Z_m is a scaling constant that forces the weights to sum to one.

The final decision rule for the boosting algorithm is to categorize the outcomes according to $d(x) = \arg\max_{y \in Y} \sum_{m:d_m(\mathbf{x})=y} \log\left(\frac{1-\hat{R}_m(d)}{\hat{R}_m(d)}\right)$. Using this decision rule and its corresponding predictions, we calculate the estimate of the misclassification rate in the same way as in step (4) of the single tree algorithm.

3.2.3 Bootstrap Aggregating (Bagging) Predictor

The bagging predictor proposed by Breiman (1996) takes random resamples $\{\mathcal{L}^{(M)}\}$ from the learning sample *with replacement* to create *M* samples using only the observations from the learning sample. Each of these samples will contain *N* observations – the same as the number of observations in the full training sample. However, in any one bootstrapped sample, some observations may appear twice (or more), others not at all.[6] The bagging predictor then adopts the rules for splitting and declaring nodes to be terminal described in the previous section to build *M* classification trees.

To complete steps (3) and (4), the bagging predictor needs a way of aggregating the information of the predictions from each of the trees. The way that bagging predictors (and other ensemble methods) do this for class variables is through *voting*. For *classification trees* (categorical target variables), the voting processes each observation[7] through all of the *M* trees that was constructed from each of the bootstrapped samples to obtain that observation's predicted class for each tree. The predicted class for the entire model, then, is equal to the mode prediction of all of the trees. For *regression trees* (continuous target variables), the voting process calculates the mean of the predicted values for all of the bootstrapped trees. Finally, the predictor calculates the redistribution estimate in the same way as it did for the single classification tree, using the predicted class based on the voting outcome for each predictor.

3.2.4 Random Forests

[6] Note that the probability that a single observation is selected in each draw from the learning set is $1/N$ Hence, sampling with replacement, the probability that it is completely left out of any given bootstrap sample is $(1 - 1/N)^N$. For large samples this tends to $1/e$. The probability that an observation will be completely left out of all *M* bootstrap samples, then, is $(1 - 1/N)^{NM}$.

[7] Note that the observations under consideration could be from the in-sample learning set or from outside the sample (the test set).

Like the bagging predictor, the random forest predictor is a tree-based algorithm that uses a voting rule to determine the predicted class of each observation. However, whereas the bagging predictor randomizes the selection of the observations into the sample for each tree, and then builds the tree using the same procedure as CART, the random forest predictor may randomize over multiple dimensions of the classifier (Breiman, 2001). The most common dimensions for randomizing the trees are the selection of the inputs for node of each tree, as well as the observations included for constructing each of the trees. We briefly describe the construction of the trees for the random forest ensemble below.

A random forest is a collection of tree decision rules, $\{d(x, \Theta_m), m = 1, \ldots, M\}$, where Θ_m is a random vector specifying the observations and inputs that are included at each step of the construction of the decision rule for that tree. To construct a tree, the random forest algorithm takes to following steps:

i. Randomly select $n \leq N$ observations from the learning sample;[8]
ii. At the "root" node of the tree, select $k \in K$ inputs from $\mathbf{x}$;
iii. Find the split in each variable selected in (ii) that minimizes the mean square error at that node and select the variable/split that achieves the minimal error;
iv. Repeat the random selection of inputs and optimal splits in (ii) and (iii) until some stopping criteria (minimum improvement, minimum number of observations, or maximum number of levels) is met.

[8] In contrast to bagging, where the number of observations selected for each tree exactly equals the total number of observations in the learning sample, and the draws always sampled with replacement, the number of observations selected for each tree of the forest can be set to be less than the total size of the learning sample, and can therefore be sampled with *or without* replacement. This also allows for slightly greater flexibility with respect to stratified or clustered sampling from the learning sample.

The bagging predictor described in the previous sub-section is in fact a special case of the random forest estimator where, for each tree, Θ_m consists of a random selection of $n = N$ observations from the learning sample with replacement (and each observation having a probability of being selected in each draw equal to $1/N$) and sets the number of inputs to select at each node, k, equal to the full length of the input vector, K so that all of the variables are considered at each node.

3.3 Validation and Testing of Predictive Accuracy

Once we have built our dataset and imputed the missing values, the next issue is to evaluate the validity of our error estimates and the predictive strength of our models. Error estimates ($R[d]$) can sometimes be misleading if the model we are evaluating is over-fitted to the learning sample. These error estimates can be tested out-of-sample and also cross-validated using the learning sample.

To test the out-of-sample validity, we simply split the full dataset into two random subsets of *countries*: the first, known as the *learning sample* (or training sample) contains the countries and observations that will build the models; the second, known as the *test sample*, will test the out-of-sample predictive accuracy of the models. The out-of-sample error rates will indicate which models and specifications perform best, and will help reveal if any of the models are over-fitted.

To validate the error rates, machine learning uses a method known as cross-validation. Cross-validation creates V random sub-samples of the data, predicts the observations in those sub-samples using the full model, and calculates the error estimate for each of them. It then averages the error estimates of all V random subsamples to give the cross-validated error estimate.

4. Data Sources and Issues

4.1 Variables and Sources

The target variable in our analysis is the five-year moving average of the yearly growth rate of real per capita gross domestic product (GDP). We have taken this variable from the World Development Indicators. We use this variable as our main proxy for increases in economic output and well-being.

In slight contrast to Sala-i-Martin's rather exhaustive search for robustly significant covariates, our goal is to focus on potential "policy levers." Thus, we omit the various fixed effects like geography, religion and colonial origin that some studies have found to matter. After all, a country cannot easily change its location, history, or religion! Instead, we begin with a long list of time-variant inputs from the World Development Indicators (and some other sources) that other studies have found to explain growth. The first two of these variables, lagged real per capita GDP growth and lagged real per capita GDP, which proxy for the persistence effects of past growth and convergence effects, respectively. From there, we add several variables relating to the composition of domestic output and expenditures: Consumption, investment, industry, total trade, imports, exports, mineral rents, fuel imports, fuel exports, foreign direct investment, government expenditure, military expenditures, and foreign aid/development assistance. Each of these variables is measured in terms of its share of GDP and is lagged in a similar way to the lagged values of GDP per capita and its growth rate. As a measure of the level of and penetration of technology in the economy, we add the number of phones as a percentage of the total population.

Next, we include lagged values of several variables to account for domestic monetary and price factors that may impact growth: the money supply (as a share of GDP), the rate of

growth in the money supply, the CPI inflation rate, the lending interest rate, the real interest rate, and the interest rate spread. We add several additional factors that capture the impacts of external forces on price levels: the terms of trade, the export price index, and the import price index. These variables also come directly from the World Development Indicators.

We also include several lagged variables pertaining to demographics and human development. The variables in the WDI from this category are: the total population, the population growth rate, the rural population as a percentage of the total population, the dependency ratio (measured as the ratio youth aged 0-15 and elderly aged 65 and over to the working-age population aged 16-64), life expectancy, and the gross secondary school enrollment rate. To these, we add the Gini coefficient measure of income inequality, which we have obtained from the Standardized World Income Inequality Database (SWIID) compiled by Solt (2014).

Finally, we consider variables that capture various aspects of institutional quality and stability from the International Country Risk Guide (ICRG), Database of Political Institutions (DPI), Political Instability Task Force (PITF), and Cross-National Time Series (CNTS) datasets. The eight variables from the ICRG that we include in our EFA are: *government stability*, which assesses "the government's ability to carry out its declared programs and … stay in office"; the *democratic accountability* index; the *investment profile* index, which captures the enforcement of contractual agreements and expropriation risk; the *corruption* index, which measures the *absence* of corruption; the index of *bureaucratic quality*, which assesses the efficiency of the bureaucracy; *internal conflict*, which captures the *absence* of internal civil war; *external conflict*, which similarly measures the absence of

foreign wars; and *ethnic tensions*, which provides an *inverse* measure of the extent to which racial and ethnic divisions lead to hostility and violence.

Next, we include nine variables from the DPI dataset. They are: legislative fractionalization; political polarization, which takes values between 0 and 2 depending on the ideological distance between the legislature and the executive; the number of years the current chief executive has served; the number of changes in the number of veto players in the government; a legislative Herfindahl-Hirschman index; the number of veto players within the government; whether allegations of fraud surfaced in the last election; the legislative index of electoral competition; and the executive index of electoral competition.

To these we add nine measures from the CNTS, which are: assassinations, strikes, government crises, demonstrations, purges, riots, major cabinet changes, changes in the effective executive and the legislative effectiveness index. Finally, we include four variables from the PITF: the Polity 2 democracy index, regime durability, ethnic wars, and nonethnic civil war.

4.2 Problems with Institutional Measures

Simply including a subset of those measures is problematic, for three reasons. First, although they purport to gauge distinct aspects of institutional character, many of them overlap substantially, and most of them are highly correlated with one another. Second, the subjective nature of these de facto indices of quality may expose them to considerable measurement error. Third, institutional quality has been shown to be multidimensional (Bang, Basu, and Mitra, 2015), and the different dimensions may have different impacts. The nonparametric methodology that we adopt partially avoids that issue in the sense that we do not need to worry about obtaining biased parameter estimates. However, we do

need to worry that similar measures of institutional quality that represent the same underlying concept might dominate our classification.

In order to purge our institutional measures from some of these problems we perform an exploratory factor analysis (EFA) on the institutional measures described above. EFA is similar in some respects to the more familiar technique of principle components analysis (PCA) in that both EFA and PCA reduce the dimensionality of the observed variables based on the variance-covariance matrix. However, in contrast to PCA, which seeks to extract the *maximum* amount of variation in the correlated variables, EFA seeks to extract the *common* sources of variation. To achieve this, EFA expresses the observed variables as linear combinations of the latent underlying factors (and measurement errors), whereas PCA expresses the latent components as combinations of the observed variables.

We report the results of the factor analysis in the appendix in Table A1. From the factor loadings, we identify seven common factors out of the list of institutional variables:

(1) **Democracy** is comprised by the Polity index; the legislative and executive indices of electoral competition; legislative fractionalization; and democratic accountability.

(2) **Violence** consists of the internal and external conflict indices, ethnic tensions, and the presence of ethnic conflict and civil war. Higher scores indicate greater *stability*.

(3) **Transparency** incorporates the corruption, bureaucratic quality, and democratic accountability indices, along with regime durability and fraud.

(4) **Protest** is constructed primarily from the numbers of demonstrations, riots, and strikes in society. Higher numbers indicate greater *unrest*.

(5) **Within-regime instability** includes legislative concentration and fractionalization, as well as political polarization. Countries with more fractious governments receive higher values.

(6) **Credibility** is formed by the investment profile and government stability indices.

(7) **Regime instability** is composed of the numbers of executive changes and major cabinet changes, along with the changes in veto players and executive tenure.

One useful feature of these results is that they bear a striking similarity to the factors previously derived by Jong-a-Pin (2009) and Bang and Mitra (2011). For this reason, we have applied the same terms in our interpretation of these factors. In this sense, our results are quite consistent with previous contributions to the literature on institutions that employ factor analysis.

4.3 Missing Data

Another problem with many empirical studies of growth is that many of the variables are missing for a substantial portion of any time sample. As an immediate consequence of this, simply cobbling together a dataset that includes a diverse range of input variables *and* covers a wide range of countries over a long period is nearly impossible. A secondary consequence, therefore, is that any study of growth must trade off bias resulting from sample selection on the one hand, against omitted variables on the other hand.

Tree-based Machine Learning techniques deal with the problem well because if data for the optimal splitting variable at any particular node is missing for an observation, the algorithm can substitute the missing information in one of two ways. First, a regression tree will attempt to complete the splits using surrogate information from other variables that track the values of the optimal splitting variables very closely. If that is not possible,

then the tree model will split the missing values based on the conditional median (or mode for categorical variables) for the observations in that node.

Thus, Machine Learning actually suggests a useful way to impute data: Replace missing values in the dataset with the median (mode) value, conditional on the observed values of both the target and input variables up until reaching the node where the model encountered the missing values. While this imputation tactic may not be ideal for a single iteration of a tree model, conditioning the imputed values on the observed inputs and outputs of a few *hundred* random trees (as would be the case with the random forest model) is likely to yield reasonably good imputed values. Studies that have tested the validity of random forest imputation using simulated missing values have found that this imputation method performs comparably, and often better than, other methods of imputation (such as multiple imputation and OLS). We report summary statistics for the learning and test samples of our raw and imputed datasets in Tables 1 and 2.

5. Results

5.1 Comparison of Prediction Methods

Column 1 of Table 2 presents the in-sample mean squared errors from the growth models we have predicted using the full dataset of 39 inputs. Column 2 presents the corresponding out-of-sample mean squared errors. Each of these models uses a random forest imputation method to ensure that we evaluate the methods based on the same samples.

For this specification, the bagging and random forest models perform the best and perform about equally well on average. While the bagging model performs slightly better out-of-sample, the random forest model performs slightly better in-sample.

Columns 3 and 4 of Table 2 report the corresponding in-sample and out-of-sample prediction errors for the models we have predicted using the subset of 20 predictors that remained after culling the variable list based first on eliminating the variables that measured essentially the same thing (eliminating import prices, terms of trade, imports, trade, interest rate spread, lending interest rate and industry). Then, second, we reduced the variable list by keeping only the 20 most important variables of the remaining list (eliminating government spending, fuel imports and exports, income inequality, secondary enrollment, aid and development assistance, life expectancy, population growth, dependency ratio, phones, money supply, and mineral rents).

The remaining variables are: lagged GDP per capita and its growth rate, export prices, export volumes, foreign investment, domestic investment, domestic consumption, military spending, money growth, inflation, real interest rates, total population, percentage of rural population, regime stability, credibility, transparency, democracy, protest, within-regime stability, and ethnic violence.

Again, we see that the bagging and random forest estimators perform the best, with the bagging model doing better out-of-sample, and the forest model doing better in-sample. Thus, the random forest model seems to do a good job of predicting. Because of the randomization of the variables selected at each node, it also has the useful property of ensuring that all of the variables have a decent chance to be incorporated into the model (unlike the bagging model which may completely ignore certain variables).

It is worth noting that the 20 variable model does not give up much in comparison to the 39-variable model in terms of predictive accuracy. In terms of in-sample accuracy, the linear regression model loses the most from the omission of the extra variables, with the

MSE performs a little less than 9% worse, with the MSE increasing from about 10.85 to about 11.80. In the case of the boosting model, however, the in-sample prediction is exactly the same (perhaps because this algorithm sets a maximum target error). Out-of-sample, we see comparable similarities in all of the models, except it is worth noting here that omitting some variables actually *improves* the performance of the linear regression model. This suggests that, at least for that type of model, a large model is prone to the problem of *over fitting* the model to the learning sample.

4.2 Variable Importance

Now, we shift focus to analyze the impacts of the individual variables that the random forest model identifies as "important" in terms of the margin by which they improve our prediction of growth rates. Tables 3 and 4 reports the variable importance levels for the 39-variable and 20-variable specifications of the random forest model, respectively. The first thing we notice from these tables is that the strongest predictors of growth are, predictably, lagged GDP per capita and lagged GDP per capita growth. This does not come as a big surprise given the strong theoretical basis and empirical support for autocorrelation and convergence effects in growth series. Beyond that, the next several predictors of growth are similar across the two specifications. This is also not too surprising, since the smaller specification has culled the variables that are least predictive. Indeed, this ability to ignore relatively unimportant predictors is a strong advantage of the tree-based learning models.

As we might expect, lagged GDP per capita and lagged GDP per capita growth rank as the top predictors of current GDP growth rates in both the larger and smaller models. This is largely attributable to the high degree of persistence in GDP and GDP growth. After that,

variables that the larger model identifies as important include: Export and import prices and the terms of trade (3rd, 6th, and 17th, respectively); various measures of institutional quality, especially regime stability, credibility, transparency, and democracy (4th, 5th, 10th, and 12th); openness to foreign flows of goods and capital (exports, FDI, imports, and total trade – 7th, 8th, 18th, and 24th), and the composition of private expenditures (consumption and investment, 9th and 13th). Some of the inputs that did *not* matter too much were the level of government expenditures, fuel imports/exports and mineral rents, income inequality, the dependency ratio, and, perhaps most surprisingly, secondary school enrollment.

The variables that the smaller model identifies similar variables as important for growth. Export prices (3rd), selected institutional variables (regime instability, 4th; credibility, 5th; transparency, 7th; and democracy, 10th), and openness to foreign flows (FDI, 6th, and exports, 9th) all make the top ten most important predictors of growth. Factors that still ail to register a high level of importance include military expenditures (20th), within regime instability (19th), size measured by population (18th), and ethnic violence (17th).

To address a potential technical concern, we are not necessarily claiming that the importance of the variables as determined by their improvement of predictive accuracy necessarily implies any sort of causal relationship. Nor do we claim that predictive accuracy should serve as a substitute for advanced regression methods that are well-suited for detecting such causal relationships. However, we do propose that the variables that do *not* predict well are also not likely to have causal effects that are large enough to care about, even if those causal effects are "statistically significant." Moreover, the results

suggests a list of variables which deserve a closer investigation of their theoretical and empirical causal links to growth.

4.3 The Impacts of Selected Input Variables on Growth

A partial dependence plot for each input variables can help to visualize the direction (and size) of the impact of each input variable on the rate of growth. The function that a partial dependence plot graphs is:

$\tilde{f}(x_k) = \frac{1}{N}\sum_{i=1}^{N} f(x_k, x_{i,-k})$,

where $f(x_k, x_{i,-k}) = d(x)$ is the predicted regression function. Therefore, the y-axis in the PDPs measures the growth rate conditional on *x* and the x-axis measures the values of the input variable. As reference points for the distribution of x-values, we have included inner tick marks for the deciles (10^{th} percentile, 20^{th} percentile, etc.) in the distribution. We report these PDPs in figures one through twelve for the top several variables in the 20-variable specification of the random forest prediction model.

The PDPs demonstrate that, in most cases, the most predictive variables for growth actually show a monotonic relationship. Export prices, credibility, transparency, and export volume all increase growth rates uniformly, while regime instability, consumption, money supply growth, protest, and, interestingly, democracy, decrease growth. Also, and somewhat surprisingly, many of the same variables exhibit a fairly linear relationship with growth over much of the distribution.

In a few cases, however, PDPs show that in some cases, values in the extreme values of the *x* variables can have dramatic, and sometimes slightly counter-intuitive, impacts on growth. Take, for example, foreign direct investment as a share of GDP. Over much of the distribution (between the highest and lowest tick marks inside the x-axis), FDI appears to

have a small but linear impact on growth, with growth rising from about 1.9 percent at around the 10th percentile and rising to about 2.1 percent at the 80th percentile. However, as investment balloons to very high values (greater than 20% of GDP), growth increases dramatically, reaching a plateau of about 4.6 percent growth at investment rates a little over 40 percent of GDP. At the other extreme, very high rates of lagged foreign *disinvestment* can also increase current predicted growth rates. We observe a similar U-shaped impact for domestic investment (capital formation as a percentage of GDP).

As another example, consider export volumes. While most studies find exports to have statistically significant effects at the mean, the tree prediction algorithms show that the impacts may not be so simple. While there exports generally increase growth, this impact remains modest through the bottom 95% of the distribution of export volumes, rising from about 1.75% to about 1.9 % over this range. But, when exports reach a level above about 30% of GDP, growth seems to take off, rising to about 12% by the time exports reach their maximum at around 60% of GDP.

One counter-intuitive example is democracy. Even though higher quality and more stable institutions tend to contribute positively to growth in most studies, our data suggest that there may be a negative (albeit small) impact of democracy on growth. This does not come completely as a surprise, since we construct the institutional factors in such a way as to extract the common variance from the observed institutional variables that is most correlated with elections and electoral competition. While electoral competition may often be a good thing, it can also lead to an increase in the number of groups competing to extract rents, and therefore may – in the absence of other complementary institutions – impede growth. Hence the impact of democracy may be ambiguous on the margin.

Regression methods make it difficult to discern the effects at the middle of the distribution of each input variable from the impacts at the mean because the impacts are assumed to be the same at all ranges of *x*. This is even true for quantile regression, which estimates the impact of the input variables on the various conditional quantiles of the *target* variable, but does so near the input variables *means*.

6. Conclusion

While more traditional models will continue to be the gold standard for teasing out causal effects of individual variables on the margin, we have found that Machine Learning models do a better job than some traditional models for the purpose of predicting. However, we do not claim that Machine Learning should be viewed as a substitute for theory and traditional regression models. Machine Learning can (and should) be used as a complement to traditional methods

One important function that Machine Learning is that it can be a useful a tool for sifting through the glut of data available to study and predict growth. With the ever-growing number of variables and indices available from governments, the World Bank, private think tanks, and individual researchers, it can be hard to tell which variables serve as the best proxies for a given economic or political concept. Data reduction techniques such as EFA, variable importance rankings, and data regularization methods such as stepwise regression and LASSO provide a more useful means for choosing which variables should be investigated more closely in economic theory and tested more carefully for causal significance.

Machine learning also deals more constructively with the problem of missing data than do many traditional models. Most studies of growth can be considered to be severely

biased by sample selection due to the fact that traditional models automatically drop any country-year observation that has a missing data point for even one of the variables specified in the model. Random forest imputation methods, on the other hand, help to deal with these missing values by first running the forest model with the missing values to impute the missing values using the full distribution of the missing variable conditional on all of the other observed input and output variables for that observation.

Finally, we are able to find the following:

1. Trade matters a lot. Growth responds favorably to an increase in export prices uniformly, but export volumes matter a lot at very high values.
2. Institutions also matter. Regime instability hurts growth, while strong property rights and an efficient, transparent bureaucracy helps growth. Interestingly, Democratic competition in isolation, seems to harm growth slightly.
3. Investment matters, and has large impacts in the extreme. Lagged values of both FDI and domestic capital formation increase growth a little through the center of the distribution, but countries with very high levels of investment and FDI see the highest rates of growth. Interestingly, countries with very low rates of investment and FDI in the previous periods also grow faster.
4. Reliance on primary commodities, inequality, schooling, and government spending – even military spending – do not matter much in either direction.

These results suggest that countries looking to promote growth would do best by focusing on trade (and in particular exports) and institutional quality and stability. However, the reforms to institutional quality that will prove the most helpful – establishing credible enforcement of

property rights and rooting out corruption in the bureaucracy – will also be the hardest and take the longest to implement.

References

Athey, S., & Imbens, G. (2015). Machine Learning Methods for Estimating Heterogeneous Causal Effects. *arXiv Preprints*, 1-9. Retrieved from http://arxiv.org/pdf/1504.01132v2.pdf

Basuchoudhary, A., & Cotting, D. (2014). Cultural Assimilation: The Political Economy of Psychology as an Evolutionary Game Theoretic Dynamic. *Evolutionary Behavioral Sciences, 8*(3), 209-22.

Basuchoudhary, A., & Razzolini, L. (2014). The Evolution of Revollution: Is splintering Inevitable. *VMI Working Paper Series*.

Bazzi, S., & Clemens, M. A. (2013). Blunt instruments: avoiding common pitfalls in identifying the causes of economic growth. *American Economic Journal: Macroeconomics, 5*(2), 152-86.

Blattman, C., & Miguel, E. (2010). Civil War. *Journal of Economic Literature, 48*(1), 3-57.

Esteban, J., & Ray, D. (2008). On the Salience of Ethnic Conflict. *American Economic Review, 98*(5), 2185-2202.

Fearon, J. D. (1995). Rationalist Explanations For War. *International Organization, 49*(3), 379-414.

Garfinkel, M. R., & Skaperdas, S. (2007). Economics of Conflict: An Overview. In T. Sandler, & K. Hartley, *Handbook of Defense Economics: Defense in a Globalized World* (Vol. 2, pp. 649-710). Amsterdam and Oxford: Elsevier, North-Holland.

Gates, S. (2002). Recruitment and Allegiance: The Microfoundations of Rebellion. *Journal of Conflict Resolution, 46*(1), 111-30.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining.* Cambridge: MIT Press.

Powell, R. (2006). War is a Commitment Problem. *International Organization, 60*(1), 169-203.

Sala-i-Martin, X. (1997). I Just Ran Four Million Regressions. *American Economic Review, 87*, 178-183.

Varian, H. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives, 28*(2), 3-28.

Tables

Table 1. Summary Statistics.

Table 1a. Summary Statistics for the Raw Data

	Learning Sample					Test Sample				
Source	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
GDP pc Growth	5,388	1.895	4.342	-42.600	51.920	2,322	2.177	4.480	-33.500	58.200
Lag GDP pc	4,762	831.697	2,099.474	-24,940.000	28,000.000	1,973	832.641	1,936.719	-3,900.000	22,100.000
Lag GDP pc Growth	4,723	0.123	5.024	-38.150	57.066	2,016	0.136	5.034	-46.686	37.587
Aid & Dev. Asst.	3,875	41,000,000	521,000,000	-4,970,000,000	7,090,000,000	1,795	64,100,000	650,000,000	-10,400,000,000	11,800,000,000
Consumption/GDP	4,364	-0.553	9.929	-133.600	181.120	1,829	-0.105	7.631	-39.925	41.980
Dependency	4,838	-2.161	4.358	-31.380	18.160	2,046	-2.979	4.606	-21.960	14.380
Export Prices	2,955	56.979	157.584	-410.200	4,820.000	1,359	51.624	97.053	-111.260	931.920
Exports/GDP	4,608	1.682	7.381	-38.366	58.000	1,881	1.390	9.014	-50.460	58.420
FDI/GDP	3,692	0.666	5.409	-56.184	97.240	1,645	0.838	3.634	-26.576	32.150
Fuel Exports/GDP	3,493	1.187	9.643	-81.061	87.770	1,510	0.517	7.992	-88.100	72.471
Fuel Imports/GDP	3,600	0.768	6.332	-52.833	23.319	1,606	0.603	5.805	-46.268	37.865
Gini Coefficient	2,690	0.055	3.030	-16.655	15.018	1,160	-0.162	3.113	-21.002	15.396
Government/GDP	4,492	-0.021	4.024	-72.920	39.933	1,795	0.253	3.313	-19.818	24.003
Growth	4,087	-20.369	369.406	-12,301.080	5,026.260	1,832	-10.415	340.905	-4,949.040	4,877.680
Import Prices	2,955	55.322	87.900	-164.933	805.600	1,359	50.389	76.799	-131.100	518.200
Imports/GDP	4,608	1.493	13.998	-180.280	273.653	1,881	1.253	9.096	-35.000	46.050
Industry/GDP	3,785	0.281	4.735	-39.440	22.360	1,594	-0.033	4.854	-17.960	47.160
Inflation	3,781	-7.922	339.607	-6,218.580	6,446.500	1,645	-12.503	228.239	-4,455.040	2,303.300
Interest Rate Spread	2,630	-1.728	25.780	-765.440	280.686	1,111	-30.362	522.007	-14,699.060	684.260
Investment	4,352	0.430	7.496	-90.480	84.850	1,832	0.010	5.749	-29.220	26.490
Lending Interest Rate	2,844	-4.420	46.467	-1,280.620	231.500	1,148	-246.255	4,355.439	-121,872.500	1,439.900
Life Expectancy	4,870	1.530	1.608	-16.140	15.140	2,079	1.433	1.471	-8.400	9.120
Military/GDP	2,020	-0.305	2.307	-35.688	32.470	882	-0.316	0.905	-5.790	5.409

Table 1a. Summary Statistics for the Raw Data (Continued)

	Learning Sample					Test Sample				
Source	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
Mineral Rents/GDP	4,216	0.113	1.675	-16.086	30.840	1,783	0.106	1.917	-18.314	16.579
Money/GDP	4,117	4.492	11.935	-129.200	89.400	1,838	-13.507	295.508	-6,916.880	93.000
Phones/Population	4,575	1.974	3.762	-18.240	20.440	2,025	1.853	3.377	-13.400	17.320
Population Growth	4,991	-0.051	1.026	-9.088	12.720	2,145	-0.075	0.747	-4.230	6.730
Real Interest Rate	2,776	1.052	10.077	-73.017	89.380	1,129	0.858	39.305	-846.991	416.320
Rural Population	4,991	-2.074	2.148	-14.660	9.900	2,144	-1.627	1.868	-9.500	2.640
Secondary Enrollment	3,453	4.440	7.237	-43.400	43.825	1,571	5.052	7.984	-44.400	54.200
Terms of Trade	2,285	0.280	28.202	-164.600	104.200	955	-1.463	33.681	-362.950	88.460
Total Population	4,991	3,020,207	10,400,000	-3,400,000	105,000,000	2,145	1,043,260	2,186,589	-2,320,000	17,200,000
Trade	4,608	3.173	18.080	-181.200	285.820	1,881	2.645	16.624	-70.000	93.800
Democracy	1,514	0.112	0.426	-2.333	2.783	604	0.159	0.470	-1.998	2.448
Transparency	1,514	-0.027	0.393	-2.217	2.043	604	-0.043	0.397	-1.831	1.711
Ethnic Violence	1,514	0.033	0.486	-1.868	2.786	604	0.063	0.472	-1.543	1.824
Protest	1,514	0.005	0.575	-5.641	4.543	604	0.027	0.316	-1.063	2.841
Regime	1,514	0.025	0.540	-3.061	3.452	604	0.045	0.560	-2.508	2.796
Within	1,514	-0.015	0.516	-2.480	2.090	604	-0.014	0.468	-3.503	1.799
Credibility	1,514	0.161	0.541	-1.535	2.277	604	0.185	0.543	-1.594	1.730

Table 1b. Summary Statistics for Random Forest Imputed Data

	Learning Sample					Test Sample				
Source	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
GDP pc Growth	5,388	1.895	4.342	-42.600	51.920	2,322	2.177	4.480	-33.500	58.200
Lag GDP pc	5,388	867.370	1,977.063	-24,940.000	28,000.000	2,322	884.323	1,790.693	-3,900.000	22,100.000
Lag GDP pc Growth	5,388	-0.186	4.794	-38.150	57.066	2,322	-0.216	4.790	-46.686	37.587
Aid & Dev. Asst.	5,388	34,400,000	442,000,000	-4,970,000,000	7,090,000,000	2,322	53,500,000	572,000,000	-10,400,000,000	11,800,000,000
Consumption/GDP	5,388	-0.342	8.981	-133.600	181.120	2,322	0.004	6.826	-39.925	41.980
Dependency	5,388	-2.159	4.131	-31.380	18.160	2,322	-2.884	4.333	-21.960	14.380
Export Prices	5,388	55.047	120.318	-410.200	4,820.000	2,322	52.791	78.542	-111.260	931.920
Exports/GDP	5,388	1.524	6.848	-38.366	58.000	2,322	1.296	8.130	-50.460	58.420
FDI/GDP	5,388	0.648	4.486	-56.184	97.240	2,322	0.768	3.073	-26.576	32.150
Fuel Exports/GDP	5,388	1.047	7.782	-81.061	87.770	2,322	0.607	6.460	-88.100	72.471
Fuel Imports/GDP	5,388	0.805	5.197	-52.833	23.319	2,322	0.701	4.852	-46.268	37.865
Gini Coefficient	5,388	0.015	2.154	-16.655	15.018	2,322	-0.098	2.210	-21.002	15.396
Government/GDP	5,388	0.071	3.684	-72.920	39.933	2,322	0.299	2.922	-19.818	24.003
Growth	5,388	-21.691	322.020	-12,301.080	5,026.260	2,322	-14.196	303.151	-4,949.040	4,877.680
Import Prices	5,388	54.235	69.346	-164.933	805.600	2,322	52.244	62.446	-131.100	518.200
Imports/GDP	5,388	1.458	12.949	-180.280	273.653	2,322	1.276	8.193	-35.000	46.050
Industry/GDP	5,388	0.211	3.995	-39.440	22.360	2,322	0.008	4.052	-17.960	47.160
Inflation	5,388	-7.488	285.602	-6,218.580	6,446.500	2,322	-11.163	193.563	-4,455.040	2,303.300
Interest Rate Spread	5,388	-6.653	23.504	-765.440	280.686	2,322	-19.629	361.443	-14,699.060	684.260
Investment	5,388	0.324	6.766	-90.480	84.850	2,322	0.017	5.134	-29.220	26.490
Lending Interest Rate	5,388	-43.949	132.578	-1,280.620	231.500	2,322	-155.796	3,065.645	-121,872.500	1,439.900
Life Expectancy	5,388	1.532	1.529	-16.140	15.140	2,322	1.446	1.393	-8.400	9.120
Military/GDP	5,388	-0.334	1.417	-35.688	32.470	2,322	-0.337	0.564	-5.790	5.409
Mineral Rents/GDP	5,388	0.113	1.485	-16.086	30.840	2,322	0.106	1.682	-18.314	16.579
Money/GDP	5,388	3.548	11.421	-129.200	89.400	2,322	-10.442	262.991	-6,916.880	93.000
Phones/Population	5,388	1.990	3.471	-18.240	20.440	2,322	1.902	3.158	-13.400	17.320
Population Growth	5,388	-0.045	0.988	-9.088	12.720	2,322	-0.067	0.719	-4.230	6.730

Table 1b. Summary Statistics for Random Forest Imputed Data

	Learning Sample					Test Sample				
Source	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
Real Interest Rate	5,388	1.001	7.325	-73.017	89.380	2,322	0.873	27.426	-846.991	416.320
Rural Population	5,388	-2.060	2.068	-14.660	9.900	2,322	-1.646	1.797	-9.500	2.640
Secondary Enrollment	5,388	4.576	5.822	-43.400	43.825	2,322	5.052	6.578	-44.400	54.200
Terms of Trade	5,388	-0.871	19.119	-164.600	104.200	2,322	-1.705	22.125	-362.950	88.460
Total Population	5,388	2,951,327	10,000,000	-3,400,000	105,000,000	2,322	1,121,556	2,119,833	-2,320,000	17,200,000
Trade	5,388	2.981	16.739	-181.200	285.820	2,322	2.574	14.982	-70.000	93.800
Democracy	5,388	0.112	0.239	-2.333	2.783	2,322	0.114	0.251	-1.998	2.448
Transparency	5,388	-0.020	0.225	-2.217	2.043	2,322	-0.011	0.217	-1.831	1.711
Ethnic Violence	5,388	0.032	0.264	-1.868	2.786	2,322	0.040	0.247	-1.543	1.824
Protest	5,388	0.008	0.313	-5.641	4.543	2,322	0.013	0.174	-1.063	2.841
Regime	5,388	0.051	0.332	-3.061	3.452	2,322	0.051	0.328	-2.508	2.796
Within	5,388	-0.016	0.278	-2.480	2.090	2,322	-0.017	0.243	-3.503	1.799
Credibility	5,388	0.151	0.336	-1.535	2.277	2,322	0.166	0.328	-1.594	1.730

Table 2. Comparison of Error Rates, Selected Predictors.

	39-Variable Model		20-Variable Model	
	Learning Sample	Test Sample	Learning Sample	Test Sample
Linear Regression	10.847	164.899	11.795	16.270
Single Tree	9.167	14.290	9.168	14.441
Bagging	9.011	12.860	9.123	13.113
Boosting	18.313	19.749	18.313	19.749
Random Forest	8.599	13.662	8.788	13.715

Table 3. Variable Importance Ranking, Random Forest Predictor, 39-Variable Specification

	% Increase in MSE
Lag GDP pc	5.553
Lag GDP pc Growth	5.063
Export Prices	0.912
Regime	0.794
Credibility	0.746
Import Prices	0.658
Exports/GDP	0.517
FDI/GDP	0.493
Consumption/GDP	0.385
Transparency	0.359
Money Growth	0.322
Democracy	0.278
Investment/GDP	0.267
Rural Population/Total	0.254
Real Interest Rate	0.209
Industry/GDP	0.207
Terms of Trade	0.202
Imports/GDP	0.196
Interest Rate Spread	0.174
Population	0.171
Inflation	0.171
Protest	0.170
Lending Interest Rate	0.158
Trade/GDP	0.150
Ethnic Violence	0.147
Within	0.146
Phones/Population	0.126
Life Expectancy	0.107
Aid & Dev. Asst/GDP	0.100
Military/GDP	0.096
Population Growth	0.090
Money/GDP	0.088
MineralRents/GDP	0.078
Gini Coefficient	0.062
Dependency	0.050
Secondary Enrollment	0.043
FuelExports/GDP	0.040
FuelImports/GDP	0.028
Government/GDP	0.026

Table 4. Variable Importance Ranking, Random Forest Predictor, 20-Variable Specification

	% Increase in MSE
GDP pc	5.403
GDP pc Growth	5.132
Export Prices	1.100
Regime	0.843
Credibility	0.796
FDI/GDP	0.585
Transparency	0.573
Consumption/GDP	0.572
Exports/GDP	0.558
Democracy	0.465
Investment	0.439
Money Growth	0.360
Rural Population/Total	0.343
Protest	0.332
Inflation	0.331
Real Interest Rate	0.271
Ethnic Violence	0.230
Population	0.213
Within	0.208
Military/GDP	0.168

Figures

Figure 1. Partial Dependence Plot for Export Prices

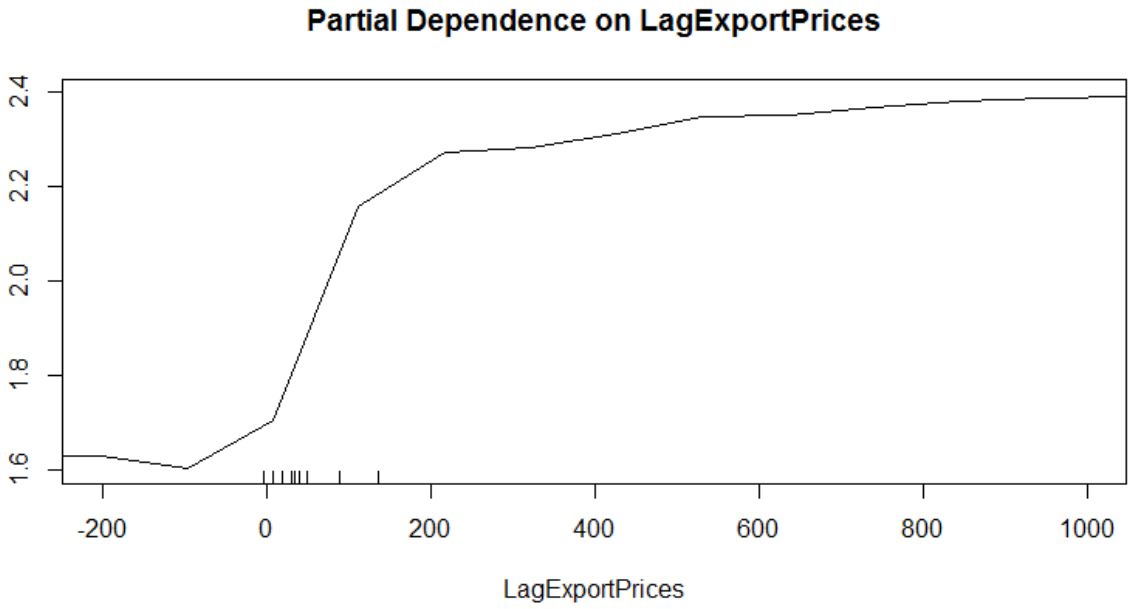


Figure 2. Partial Dependence Plot for Regime Instability

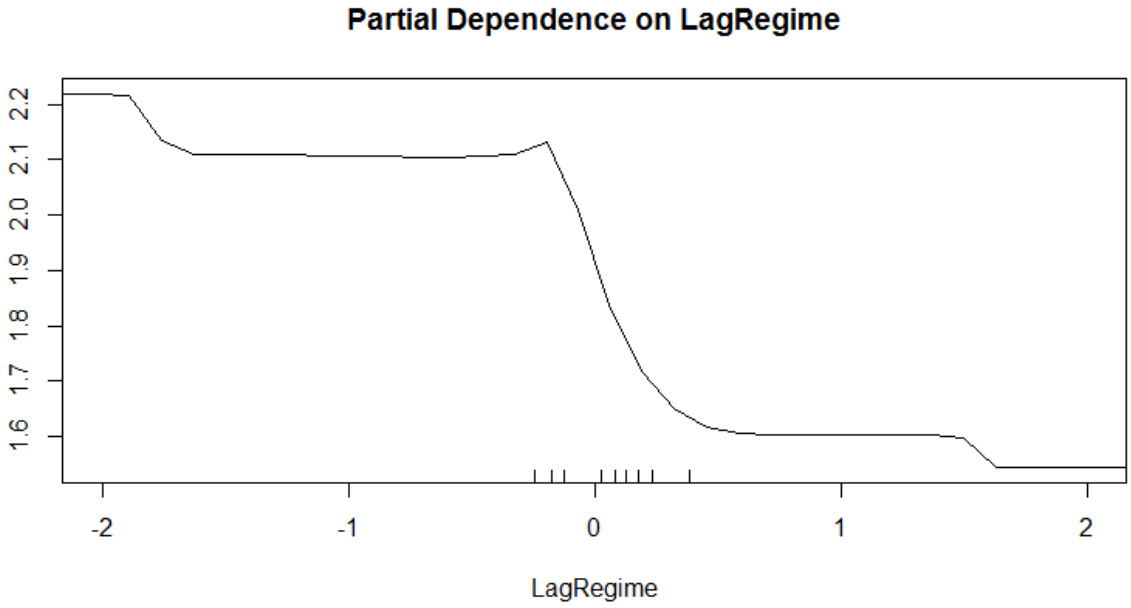


Figure 3. Partial Dependence Plot for Credibility

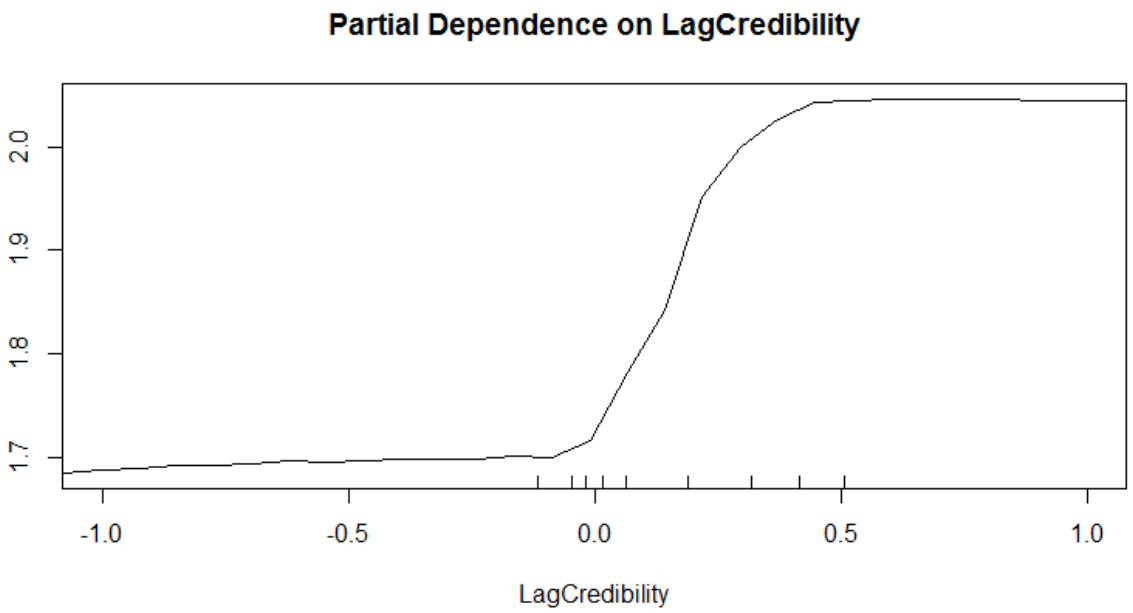


Figure 4. Partial Dependence Plot for FDI

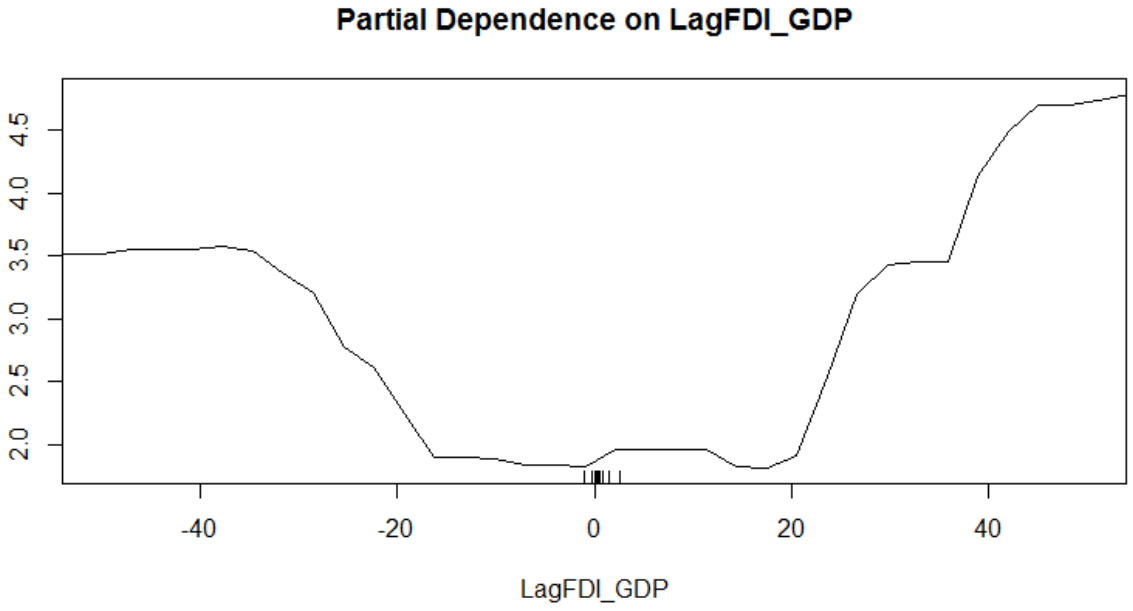


Figure 5. Partial Dependence Plot for Transparency

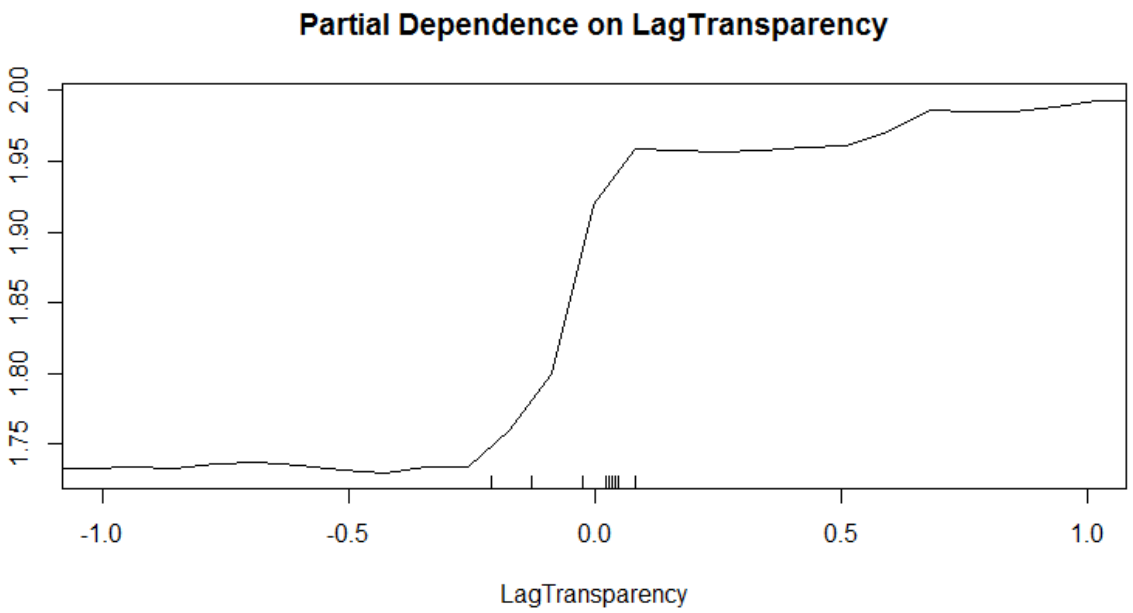


Figure 6. Partial Dependence Plot for Consumption

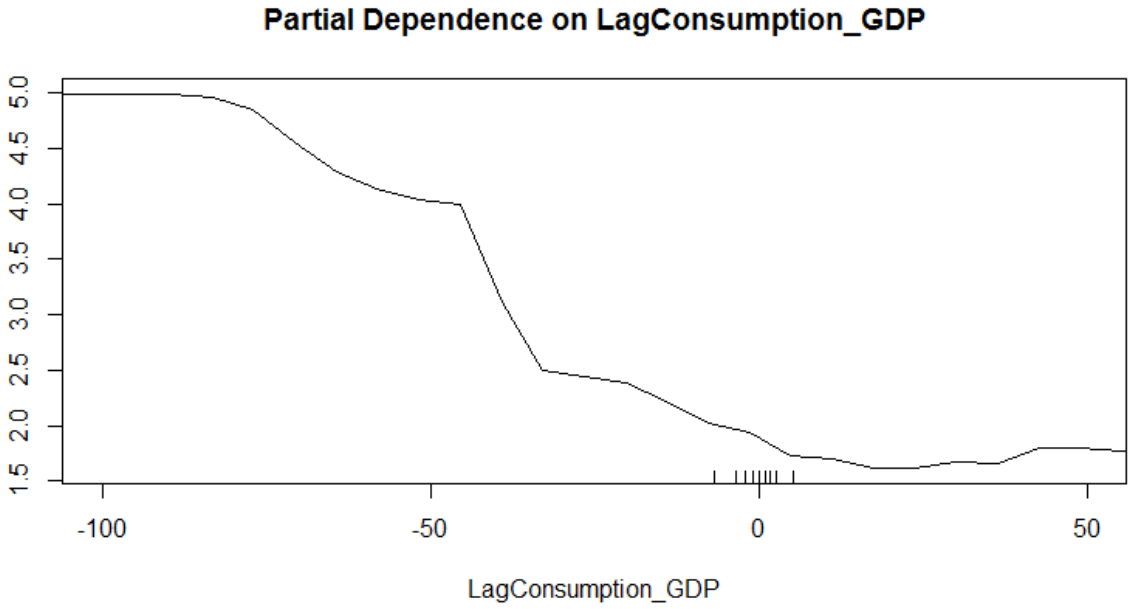


Figure 7. Partial Dependence Plot for Exports

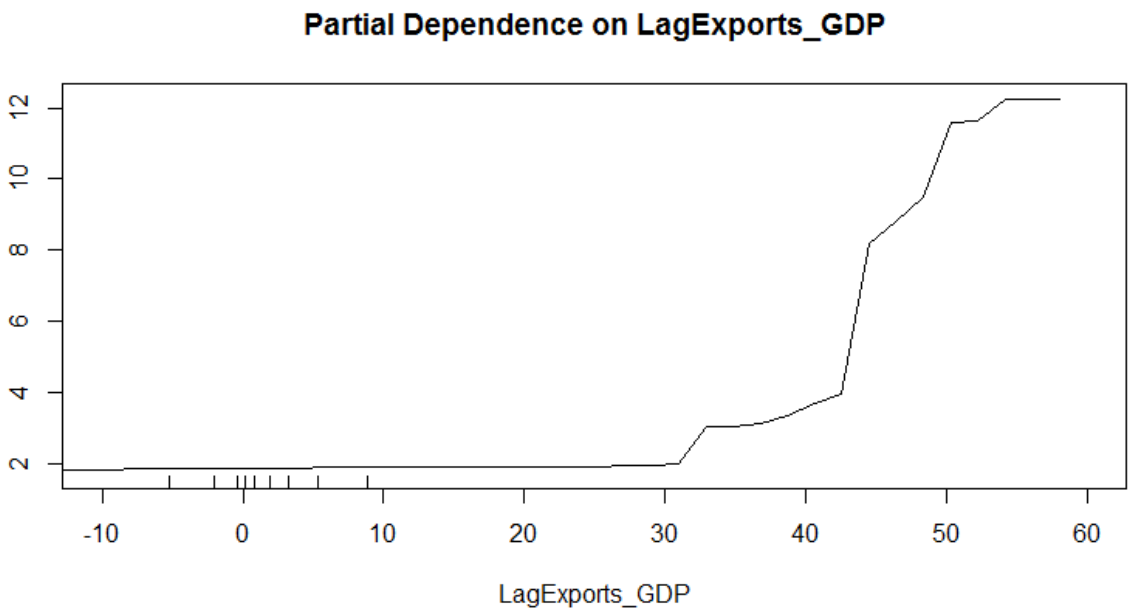


Figure 8. Partial Dependence Plot for Democracy

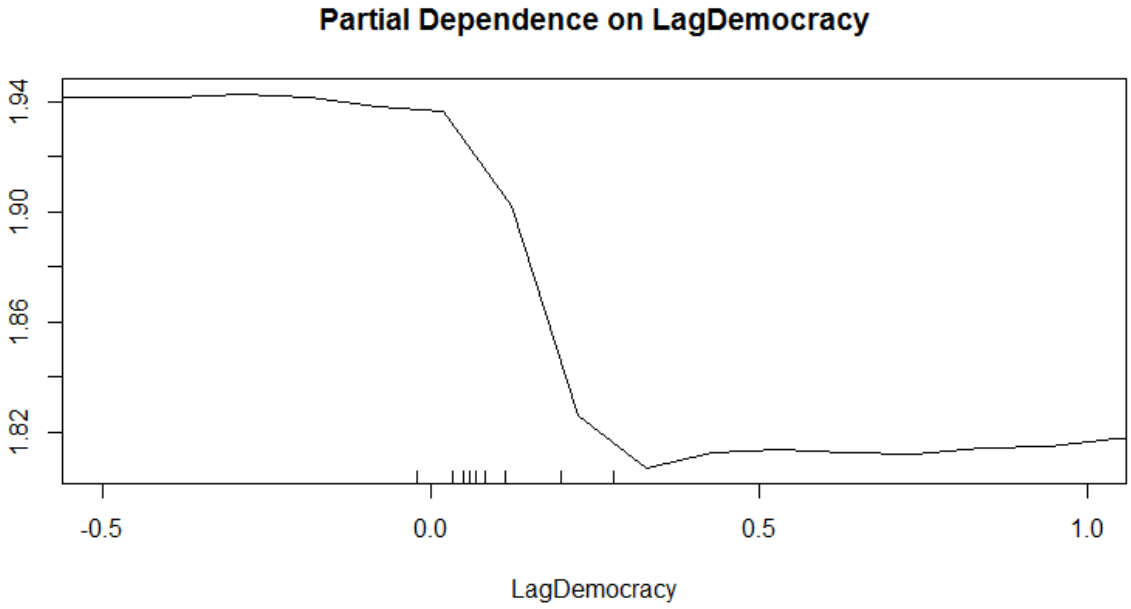


Figure 9. Partial Dependence Plot for Investment

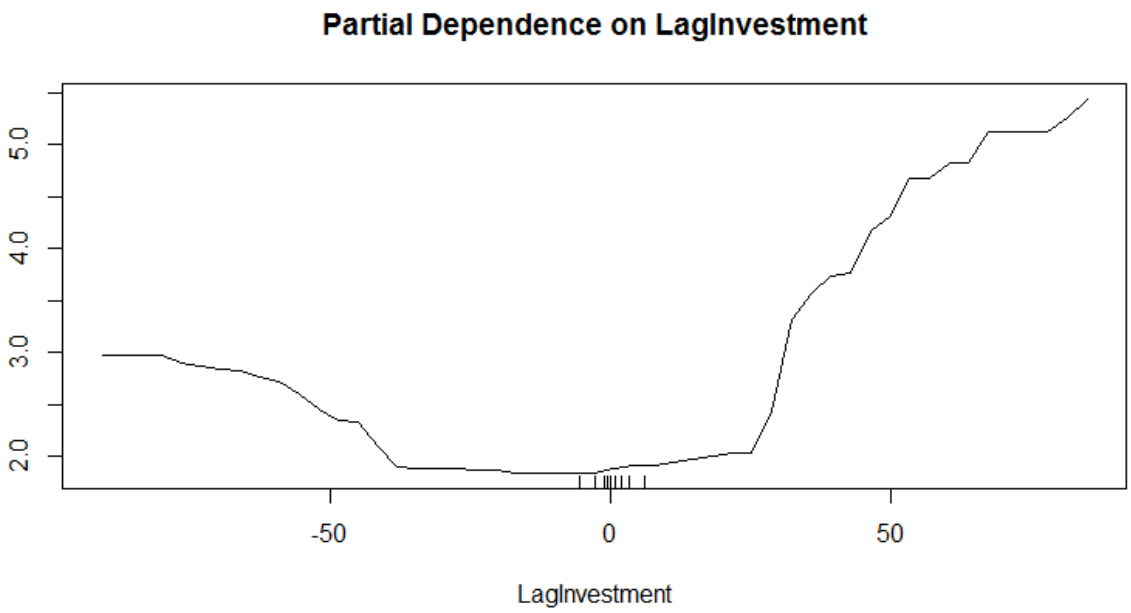


Figure 10. Partial Dependence Plot for Money Growth

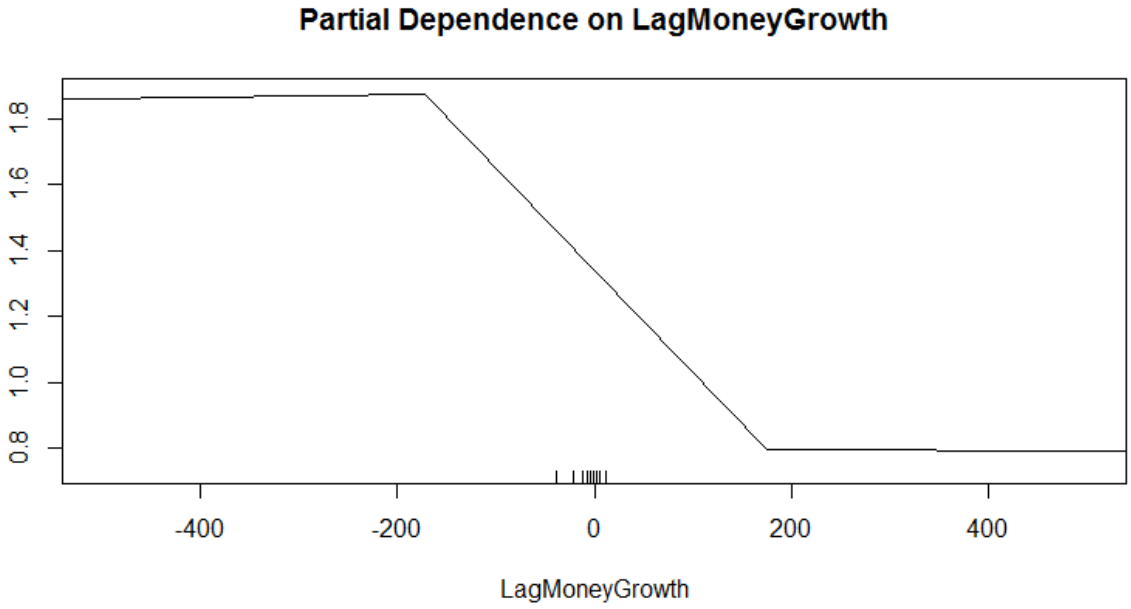


Figure 11. Partial Dependence Plot for Rural Population

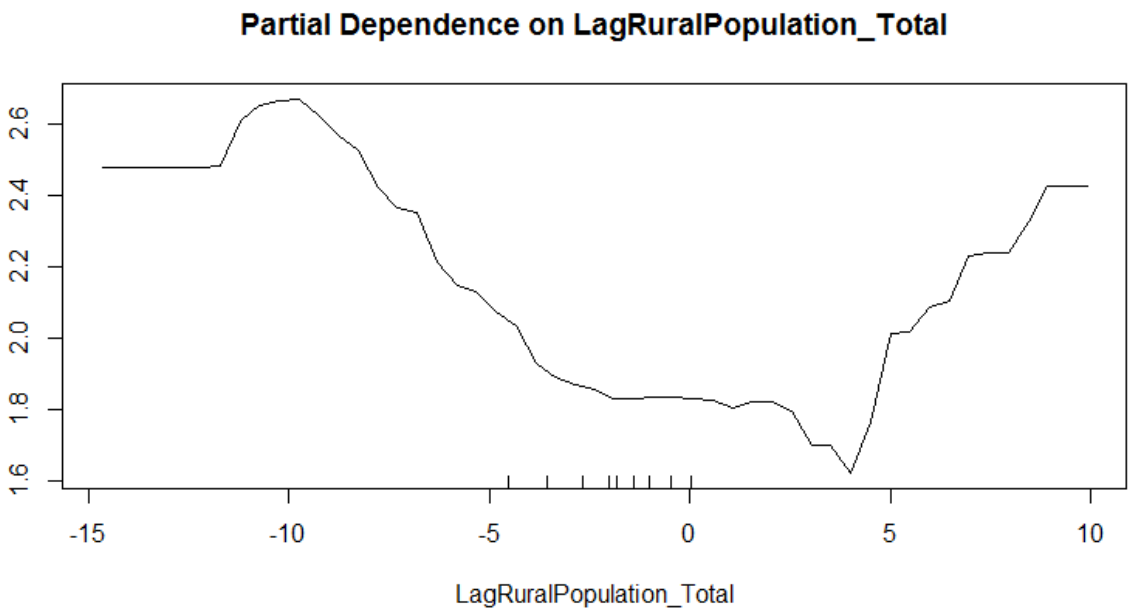
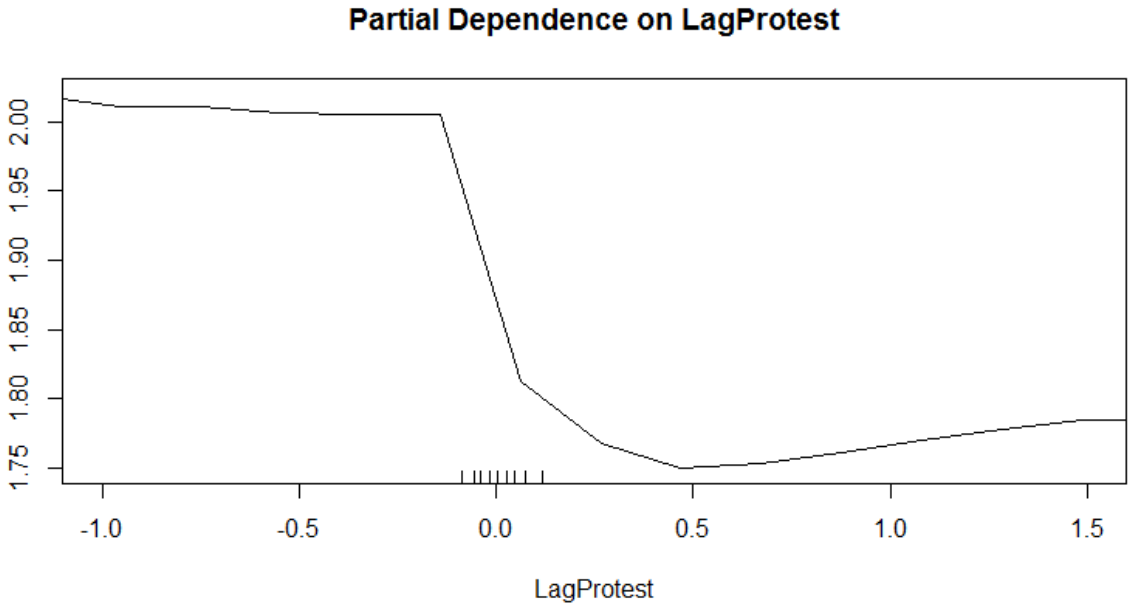


Figure 12. Partial Dependence Plot for Protest



Appendix

Table A1. Exploratory Factor Analysis.

Democracy	5.906
Violence	4.543
Protest	1.301
Regime	1.176
Transparency	0.885
Within	0.821
Credibility	0.465
Factor8	0.465

Observations	10,806
Retained Factors	8
Parameters	212
LR Stat	173,460
P-Value	0.000

	Democ-racy	Violence	Protest	Regime	Trans-parency	Within	Credi-bility	Factor8	Unique-ness
Leg Frac	0.777	0.006	0.003	0.085	-0.084	0.430	0.029	0.025	0.195
Pol Polariz	0.613	0.146	0.010	-0.005	0.166	0.344	-0.048	-0.033	0.454
Exec Tenure	-0.473	0.064	0.028	-0.340	-0.218	0.006	-0.127	-0.093	0.583
Δ Vetoes	0.131	-0.074	-0.012	0.300	0.131	-0.008	0.083	0.073	0.858
Gov Herf	-0.403	0.009	0.055	-0.047	-0.017	-0.660	-0.016	0.010	0.395
Checks	0.820	0.087	0.020	0.010	0.038	0.123	-0.027	-0.045	0.300
Leg Elec Com	0.857	0.120	0.027	-0.060	-0.210	-0.119	-0.039	-0.040	0.185
Exec Elec Com	0.916	0.111	0.008	-0.017	-0.147	-0.142	-0.012	0.016	0.105
Fraud	-0.128	-0.235	-0.032	-0.029	-0.377	0.010	-0.041	0.104	0.772
Polity2	0.860	0.113	0.008	0.050	0.167	-0.018	0.049	0.020	0.213
Reg Dur	0.200	0.241	0.107	-0.274	0.324	-0.015	0.044	-0.054	0.705
Eth Conf	-0.027	-0.625	0.065	-0.025	0.090	0.087	0.228	-0.184	0.502
Civil War	-0.029	-0.445	0.060	0.006	0.023	-0.011	-0.000	0.483	0.563
Assassin	0.059	-0.198	0.173	0.089	-0.039	-0.021	-0.007	0.249	0.855
Strikes	0.089	-0.110	0.407	0.120	-0.003	-0.006	-0.112	0.029	0.786
Gov Crises	0.108	-0.125	0.227	0.301	-0.032	0.055	-0.106	0.065	0.811
Purges	-0.058	-0.038	0.182	0.133	-0.078	0.012	-0.065	0.033	0.933
Riots	0.060	-0.090	0.708	0.033	0.008	-0.020	0.011	0.001	0.485
Demonstr	0.052	-0.088	0.695	-0.007	0.018	-0.024	0.007	0.013	0.506
Cab Changes	0.073	-0.195	0.052	0.579	-0.120	0.056	-0.041	-0.013	0.599
Exec Changes	0.243	-0.061	0.039	0.587	0.049	0.041	-0.028	-0.016	0.586
Leg Elections	0.334	0.099	0.026	0.175	0.027	-0.076	-0.071	-0.080	0.830
Gov Stab	0.027	0.673	-0.036	-0.129	-0.136	0.046	0.372	0.009	0.369
Dem Acct	0.737	0.446	0.000	0.051	0.216	-0.006	0.051	0.020	0.206
Invest Profile	0.290	0.716	-0.019	-0.095	0.012	0.030	0.380	0.000	0.249
Corruption	0.480	0.486	0.014	-0.028	0.462	0.020	-0.145	0.083	0.290
Bur Qual	0.503	0.605	0.045	-0.067	0.357	0.051	0.065	-0.022	0.240
Eth Tensions	0.119	0.743	-0.026	0.010	-0.004	-0.039	-0.148	0.212	0.364
Int Conflict	0.201	0.898	-0.045	0.007	0.018	0.001	-0.077	-0.190	0.109
Ext Conflict	0.322	0.708	0.001	0.049	-0.040	-0.000	-0.062	-0.014	0.387