BROOKINGS

December 2022

Center on Regulation and Markets Working Paper #5

Non-Euclidean statistics beyond linear regression

Aaron Klein and Joel Levine

This working paper is available online at: https://www.brookings.edu/series/center-on-regulation-and-markets-working-papers/

Center on Regulation and Markets at BROOKINGS

The Center on Regulation and Markets at Brookings creates and promotes rigorous economic scholarship to inform regulatory policymaking, the regulatory process, and the efficient and equitable functioning of economic markets. The Center provides independent, non-partisan research on regulatory policy, applied broadly across microeconomic fields.

Disclosure

The Brookings Institution is financed through the support of a diverse array of foundations, corporations, governments, individuals, as well as an endowment. A list of donors can be found in our annual reports published online here. The findings, interpretations, and conclusions in this report are solely those of its author(s) and are not influenced by any donation.

Non-Euclidean statistics beyond linear regression[1]

Aaron Klein
Senior Fellow
Economic Studies
Brookings Institution

Joel Levine
Professor Emeritus
Quantitative Social Sciences
Dartmouth College

This paper considers the application of one type of advanced statistical model using three examples to demonstrate its ability to detect relationships in data that does not initially present either a numerical or ordinal scale. These three examples, socioeconomic status, movement of stock prices over time, and voting patterns within Congress, do not have any inherent relationship to each other. The three have been studied and analyzed together to show the ability of this model to detect interesting relationships.

This AI model uses non-Euclidean statistical techniques to infer quantitative scales for variables, to estimate metric distances among attributes, and to evaluate the fit of the model in every cell of the tabular data by chi-square or other devices appropriate to the data. The model does not use linear regression as part of its techniques. Linear regression has been the traditional workhorse of econometrics and other empirical fields in social science. By avoiding the use of regression techniques and statistics, this model has the potential to uncover new insights.[2] It also opens up a broader question of where future analysis can migrate as mathematical and computational abilities create opportunities for more sophisticated analytical tools than regression analysis.

[1] Computer code and assistance are available from the authors.

[2] A linear model with 12 columns would be asked to fit 12 column means. By contrast, for data with 12 columns and 6 rows this model is asked to fit 74 cell values, using every detail of the tabulation. This allows the model to extract more information from the data and, perhaps counter-intuitively, this allows the model to infer a more-orderly-than-anticipated overview of the data.

Example 1: Social economic status

We begin with two well-researched and correlated elements of social economic status: educational attainment and income. Our data are shown in Exhibit 1, without usable labels (by which a human would be able to guess at order) and without numbers that a human could identify as a scale.

Exhibit 1:
Joint Distribution of Two Sets of Attributes — In Random Order with Labels Removed

Data with labels and order suppressed	l	b	a	d	e	c	j	i	f	k	g	h
A	0	41	22	6	6	120	0	2	11	0	3	0
D	1	32	20	25	34	379	1	26	131	2	30	1
E	1	22	10	28	29	253	0	77	195	8	53	3
C	0	45	27	20	20	360	0	17	62	0	16	1
B	1	34	20	12	11	173	0	4	30	0	3	1
F	5	5	1	3	5	32	1	34	54	5	17	2

Our assumptions, however, will make statistic-like assumptions about the structure of these data. Specifically, while making sure that this hypothesis is falsifiable, we hypothesize that there exist x's for the rows and x's for the columns setting up a standard matrix. We also hypothesize that these numbers exist in a one-dimensional space, and we define distances (row *i* to column *j*) as the absolute value of the difference between the unknown x's for row i and column j, Equation 1.[3]

$$distance_{row_i, column_j} = \left| x_{row_i} - x_{col_j} \right| \qquad (1)$$

Mathematically, this initial model is simple. But as a hypothesis about the world behind the data it requires that the structure of the real world itself be simple—as the 72 row-to-column distances will have to be derived at the cost of only 15 degrees of freedom (the 15 intervals among the 16 objects in the space).[4] We hypothesize that such distances exist but are unknown.

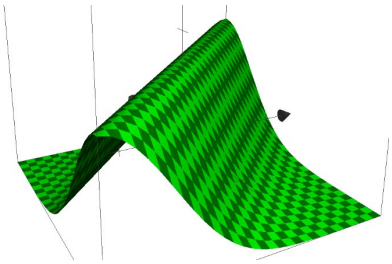
If these distances and their space exist, we can find them by asking how they would manifest themselves in the data? Initially we suggest that when the distance between a row object and a column object is short, the joint frequency of that row and that column will be high and, more specifically, that the frequencies will attenuate as a negative exponential of distance with attenuation parameter *a*.

[3] The second and third examples generalize to multiple dimensions and not-necessarily-Euclidean metrics.

[4] The number of rows minus 1, plus the number of columns minus 1, minus 1.

$$Frequency_{ij} \approx 2^{-distance_{ij}{}^{a}} \quad (2)$$

shown with $a = 2$



If Equation 2 'worked' it would be elegant: It would mean that the row attributes and column attributes could be quantified. It would mean that the initially unquantified attributes are linked by a linear central tendency (the ridge of Equation 2). And it would mean that even the deviations from the central tendency were orderly. We would give our program the ability to use this model by encoding the model as a function and allowing the program to move rows and columns of Exhibit 1 (i.e., change the x's) until the re-arranged data matched the pattern of Equation 2 (with the match evaluated by chi-square error).

But Equation 2 does not work. All six frequencies in column c (Exhibit 1) are high compared to all six frequencies in column l, making it apparent that Equation 2 cannot match.

However, Equation 2 can be augmented by row and column effects used as multipliers, R_i and C_j. $Equation\ 3$. In this form the model has five parts—a constant effect, row effects, column effects, interaction effects, and the inevitable residuals. The program can estimate these effects using the chi-square statistic as an objective function, Equation 4.

$$F(i,j) \approx R_i C_j 2^{-|\,distance_{i,j}|^{a}} \quad (3)$$

$$\text{with } distance_{i,j} = \left|x_{row\ i} - x_{col\ j}\right|$$

$$\text{Chi-Square Error} = \sum_{i,j} \frac{\left(F_{ij} - \hat{F}_{ij}\right)^2}{\hat{F}_{ij}} \quad (4)$$

We hypothesize that this augmented relationship is appropriate to the data and let the program search for the parameters that create the best fit. Through computational brute force, it evaluates the goodness of fit with which the initial parameters fit the data.[5] It increases or decreases the parameters, one at a time—deciding which is the better value. By repeating these small steps tens of thousands of times, within seconds of real time, it arrives at a solution.[6]

[5] Demonstrated in Appendix 1.

[6] Even so, the run time of large examples runs to hours and days

Executing the search, the model reaches a least chi-square of 37.24. The result is good:[7] Using the standards for null hypotheses as a rule of thumb by which to evaluate the hypothesis, *if* chi-square with 39 degrees of freedom is within the interval 39 ± 8.83 (± one standard deviation) or 39 ± 17.66 (± two standard deviations)—which it is—then there is no compelling evidence of error.

Because positive hypotheses cannot be accepted or rejected (as the mirror image of null hypotheses) this close fit does not *prove* the positive hypothesis. But the close fit demonstrates that the positive hypothesis is sufficient to exhaust the information present in these data. The solution is a good match.

The fitted values for the data displayed in Exhibit 2 show how the model has quantified the relationships between income and education. The data on top of the cell shows the frequency of the observed pairing with the model generated fitted frequency below it. Next to each row and column are shown the plot point in a uni-dimensional axis. Those points are then shown in Exhibit 3 which displays the model's solution on a uni-dimensional axis that effectively pairs income with educational attainment.[8] Properly labelled, the augmented model has 'figured out' that education and income are two scalable variables and that they have a strongly correlated linear relation whose common attribute is social economic status.

Ordinarily the overview of a study of education and income begins with the overview already in hand, prior to the analysis. The variables are in hand and the research adds detail and assess the strength of the relation.

By contrast, here our advanced model, with its data-fitting metric space, infers much of the structure of the variables and their attributes directly from the data.

It is the attempt to fit all of the cells that allows us to extract information for which regression requires exogenous sources or assumptions regarding the relationship between the data (integers, ordinality, etc.). It is the goodness of fit achieved that demonstrates that what the model has inferred is consistent with the data.

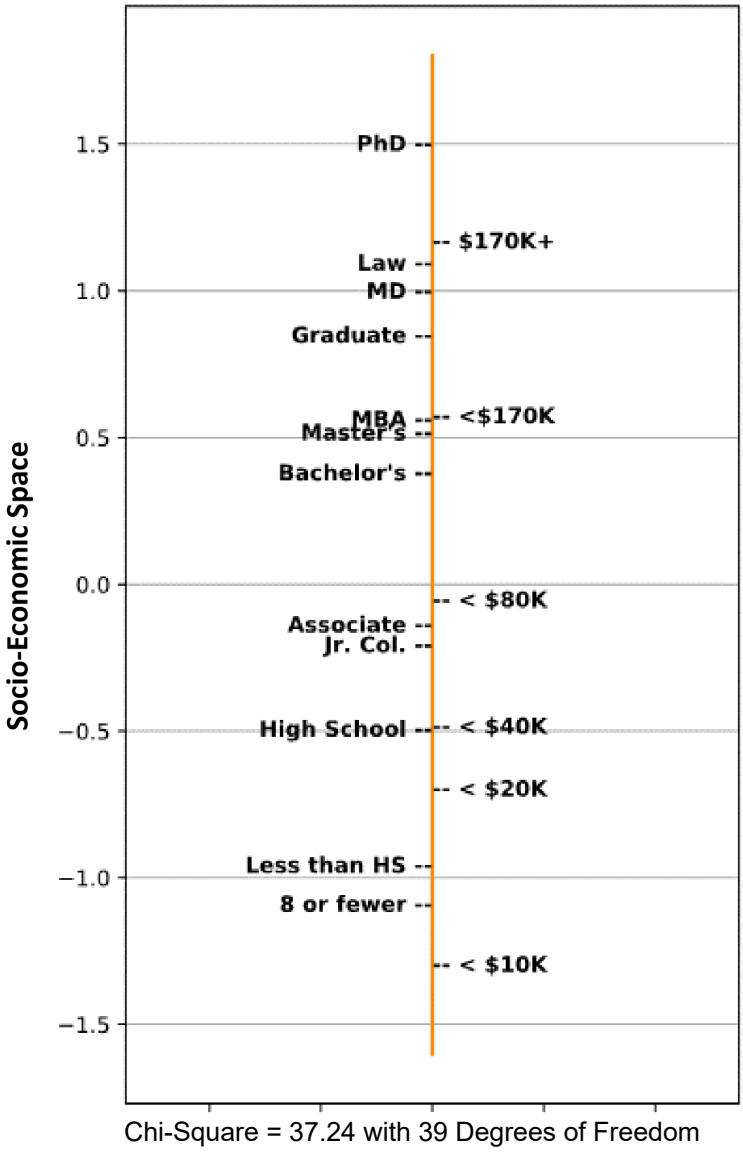
[7] For simplicity, we have started with a = 2, the power of distance as it would be if he data were bivariate normal. The resulting makes it clear that either "2" is correct or else the remaining information in the residuals is insufficient for a better estimate.

[8] For simplicity we started with the power of distance, a, set equal to 2, as it would be if the data were bivariate normal. Having then discovered that the close fit achieved with $a = 2$ exhausts the information in the data (low chi-square) , we have left the estimate at *a* = 2.

Exhibit 2:
Frequencies, Fitted Frequencies, Estimated Multipliers and Estimated Coordinates

All multi-plier														
4.25		Col Multipli-ers	5.771	7.806	23.818	1.186	1.231	4.176	1.049	0.077	1.503	0.032	0.138	0.286
		X Coordi-nates	1.677	1.510	0.931	0.572	0.486	-0.160	-0.331	-0.388	-0.747	-0.934	-1.054	-1.562
Row Multi-pliers	X coor-dinates	Observed and Fitted Values	8 or less	Less than high school	High school	Junior college	Associ-ate's	Bachelor's	Master's	MBA	Graduate	MD	Law	PhD
0.307	0.932	Under $10 thousand	22	41	120	6	6	11	3	0	2	0	0	0
			21.73	34.23	131.66	5.99	5.93	10.09	1.92	0.13	1.18	0.02	0.05	0.02
0.492	0.453	Under $20 thousand	20	34	173	12	11	30	3	1	4	0	1	0
			18.12	31.88	180.25	10.41	10.90	28.51	6.07	0.42	4.91	0.07	0.25	0.15
1.068	0.282	Under $40 thousand	27	45	36	20	20	62	16	1	17	0	0	0
			28.86	52.79	342.52	21.53	23.02	70.16	15.55	1.09	13.89	0.22	0.77	0.52
1.691	-0.063	Under $80 thousand	20	32	379	25	34	131	30	1	26	1	1	2
			21.62	42.81	366.55	27.34	30.46	126.50	30.43	2.19	33.14	0.57	2.13	1.84
2.853	-0.564	Under $170 thousand	10	22	253	28	29	195	53	3	77	0	1	8
			9.15	20.33	260.47	24.93	29.47	191.83	51.96	3.89	75.53	1.49	6.00	7.37
1.288	-1.041	$170 thou-sand or over	1	5	32	3	5	54	17	2	34	1	5	5
			0.80	1.99	37.37	4.53	5.67	56.64	17.18	1.34	32.87	0.73	3.20	5.50

Exhibit 3: Social Economic Status Data Space Educational Degree by Family Income (6x12)



Chi-Square = 37.24 with 39 Degrees of Freedom

The model's ability to recreate the known rank order of impact with unknown data is a significant validation of the overall approach. It is also a reminder that rank order data is commonly misused in linear regression. Linear regression analysis requires data that are integers. Assigning rank order values to numbers does not imbue integer status. Integers are by definition evenly spaced between each other. If a college degree is given the number 3, high school 2, and graduated school 4, the assignment does not imbue integer status on the actual relationships. However, it forces the mathematical models to treat the differences as having constant value between accomplishments.

However, perhaps for lack of a solution, many social scientists seem to wave away this problem and treat ordinal data as integers.[9] For example, ordinal data is represented by integers but using the properties of real numbers. This can be seen in the survey process of assigning a rank order value of 1-10 in answer to a standardized question produces a set of rank order data. For each person, an answer of eight is greater than that of seven and less than that of nine. However, the difference between seven and eight is not necessarily equal to that between eight and nine. Thus, the mathematical operation of averaging, or computing the sum of squares of error produced by the line of best fit, is not valid. It can be done; many mathematical operations can be performed, regardless of whether they make sense. Ask a computer to average 'satisfied,' 'happy,' 'ecstatic,' 'apoplectic,' and there will be no answer. Assign those values five, seven, nine, and two, and the computer will spit back an average of 5.75. That hardly makes the average experience of the group that level.

Put another way, an exciting application of this solution is that it is no longer necessary to assume that 'happy' is one unit greater than 'satisfied' and one unit less than 'ecstatic'. That is not in the data. And the solution allows us to estimate the scale from the evidence as people experience it.

Example 2: Time series without time

For a second demonstration, we move to a different target: *Time*. Whereas the prior variables of education were not ordinal, time is. Time is a variable on a uniform scale, the distance between days remains constant. We withhold this information from our model, which is challenged with the data and the requirement to order it, without the information that the variable being searched exists on an interval scale.

The data are shown schematically in Exhibit 4. They provide twenty years (5,033 dated columns) of prices for 352 components of the S&P 500 over 20 years. These are the 352 members of the present 'Standard & Poor's 500'—for which there is a 20-year record. For these data, stripped of their labels, time will be treated as an unknown that needs to be inferred from the data.

[9] Liddell, Kurschke. "Analyzing ordinal data with metric models: What could possibly go wrong?" *Journal of Experiment Social Psychology* 79 (November 2018).

Exhibit 4: Excerpt for stock prices in time series

S&P 500	1498.58	1505.97	1494.73	1487.37	1501.34	…	$2526.90
A	$74.39	$70.10	$66.88	$69.34	$75.10	…	$72.29
AAPL	$4.85	$4.76	$4.55	$4.66	$4.47	…	$244.93
ABC	$3.75	$3.66	$3.89	$4.01	$4.20	…	$83.89
ABT	$15.75	$16.56	$17.65	$18.35	$16.92	…	$79.44
ADBE	$27.83	$26.39	$25.44	$26.80	$28.66	…	$303.96
ADI	$80.50	$74.19	$71.00	$72.97	$78.81	…	$87.70
ADM	$9.82	$10.00	$10.06	$10.12	$10.24	…	$34.33
ADP	$38.06	$39.44	$40.23	$39.59	$41.36	…	$131.55
…	…	…	…	…	…	…	…

Withholding the known dates and applying our model calibration to the assumed-to-be-unknown variable time, the procedure is identical to that of the first example except for goodness of fit. For price data, goodness of fit has been measured by least squares applied to natural log prices and fitted log prices. With this change the result is shown in Exhibit 5. (For these data the procedure demonstrated in Appendix I continues to apply.)

For this result the point of interest is the overview. The remarkable thing about Exhibit 5 is that it exists at all: Instead of using physical time to mark the data prior to analysis, “market time” has been inferred directly from behavior. As ordinary physical time is inferred from physical behavior, the graph indicates that behavioral time—the time period it takes for behavior to change—exists, and that behavioral time can be inferred from behavioral data. While market averages moved up and down during these years (Exhibit 5), market time (generally) moved ‘forward,’ albeit at different rates relative to physical time.

This result shows that market time exists, and it is quantifiable. What “market time” *is* is a rate of differentiation: If stock #1 is moving up by at a certain rate per unit of time and stock #2 is moving down by a certain rate per unit time, then the spread between the two increases over time. The search has extracted “time” from the spread. If “market time” increases positively with physical time, then the differentiation is increasing, on average, across the data.

In 2008-2009, as the market average fell, while “time” or “differentiation” accelerated compared to the rate of differentiation that had prevailed for the previous eight years. For a few months in 2009, the differentiation reversed—it went backward as shown by the decline just after the local peak during the middle period in the graph—after which the long-term trend of differentiation resumed the trend of the previous nine years. This demonstrates the model’s ability to detect not only the shock of the financial crisis but the change in the speed of the market itself, which shifted during the financial crisis. The AI has inferred from data in an objective process something that market participants speak of as

anecdotal inference: during the 2008 financial panic days felt like months and months felt like years.

Exhibit 5:
Market Time versus Clock Time

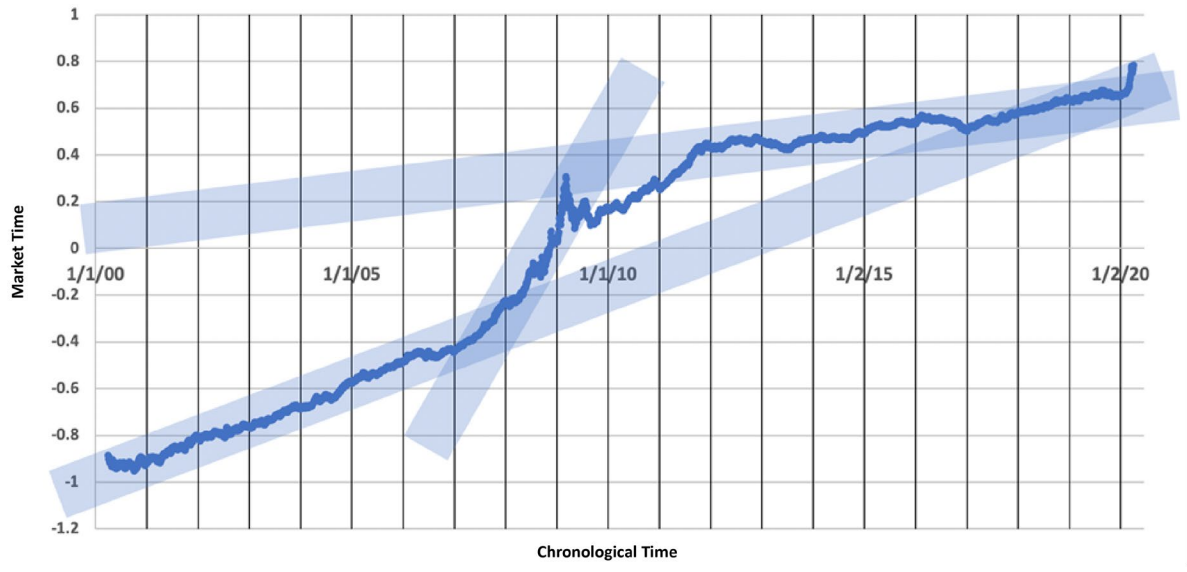
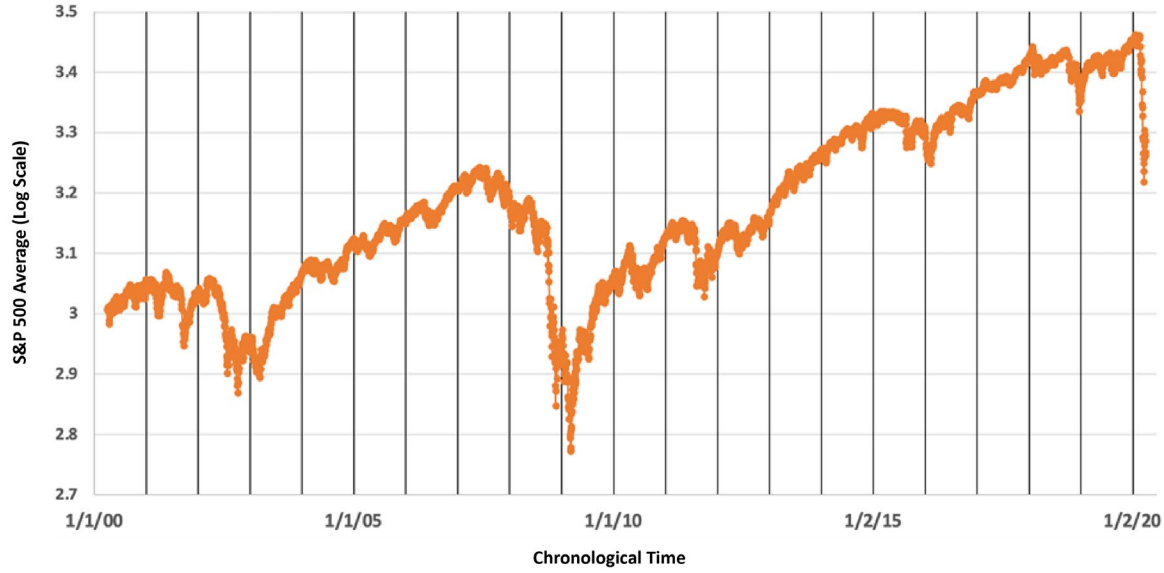


Exhibit 6:
Standard & Poor's 500 Average (Log Scale) versus Clock Time



Two-dimensional time series

For initial exploration it is useful to demonstrate that AI-based quantification is consistent with or not in conflict with traditional quantification. But it is also useful to take an initial peek into the future, into what mathematics and data analytics might discover that traditional quantification cannot. For these data, they could expand time series into two dimensions—with interesting initial results.

Expanding time series to 2 dimensions and generalizing distance to Minkowski metrics (that include the Euclidean metric), we generalize distance to the definition in Equation 5.

$$distance_{row\ i,\ col\ j} = \left[\sum_{d=1}^{D} \left| x_{row\ i,d} - x_{col\ j,d} \right|^{M}\right]^{1/M} \qquad (5)$$

For dimensions 1 through dim, and Minkowski parameter M > 0.[10]

With this extension, the same model using the same equations as earlier, with the same chi-square optimization function, is asked to search the data for empirical evidence of the two-dimensional theoretical pattern described by the equation. And the result is extraordinary, as shown in Exhibit 7. On the time scale of 20 years, the market has redistributed value with great regularity, with long trends punctuated by orthogonal change.

The direction of the market can be thought of as the set of public companies whose equity prices are appreciating at different speeds relative to the market average. This is constant over long periods of time. The model depicts these as nearly straight lines on the graph. For two-plus years, from March 2000 into the third quarter of 2002 stocks shown on the left side of the space were gaining relative to stocks on the right (second and third quadrants over the first and the fourth). Whether by coincidence or by cause, the market continued in this direction until it was interrupted, after which the market turned abruptly in a new direction (with almost orthogonal sets of winners and losers).

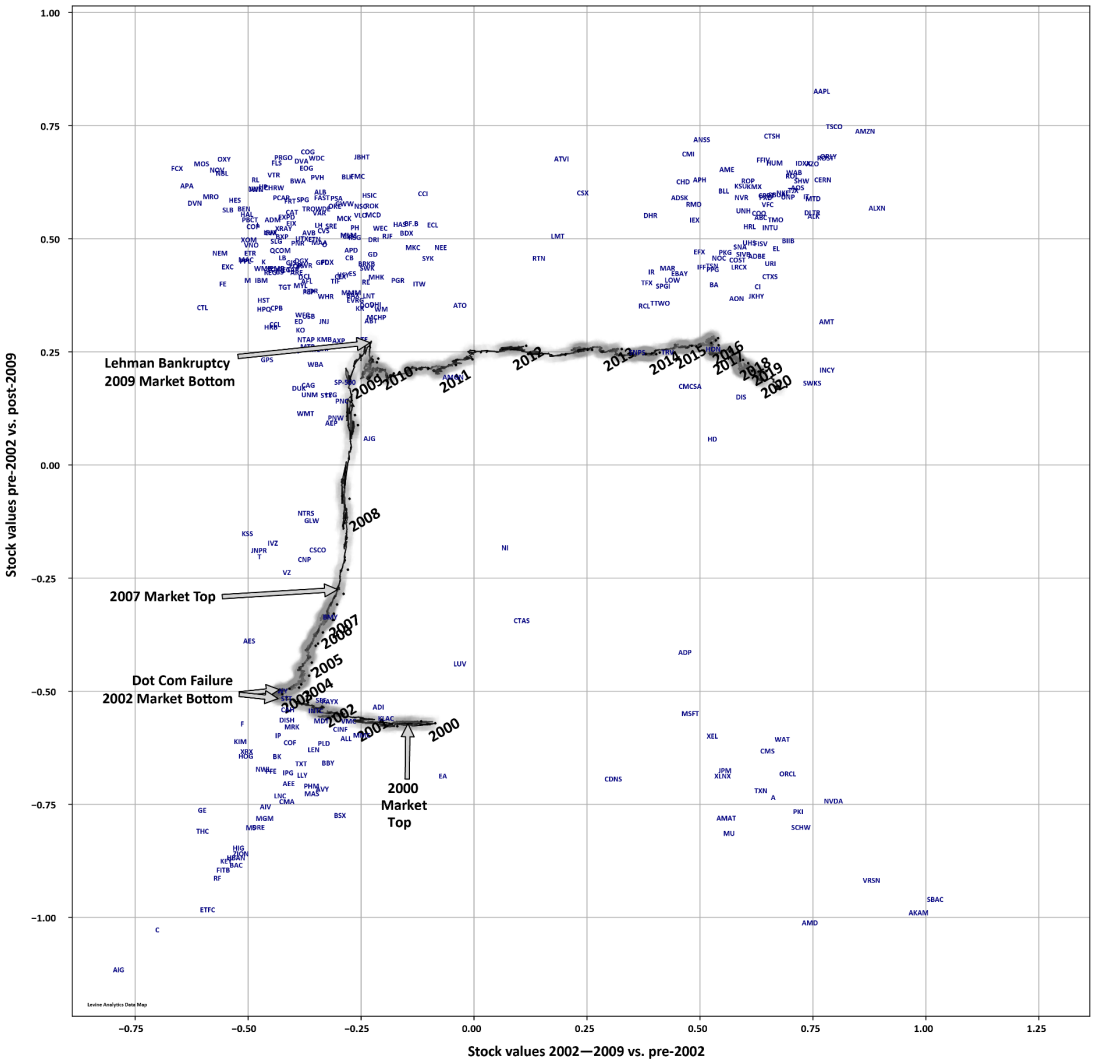
This period corresponds with the dot-com crash and accounting scandals, both of which had major impacts on equity prices. The magnitude of the accounting scandals, sometimes referenced by two of the largest accounting failures of Enron and WorldCom, is often overlooked. However, the decline in the S&P during the accounting scandal period is comparable to that of the other major equity crashes during the twenty-year period the data covers.

After these corrections, the market maintained a new direction for 7 years until again it was interrupted by a crash (the financial panic of 2008), after which it abruptly executed another sharp turn. Post financial panic, the market then changed again, engaging in a uni-directional movement for the next 11 years.

[10] Technically, for M>=1 the result is a distance, while 0<M<1 is a semi-distance for which the triangle inequality fails.

With a nod to Stephen Jay Gould's 'punctuated evolution,' the evidence suggests that market 'evolution' during these years has been punctuated: For long intervals, 2 years, 7 years, and 11 years, market movement was stable in one direction until, abruptly and coincident with a crisis, it changed direction, and returned to gradual 'evolution' in a new almost-orthogonal direction.[11]

Exhibit 7: The Time Line
Estimate Coordinates Describing the Twenty Year Trend of Daily Prices or S&P 500 Stocks that have a Twenty Year Record



Sum of squared log errors = 87,315.646, $a \approx 5.523$, $M \approx 325.651$

[11] The sum of squared errors is 87,315.6, (with attenuation $a \approx 5.523$ and Minkowski metric M ≈ 325.651. For comparison, the sum of squared errors from 352 regressions (one for each stock) is 204,719.3. If the data are 'corrected' for changes in the S&P average (dividing stock prices by the S&P) — the sum of squared errors from the 352 regressions is reduced to 179,700.6. Thus, the error shown here is approximately half (49%) of the comparable regression error.

Example 3: The two-dimensional Senate

The third demonstration is applied to the quantification of partisanship in the current U.S. Senate: The U.S. Senate and, presumably, the data for the Senate will show partisan polarization, Republicans versus Democrats—a one-dimensional split.

It would hardly be an accomplishment should our analysis concur with this partisan split. The question for this example is whether AI quantification can show more.[12]

Our data for this question is the record of all 241 roll call votes for the 2021 U.S. Senate, January 6, 2021, through June 17, 2021, shown schematically in Exhibit 8.

Exhibit 8:
U.S. Senate Roll Call Votes ("Yea", "Nay", or not voting)
January 6, 2021 through June 17, 2021

	Senate Vote #1		Senate Vote #2		Senate Vote #3		Senate Vote #4		• • •	Senate Vote #241	
	Yea	Nay	Yea	Nay	Yea	Nay	Yea	Nay		Yea	Nay
Sen. Maria Cantwell [D]	0	1	0	1	1	0	1	0	....	1	0
Sen. Thomas Carper [D]	0	1	0	1	1	0	1	0	....	1	0
Sen. Susan Collins [R]	0	1	0	1	1	0	0	1	....	1	0
Sen. John Cornyn [R]	0	1	0	1	1	0	1	0	....	0	1
Sen. Michael "Mike" Crapo [I	0	1	0	1	NA	NA	1	0	....	0	1
Sen. Richard Durbin [D]	0	1	0	1	1	0	1	0	....	1	0
Sen. Dianne Feinstein [D]	0	1	0	1	1	0	1	0	....	1	0
• • • •	....	....	....	....	....	....	....	....	....	....	....
Vice President Kamala Harris	NA	NA	NA	NA	NA	NA	NA	NA	....	NA	NA
Sen. Alejandro "Alex" Paddilla	NA	NA	NA	NA	1	0	1	0	....	1	0
Sen. Jon Ossoff [D]	NA	NA	NA	NA	1	0	1	0	....	1	0
Sen. Raphael Warnock [D]	NA	NA	NA	NA	1	0	1	0	....	1	0

Source: Civic Impulse, LLC: www.govtrack.us, June 17, 2021

For this analysis we think of each roll call as a split into two coalitions, a coalition of the "yeas" and a coalition of the "nays," forming 482 coalitions for the 241 roll calls. We think of each senator as a coalition of size 1 that intersects with some of these 482 coalitions. We think of the data as frequencies—as the number of links between each senator and each of the 482 coalitions. With these 0's and 1's as frequencies we continue to use chi-square as the objective functions, as it was used for the frequencies in the social economic status data. The same equations from the prior models are utilized, just changing the data to senators and their votes.

[12] See Hanson, 1998 for an earlier attempt at similar analysis using a more primitive AI given the computing power available at that time.

Moving past the obvious (that the Senate is polarized) the two-dimensional quantification is shown in Exhibit 9. This breaks down the senators on the basis of their votes, showing clear cleavages into two major partitions, Democrats and Republicans.

Here as in the previous example, the point of interest is the overview: The overview is that there *is* a second dimension. The first dimension (horizontally) suggests partisanship dividing the Senate. One can clearly see the Democrats clustering on one side and the Republicans on the other. The second dimension (vertically) suggests factionalism that divides the parties, with differences observed among each party on that dimension.

Between parties—between an approximate middle of the Democrat's cluster and an approximate middle of the Republican's cluster—the distance is roughly two units (in this metric). Within parties, and particularly within the Republican, the distances between the outlying factions and the party leaders is of the same order of magnitude.

Among Democrats, Exhibit 10 (right-hand side of the first map) factionalism is clear but limited when it comes to a vote. Factions are clear, with two well-known liberal senators, Warren (D-RI) and Sanders (I-VT), at one end and party leadership near the other. But when it comes to these recorded votes, the division is limited (with comparatively little variation).

Among Republicans, Exhibit 11 (left-hand side of the first map) the factional distance between the self-described libertarian[13] Senator Paul (R-KY) and Republican leadership is of the same order of magnitude as the distance between Republicans and Democrats.[14]

[13] Hartsoe, Steve. "Rand Paul on How Libertarian Philosophy Can Connect Divided Partisans." *Duke Today*, November 9, 2018. https://today.duke.edu/2018/11/rand-paul-how-libertarian-philosophy-can-connect-divided-partisans

[14] The Senate identification number of each vote and each coalition of "yeas" and "nays" is also indicated on the map. At the cost of legibility, these have been shown in their true positions, showing the densely packed clumps of apparently politically equivalent roll calls.

Exhibit 9:
Partisanship and Factionalization: The Two Dimensional Senate

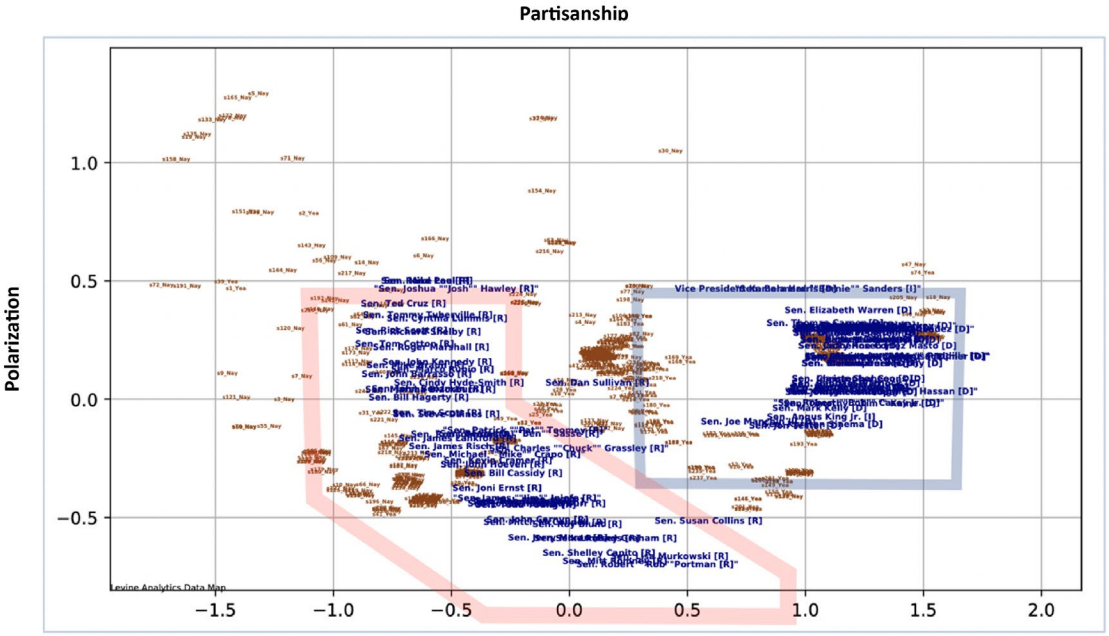


Exhibit 10: Democrats

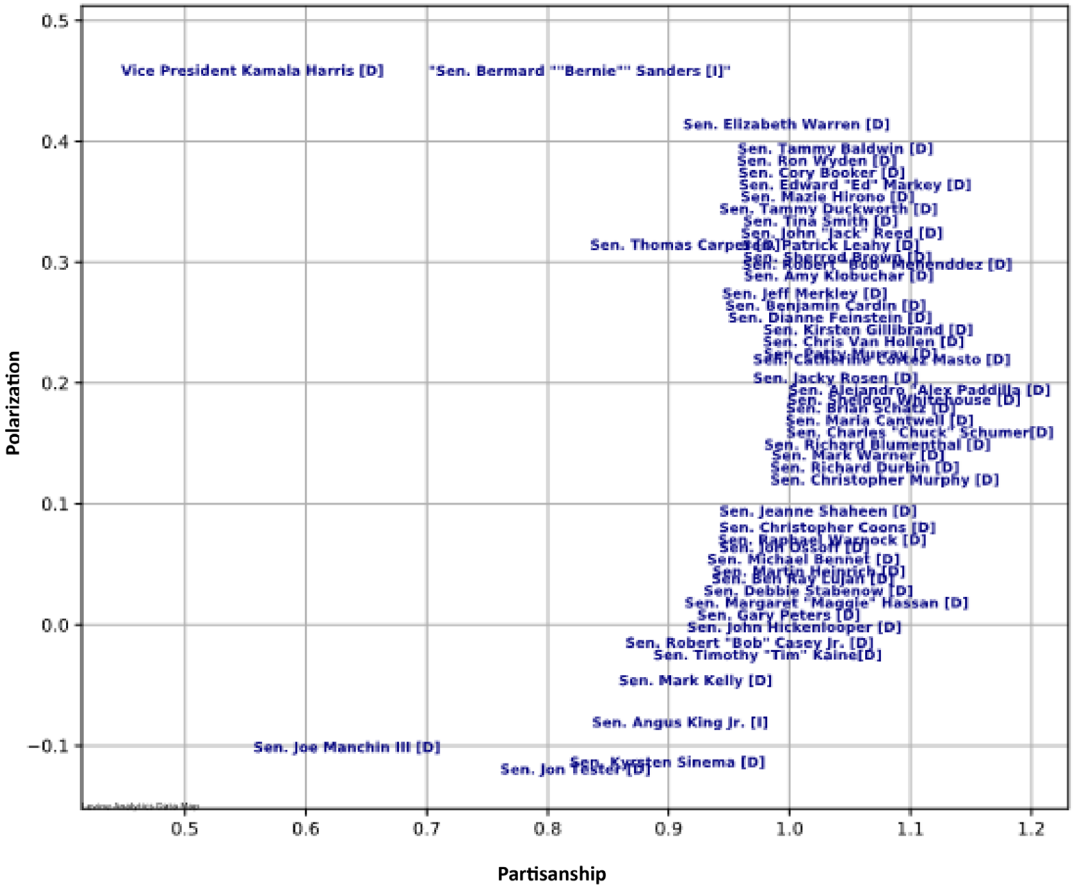
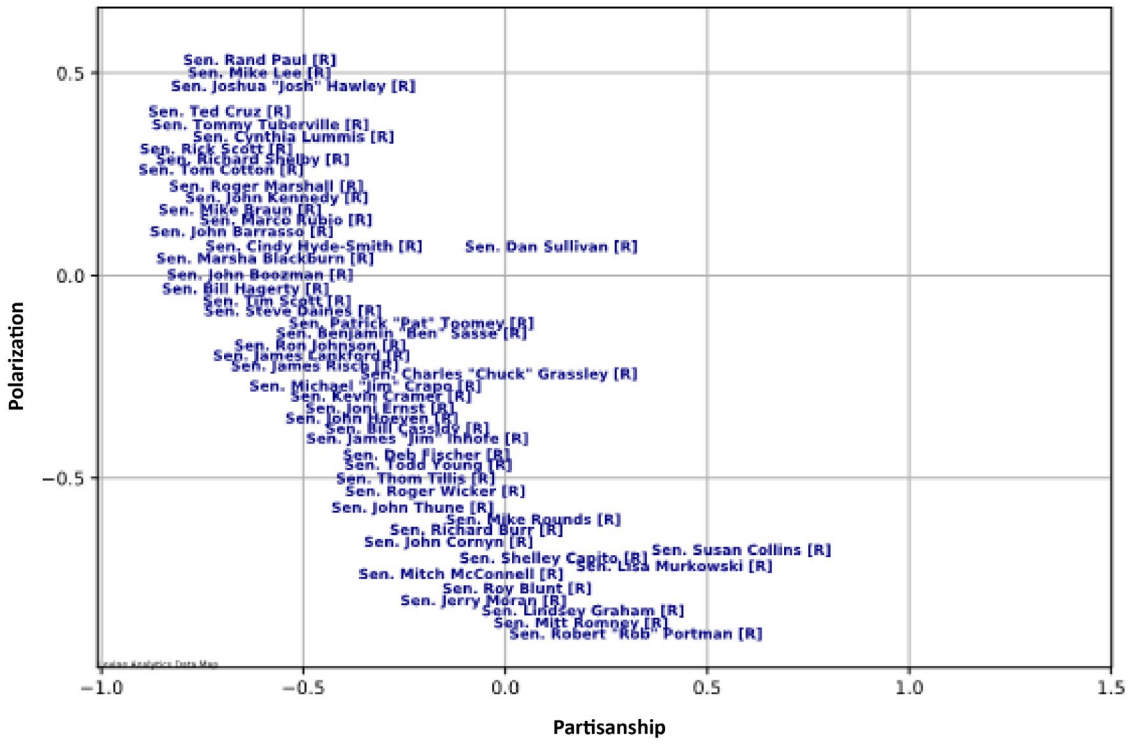


Exhibit 11: Republicans



The model has sorted the entire U.S. Senate along two axes, picking up both the partisan sorting as well as clustering within parties. The within-party differentiation on *two* axes produces two different set of end points. Among Republican senators, Sen. Collins (R-ME) is at one end, closest to the Democrats on the more dominant x-axis. However, along the less dominant but still important y-axis, Sen. Portman (R-Oh) becomes the Republican closest to the Democrats.

A close examination of the y-axis differentiation yields substantial overlap of the Republican leaders who formed the so-called 'Gang of 20' to negotiate the bipartisan infrastructure legislation the Senate recently passed. This group, led by Senator Portman, including Senators Moran (R-Kansas), Blunt (R-MO), and Graham (R-SC), was in essence predicted by the model using a set of roll call data produced before consideration of the infrastructure legislation.

This is quite interesting as traditional classification would not have labeled some of these senators as the most likely to find bi-partisan consensus. The traditional metrics listed Senator Blunt as the 25th most conservative senator and Senator Moran ranks 30th.[15]

Similarly, for the Democrats, while Senator Machin (D-WV) is clearly the most conservative Democrat on the dominant x-axis, Senator Tester (D-MT) becomes the most

[15] Govtrack ideological rankings for 2019 legislative year. https://www.govtrack.us/congress/members/report-cards/2019/senate/ideology

‘Republican’ Democrat on the y-axis. Both were leaders of the Gang of 20 for the Democrats, indicating that the AI is able to determine multiple dimensions of partisanship.

Discussion: New science / old principles

New advances in data analysis enable social scientists to make progress by using data in novel ways. It is Newton and Leibnitz’ calculus that made least squares curve fitting practical. It was tables of probability distributions, hand-computed over decades in the 19th century, that enabled 20th century scholars to test results against null hypotheses.

But “soft science” as it is now practiced with these methods is a struggle against at least two problems. One is a paucity of valid quantitative variables. The other is the basic problem of exacting positive information from the rejection of null hypotheses. As B.F. Skinner might have put it in describing the ineffectiveness of punishment as a device for learning: There are just too many ways to be wrong. Basically it is only binary tests—Yes/No, True/False, Same/Different—for which disproving a null has the effect of “proving” a positive result.

But when a scientist is ‘exploring’ the data, when a scientist has twenty possibly ‘causal’ variables to consider, a clean implementation of a test against a clear null hypothesis becomes difficult. If for each variable the probability of a false positive is .05, then probability of one or more false positives is 0.65, i.e., greater than even. If any one of the twenty variables has a strong result, the scientist should follow up (and publish the preliminary result). But the scientist cannot (yet) claim statistical “proof.” The problem is that while the strategy of inventing and testing against null hypotheses is brilliant in its place, its place is limited.

The solution would appear to be simple, Above we have used simple hypotheses: That numbers exist, that relations exist, that row and column effects exist, and that there is an overall distribution to which a joint distribution of data conforms. The trouble is that the simplicity of an idea can be quite unrelated to the computational difficulty of putting it to work.

Computing power and advanced data analytics together are cracking this barrier at the speed of Moor’s Law. In this work we have addressed the problem of quantifying variables. We have reverse-engineered quantity from the pattern it imposes on data. The mathematics of a linear relation sometimes hold. The mathematics of negative exponentials has been practical for about two centuries. The mathematics of separating effects into row effects, column effects, and constants, dates back at least a century. But the technology for putting these technologies together is new.

Our emphasis has been on validating the results achievable with advanced methods. The first example acts as if categories of income and education had no known scale or order and then reverse-engineers their numbers from the data. The second example reverse-engineers time from time series data of prices. The third example reverse-engineers partisanship and factionalism from unlabeled votes in the U.S. Senate. In each case data

values implied by the constructed quantities are a good fit to the actual data. And both of the latter examples show preliminary evidence that quantification may lead beyond numbers to new understandings of data.

This paper demonstrates the capacity of non-parametric statistical models to go beyond the capacity of linear regression to facilitate understanding of the relations present in well-studied data and to reveal order that, in some data, evades ordinary regression. New statistical tools combined with new computational power allow for new methods of analysis facilitating an understanding that linear regression does not always allow. It has the potential to overcome some commonly used assumptions of linearity and ordinal relationships in data that are not true.

Appendix

Figures A1 and A2 operationalize the model in Excel format

In A1 there are three tables: Table 1 is the data for education and family income, showing labels and frequencies.

In A1 Table 2 is a starting configuration from which to minimize error. Row and column multipliers are initialized with the row and column multipliers of an ordinary "null model" for frequency data. In the null model row multipliers are row sums divided by the square root of the "n" for the full table. Column multipliers use the corresponding computation for columns.

In this starting configuration the coordinates are initialized randomly, using Excel's rand() fuction at for all x's.

In A1 Table 3 shows the components of the chi-square, cell by cell, with total total chi-square of 664558

Figure A2 is identical to Figure A1 with a single exception: One coordinate has been increased resulting in a modest improvement of goodness of it. The search for order has begun.

Figure A1

	8 or less	Less than high school	High school	Junior college	Associate's	Bachelor's	Master's	MBA	Graduate	MD	PhD	Law
$170K +	1	5	32	3	5	54	17	2	34	1	5	5
<$170K	10	22	253	28	29	195	53	3	77	0	8	1
<$80K	20	32	379	25	34	131	30	1	26	1	2	1
<$40K	27	45	360	20	20	62	16	1	17	0	0	0
<$20K	20	34	173	12	11	30	3	1	4	0	0	1
<$10K	22	41	120	6	6	11	3	0	2	0	0	0

Sum: 2593

			8 or less	Less than high school	High school	Junior college	Associate	Bachelor's	Master's	MBA	Graduate	MD	PhD	Law
		Col Mult	1.964	3.515	25.863	1.846	2.062	9.485	2.396	0.157	3.142	0.039	0.295	0.157
		Col x's	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4.25														
	Row Mult	Row x's												
$170K +	3.221	0.000	6.32	11.32	83.30	5.95	6.64	30.55	7.72	0.51	10.12	0.13	0.95	0.51
$170K	13.334	0.000	26.19	46.87	344.87	24.61	27.50	126.48	31.95	2.09	41.90	0.52	3.93	2.09
$80K	13.393	0.000	26.30	47.08	346.39	24.72	27.62	127.04	32.09	2.10	42.08	0.53	3.95	2.10
$40K	11.154	0.000	21.91	39.21	288.49	20.59	23.00	105.80	26.72	1.75	35.05	0.44	3.29	1.75
$20K	5.675	0.000	11.15	19.95	146.78	10.48	11.70	53.83	13.60	0.89	17.83	0.22	1.67	0.89
$10K	4.144	0.000	8.14	14.57	107.17	7.65	8.54	39.30	9.93	0.65	13.02	0.16	1.22	0.65
		Chi-Sq by cell												
			4.483	3.529	31.590	1.459	0.405	18.003	11.170	4.411	56.354	6.032	17.300	39.915
			10.005	13.199	24.472	0.466	0.082	37.123	13.874	0.391	29.410	0.524	4.222	0.572
			1.510	4.830	3.070	0.003	1.475	0.124	0.136	0.579	6.146	0.427	0.959	0.579
			1.185	0.855	17.725	0.017	0.391	18.134	4.304	0.323	9.294	0.438	3.286	1.752

			7.035	9.894	4.682	0.221	0.042	10.551	8.259	0.013	10.730	0.223	1.672	0.013
			23.617	47.973	1.536	0.356	0.758	20.382	4.834	0.651	9.327	0.163	1.221	0.651
		Chi-sq	571.3											

Where the initial (null model) row multiplier in cell B15 is

$$fx = SUM(D2{:}O2/\$B\$8\text{^}0.5)$$

Where the initialized (null model) column multiplier in cell D12 is

$$fx = SUM(D2{:}D7)/\$B\$8\text{^}0.5$$

With fitted values as specified by Equation 2. For cell D16 the Excel expression of Equation 2 is

$$fx = \$B16 * D\$12 * 2\text{^} - (ABS(\$C16 - D\$13)^2)$$

The Excel expression for the chi-square contribution from cell D24 is

$$fx = ((D2 - D16)^2)/D16)$$

And the table chi-square in cell D31 is

$$fx = SUM(D24{:}O29)$$

Figure A2

	8 or less	Less than	High scho	Junior coll	Associ-ate	Bache-lor's	Mas-ter's	MBA	Gradu-ate	MD	PhD	Law
$170K+	1	5	32	3	5	54	17	2	34	1	5	5
<$170K	10	22	253	28	29	195	53	3	77	0	8	1
<$80K	20	32	379	25	34	131	30	1	26	1	2	1
<$40K	27	45	360	20	20	62	16	1	17	0	0	0
<$20K	20	34	173	12	11	30	3	1	4	0	0	1
<$10K	22	41	120	6	6	11	3	0	2	0	0	0

			8 or less	Less than high school	High school	Junior college	Associate	Bachelor's	Master's	MBA	Graduate	MD	PhD	Law
		Col Mults	1.964	3.515	25.863	1.846	2.062	9.485	2.396	0.157	3.142	0.039	0.295	0.157
		Col x's	0.163	0.464	0.226	-0.265	-0.248	0.034	0.179	0.360	0.341	-0.295	0.288	0.292
4.25														
	Row Mults	Row x's												
$170K +	3.221	0.409	6.07	11.30	81.39	4.34	4.92	27.71	7.44	0.51	10.09	0.09	0.94	0.50

<$170K	13.334	-0.069	25.22	38.49	324.68	23.97	26.89	125.56	30.61	1.84	37.30	0.51	3.60	1.91
<$80K	13.393	-0.428	20.65	27.13	257.59	24.27	27.01	109.61	24.86	1.37	27.95	0.52	2.77	1.47
<$40K	11.154	-0.096	20.91	31.54	268.47	20.19	22.63	104.57	25.36	1.52	30.71	0.43	2.97	1.58
<$20K	5.675	-0.146	10.43	15.42	133.39	10.37	11.62	52.65	12.64	0.75	15.14	0.22	1.47	0.78
<$10K	4.144	0.199	8.13	13.87	107.12	6.59	7.44	38.56	9.92	0.64	12.84	0.14	1.21	0.65
		Chi-Sq by cell												
			4.231	3.510	29.974	0.414	0.001	24.937	12.286	4.424	56.689	9.235	17.557	40.379
			9.189	7.064	15.823	0.678	0.166	38.408	16.373	0.724	42.257	0.505	5.396	0.436
			0.020	0.876	57.225	0.022	1.811	4.174	1.062	0.099	0.137	0.444	0.212	0.150
			1.776	5.740	31.204	0.002	0.307	17.333	3.453	0.176	6.121	0.426	2.966	1.579
			8.775	22.394	11.763	0.255	0.033	9.742	7.352	0.086	8.196	0.219	1.467	0.062
			23.663	53.034	1.550	0.052	0.278	19.701	4.832	0.639	9.152	0.137	1.214	0.647
			47.653	92.618	147.539	1.423	2.595	114.294	45.358	6.149	122.552	10.968	28.812	43.252
		Chi-Square	663.2136043											
		Chi-Square	664.5583589											

Bibliography

Duncan, Bruce, Daniel Kim, and Joel Levine. “E.T.A. Hoffmann and Schwester Monika: A Stylometric Analysis.” *Mitteilungen der E.T.A. Hoffmann-Gesellschaft*, 19 (2011): 113-124.

Hanson, Chris. “A statistical analysis votes in the U.S. Senate.” *Dartmouth College Math Social Science Papers*, Dartmouth College, 1998.

Klein, Aaron. “Text Analysis Without Coding: It can be Done.” *Dartmouth College Math Social Science Papers*, Dartmouth College, 1999.

Liddell, Kurschke. “Analyzing ordinal data with metric models: What could possibly go wrong?” *Journal of Experiment Social Psychology* 79 (November 2018).

Levine, Joel H. “New Domains for Structural Analysis: Reformulating standard data analysis as structural analysis.” *Information, Communication & Society* 16, No. 4 (2013): 613-629.

Levine, Joel H. “Extended Correlation: Not Necessarily Quadratic, Not Necessarily Quantitative.” *Sociological Methods and Research* 34, No. 1 (Auguse 2005): 31-75.

Exceptions are the Rule: Inquiries on Method in the Social Sciences: A critique of sociological methodology, with structuralist solutions, Westview Press, (Tilly and McNall, series editors), May, 1993. (http://www.dartmouth.edu/~jlevine/)

Levine, Joel H., Aaron Klein, and James Mathews. “Data Without Variables.” *Journal of Mathematical Sociology* 23, no. 3 (2001): 225-273.

Levine, Joel H. and William S. Roy. "A Study of Interlocking Directorates: Vital Concepts of Organization." *Perspectives on Social Network Research* (1979).

Computer code and instructions are available from the authors.

The Center on Regulation and Markets at Brookings provides independent, non-partisan research on regulatory policy, applied broadly across microeconomic fields. It creates and promotes independent economic scholarship to inform regulatory policymaking, the regulatory process, and the efficient and equitable functioning of economic markets.

Questions about the research? Email communications@brookings.edu.
Be sure to include the title of this paper in your inquiry.