V-Dem
INSTITUTE

Correlation of Democracy Indicators and Markets Returns

Scott Axelrod
James Leitner

December 2016

Users Working Paper

SERIES 2016:04

THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

Please address comments and/or queries for information to:

V–Dem Institute

Department of Political Science

University of Gothenburg

Sprängkullsgatan 19, PO Box 711

SE 40530 Gothenburg

Sweden

E-mail: contact@v-dem.net

V–Dem Working Papers are available in electronic format at www.v-dem.net.

Correlation of Democracy Indicators and Markets Returns

Scott Axelrod, Falcon Management, axelrod@falconmgt.com
James Leitner, Falcon Management, jleitner@falconmgt.com

Abstract

We perform various experiments correlating past changes of social indicators about a country with future stock market returns for that country. The 169 social indicators we use, which go back as far as the year 1900, are available from the Varieties of Democracy Project. We use two sets of data for country-wide stock market returns: data compiled by Dimson, Marsh, and Staunton covers 17 countries going back to 1900, and data from the MSCI data analytics and index service covering 45 countries going back as far as 1970. We consider five and ten year time windows. This gives us four different "studies": MSCI 10 year, DMS 10 year, MSCI 5 year, and DMS 5 year.

We find the striking result that good changes of the social indicators have a positive mean (averaged over studies) total correlation (correlation of change vectors indexed by country-year pairs) with future stock market returns in 157 out of 158 cases in which the indicator measures something good or bad for society. We obtain a result almost as strong when the correlation is aggregated differently using the separate country and year groupings. We perform statistical hypothesis testing to show that, even though the social indicators are not all independent, these result are exceedingly unlikely to be the result of random (white noise) stock market returns.

We also perform "positive linear regression" of stock market return on all 158 indicators, which means that the sign of the regression coefficient for an indicator is constrained to be positive or negative according to whether a positive change of the indicator is good or bad. The fraction of data explained by positive regression is shown to be extremely statistical significant. We calculate a confidence interval for the percentage of data genuinely explained by regression, not just by fitting to noise. The lower end of the confidence window for the four studies is 11%, 14%, 6%, and 9%.

We include a long appendix on the statistical theory of correlation and (unconstrained) regression. This provides background to the novel applications of hypothesis testing and confidence interval calculation in the body of the paper.

Contents

1 Introduction and Overview

This paper reports on analysis to address the question: How are changes in social indicators for a country correlated with future returns in that country's stock markets? Although there is a huge amount of work addressing stock market predictability, we have not been able to find any systematic studies that address the obvious and important question above. In this paper we perform such a study for the social indicators available in the Varieties of Democracy (V-Dem) database (Coppedge, Gerring, Lindberg, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Knutsen, McMann, Pemstein, Skaaning, Staton, Tzelgov, Wang & Zimmerman 2015a, Coppedge, Gerring, Lindberg, Skanning, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Knutsen, McMann, Paxton, Pemstein, Staton, Zimmerman, Andresson, Mechkova & Mri 2015b, Coppedge, Gerring, Lindberg, Skaaning, Teorell, Andersson, Marquardt, Mechkova, Miri, Pemstein, Pernes, Stepanova, Tzelgov & Wang 2015c). This is a large database which covers many indicators for many countries over many years. In this report we filter to a set of 169 different indicators which have ample data. Since our intent is to focus on robust long term trends, we look at correlation of changes over either five or ten year windows. Our main result is that the correlations between "good" changes of V-Dem indicators and future stock market returns are robustly positive.

We see this work as part of the corporate social responsibility (CSR) and environmental, social, and government (ESG) movements. A preliminary version of the results here was presented in this broader context in (Leitner & Axelrod 2016). Although our hope is that this work might ultimately help sway political leaders and investors that social improvements are good for markets, the focus of this paper is not to push a sociological agenda, but rather to provide a detailed analysis in the framework of statistical hypothesis testing to show that the positive correlation between good changes for society and future market improvements is very consistent and not simply a result of random noise in the data. In this paper we keep focus on overall summary results, avoiding delving into analyzing results for specific countries or specific years.

One mistake that is often made when people declare correlation results interesting is to not account for the fact that that objects of interest are pulled from a large variety of candidates. This is a mistake because one would expect just by random fluctuation that some indicators would have larger correlations. So it would be wrong for us to "cherry-pick" one or two indicators out of 169 which, when looked at in isolation by some standard statistical test, appear to have a one in a hundred chance of being reproducible by random noise.

The worry of studying a single indicator in isolation was, in fact, at the genesis of the current paper. We had noticed V-Dem's new Women's Political Empowerment Index (Sundström, Paxton, Wang & Lindberg December 2015) and we were interested in studying correlation of that index with financial markets. To put this in context, we decided to systematically look at the data for V-Dem indicators and their correlations with the financial markets. We were surprised by the consistency of the result we found: For 157 our of 158 indicators that measured something good or bad for society, for example the Women's Political Empowerment and Political Corruption indicators, good changes of the indicators have a positive mean total correlation with future stock market returns. Additionally, the correlation to markets was very close to zero for the 7 indicators that measure things that are not clearly good or bad, such as the Urbanization indicator. The other 4 of the 169 indicators we looked at were related to measures of GDP whose past changes correlate negatively with future stock market returns. GDP behaves similarly in this respect to this stock markets themselves, which are negatively auto-correlated at the time scales we look at.

In Section 2, we describe the data we use in this paper. The stock data we use comes from two different databases of country-wide total return stock markets indices: "DMS data", studied in (Dimson, Marsh & Staunton 2002), and "MSCI" data, coming from (*MSCI data index and analytics service*). The "DMS" data we use is inflation adjusted, total return country-wide stock index data for seventeen countries from 1900 through 2004. The "MSCI" data is total return country-wide stock data for forty-five countries starting on (or after) 1970 and going through to the present. The four studies come about by choosing (i) either a five or a ten year window to measure change and (ii) either the MSCI or DMS as the source of stock data. The source of data determines the collection of possible country-years pairs. Each V-Dem indicator provides data for a subset of these pairs. Section 3 explains how the 169 indicators we look at are chosen from the full database of V-Dem indicators solely on the basis that the amount of data for the chosen indicators exceeds various threshold tests.

In Section 4, we present in detail results for the total correlation between past indicator change data and future stock return data. Total correlation here refers to the correlation of vectors indexed by country-year pairs for which data exists. The "mean total correlation" mentioned above is the mean over the four different studies of the total correlation. We find that all indicators for which a positive change is a "good" thing have positive mean total correlation, and all but one of those for which positive change is a bad thing have negative mean total correlation. We call consistency with the sign of the correlation with whether an indicator increase is good or bad, the "good-is-good" rule. Table 4 summarizes the amount of data used in each study. Detailed result for all indicators of the total correlation for each study and the mean of these over all studies are given in Appendix A. This is a lot of information, which we provide for readers interested in exploring particular indicators. However, this information is not all independent. Many of the indicators, for example, are defined as combination of other indicators. For this reason, we have chosen, by looking through the V-Dem codebook (Coppedge et al. 2015b), a collection of indicators (also called "codes"), which try to address a "high level" concept by combining together lower level indicators. Table 5 gives these indicators, their brief descriptions, the question they address, and their mean total correlation with stock markets.

In Section 5, we address the question of whether the strong results we obtained by looking at total correlation is dependent upon how we examine the data. We do this by considering different approaches to looking at country-year data when one makes use of the separate country and year groupings. One approach to this kind of study goes under the moniker of "panel data analysis" with "fixed effect models" and "random effect models". See for example (Yeşin 2016), which looks at country-year data for predicting foreign exchange rates (rather than stock market returns). Rather than rely on this technical terminology, we keep focus directly on our main object of study, correlation, and consider the four natural method of defining aggregate correlation that use the grouping of the data by countries and by years. For grouping by countries, one approach is to take the average of the correlation over time obtained for each country separately. Similarly, one can take the average over years of the correlation across countries. We call these correlations the "within-country" and "within-year" correlations. The other two methods of aggregation are obtained by similarly averaging correlation for separate countries or years, but by "tying" the standard deviations together. Tying of standard deviations can avoid the effects of bad standard deviation estimations for individual countries that do not have many years worth of data. Appendix B compares the results of the different methods of aggregation of correlation for all V-Dem indicators. Table 6 compares the counts of codes that violate the "good-is-good" rule when the computation is done by different methods of aggregation. The table show that the exception rate of 0.6% for total correlation is reproduced for within-year correlations with tied standard deviations. It goes up slightly to 1.3% for within-country correlation with tied standard deviation. For the case of untied standard deviation, the exception rates go up a little more, to 4.4% and 8.9% for within-country and within-year correlations, which is not surprising because of the difficulty of estimating standard deviation for groups with small amount of data.

In any case, these exception rates are all far below the rate of 50% one would expect by random chance.

While the "good-is-good" result is quite consistent across indicators, the correlation for individual indicators is modest. In Section 6, we see how additive the "good-is-good" result is, i.e. how much stronger of a result one can get by taking linear combination of indicators. The average of the correlations for all 158 good or bad indicators (times minus one for "bad" indicators) ranges from 8% to 17% in our four different studies, as can be seen on the bottom line of Table 7. The bottom line of that table also shows that a slightly stronger result (ranging from 14% to 23%) is obtained by considering the correlation of future stock return with the "overall good index", which is the average of the (Z-scores of) all the good-or-bad indicators (multiplied by a sign if necessary so that good changes are positive). The square of the correlation with this index is equal to the fraction of the variance (the sum of squares of the difference from the mean) of future stock data which is explained by correlation with the index. This value, known as the R-squared, ranges from 2% to 5% over the four studies.

In Section 6.2, we find the linear combination of good-or-bad indicators whose coefficients have the same sign as the overall good index that maximize the value of R-squared, i.e. we regress future stock return against all good or bad indicators subject to the sign constraint that the regression coefficient for each indicator either vanishes or else has sign which agrees with the sign for which changes of the indicator are socially good. We call such constrained regression *positive linear regression*, or simply *positive regression.* The bottom line of Table 8 show that this optimal R-squared varies from 11% to 18%, and that the number of active (non-zero) coefficients ranges from 16 to 24, depending on the study. This corresponds to an adjusted R-squared, which is a better estimate than R-squared itself of the fraction of variance explained out of sample, ranging from 9% to 17% (see column `real` in Table 15).

In Section 7, we show that all of our results are extremely statistically significant in the sense that they would be very unlikely to be produced by *white noise* – random stock returns generated independently from identical normal distributions. We call the assumption that the stock returns are white noise, our *null hypothesis.* To begin, Table 9 gives the number and percentage of the good or bad V-Dem indicators whose correlation with future stock returns have the wrong sign for the good-is-good rule, broken down by study type and type of correlation aggregation. This is always below 15%, except for a few studies when looking at within group correlation without tied standard deviations. We have already remarked that the weaker results when standard deviations are not tied together is due to the difficulty of estimating standard deviation for groups with small amounts of data. Focusing on the case of total correlation in Figure 2 and Table 10, we see that the probability of random chance generating so few (or fewer) exception is less than a few hundredths of a percent. Note that a "few hundredths of a percent" is much higher than the probability one would obtain if the indicators were independent (e.g. the probability of flipping heads 157 or more out of 158 times is $159 * 2^{-158}$).

Table 12 addresses the question of how many individual indicators have correlations that appear statistically significant at various levels if looked at individually[1]. The table shows, for example, the mean total correlation is statistically significant for 84% of codes at the traditional five-percent significance level, for 73% at the one-percent significance level, and for 67% at a level of a tenth of a percent.

Figure 3 and Tables 13 do the same thing for the value of total correlation of future stock returns with the overall good index that Figure 2 and Tables 10 did for the number of exception to the good-is-good rule. They show that it is exceedingly improbable for this total correlation to be produced by white noise. Table 14 shows that the P-values for the within-country and within-year (with tied standard deviations) versions of this correlation are also all minuscule, with the exception of the DMS 5 year study where the P-value of 4% is merely small. As usual, the within-group correlations with untied standard deviations are not as significant.

In Section 7.3, we consider the statistical significance of the adjusted R-squared for positive regression, which we mentioned above ranges from 9% to 17%. In the unconstrained case, it is natural to look at adjusted R-squared for significance testing against the null hypothesis since it has mean zero, as opposed to the mean of unadjusted R-squared which increases with the number of codes being regressed against. In the constrained case, there is no simple solution for the distribution of R-squared, although (Grömping 2010) gives a nice overview of the theory that is known and a computational package for computing the distribution. No one seems to have studied the appropriate adjustment to R-squared in the case of constrained regression. The adjustment we use depends on the number of active codes (rather than all code regressed against). This seems to be the correct adjustment to make because the mean value of this adjusted R-squared vanishes, to within the precision of simulation we perform which use 10,000 randomly generated stock return series for each of our four studies. As can be seen in Figure 4 and Table 15, the distribution of adjusted R-squared when the null hypothesis is true is strongly peaked around zero and has standard deviation of about 1%. So the observed adjusted R-squared values are at least nine standard deviations above the mean, extremely statistically significant!

We mentioned above that adjusted R-squared is “a better estimate than R-squared itself of the fraction of variance explained out of sample”. The issue of finding the best

[1] Here is a lightning summary of statistical significance and hypothesis testing, which are discussed in more detail in Appendix E.6: The probability that random chance would generate a result (e.g. the total correlation with a particular indicator) at least as extreme as that actually observed is called the P-value. The lower the P-value, the less likely it is that the phenomenon occurred by chance. The P-value is sometimes called the statistical significance level. Formally, one says that one can reject the null hypothesis at a given significance level p if the P-value is below p.

estimate of R-squared for an underlying population, based only on observed sampling of data, has a long history and is the subject of active research in the unconstrained case, see for example (Yin & Fan 2001, Salh 2015, Nimon, Zientek & Thompson 2015). In Appendix E, we give a thorough accounting of adjusted R-squared, including an explanation of why it is an approximately unbiased estimate of the *population R-squared* – the percentage of data genuinely explained by regression, not just by fitting to noise. We also explain how to calculate *confidence intervals* giving a range of likely values for population R-squared for ordinary linear regression.

In Section 8, we calculate confidence intervals for population R-squared for our positive regression problem. In summary, because we have enough data, the widths of the confidence intervals are fairly narrow. So the estimate that on the order of 10% of future stock returns is explained by positive regression on past indicators is likely to hold up if the future is like the past.

In Section 9, we give some concluding remarks discussing limitations of, and possible future directions for, this work.

In the body of the paper we attempt to provide as much detail and require as little background as possible. In addition, Appendix E is essentially a little monograph on the mathematics of the probability and statistics of correlation and regression. We attempt to achieve the contradictory goals of being self-contained, brief, cogent, thorough, and precise, while providing full mathematical details including some derivations not seen elsewhere. We assume, to varying degrees throughout the appendix, that the reader is comfortable with some basic mathematical notation and concepts and has some knowledge of multivariable calculus. Some terminology from linear algebra that is used toward the end of the Appendix E is summarized in Appendix D.

All simulation results and figures in this paper were produced by custom code written in Matlab (MathWorks Inc. 2015), with the exception of Figure 5, which was produced using Mathematica (Wolfram Research Inc. 2014).

As a convenience for readers of the PDF version of this paper, section, footnote, figure, table, and equation numbers should be clickable hyper-references in most viewers.

2 Data

Data for all results in the paper comes from three source: V-Dem, DMS, and MSCI. We describe each of these data sources in turn.

2.1 V-Dem Data

The Varieties of Democracy (V-Dem) institute is "a team of fifteen social scientists on three continents" who "work with more than 2,500 country experts and a truly global international advisory board" with the aim of producing and studying better indicators of democracy. A wealth of information about the project is available at their website: `https://www.v-dem.net/en`.

In this paper, we use the V-Dem data release version 5 published in January of 2016, which can be freely downloaded from the V-Dem web site. Specifically, we use the `Country-Year-V-Dem_other` archive and a few other descriptive documents, which include:

- A large database (Coppedge et al. 2015a) of 585 indicators (described at the V-Dem web site as "over 350 V-Dem indicators and indices and over 300 other indicators from other data sources"). For each indicator and each of 173 different countries, the database contains an annual data series for a (country-dependent) subset of the years 1900 to 2012. Each indicator has a code name. For example the "Women political empowerment index" has code `v2x_gender`. We will often use the term "V-Dem code" synonymously with "V-Dem indicator".

- A "codebook" (Coppedge et al. 2015b) describing each of the V-Dem codes and also giving background materials and an overview of the structure of all the indicators. Note that the indicators are not all independent, for example many "higher level" indicators are built from more basic indicators. Much of this is summarized in the appendix "Structure of Aggregations - All Indices and Indicators".

- The document (Coppedge et al. 2015c) describing the methodology used to collect the data, which includes sections on the "Conceptual Scheme", "Data Collection", and "Measurement".

- The document (Coppedge, Gerring, Lindberg, Skaaning, Teorell & Ciobanu 2015) which "lists (a) every country in the envisioned V-Dem database, (b) the identities of each polity that comprises a country's history through the twentieth and twenty-first centuries (e.g., Russia-USSR); (c) the years for which we have collected data or plan to collect data (in parentheses next to the entry); and (d) the borders of each country (wherever this might be unclear)."

- The document (Coppedge, Gerring, Lindberg, Skaaning, Teorell, Andersson, Mechkova, Pernes & Stepanova 2015) which “contains an overview of the organization and management of the Varieties of Democracy (V-Dem). It provides information about the team working on the project, the V-Dem infrastructure and the website, as well as the outreach and policy-oriented activities, our funding, and the progress of data collection so far. It also presents the plans for sustainability of our activities and benchmarks by which the impact of the V-Dem project on the development community (encompassing both policymakers and academics) can be monitored in the coming years.”
- The document (Coppedge, Gerring, Lindberg, Skaaning & Teorell 2015) which provides a “critical review of the field of democracy indices” and “discusses in general terms how the Varieties of Democracy (V-Dem) project differs from extant indices and how the novel approach taken by V-Dem might assist the work of activists, professionals, and scholars.”

2.2 DMS Data

The book *Triumph of the Optimists* (Dimson, Marsh & Staunton 2002) presents a “comprehensive and consistent analysis of investment returns for equities, bonds, bills, currencies and inflation, spanning sixteen countries, from the end of the nineteenth century to the beginning of the twenty-first.” The “DMS data” associated with the book is proprietary, but available for sale. For our study, we use an update of the original data set, which has yearly data series for the years 1900 through 2004 and covers the following seventeen countries, which include Norway in addition to the sixteen countries studied in *Triumph of the Optimists*:

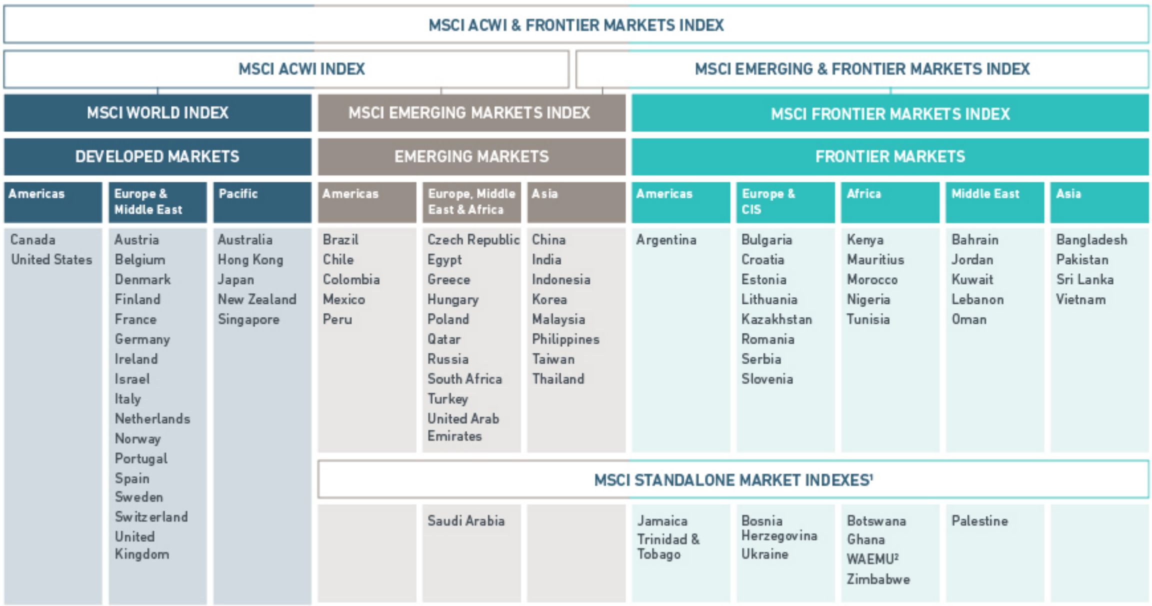
> Australia, Belgium, Canada, Denmark, France, Germany, Ireland, Italy, Japan, Netherlands, Norway, South Africa, Spain, Sweden, Switzerland, United Kingdom, United States

We only use the real equity total return series. This is a carefully constructed stock index for each country, which is inflation adjusted (real) and includes the (total return) effect of reinvesting dividends.

The data set we use was purchased from Ibbotson Associates. More recent updates to the DMS data are available from Morningstar. Rather than use these updates, we consider studies with the DMS data to be historical “twentieth century” (plus a few years) studies. In the next subsection, we describe the data set we use for studies with many more countries and more recent years.

2.3 MSCI Data

The other source of stock data we use is the MSCI data analytics and index service. There are 78 countries listed as having "market cap indices" in Figure 1, downloaded from the MSCI web page `https://www.msci.com/en/market-cap-weighted-indexes`.



MSCI ACWI & FRONTIER MARKETS INDEX

MSCI ACWI INDEX

MSCI EMERGING & FRONTIER MARKETS INDEX

MSCI WORLD INDEX

MSCI EMERGING MARKETS INDEX

MSCI FRONTIER MARKETS INDEX

DEVELOPED MARKETS

EMERGING MARKETS

FRONTIER MARKETS

Americas	Europe & Middle East	Pacific	Americas	Europe, Middle East & Africa	Asia	Americas	Europe & CIS	Africa	Middle East	Asia
Canada United States	Austria Belgium Denmark Finland France Germany Ireland Israel Italy Netherlands Norway Portugal Spain Sweden Switzerland United Kingdom	Australia Hong Kong Japan New Zealand Singapore	Brazil Chile Colombia Mexico Peru	Czech Republic Egypt Greece Hungary Poland Qatar Russia South Africa Turkey United Arab Emirates	China India Indonesia Korea Malaysia Philippines Taiwan Thailand	Argentina	Bulgaria Croatia Estonia Lithuania Kazakhstan Romania Serbia Slovenia	Kenya Mauritius Morocco Nigeria Tunisia	Bahrain Jordan Kuwait Lebanon Oman	Bangladesh Pakistan Sri Lanka Vietnam

MSCI STANDALONE MARKET INDEXES[1]

				Saudi Arabia		Jamaica Trinidad & Tobago	Bosnia Herzegovina Ukraine	Botswana Ghana WAEMU[2] Zimbabwe	Palestine	

Figure 1: MSCI market cap indices

We found forty-five countries which have country-wide MSCI total return stock indices starting before the year 2000 and which also have VDEM codes[2][3]. We downloaded this data using the financial data vendor Bloomberg L.P. The start date for the data depends on the country. MSCI data for all the countries we consider is available up to the present; however, since the V-Dem database ends in 2012, we only use MSCI data up until that year[4]. Table 1 is a list of all countries we study, together with the starting year of the MSCI data and the Bloomberg "ticker" used to download the MSCI data. Table 2 is a list of the different starting years for country return data and the number of countries with that starting year.

The earliest start year for MSCI returns is 1970. The countries starting that year are Austria and all of the DMS countries except for Ireland (which starts in 1988, along with

[2] In the presentation (Leitner & Axelrod 2016), we studied all countries in the VDEM data base which were also on the list of Bloomberg tickers from MSCI found at `https://www.msci.com/zh/bloomberg-tickers-end-of-day#bbgeodcountryusd`. That study included all the countries included here plus Qatar, which has MSCI data starting in 2006. The difference between inclusion of Qatar or not is on the order of rounding error, since the only effect of including Qatar is to include two more data points in the study of MSCI data with five year windows.

[3] The MSCI data is not adjusted for inflation, although the DMS data we use is.

[4] This choice was made so they we could treat stock and V-Dem data on an equal footing. However, when we only consider the future variable to be stock returns, using MSCI data up until 2015 would allow us to gain three extra data points for most choices of country and V-Dem code.

country	**history start date**	**Bloomberg ticker**
Argentina	1998-12-31	GDUESAG Index
Australia	1969-12-31	GDDUAS Index
Austria	1969-12-31	GDDUAT Index
Belgium	1969-12-31	GDDUBE Index
Brazil	1998-12-31	GDUEBRAF Index
Canada	1969-12-31	GDDUCA Index
Chile	1998-12-31	GDUESCH Index
China	1992-12-31	GDUETCF Index
Colombia	1998-12-31	GDUESCO Index
Czech Republic	1998-12-31	GDUESCZ Index
Denmark	1969-12-31	GDDUDE Index
Egypt	1994-12-30	GDUESEG Index
Finland	1987-12-31	GDDUFI Index
France	1969-12-31	GDDUFR Index
Germany	1969-12-31	GDDUGR Index
Greece	1987-12-31	GDUESGE Index
Hungary	1998-12-31	GDUESHG Index
India	1992-12-31	GDUESIA Index
Indonesia	1998-12-31	GDUESINF Index
Ireland	1987-12-31	GDDUIE Index
Israel	1998-12-31	GDUESIS Index
Italy	1969-12-31	GDDUIT Index
Japan	1969-12-31	GDDUJN Index
Jordan	1998-12-31	GDUESJO Index
Malaysia	1998-12-31	GDDUMAF Index
Mexico	1987-12-31	GDUETMXF Index
Morocco	1998-12-31	GDUESMO Index
Netherlands	1969-12-31	GDDUNE Index
New Zealand	1987-12-31	GDDUNZ Index
Norway	1969-12-31	GDDUNO Index
Pakistan	1998-12-31	GDUESPF Index
Peru	1998-12-31	GDUESPR Index
Philippines	1998-12-31	GDUESPHF Index
Poland	1998-12-31	GDUESPO Index
Portugal	1987-12-31	GDDUPT Index
Russia	1998-12-31	GDUESRUS Index
South Africa	1998-12-31	GDUESSA Index
Spain	1969-12-31	GDDUSP Index
Sweden	1969-12-31	GDDUSW Index
Switzerland	1969-12-31	GDDUSZ Index
Taiwan	1998-12-31	GDUESTW Index
Thailand	1998-12-31	GDUESTHF Index
Turkey	1998-12-31	GDUESTK Index
United Kingdom	1969-12-31	GDDUUK Index
United States	1969-12-31	GDDUUS Index

Table 1: MSCI countries with start dates and Bloomberg tickers.
DMS countries are in bold.

start year of returns	**number of countries**
1970	16
1988	6
1993	2
1995	1
1999	20

Table 2: Number of countries for various start years of MSCI return data.

Finland, Greece, Mexico, New Zealand, and Portugal) and South Africa (which starts in 1999, along with Argentina, Brazil, Chile, Colombia, Czech Republic, Hungary, Indonesia, Israel, Jordan, Malaysia, Morocco, Pakistan, Peru, Philippines, Poland, Russia, Taiwan, Thailand, and Turkey). In addition there is one country (Egypt) with returns starting in 1995 and two countries (China and India) with returns starting in 1993.

3 Pruning to a List of 169 V-Dem Codes with Enough Data

We shall be looking at correlations between past changes on either a five or ten year time scale and future changes on the same time scale. The changes we look at are either the difference of a V-Dem indicator or a continuously compounded (i.e. logarithmic) annualized stock market return over the time window of interest. The stock market returns come from either the DMS or MSCI database. The four studies we look at are described in Table 3.

Study 1	MSCI	ten year window
Study 2	DMS	ten year window
Study 3	MSCI	five year window
Study 4	DMS	five year window

Table 3: Four studies examined. The studies with MSCI data cover the forty-five countries in Table 1 with returns starting between 1970 and 1999 (and with an average start year of 1987). The MSCI studies use V-Dem data back to as early as 1960 to provide a ten year window before the start of stock data. The studies with DMS data cover the seventeen DMS countries and the years 1900 through 2004.

The DMS data is complete (no missing data) for all seventeen DMS countries and for the years 1900 through 2004. The MSCI data is also complete for all the MSCI countries from their starting year through 2012. However, many of the yearly time series associated with a choice of V-Dem indicator and country either start late, end early, or have several years in which data is missing. We calculate correlations even when data is missing by simply ignoring the missing years. But for the correlation results to be meaningful, we don't want there to be too many years with missing data or to have data series that are constant or near constant. We apply three filters which together reduce the original set of 585 V-Dem indicators to a set of 169 kept indicators which we will focus on.

We add the code "**stock**" to the list of 169 kept V-Dem indicators; so the full set of codes we look at has 170 codes.

3.1 Filter Step 1: Reduce to 251 V-Dem Codes that Have Enough Data for DMS Countries and Years

Our first step is to filter to a subset of the V-Dem codes which don't have too much missing data for any of the DMS countries.

For each of the countries in the V-Dem database, there is specified a starting and ending year of data and also a starting and ending years for a gap in the middle of data. All of the individual indicators have data for a subset (sometimes a small subset) of the years between the start and end years, excluding the gap year.

The structure of the V-Dem database encodes the fact that data (for the years 1900-2012) is missing for all codes for the following DMS countries and years:

- the start date for V-Dem data for Australia is 1901;
- the start date for V-Dem data for Ireland is 1919; and
- V-Dem data for Germany is missing for the years 1945 through 1949.

This determines a list of "possible ten year past change data years" for each country. No indicator can have ten year past change data except for the years in this list, although individual indicators may be missing such data on many other years. The possible ten year past change data years for all DMS countries are 1910-2012, with the following exceptions: Australia and Ireland have earliest start years 1911 and 1929, respectively, and Germany has a gap from 1945 to 1959 (since we require that there be no missing data between the start and end year of the ten-year window).

For filter step 1, we use the two criteria below as a simple general rule that is flexible to the fact V-Dem data in the early years is more gappy than data for the later years. This keeps 251 codes, i.e. about half of the original 585 codes. A V-Dem code is kept if:

- all DMS countries have ten year past change data for that code for at least 50% of the possible data years in the range 1919 to 2004; and
- all DMS countries have ten year past change data for that code for at least 85% of possible DMS data years in the range 1949 to 2004.

3.2 Filter Step 2: Reduce to 218 V-Dem Codes that have Enough Data for MSCI Countries and Years

In filter step 2, we keep 218 codes (out of the 251 codes that survived filter step 1) which satisfy the following criteria list. A V-Dem code is kept if:

- for all MSCI countries, data for the code ends on or after 2010;

- for all MSCI countries, data for the code starts on or before 2006; and
- for all MSCI countries, the fraction of years (after the first data year[5]) with missing data for the code is less than one third.

3.3 Filter Step 3: Reduce to 169 V-Dem Codes with Enough Unique Values

For each study, country-year pair in the study, and V-Dem code, we calculate a value,

$$xDiff = xDiff(\mathit{study}, \mathit{country}, \mathit{year}, \mathit{code}),$$

which is the difference of the V-Dem indicator for the country across a window ending in the specified year and of size equal to the window for the study. We are not interested in considering indicators that remain constant. To measure this, we define

$$NUV = NUV(\mathit{study}, \mathit{country}, \mathit{code}),$$

the number of different (i.e. unique) value that $xDiff$ takes across all years of the study for the given country, and code[6]. We would like to focus on indicators that have the ability to go up, go down, and stay more or less constant.

In filter step 3, we keep 169 codes (out of the 218 codes that survived filter step 2) which satisfy the following criteria list. A V-Dem code is kept if the following is true for all four studies in Table 3:

- the mean of $NUV(\mathit{study}, \mathit{country}, \mathit{code})$ over all countries in the study should be at least 2; and
- at least 2/3 of the countries in the study have $NUV(\mathit{study}, \mathit{country}, \mathit{code}) \geq 3$.

4 Basic Correlation Results

In this section we look at correlation of changes in past and future codes (in the set of 169 kept V-Dem codes plus the code **stock**) for the four studies in Table 3. In this paper we report results when the future code is **stock**, but will keep the discussion general when we can. For each separate country in each study, we can calculate a simple correlation of time series for the years in which we have past and future data (change/return data over the window of the study). The set of all these correlation is a lot of information. We

[5] Since we look at V-Dem data in a window of up to ten years in the past of stock return data and the MSCI annual return data starts in 1970, we only look here at V-Dem data in years on or after 1960. The "first data year" above refers to the first year data exists for the given code and country.

[6] This doesn't include that value NaN (not a number) indicating missing data.

would like to obtain some sort of comprehensible aggregate correlation which combines the data for separate countries and studies together. A primary interest for us is whether past changes of V-Dem codes that measure “good” things have positive correlation with future stock market returns. Although we don’t have an objective reference for what a “good change” is, we feel it is pretty clear for most codes what would generally be considered good.

In this section, we calculate the *total* correlation with future stock returns for each study and past code by combining data for all countries and years in which it is available. That is, we calculate the correlation of the vectors X_{cy} and Y_{cy}, where cy is an index that runs over the country-year pairs for which there is both past and future change data. Table 4 gives a feel for the amount of data used in each study.

	type	**win**	**nCountries**	**minY**	**maxY**	**nYears**	**nValsMin**	**nValsMean**	**nValsMax**
1	MSCI	10	45	1972	2003	32	398 (8.8)	736 (16.4)	739 (16.4)
2	DMS	10	17	1911	1995	85	1313 (77.2)	1395 (82.0)	1445 (85.0)
3	MSCI	5	45	1970	2008	39	726 (16.1)	994 (22.1)	996 (22.1)
4	DMS	5	17	1906	2000	95	1490 (87.6)	1565 (92.0)	1615 (95.0)

Table 4: Amount of data used for each study to calculate total correlation of changes in a past code with future stock returns. The middle columns give the number of countries and the range and number of years for which there is some data. The last columns give the minimum, mean, and maximum statistics (over all past codes) of the number of data points, i.e. the number of country-year pairs. Values in parentheses are these statistics divided by the number of countries in the study.

In Appendix A, we present a table giving the aggregate correlation with future stock return for all codes and all studies. The column `meanCorr` gives the average correlation over all four studies. Rows of the table are sorted in order of decreasing `meanCorr`.

A glance at the table reveals a remarkable consistency: Codes that measure “good” things have positive correlation and those that measure “bad” things have negative correlation with future stock market returns. We will make this statement more precise in a moment. But first we must point out that not every line of this large table is independent. For example, there are several “high-level” V-Dem codes that are built from other V-Dem codes. Table 5 gives a subset of the result of the big table in the Appendix for a collection of these “high-level” codes.

4.1 Summary of Results: What’s Good for Society is Good for Markets

We are now ready to examine our basic question of whether past changes of codes that measure “good” (or “bad”) things have positive (or negative) correlation with future

R	code	des	question	meanCorr
5	`v2x_freexp_thick`	Expanded freedom of expression index	To what extent does government respect press & media freedom, the freedom of ordinary people to discuss political matters at home and in the public sphere, as well as the freedom of academic and cultural expression?	0.19
6	`e_rol_free`	Civil liberties and rule of law index	To what extent are civil liberties protected, and rule of law observed in a country?	0.18
8	`v2xcs_ccsi`	Core civil society index	How robust is civil society?	0.18
13	`v2x_liberal`	Liberal component index	To what extent is the liberal principle of democracy achieved?	0.17
18	`v2x_partip`	Participatory component index	To what extent is the participatory principle achieved?	0.17
21	`v2xcl_rol`	Equality before the law and individual liberty index	To what extent are laws transparent and rigorously enforced and public administration impartial, and to what extent do citizens enjoy access to justice, secure property rights, freedom from forced labor, freedom of movement, physical integrity rights, and freedom of religion?	0.17
40	`v2x_libdem`	Liberal democracy index	To what extent is the ideal of liberal democracy achieved?	0.16
48	`v2x_gender`	Women political empowerment index	How politically empowered are women?	0.15
49	`v2x_jucon`	Judicial constraints on the executive index	To what extent does the executive respect the constitution and comply with court rulings, and to what extent is the judiciary able to act in an independent fashion?	0.15
55	`v2x_egaldem`	Egalitarian democracy index	To what extent is the ideal of egalitarian democracy achieved?	0.14
60	`v2x_polyarchy`	Electoral democracy index	To what extent is the ideal of electoral democracy in its fullest sense achieved?	0.14
79	`v2x_EDcomp_thick`	Electoral component index	To what extent is the electoral principle of democracy achieved?	0.12
165	`v2x_corr`	Political corruption	How pervasive is political corruption?	-0.11

.

Table 5: High level V-Dem codes. The first column gives the row of the code in Table of Appendix A. The last column give `meanCorr`, the average over studies of the correlation (of vectors indexed by country-year pairs) between the past change of the code and future stock return. The `question` column is the question associated with the code as given in the V-Dem codebook.

stock market returns. A detailed look at the Table in Appendix A reveal that the answer is an emphatic yes. This means that, for the data we look at, good or bad changes in social indicators tend to be followed by good or bad market returns[7].

We must admit that, lacking an objective reference, we used our own judgment to decide that a positive change is good for 144 indicators, bad for 14 indicators, and not clearly good or bad for 7 indicators. In addition, we note that 4 indicators are related to measures of GDP (gross domestic product). We feel that a poll would confirm our judgments, but we invite readers to download the V-Dem codebook (Coppedge et al. 2015b) and judge for themselves.

Our main result is that every past code $\mathbf{x}$ in the table satisfies one of the following conditions.

- When $\mathbf{x} = \mathbf{stock}$, $meanCorr = -0.20$.
  This mean reversion effect is the strongest mean correlation in the table. It says that periods when stocks outperform their mean tend to be followed by period when they under-perform, and vice-versa.

- When $\mathbf{x}$ is about GDP, $meanCorr < 0$.
  Viewing GDP (gross domestic product) as closely coupled with stock market, this can be seen as a weaker version of the above mean reversion effect.

- When positive change of $\mathbf{x}$ is "good", $meanCorr > 0$.

- With one exception, when positive change of $\mathbf{x}$ is "bad", $meanCorr < 0$.
  The exception is the code for "Institutionalized autocracy"
  (`e_autoc`, $R = 142$), which has $meanCorr = 0.02$.

- When the meaning of change of $\mathbf{x}$ is "questionable", $|meanCorr| < 0.025$,
  with the one exception for "HOS proposes legislation in practice"
  (`v2exdfpphs`, $R = 77$), which has $meanCorr = 0.12$.

5 Within-Group Correlation

In the previous section, we computed, for each study and past code $\mathbf{x}$, the *total* correlation of the vector of past changes (over a past window of five or ten years, depending on the study) with the vector of future changes for the code $\mathbf{y} = \mathbf{stock}$. These vectors each have one component for every country-year pair for which there is both past and future change data. In this section, we compare these total correlation results with four methods of

[7] A positive (resp. negative) correlation of two variables usually implies the signs of the deviation from the mean of the variables tend to be the same (resp. the opposite). So when we talk about good and bad changes above, we are referring to the change relative to the mean change.

computing an aggregate "within-group" correlation which use the fact that the country-year pairs natural divide up into groups by either country or year. In Appendix E.7, we discuss within group correlation in the abstract. In Section 5.1, we give the formulas used as applied to our application. In Section 5.2, we give details of the results.

The upshot is that the different methods of aggregation give similar results. This is not surprising, since the different methods are essentially equivalent up to nuances of how data is weighted. One reason we include these results here is for completeness and to encourage the reader to think about different ways to "play" with the data. The other reason is to respond to the question the reader might possibly have in mind: *How much does the strength of the claim "What's Good For Society is Good For Markets" depend on how the data was processed*. The answer is that the strikingly strong result of the previous section is fairly robust to changes (at least of the type we consider in this section), although it does look a little weaker.

To be specific, we say that a correlation for a code is "consistent with good-is-good" if it is positive, negative, or about zero (i.e. has absolute value smaller than 0.025) according to whether the code is deemed "good", "bad", or "unclear" (respectively). In the last section, we saw that the mean of total correlation was "consistent with good-is-good" in all but 2 out of 165 cases. In this section, we see that for intra-group correlation the number of exception goes as high as 17 out of 165. So the percentage of exceptions goes up from about 1% to about 10%. This is still a very low percentage.

5.1 Definitions of Different Types of Correlation

In this subsection we give explicit formulas for total correlation and two different types of within-group correlation (which we will apply to grouping by country and by year). We include this detail because we don't know of a standard reference with our exact formulation (although books on analysis of variance have closely related descriptions) and we want this paper to be self-contained to those mathematically inclined readers that are interesting in the precise definitions. In this paper we focus on correlation between past changes of a V-Dem code $\mathbf{x}$ and future changes of stock returns, but to be general we can consider future changes of any code $\mathbf{y}$, which includes the code **stock** as a special case.

For each study $\mathbf{s}$, let Countries($\mathbf{s}$) and Years($\mathbf{s}$) be the set of all countries and all data years of the study. Given codes $\mathbf{x}$ and $\mathbf{y}$, let $\mathcal{I}(\mathbf{x}, \mathbf{y}, \mathbf{s})$ be the set of pairs cy of a country c (in Countries($\mathbf{s}$)) and year y (in Years($\mathbf{s}$)) for which there is both past change data for $\mathbf{x}$ and future change data for $\mathbf{y}$ available for the study $\mathbf{s}$. Also let $\mathcal{I}_c(\mathbf{x}, \mathbf{y}, \mathbf{s})$ ($\mathcal{I}_y(\mathbf{x}, \mathbf{y}, \mathbf{s})$) be the set of years (respectively countries) for which there is data for the country c (year y). Identifying a year y in $\mathcal{I}_c(\mathbf{x}, \mathbf{y}, \mathbf{s})$ with the pair cy, $\mathcal{I}(\mathbf{x}, \mathbf{y}, \mathbf{s})$ equals the disjoint union of the $\mathcal{I}_c(\mathbf{x}, \mathbf{y}, \mathbf{s})$ over all c in Countries($\mathbf{s}$). Similarly, $\mathcal{I}(\mathbf{x}, \mathbf{y}, \mathbf{s})$ is equal to the disjoint union of the $\mathcal{I}_y(\mathbf{x},\mathbf{y},\mathbf{s})$ over all y in Years($\mathbf{s}$). Let $\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s})$, $\mathcal{N}_c(\mathbf{x},\mathbf{y},\mathbf{s})$, and $\mathcal{N}_y(\mathbf{x},\mathbf{y},\mathbf{s})$ be the number of elements in $\mathcal{I}(\mathbf{x},\mathbf{y},\mathbf{s})$, $\mathcal{I}_c(\mathbf{x},\mathbf{y},\mathbf{s})$, and $\mathcal{I}_y(\mathbf{x},\mathbf{y},\mathbf{s})$. So

$$\begin{aligned}\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s}) &= \sum_{c\in\text{Countries}(\mathbf{s})} \mathcal{N}_c(\mathbf{x},\mathbf{y},\mathbf{s}) \\ &= \sum_{y\in\text{Years}(\mathbf{s})} \mathcal{N}_y(\mathbf{x},\mathbf{y},\mathbf{s}). \qquad (1)\end{aligned}$$

$\mathcal{I}(\mathbf{x},\mathbf{y},\mathbf{s})$ is the set of indices for the past and future change vectors X_{cy} and Y_{cy}. The total correlation of these vectors is

$$\text{corr}(X,Y) = \frac{1}{\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{cy\in\mathcal{I}(\mathbf{x},\mathbf{y},\mathbf{s})} \frac{(X_{cy}-\bar{X})(Y_{cy}-\bar{Y})}{\text{std}(X)\ \text{std}(Y)}. \qquad (2)$$

Here $\bar{X}$ and std(X) are the mean and standard deviation of X. The standard deviation is the square root of the variance.

$$\begin{aligned}\bar{X} &= \frac{1}{\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{cy\in\mathcal{I}(\mathbf{x},\mathbf{y},\mathbf{s})} X_{cy}, \\ \text{var}(X) &= \frac{1}{\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{cy\in\mathcal{I}(\mathbf{x},\mathbf{y},\mathbf{s})} (X_{cy}-\bar{X})^2, \\ \text{std}(X) &= (\text{var}(X))^{1/2}. \qquad (3)\end{aligned}$$

Similar definitions hold for $\bar{Y}$ and and std(Y)[8].

Now we will define the different versions of within-group correlation. We will describe the case of grouping by country; grouping by year works similarly. There are two slightly different natural definitions of within-group correlation. To define them, we first introduce the mean, variance, standard deviation, and correlation for a given country c,

$$\begin{aligned}\bar{X}_c &= \frac{1}{\mathcal{N}_c(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{Y\in\mathcal{I}_c(\mathbf{x},\mathbf{y},\mathbf{s})} X_{cy}, \\ \text{var}(X_c) &= \frac{1}{\mathcal{N}_c(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{Y\in\mathcal{I}_c(\mathbf{x},\mathbf{y},\mathbf{s})} (X_{cy}-\bar{X}_c)^2, \\ \text{std}(X_c) &= (\text{var}(X_c))^{1/2}, \qquad (4) \\ \text{corr}(X_c,Y_c) &= \frac{1}{\mathcal{N}_c(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{Y\in\mathcal{I}_c(\mathbf{x},\mathbf{y},\mathbf{s})} \frac{(X_{cy}-\bar{X}_c)(Y_{cy}-\bar{Y}_c)}{\text{std}(X_c)\ \text{std}(Y_c)}. \qquad (5)\end{aligned}$$

The within-group variance of X is the average of the squared differences of X_{cy} from the

[8] For statisticians, the standard deviation here is is normalized as the *uncorrected* standard deviation. See Appendix E.5.

country mean, and the within-group standard deviation is the square root of that,

$$\begin{aligned}
\mathrm{var}_{within-country}(X) &= \frac{1}{\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{cY \in \mathcal{I}(\mathbf{x},\mathbf{y},\mathbf{s})} (X_{cy} - \bar{X}_c)^2, \\
\mathrm{std}_{within-country}(X) &= \left(\mathrm{var}_{within-country}(X)\right)^{1/2}. \qquad (6) \\
& \qquad (7)
\end{aligned}$$

The within-country variance equals a weighted average of the variances for the individual countries,

$$\begin{aligned}
\mathrm{var}_{within-country}(X) &= \sum_{c \in \mathrm{Countries}(s)} w_c \, \mathrm{var}(X_c), \qquad (8) \\
& \qquad (9)
\end{aligned}$$

where the weighting for a country is just the fraction of the data due to that country,

$$w_c = \frac{\mathcal{N}_c(\mathbf{x},\mathbf{y},\mathbf{s})}{\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s})}. \qquad (10)$$

One natural definition of within-country correlation is the weighted sum of the correlations per country:

$$\begin{aligned}
\mathrm{corr}_{within-country}(X,Y) &= \sum_{c \in \mathrm{Countries}(s)} w_c \, \mathrm{corr}(X_c, Y_c) \\
&= \frac{1}{\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{cy \in \mathcal{I}(\mathbf{x},\mathbf{y},\mathbf{s})} \frac{(X_{cy} - \bar{X}_c)(Y_{cy} - \bar{Y}_c)}{\mathrm{std}(X_c)\ \mathrm{std}(Y_c)}. \qquad (11)
\end{aligned}$$

The first formula shows that the within-country correlation lies between minus one and one, since it is the weighted average of quantities that do. The second formula for within-country correlation is just like the formula for total correlation except that the means subtracted and standards deviation divided by are country dependent.

The other natural definition of within-country correlation is what can be called the definition with "tied standard deviations". This is just the same definition as Eq. 11, but the per-country standard deviations in the denominator are replaced by the within-country standard deviations:

$$\mathrm{corr}^{std-tied}_{within-country}(X,Y) = \frac{1}{\mathcal{N}(\mathbf{x},\mathbf{y},\mathbf{s})} \sum_{cy \in \mathcal{I}(\mathbf{x},\mathbf{y},\mathbf{s})} \frac{(X_{cy} - \bar{X}_c)(Y_{cy} - \bar{Y}_c)}{\mathrm{std}_{within-country}(X)\,\mathrm{std}_{within-country}(Y)}. \qquad (12)$$

This std-tied, within-country correlation lies between minus one and one because it can be interpreted as the dot product of two unit vectors, i.e. vectors of length one (see

Appendix E.7).

5.2 Results for Different Types of Correlation

The table in Appendix B compares the average (over all four studies) of the correlation between past changes of data for a code and future stock returns. (To keep the results manageable, we do not report within group correlations for the individual studies as we did for total correlation in Appendix A.) The different methods of calculating an aggregate correlation between past and future data vectors, X_{cy} and Y_{cy}, indexed by country-year pairs for which there is data are:

- **total:** This is just the correlation between data vectors ignoring group labeling by either year of country. The means subtracted and standard deviations divided by are statistics summarizing all country-year data. We say the means and standard deviations are "tied" (to have the same value independent of country or year).
- **within-country:** The means subtracted and standard deviations divided by summarize data over all years for a given country. This correlation is a weighted average over all countries in the study of the country-specific correlation across years.
- **within-country, std-tied:** The means subtracted depend on the country, but the standard deviations divided by are independent of country (and year).
- **within-year:** The means subtracted and standard deviations divided by summarize data over all countries for a given year. This correlation is a weighted average over years in the study of the year-specific correlation across countries.
- **within-year, std-tied:** The means subtracted depend on the year, but the standard deviations divided by are independent of year (and country).

In Section 4.1, we saw that codes for which a positive change is "good" tended to have a positive correlation with future stock returns, codes for which a positive change is "bad" tended to have a negative correlation, and codes for which it is "unclear" whether a positive change is good or not tended to have correlation about zero (within 0.025 of zero). In addition, codes that had to do with GDP tended to have a negative correlation with future stock returns. Stocks themselves tend to be mean reverting (i.e. have a negative correlation with their future change). In the last section, where we looked at the mean over studies of the total correlation, there were only two exception to the tendencies just described. In Table 6 we tabulate the number of exception to these tendencies by type of code and type of correlation calculation. This table can be derived from the table in Appendix B, which gives results for individual codes. The exception rates for all good or bad codes are far below the rate of 50% one would expect by random chance.

corrType	**GDP** **4 codes**		**stock** **1 code**		**bad** **14 codes**		**good** **144 codes**		**unclear** **7 codes**		**good or bad** **158 codes**	
total	0	0.0%	0	0.0%	1	7.1%	0	0.0%	1	14.3%	1	0.6%
within-country	0	0.0%	0	0.0%	2	14.3%	5	3.5%	5	71.4%	7	4.4%
within-country, std-tied	0	0.0%	0	0.0%	1	7.1%	1	0.7%	3	42.9%	2	1.3%
within-year	0	0.0%	0	0.0%	0	0.0%	14	9.7%	3	42.9%	14	8.9%
within-year, std-tied	2	50.0%	0	0.0%	0	0.0%	1	0.7%	2	28.6%	1	0.6%

Table 6: Count of codes for which the mean over all studies of the correlation between past change of the code with future stock returns is not "consistent with good-is-good". Rows correspond to different methods of computing correlations from data indexed by country-year pairs. Columns correspond to different types of codes.

6 Multi-Factor Regression

We have seen that the signs of the correlation of past changes of individual V-Dem codes with future stock returns are most often "consistent with good-is-good", i.e. changes of indicators for which a positive change is "good" (or "bad") correlate positively (or negatively) with future stock returns. This tells us that periods with above average increase of a good (or bad) indicator tend to be followed by periods with above (or below) average stock returns. The magnitude of the correlation tells us how large this co-movement effect is. Specifically the square of the correlation coefficient is the proportion of the deviation (from the mean) of future stock market returns that is accounted for by a simple linear regression on the past V-Dem indicator being considered. This is known as the R-squared of the regression. It is reviewed in detail in Appendix E.8.

For a typical correlation of ten percent (0.1), the R-squared is only one percent (0.01). So, although the effect we are seeing is quite consistent, it is a fairly small effect when judged in terms of power to explain individual observations. This does not mean it is an inconsequential effect. Just ask any citizen to judge whether they prefer the long term compounded effects of degraded stock markets following collapses in civil society vs compounded enhanced stock returns following social improvements.

It is natural for us to ask: *How additive are the "good-is-good" improvements "predicted" by different indicators?* Our first approach to addressing this question is discussed in Section 6.1, where we compare two things. The first thing is the average correlation with stocks across all "good or bad" indicators (the 158 V-Dem indicators that we consider "good" or "bad"). The second thing is the correlation of stocks with a "good index" created from those indicators. The good index is essentially an average of the indicators multiplied by their "goodness sign" ("+" or "−" as specified in Appendix A).

The second approach to addressing additivity is discussion in Section 6.2. There we look at positive linear regression of future stock changes on past changes of all "good or bad" indicators (and various subsets as well). This is just like ordinary linear regression except that the sign of the regression coefficient for a code is constrained to agree with its goodness sign.

6.1 Average Correlation and Correlation of Average

Before reporting the results of constrained multiple regression, let us consider a simpler way to combine a set of V-Dem indicators, which we call the "good index". This is just the average of the Z-scores[9] for the indicators for which data is available, multiplied by their goodness signs. Let us define this formally.

To begin, we denote the goodness sign for a code $\mathbf{x}$ by goodsign($\mathbf{x}$). This is either $+1$ or -1, according to the sign in Appendix A.

Given a set, $\{\mathbf{x}^1, ..., \mathbf{x}^k\}$, of indicators and a study $\mathbf{s}$, let $\{Y_i, X_i^1, ...X_i^k\}_{i=1}^N$ be the past and future change data, where i labels the country-year pair. We include all years for which there is past change data for $\mathbf{y}$ and for at least one of the $\mathbf{x}^l$. Unavailable data is imputed (filled in) by the mean of available data. N is the size of the index set

$$\mathcal{I}(\mathbf{x}^1, ..., \mathbf{x}^k, \mathbf{y}, \mathbf{s}) = \bigcup_{l=1}^{k} \mathcal{I}(\mathbf{x}^l, \mathbf{y}, \mathbf{s}). \tag{13}$$

Let A_i be the set of l for which the data X_i^l is available (not just filled in by a mean of available data). Let $nAvail_i$ be the number of elements in A_i. Then

$$goodIndex_i = \frac{1}{nAvail_i} \sum_{l \in A_i} \text{goodsign}(\mathbf{x}^l) Z_i^l, \tag{14}$$

$$Z_i^l = \frac{X_i^l - \bar{X}^l}{\text{std}(X^l)}. \tag{15}$$

Each row in Table 7 gives results for a different set of V-Dem codes. For each of the studies `MSCI 10`, `DMS 10`, `MSCI 5`, and `DMS 5`, we report: (i) the correlation between future stock returns and the *goodIndex* associated with a code set (called `corr_index`), and (ii) the average over all codes in the code set of the correlation of future stock returns with past change of the code times the good sign of the code (called `ave_corr`).

The first five lines of the table just repeat the correlation in the Appendix A for the five codes with highest `meanCorr`. `corr_index` equals `ave_corr` for these lines since the code set is just a single code.

The bottom four lines of the table show results for the code sets being: the top two

[9] The Z-score for a data vector is computed by first subtracting the mean of the components of the vector from the vector and then dividing by the standard deviation. Subtraction of the means changes the index by an overall constant, which doesn't effect the correlation coefficient. Normalizing by dividing by standard deviations makes the size of the contribution of different indicators similar. To be precise, we should say that we divide by the standard deviation of the available data (as opposed to the standard deviation of the series with missing data filled in with the mean, which can be smaller).

codes, the top five codes, the high level codes in Table 5, and all V-Dem codes considered either “good” or “bad”. We see that `corr_index` consistently improves upon `ave_corr`. The improvement is significant for the bottom line, which reports overall results (for the code set being all good or bad indicators).

		MSCI 10 year		DMS 10 year		MSCI 5 year		DMS 5 year	
code set	*num codes*	*corr index*	*ave corr*	*corr index*	*ave corr*	*corr index*	*ave corr*	*corr index*	*ave corr*
rank 1	1	0.15	0.15	0.26	0.26	0.15	0.15	0.20	0.20
rank 2	1	0.16	0.16	0.26	0.26	0.15	0.15	0.18	0.18
rank 3	1	0.10	0.10	0.24	0.24	0.18	0.18	0.24	0.24
rank 4	1	0.17	0.17	0.22	0.22	0.14	0.14	0.21	0.21
rank 5	1	0.15	0.15	0.26	0.26	0.13	0.13	0.20	0.20
top 2	2	0.16	0.16	0.26	0.26	0.15	0.15	0.20	0.19
top 5	5	0.15	0.15	0.26	0.25	0.16	0.15	0.22	0.21
high level	13	0.16	0.15	0.24	0.21	0.13	0.11	0.19	0.16
all good and bad	158	0.16	0.10	0.23	0.17	0.14	0.08	0.18	0.12

Table 7: Average of correlations and correlations of averages for following sets of V-Dem codes: the singleton sets with the top five ranked codes in the Table in Appendix A; the top two and top five codes; the thirteen high level codes considered in Table 5; and all V-Dem codes which are considered either “good” or “bad”. Column *num_codes* gives total number of codes in each code set. Columns *corr_index* (one for each study) give the total correlation of future stock returns with the *goodIndex* of each code set. The *goodIndex* is the average of the data for the codes (multiplied by the code’s goodsign and divided by the standard deviation of the code’s data). Columns *ave_corr* give an average (over codes in a code set) of the total correlation of future stock returns with each code’s data multiplied by the code’s goodsign.

6.2 Regression with Sign Constraint

The good index (for any given code set) can be written as a linear combination of the separate indicators. The good index has the property that the sign of the coefficient of each indicator agrees with its goodsign. We will call a linear combination with this property a *positive linear combination.* In this section, we report on a constrained version of linear regression which gives the positive linear combination that optimally fits the data.

For each choice of a study and a code set with k codes, we have a data set consisting of future stock return data Y_i and past change data $X_i^1, ..., X_i^k$, where i labels a country-year pair. Our regression chooses the coefficients β_i in the model[10] (16) which minimize the

[10] The usual way of writing a regression models include a constant term. In Eq. 16, this term has been solved for and equals the sum of the terms involving means. When there is no missing data, the good index has $\beta_l = \text{goodsign}(\mathbf{x}^l)/(k\ \text{std}(X^l))$.

sum of squared errors (17) subject to the constraint (18).

$$\hat{Y} \;=\; \bar{Y} + \beta_1(X^1 - \bar{X}^1) + ...\beta_k(X^k - \bar{X}^k), \tag{16}$$

$$||Y - \hat{Y}||^2 = \sum_{i=1}^{N} \left| Y_i - \left[\bar{Y} + \beta_1(X_i^1 - \bar{X}^1) + ...\beta_k(X_i^k - \bar{X}^k)\right] \right|^2, \tag{17}$$

$$\beta_i = 0 \qquad \text{or} \qquad \text{sign}(\beta_i) = \text{goodsign}(\mathbf{x}^i) \qquad\quad \text{for } i = 1, ..., k. \tag{18}$$

A standard measure of the goodness of fit is called the R-squared. It measures the fraction of (the variation from the mean of) the Y data that is explained by (the variation from the mean of) the model $\hat{Y}$. To be precise, R-squared is the ratio of the sum of squares of the "predicted" deviations from the mean to the sum of squares of the actual deviations, see Eq. 106.

Table 8 gives the number of non-zero coefficients and the R-squared for positive regression of future stock versus various sets of V-Dem codes. The sets of V-Dem codes considered are the same as in Table 7, to which Table 8 is comparable.

We observe the following:

- When regressing against a single indicator, as done in the first five rows, the R-squared values are just the square of the correlation coefficients reported in Table 7 and Appendix A.
- When regressing against a set of indicators the R-squared values must be greater than or equal to the maximum of the R-squared values for the indicators considered one-by-one.
- For row six through eight, the amount by which the R-squared for positive regression exceeds the maximum for the separate codes is negligible.
- The final row does show a large increase in R-squared, to on the order of 15%, when all 158 (good or bad) V-Dem codes are regressed against.

In Appendix E.8, we give a detailed account of hypothesis testing and the role of R-squared for unconstrained linear regression. One fact we discuss, which is known to anyone familiar with multiple regression, is that the R-squared values become large when many predictors (X variables) are included and that this is not necessarily significant. One standard practice is to calculate an "adjusted R-squared", which only gets larger when new predictors are added if they reduce the squared error by more than what one would be expected by chance. For ordinary regression, the adjustment depends on the number of predictors and the number of data points. In our constrained case, it is not a priori clear whether to adjust based on the number of codes in the full code set or to adjust based on the number of active codes (codes with a non-zero coefficient). Another standard practice to evaluate regression results is to apply a formal hypothesis test. But, as with the adjustment of R-squared, it is unclear how to account for active vs inactive codes when hypothesis testing. In the next section, we will see how the proper thing to do in hypothesis testing and R-squared adjustment for our problem is to adjust by the number of active coefficients. We will see that the overall positive regression results in the last row of Table 8 are indeed significant.

	num	MSCI 10 year		DMS 10 year		MSCI 5 year		DMS 5 year	
code set	*codes*	$k_{\neq 0}$	R^2	$k_{\neq 0}$	R^2	$k_{\neq 0}$	R^2	$k_{\neq 0}$	R^2
rank 1	1	1	0.02	1	0.07	1	0.02	1	0.04
rank 2	1	1	0.03	1	0.07	1	0.02	1	0.03
rank 3	1	1	0.01	1	0.06	1	0.03	1	0.06
rank 4	1	1	0.03	1	0.05	1	0.02	1	0.04
rank 5	1	1	0.02	1	0.07	1	0.02	1	0.04
top 2	2	1	0.03	2	0.07	2	0.02	1	0.04
top 5	5	2	0.03	2	0.07	3	0.03	2	0.06
high level	13	3	0.04	4	0.07	5	0.02	4	0.04
all good and bad	158	22	0.18	19	0.18	24	0.11	16	0.13

Table 8: Results of positive linear regression of future stock returns versus the various sets of V-Dem codes considered in Table 7. Non-zero regression coefficients are constrained to be positive for "good" codes or negative for "bad" codes. Column `numCodes` gives total number of codes in each code set. For each of the four studies considered here, columns $k_{\neq 0}$ and R^2 give the number of non-zero coefficients and the R-squared of the regression for that study. These active codes are displayed in Appendix C on p. 49.

7 Statistical Significance

In the following subsections we explore the statistical significance of the following results in this paper:

§7.1 The percentage of indicators which follow the "good-is-good" pattern.

§7.2 The value of the correlation of individual indicators and the overall good index.

§7.3 The amount of data accounted for by regression with coefficients constrained to equal the goodsign.

Significance testing for our circumstances asks the following question:

> *Suppose the past data is given (i.e., the actual V-Dem data is used) and the future stock return data equals a constant return plus white noise. What is the P-value, that is the probability, that this random process will produce results as least as extreme as the real-life observed results?*

By white noise we mean that each observation is sampled independently from normal distributions with mean zero and the same standard deviation.

In formal hypothesis testing of a scientific experiment, one often specifies, prior to conducting the experiment, a *significance level* (typically five percent). If the probability value of obtaining the observed result (or more extreme) by random chance is less than the significance level, one is said to be entitled to *reject the null hypothesis.* One is technically only entitled to reject the specific form that was assumed for the null hypothesis, which, in our case, is the assumption that stock returns are a constant plus white noise. But a low probability for the simplest model of random chance often implies a low probability for related, more subtle, versions of random chance.

We will not take the formal step of deciding whether or not to reject the null hypothesis, but simply report that the probability value for the null hypothesis for various versions of results R1-R3 above are negligibly small.

7.1 Significance of the Percentage of Indicators which Follow the "Good-Is-Good" Pattern

Let $\mathbf{x}$ be one of the 158 "good or bad codes", which have goodsign of $+1$ (positive changes are "good") or -1 (positive changes are "bad"). Also let $meanCorr(\mathbf{x})$ be the mean over the four studies considered here of the total correlation between past changes of $\mathbf{x}$ and future stock returns. One of our most striking results is that the sign of $meanCorr(\mathbf{x})$ agrees with goodsign($\mathbf{x}$) for 157 out of 158 cases. Another way to put this result is that the number of exceptions to the "good-is-good" pattern is 1 in 158, which is less than one percent. If the data for every pair of codes was always uncorrelated and we didn't have to worry about missing data, then, under the null hypothesis of random stock returns, the probability that $meanCorr(\mathbf{x})$ agrees with goodsign($\mathbf{x}$) would be an independent unbiased coin flip for each code. Clearly, the odds of flipping one (or fewer) heads out of 158 coin flips is astronomically small.

7.1.1 Number of Exceptions to "Good-Is-Good" Rule for Within-Group Correlation

In a moment, we will calculate some P-values properly by simulation. The numbers are not as astronomically small as the simple coin flip calculation would dictate, but they still show the results are extremely significant. Before doing this, we will purposely throw a little shade on our result by showing how the number of exceptions varies when measured in slightly different ways. Table 9 presents the number of exceptions (and the percentage of exceptions) for the five different types of correlation reported on in Section 5.2. It reports on each study separately, as well as for the mean over studies. The number of exceptions in the mean column is just the sum of the "minus" and "plus" columns in Table 6. The results for the individual studies clearly have more exceptions than the mean over studies. Similarly, using any type of within-group correlation leads to more exceptions than total correlation. This is particularly so for experiments with untied standard deviations, which is not surprising given the difficulty of estimating standard deviations for groups with small amounts of data. But the bottom line is that, apart from some expected exceptions when standard deviations are not tied, the exception rates are all well below fifty percent.

corrType	**MSCI 10**	**DMS 10**	**MSCI 5**	**DMS 5**	**mean**
total	13 (8%)	5 (3%)	14 (9%)	6 (4%)	1 (1%)
within-country	50 (32%)	6 (4%)	26 (16%)	11 (7%)	7 (4%)
within-country, std-tied	20 (13%)	5 (3%)	20 (13%)	6 (4%)	2 (1%)
within-year	13 (8%)	48 (30%)	25 (16%)	83 (53%)	14 (9%)
within-year, std-tied	13 (8%)	19 (12%)	15 (9%)	22 (14%)	1 (1%)

Table 9: Number of codes $\mathbf{x}$ which are exception to "good is good" rule that the sign of the correlation for $\mathbf{x}$ agrees with goodsign($\mathbf{x}$). Columns of the table are different studies and mean over all studies. Rows of the table are different methods of calculating correlation.

7.1.2 P-Value of Number of Exceptions to "Good-Is-Good" Rule for Total Correlation

We will now present the P-values for the number of exceptions to the "good is good" rule for the individual study results for total correlation (see top row of Table 9). We restrict to looking at total correlation, which was our original method of calculation and which, as we have just seen, gives stronger results than within-group correlation.

For each study, we calculate a P-value by performing 100,000 trials of simulating stock-return series by independent sampling from a normal distribution[11]. For each series, we calculate the total correlation with each of the V-Dem indicators and calculate the number of exceptions to the good-is-good rule. Table 10 gives the statistics of the number of exceptions for the simulated distribution. The column **real** gives the number of exception for the real-world stock data. The column **pValue** gives the frequency with which the simulation gave the same or fewer exceptions than the real-life stock data. Since we use a large number of trials, this is a good approximation to the theoretical P-value. Additional columns of the table give more information about the distribution of the number of exceptions.

Note that the mean of the simulated number of exceptions is just half of the total number of "good or bad" codes. This is as expected from symmetry even though the codes are not independent. If the codes were independent, the standard deviation of the number of exceptions would be $\sqrt{158}/2 = 6.3$. On the other hand, if the codes were all identical, the standard deviation would be $158/2 = 79$ (because the number of exceptions would be 158 half of the time and 0 the other half the time). The fact that values in the **std** column are much larger than 6.3 indicates that the codes are not all independent.

Figure 2 plots the simulated null-hypothesis distribution, the normal approximation to the simulation distribution, and the real-world value of the number of exceptions. The simulated null-hypothesis distribution is concentrated at the top of its normal approximation, whose much longer tails have been cutoff in the figure. This means that the null-hypothesis distribution of the number of exceptions is very thin-tailed compared to a normal distribution. This is reflected in Table 10 by the fact that the simulated **pValue** is much smaller than the value, **pZ**, that the normal approximation would yield. Another measure of thin-tailedness is that the (absolute value of the) Z-score of the minimum of the distribution, $\frac{\textbf{mean}-\textbf{min}}{\textbf{std}}$, is under 2 for all the studies, whereas it would be about$\sqrt{2 \log 100000} = 4.3$ if the null-hypothesis distribution were a normal distribution.

[11] The mean and standard deviation of the normal distribution which we sample from are immaterial since they are subtracted off and divided out in the calculation.

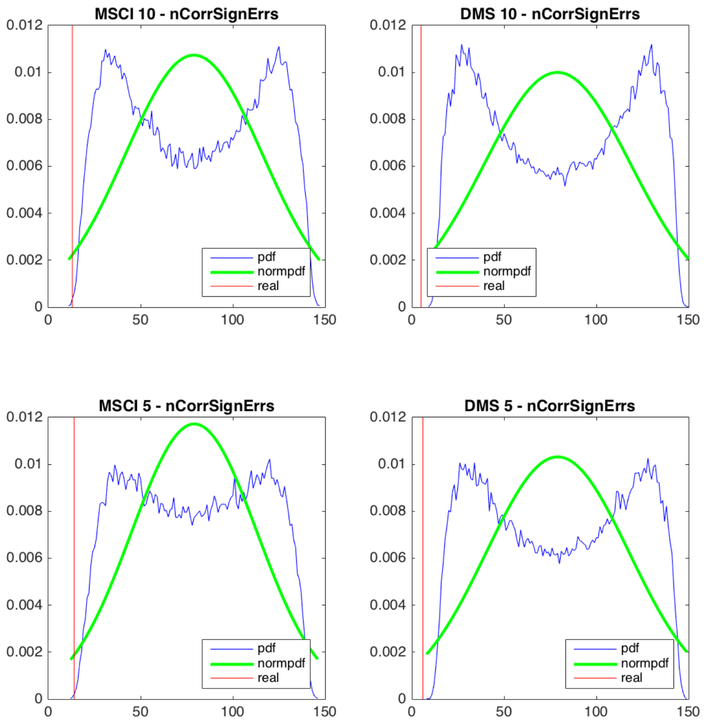


Figure 2: Distribution of the number of indicators whose total correlation with future stock returns disagree with the “good-is-good” rule. Blue curve is based on 100,000 simulations of normally distributed stock returns series; green curve is normal approximation to the distribution); and the vertical red line shows the real observed value.

study	real	pValue	nSmall	min	mean	std	pZ
MSCI 10	13	0.0002	16	11	79	37.2	0.038
DMS 10	5	0	0	9	79	39.9	0.032
MSCI 5	14	0.0001	10	12	79	34.0	0.028
DMS 5	6	0	0	8	79	38.7	0.029

Table 10: Test, for each study, of significance of the number of indicators whose total correlation with future stock returns disagree with the “good-is-good” rule. Column **real** gives the actual number of exceptions (count of codes for which the sign of the correlation disagrees with the code’s goodsign). Remaining columns give statistics based on 100,000 simulations with randomly generated stock returns. **nSmall** gives the number of simulations for which the number of exceptions (sign disagreements) is small, i.e. no larger than the actual **real** value. **pValue** gives the fraction **nSmall**/100000 of trials with a small number of exceptions. **min**, **mean**, and **std** give the minimum, mean, and standard deviation of the simulated number of exceptions. Finally, **pZ** gives the probability, under a normal distribution with the calculated mean and standard deviation, that the number of exceptions is at least as small as the **real** value. **pZ** equals the cumulative normal distribution of the Z-score $\frac{\textbf{real}-\textbf{mean}}{\textbf{std}}$.

7.2 Significance of the Values of Correlation

7.2.1 Distribution of Correlation under Null Hypothesis

Under the null hypothesis that stock returns are sampled randomly and independently from a normal distribution, the theoretical distribution of the correlation c between the change data for a given index and stock returns can be calculated from the well-known fact that the probability distribution for the *t-statistic* (Eq. 67) is the t-distribution with $\nu = n - 2$ degrees of freedom. In Appendix E.6 we derive the distribution of c directly (as Eq. 65) and show that is equivalent to the t-distribution of the t-statistic. The probability density function, $pdf(c)$, for c is, up to a normalization constant, equal to $1 - c^2$ raised to the power $(\nu - 2)/2$. The mean of c is zero and its standard deviation is the square root of $1/(1 + \nu)$.

The P-value that the correlation is greater than an observed real-world value c_{real} equals the integral of $pdf(c)$ from c_{real} to 1. For large n (which we have here), this P-value can be calculate using a normal distribution. If c_{real} is *nStdOut* standard deviations above the mean (i.e. $c_{real} = nStdOut/\sqrt{1 + \nu}$), the P-value is the tail probability of being *nStdOut* or more standard deviations out on a bell curve. The tail probabilities for various *nStdOut* are given in the following table.

nStdOut	0	1	2	3	4	5
tail prob	50%	16%	2.3%	0.1%	0.003%	0.00003%

Table 11: Tail probabilities of a normal distribution for various numbers of standard deviations above the mean.

For example, if we take n to be 398, which is the minimum number of data points used to compute the correlation for any of the indicators in any of the studies (see Table 4), $1/\sqrt{n - 1}$ equals 0.05. So, even when we have minimal data, a typical correlation of 0.1 is two standard deviations above the mean of zero and so has a P-value of 2.3%.

7.2.2 P-Values of Total Correlation for Individual Codes

In a moment we will focus on the P-value for the *overall good index* (the *goodIndex* associated to the set of all 158 good or bad codes). Before doing so we give a feel of the P-values for each of these indicators separately in Table 12. The top row of the table gives the percentages of codes which have P-value above 50%, i.e. the percentage of codes whose correlation has the wrong sign. These are the same percentages given in the first row of Table 9. Subsequent rows of Table 12 give the percentages of codes which have P-values above smaller thresholds. The lower right corner of the table shows that for 57% of codes, the chance of randomly obtaining a mean total correlation at least as big as seen in the real-world data is (by a conservative calculation) below one in a thousand.

threshold pVal	MSCI 10	DMS 10	MSCI 5	DMS 5	meanCorr
50%	8%	3%	9%	4%	1%
5%	23%	11%	27%	8%	16%
1%	37%	14%	36%	14%	27%
0.1%	54%	16%	60%	20%	43%

Table 12: Percentages of good or bad codes whose total correlation with stocks have P-value above various thresholds. Results are given for each study separately and for the mean over studies. In the case of the mean over studies, the number of data points is taken to be the most conservative value, i.e. the minimum of the number of data points for each study separately.

7.2.3 P-Values of Total Correlation with the Overall Good Index

The P-values for the total correlation of stock returns with the overall good index are easy to calculate using the exact distribution (Eq. 65). In Table 13, we give both the exact P-values and those calculated by simulation. The closeness of the theoretically exact and simulated P-values in Table 13 should give the reader confidence that the simulated values in Tables 10 and 15 are close to the theoretically exact, but not easily calculable, values. Figure 3 shows that the simulated distribution is close to a normal distribution, which is to be expected (as proved in Appendix E.6) since the number of data points is large. A numerical confirmation of this closeness is that the Z-score of the maximum of the distribution, $\frac{\mathbf{max-mean}}{\mathbf{std}}$, is 4.3, 4.6, 4.4, and 4.4 for the different studies, all of which are close to the value $\sqrt{2\log 100000} = 4.3$ for a normal distribution.

The table shows that the P-values calculated exactly, by simulation, or using the normal distribution based on simulated statistics are all extremely small. *In other words, the positivity of the total correlation of stock returns with the overall good index is extremely statistically significant.* As one might expect considering that fact that tail events are so rare, the normal distribution based on simulation statistics gives a more precise estimate than the simply frequency count of large correlations in simulation.

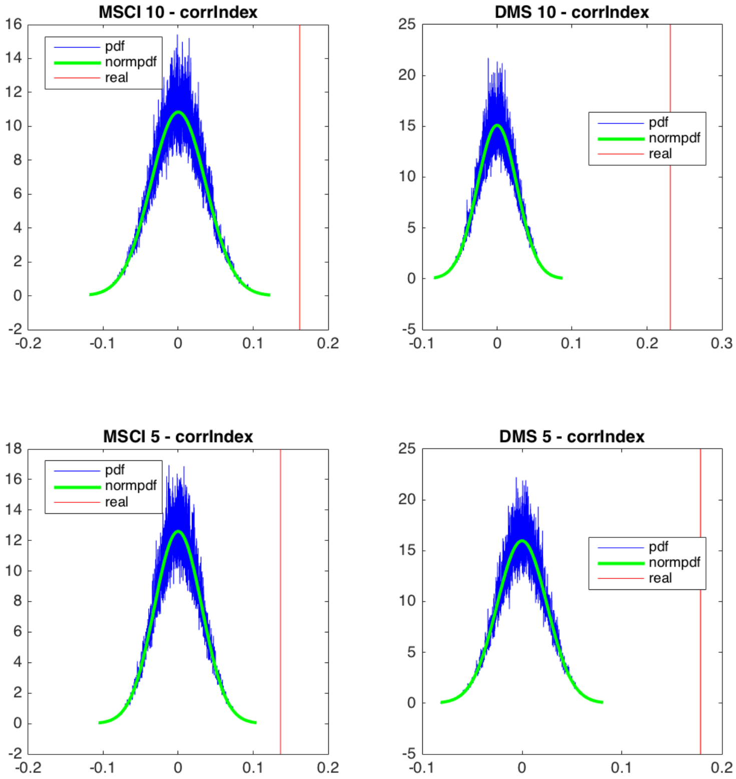


Figure 3: Distribution of the total correlation of the overall good index with normally distributed random stock returns series (blue); normal approximation to the distribution (green); and the real-world value (red).

study	real	pValue	nBig	max	mean	std	pZ	pNull
MSCI 10	0.16	0	0	0.16	7e-05	0.037	5e-06	4e-06
DMS 10	0.23	0	0	0.12	5e-05	0.026	1e-18	5e-20
MSCI 5	0.14	2e-05	2	0.14	-1e-04	0.032	8e-06	6e-06
DMS 5	0.18	0	0	0.11	-6e-05	0.025	4e-13	1e-13

Table 13: Test, for each study, of significance of the total correlation of the overall good index with future stock returns. Column **real** gives the correlation observed. Remaining columns give statistics based on 100,000 simulations with randomly generated stock returns series. **nBig** gives the number of simulations for which the simulated correlation is bigger than the **real** value. **pValue** gives the fraction **nBig**/100000 with big values. **max**, **mean**, and **std** give the maximum, mean, and standard deviation of the simulated correlations. **pZ** gives the probability, under a normal distribution with the calculated mean and standard deviation, of a correlation being at least as big as the **real** value. Finally, column **pNull** gives the precise value of the P-value for the null hypothesis (i.e. the above probabilities calculated theoretically, rather than by simulation.

7.2.4 P-Values for Within-Group Correlation

As we have seen, for example in Table 9, within-group correlations are weaker than total correlation. P-values for within-group correlations can be calculated by the method described in Appendix E.7. When there is enough data, P-values can still be calculated using a normal distribution as they were in the case of total correlation. For the case of tied standard deviations, the correlation distribution still takes the same form (Eq. 65) although now the number of degrees of freedom, ν, equals $n - 1 - (\#countries)$ and $n - 1 - (\#years)$ for the cases of `within-country, std-tied` and `within-year, std-tied`, respectively, whereas it was $n - 2$ for total correlation. The variance of the correlation distribution is still $1/(\nu+1)$. The PDF is more complicated when the standard deviations are not tied. We refer the reader to the appendix.

Table 14 presents results for the four studies for various types of aggregate correlation between the overall good index and future stock returns, along with their P-values. To check how the tables we present "hang together", we note the following: The total correlation in Table 14 is just another presentation of data we have seen twice already, in (i) the `corr_index` entries in the last row of Table 7, and (ii) the column `real` in Table 13. Also, the `pNull` entries in the total row of Table 14 are extremely close to the column `pZ` in Table 13. The former is calculated by the normal distribution approximation to the theoretically exact P-value (see Section E.7) and the latter is calculated by the normal distribution approximation to the simulated distribution.

We note also the following consistency, which does not follow automatically on mathematical grounds but is a good illustration of the data hanging together: The `DMS 5` study with `within-year` group correlation (with standard deviation not tied) is the only problematic case in both Table 14 and Table 9. (For Table 14, "problematic" means that there is a negative correlation between future stock returns and overall good index change; for Table 9, "problematic" means that the percentage of good-or-bad V-Dem codes with a negative correlation is above 50%.) As we have seen before, that these cases are "problematic" is not surprising given the difficulty of estimating standard deviations for groups with small amounts of data.

7.3 Significance of Adjusted R-Squared for Constrained Regression

In Section 6.2, we reported results of positive regression of future stock returns against past changes of a collection of indicators, where the regression coefficients are subject to the constraint that their signs agree with the goodsign of their corresponding codes. In Table 8, we saw that the fraction of data explained, called R-squared, goes up from between two and seven percent for regression against a single top ranking code (in which case the R-squared is simply the square of the correlation coefficient), to between eleven

correlation	MSCI 10 year		DMS 10 year		MSCI 5 year		DMS 5 year	
types	*corr*	*pNull*	*corr*	*pNull*	*corr*	*pNull*	*corr*	*pNull*
total	0.16	5e-06	0.23	1e-18	0.14	8e-06	0.18	5e-13
within-country	0.05	1e-01	0.12	5e-06	0.12	7e-05	0.14	2e-08
within-country, std-tied	0.14	8e-05	0.25	1e-21	0.11	4e-04	0.18	2e-13
within-year	0.17	5e-06	0.02	3e-01	0.10	1e-03	-0.05	2e-02
within-year, std-tied	0.18	5e-07	0.12	5e-06	0.11	3e-04	0.04	4e-02

Table 14: Real-world values of various types of aggregate correlation between the overall good index and future stock returns, along with their P-Values. Each P-value is the probability that randomly generated stocks return (independently sampled from a normal distribution) will generate a correlation with the same sign and magnitude at least as big as the real-world value.

and eighteen percent for constrained regression against all 158 good or bad codes. In this subsection, we will show that these result are very statistically significant.

A crucial issue in understanding the significance of a regression is that the R-squared of the regression necessarily gets larger when new independent variables (predictors) are added. A common practice is to report an *adjusted R-squared* which depends on the number k of predictors:

$$R^2_{adj} = \frac{(n-1)R^2 - k}{n-1-k}. \tag{19}$$

As we show for unconstrained regression in Appendix E.8, the adjusted R-squared only gets larger when new predictors are added if they reduce the squared error by more than what would be expected by chance.

As pointed out at the end of Section 6.2, for constrained regression it is unclear *a priori* which value of k one should use to define adjusted R-squared in Eq. 19. Should we take k to be the total number of predictors, 158, or should we take k to be the number of active predictors, i.e. the number of codes with non-zero regression coefficient, which ranges from 16 to 22 in Table 8? We will justify below that the proper choice to make is the number of active predictors.

Figure 4 and Table 15 on page 38 display for adjusted R-squared the kind of figure and table pair we have seen twice before in this paper. The first pair was Figure 2 and Table 10 on page 31, which displayed results for the count of errors to the "good-is-good" rule. The second pair was Figure 3 and Table 13 on page 34, which displayed results for total correlation with the overall good index.

Specifically, Figure 4 shows: the simulated distribution of adjusted R-squared, adjusted by $k_{\neq 0}$; the normal approximation to this simulated distribution; and the real-world value of the adjusted R-squared. Table 15 gives numerical data about these plots.

We observe the following.

- The mean of the null-hypothesis distribution of adjusted R-squared values is very close to zero for all studies. This verifies that we have chosen the proper choice of adjustment so that, on average, random chance makes the adjusted R-square vanish.

- The distribution of adjusted R-squared values is somewhat thin-tailed compared to its normal approximation, at least for three out of four studies. This can be seen simply by looking at Figure 4. It can also be seen from the fact that Z-score of the maximum of the distribution, $\frac{\mathbf{max-mean}}{\mathbf{std}}$, is 6.7, 3.0, 4.0, and 3.0 for the different studies, which is, with the exception of the first study, a bit bigger than the value $\sqrt{2 \log 100000} = 4.3$ if the null-hypothesis distribution were normal.

- The P-values are extremely small. There were no instances in 100,000 trials in which the simulated adjusted R-squared was as big as the real-world adjusted R-squared. The normal approximations, **pZ** to the P-values are vanishingly small even though they are probably overestimates because the distributions are thin-tailed. *In other words, the fraction of real-world stock return data that can be modeled as a positive linear combinations of past changes of V-Dem indicators is extremely statistically significant.*

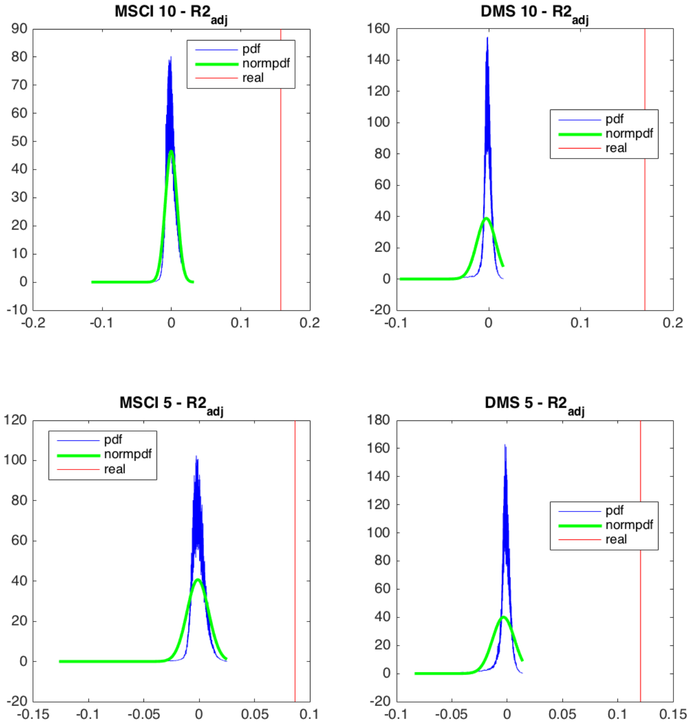


Figure 4: Distribution of the adjusted R-squared for positive regression of future stock returns on all good or bad indicators. Blue shows distribution based on 100,000 simulations of white noise for the stock return series; green shows normal approximation to the distribution; and red shows the real-world value.

study	**real**	**pValue**	**nBig**	**max**	**mean**	**std**	**pZ**
MSCI 10	0.16	0	0	0.06	-6e-04	0.009	7e-77
DMS 10	0.17	0	0	0.03	-3e-03	0.010	4e-63
MSCI 5	0.09	0	0	0.04	-1e-03	0.010	2e-19
DMS 5	0.12	0	0	0.03	-3e-03	0.010	6e-36

Table 15: Significance of the adjusted R-squared for sign-constrained regression of future stock return on all good or bad indicators. Column **real** gives the adjusted R-squared observed. Remaining columns give statistics based on 100,000 simulations with randomly generated stock returns series. **pValue** (**nBig**) gives the number (fraction) of simulations for which the simulated adjusted R-squared is bigger than the **real** value. **max**, **mean**, and **std** give the maximum, mean, and standard deviation of the simulated adjusted R-squared values. **pZ** gives the probability, under a normal distribution with the calculated mean and standard deviation, of an adjusted R-squared being at least as big as the **real** value.

8 Confidence Intervals for Population R-squared for Positive Regression

Having established the statistical significance of our positive regression results, we turn to calculating a 95% confidence interval for the true population R-squared for the model

$$Y = \mu + \tilde{X}\beta + \epsilon. \tag{20}$$

In this probabilistic model, future stock returns Y are equal to an overall mean value, plus a linear combination of the variation of the indicators from their means, plus white noise. The null hypothesis is the special case when β vanishes. This is the same form as the population model described in Appendix E.9.2, although here the model regression coefficients must obey our sign constraints. The definition of population multiple correlation coefficient, ρ, and population R-squared, ρ^2, remains the same as in the appendix: ρ is the correlation (for the population distribution) between Y and the population model "prediction" $\hat{Y} = \mu + \tilde{X}\beta$. ρ^2 is the fraction of the variance of Y (in the population distribution) that is explained by the model prediction.

Confidence intervals for unconstrained regression are discussed in detail in Appendix E.10. Calculations there are straight-forward to perform using the fact that the distribution of R^2_{adj} is a simple transformation of a non-central F-distribution, which depends only on the length of the coefficient vector β (which is determined ρ^2) and not on the direction of β.

The definition of confidence interval we use now is the same as described in the appendix: The confidence interval for ρ^2 is the range of ρ^2 values for which the adjusted R-squared observed for positive regression does not belong to the extreme wings (containing less than 2.5% in each wing) of the probability distribution. However, matters are more complicated in the positivity constrained case because the model distribution of adjusted R-squared is highly non-trivial and depends on the direction of β.

We will calculate confidence intervals using simulation. To simplify, we restrict ourselves to consider β vectors which are multiples of the observed regression coefficient vector, where the multiplicative constant is determined by ρ^2. In other words, we refrain from a computationally intractable scan over all directions of β. Based on a few additional tests, we do not believe that this simplifying restriction on β substantially effects our results. With the simplification, we were able to proceed in a computationally feasible way by implementing a binary search for the upper and lower bounds of the confidence interval, ρ^2_{lo} and ρ^2_{hi}. For each choice of ρ encountered in the binary search, we estimate the probability distribution of R^2_{adj} using ten thousand randomly generated Y vectors.

Table 16 presents the confidence intervals calculated for positive regression and, for comparison, unconstrained regression using k_{active} indicators. Note that the observed adjusted R-squared values lie approximately in the center of the calculated confidence intervals for ρ^2. This is a reflection of the fact that adjusted R-squared is an almost unbiased estimate of ρ^2. Also note the number of data points is large enough so that the confidence intervals are fairly tight. Two out of four of the confidence intervals lie entirely above the value 10%; one lies mostly above 10%; and the fourth lies partially above 10%. A simple summary is that a value of 10% for population R-squared lies in the confidence interval for all four studies.

One final point to note is that the confidence intervals for unconstrained regression using k_{active} indicators are almost identical to that of the positivity constrained confidence intervals. This is a confirmation of the fact that k_{active} was the proper value to use when applying the formula for adjusted R-square in the constrained case, rather than using the total number of regressors, 158, which is much larger.

				positive regression		unconstrained regeression	
study	n	k_{active}	$R2_{adj,obs}$	ρ^2_{lo}	ρ^2_{hi}	ρ^2_{lo}	ρ^2_{hi}
MSCI 10	739	22	0.16	0.11	0.21	0.11	0.21
DMS 10	1428	19	0.17	0.14	0.21	0.13	0.22
MSCI 5	996	24	0.087	0.055	0.12	0.051	0.12
DMS 5	1598	16	0.12	0.093	0.15	0.094	0.17

Table 16: For each study: the number of data points used, the number of non-zero regression coefficients and the observed value of adjusted R-squared for positive regression, and the confidence intervals for population R-squared for regression with our without the positivity constraint.
For the DMS studies, the column `n` specifying the number of data points used is less than the `nValsMax` column in Table 4 which specifies how many country-year pairs have some V-Dem indicator data. This is because the DMS data ends earlier than the V-Dem data.

9 Conclusion

We have documented that, over a five or ten year time scale, there has historically been, on average, a consistent positive correlation between future returns of a country's stock market and past changes of the same country's indicators that are socially "good". Of course the broadness of our claim is limited by the data we tested it with. For social indicators, we only use data from the V-Dem database (Coppedge et al. 2015a) which ends in the year 2012 and contains data for 173 countries, some of which goes back as far as 1900. Our work is by definition limited to look only at those countries which have available stock market data. Within this constraint, we were able to get a fairly wide coverage of countries and years by looking at two different databases of stock market data: DMS data (Dimson, Marsh & Staunton 2002, and data updates) covering 17 countries from 1900 to 2004; and MSCI data (*MSCI data index and analytics service*) covering 45 countries going back as far as 1970 and ending in 2012 (since that is when the V-Dem data ends).

Our results break down into four studies. Each study corresponding to a choice of either a five or ten year time scale and either the DMS or MSCI data (and the set of countries and years associated with them). Our initial, basic measure of consistency of correlation was striking: *The average over all four studies of the total correlation (across country-year pairs) between "good" past democracy indicators changes and future stock market returns is positive for* 157 *out of the* 158 *indicators that were selected solely based on whether they had enough data.* Robustness of this result is shown by its consistency across all four studies. However, a second limitation of our work is that we have not divided the data up in a way to do proper in-sample vs out-of-sample testing.

The third limitation of this work that we will mention is that there is no objective definition of the "good" direction of change of a V-Dem indicator, although it is somewhat implicit in the V-Dem codebook (Coppedge et al. 2015b). We have provided details results for individual indicators in appendices in the hope that some readers might want to delve into details. Although it is not objective, we have deliberately use the word "good" so that we can state the result in a manner that we hope will sway the powers that be: *What's good for society tends to be good for its markets.*

Although it seems obvious to the "naked eye" that our basic result is statistically significant, one must account for the fact that the V-Dem indicators are not at all mutually independent. Since the consistency of the effect seems striking and since we know of no comparable results, we have presented here a detailed statistical analysis of our basic result as well as related experiments. We limit ourselves to formal hypothesis testing, foregoing, for example, detailed factor analysis. The extreme statistical significance of our initial result is confirmed by hypothesis testing of both the percentages of exception to the "good-is-good" rule as well as the values of the total correlation for individual codes.

We also saw extreme statistical significance for all four studies of the total correlation of future stock returns with the “overall good index”, which is a signed average of all of the (suitable normalized) indicators. The within-year and within-group (with tied standard deviations) versions of this correlation were also very significant.

Although the sign of the correlations is robustly significant, the magnitude is not very big. It is natural to ask how much these small effects add together. To address this, we looked at positive regression of future stock returns on all the indicators, i.e. linear regression with coefficients constrained to have the “good” sign. We again found that the result were extremely statistically significant. For the positive regression results, we went beyond hypothesis testing against a null hypothesis and calculated 95% confidence intervals for what is known as the population R-squared, which is the percentage of data genuinely explained by regression, not just by fitting to noise. The lower end of the confidence window for the four studies was 11%, 14%, 6%, and 9%.

We have emphasized several limitations to our work in this conclusion in the hopes of stimulating critical thought and future work. One simple, concrete thing to do in the future would be to see if our results carry over to updates of the V-Dem database and to other countries and indicators. Another approach would be to use more complex models for both the effect of interest and the noise. It would be natural to apply techniques such as Granger causality testing discussed in econometric textbooks (Hamilton 1994, for example) or the theory of causal models discussed in statistic textbooks (Darlington & Hayes 2017, for example). Or one could focus on effects associated with particular sets of indicators, time periods, and countries. This would amount to adding a market focus to an already large literature relating social and economic effects (Ahlerup, Baskaran & Bigsten 2016, Knutsen 2014, just to pick two examples).

Most ambitiously, we hope that future work will understand how the broad brush pattern that social improvements are correlated with future market improvements fits in as part of the research community’s evolving understanding of the complex, dynamical, non-linear relation between social and economic changes.

A Table of All Total Correlations

This appendix contains a table giving the total correlation with future stock return for all codes and all four studies in Table 3. The column `meanCorr` gives the average correlation over all four studies. Rows of the table are sorted in order of decreasing `meanCorr`. The first column, `R`, specifies the rank order. The `notes` column specifies additional information about the code. For the code **stock**, the note is simply '`stock`'. For V-Dem codes that are derived from measures of GDP, the note is simply '`GDP`'. For all other codes, the first character of the note specifies whether a positive change in the code is: '`+`' – a good thing ; '`-`' – a bad thing , or '`?`' – unclear. The symbol '`f`' indicates that the wording in the description of the code sounds funny, but upon checking the definition in the V-Dem codebook (Coppedge et al. 2015b), positive is good. Finally, the symbol '`h`' indicates that the code is one of the "high-level" codes looked at in Table 5.

R	**code**	**des**	**mean corr**	**corr MSCI 10**	**corr DMS 10**	**corr MSCI 5**	**corr DMS 5**	**notes**
1	`v2x_freexp`	Freedom of expression index	0.19	0.15	0.26	0.15	0.20	+
2	`v2xcl_disc`	Freedom of discussion	0.19	0.16	0.26	0.15	0.18	+
3	`v2meslfcen`	Media self-censorship	0.19	0.10	0.24	0.18	0.24	+ f
4	`v2mecenefm`	Government censorship effort - Media	0.19	0.17	0.22	0.14	0.21	+ f
5	`v2x_freexp_thick`	Expanded freedom of expression index	0.19	0.15	0.26	0.13	0.20	+ h
6	`e_rol_free`	Civil liberties and rule of law index	0.18	0.16	0.24	0.14	0.19	+ h
7	`v2clrelig`	Freedom of religion	0.18	0.20	0.25	0.09	0.19	+
8	`v2xcs_ccsi`	Core civil society index	0.18	0.18	0.22	0.12	0.19	+ h
9	`v2clslavef`	Freedom from forced labor for women	0.18	0.20	0.26	0.09	0.16	+
10	`v2csrlgrep`	Religious organization repression	0.18	0.21	0.25	0.09	0.16	+
11	`v2cldiscw`	Freedom of discussion for women	0.18	0.16	0.24	0.14	0.16	+
12	`v2xme_altinf`	Alternative sources of information index	0.17	0.15	0.25	0.09	0.19	+
13	`v2x_liberal`	Liberal component index	0.17	0.15	0.23	0.14	0.16	+ h
14	`v2mecrit`	Print/broadcast media critical	0.17	0.19	0.24	0.10	0.15	+
15	`v2cldiscm`	Freedom of discussion for men	0.17	0.10	0.25	0.15	0.18	+
16	`v2clacfree`	Freedom of academic and cultural expression	0.17	0.10	0.26	0.15	0.18	+
17	`v2x_gencl`	Women civil liberties index	0.17	0.20	0.26	0.09	0.14	+
18	`v2x_partip`	Participatory component index	0.17	0.16	0.24	0.09	0.18	+ h
19	`v2meaccess`	Media access	0.17	0.14	0.22	0.15	0.16	+
20	`v2cldmovew`	Freedom of domestic movement for women	0.17	0.14	0.25	0.12	0.17	+
21	`v2xcl_rol`	Equality before the law and individual liberty index	0.17	0.17	0.22	0.12	0.16	+ h
22	`v2juhcind`	High court independence	0.17	0.15	0.25	0.13	0.15	+
23	`v2psparban`	Party ban	0.17	0.17	0.22	0.11	0.17	+ f
24	`v2xcl_slave`	Freedom from forced labor	0.17	0.20	0.23	0.09	0.14	+
25	`v2cseeorgs`	CSO entry and exit	0.16	0.15	0.21	0.10	0.20	+
26	`v2xdl_delib`	Deliberative component index	0.16	0.15	0.23	0.11	0.17	+
27	`v2xcl_dmove`	Freedom of domestic movement	0.16	0.15	0.23	0.12	0.15	+
28	`v2x_cspart`	Civil society participation index	0.16	0.15	0.26	0.10	0.15	+
29	`v2juncind`	Lower court independence	0.16	0.14	0.21	0.17	0.13	+
30	`v2mebias`	Media bias	0.16	0.07	0.24	0.11	0.23	+ f
31	`v2x_gencs`	Women civil society participation index	0.16	0.17	0.24	0.10	0.13	+
32	`v2clstown`	State ownership of economy	0.16	0.07	0.27	0.05	0.25	+
33	`v2xlg_legcon`	Legislative constraints on the executive index	0.16	0.12	0.21	0.17	0.14	+
34	`v2mecorrpt`	Media corrupt	0.16	0.10	0.21	0.16	0.16	+ f
35	`v2csreprss`	CSO repression	0.16	0.13	0.22	0.11	0.18	+
36	`v2dlconslt`	Range of consultation	0.16	0.17	0.21	0.12	0.13	+
37	`v2x_frassoc_thick`	Freedom of association (thick) index	0.16	0.15	0.19	0.12	0.16	+
38	`v2jureview`	Judicial review	0.16	0.18	0.16	0.14	0.13	+
39	`v2merange`	Print/broadcast media perspectives	0.16	0.13	0.25	0.05	0.19	+ f
40	`v2x_libdem`	Liberal democracy index	0.16	0.12	0.19	0.12	0.19	+ h

R	code	des	mean corr	corr MSCI 10	corr DMS 10	corr MSCI 5	corr DMS 5	notes
41	`v2dlreason`	Reasoned justification	0.15	0.20	0.20	0.06	0.15	+
42	`v2psoppaut`	Opposition parties autonomy	0.15	0.12	0.22	0.11	0.16	+
43	`v2mefemjrn`	Female journalists	0.15	0.05	0.25	0.16	0.15	+
44	`v2x_partipdem`	Participatory democracy index	0.15	0.13	0.19	0.11	0.18	+
45	`v2clacjstw`	Access to justice for women	0.15	0.16	0.24	0.09	0.12	+
46	`v2clkill`	Freedom from political killings	0.15	0.11	0.22	0.11	0.17	+
47	`v2lginvstp`	Legislature investigates in practice	0.15	0.10	0.23	0.16	0.12	+
48	`v2x_gender`	Women political empowerment index	0.15	0.18	0.21	0.08	0.13	+ h
49	`v2x_jucon`	Judicial constraints on the executive index	0.15	0.13	0.22	0.12	0.13	+ h
50	`v2clslavem`	Freedom from forced labor for men	0.15	0.19	0.22	0.04	0.14	+
51	`v2x_delibdem`	Deliberative democracy index	0.15	0.13	0.16	0.11	0.18	+
52	`v2cltrnslw`	Transparent laws with predictable enforcement	0.15	0.12	0.22	0.09	0.16	+
53	`v2lgqstexp`	Legislature questions officials in practice	0.15	0.17	0.21	0.12	0.09	+
54	`v2dlengage`	Engaged society	0.14	0.11	0.20	0.12	0.15	+
55	`v2x_egaldem`	Egalitarian democracy index	0.14	0.14	0.18	0.10	0.16	+ h
56	`v2psbars`	Barriers to parties	0.14	0.15	0.18	0.11	0.13	+ f
57	`v2jureform`	Judicial reform	0.14	0.08	0.25	0.15	0.08	+
58	`v2xcl_acjst`	Access to justice	0.14	0.17	0.20	0.10	0.10	+
59	`v2pepwrgen`	Power distributed by gender	0.14	0.13	0.21	0.09	0.13	+
60	`v2x_polyarchy`	Electoral democracy index	0.14	0.12	0.16	0.09	0.18	+ h
61	`v2cldmovem`	Freedom of domestic movement for men	0.14	0.10	0.19	0.10	0.15	+
62	`v2lgotovst`	Executive oversight	0.14	0.15	0.14	0.12	0.14	+
63	`v2meharjrn`	Harassment of journalists	0.14	0.03	0.21	0.12	0.18	+
64	`v2clfmove`	Freedom of foreign movement	0.14	0.15	0.15	0.10	0.14	+
65	`v2csprtcpt`	CSO participatory environment	0.13	0.11	0.23	0.06	0.14	+
66	`v2clacjstm`	Access to justice for men	0.13	0.10	0.18	0.13	0.12	+
67	`v2cltort`	Freedom from torture	0.13	0.06	0.23	0.09	0.15	+
68	`v2dlcountr`	Respect counterarguments	0.13	0.10	0.21	0.06	0.15	+
69	`v2lgoppart`	Legislature opposition parties	0.13	0.10	0.20	0.10	0.12	+
70	`v2x_egal`	Egalitarian component index	0.13	0.17	0.19	0.09	0.08	+
71	`v2clrspct`	Rigorous and impartial public administration	0.13	0.07	0.19	0.13	0.13	+
72	`v2lgcomslo`	Lower chamber committees	0.13	0.09	0.27	0.06	0.10	+
73	`v2clprptyw`	Property rights for women	0.13	0.21	0.17	0.04	0.09	+
74	`v2xeg_eqdr`	Equal distribution of resources index	0.13	0.16	0.17	0.08	0.10	+
75	`v2xcl_prpty`	Property rights	0.13	0.17	0.22	0.02	0.09	+
76	`v2csgender`	CSO women's participation	0.13	0.13	0.21	0.07	0.10	+
77	`v2exdfpphs`	HOS proposes legislation in practice	0.12	0.14	0.17	0.07	0.12	?
78	`v2cscnsult`	CSO consultation	0.12	0.10	0.23	0.06	0.11	+
79	`v2x_EDcomp_thick`	Electoral component index	0.12	0.11	0.14	0.07	0.17	+ h
80	`v2jucomp`	Compliance with judiciary	0.12	0.09	0.19	0.11	0.10	+
81	`v2elffelr`	Subnational elections free and fair	0.12	0.09	0.20	0.08	0.12	+
82	`v2pssunpar`	Subnational party control	0.12	0.14	0.16	0.08	0.09	+
83	`v2pscnslnl`	Candidate selection-national/local	0.12	0.05	0.18	0.12	0.12	+
84	`v2exembez`	Executive embezzlement and theft	0.12	0.09	0.16	0.13	0.08	+ f
85	`v2clsocgrp`	Social group equality in respect for civil liberties	0.12	0.11	0.20	0.08	0.07	+
86	`v2juhccomp`	Compliance with high court	0.11	0.09	0.20	0.07	0.11	+
87	`v2csrlgcon`	Religious organization consultation	0.11	0.11	0.17	0.08	0.10	+
88	`v2pehealth`	Health equality	0.11	0.09	0.18	0.06	0.12	+
89	`v2lgfunds`	Legislature controls resources	0.11	0.17	0.12	0.10	0.06	+
90	`v2juaccnt`	Judicial accountability	0.11	0.13	0.12	0.10	0.09	+
91	`v2psplats`	Distinct party platforms	0.11	0.02	0.22	0.03	0.16	+
92	`v2pepwrsoc`	Power distributed by social group	0.11	0.09	0.18	0.04	0.12	+
93	`v2x_genpp`	Women political participation index	0.11	0.08	0.19	0.04	0.11	+
94	`v2exrescon`	Executive respects constitution	0.10	0.07	0.17	0.08	0.10	+
95	`v2psswitch`	Party switching	0.10	0.18	0.13	0.06	0.05	+
96	`v2psprlnks`	Party linkages	0.10	0.10	0.13	0.08	0.08	+
97	`v2psorgs`	Party organizations	0.10	0.07	0.14	0.03	0.14	+
98	`v2pepwrort`	Power distributed by sexual orientation	0.10	0.02	0.24	0.02	0.11	+
99	`e_pelifeex`	Life expectancy	0.10	0.07	0.06	0.02	0.22	+
100	`v2exbribe`	Executive bribery and corrupt exchanges	0.09	0.08	0.14	0.13	0.02	+ f
101	`v2xel_frefair`	Clean elections index	0.09	0.06	0.10	0.04	0.17	+
102	`v2jupurge`	Judicial purges	0.09	0.01	0.19	0.03	0.13	+ f
103	`v2dlcommon`	Common good	0.09	0.10	0.12	0.07	0.07	+
104	`v2jucorrdc`	Judicial corruption decision	0.09	0.08	0.09	0.14	0.04	+
105	`v2xps_party`	Party system institutionalization index	0.09	0.04	0.15	0.05	0.11	+
106	`v2lgcrrpt`	Legislature corrupt activities	0.09	0.07	0.15	0.04	0.08	+
107	`v2exthftps`	Public sector theft	0.09	0.10	0.14	0.05	0.05	+ f
108	`v2elffelrbin`	Subnational elections binary	0.09	0.04	0.17	0.02	0.12	+
109	`v2ellocons`	Lower chamber election consecutive	0.08	0.08	0.06	0.05	0.14	+
110	`v2x_accex`	Elected executive index	0.08	0.07	0.16	-0.03	0.13	+

R	code	des	mean corr	corr MSCI 10	corr DMS 10	corr MSCI 5	corr DMS 5	notes
111	v2lgstafflo	Lower chamber staff	0.08	0.04	0.19	-0.00	0.09	+
112	v2lgdsadlo	Representation of disadvantaged social groups	0.08	0.06	0.13	0.03	0.08	+
113	v2clprptym	Property rights for men	0.08	0.08	0.16	-0.01	0.08	+
114	v2asuffrage	Suffrage	0.08	0.06	0.13	-0.07	0.19	+
115	v2svstterr	State authority over territory	0.07	0.05	0.06	0.05	0.14	+
116	v2lgfemleg	Lower chamber female legislators	0.07	-0.04	0.30	-0.07	0.11	+
117	v2ellocumul	Lower chamber election cumulative	0.07	0.15	-0.01	0.07	0.09	+
118	v2exremhsp	HOS removal by legislature in practice	0.07	0.05	0.12	0.11	0.02	+
119	v2msuffrage	Male suffrage	0.07	0.07	0.10	-0.08	0.20	+
120	v2pepwrses	Power distributed by socioeconomic position	0.07	0.14	0.11	-0.01	0.04	+
121	v2dlunivl	Means-tested v. universalistic policy	0.07	0.10	0.12	0.03	0.02	+
122	v2fsuffrage	Female suffrage	0.06	0.06	0.12	-0.06	0.15	+
123	v2lgello	Lower chamber elected	0.06	0.03	0.14	-0.06	0.14	+
124	v2svstpop	State authority over population	0.06	0.02	0.05	0.05	0.13	+
125	v2ex_elecleg	Legislature directly elected	0.06	0.02	0.16	-0.09	0.15	+
126	v2jupoatck	Government attacks on judiciary	0.06	-0.06	0.15	0.05	0.10	+ f
127	e_polity	Combined POLITY score	0.06	-0.01	0.04	0.08	0.12	+
128	v2jupack	Court packing	0.06	-0.04	0.11	0.09	0.07	+
129	v2psprbrch	Party branches	0.06	0.03	0.11	0.04	0.05	+
130	v2dlencmps	Particularistic or public goods	0.05	0.05	0.07	0.08	0.01	+
131	v2xel_regelec	Regional government index	0.05	-0.00	0.16	-0.01	0.06	+
132	v2excrptps	Public sector corrupt exchanges	0.05	0.01	0.13	0.03	0.03	+ f
133	v2peedueq	Educational equality	0.05	0.08	0.08	0.01	0.03	+
134	e_democ	Institutionalized democracy	0.05	-0.03	0.02	0.08	0.11	+
135	e_polcomp	Political competition	0.05	-0.02	0.02	0.08	0.11	+
136	e_exconst	Executive constraints	0.04	-0.04	0.02	0.08	0.11	+
137	e_exrec	Executive recruitment	0.04	-0.04	0.01	0.08	0.11	+
138	e_parcomp	The competitiveness of participation	0.04	-0.04	0.01	0.08	0.11	+
139	e_xrcomp	Competitiveness of executive recruitment	0.04	-0.04	0.01	0.08	0.11	+
140	e_parreg	Regulation of participation	0.04	-0.05	0.01	0.08	0.11	+
141	v2xel_elecparl	Legislative or constituent assembly election	0.02	0.02	0.05	-0.01	0.04	+
142	e_autoc	Institutionalized autocracy	0.02	-0.06	-0.01	0.07	0.10	-
143	v2psnatpar	National party control	0.02	0.11	-0.08	0.10	-0.04	+
144	v2xdd_dd	Direct popular vote index	0.02	0.08	-0.03	-0.03	0.07	+
145	e_peaveduc	Education 15+	0.02	0.01	0.01	0.08	-0.02	+
146	v2elsnlsff	Subnational election unevenness	0.00	0.08	-0.05	0.05	-0.07	+
147	v2ddnumvot	Number of popular votes this year	0.00	-0.05	0.03	0.02	0.01	?
148	v2ddplebyr	Occurrence of plebiscite this year	-0.00	-0.00	0.02	0.00	-0.02	?
149	e_miurbani	Urbanization	-0.00	0.03	-0.02	0.01	-0.03	?
150	e_mipopula	Population total	-0.01	-0.06	0.02	-0.05	0.03	?
151	e_miurbpop	Urban population	-0.02	-0.09	0.04	-0.06	0.03	?
152	v2pscomprg	Party competition across regions	-0.02	-0.15	0.02	-0.01	0.05	?
153	v2clrgunev	Regional unevenness in respect for civil liberties	-0.03	0.03	-0.03	-0.04	-0.07	-
154	e_migdpgrolns	GDP Growth (rescaled)	-0.04	-0.14	0.01	-0.12	0.06	GDP
155	e_migdpgro	GDP Growth	-0.05	-0.15	0.00	-0.11	0.06	GDP
156	e_miinflat	Inflation	-0.05	-0.14	-0.03	-0.05	0.01	-
157	v2csantimv	CSO anti-system movements	-0.06	-0.02	0.03	-0.17	-0.08	-
158	e_peedgini	Educational inequality, Gini	-0.08	-0.20	-0.05	-0.09	0.00	-
159	v2pscohesv	Legislative party cohesion	-0.09	-0.18	-0.09	-0.05	-0.05	-
160	v2lgdsadlobin	Representation of disadvantaged social groups binary	-0.10	-0.04	-0.14	-0.07	-0.14	-
161	e_migdppc	GDP per capita	-0.10	-0.23	0.06	-0.24	0.00	GDP
162	v2x_pubcorr	Public sector corruption index	-0.10	-0.11	-0.18	-0.08	-0.04	-
163	v2exdfvths	HOS veto power in practice	-0.11	-0.14	-0.15	-0.08	-0.07	-
164	v2x_execorr	Executive corruption index	-0.11	-0.12	-0.15	-0.14	-0.04	-
165	v2x_corr	Political corruption	-0.11	-0.11	-0.15	-0.14	-0.06	- h
166	e_migdppcln	GDP per capita, logged, base 10	-0.13	-0.16	-0.12	-0.23	-0.02	GDP
167	v2exdfdmhs	HOS dismisses ministers in practice	-0.14	-0.14	-0.22	-0.05	-0.14	-
168	v2exdfdshs	HOS dissolution in practice	-0.14	-0.13	-0.19	-0.16	-0.09	-
169	v2exdfcbhs	HOS appoints cabinet in practice	-0.17	-0.19	-0.21	-0.13	-0.15	-
170	stock	country stock returns	-0.20	-0.18	-0.23	-0.21	-0.16	stock

B Table of Mean Correlations for Different Aggregation Methods

This appendix contains a table comparing different methods of calculating correlation between past changes of data for a code and future stock returns. For a given study and code, the past and future data are vectors, X_{cy} and Y_{cy}, indexed by country-year pairs for which there is data. As in Appendix A, rows of the table are sorted in order of decreasing mean over studies of the total correlation. In Appendix A, the mean total correlation had the column heading "meanCorr". Here it has the heading "total", to distinguish it from the columns "within-country", "within-country, std-tied", "within-year", and "within-year, std-tied", which give the mean over studies of the different version of group correlation discussed in Section 5. As in Appendix A, the first two columns specify the rank order and the code, and the last column specifies additional notes.

R	code	total	within-country	within-country std-tied	within-year	within-year std-tied	notes
1	v2x_freexp	0.19	0.14	0.19	0.09	0.13	+
2	v2xcl_disc	0.19	0.13	0.18	0.11	0.14	+
3	v2meslfcen	0.19	0.13	0.19	0.06	0.12	+ f
4	v2mecenefm	0.19	0.13	0.19	0.08	0.09	+ f
5	v2x_freexp_thick	0.19	0.14	0.18	0.06	0.12	+ h
6	e_rol_free	0.18	0.13	0.17	0.06	0.13	+ h
7	v2clrelig	0.18	0.08	0.16	0.03	0.14	+
8	v2xcs_ccsi	0.18	0.12	0.17	0.04	0.13	+ h
9	v2clslavef	0.18	0.15	0.18	0.07	0.13	+
10	v2csrlgrep	0.18	0.08	0.15	0.07	0.15	+
11	v2cldiscw	0.18	0.12	0.17	0.10	0.11	+
12	v2xme_altinf	0.17	0.10	0.17	0.03	0.11	+
13	v2x_liberal	0.17	0.10	0.16	0.05	0.13	+ h
14	v2mecrit	0.17	0.11	0.17	0.08	0.10	+
15	v2cldiscm	0.17	0.11	0.18	0.07	0.11	+
16	v2clacfree	0.17	0.10	0.17	0.08	0.12	+
17	v2x_gencl	0.17	0.12	0.17	0.03	0.11	+
18	v2x_partip	0.17	0.06	0.17	0.04	0.13	+ h
19	v2meaccess	0.17	0.12	0.18	0.02	0.09	+
20	v2cldmovew	0.17	0.11	0.19	0.03	0.08	+
21	v2xcl_rol	0.17	0.09	0.15	0.04	0.12	+ h
22	v2juhcind	0.17	0.10	0.16	0.06	0.10	+
23	v2psparban	0.17	0.09	0.17	0.07	0.11	+ f
24	v2xcl_slave	0.17	0.12	0.15	0.07	0.12	+
25	v2cseeorgs	0.16	0.11	0.16	0.02	0.10	+
26	v2xdl_delib	0.16	0.07	0.17	0.04	0.11	+
27	v2xcl_dmove	0.16	0.09	0.16	0.01	0.09	+
28	v2x_cspart	0.16	0.06	0.16	0.05	0.12	+
29	v2juncind	0.16	0.09	0.15	0.10	0.11	+
30	v2mebias	0.16	0.10	0.16	-0.00	0.09	+ f
31	v2x_gencs	0.16	0.09	0.16	0.07	0.10	+
32	v2clstown	0.16	0.07	0.14	0.06	0.12	+
33	v2xlg_legcon	0.16	0.11	0.16	0.07	0.13	+
34	v2mecorrpt	0.16	0.10	0.15	0.04	0.10	+ f
35	v2csreprss	0.16	0.13	0.17	0.03	0.09	+
36	v2dlconslt	0.16	0.06	0.16	0.04	0.11	+
37	v2x_frassoc_thick	0.16	0.12	0.15	0.06	0.11	+
38	v2jureview	0.16	0.08	0.14	0.07	0.12	+
39	v2merange	0.16	0.07	0.15	0.03	0.10	+ f
40	v2x_libdem	0.16	0.10	0.15	0.06	0.10	+ h

R	**code**	**total**	**within-country**	**within-country std-tied**	**within-year**	**within-year std-tied**	**notes**
41	`v2dlreason`	0.15	0.07	0.15	0.05	0.09	+
42	`v2psoppaut`	0.15	0.07	0.15	0.06	0.11	+
43	`v2mefemjrn`	0.15	0.16	0.17	0.07	0.09	+
44	`v2x_partipdem`	0.15	0.09	0.15	0.04	0.09	+
45	`v2clacjstw`	0.15	0.11	0.15	0.01	0.06	+
46	`v2clkill`	0.15	0.07	0.13	0.06	0.10	+
47	`v2lginvstp`	0.15	0.09	0.15	0.06	0.12	+
48	`v2x_gender`	0.15	0.11	0.16	0.03	0.08	+ h
49	`v2x_jucon`	0.15	0.05	0.13	0.03	0.12	+ h
50	`v2clslavem`	0.15	0.07	0.13	0.05	0.10	+
51	`v2x_delibdem`	0.15	0.09	0.15	0.05	0.08	+
52	`v2cltrnslw`	0.15	0.09	0.16	0.07	0.09	+
53	`v2lgqstexp`	0.15	0.06	0.13	0.09	0.12	+
54	`v2dlengage`	0.14	0.04	0.13	0.05	0.09	+
55	`v2x_egaldem`	0.14	0.07	0.14	0.04	0.08	+ h
56	`v2psbars`	0.14	0.08	0.13	0.04	0.09	+ f
57	`v2jureform`	0.14	0.07	0.15	0.06	0.10	+
58	`v2xcl_acjst`	0.14	0.09	0.12	0.02	0.10	+
59	`v2pepwrgen`	0.14	0.08	0.14	0.04	0.07	+
60	`v2x_polyarchy`	0.14	0.09	0.13	0.04	0.08	+ h
61	`v2cldmovem`	0.14	0.06	0.14	-0.00	0.08	+
62	`v2lgotovst`	0.14	0.07	0.13	0.07	0.10	+
63	`v2meharjrn`	0.14	0.09	0.14	0.02	0.07	+
64	`v2clfmove`	0.14	0.09	0.15	0.02	0.06	+
65	`v2csprtcpt`	0.13	0.08	0.12	0.04	0.09	+
66	`v2clacjstm`	0.13	0.06	0.12	0.03	0.08	+
67	`v2cltort`	0.13	0.07	0.12	0.05	0.09	+
68	`v2dlcountr`	0.13	0.07	0.14	0.03	0.09	+
69	`v2lgoppart`	0.13	0.07	0.11	0.05	0.12	+
70	`v2x_egal`	0.13	0.05	0.10	0.02	0.10	+
71	`v2clrspct`	0.13	0.06	0.12	0.05	0.07	+
72	`v2lgcomslo`	0.13	0.08	0.13	0.05	0.09	+
73	`v2clprptyw`	0.13	0.06	0.14	0.03	0.07	+
74	`v2xeg_eqdr`	0.13	0.06	0.12	0.03	0.10	+
75	`v2xcl_prpty`	0.13	0.06	0.13	0.03	0.09	+
76	`v2csgender`	0.13	0.05	0.13	0.04	0.08	+
77	`v2exdfpphs`	0.12	0.04	0.14	0.08	0.07	?
78	`v2cscnsult`	0.12	0.09	0.15	0.01	0.07	+
79	`v2x_EDcomp_thick`	0.12	0.06	0.11	0.04	0.07	+ h
80	`v2jucomp`	0.12	0.04	0.10	0.04	0.10	+
81	`v2elffelr`	0.12	0.07	0.12	0.03	0.09	+
82	`v2pssunpar`	0.12	0.02	0.10	0.06	0.10	+
83	`v2pscnslnl`	0.12	0.01	0.08	0.09	0.13	+
84	`v2exembez`	0.12	0.02	0.12	0.04	0.07	+ f
85	`v2clsocgrp`	0.12	0.04	0.10	0.02	0.08	+
86	`v2juhccomp`	0.11	0.03	0.10	0.03	0.09	+
87	`v2csrlgcon`	0.11	0.07	0.08	0.02	0.09	+
88	`v2pehealth`	0.11	0.10	0.14	0.04	0.06	+
89	`v2lgfunds`	0.11	0.05	0.12	0.05	0.08	+
90	`v2juaccnt`	0.11	0.00	0.11	0.03	0.09	+
91	`v2psplats`	0.11	0.05	0.12	0.04	0.10	+
92	`v2pepwrsoc`	0.11	0.04	0.10	0.02	0.08	+
93	`v2x_genpp`	0.11	0.10	0.13	-0.01	0.02	+
94	`v2exrescon`	0.10	0.04	0.09	0.01	0.07	+
95	`v2psswitch`	0.10	0.05	0.05	0.05	0.10	+
96	`v2psprlnks`	0.10	0.06	0.11	0.01	0.05	+
97	`v2psorgs`	0.10	0.07	0.12	0.02	0.04	+
98	`v2pepwrort`	0.10	0.07	0.11	0.01	0.04	+
99	`e_pelifeex`	0.10	0.10	0.13	-0.00	0.05	+
100	`v2exbribe`	0.09	0.04	0.10	0.03	0.05	+ f
101	`v2xel_frefair`	0.09	0.04	0.07	-0.02	0.07	+
102	`v2jupurge`	0.09	0.07	0.09	0.03	0.05	+ f
103	`v2dlcommon`	0.09	0.05	0.10	0.03	0.05	+
104	`v2jucorrdc`	0.09	0.02	0.08	0.05	0.08	+
105	`v2xps_party`	0.09	0.07	0.11	0.05	0.05	+
106	`v2lgcrrpt`	0.09	0.00	0.08	0.06	0.07	+
107	`v2exthftps`	0.09	0.02	0.09	0.06	0.09	+ f
108	`v2elffelrbin`	0.09	0.06	0.07	0.04	0.06	+
109	`v2ellocons`	0.08	0.10	0.09	0.08	0.06	+
110	`v2x_accex`	0.08	0.04	0.09	0.03	0.04	+

R	code	total	within-country	within-country std-tied	within-year	within-year std-tied	notes
111	v2lgstafflo	0.08	0.10	0.09	0.02	0.05	+
112	v2lgdsadlo	0.08	0.04	0.10	-0.01	0.02	+
113	v2clprptym	0.08	0.01	0.05	0.00	0.07	+
114	v2asuffrage	0.08	0.06	0.09	-0.00	0.02	+
115	v2svstterr	0.07	0.07	0.04	0.04	0.03	+
116	v2lgfemleg	0.07	0.11	0.08	-0.02	-0.02	+
117	v2ellocumul	0.07	0.08	0.08	0.08	0.06	+
118	v2exremhsp	0.07	0.08	0.11	0.02	0.02	+
119	v2msuffrage	0.07	0.05	0.08	0.00	0.03	+
120	v2pepwrses	0.07	-0.02	0.04	0.02	0.06	+
121	v2dlunivl	0.07	0.00	0.08	0.05	0.04	+
122	v2fsuffrage	0.06	0.05	0.07	-0.01	0.02	+
123	v2lgello	0.06	0.04	0.07	-0.00	0.03	+
124	v2svstpop	0.06	0.03	0.04	0.06	0.03	+
125	v2ex_elecleg	0.06	0.05	0.07	0.00	0.03	+
126	v2jupoatck	0.06	0.03	0.06	0.00	0.01	+ f
127	e_polity	0.06	0.06	0.06	0.06	0.02	+
128	v2jupack	0.06	0.03	0.06	-0.01	0.02	+
129	v2psprbrch	0.06	0.06	0.08	0.02	0.01	+
130	v2dlencmps	0.05	-0.03	0.04	0.01	0.06	+
131	v2xel_regelec	0.05	0.01	0.05	0.02	0.03	+
132	v2excrptps	0.05	-0.02	0.04	0.01	0.03	+ f
133	v2peedueq	0.05	0.03	0.07	-0.02	0.02	+
134	e_democ	0.05	0.05	0.05	0.06	0.01	+
135	e_polcomp	0.05	0.05	0.05	0.06	0.01	+
136	e_exconst	0.04	0.04	0.05	0.04	0.01	+
137	e_exrec	0.04	0.04	0.05	0.05	0.01	+
138	e_parcomp	0.04	0.05	0.05	0.05	0.01	+
139	e_xrcomp	0.04	0.03	0.04	0.05	0.01	+
140	e_parreg	0.04	0.04	0.04	0.01	0.00	+
141	v2xel_elecparl	0.02	0.02	0.02	0.01	0.02	+
142	e_autoc	0.02	0.03	0.03	-0.03	-0.00	-
143	v2psnatpar	0.02	-0.01	0.02	0.03	0.02	+
144	v2xdd_dd	0.02	0.01	0.03	-0.02	0.01	+
145	e_peaveduc	0.02	0.06	0.05	0.03	0.01	+
146	v2elsnlsff	0.00	-0.02	-0.01	-0.01	0.01	+
147	v2ddnumvot	0.00	0.00	0.00	-0.02	-0.02	?
148	v2ddplebyr	-0.00	-0.00	0.00	-0.01	0.00	?
149	e_miurbani	-0.00	-0.06	0.01	0.06	0.05	?
150	e_mipopula	-0.01	-0.11	-0.05	0.02	-0.00	?
151	e_miurbpop	-0.02	-0.04	0.06	0.01	-0.01	?
152	v2pscomprg	-0.02	-0.03	-0.02	-0.04	-0.01	?
153	v2clrgunev	-0.03	-0.02	-0.03	-0.02	-0.05	-
154	e_migdpgrolns	-0.04	-0.08	-0.06	-0.03	0.00	GDP
155	e_migdpgro	-0.05	-0.09	-0.06	-0.03	0.00	GDP
156	e_miinflat	-0.05	-0.04	-0.01	-0.03	-0.08	-
157	v2csantimv	-0.06	-0.06	-0.06	-0.04	-0.04	-
158	e_peedgini	-0.08	-0.13	-0.12	-0.09	-0.12	-
159	v2pscohesv	-0.09	-0.05	-0.06	-0.03	-0.08	-
160	v2lgdsadlobin	-0.10	-0.07	-0.13	-0.01	-0.06	-
161	e_migdppc	-0.10	-0.02	-0.05	-0.15	-0.19	GDP
162	v2x_pubcorr	-0.10	-0.01	-0.09	-0.05	-0.10	-
163	v2exdfvths	-0.11	-0.08	-0.13	-0.05	-0.04	-
164	v2x_execorr	-0.11	-0.02	-0.10	-0.05	-0.07	-
165	v2x_corr	-0.11	0.02	-0.11	-0.06	-0.09	- h
166	e_migdppcln	-0.13	-0.06	-0.06	-0.13	-0.15	GDP
167	v2exdfdmhs	-0.14	-0.10	-0.16	-0.07	-0.07	-
168	v2exdfdshs	-0.14	-0.08	-0.15	-0.10	-0.09	-
169	v2exdfcbhs	-0.17	-0.09	-0.18	-0.10	-0.10	-
170	stock	-0.20	-0.27	-0.26	-0.08	-0.11	stock

C Table of High Level and Active Codes

The table below lists all codes which are either in the list of high level codes given in Table 5 or are active codes (codes with non-zero regression coefficient) for the constrained multiple regression of stocks again all good or bad codes, reported on in the last line of Table 8. For each code, we give the description and indicate with asterisks whether the code is high level and which studies it is active for.

N	code	des	inHL	in1	in2	in3	in4
1	v2meslfcen	Media self-censorship				*	*
2	v2x_freexp_thick	Expanded freedom of expression index	*				
3	e_rol_free	Civil liberties and rule of law index	*				
4	v2xcs_ccsi	Core civil society index	*				
5	v2clslavef	Freedom from forced labor for women		*	*		
6	v2x_liberal	Liberal component index	*				
7	v2x_partip	Participatory component index	*				
8	v2cldmovew	Freedom of domestic movement for women			*		
9	v2xcl_rol	Equality before the law and individual liberty index	*				
10	v2mebias	Media bias					*
11	v2clstown	State ownership of economy			*	*	*
12	v2mecorrpt	Media corrupt				*	
13	v2jureview	Judicial review		*		*	
14	v2x_libdem	Liberal democracy index	*				
15	v2dlreason	Reasoned justification		*			
16	v2mefemjrn	Female journalists			*	*	*
17	v2clacjstw	Access to justice for women		*			
18	v2lginvstp	Legislature investigates in practice				*	
19	v2x_gender	Women political empowerment index	*				
20	v2x_jucon	Judicial constraints on the executive index	*				
21	v2x_egaldem	Egalitarian democracy index	*				
22	v2jureform	Judicial reform			*	*	
23	v2pepwrgen	Power distributed by gender		*			
24	v2x_polyarchy	Electoral democracy index	*				
25	v2lgcomslo	Lower chamber committees			*		
26	v2clprptyw	Property rights for women		*			
27	v2csgender	CSO women's participation			*		
28	v2x_EDcomp_thick	Electoral component index	*				
29	v2csrlgcon	Religious organization consultation		*	*	*	
30	v2pehealth	Health equality			*		*
31	v2psswitch	Party switching		*		*	
32	e_pelifeex	Life expectancy		*		*	*
33	v2exbribe	Executive bribery and corrupt exchanges		*		*	
34	v2ellocons	Lower chamber election consecutive					*
35	v2lgstafflo	Lower chamber staff			*		
36	v2svstterr	State authority over territory		*		*	*
37	v2lgfemleg	Lower chamber female legislators		*	*		*
38	v2ellocumul	Lower chamber election cumulative		*	*	*	*
39	v2msuffrage	Male suffrage					*
40	v2jupoatck	Government attacks on judiciary			*	*	
41	v2peedueq	Educational equality		*			
42	e_parreg	Regulation of participation			*	*	*
43	e_autoc	Institutionalized autocracy		*	*		
44	v2psnatpar	National party control		*		*	
45	v2xdd_dd	Direct popular vote index		*			*
46	e_peaveduc	Education 15+		*		*	
47	v2clrgunev	Regional unevenness in respect for civil liberties					*
48	e_miinflat	Inflation		*	*	*	
49	v2csantimv	CSO anti-system movements				*	*
50	e_peedgini	Educational inequality, Gini		*	*	*	
51	v2pscohesv	Legislative party cohesion		*			
52	v2lgdsadlobin	Representation of disadvantaged social groups binary			*	*	*
53	v2x_execorr	Executive corruption index				*	
54	v2x_corr	Political corruption	*				
55	v2exdfdshs	HOS dissolution in practice				*	
56	v2exdfcbhs	HOS appoints cabinet in practice		*	*	*	

D Definitions from Linear Algebra

In this appendix we collect a few basic definition we will need from the subject of linear algebra. Our goal is to give a heads up of terminology for readers who already have some familiarity with basic operations with matrices and vectors. The only place in this paper where we use anything but the simplest terminology from this appendix is in Appendix E.8 through E.10.

A *vector space* is a set of elements, called vectors, which can be added together or multiplied by a real number to produce another element of the set. The definition of a vector space requires that the operations of addition and multiplication obey certain natural axioms. The general definition of a vector space allows the real numbers to be replaced by a general *field* of numbers, called the field of *scalars*. The act of multiplication of a scalar by a vector is called *scalar multiplication*. For example, the space $\mathbb{R}^n$ of column vectors with n components is a vector space.

A *subspace* V of larger vector space (which will always be $\mathbb{R}^n$ in this paper) is a subset of that space which has the property that all multiples of a vector in V by a real number and all sums of vectors in V belong to V.

A *linear map* from a vector space V to a vector space W is a function L from V to W which takes addition to addition and multiplication to multiplication, i.e. so that $L(v+v') = L(v)+L(v')$ and $L(rv) = rL(v)$ for v and v' vectors in V and r a real number. A linear map from $\mathbb{R}^k$ to $\mathbb{R}^n$ can be identified with an $n \times k$ matrix X so that $L(\beta) = X\beta$ for any vector β in $\mathbb{R}^k$. A linear combination of a set of vectors X^1, ..., X^k in $\mathbb{R}^n$ with weights β_1, ... β_k in $\mathbb{R}$ is the sum of the vectors multiplied by their corresponding weights. Letting X be the $n \times k$ matrix whose columns are the vectors X^1, ... , X^k in $\mathbb{R}^n$ and β be the column vector with components β_1, ..., β_k, the linear combination may be written as a matrix multiplication:

$$X\beta = \beta_1 X^1 + ... + \beta_k X^k. \tag{21}$$

The set of all such linear combination of the vectors $\{X^l\}_{l=1}^k$ is a subspace of $\mathbb{R}^n$ called the *span* of the vectors:

$$\text{Span}(\{X^l\}_{l=1}^k) \quad = \quad \{\beta_1 X^1 + ... + \beta_k X^k \in \mathbb{R}^n;\ \beta_1, ..., \beta_k \in \mathbb{R}\}. \tag{22}$$

The vectors $\{X^1\}_{l=1}^k$ are called *linearly independent* if no non-zero linear combination of them vanishes, i.e. if the only solution of Eq. 21 is the vector β of all zeros. If the vectors are linearly independent, then every vector in their span can be written as a linear combination for a unique choice of β. In that case, $\{X^l\}_{l=1}^k$ is called a *basis* for the subspace that they span.

A vector space is finite dimensional if it has a finite basis. Any two bases of a finite dimensional vector space have the same number of elements, called its *dimension*.

The *sum of subspaces* V and W of $\mathbb{R}^n$ is the subspace, $V + W$ consisting of all sums of a vector in V and a vector in W. This sum is called a *direct sum*, written $V \oplus W$, if the zero vector is the only vector in common to V and W. Any vector in the direct sum $V \oplus W$ can be written uniquely as $v + w$, where v is in V and w is in W.

D.1 Orthogonality

The *dot product* of vectors v and w in $\mathbb{R}^n$ is the sum of the product of their components:

$$v \cdot w \;=\; \sum_{i=1}^{n} v_i\, w_i. \tag{23}$$

The length of v is the square root of the dot product of v with itself,

$$||v|| \;=\; \sqrt{v \cdot v}. \tag{24}$$

The dot product of v and w is the product of their lengths times the cosine of the angle between them:

$$v \cdot w = \cos(\theta)||v||\,||w||. \tag{25}$$

Vectors v and w in $\mathbb{R}^n$ are called *orthogonal* if their dot product is zero. This means that either one of the vectors is the zero vector or the two vectors are perpendicular.

A linear map from one vector space to another that preserves length is call an *orthogonal transformation.* This generalizes the notion of rotation.

The *orthogonal projection* of $\mathbb{R}^n$ onto a subspace V is the *linear map* Π_V from $\mathbb{R}^n$ to V which leaves vectors in V unchanged and takes vectors orthogonal V to the zero vector. For a given vector $Y \in \mathbb{R}^n$, $\Pi_V(Y)$ is called the orthogonal projection of Y onto V. If V is the span of the vectors $\{X^l\}_{l=1}^k$, then the projection matrix may be written[12]

$$\Pi_{\mathrm{Span}(\{X^l\}_{l=1}^k)} = X(X^TX)^{-1}X^T. \tag{26}$$

Every finite dimensional subspace V of $\mathbb{R}^n$ has an *orthonormal basis* $\{u^l\}_{l=1}^k$. This means that the vectors each have unit length and are orthogonal to each other. This can be summarized by saying that the matrix U^TU is the $k \times k$ identity matrix. Multiplication by the matrix $U = [u^1 ... u^k]$ is an orthogonal transformation from $\mathbb{R}^k$ to V. The inverse

[12] Eq. 26 assumes that the matrix X^TX is invertible. This is equivalent to the condition that the columns of X are *linearly independent*, i.e. $X\beta$ only vanishes when β is the zero vector. Our discussion of regression below will assume independence.

of this transformation is multiplication by the transpose U^T.

An orthonormal basis $\{u^l\}_{l=1}^k$ can be constructed from a generic basis $\{v^l\}_{l=1}^k$ by the *Gramm-Schmidt orthonormalization process.* This Gramm-Schmidt basis is uniquely characterized by the fact that u^1 is a positive multiple of v^1 and the dot product of u^k with v^k is positive for all k.

E Elements of Probability and Statistical Theory

For an individual flip of a coin, it would be an incredibly complex problem to decide in advance whether it comes up heads or tails. But it is common sense notion that there is about a fifty percent chance of heads and a fifty percent chance of tails. We think of a coin as "fair" if those chances are exactly fifty-fifty. We can can test a coin by observing a large number of flips. We modify our assumption that the coin is fair if the observed deviation from fifty-fifty is substantial. Similarly, we can assume, test, and modify the assumption that each coin flip is independent of environmental factors and the result of previous flips.

Probability theory provides a general framework to model our assumptions. We may model anything. Our "coin" can metaphorically ask the discrete question of whether the change of an indicator of the return of a stock market is positive or negative. Or we can ask, as we have in this paper, how the (future) returns of stock markets are related to past changes of indicators.

A probability model, also called a probability "distribution", is a theoretical construct from which we think observed data is *drawn.* The word "drawn" comes from the paradigm example of drawing balls from an urn, which generalizes a simple coin flip to the case of several discrete outcomes. The model specifies the probability of possible outcomes. For example, the probability of drawing a white ball from an urn with three white and five black balls is 3/8.

We need not believe that a model is a full description of reality, just that it is provides a useful approximation to salient elements of the real world. We sometimes call a theoretical distribution the *population* distribution and data selected from that distribution as *samples.* (In the medical and social science fields, the "population" is often envisaged as a large cohort of people from which the researcher has taken multiple samples.) The population distribution model may be described by and depend on one or more unknown parameters, for example the probability that a coin flip will be heads.

Statistical theory provides a framework for making precise statements (i.e. educated guesses) about the parameters of a population model based on a sample of data. The statements may be: *point estimates* – specific estimates of population parameters; *confidence intervals* – ranges of values for parameter values; or *hypothesis tests* – tests of specific statements about the model.

E.1 Basics of Probability Distributions

A *probability distribution* on a *random variable* X taking value in a set $\mathcal{S}$, called the *sample space*, assigns a probability, $P(X \in \mathcal{S})$, which is between 0 and 1, to each subset of $\mathcal{S}$. If $\mathcal{S}$ is a *discrete set*, a probability distribution is determined by a *probability mass function*, which specifies the probability of each element of $\mathcal{S}$. The probability of a general subset of $\mathcal{S}$ is then just the sum of the probabilities of the individual elements in the subset. If X takes continuous values, the probability distribution may be specified by a *probability distribution function* (PDF), which is a function on the sample space. The probability of a general subset of $\mathcal{S}$ in this case is the integral of the PDF over the set.

$$P(X \in \mathcal{S}) \;=\; \begin{cases} \sum_{x\in\mathcal{S}} p(x) & \textit{for } \mathcal{S} \textit{ discrete with probability mass function } p \\ \int_{x\in\mathcal{S}} p(x) & \textit{for } \mathcal{S} \textit{ continuous with PDF } p. \end{cases} \tag{27}$$

Above, we have followed the convention of using an upper case letter when referring to the random variable in the abstract and a lower case letter to represent an individual *sample* from the distribution.

There is an extended definition of the integral for which the second expression in Eq. 27 (and Eq. 28 below) applies to both the discrete and continuous cases. With this extended definition in mind, we can refer to the probability mass function p or the probability density function p above simple as the *distribution function*.

For f a function on $\mathcal{S}$, the *expectation value*, or *mean*, of $f(X)$ is the probability-weighted average of f over all values of a *sample* x.

$$E[f(X)] \;=\; \begin{cases} \sum_{x\in\mathcal{S}} p(x)f(x) & \textit{for } \mathcal{S} \textit{ discrete with probability mass function } p \\ \int_{x\in\mathcal{S}} p(x)f(x) & \textit{for } \mathcal{S} \textit{ continuous with PDF } p. \end{cases} \tag{28}$$

The *variance* of X is the expectation value of the square of the difference of X from its mean; and the *standard deviation* of X is the square root of the variance:

$$\begin{aligned} \mathrm{var}(X) &= E[(X-\mu_X)^2] = \int_{x\in\mathcal{S}} p(x)(x-\mu_X)^2, \\ \mathrm{std}(X) &= \sqrt{\mathrm{var}(X)}, \end{aligned} \tag{29}$$

where

$$\mu_X = E[X] \tag{30}$$

is the mean of X.

If the value of X are real numbers, we define the *cumulative probability distribution* to be the function of a top value x, which gives the probability that X is below x:

$$P(X \leq x) \;\; = \;\; \int_{-\infty}^{x} dx' \; p(x'). \tag{31}$$

Here we have written the differential dx' of the dummy variable explicitly.

E.2 Joint Distributions

A *joint distribution* for two random variables X and Y, taking values in $\mathcal{S}_X$ and $\mathcal{S}_Y$ respectively, is just a probability distribution on the product space $\mathcal{S}_X \times \mathcal{S}_Y$. It is determined by a *joint distribution function* $p(x, y)$.

We say that X and Y are *independent* if the joint distribution is just the product of separate distributions for X and Y[13]:

$$p(x, y) \;\; = \;\; p(x)\, p(y). \tag{32}$$

An opposite extreme to independence is when Y is a function g of X. In that case[14]

$$p(x, y) \;\; = \;\; \begin{cases} p(x) & \text{for } y = g(x) \\ 0 & \text{for } y \neq g(x). \end{cases} \tag{33}$$

The *marginal distribution* for Y is the distribution of the variable Y by itself which has the property that, for any function F of Y, the expectation value of F using the marginal distribution is the same as the expectation value using the joint distribution of X and Y:

$$E[F] = \int_y p(y) F(y) = \int_{(x,y)} p(x, y) F(y). \tag{34}$$

Above, we have indicated the variables being integrated over in the subscripts to the integral signs, but we have not written down the differentials because the appropriate symbols would vary depending on circumstances. We know spell out some examples.

If X takes values in $\mathbb{R}^k$ and Y in $\mathbb{R}^l$ and both variables are unconstrained (so samples (x, y) may range over the whole of $\mathbb{R}^k \times \mathbb{R}^l$, not just a lower dimensional subset as would

[13] In Eq. 32, we have abused notation and used the same symbol p for functions that are distinguished by the variables that they depend upon.

[14] A subtle explanation can be given to spell out how Eq. 33 is not an abuse of notation. This would include, for example, a discussion of how the probability distribution $p(x, y)$ has *support* on a set that is in one-one correspondence with $\mathcal{S}_X$, namely the graph of g, which is the subset of $\mathcal{S}_X \times \mathcal{S}_Y$ consisting of all pairs $(x, g(x))$.

be the case if, for example, one was a function of the other), then

$$E[F] \;=\; \int d^l y\, p(y) F(y) = \int d^k x\, d^l y\, p(x,y) F(y). \tag{35}$$

So

$$p(y) \;=\; \int d^k x\, p(x,y). \tag{36}$$

If X takes values in $\mathbb{R}^k$ and y in $\mathbb{R}^l$ is a function of g of x, then:

$$E[F] = \int d^l y\, p(y) F(y) = \int d^k x\, p(x) F(g(x)). \tag{37}$$

In particular, if g is invertible, i.e. y is a change of variables of x, then l must equal k. If k is one, then

$$p(y) \;=\; p(x) \left(\frac{dx}{dy} \right) = p(x) \left(\frac{dg(x)}{dx} \right)^{-1}. \tag{38}$$

When k is greater than one, the derivative on the right is replaced by the (absolute value of the) determinant of the Jacobian matrix of partial derivatives, $\left(\frac{\partial g_i}{\partial x^j}\right)$.

E.3 Covariance and Correlation

Given two random variables X and Y, the *covariance* of X and Y is the expectation value of the product of their displacements from their respective means:

$$\mathrm{cov}(X,Y) \;=\; E[(X-\mu_X)(Y-\mu_Y)] = E[X\,Y] - E[X]\,E[Y]. \tag{39}$$

When X equals Y, the covariance of X and Y reduces to the variance of X. When X and Y are independent, the covariance of X and Y equals zero.

The *correlation*, or more formally the Pearson product-moment correlation coefficient, of X and Y is their covariance divided by the product of their standard deviations:

$$\mathrm{corr}(X,Y) \;=\; \frac{\mathrm{cov}(X,Y)}{\mathrm{std}(X)\,\mathrm{std}(Y)} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\mathrm{std}(X)\,\mathrm{std}(Y)}. \tag{40}$$

The correlation is undefined when the standard deviation of X or Y vanishes, i.e. if either variable is constant. If X and Y are independent (and neither is constant), the correlation vanishes.

The *Z-score* of X is the random variable that is a linear transformation of X having mean zero and standard deviation one:

$$Z_X = \frac{X - \mu_X}{\text{std}(X)}. \tag{41}$$

The correlation of X and Y equals that expectation of the product of their Z-scores:

$$\text{corr}(X, Y) \quad = \quad E[Z_X \, Z_Y]. \tag{42}$$

We will now explain the geometric reason why the correlation always lies in the range between -1 and 1. The picture here will be useful later when we derive a probability distribution for hypothesis testing correlation.

The Cauchy-Schwarz inequality for vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ says that the square of their dot product is smaller than or equal to the product of their lengths squared:

$$\begin{aligned} (x \cdot y)^2 &\leq ||x||^2 \, ||y||^2, \text{ where} && (43) \\ x \cdot y &= \sum_{k=1}^{n} x_k \, y_k = \text{ dot product of } x \text{ and } y, \text{ and} && (44) \\ ||x|| &= \sqrt{x \cdot x} = \left(\sum_{k=1}^{n} x_k^2 \right)^{1/2} = \text{length of } x. && (45) \end{aligned}$$

The geometric interpretation of this is that the dot product of x and y equals the product of the lengths of x and y times the cosine of the angle between them (in n-dimensional space):

$$x \cdot y \quad = \quad \cos(\theta) \, ||x|| \, ||y||. \tag{46}$$

The generalization of the Cauchy-Schwarz inequality to probability distributions says that the square of the expectation of a product is smaller than or equal to the product of the expectation of squares:

$$(E[X \, Y])^2 \leq E[X^2] \, E[Y^2]. \tag{47}$$

The fact that the absolute value of correlation is less than or equal to one follows by applying the previous equation to Eq. 42 for correlation and using the fact the expectation value of the squared Z-scores Z_X^2 and Z_Y^2 are both one.

E.4 Multiple Sampling and The Normal Distribution

Assuming that samples are all drawn independently from a probability distribution for a single sample, the probability distribution for a *multiple sample* $x = (x_1, ..., x_n)$ with n

samples is the product of the distribution functions for each individual sample:

$$p(x_1, ..., x_n) \quad = \quad p(x_1)p(x_2)...p(x_n) \tag{48}$$

The mean and standard deviation of the average of the random variables $X_1, ..., X_n$ with this distribution are

$$\begin{aligned} E\left[(X_1 + ... + X_n)/n\right] &= E[X_1] \\ \text{std}\left((X_1 + ... + X_n)/n\right) &= \frac{1}{\sqrt{n}}\ \text{std}(X_1). \end{aligned} \tag{49}$$

The *Central Limit Theorem* says that quite generally the distribution of the average tends to a *normal distribution*, otherwise known as a bell curve, for large sample size n. Even more strongly, whenever some very mild assumptions hold, the average of a large number of different random effects tends to look like a normal distribution. This is why the normal distribution is often a reasonable model for background "noise" in statistical estimation and hypothesis testing.

The normal distribution with mean μ and standard deviation σ is the probability distribution on the real line, with PDF

$$N(x; \mu, \sigma) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{50}$$

Since it is a PDF, the integral of $N(x; \mu, \sigma)$ over x is equal to 1. The integral of $xN(x; \mu, \sigma)$ is the mean μ and the integral of $(x - \mu)^2 N(x; \mu, \sigma)$ equals the variance σ^2. The *standard normal distribution* is the special case when the mean vanishes and the standard deviation is one.

The PDF for n independent samples from the above normal distribution is

$$N(x_1, ..., x_n; \mu, \sigma) \quad = \quad \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{\sum_{k=1}^{n}(x_k-\mu)^2}{2\sigma^2}}. \tag{51}$$

In our derivation below of the distribution of sample correlations, we will use the fact that the exponent in Eq. 51 is proportional to the squared length of $(\tilde{x}_1, ..., \tilde{x}_n)$, where $\tilde{x}_k$ is just x_k shifted by μ.

The above is a special case of the multivariate normal distribution for an n-dimensional vector x with mean μ and covariance Σ (a positive definite symmetric $(n \times n)$ matrix), which has PDF

$$N(x; \mu, \Sigma) \quad = \quad \det\left(2\pi\, \Sigma\right)^{-1/2} e^{-(x-\mu)^T \Sigma^{-1}(x-\mu)/2}. \tag{52}$$

E.5 Sample Statistics

Suppose we draw a multiple sample from some underlying population distribution whose parameters are unknown. A *sample statistic* is a function of sample data which does not depend on the population parameters. We say that the statistic is an *estimator* if it is designed to allow us to make a guess at a parameter describing the underlying population.

One fundamental sample statistic is the *sample mean*. Given a sample $x = (x_1, ...x_n)$ from some underlying population distribution on some set of real numbers, the sample mean is just the average

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k. \tag{53}$$

$$\tag{54}$$

This is an *unbiased estimator* of the population mean, which means that the expectation value over all samples of the sample mean equals the mean of the true population. Note that the sample mean equals the expectation of the values x_k, considered as a function of a random variable k, which has equally weighted probability to take the values 1 through k. We can think of this discrete distribution as a crude estimate of the population distribution.

Another fundamental statistic is the *uncorrected sample standard deviation*:

$$s_x^U = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2}. \tag{55}$$

The square of this is the uncorrected sample variance, which is a biased estimator of population variance because the expectation value of it over all samples equals the variance of the population distribution times a bias factor of $(n-1)/n$. *Bessel's correction* to the sample variance is to multiply the uncorrected sample variance by $n/(n-1)$. The sample variance with this correction is an unbiased estimate of the population variance.

The common definition of (corrected) sample standard deviation is the uncorrected sample standard deviation time the square root of $n/(n-1)$. The corrected version generally comes much closer to being an unbiased estimator of population standard deviation than the uncorrected sample standard deviation.

The (uncorrected) *Z-score* is the vector with mean zero and standard deviation one obtained from x by subtracting the mean and dividing by (uncorrected) standard deviation:

$$(Z^{(x)})_k = (x_k - \bar{x})/s_x^U. \tag{56}$$

It is will be convenient for us to define the unit-length Z score to be normalized to have length one:

$$\tilde{Z}^{(x)} \;=\; \frac{1}{\sqrt{n}} Z^{(x)}. \tag{57}$$

In addition to having unit length, the dot product of $\tilde{Z}^{(x)}$ with the vector $\underline{\mathbf{1}} = (1, 1, ..., 1)$ vanishes. So $\tilde{Z}^{(x)}$ belongs to the $(n-2)$-dimensional sphere, $\mathbf{S}^{n-2}_{\underline{\mathbf{1}}}$, consisting of vectors of length one in the $(n-1)$-dimensional subspace of $\mathbb{R}^n$ orthogonal to the vector $\underline{\mathbf{1}}$.

Our final example of a sample statistic is the *sample correlation* between two data vectors. In this paper, we are interested in the case where one vector is a vector of past changes of a V-Dem indicator and the other is a vector of future stock returns. Let x and y be vectors in $\mathbb{R}^n$. The *sample correlation* is the dot product of the unit-length Z-scores for x and y:

$$\operatorname{corr}(x, y) = \tilde{Z}^{(x)} \cdot \tilde{Z}^{(y)} \;=\; \frac{\sum_{k=1}^{n} (x_k - \bar{y})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^{n} (y_k - \bar{y})^2}}. \tag{58}$$

Sample correlation is a biased estimator of population correlation, although the bias decreases with the number of samples and vanishes in the case when population correlation vanishes, which is the case we consider for hypothesis testing.

E.6 Hypothesis Testing of Correlation

Hypothesis testing of the correlation between n-dimensional samples x_{obs} and y_{obs} asks the question: *Is the sample correlation, $c_{obs} = \operatorname{corr}(x_{obs}, y_{obs})$, likely to represent a real phenomenon or does it just amount to random noise?* Our model of "random noise" is the *null hypothesis* that the x vector is the given value[15] x_{obs} and that y consists of n independent samples drawn from a normal distribution. The *P-value* of the observed correlation is the probability that the correlation of a random sample is at least as big as the observed correlation[16]. If that probability is very low, we say that we are justified in *rejecting* the null hypothesis; i.e., we have good reason to doubt that the observation was just a random effect.

[15] The distribution of correlation turns out to be independent of x_{obs}. So it is the same whether we fix x or allow x to be chosen randomly from some distribution.

[16] What we have described above is a *one-sided* test appropriate for the case when we are asking if a correlation is genuinely positive. In case we are interested in negative correlation, we would look at the probability that the random sample had correlation smaller than the observed correlation. If the focus of our test was the absolute value of the correlation, we would perform a two-sided test.

E.6.1 Hypothesis Testing Background

It is now standard in the statistical literature to perform hypothesis testing of correlation by calculating a transform of it called a *t-statistic*. When the null hypothesis is true, the t-statistic is distributed by Student's *t-distribution*. Introducing the t-statistic obscures the meaning in terms of correlation itself. Below we will derive the distribution of correlation under the null hypothesis directly by a simple geometric argument. We also prove that the large n limit of this distribution is a normal distribution.

The earliest paper to present the distribution of sample correlation when the population correlation vanishes is actually over a century old (Student 1908). The case when population correlation is non-zero was dealt with in (Soper 1913)). R.A. Fisher subsequently published several papers that put the subject on a firm foundation (Fisher 1915, Fisher 1921, Fisher 1924). In (Hotelling 1953), Hotelling states that "the best present-day usage in dealing with correlation coefficients is based on R.A. Fisher's chapter on the subject (Fisher 1950)." Hotelling simplifies and makes the theory more exact and rigorous.

The distribution of sample correlation is now embedded inside of a collection of tools that it is now standard to apply by rote. That machine is designed to handle not just testing against the null hypothesis of zero population correlation, but also parameter estimation of the value of population correlation, for which one need to use the distribution of sample correlation for non-zero population correlation. Reference (Hogben 1968) explains that this distribution is a Q distribution and mentions that "J. N. K. Rao and an unidentified person have pointed out that the distribution of [the correlation coefficient squared] can be obtained as a special case of the conditional distribution of the multiple correlation coefficient for the multivariate normal (Rao 1965, p. 509)." See also (Rao 2001). Exploring distributions for multiple correlation coefficients is still a subject of active research. We will have more to say about it when we discuss regression later in the appendix.

E.6.2 Derivation of the Distribution of Correlation Under the Null Hypothesis

We now present the promised simple geometric argument to derive the distribution of sample correlation under the null hypothesis. We fix an n-dimension vector x and let y be an n-dimensional vector whose components are independently sampled from a normal distribution with some mean μ_Y and standard deviation σ_Y.

The argument starts by observing that the unit-length Z-score $\tilde{Z}^{(y)}$ is *uniformly distributed* over the $(n-2)$-sphere $\mathbf{S}^{n-2}_{\underline{\mathbf{1}}}$ to which it belongs, i.e. the PDF for $\tilde{Z}^{(y)}$ is independent of μ_Y and σ_Y and equal to a constant on the sphere. (The constant is the inverse of the volume of the sphere). This follows because, with the mean subtracted off, the PDF in Eq. 51 is rotationally invariant[17].

The dimensionality of the sphere to which the Z-score $\tilde{Z}^{(y)}$ belongs is the number of pieces of information in y, other than the mean and standard deviation, which the correlation does not depend on. This dimension is called the number of *degrees of freedom*,

$$\nu = n - 2. \tag{59}$$

Now we are ready to derive the distribution for the sample correlation. By rotational invariance, we may fix $\tilde{Z}^{(x)}$ to the north pole. By Eq. 58, the correlation of x and y is the cosine of the angle θ between y and the north pole, which we call c:

$$\mathrm{corr}(x, y) = \tilde{Z}^{(x)} \cdot \tilde{Z}^{(y)} = \cos(\theta) = c. \tag{60}$$

It is now useful to introduce generalized spherical coordinates on the sphere $\mathbf{S}^{\nu}$ [18]. The first coordinate is the latitude θ, which goes from 0 (the north pole) to π (the south pole). The other coordinates specify a point on the $(n-3)$-dimensional sphere at latitude θ, which has radius $\sin(\theta)$.

As a warm up exercise, let us show how to calculate the (ν-dimensional) volume of $\mathbf{S}^{\nu}_{s}$, the ν-sphere of radius s. This volume scales as the ν'th power of s:

$$\mathrm{Vol}(\mathbf{S}^{\nu}_{s}) = s^{\nu}\, \mathrm{Vol}(\mathbf{S}^{\nu}_{1}). \tag{61}$$

We can calculate $\mathrm{Vol}(\mathbf{S}^{\nu}_{1})$ using generalized spherical coordinates:

$$\begin{aligned} \mathrm{Vol}(\mathbf{S}^{\nu}_{1}) &= \int_0^{\pi} d\theta\, \mathrm{Vol}\left(\mathbf{S}^{\nu-1}_{\sin(\theta)}\right) \\ &= \mathrm{Vol}(\mathbf{S}^{\nu-1}_{1}) \int_0^{\pi} d\theta \sin(\theta)^{\nu-1} \\ &= \mathrm{Vol}(\mathbf{S}^{\nu-1}_{1}) \int_{c=-1}^{1} dc\, (1-c^2)^{(\nu-2)/2}. \end{aligned} \tag{62}$$

The last equality follows by changing variable to $c = \cos(\theta)$, which introduces an extra factor of

$$\left|\frac{d\theta}{dc}\right| = \left|\frac{dc}{d\theta}\right|^{-1} = |\sin(\theta)|^{-1} = (1-c^2)^{-1/2}. \tag{63}$$

Using generalized spherical coordinates as we did for the calculation of the volume

[17] The discussion around Eq. 89 spells out the rotational invariance argument explicitly in the more general context of group correlation.

[18] We drop the subscript $\underline{\mathbf{1}}$ on the sphere at this point since we are really just doing a calculation for a general sphere and because we want to introduce another subscript for the radius of the sphere. The radius defaults to one if not specified.

of the sphere, we can calculate the expectation value of any function F of the sample correlation c:

$$E[F(c)] \quad = \quad \int_{c=-1}^{1} dc\, F(c) \left[(1-c^2)^{(\nu-2)/2} \frac{\mathrm{Vol}(\mathbf{S}_1^{\nu-1})}{\mathrm{Vol}(\mathbf{S}_1^{\nu})} \right]. \tag{64}$$

By the definition of marginal distribution (see Eq. 37 with $y = c$, $l = 1$, and $x \in \mathbb{R}^k$ replaced by $\tilde{Z}^{(y)} \in \mathbf{S}^{\nu}$), the PDF for c is

$$pdf(c) \quad = \quad \frac{(1-c^2)^{(\nu-2)/2}}{B(\nu/2, 1/2)}. \tag{65}$$

Here we have written the normalization constant as a special case of the beta function,

$$B(a,b) = \int_0^1 dt\; t^{a-1}(1-t)^{b-1} = B(b,a). \tag{66}$$

The correlation distribution has mean zero and standard deviation $1/\sqrt{1+\nu}$.

E.6.3 Relation of Correlation to T-Statistic

As mentioned previously, it is standard in statistics to work with a transformation of the correlation c called the t-statistic:

$$t \quad = \quad \sqrt{\nu} \frac{c}{\sqrt{1-c^2}}. \tag{67}$$

The t-statistic ranges over the whole real line. It is easy to check that the distribution Eq. 65 is equivalent to Student's t-distribution with ν degrees of freedom

$$pdf(t) = \frac{\left(\frac{\nu}{\nu+t^2}\right)^{\frac{\nu+1}{2}}}{\sqrt{\nu} B(\nu/2, 1/2)}. \tag{68}$$

E.6.4 Large n Limit of Distribution of Correlation

Although the distribution of correlation is only non-zero on the interval $[-1, 1]$, for large n it looks like a bell curve concentrated tightly around the origin. If one rescales by the standard deviation, $\sqrt{1+\nu}$, the limit is in fact a bell curve. This mean that, for large n, P-values for correlation can be calculated as tail probabilities of a normal distribution (of the sort given in Table 11 on p. 32). Specifically, we will now show that

$$Lim_{n\to\infty}\, \mathrm{Prob}(c > \frac{nStdOut}{\sqrt{1+\nu}}) = \int_{x=nStdOut}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{69}$$

In words, the limit of the probability that the correlation with a random sample is *nStdOut* or more standard deviations above zero equals the probability of choosing a number greater than *nStdOut* from a standard normal distribution.

To prove Eq. 69, let

$$\begin{aligned} M &= \nu - 2 = n - 4, && (70)\\ a &= \sqrt{M}, \text{ and} && (71)\\ \mathcal{N} &= \frac{1}{B(\nu/2, 1/2)} && (72)\\ & && (73) \end{aligned}$$

Next, we change variables to $x = a * c$ in the integral to calculate the following tail probability of the correlation distribution:

$$\begin{aligned} \text{Prob}(c > nStdOut/a) &= \mathcal{N} \int_{c=nStdOut/a}^{\infty} dc\ \left(1 - c^2\right)^{(\nu-2)/2} && (74)\\ &= (\mathcal{N}/a) \int_{x=nStdOut}^{\infty} dx\ \left(1 - x^2/M\right)^{M/2}. && (75) \end{aligned}$$

Direct calculation verifies that the limit for large n of $\mathcal{N}/a$ is $1/\sqrt{2\pi}$. Using this and the fact that the large M limit of $(1 - y/M)^M$ is e^{-y} shows that

$$Lim_{n\to\infty}\, \text{Prob}(c > nStdOut/a) = \int_{x=nStdOut}^{\infty} dx\ \frac{1}{\sqrt{2\pi}}\ e^{\frac{-x^2}{2}}. \qquad (76)$$

The right hand side of Eq. 76 equals the right hand side of Eq. 69. Our proof would be done except that the left hand sides of these equations differ because $1/a$ is not the standard deviation of the correlation distribution. The difference between the two left hand sides vanishes because it is the large n limit of

$$\text{Prob}(nStdout/\sqrt{n-1} < c < nStdout/\sqrt{n-4}).$$

But this is smaller than $(nStdout * \mathcal{N}/a)$ times $\left(1 - \sqrt{\frac{n-4}{n-1}}\right)$. The former factor limits to $nStdout/\sqrt{2\pi}$ and the latter factor drop off faster than $2/n$ for large n.

E.7 Group Correlation

In Section 5, we compare the result of total correlation with two types of in-group correlation, where the grouping was either by country or by year. In this subsection, we give the general definitions of the two types of in-group correlation and derive the distribution of those correlations under the null hypothesis that the Y variable is generated by independent normal sampling.

Suppose X and Y are vector with components X_i, Y_i and that the index set $\{i\}_{i=1}^{n}$ is divided into G different groups. Let $\mathcal{G}$ be the set of possible group labels; in the application of this paper, this would either be the set of all countries or the set of all years being considered for a particular study. Let $g_i \in \mathcal{G}$ be the label for the group to which the index i belongs. For each group label g, let $\mathcal{I}_g$ be the set of indices with group label g, and n_g be the number of element of $\mathcal{I}_g$. We let $\bar{X}^g$, $\bar{Y}^g$, $\mathrm{std}(X^g)$, $\mathrm{std}(Y^g)$, and $\mathrm{corr}(X^g, Y^g)$ be the means, (uncorrected) standard deviations, and correlations of the vectors X^g and Y^g consisting of the components of X and Y which have label g.

The within-group variance of X (or Y) is the average of the squared difference of the components of X (resp. Y) from their group mean. This equals a weighted average of the variance for each group separately. The within-group standard deviation is the square root of the within-group variance:

$$\begin{aligned}
\mathrm{var}_{in}(X) &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}^{g_i})^2 = \sum_{g\in\mathcal{G}} w_g \, \frac{1}{n_g}\sum_{i\in\mathcal{I}_g}(X_i - \bar{X}^g)^2 && (77)\\
&= \sum_{g\in\mathcal{G}} w_g \, \mathrm{var}(X^g) && (78)\\
\mathrm{std}_{in}(X) &= \sqrt{\mathrm{var}_{in}(X)} && (79)\\
w_g &= \frac{n_g}{n}. && (80)
\end{aligned}$$

It is natural for us to define the within-group Z-score associated to X as the vector obtained by subtracting off the group means and dividing by the within-group standard-deviation:

$$\left(Z^X_{in}\right)_i = \frac{X_i - \bar{X}^{g_i}}{\mathrm{std}_{in}(X)}. \tag{81}$$

Similarly to Eq. 57, we define the unit-length, within-group Z-score to be

$$\tilde{Z}^X_{in} = n^{-1/2} Z^X_{in}. \tag{82}$$

There are two variants of the definition of within-group correlation, depending on whether the standard deviations we divide by are allowed to depend on the group or are taken to be a common "tied" value common to all indices. The first definition, in which the standard deviations are not tied, is just a weighted[19] average of the correlations for

[19] There is a variant of the definition of within-group correction using untied standard deviations that is the simple (unweighted) average of the correlation for each group label. This variant equals the dot product of "unit-length, within-group Z-scores with untied standard deviations", defined as:

$$\left(\tilde{Z}^X_{in}\right)^{untied}_i = \frac{1}{G^{1/2}} \frac{X_i - \bar{X}^{g_i}}{n_g^{1/2}\,\mathrm{std}(X^g)}. \tag{83}$$

each group label:

$$\begin{aligned} \mathrm{corr}_{in}(X,Y) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}^g}{\mathrm{std}(X^g)}\right)\left(\frac{Y_i - \bar{Y}^g}{\mathrm{std}(Y^g)}\right) && (84)\\ &= \sum_{g=1}^{G} w_g \,\mathrm{corr}(X^g, Y^g). && (85) \end{aligned}$$

In the second definition, we take the common, tied standard deviations to be the within-group standard deviations. This definition is equivalently defined as the dot product of unit-length, within-group Z-scores:

$$\begin{aligned} \mathrm{corr}_{in}^{std-tied}(X,Y) &= \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}^g}{n^{1/2}\,\mathrm{std}_{in}(X)}\right)\left(\frac{Y_i - \bar{Y}^g}{n^{1/2}\,\mathrm{std}_{in}(Y)}\right) && (86)\\ &= \tilde{Z}_{in}^{X}\cdot\tilde{Z}_{in}^{Y}. && (87) \end{aligned}$$

Both versions of within-group correlation always lie between minus one and one. For the case when standard deviation are not tied, this is because within-group correlation is a weighted average of numbers in this range. When the standard deviations are tied, within-group correlation lies in this range because it is the dot product of unit vectors.

We now derive the distribution of within-group correlation when the vector X is fixed and the components of Y are selected independently from a normal distribution. We first look at the case of tied standard deviations and then look at the untied case.

E.7.1 Distribution for Within-Group Correlation with Tied Standard Deviations Under the Null Hypothesis

For the case when standard deviations are tied, the derivation of the distribution of within-group correlation is a direct generalization of the derivation in Section E.6 for correlation not using group labels (which we have called total correlation in the main text. The within-group correlation with tied standard deviations is again the dot product of the Z-scores of X and Y, i.e. the cosine of the angle between two vectors on a unit sphere. The only difference is that now a subtraction of a separate mean for each group label means that the vector $Y_i - \bar{Y}^{g_i}$ belongs to the $(n-G)$-dimensional subspace of $\mathbb{R}^n$ consisting of vectors whose mean for each group label vanishes. So the dimension of the sphere to which within-group Z-scores belong, a.k.a. the number of degrees of freedom, now equals

$$\nu_{in} = n - G - 1. \qquad (88)$$

The demonstration that the correlation distribution has the form Eq. 65 and has large n limit Eq. 69 goes through as before.

In footnote 17, we promised that we would spell out explicitly here the argument that the Z-score for Y has uniform density. We continue to assume that the components of Y are picked independently from a normal distribution with group-independent standard deviation σ and mean μ (although we could allow the mean to depend on group as well). A simple calculation shows that the PDF for Y only depends on the group means and the within-group standard deviation:

$$\begin{aligned}
\mathrm{Prob}(Y) &= (2\pi\sigma)^{-n/2}\, e^{-Q/(2\sigma^2)}, \text{where} \\
Q &= n\,\mathrm{std}_{in}(Y)^2(\tilde{Z}^Y_{in}\cdot\tilde{Z}^Y_{in}) + \sum_g n_g\left(\bar{Y}^g-\mu\right)^2 \\
&= n\,\mathrm{std}_{in}(Y)^2 \qquad\qquad + \sum_g n_g\left(\bar{Y}^g-\mu\right)^2. \qquad (89)
\end{aligned}$$

The marginal distribution for $\tilde{Z}^Y_{in}$ is constant since Prob(Y) does not depend on it.

E.7.2 Distribution for Within-Group Correlation with Untied Standard Deviations Under the Null Hypothesis

As we saw in Eq. 84, the within-group correlation when the standard deviations are not tied is just a weighted average of the correlation for each group label,

$$\begin{aligned}
c_{in} &= \sum w_g c_g \qquad (90)\\
c_g &= \mathrm{corr}(X^g, Y^g). \qquad (91)
\end{aligned}$$

There is no simple closed form solution for the distribution of this correlation under the null hypothesis. However, things do simplify because the c_g are independent random variables. For example, the expectation value of c_{in} vanishes because it equals the weighted average of the expectation values of the c_g, which are all zero under the null hypothesis. Since the c_g are independent, the variance of c_{in} is

$$\mathrm{var}(c_{in}) = \sum_{g\in\mathcal{G}} {w_g}^2\left(\frac{1}{n_g-1}\right). \qquad (92)$$

We may write this in the form

$$\mathrm{var}(c_{in}) = \frac{1}{1+\nu^{eff}}, \qquad (93)$$

where the *effective* number of degrees of freedom, ν^{eff}, is

$$\nu^{eff} \;=\; \left[\sum_{g \in G} {w_g}^2 \left(\frac{1}{n_g - 1}\right)\right]^{-1} - 1. \tag{94}$$

If the n_g are independent of g, then ν^{eff} equals $n - G - 1$, the same form as found for tied standard deviations in Eq. 88. Note that this differs from the naive count $n - 2G$ obtained by multiplying the number of groups by the number of degrees of freedom per group.

There are two separate reasons that the distribution (of within-group correlation with untied standard deviations) is approximately normally distributed for large n. The first reason is that the central limit theorem applies when the number of group labels, G, is large to tell us that c_{in}, which is an average of G independent variables, is approximately normal. The second reason is that the c_g are approximately normal for large n_g, and the weighted average of normal distributions is normal. Thus Eq. 69 holds with c and ν replaced by c_{in} and ν^{eff}.

E.8 Regression

In this subsection of this appendix, we will discuss the mathematics of (ordinary unconstrained) linear regression, with a focus on significance testing. We explain how four different measures of a regression – the multiple correlation coefficient, R-squared, adjusted R-squared and F statistics – fit together in a hypothesis testing framework.

E.8.1 Formulas for Regression

Let us consider least squares linear regression for a data set, $\{Y_i, {X_i}^1, ...{X_i}^k\}_{i=1}^n$, of samples of a variable Y, called the *dependent* variable, and variables X^1, ..., X^k, called the *independent*, *regressor*, *predictor*, or *explanatory* variables.

The regression model is the best linear estimate of Y given X:

$$\hat{Y} \;=\; \alpha + \beta_1 X^1 + ... + \beta_k X^k, \tag{95}$$

where α and β_1, ..., β_k are constants. The model is "best" in the sense that it minimizes the sum of squared errors,

$$SSE \;=\; \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \tag{96}$$

$$\hat{Y}_i \;=\; \alpha + {X_i}^1 \beta_1 + ... + {X_i}^l \beta_l. \tag{97}$$

The term alpha can be thought of as the constant associated with a regressor that has constant value one. The minimization with respect to α is easy to solve for and allows us to rewrite the model as

$$\hat{Y}_i \;=\; \bar{Y} + ({X_i}^1 - \bar{X}^1)\beta_1 + ... + ({X_i}^k - \bar{X}^k)\beta_k. \tag{98}$$

We identify the variable names Y, X^l, and $\hat{Y}$ with the n-dimensional column vectors of their instances. The $n \times k$ matrix with entries ${X_i}^l$ will be denoted X and the k-dimensional column vector of coefficients will be written as β. We also let $\tilde{X}$ be the matrix whose columns are the columns of X with their means subtracted. With these definitions, Eq. 95 and Eq. 98 become vector equations using matrix multiplication:

$$\hat{Y} \;=\; \alpha\underline{\mathbf{1}} + X\beta = \bar{Y}\underline{\mathbf{1}} + \tilde{X}\beta. \tag{99}$$

(Recall the $\underline{\mathbf{1}}$ is the column vector whose components are all one.)

The *estimation error vector* is the difference between Y and its estimate:

$$e \;=\; Y - \hat{Y}. \tag{100}$$

The sum of squared errors, Eq. 96, equals the squared length of the error vector.

$$SSE \;=\; ||e||^2 = \sum_{i=1}^{n} e_i^2. \tag{101}$$

An estimate of the variance of the error is

$$MSE \;=\; \frac{SSE}{n-k-1}. \tag{102}$$

We have divided by $n-k-1$ because that is the number which makes MSE an unbiased estimate of population error variance. Another convention for defining the mean squared error takes it to be the naive average of the squared errors. We will refer to this as the uncorrected mean squared error,

$$MSE^U \;=\; \frac{SSE}{n}. \tag{103}$$

A little calculus with matrices shows that the solution for β is

$$\beta = \left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}^TY. \tag{104}$$

The Y vector is the sum of three pieces which are orthogonal to each other:

$$Y = \bar{Y}\underline{\mathbf{1}} \;+\; \tilde{X}\beta \;+\; e. \tag{105}$$

The R-squared of the regression is the fraction of the the sum of the squares of the deviations of the components of Y from its mean which is "predicted" by the model. That is, it is the ratio of the sum of squares of values from regression (SSR) to the total sum of squares (SST):

$$\begin{aligned} R^2 &= \frac{SSR}{SST}, && (106)\\ SSR &= ||\hat{Y}-\bar{Y}||^2, && (107)\\ SST &= ||Y-\bar{Y}||^2. && (108) \end{aligned}$$

Since the decomposition in Eq. 105 is orthogonal we have

$$SST = ||Y-\bar{Y}||^2 \;=\; ||Y-\hat{Y}||^2 + ||\hat{Y}-\bar{Y}||^2 = SSE + SSR. \tag{109}$$

So we may rewrite R-squared as:

$$R^2 \;=\; 1-\frac{SSE}{SST}. \tag{110}$$

E.8.2 For Simple Regression, R-squared is Correlation Squared

Simple regression is the case when there is only one independent variable, i.e. $k = 1$. In that case, the R-squared of the regression equals the square of the correlation of X and Y. This follows as a special case of the demonstration we give in a moment for multiple regression ($k \geq 1$).

A direct proof for simple regression follows from the fact that, in that case, the matrix X is just a column vector, and X^TY equals the dot product of X and Y. The regression model still has the form in Eq. 99, but now $\tilde{X}$ is a column vector. The regression coefficient β is

$$\begin{aligned} \beta &= \left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}^TY = \frac{\tilde{X}^T\tilde{Y}}{||\tilde{X}||^2} = \frac{||\tilde{Y}||}{||\tilde{X}||}(\tilde{Z}^{(x)}\cdot\tilde{Z}^{(y)}) && (111)\\ &= \frac{s_y^U}{s_x^U}\,\mathrm{corr}(X,Y). && (112) \end{aligned}$$

In this case, R-squared can be written in several equivalent ways:

$$R^2 \;=\; \frac{||\hat{Y}-\bar{Y}||^2}{||Y-\bar{Y}||^2} = \beta^2\frac{||X-\bar{X}||^2}{||Y-\bar{Y}||^2} = \mathrm{corr}(X,Y)^2. \tag{113}$$

E.8.3 For Multiple Regression, R-squared Is Square of Multiple Correlation Coefficient

The *multiple correlation coefficient* is the correlation between the Y data and the estimate $\hat{Y}$:

$$r \quad = \quad \text{multiple correlation coefficient} = \text{corr}(Y, \hat{Y}). \tag{114}$$

The multiple correlation coefficient is the square root of R-squared, so R-squared is sometimes called the "squared multiple correlation coefficient". This follows because

$$r = \frac{(Y - \bar{Y}) \cdot (\hat{Y} - \bar{Y})}{||Y - \bar{Y}||\,||\hat{Y} - \bar{Y}||} = \frac{(\hat{Y} - \bar{Y}) \cdot (\hat{Y} - \bar{Y})}{||Y - \bar{Y}||\,||\hat{Y} - \bar{Y}||} = \frac{||\hat{Y} - \bar{Y}||}{||Y - \bar{Y}||} = \sqrt{R^2}. \tag{115}$$

The second equality follows because the error $e = Y - \hat{Y}$ is orthogonal to $\hat{Y} - \bar{Y}$.

Note that the multiple correlation coefficient is always positive. For the case of simple regression, $\hat{Y} - \bar{Y}$ is a multiple of X, and so the multiple correlation coefficient is equal to the absolute value of the correlation between Y and X.

E.8.4 Split of $\mathbb{R}^n$ into Three Orthogonal Pieces: constant vectors, linear combinations of the independent variables, and estimation error vectors

That any vector Y in $\mathbb{R}^n$ can be uniquely decomposed as a sum of three orthogonal pieces as in Eq. 105 means that that $\mathbb{R}^n$ is the *orthogonal direct sum* of three pieces: the vectors proportional to the vector $\underline{\mathbf{1}}$ of all ones, the space of all linear combination of the $\tilde{X}^l$, and the space E of all possible error vectors. E is the subspace of $\mathbb{R}^n$ consisting of all vectors orthogonal to the sum of the other two spaces. This decomposition is denoted as follows:

$$\mathbb{R}^n = \text{Span}(\{\underline{\mathbf{1}}\}) \oplus \text{Span}(\{\tilde{X}^l\}) \oplus E. \tag{116}$$

The matrices for the orthogonal projection onto the first two pieces in Eq. 116 are

$$\begin{aligned} \Pi_{\text{Span}(\{\underline{\mathbf{1}}\})} &= n^{-1}\, \underline{\mathbf{1}}\,\underline{\mathbf{1}}^T, \text{ and} && (117) \\ \Pi_{\text{Span}(\{\tilde{X}^l\})} &= \tilde{X} \left(\tilde{X}^T \tilde{X}\right)^{-1} \tilde{X}^T. && (118) \\ &. && (119) \end{aligned}$$

The matrix for orthogonal projection onto the third piece is just the identity matrix minus the other two projection matrices.

For simple regression, we saw in Eq. 113 that the R-squared statistic equals the square of the correlation of the independent and dependent data, which equal the square of the dot product of their unit-length Z-scores. For multivariable regression in general, the

R-squared equals the length squared of the projection of the unit-length Z-score for the dependent variable onto the space spanned by the mean subtracted instance vectors of the independent variables. This translates into the following equation:

$$R^2 \;\; = \;\; ||\Pi_{\mathrm{Span}(\{\tilde{X}^l\})}\tilde{Z}^{(y)}||^2. \tag{120}$$

E.8.5 Distribution of R-squared Under the Null Hypothesis

The description we have just given of R-squared is a mouthful, but it will allow us to easily derive the the distribution of R-squared under the null hypothesis that the components of the vector Y are sampled independently from a normal distribution with some mean μ and standard deviation σ. We now present this derivation in a way generalizing how we derived the distribution of correlation in Appendix E.6.2. The distribution we find only depends on the number of independent variables and instances.

We start by letting r be the square root of R-squared, i.e. the multiple correlation coefficient. For simple regression, this is just the absolute value of the correlation of X and Y. In general, r is the norm of the projection of $\tilde{Z}^{(y)}$ onto the space spanned by $\{\tilde{X}^l\}$. So the distribution of r can be deduced from the distribution of $\tilde{Z}^{(y)}$, which we saw in Appendix E.6.2 is the uniform distribution on $\mathbf{S}^{n-2}_{\underline{\mathbf{1}}}$, the $(n-2)$-dimensional sphere of unit-vectors in $\mathbb{R}^n$ which are orthogonal to the vector of all ones.

Since the distribution is rotationally invariant, we are free to apply a rotation that rotates the subspace spanned by $\{\tilde{X}^l\}$ into the subspace $\mathbb{R}^k$ of $\mathbb{R}^n$ consisting of vectors whose last $n-k$ components vanish. This step is a generalization of the step in Appendix E.6 where we rotated $\tilde{Z}^{(x)}$ to the north pole. At the same time, we rotate the space of vectors orthogonal to $\underline{\mathbf{1}}$ into $\mathbb{R}^{n-1}$.

We now define a useful variant of spherical coordinates for the vector $\tilde{Z}^{(y)}$. Let v be the unit vector in the direction of $\hat{Y}-\bar{Y}$, which we can now say is the direction of the projection of $\tilde{Z}^{(y)}$ onto $\mathbb{R}^k$. Also let w be the unit vector in the direction of e, which is now the direction of the projection of $\tilde{Z}^{(y)}$ onto of the copy of $\mathbb{R}^{n-k-1}$ consisting of vectors in $\mathbb{R}^{n-1}$ whose first k coordinates vanish. Then

$$\tilde{Z}^{(y)} = r\,v + \sqrt{1-r^2}\,w. \tag{121}$$

The PDF for r is proportional to the $(n-3)$-dimensional volume of the set of $\tilde{Z}^{(y)}$ which have the given r, times a change of variables factor dependent on r. The volume of the space for fixed r is the product of: (i) the volume of the sphere in $\mathbb{R}^k$ of radius r to which $r\,v$ belongs and (ii) the volume of the sphere in $\mathbb{R}^{n-k-1}$ of radius $\sqrt{1-r^2}$ to which $\sqrt{1-r^2}\,w$ belongs. This product is proportional to r^{k-1} times $\left(\sqrt{1-r^2}\right)^{n-2-k}$. For those comfortable with change of variables in multivariable calculate, the change of variables factor can be calculated by computing the appropriate Jacobian determinant. The result is $(1-r^2)^{-1/2}$, the direct generalization of the factor in Eq. 63. As we said, the PDF for r is proportional to the product of the volume factor and the change of variables factor we just calculated:

$$pdf(r) \propto r^{k-1}\left(\sqrt{1-r^2}\right)^{n-3-k}. \tag{122}$$

The PDF for R^2 follows by changing variable using[20] $R^2 = r^2$ and including an overall factor so that the PDF integrates to one:

$$pdf(R^2) \;=\; \frac{(R^2)^{-1+k/2}(1-R^2)^{(n-3-k)/2}}{B(k/2,(n-1-k)/2)}, \tag{123}$$

where B is the beta function defined in Eq. 66.

E.8.6 Adjusted R-squared and Its Statistics Under the Null Hypothesis

The mean and variance of R-squared under the null hypothesis can easily be calculated from the distribution Eq. 123:

$$\begin{aligned} \text{mean}(R^2) &= \frac{k}{(n-1)}, & (124)\\ \text{var}(R^2) &= \frac{2k(n-1-k)}{(n+1)(n-1)^2}. & (125)\end{aligned}$$

So the mean of R-squared increases linearly with the number of independent variables. This is a precise version of the statement made in Section 6.2 that the R-squared values become large when many independent variables are included. A common practice to correct for this is to report an adjusted value for R-squared which correctly accounts for the fact that the error vector $Y-\hat{Y}$ belong to the $n-1-k$ dimensional space E of possible error vectors, and the deviation of Y from the mean, $Y-\bar{Y}$, belongs to the $n-1$ dimensional space of noise vectors with mean zero. R^2 may be written in terms of uncorrected mean squares as

$$1-R^2 \;=\; \frac{MSE^U}{MST^U} = \frac{||Y-\hat{Y}||^2/n}{||Y-\bar{Y}||^2/n}. \tag{126}$$

Adjusted R-squared can be written in term of corrected mean squares as

$$1-R^2_{adj} \;=\; \frac{MSE}{MST} = \frac{||Y-\hat{Y}||^2/(n-1-k)}{||Y-\bar{Y}||^2/(n-1)}. \tag{127}$$

[20] Note that R^2 is the name of a variable, whereas r^2 is the square of the variable r.

The formula transforming R-squared to adjusted R-squared is

$$R^2_{adj} \quad = \quad 1 - (1 - R^2)\left(\frac{n-1}{n-1-k}\right). \tag{128}$$

This can be written in a way that make it manifest that adjusted R-squared has mean zero:

$$R^2_{adj} = \left(\frac{n-1}{n-1-k}\right)\left(R^2 - \mathrm{mean}(R^2)\right). \tag{129}$$

The fact that the mean is zero is the precise version of the statement made in Section 6.2 that the adjusted R-squared only gets larger when new predictors are added if they reduce the squared error by more than what would be expected by chance.

Using the mean and variance of R-squared in Eq. 124, we find that the adjusted R-squared has mean zero and variance

$$\mathrm{var}(R^2_{adj}) \quad = \quad \frac{2k}{(n+1)(n-k-1)}. \tag{130}$$

E.8.7 Relation of R-squared to F-statistic

For completeness we now show that the distribution for R^2, given by Eq. 123, is equivalent to the F-distribution of the F-statistic. The F-statistic is the statistic commonly used for hypothesis testing the overall significance of a regression, despite the fact that it is the R-squared statistic that is the one usually reported to measure how well the regression explains the data.

The F-statistic is the ratio of the (corrected) mean square of values from regression to the (corrected) mean square error,

$$\begin{aligned} f \quad &= \quad \frac{MSR}{MSE} = \frac{||\hat{Y} - \bar{Y}||^2/k}{||Y - \hat{Y}||^2/(n-k-1)} \qquad (131)\\ &= \quad \frac{R^2}{1-R^2}\frac{n-1-k}{k} \qquad (132)\\ &= \quad \frac{n-1}{k(1-R^2_{adj})} - \frac{n-1-k}{k}. \qquad (133)\end{aligned}$$

Under the null hypothesis, the probability distribution for the F-statistic is the F distribution with parameters $(k, n-1-k)$:

$$\begin{aligned}
\mathrm{Prob}(f) &= \frac{{\nu_2}^{\nu_2/2}(f\nu_1)^{\nu_1/2}(f\nu_1+\nu_2)^{\frac{1}{2}(-\nu_1-\nu_2)}}{fB\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \\
&= \frac{(fk)^{k/2}(-k+n-1)^{\frac{1}{2}(-k+n-1)}((f-1)k+n-1)^{\frac{1-n}{2}}}{fB\left(\frac{k}{2},\frac{1}{2}(-k+n-1)\right)}, \text{ where} \qquad (134) \\
\nu_1 &= k, \qquad (135) \\
\nu_2 &= n-1-k. \qquad (136) \\
& \qquad (137)
\end{aligned}$$

This has mean and variance

$$\begin{aligned}
\mathrm{mean}(f) &= \frac{\nu_2}{\nu_2-2}, \text{ and} \qquad (138) \\
\mathrm{var}(f) &= \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)(\nu_2-2)^2}. \qquad (139)
\end{aligned}$$

The distribution for f can be derived simply from the distribution for R-squared (Eq. 123) by the change of variables Eq. 132. By contrast, the standard derivation of the distribution for f begins by noting that the squared lengths in the numerator and denominator in Eq. 131 are independent random variables and each is distributed as a chi-squared distribution (the distribution of a sum of squares of independent standard normal variables). Our derivation seems simpler to us and highlights the role of multiple correlation, whereas the standard derivation has the advantage that it can be more easily generalized to the case when population R-squared is non-zero, which we will spell out in Appdendix E.9.5.

E.9 Adjusted R-Squared and the Population Model for Regression

E.9.1 Adjusted R-Squared is an Almost Unbiased Estimator of Population R-Squared

We introduce below a population model in which some fraction of the variation of Y is "truly explained" by regression. This fraction is called the *population R-squared*, denoted ρ^2. The null hypothesis is the special case when ρ^2 vanishes so that no fraction of the deviation of Y from its mean is "truly explained" by regression. As we have seen, adjusted R-squared has mean zero when the null hypothesis is true, correcting for the problematic fact that R-squared has positive expectation value because it is positive for any sample[21].

Reference (Yin & Fan 2001) reviews the literature on many alternative formulas for

[21] Technically, R-squared vanishes on a set of Y with probability zero.

adjusted R-squared that have been proposed over many decades. The formula we use in this paper is the one proposed by Wherry in (Wherry 1931), which Yin and Fan point out is “the most widely used”, although they conclude that it is “probably not the most effective analytical formula for estimating ρ^2.” In his 1931 paper, Wherry claims that: “It has been demonstrated that the new Wherry formula, both by a least squares criterion and by actual application, is more nearly true than [a previous effort at adjusting R-squared]”.

Yin and Fan systematically compare six different formulas under a range of conditions (various n, k, ρ^2, and the degree of dependence between regressors[22]) to see which one is the most “operationally unbiased” estimator of ρ^2. For each condition, they calculate an approximation to the expectation value of population R-squared based on 500 random trials[23] from the population. They follow the convention by which “researchers have operationally defined an unbiased estimate as having means based on the 500 replications to be within .01 of the corresponding population parameters” (Yin & Fan 2001, p. 214). A formula for R-squared can fail to be operationally unbiased by this definition either because its expectation value (the limit of the mean as the number of trials becomes infinite) is biased or because its standard deviation is too high (so that the mean based on 500 trials is substantially different from the full expectation value). They find that all 6 formulas are operationally unbiased under the majority of conditions, with the Wherry formula being unbiased 77% of the time.

E.9.2 Population Model and Population R-squared

The population model for regression specifies that the n-dimensional vector Y is a random variable equal to a constant, plus a linear combination of the columns of an $n \times k$ regressor matrix X plus an error vector. We will restrict ourselves to a so-called “fixed-effects” model in which X is constant, as opposed to a “random-effects” model in which the matrix X is a random variable. Population R-squared is the fraction of the variance of Y explained as a linear combination of the columns of X.

The parameters of the model in addition to X are a constant μ, a $k \times 1$ vector β of “population regression coefficients”, and a noise scale σ. The model for Y is

$$Y = \mu \underline{\mathbf{1}} + \tilde{X}\beta + \epsilon. \tag{140}$$

In order for the constant μ to have an interpretation as a mean, we have used the matrix

[22] We will derive the population distribution of adjusted R-squared below. The exact distribution only depends on n, k, and ρ^2. It does not depend on the matrix $\tilde{X}$, as long as it is invertible. The modest dependence on the intercorrelation of independent variables that Yin and Fan report presumably comes about due to numerical approximation.

[23] A single random trial, or “replication” in the wording of Yin and Fan, is a sampling of the n-dimensional vector Y from the population distribution.

$\tilde{X}$, which equals the matrix X with the mean subtracted from each column. The "noise" vector ϵ is an $n \times 1$ vector whose components are independent and identically distributed by a normal distribution with mean zero and standard deviation σ. Equivalently, we could say that Y is distributed as a multivariate normal distribution with mean vector $\mu\underline{\mathbf{1}} + \tilde{X}\beta$ and covariance matrix equal to σ^2 time the identity matrix.

We define the population regression estimate to be the first two terms of Eq. 140,

$$\hat{Y}_{pop} = \mu\underline{\mathbf{1}} + \tilde{X}\beta. \tag{141}$$

The population version of the sum of squares of values from the regression defined in Eq. 107 is

$$SSR_{pop} = ||\hat{Y}_{pop} - \bar{Y}_{pop}||^2 = ||\tilde{X}\beta||^2. \tag{142}$$

The population total sum of squares is the expectation value of the length squared of Y minus its mean, $\bar{Y}$. It equals the population sum of squares of values from regression plus the sum of squares from noise,

$$\begin{aligned} SST_{pop} &= E(||Y - \bar{Y}||^2) = SSR_{pop} + SSE_{pop}, && (143) \\ SSE_{pop} &= E(||\epsilon - \bar{\epsilon}||^2) = (n-1)\sigma^2. && (144) \end{aligned}$$

Now we can define the population R-squared,

$$\rho^2 = R^2_{pop} = \frac{SSR_{pop}}{SST_{pop}} = \frac{||\tilde{X}\beta||^2}{||\tilde{X}\beta||^2 + (n-1)\sigma^2}. \tag{145}$$

E.9.3 Refined Population Model and Population Multiple Correlation

In the literature, population R-squared is sometimes called a square of the population multiple correlation coefficient, but it rarely if ever seems to be spelled out what that means in terms of a precise population model. For completeness we explain one way to spell this out in the subsubsection, although we shall not make use of this formulation elsewhere in this paper.

The model now is for a scalar random variable Y^s, which takes a form similar the population model in Eq. 140 for the column vector Y. The parameters of the model are the same as before: an $n \times k$ matrix X; a constant μ; a $k \times 1$ vector β; and a noise scale σ. We let $\tilde{X}^s$ be a random variable which is uniformly distributed on the rows of the matrix $\tilde{X}$. In other words, the value of $\tilde{X}^s$ has a $1/n$ probability of equalling the i'th row of $\tilde{X}$. The model for Y^s is

$$Y^s = \mu + \tilde{X}^s\beta + \epsilon^s, \tag{146}$$

where ϵ^s is distributed normally (and independently from X^s) with mean 0 and standard deviation σ.

The model parameters μ and β are the choice of μ^* and β^* that mininimize the expectatation value of the squared error $(Y^s - \hat{Y}^s)^2$ in the linear model $\hat{Y}^s = \mu^* + \tilde{X}^S\beta^*$. So the scalar version of the population regression estimate (in Eq. 141) is

$$\hat{Y}^s = \mu + \tilde{X}^s\beta, \tag{147}$$

This is the least squares estimate for regression over the whole population.

It is easy to calculate the following:

$$\begin{aligned}
E(\tilde{X}^s) &= 0, & (148)\\
E(Y^s) &= E(\hat{Y}^s) = \mu, & (149)\\
\mathrm{var}(\hat{Y}^s) &= E\left((\tilde{X}^s\beta)^2\right) = \frac{1}{n}||\tilde{X}\beta||^2, & (150)\\
\mathrm{var}(Y^s) &= E\left((\tilde{X}^s\beta + \epsilon)^2\right) = \frac{1}{n}||\tilde{X}\beta||^2 + \sigma^2, & (151)\\
\mathrm{cov}(Y^s, \hat{Y}^s) &= E\left((\tilde{X}^s\beta)(\tilde{X}^s\beta + \epsilon)\right) = \frac{1}{n}||\tilde{X}\beta||^2. & (152)
\end{aligned}$$

Finally, we may calculate the population multiple correlation coefficient,

$$\begin{aligned}
\rho^s &= corr(Y^s, \hat{Y}^s) = \left(\frac{\mathrm{cov}(Y^s, \hat{Y}^s)^2}{\mathrm{var}(Y^s)\,\mathrm{var}(\hat{Y}^s)}\right)^{1/2}\\
&= \left(\frac{\frac{1}{n}||\tilde{X}\beta||^2}{\frac{1}{n}||\tilde{X}\beta||^2 + \sigma^2}\right)^{1/2} & (153)
\end{aligned}$$

So the population R-squared is:

$$(\rho^s)^2 = \frac{||\tilde{X}\beta||^2}{||\tilde{X}\beta||^2 + n\sigma^2}. \tag{154}$$

Note that this formula for population R-squared differs from Eq. 145 in that the factor $n-1$ appearing in the denominator is replaced by n.

E.9.4 "Derivation" of Formula for Adjusted R-Squared

In this subsection we shall derive the formula for adjusted R-Squared in a way that makes manifest that it is an almost unbiased estimator of the square of the population multiple correlation coefficient.

The R-squared for regression of Y on X produces a biased estimate of ρ^2 because the least square projection of Y onto the space spanned by the columns of $\tilde{X}$ includes the projection of the noise vector. To spell this out, we recall that the orthogonal decomposition in Eq. 116 allows us to write any vector Y as a sum of three orthogonal pieces: a constant, a combination of the columns of $\tilde{X}$, and an error term:

$$Y = \bar{Y}\underline{\mathbf{1}} + (\hat{Y} - \bar{Y}\underline{\mathbf{1}}) + e. \tag{155}$$

The error term e is the projection of ϵ onto the error space E, which is the space orthogonal to constant vectors and the columns of $\tilde{X}$. The constant $\bar{Y}$ is the average of the components of Y,

$$\bar{Y} = \mu + \bar{X}\beta + \bar{\epsilon}, \tag{156}$$

where $\bar{\epsilon}$ is the average of the components of ϵ and $\bar{X}$ is the $1 \times k$ vector of averages of the columns of X.

The deviation of the regression estimate $\hat{Y}$ from its mean is

$$\hat{Y} - \bar{Y} = \tilde{X}\beta + p, \tag{157}$$

where p is the projection of ϵ onto the span of the columns of $\tilde{X}$. The regression sum of squares is the squared length $\hat{Y} - \bar{Y}$,

$$SSR \;=\; ||\tilde{X}\beta||^2 + ||p||^2 + 2p \cdot (\tilde{X}\beta). \tag{158}$$

The expectation under the population distribution of the cross term vanishes and the expectation of $||p||^2$ is the dimension k of the subspace to which it belong times the variance for each component of the noise. So

$$E(SSR) \;=\; ||\tilde{X}\beta||^2 + k\sigma^2. \tag{159}$$

We have already given a formula (Eq. 143) for the expectation value of the total sum of squares when it went under the name of population total sum of squares,

$$E(SST) \;=\; E(||Y - \bar{Y}||^2) = ||\tilde{X}\beta||^2 + (n-1)\sigma^2. \tag{160}$$

The expectation of the sum of squared errors equal σ^2 times the dimensional of the error space,

$$E(SSE) \;=\; E(SST) - E(SSR) = (n-1-k)\sigma^2. \tag{161}$$

Now we are ready to encapsulate in one line why the expectation value of the Wherry formula for adjusted R-squared is approximated equality to the population R-squared:

$$E(R^2_{adj}) = E(1 - \frac{MSE}{MST}) \approx 1 - \frac{E(MSE)}{E(MST)} = 1 - \frac{\sigma^2}{||\tilde{X}\beta||^2/(n-1) + \sigma^2} = \rho^2. \qquad (162)$$

The approximation used is that the expectation value of the ratio MSE/MST is approximately equal to the ratio of the expectation values $E(MSE)/E(MST)$.

E.9.5 Distribution of Adjusted R-squared for General Population Distribution

We now describe the distribution of adjusted R-squared for a general regression population distribution. To do so, we will first go through the derivation of how the F-statistic is distributed as an F-distribution.

We use the orthogonal decomposition (Eq. 116) of $\mathbb{R}^n$ as an orthogonal sum of three spaces of dimensions 1, $\nu_1 = k$, and $\nu_2 = n - k - 1$. This determines a breakup of Y as the some of three pieces, which are distributed independently: $\bar{Y}\underline{\mathbf{1}}$, $\hat{Y} - \bar{Y}$, and e. For the distribution of the first component, we can ignore the vector $\underline{\mathbf{1}}$ and just provide a distribution for the scalar $\bar{Y}$. To write down the distribution of e, we identify the ν_2-dimensional space E to which it belongs with $\mathbb{R}^{\nu_2}$ via a norm preserving, invertible linear map from $\mathbb{R}^{\nu_2}$ to E. Such a linear map takes the form multiplication by an $n \times \nu_2$ matrix whose columns are an orthonormal basis of E. Similarly, to write down the distribution of $\hat{Y} - \bar{Y}$ we need to identify the k-dimensional Span($\{\tilde{X}^l\}$) with $\mathbb{R}^k$. Such a map is determined by an orthonormal basis of the latter space. Using the Gramm-Schmidt orthonormalization process, we can choose a basis so that the first basis vector is a unit vector pointing in the direction of $\tilde{X}\beta$. Then the vector $\tilde{X}\beta$ is identified with the vector $||\tilde{X}\beta||e_{1,k}$, where $e_{1,k}$ is the vector in $\mathbb{R}^k$ whose components all vanish except for the first, which is one.

Using the above identification, the three pieces of Y are distributed as follows:

$$\begin{aligned} \bar{Y} &\sim N(\mu, \sigma^2), & (163)\\ \hat{Y} - \bar{Y} &\sim N(||\tilde{X}\beta||e_{1,k}, \sigma^2 I_k), & (164)\\ e &\sim N(0_{n-k-1}, \sigma^2 I_{n-k-1}). & (165) \end{aligned}$$

Here, the notation

$$X \sim N(v, C) \qquad (166)$$

indicates that the random variable X is distributed by a normal distribution with mean vector v and covariance matrix C; 0_m denotes the m-dimensional vector whose components all vanish; and I_m denotes the $m \times m$ identity matrix.

It is helpful to introduce the *population signal-to-noise ratio*,

$$\lambda = \frac{||\tilde{X}\beta||^2}{\sigma^2}. \tag{167}$$

The population R-squared and the signal-to-noise ratio are related by

$$\begin{aligned} \rho^2 &= \frac{\lambda}{(n-1)+\lambda}, & (168)\\ \lambda &= (n-1)\frac{\rho^2}{1-\rho^2}. & (169) \end{aligned}$$

We introduce the following rescaled vectors which are distributed normally with identity covariance:

$$\begin{aligned} \frac{\hat{Y}-\bar{Y}}{\sigma} &\sim N(\lambda^{1/2}e_{1,k}, I_k), & (170)\\ \frac{e}{\sigma} &\sim N(0_{\nu_2}, I_{\nu_2}). & (171) \end{aligned}$$

The norm squared of the second vector is distributed as an ordinary chi-squared distribution, and that of the first vector is distributed by a non-central chi-squared distribution.

The ordinary chi-squared distribution with ν degrees of freedom, denoted χ^2_ν, is a distribution on the positive real line of a variable which equals the sum of the squares of ν independent random variables distributed by the standard normal distribution. Equivalently, it is the distribution of the length squared of a vector which is distributed by $N(0_\nu, I_\nu)$. A formula for the PDF of this distribution is easily derived from the formula for a normal distribution or looked up in many standard references.

The non-central chi-squared distribution, $\chi^2_\nu(\lambda)$, with ν degrees of freedom and non-centrality parameter λ, is the distribution of the length squared of a vector X distributed as $N(\mu, I_\nu)$, where the length squared of μ is equal to λ. Equivalently, X is distributed as the sum of squares of independent normally distributed random variables $X_1, ..., X_k$ with unit variance and means $\mu_1, ..., \mu_\nu$ whose squares sum to λ. The ordinary chi-squared distribution is the special case when the non-centrality parameter vanishes, $\chi^2_\nu = \chi^2_\nu(0)$.

Let SSR_1 and SSE_1 be the norm-squared of the vectors in Eq. 170. Then

$$\begin{array}{lllllll} SSR_1 & = & SSR/\sigma^2 & = & \left|\left|\frac{\hat{Y}-\bar{Y}}{\sigma}\right|\right|^2 & \sim & \chi^2_{\nu_1}(\lambda), \\ SSE_1 & = & SSE/\sigma^2 & = & \left|\left|\frac{e}{\sigma}\right|\right|^2 & \sim & \chi^2_{\nu_2}(0). \end{array} \tag{172}$$

The F-statistic defined in Eq. 131 equals the ratio of SSR_1 to SSE_1 time the normalization factor ν_2/ν_1. By definition, the non-central F-distribution with dimensions ν_1, ν_2, and non-centrality parameter λ is the distribution of a ratio such as this:

$$f \quad = \quad \frac{SSR_1/\nu_1}{SSE_1/\nu_2} \sim \text{non-central } F(\nu_1, \nu_2, \lambda). \tag{173}$$

The mean and variance of the F-statistic under the population distribution are straightforward to calculate in general. Rather than go through the derivation here, we present in Figure 5 a Mathematica (Wolfram Research Inc. 2014) notebook which calculates the mean in a single line of code. The result is

$$E[f] \quad = \quad \frac{\nu_2}{\nu_2 - 2}\left(1 + \frac{\lambda}{\nu_1}\right). \tag{174}$$

Let us bring out focus back to the adjusted R-square statistic. We can now say that it is distributed as a transformation of the non-central F-distribution of the F-statistic by the change of variables:

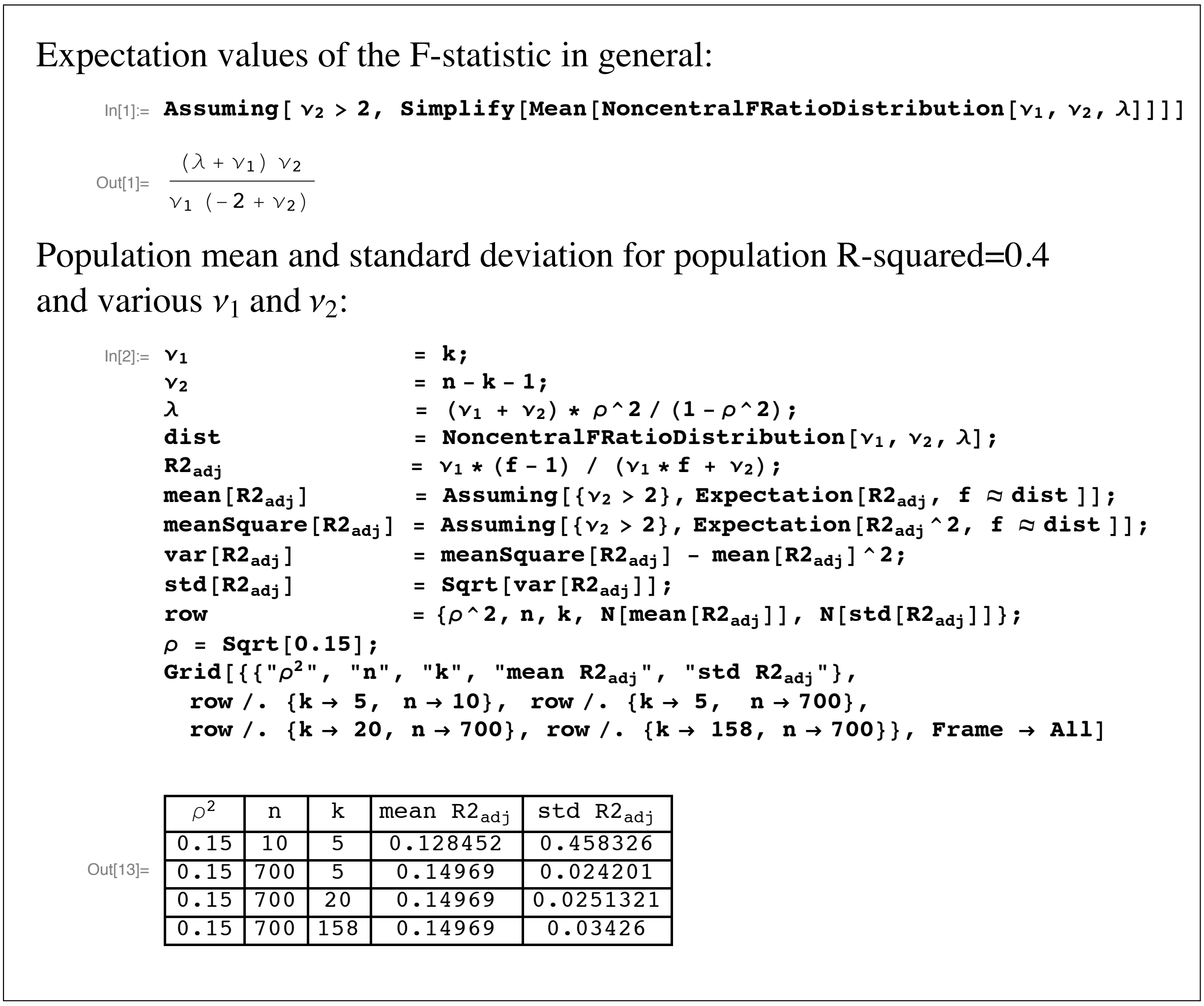
$$R^2_{adj} \quad = \quad \frac{\nu_1(f-1)}{\nu_1 f + \nu_2} = 1 - \frac{\nu_1 + \nu_2}{\nu_1 f + \nu_2}. \tag{175}$$

We know of no closed form solution for the mean of adjusted R-squared in general, but in Figure 5 we show how the mean and standard deviation of adjusted R-squared for particular parameter values can be calculated in just a few lines of Mathematica code. The parameter values we pick have relevance to understanding our results for confidence intervals for the positive regression considered in the body of the paper.

The examples in Figure 5 illustrate that that adjusted R-squared is indeed approximately unbiased, i.e. the mean of adjusted R-squared is close to the population R-squared. Also, as one would expect, the standard deviation decreases as the number of data points increases. When $\rho^2 = 0.15$, $k = 5$, and n increases from 10 to 700 as in the first two rows of the table in the figure, the standard deviation decreases as $n^{-0.7}$, i.e. a little faster than the inverse square root of n. Further, the standard deviation increases very slowly with the number of regressors k. Not shown is that the standard deviation is almost constant as k varies for ν_2 held fixed. Contrast all this with the exact formula for the dependence on n and k when $\rho^2 = 0$, which is given in Equation 130.

E.10 Confidence Intervals for Adjusted R-squared

The distribution of adjusted R-squared for the regression model depends on the sample size n, the number of regressors k, and the population R-squared, ρ^2. Suppose we observe a value, $R^2_{adj,obs}$, of adjusted R-squared for some regression data. One thing we can do is to compute the P-value for significance testing the observed R-squared against the null hypothesis, which is the special case of our population model when ρ^2 vanishes.



Expectation values of the F-statistic in general:

```
In[1]:= Assuming[ ν2 > 2, Simplify[Mean[NoncentralFRatioDistribution[ν1, ν2, λ]]]]
```

Out[1]= $\frac{(\lambda + \nu_1)\ \nu_2}{\nu_1\ (-2 + \nu_2)}$

Population mean and standard deviation for population R-squared=0.4 and various ν_1 and ν_2:

```
In[2]:= ν1                  = k;
        ν2                  = n - k - 1;
        λ                   = (ν1 + ν2) * ρ^2 / (1 - ρ^2);
        dist                = NoncentralFRatioDistribution[ν1, ν2, λ];
        R2adj               = ν1 * (f - 1) / (ν1 * f + ν2);
        mean[R2adj]         = Assuming[{ν2 > 2}, Expectation[R2adj, f ≈ dist ]];
        meanSquare[R2adj]   = Assuming[{ν2 > 2}, Expectation[R2adj^2, f ≈ dist ]];
        var[R2adj]          = meanSquare[R2adj] - mean[R2adj]^2;
        std[R2adj]          = Sqrt[var[R2adj]];
        row                 = {ρ^2, n, k, N[mean[R2adj]], N[std[R2adj]]};
        ρ = Sqrt[0.15];
        Grid[{{"ρ2", "n", "k", "mean R2adj", "std R2adj"},
          row /. {k → 5,  n → 10},  row /. {k → 5,  n → 700},
          row /. {k → 20, n → 700}, row /. {k → 158, n → 700}}, Frame → All]
```

Out[13]=

ρ^2	n	k	mean $R2_{adj}$	std $R2_{adj}$
0.15	10	5	0.128452	0.458326
0.15	700	5	0.14969	0.024201
0.15	700	20	0.14969	0.0251321
0.15	700	158	0.14969	0.03426

Figure 5: Mathematica notebook which calculates:
the expectation value of F-statistic for generic regression population parameters, and examples of the mean and standard deviation of sample adjusted R-squared for population R-squared 0.15.

Another thing we can do is estimate the population R-squared value based on the observed adjusted R-squared. $R^2_{adj,obs}$ itself is a reasonable choice of estimate of ρ^2 since adjusted R-squared is approximately unbiased. However, there are problems with this choice. For example, $R^2_{adj,obs}$ might be negative, whereas ρ^2 is always non-negative.

One approach to getting some feel for a range of possible values for the population parameter, ρ^2, which may have generated the observed value, $R^2_{adj,obs}$, is to use the Neyman-Pearson theory of confidence intervals (Neyman 1937). This approach is standard when estimating simple parameters like individual regression coefficients, but is not so standard when it comes to estimating population R-squared. In the paper "Correct Confidence Intervals for Various Regression Effect Sizes and Parameters: The Importance of Noncentral Distributions in Computing Intervals" (Smithson 2001), Smithson states that "Until recently, these techniques have not been widely available due to their neglect in popular statistical textbooks and software." These techniques are now more accessible because confidence intervals can now be computed both theoretically and by simulation. See for example (Young 2010).

Given a *significance level* α between 0 and 1 and an observed value $R^2_{adj,obs}$, the (two-sided) *confidence interval* for ρ^2 at *confidence level* $100 * (1 - \alpha)\%$ is the range of values of ρ^2 for which $R^2_{adj,obs}$ is in the middle of the probability distribution in the sense that the cumulative probability below $R^2_{adj,obs}$ is in the range $[\alpha/2, 1 - \alpha/2]$. In formulas, this condition reads

$$\begin{aligned} P(R^2_{adj} < R^2_{adj,obs};\ k, n, \rho^2) &>= \alpha/2, \text{ and} \quad (176)\\ P(R^2_{adj} > R^2_{adj,obs};\ k, n, \rho^2) &>= \alpha/2. \quad (177) \end{aligned}$$

The confidence interval is:

$$\begin{aligned} CI(k, n, R^2_{adj,obs}, \alpha) &= [\rho^2_{lo}, \rho^2_{hi}], \quad (178)\\ \rho^2_{lo} &= \min(\{\rho^2; P(R^2_{adj} > R^2_{adj,obs};\ k, n, \rho^2) >= \alpha/2\}), \quad (179)\\ \rho^2_{hi} &= \max(\{\rho^2; P(R^2_{adj} < R^2_{adj,obs};\ k, n, \rho^2) >= \alpha/2\}). \quad (180) \end{aligned}$$

The precise meaning of a confidence interval is a little subtle. One way to state it is as follows:

> For any population parameter ρ^2, there is a 95% probability that ρ^2 will belong to the 95% confidence interval computed from a sample generated by the distribution with parameter ρ^2.

Another way to look at this is to ask the question: What is the range of R^2_{adj} values whose confidence intervals contain a given ρ^2? The low end of this range, $R^2_{adj,lo}(\rho^2)$, is the value of R^2_{adj} which has cumulative probability $\alpha/2$. In other words $R^2_{adj,lo}(\rho^2)$ is the $(100 * \alpha/2)$'th percentile of the distribution with population parameter ρ^2. The high end of the range, $R^2_{adj,hi}(\rho^2)$, is the $(100 * (1 - \alpha/2))$'th percentile. This is illustrated for two different choice of the pair (k, n) in Figure 6.

Figure 7 illustrates the relationship between the significance ranges of R^2_{adj} as ρ^2 varies and the confidence intervals of ρ^2 as R^2_{adj} varies.

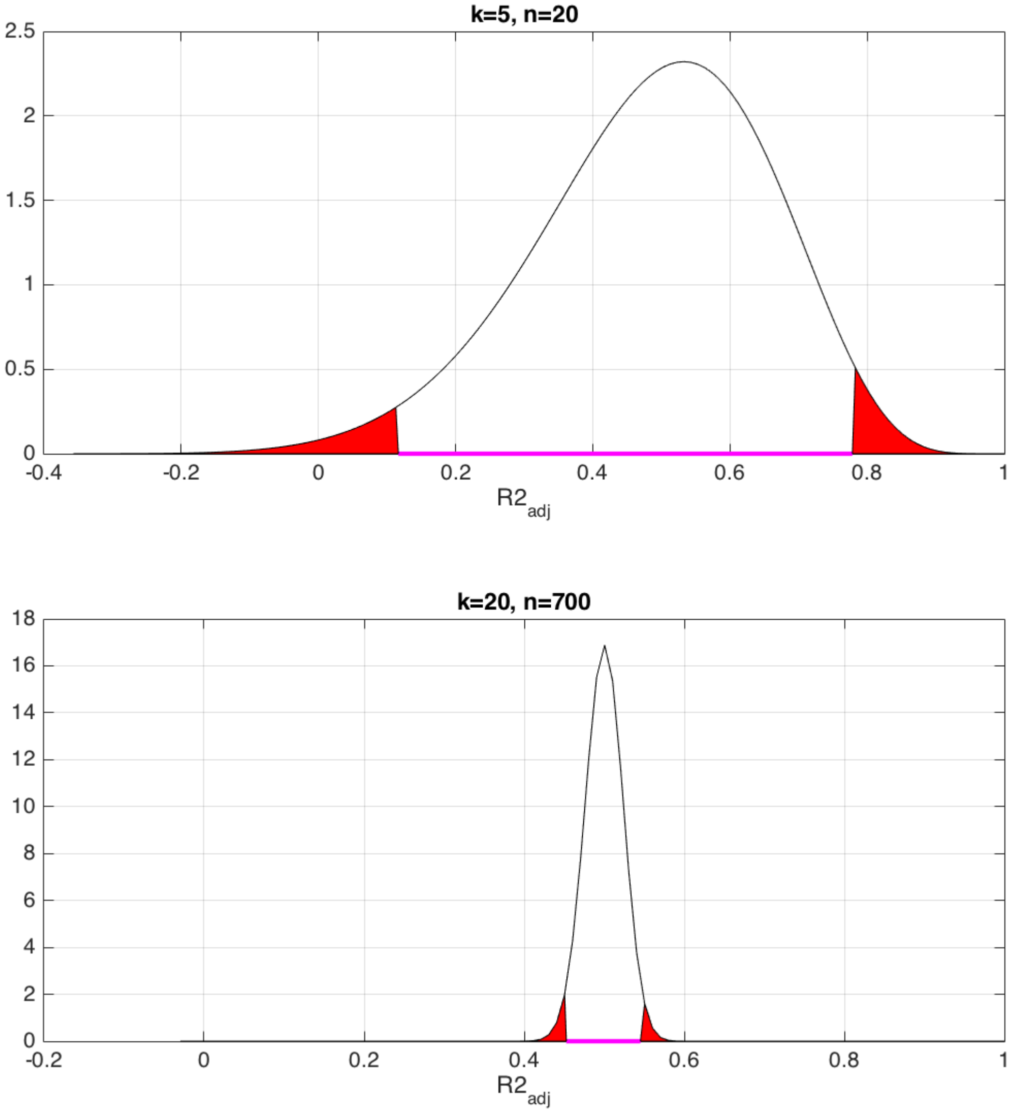


Figure 6: Probability density function for adjusted R-squared when population R-squared parameter ρ^2 is fixed to equal to 0.5. Top plot has $(k, n) = (5, 20)$; bottom plot has $(k, n) = (20, 700)$. Total area in red is $\alpha = 5\%$ of the area under the full curve. The magenta line (on the horizontal axis going from left red region to the right red region) covers the values of R^2_{adj} in the middle of the probability distribution. $R^2_{adj,lo}(\rho^2)$ and $R^2_{adj,hi}(\rho^2)$ are coordinate of the left and right ends of the magenta line.

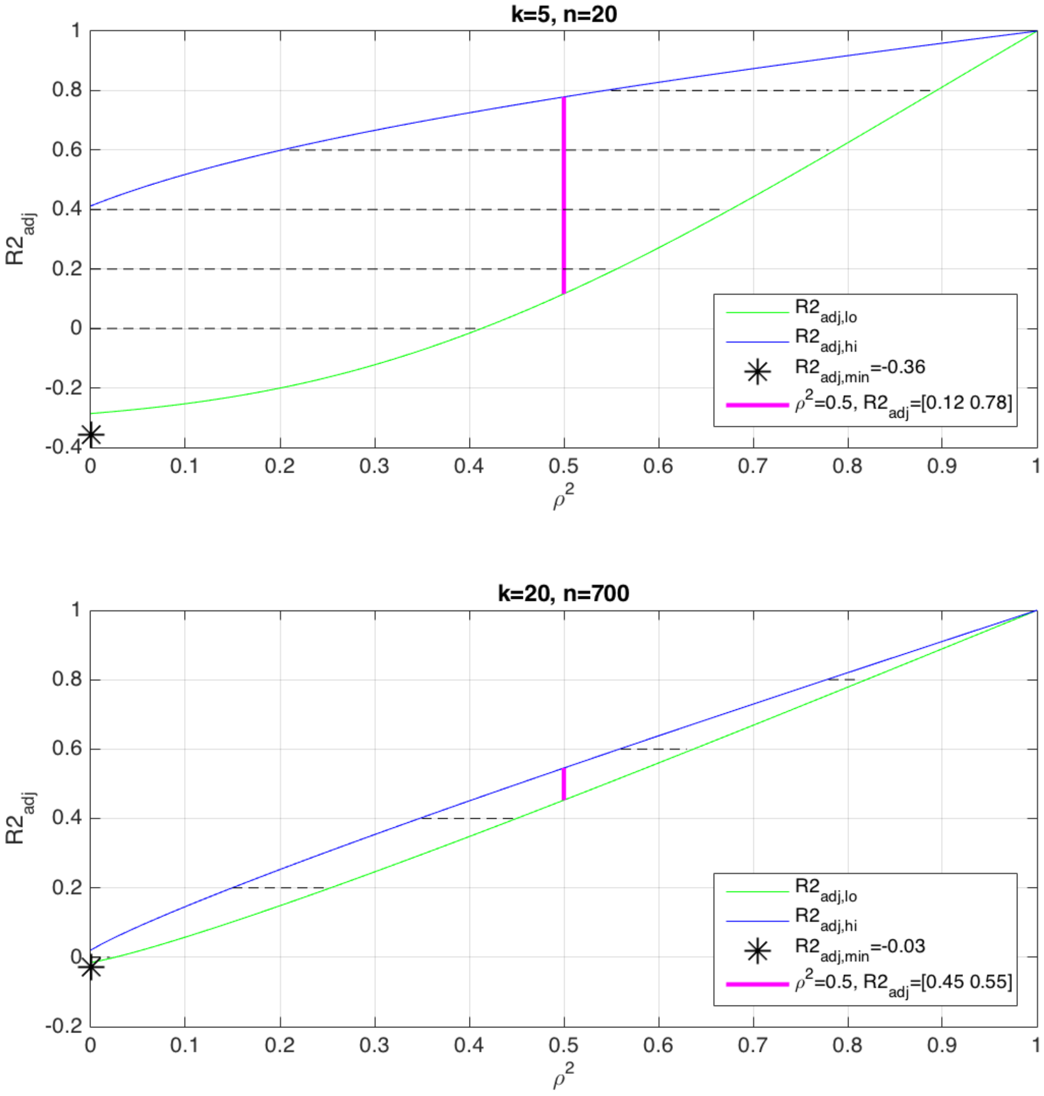


Figure 7: Lower and upper percentile curves at significance level $\alpha = 5\%$ as a function of population R-squared. $R^2_{adj,lo}$ (green curve) is the $\alpha/2$ percentile and $R^2_{adj,hi}$ (blue curve) is the $1 - \alpha/2$ percentile. 95% confidence interval for ρ^2 for a given value of R^2_{adj} is the horizontal range of the line segment at height R^2_{adj} going from the blue curve to the green curve. Examples are drawn as dashed lines. R^2_{adj} values for vertical magenta line at $\rho^2 = 0.5$ are the same as values for the magenta segment in Figure 6.

References

Ahlerup, Pelle, Thushyanthan Baskaran & Arne Bigsten. 2016. "Government Impartiality and Sustained Growth in Sub-Saharan Africa." *Germany World Development* 83:54–69.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kelly McMann, Daniel Pemstein, Svend-Erik Skaaning, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang & Brigitte Zimmerman. 2015a. "Varieties of Democracy: Dataset v5." Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning & Jan Teorell. 2015. "V-Dem Comparisons and Contrasts with Other Measurement Projects." Varieties of Democracy (V- Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Frida Andersson, Kyle Marquardt, Valeriya Mechkova, Farhad Miri, Daniel Pemstein, Josefine Pernes, Natalia Stepanova, Eitan Tzelgov & Yi-ting Wang. 2015c. "V-Dem Methodology v5." Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Frida Andersson, Valeriya Mechkova, Josefine Pernes & Natalia Stepanova. 2015. "V-Dem Organization and Management v5." Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell & Vlad Ciobanu. 2015. "V-Dem Country Coding Units v5." Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skanning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kelly McMann, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Brigitte Zimmerman, Frida Andresson, Valeriya Mechkova & Farhad Mri. 2015b. "V-Dem Codebook v5." Varieties of Democracy (V-Dem) Project.

Darlington, Richard B. & Andrew F. Hayes. 2017. *Regression Analysis and Linear Models Concepts, Applications, and Implementation.* New York, NY: Guilford Press.

Dimson, E., Paul Marsh & Mike Staunton. 2002. *Triumph of the Optimists: 101 Years of Global Investment Returns.* Princeton, NJ: Princeton University Press.

Fisher, R.A. 1915. "Frequency distribution of the value of the correlation coefficient in samples from an indefinitely large population." *Biometrika* 10:507–521.

Fisher, R.A. 1921. “On the ’probable error’ of a coefficient of correlation deduced from a small sample.” *Metron* 1:1–32.

Fisher, R.A. 1924. “The distribution of the partial correlation coefficient.” *Metron* 3:329–332.

Fisher, R.A. 1950. *Statistical Method for Research Workers, 11th ed.* London: Oliver & Boyd.

Grömping, Ulrike. 2010. “Inference with Linear Equality and Inequality Constraints Using R: The Package ic.infer.” *Journal of Statistical Software* 33(1):1–31.

Hamilton, James D. 1994. *Time Series Analysis.* Princeton, NJ: Princeton University Press.

Hogben, David. 1968. “The Distribution of the Sample Correlation Coefficient With One Variable Fixed.” *Journal of Research of the National Bureau of Standards* 72B(1).

Hotelling, Harold. 1953. “New Light on the Correlation Coefficient and its Transforms.” *Journal of the Royal Statistical Society* 15(2):193–232.

Knutsen, Carl Henrik. 2014. “Income Growth and Revolutions.” *Social Science Quarterly* 95(4):920–937.

Leitner, James & Scott Axelrod. 2016. “Perspectives on Corporate Social Responsibility.”. Presentation at the UN Global Compact Office, New York, NY.
URL: `http://www.falconmgt.com/papers/csr_vdem_un2016.pdf`

MathWorks Inc. 2015. *MATLAB 8.6.* Natick, MA.
URL: `http://www.mathworks.com`

MSCI data index and analytics service. .
URL: `https://www.msci.com`

Neyman, J. 1937. “Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability.” *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236(767):333–380.

Nimon, Kim F., Linda R. Zientek & Bruce Thompson. 2015. “Investigating bias in squared regression structure coefficients.” *Frontiers in Psychology* 6:949–959.

Rao, C.R. 1965. *Linear Statistical Inference and Its Applications.* John Wiley and Sons.

Rao, C.R. 2001. *Linear Statistical Inference and Its Applications, 2nd Edition.* John Wiley and Sons.

Salh, Dr. Samira Muhamad. 2015. “Estimating R^2 Shrinkage in Regression.” *International Journal of Technical Research and Applications* 3(2):1–6.

Smithson, Michael. 2001. “Correct Confidence Intervals for Various Regression Effect Sizes and Parameters: The Importance of Noncentral Distributions in Computing Intervals.” *Educational and Psychological Measurement* 61(4):605–632.

Soper, H.E. 1913. “On the probable error of the correlation coefficient to a second approximation.” *Biometrika* 9:91–115.

Student. 1908. “Probable Error of a Correlation Coefficient.” *Biometrika* 6:302–310.

Sundström, Aksel, Pamela Paxton, Yi-ting Wang & Staffan I. Lindberg. December 2015. “Women’s Political Empowerment: A New Global Index, 1900-2012.” Varieties of Democracy (V-Dem) Project.

Wherry, R.J. 1931. “A new formula for predicting the shrinkage of the coefficient of multiple correlation.” *Annals of Mathematical Statistics* 2:203–224.

Wolfram Research Inc. 2014. *Mathematica 10.0.* Champaign, IL.
URL: `http://www.wolfram.com`

Yeşin, Pinar. 2016. “Exchange Rate Predictability and State-of-the-Art Models.” Swiss National Bank, SNB Working Paper 16-02.

Yin, Ping & Xitao Fan. 2001. “Estimating R^2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods.” *The Journal of Experimental Education* 69(2):203–224.

Young, Derek S. 2010. “tolerance: An R Package for Estimating Tolerance Intervals.” *Journal of Statistical Software* 36(5).