the journal of
# financial data science

# Relevance-Based Prediction:
## A Transparent and Adaptive Alternative to Machine Learning

Megan Czasonis, Mark Kritzman, and David Turkington
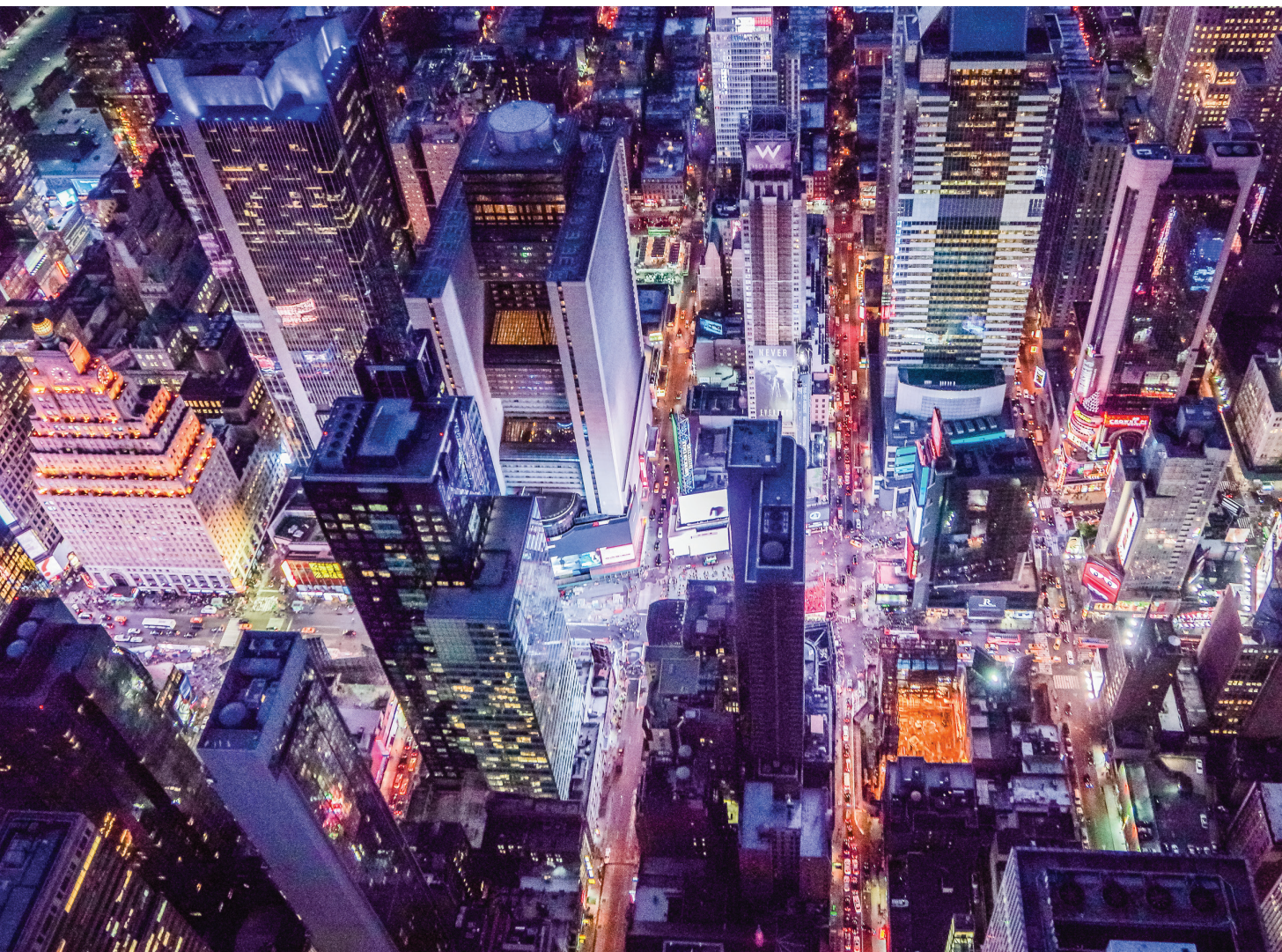
## Megan Czasonis

Megan Czasonis is a Managing Director and Head of Portfolio Management Research at State Street Associates. Her team collaborates with academic partners to develop new research on asset allocation, risk management, and quantitative investment strategy. The Portfolio Management Research team delivers this research to institutional investors through indicators, advisory projects, and thought leadership pieces. Megan has published numerous journal articles and is a co-author of *Prediction Revisited: The Importance of Observation*. Megan graduated Summa Cum Laude from Bentley University with a B.S. in Economics / Finance.

## Mark Kritzman, CFA

Mark Kritzman, CFA is a Founding Partner and CEO of Windham Capital Management, LLC and the Chairman of Windham's investment committee. He is responsible for managing research activities and investment advisory services. He is also a Founding Partner of State Street Associates, and he teaches a graduate finance course at the Massachusetts Institute of Technology. Mark served as a Founding Director of the International Securities Exchange and as a Commissioner on the Group Insurance Commission of the Commonwealth of Massachusetts. He has also served on the Advisory Board of the Government Investment Corporation of Singapore (GIC) and the boards of the Institute for Quantitative Research in Finance, The Investment Fund for Foundations, and State Street Associates. He is currently a member of the Board of Directors of Protego Trust Company, the Advisory Board of the MIT Sloan Finance Group, the Board of Governors of St. John's University, the *Emerging Markets Review*, *The Journal of Alternative Investments*, *The Journal of Derivatives*, *The Journal of Investment Management*, where he is Book Review Editor, and *The Journal of Portfolio Management*. He has written more than 100 articles for peer-reviewed journals and is the author or co-author of eight books including *Prediction Revisited: The Importance of Observation, Asset Allocation: From Theory to Practice and Beyond, Puzzles of Finance*, and *The Portable Financial Analyst*. Mark won Graham and Dodd scrolls in 1993, 2002, and 2010 the Research Prize from the Institute for Quantitative Investment Research in 1997, the Bernstein Fabozzi/Jacobs Levy Award 10 times, the Roger F. Murray Prize from the Q-Group in 2012, and the Peter L. Bernstein Award in 2013 for Best Paper in an Institutional Investor Journal. In 2004, Mark was elected a Batten Fellow at the Darden Graduate School of Business Administration, University of Virginia. Mark has a BS in economics from St. John's University, an MBA with distinction from New York University, and a CFA designation.

## David Turkington

David Turkington is Senior Managing Director and Head of State Street Associates, State Street Global Markets' decades-long partnership with renowned academics that produces innovative research on markets and investment strategy. David is a frequent presenter at industry conferences, has published more than 30 research articles in a range of journals, and currently serves on the editorial board of *The Journal of Alternative Investments*. He is the co-author of three books including *Asset Allocation: From Theory to Practice and Beyond* and *Prediction Revisited: The Importance of Observation*. His published research has received the 2010 Graham and Dodd Scroll Award, five Bernstein-Fabozzi/Jacobs-Levy Outstanding Article Awards, the 2013 Peter L. Bernstein Award for best paper in an Institutional Investor journal, and the 2021 Roger F. Murray First Prize for outstanding research presented at the Institute for Quantitative Research in Finance (Q Group) seminars. David graduated summa cum laude from Tufts University with a BA in mathematics and quantitative economics, and he holds the CFA designation.

# Relevance-Based Prediction:
## A Transparent and Adaptive Alternative to Machine Learning

### Megan Czasonis, Mark Kritzman, and David Turkington

**Megan Czasonis**
is a managing director at State Street Associates in Cambridge, MA.
**mczasonis@statestreet.com**

**Mark Kritzman**
is the chief executive officer at Windham Capital Management in Boston, MA, and a senior lecturer at the MIT Sloan School of Management in Cambridge, MA.
**kritzman@mit.edu**

**David Turkington**
is a senior managing director at State Street Associates in Cambridge, MA.
**dturkington@statestreet.com**

## KEY FINDINGS

- This article proposes a new approach to prediction based on relevance, which gives a measure of the importance of an observation to a prediction, and fit, which measures the reliability of a specific prediction task.

- Relevance-based prediction addresses the codependence of observations and variables by identifying the optimal combination of observations and predictive variables for any given prediction task, thereby presenting a single alternative to both kernel regression and lasso regression.

- Relevance-based prediction has important advantages with respect to machine learning because it is more transparent, more flexible, and less arbitrary than commonly used machine learning algorithms.

## ABSTRACT

The authors describe a new prediction system based on relevance, which gives a mathematically precise measure of the importance of an observation to forming a prediction, as well as fit, which measures a specific prediction's reliability. They show how their relevance-based approach to prediction identifies the optimal combination of observations and predictive variables for any given prediction task, thereby presenting a unified alternative to both kernel regression and lasso regression, which they call CKT regression. They argue that their new prediction system addresses complexities that are beyond the capacity of linear regression analysis but in a way that is more transparent, more flexible, and less arbitrary than widely used machine learning algorithms.

We propose a new prediction system based on the concept of statistical relevance. This new approach to prediction requires one to identify a subsample of relevant observations from which to form predictions, in which relevance has a precise mathematical meaning. Additionally, we introduce the notion of fit, which enables one to assess the unique reliability of each individual prediction task. Fit also enables one to identify the optimal combination of observations and predictive variables for any given prediction task.

This relevance-based approach to prediction addresses complexities that are beyond the capacity of linear regression analysis but in a way that is more transparent, more flexible, and less arbitrary than widely used machine learning algorithms. Moreover, this new prediction system is mathematically cohesive in the following sense. It converges to the same prediction result as produced by a linear regression

model when its key elements are properly aggregated over the full sample of observations, and the aggregation of fit across all prediction tasks converges to the classical *R*-squared statistic.

We motivate our new approach to prediction within the context of common prediction practices from the past and the present. We begin with the past by discussing linear regression analysis and its limitations when applied to data in which conditionality leads to asymmetry[1] between the predictive variables and the outcomes. Then we proceed to the present. We first discuss several model-based machine learning algorithms that are best characterized as enhancements to linear regression analysis. Next, we describe two model-free machine learning algorithms that serve as a bridge to our relevance-based prediction system. We proceed to describe our relevance-based approach to prediction, and we show how it identifies the optimal combination of observations and predictive variables for any given prediction task. We refer to this procedure as CKT regression. Finally, we conclude with a summary of how relevance-based prediction compares to linear regression analysis and machine learning.

## THE PAST: LINEAR REGRESSION ANALYSIS

The main approach to statistical prediction in the past was linear regression analysis, and it is still the dominant approach today for most simple prediction tasks. In a very general sense, linear regression analysis focuses on the selection of predictive variables that are weighted based on an assumed linear relationship between the values for the predictive variables and the outcomes to give a prediction of a new outcome given a new set of values for the predictive variables. The weights that are applied to the predictive variables are derived by fitting a line through a scatter plot of values for the predictive variables and outcomes such that the sum of the squared distances of the observations from the line is minimized. Carl Friedrich Gauss, who originated this method of least squares circa 1795, proved that it gives a prediction whose expected variance from the truth is lower than any other linear and unbiased prediction.

The prediction of a linear regression model with a single predictive variable is given by Equation 1.

$$y_t = \alpha + x_t\beta + \epsilon_t \tag{1}$$

In Equation 1, $\epsilon_t$ is an error term. The errors are assumed to be centered on zero, to have constant variance, and to be uncorrelated, which means that they are spherically distributed. Although it is commonly assumed that the errors must be normally distributed, this assumption is unnecessarily strict.

The method of least squares solves for the parameters $\alpha$ and $\beta$ that minimize the sum of squared prediction errors for the available observations. By considering $x_t$ to be a row vector containing many variables and $\beta$ to be a column vector containing an equal number of coefficients, this linear regression model can be extended to include multiple predictive variables. And by including transformations of the original variables as new variables, it can be used to linearize nonlinear relationships.

Linear regression analysis is a remarkably elegant approach to prediction, but it is limited in a significant way. It assumes that the relationship between the predictive

---

[1]Asymmetry is present when the outcomes associated with a set of circumstances do not tend to equal the opposite of the outcomes associated with the opposite circumstances. See Czasonis, Kritzman, and Turkington (2022) for further details.

variables and the outcomes is static across all observations. We now turn to machine learning, which has emerged as the standard approach to prediction in the face of more complex relationships.

## THE PRESENT: MACHINE LEARNING

To set the stage for our discussion of relevance-based prediction, it is convenient to stratify machine learning algorithms into two types: model-based algorithms that, for all intents and purposes, are enhancements to linear regression analysis, and model-free algorithms, which serve as a bridge to relevance-based prediction. We consider three model-based algorithms: lasso regression, tree-based algorithms, and neural networks. We also consider two model-free algorithms: near neighbors and Gaussian kernels.

### Model-Based Machine Learning Algorithms

Lasso regression focuses on the selection of predictive variables instead of observations. Introduced by Tibshirani (1996), its name stands for *least absolute shrinkage and selection operator.* Its goal is to remove or minimize the influence of variables in multivariate linear regression that are the least effective in forming the prediction. Whereas traditional linear regression analysis seeks coefficients that minimize the sum of squared prediction errors, lasso regression also seeks to minimize the sum of squared prediction errors but is subject to an additional penalty term that is proportional to the sum of the absolute values of the coefficients. Typically, the input variables are standardized so that the coefficients represent equivalent units. For a variable with weak predictive capacity, the penalty associated with a nonzero coefficient may outweigh its ability to reduce prediction errors, in which case the variable is deselected. Or the penalty may justify a smaller coefficient than unconditional linear regression, in which case the value of the coefficient is reduced. The extent of the penalty is determined using an additional analysis. A closely related approach called ridge regression penalizes the sum of squared coefficient values instead of their absolute values.

Tree-based algorithms categorize the multivariate input values to a prediction using a series of yes/no questions that branch into a structure that looks like a tree. Each question asks whether a given input variable is above or below a particular value. A prediction is then formed from the simple average of the outcomes of prior observations in the same category. The complexity of the tree structure must be determined using an additional analysis. The localized nature of tree-based predictions gives them the ability to reflect nonlinear relationships. Some popular algorithms such as random forest (Ho 1995; Breiman 2001) and gradient boosted trees (Friedman 2001) implement rules that calibrate multiple trees using subsets of data and aggregate their results to form predictions.

Neural networks are models that process predictive input values through multiple layers of aggregation and transformation to render a prediction. Inspired by the structure of the human brain, the nodes in each layer of a neural network apply their own weights to combine inputs from the prior layer and further transform the result using a nonlinear activation function, which defines the output of a node given a set of inputs. Each node's output then feeds the next layer of nodes, which eventually ends in a final layer that aggregates all the prior information into a prediction. These models are capable of processing large amounts of complexity if their structure is suitably wide (many nodes inside layers) or deep (many layers). The size and structure of the network is determined using an additional analysis, and there are many

calibration parameters to consider. Neural networks have achieved impressive performance across a range of practical applications. Unfortunately, their complexity inhibits transparency; it is therefore extremely difficult to understand the underlying logic of a neural network's prediction.

The common feature of these model-based algorithms is that they all rely on a general iterative process, which is first to specify a decision rule, then to calibrate the decision rule, and finally to test the decision rule based on the quality of the resultant predictions. These steps are repeated multiple times until one is satisfied with the results.

One can construct these model-based algorithms to be extraordinarily flexible in how they approach a wide range of complexities, but they are very rigid in one critical respect. After their final calibration, they are incapable of adapting to changing circumstances. The only way to update these model-based algorithms to respond to new circumstances is to repeat the iterative process with the updated sample.[2] Model-free algorithms, which we next discuss, automatically adapt to changing circumstances.

### Model-Free Algorithms

A distinguishing feature of model-free algorithms is that they form their predictions as weighted averages of prior outcomes. Moreover, the weights that are used to form the predictions are revised with each new prediction task. This approach to prediction is referred to as kernel regression. Whereas model-based algorithms focus on the selection of predictive variables, model-free algorithms focus on the selection of observations.

The near neighbors approach, for example, forms its predictions from localized sets of observations with respect to the current values of the predictive variables. These algorithms simply exclude those observations that fall outside a chosen range around the current values of the predictive variables and weight the surviving observations of the outcomes equally.

A slightly more advanced approach to kernel regression is a procedure known as a Gaussian kernel. This approach first requires one to calculate the Euclidean distance of each observation of the predictive variables from their current observation. Then, assuming a normal distribution for these Euclidean distances, it assigns weights in proportion to where the distances fall within the distribution. Smaller distances near the center of the distribution receive larger weights than distances that lie in the tails. The final step is to scale these weights to sum to one and use them to compute a weighted average of the associated outcomes to form the prediction.

These model-free machine learning algorithms share a common feature with relevance-based prediction; both approaches form predictions as weighted averages of prior outcomes. A key distinction of relevance-based prediction, however, is that the weights are theoretically justified.

## THE FUTURE: RELEVANCE-BASED PREDICTION

Our relevance-based approach to prediction is guided by three principles.

1. A prediction system should be transparent. Transparency promotes intuition and facilitates interpretation.

---

[2] Some machine learning algorithms allow for online learning, which continually refines the calibrated parameters based on new training observations. This process does not require repeating the entire training routine. Still, online learning is concerned with incremental updates to the values of model parameters, and those parameters are considered to be fixed for the purposes of a given prediction task.

2. It should be flexible, by which we mean it should be responsive to changing circumstances.
3. And it should be nonarbitrary, by which we mean it should be theoretically justified and mathematically unified. Nonarbitrariness confers legitimacy to the choices made by a prediction system, so the results are more likely to reflect underlying relationships as opposed to happenstance.

Linear regression analysis satisfies our third principle, nonarbitrariness, but it falls short with respect to our second principle, flexibility, and, to some extent, our first principle, transparency. Model-based machine learning algorithms, despite their obvious sophistication and ample evidence of empirical success, fail to satisfy any of our proposed principles. We do acknowledge, though, that these principles are the ones we value. Others may rely on different criteria to judge the merits of alternative prediction systems. Or one's focus may be descriptive rather than predictive, in which case model-based machine learning would likely fare better.

There are three key elements to our relevance-based prediction system: relevance, fit, and situational learning.

### Relevance

Like model-free algorithms, we proceed from the general perspective that a prediction is a weighted average of prior outcomes, as shown in Equation 2.

$$\hat{y}_t = \sum_i w_{it} y_i \tag{2}$$

Our challenge is to produce weights that render the prediction effective. As we just discussed regarding kernel regression, we could use proximity to weight the outcomes, which is a step in the right direction. Nonetheless, these localization algorithms are arbitrary. We propose, instead, a nonarbitrary solution for computing the weights to be applied to prior outcomes, which is relevance. Like kernel regression, relevance also considers the similarity of prior observations to current circumstances, but it defines similarity more broadly than simple proximity. Moreover, relevance measures similarity in a way that is theoretically justified. And of critical importance, relevance also considers the informativeness of prior observations as well as current circumstances, again in a way that is theoretically justified. Equation 3 shows how similarity and informativeness are combined to determine relevance. We denote current circumstances as a row vector, $x_t$, and the circumstances of a given prior observation as a row vector, $x_i$.

$$r_{it} = sim(x_i, x_t) + \frac{1}{2}(info(x_i, \overline{x}) + info(x_t, \overline{x})) \tag{3}$$

In Equation 3, similarity and informativeness are computed as Mahalanobis distances rather than absolute distances or Euclidean distances:[3]

$$sim(x_i, x_t) = -\frac{1}{2}(x_i - x_t)\Omega^{-1}(x_i - x_t)' \tag{4}$$

$$info(x_i, \overline{x}) = (x_i - \overline{x})\Omega^{-1}(x_i - \overline{x})' \tag{5}$$

---

[3] For information about the derivation of the Mahalanobis distance, see Mahalanobis (1927) and Mahalanobis (1936).

Here, $\overline{x}$ is the average of $X$, and $\Omega^{-1}$ is the inverse covariance matrix of $X$. All else being equal, observations that are like current circumstances and different from average are more relevant than those that are not.

It is very important to note that this measure of relevance, and thus the weights of prior observed outcomes for $Y$, does not yet incorporate any information from $Y$. Relevance is determined using only the set of predictive variables contained in $X$.

As we emphasized earlier, this definition of relevance is not arbitrary. The use of the Mahalanobis distance follows from two keystones of statistical reasoning: the central limit theorem and information theory. The central limit theorem holds that the sum or average of many independent random events is approximately normally distributed, so long as each underlying event comes from a distribution with finite variance. Therefore, even nonnormal processes give rise to aggregate phenomena that are normally distributed. The normal distribution is also the best-motivated starting point from the perspective of information theory because it has the maximum entropy (the least amount of assumed prior knowledge) of any distribution given a specified variance. The relative likelihood of an observation $x_i$ from a multivariate normal distribution is proportional to the exponential of a negative Mahalanobis distance:

$$likelihood(x_i) \propto e^{-\frac{1}{2}(x_i-\overline{x})\Omega^{-1}(x_i-\overline{x})'} \tag{6}$$

Information theory, created by Claude Shannon (1948), holds that the information contained in an observation is the negative logarithm of its likelihood. It therefore follows that the information contained in a point from a normal distribution is proportional to a Mahalanobis distance:

$$information(x_i) \propto (x_i-\overline{x})\Omega^{-1}(x_i-\overline{x})' \tag{7}$$

We can also justify the nonarbitrariness of relevance in the following sense. As we show in the appendix, a relevance-weighted average of prior outcomes for the full sample yields a prediction that is precisely equivalent to the prediction that results from a fitted ordinary least squares linear regression model applied to the circumstances, $x_t$.

Just as kernel regression refines a prediction by focusing on local observations, relevance-based prediction refines a prediction by focusing on the most relevant observations, as determined by $X$. Forming a prediction from a subsample of the most relevant observations is called partial sample regression. As we show in the appendix, these weights always sum to one.

$$w_{it,psr} = \frac{1}{N} + \frac{\lambda^2}{n-1}(\delta(r_{it})r_{it} - \varphi\overline{r}_{sub}) \tag{8}$$

In Equation 8, $\delta(r_{it})$ is a censoring function that equals 1 if $r_{it} \geq r^*$ and 0 otherwise. For notational concision, we write the number of observations for which $\delta(r_{it}) = 1$ as $n = \Sigma_i\delta(r_{it})$ and the proportion of all observations for which $\delta(r_{it}) = 1$ as $\varphi = \frac{n}{N}$. In addition, we write the subsample average of relevance over the retained observations as $\overline{r}_{sub} = \frac{1}{n}\Sigma_i\delta(r_{it})r_{it}$. Finally, we include a term $\lambda^2 = \sigma^2_{r,full}/\sigma^2_{r,partial} = \frac{1}{N-1}\Sigma_i r_{it}^2 / \frac{1}{n-1}\Sigma_i\delta(r_{it})r_{it}^2$, which, to the extent it differs from 1, compensates for a bias that would otherwise arise from focusing on a small subsample of highly relevant observations. The predictions that result from these partial sample regression weights share the theoretical justification of linear regression analysis, but unlike linear regression analysis, they allow for multivariate asymmetry.

To summarize:

- In a very general sense, we should think of a prediction as a weighted average of prior outcomes.
- Rather than weight the prior outcomes based on proximity to current circumstances, as in a kernel regression, relevance-based prediction weights them based on a theoretically justified quantity called relevance.
- An observation's relevance equals its similarity to current circumstances measured as the negative of half the Mahalanobis distance from current circumstances plus the average of its informativeness and the informativeness of current circumstances, both measured as Mahalanobis distances from the average of the observations.
- Relevance is not arbitrary. Multivariate normal distributions are well motivated by the central limit theorem, and we know that the relative likelihood of an observation from a multivariate normal distribution is proportional to the exponential of a negative Mahalanobis distance. And from information theory, we know that the information contained in an observation is the negative logarithm of its likelihood. Therefore, the information contained in a point on a normal distribution is proportional to a Mahalanobis distance. Moreover, a weighted average of prior outcomes across a full sample, in which the weights are relevance, gives the same prediction as a linear regression equation. Relevance-based prediction and linear regression analysis, therefore, are mathematically unified.
- When faced with asymmetry, we may be able to improve a prediction's reliability by censoring observations that are less relevant. Forming a prediction from a subsample of the most statistically relevant observations is called partial sample regression.

We now turn to the second key element of relevance-based prediction, which is fit.

### Fit

Fit is a critical component of relevance-based prediction. It reveals how much confidence we should assign to a specific prediction task, separately from the overall confidence we have in the associated prediction model, and it enables us to determine the optimal threshold for the subsample of relevant observations for each prediction task.

Here is how to think about fit. Consider a pair of observations that go into a prediction task. Each observation comprises a relevance weight and an outcome. We are interested in the alignment of the weights of the two observations with their outcomes. But we must standardize them by subtracting from them the average value and dividing by variance—in essence, converting them to $z$-scores. We then measure their alignment by taking the product of the standardized values. If this product is positive, the weights are aligned with the outcomes, and the larger the product, the stronger the alignment. We perform this calculation for every pair of observations in our sample of size $N$. Also, it is important to note that all the formulas we have thus far considered for the weights rely only on $X$; they do not make use of any of the information in previously observed outcomes, $Y$. To determine fit, however, we must also consider $Y$. We express fit as a pairwise sum that involves the relevance of weights and outcomes for both observations in a pair.

$$fit_t = \frac{1}{(N-1)^2} \sum_i \sum_j r(w_{it}, w_{jt}) r(y_i, y_j) \qquad (9)$$

In Equation 9:

$$r(w_{it}, w_{jt}) = \frac{(w_{it} - \bar{w})(w_{jt} - \bar{w})}{\sigma_w^2} \qquad (10)$$

$$r(y_i, y_j) = \frac{(y_i - \bar{y})(y_j - \bar{y})}{\sigma_y^2} \qquad (11)$$

From Equations, 9, 10, and 11, we can restate fit in terms of normalized $z$-scores:

$$fit_t = \frac{1}{(N-1)^2} \sum_i \sum_j z_{w_{it}} z_{w_{jt}} z_{y_i} z_{y_j} \qquad (12)$$

Or as follows in Equation 13:

$$fit_t = \frac{1}{(N-1)^2 \sigma_{w_t}^2 \sigma_y^2} \sum_i \sum_j (w_{it} - \bar{w})(w_{jt} - \bar{w})(y_i - \bar{y})(y_j - \bar{y}) \qquad (13)$$

If we reference each observation's predictive contribution as $\psi_i = (w_{it} - \bar{w})(y_i - \bar{y})$, we can also express fit as

$$fit_t = \frac{\psi' 1_N 1_N' \psi}{(N-1)^2 \sigma_{w_t}^2 \sigma_y^2} \qquad (14)$$

or as

$$fit_t = \rho(w_t, y)^2 \qquad (15)$$

or as

$$fit_t = \frac{info(\hat{y}_t)}{\sigma_{w_t}^2} \qquad (16)$$

Although we compute fit from the full sample of observations, the weights that determine fit vary with the threshold we choose to define the relevant subsample. Because we focus the subsample on more relevant observations, we should expect the fit of the subsample to increase, but we should also expect more noise as we shrink our number of observations. The fit across pairs of all observations in the full sample $N$ implicitly captures this trade-off by overweighting more relevant observations and underweighting less relevant observations accordingly.

Like relevance, fit is not at all arbitrary. In the appendix, we show from first principles why the sum must be normalized by a divisor of $(N-1)^2$. Moreover, Czasonis, Kritzman, and Turkington (2022) showed that the informativeness-weighted average fit across all prediction tasks in a sample equals precisely the classical $R$-squared statistic in the case of full-sample linear regression.

$$R^2 = \frac{1}{T-1} \sum_t info(x_t) fit_t \qquad (17)$$

This convergence of fit to *R*-squared reveals an intriguing insight. Without any knowledge of the success of a model's predictions, fit reveals beforehand the degree of confidence we should attach to a specific prediction task. Without fit, our only indication of confidence is the average quality of all predictions, which is determined after the predictions are made and which across the full sample aggregates to *R*-squared. Therefore, not only can we use fit to assess how much confidence we should attach to a prediction task individually; we can use fit to compute *R*-squared.

To summarize:

- The fit for a given prediction task equals the average of the alignment of standardized weights and standardized outcomes expressed as products, across every pair of observations in a sample *N*. Equivalently, it equals the squared correlation between weights and outcomes.
- Fit reveals, in advance of forming a prediction, how much confidence we should attach to a given prediction task.
- Fit is not arbitrary. The information-weighted average fit across all prediction tasks in a sample *N* converges to the classical *R*-squared statistic.

### Situational Learning

We have thus far shown how to form a prediction, given asymmetric data, as a relevance-weighted average of prior outcomes. And we have shown how we can use fit to guide our confidence in a specific prediction task. But we have left unanswered how to determine the threshold for the subsample of relevant observations. We have only noted that a partial sample regression prediction depends on the choice of a parameter, *r\**, which is the censoring threshold for relevance.

$$w_{it,psr} = g_{psr}(x_t, X, r^*) \tag{18}$$

We, therefore, turn to the third key feature of relevance-based prediction, which is situational learning, to show how to select the censoring threshold for relevance.

Rather than choose *r\** arbitrarily, we use fit to learn the value of *r\** that optimizes a trade-off between decreasing the fraction of observations included in the prediction, $\varphi$, and increasing subsample fit among the *n* noncensored observations, $fit_{t,sub}$. In other words, we iteratively raise the parameter *r\**, which is equivalent to shrinking the relevant subsample to size *n*, until we have maximized fit. From our earlier discussion, we know that fit measured from a sample *N* automatically captures the trade-off between subsample fit $fit_{t,sub}$ and noise, but it may be useful to consider this trade-off in more detail.

Recall from Equation 8 that we write the weights for partial sample regression as $w_{it,psr} = \dfrac{1}{N} + \dfrac{\lambda^2}{n-1}(\delta(r_{it})r_{it} - \varphi\bar{r}_{sub})$. Fit, as defined earlier, is the fit between weights and outcomes for a given sample of size *N*. However, expressing fit in terms of key subcomponents lends further intuition and allows us to make more general use of the concept. Note that $w_{psr}$, *y*, $r_t$, and $\delta(r_t)$ are all series that are defined for all *N* observations. As the threshold *r\** rises to focus on a highly concentrated subsample, *n*, $\varphi$, $\lambda^2$, and $c^2$ all decrease, which unambiguously penalizes fit. The fit measured over a sample *N* will rise if the subsample fit increases by more than this sample size reduction penalty.

$$fit_t(w_{psr}, y) = \left(\frac{\varphi\lambda^2}{1-\varphi c^2}\right)\frac{N}{N-1}\frac{n-1}{n}fit_t(\delta(r_t)r_t, y) \tag{19}$$

Note that we assume the subsample fit in our notation in Equation 19 is normalized according to the subsample size, $n$.

$$fit_t(\delta(r_t)r_t, y) = \frac{1}{(n-1)^2} \sum_i \sum_j r(\delta(r_{it})r_{it}, \delta(r_{jt})r_{jt})r(y_i, y_j) \tag{20}$$

The term $c^2$ is defined as follows:

$$c^2 = \frac{\overline{r}_{sub}^2}{\sigma_{r,partial}^2} \frac{n}{n-1} \tag{21}$$

The term $c^2$ is related to the variance of $r_t$ in the subsample around the subsample average, $\overline{r}_{sub}$. In fact, that variance equals precisely $\sigma_{r,partial}^2 - \frac{n}{n-1}\overline{r}_{sub}^2$, and because this quantity must always be greater than zero, it follows that $\sigma_{r,partial}^2 > \frac{n}{n-1}\overline{r}_{sub}^2$ and, therefore, $c^2 < 1$. We ignore the pathological case where $r_t$ has zero variance.

### CKT Regression and the Codependence of Observations and Variables

Kernel regression and partial sample regression both focus on the selection of observations and form predictions as a weighted average of prior observations. They differ only in how the weights are selected. It is, therefore, straightforward to show how we could modify kernel regression to converge to relevance-based prediction.

Let us consider a Gaussian kernel. With this approach, we form the weights as a function of proximity to current circumstances. Specifically, we measure each observation's Euclidean distance from current circumstances. Then, assuming these distances are normally distributed, we assign weights in proportion to where the distances fall within the distribution.

With the following adjustments, we convert a kernel regression to a partial sample regression.

1. First, we substitute similarity for simple proximity by using the Mahalanobis distance instead of the Euclidean distance to measure differences from current circumstances.
2. Then we add the informativeness of the observations and the informativeness of the current circumstances, both measured as Mahalanobis distances from the average, and we properly weight these three components.
3. Finally, we identify the optimal subsample of relevant observations by iteratively raising the relevance threshold $r*$ until we have maximized fit.

Lasso regression is used to select predictive variables. Just as fit allows us to identify the optimal subsample of relevant observations for a given set of predictive variables, we can also use fit to select variables. Rather than fixing the predictive variables and using fit to select the best observations for a given prediction task, we can fix our sample of observations and iteratively substitute different combinations of predictive variables to maximize fit. By doing so, we present an alternative to lasso regression. Like lasso regression, we can consider every combination of predictive variables and choose the combination that maximizes fit. Or we can construct predetermined sets of predictive variables based on experience and intuition.

It is important to note that fit is specific to each prediction task, so it may recommend some collections of variables only in rare circumstances. Lasso, by contrast, must decide whether to always or never use a given variable. This fit-based alternative to lasso regression, therefore, has the flexibility to adapt to changing circumstances.

Thus far, we have described these methods as though they would be applied independently. But why should we expect a given sample of observations to be most suitable for all combinations of predictive variables, and why should we expect a fixed set of predictive variables to be equally effective across all observations? Each specific prediction task has an optimal combination of observations and predictive variables. Given this codependence, the selection of observations and variables should be carried out jointly by maximizing fit as a simultaneous function of observations and variables. By doing so, we replace both kernel regression and lasso regression with a single process, which we refer to as CKT regression.

To summarize:

- Partial sample regression requires us to determine a threshold value for $r*$ to determine the size of the relevant subsample from which to form our prediction.
- As we raise the threshold value for $r*$, we increase the fit of the subsample, which all else being equal, should increase our confidence in our prediction task. However, by reducing the subsample size, we introduce noise which counters the benefit of better fit.
- We iteratively raise the relevance threshold $r*$ to maximize fit, which captures the penalty associated with the decrease in sample size from $N$ to $n$.
- We also use fit in the same way to identify the optimal collection of variables. Because the choice of observations and choice of variables are codependent, we should vary them simultaneously to maximize fit. This procedure is called CKT regression and presents a unified alternative to both kernel regression and lasso regression.

## EMPIRICAL ILLUSTRATION

We conduct a simulation to demonstrate the intuition and efficacy of CKT regression in a case in which data emanate from multiple regimes. In general, we assume that one does not know the form of the true data-generating process for the prediction problem at hand, but for the purposes of this simulation we specify a known data-generating process. We specify a two-regime Markov model in which the multivariate distribution for $X$ and the function $Y = f(X)$ are regime dependent. We assume two regimes, with persistence probabilities equal to 80% and 60%, respectively, meaning that the probability of regime 1 in the next period is 80% when currently in regime 1, and the probability of regime 2 is 60% when currently in regime 2. We express the transition matrix of the state variable as:

$$Transition = \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{pmatrix} \tag{22}$$

We define four individual variables that comprise our multivariate predictive observations $X$, and we refer to them as $A,B,C,D$. Variables $\{A,B\}$ comprise a group, and variables $\{C,D\}$ comprise a group. The variables are uncorrelated across groups and 50% correlated within groups. The correlation matrix does not depend on regimes.

$$\rho = \begin{pmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{pmatrix} \tag{23}$$

In regime 1, variables {$A,B$} are tightly distributed around a positive average, and they determine the outcome for $Y$ (they are causally important), whereas variables {$C,D$} are broadly distributed around an average of zero and do not determine the outcome for $Y$ (they contribute noise). In regime 2, the variables take on the opposite roles: {$C,D$} are causally important, and {$A,B$} contribute noise. We denote these assumptions as follows, where $\Omega$ is the covariance matrix formed from the appropriate $\sigma$ and $\rho$:

$$\mu_1 = (3 \quad 3 \quad 0 \quad 0); \quad \sigma_1 = (1 \quad 1 \quad 3 \quad 3); \quad \beta_1 = (1 \quad 2 \quad 0 \quad 0) \qquad (24)$$

$$\mu_2 = (0 \quad 0 \quad 3 \quad 3); \quad \sigma_2 = (3 \quad 3 \quad 1 \quad 1); \quad \beta_2 = (0 \quad 0 \quad 1 \quad 2) \qquad (25)$$

$$\begin{cases} X \sim N(\mu_1, \Omega_1) & \textit{if } regime = 1 \\ X \sim N(\mu_2, \Omega_2) & \textit{if } regime = 2 \end{cases} \qquad (26)$$

$$\begin{cases} y_t = \beta_1 x_t' & \textit{if } regime = 1 \\ y_t = \beta_2 x_t' & \textit{if } regime = 2 \end{cases} \qquad (27)$$

To summarize, at any given point in time, two of the variables are causally important in determining $Y$, but it is not always the same two. The formula for $Y$ is deterministic if we know the current regime, but the current regime is not known to the observer; it must be inferred from the observations of $X$, which are random. The prediction task is to form predictions based on a training sample of realizations for $X$ and $Y$, together with a new set of circumstances $x_t$ drawn from a testing sample. We simulate 500 observations for training. We also simulate 500 separate observations for testing, but for ease of illustration we extract every 10th observation to obtain a more compact set of 50 test samples that span both regimes.

For each prediction task, we compute the fit for CKT regression predictions based on different collections of variables (all four, {$A,B$} only, and {$C,D$} only) and different thresholds for filtering observations according to relevance (setting $r*$ equal to a percentile value of relevance in increments of 0.1 from 0 to 0.9). We select the specification with the greatest fit and evaluate this prediction against the actual testing outcome.

In our out-of-sample test, the correlation between CKT regression predictions and actual outcomes was 89%. This result was notably superior to traditional linear regression, which had a correlation of 54%.

Exhibit 1 shows which variables were included and what fraction of observations were included in each of the 50 test predictions. It also reports the fit for each prediction.

The behavior of the prediction routine is intuitive, and it highlights both the transparency and the flexibility of our approach. Regime 1 is more persistent and therefore occurs more frequently (76% compared to 24% for regime 2). The predictions more often rely on variables {$A,B$}, which are important in regime 1. Moreover, these predictions typically include 50% to 100% of the observations, which is in line with the proportion that emanate from regime 1 in the training sample. By contrast, the predictions rely on variables {$C,D$} less often, and when they do, they usually focus their attention on 20% of the observations, in line with the actual occurrence of regime 2 in the training sample. The predictions occasionally resort to a full-sample regression prediction, including every variable and every observation. It makes sense to use this approach when the randomly observed values of $x_t$ are ambiguous.

Exhibit 1 shows that fit varies substantially across predictions. Its largest values coincide with predictions from regime 1, which is intuitive because this regime has the benefit of more evidence in the historical sample. It is important to remember that we compute fit in advance of observing an actual outcome for $Y$ in the testing sample.

**EXHIBIT 1**
Fraction of Observations Included, Variables Included, and Fit
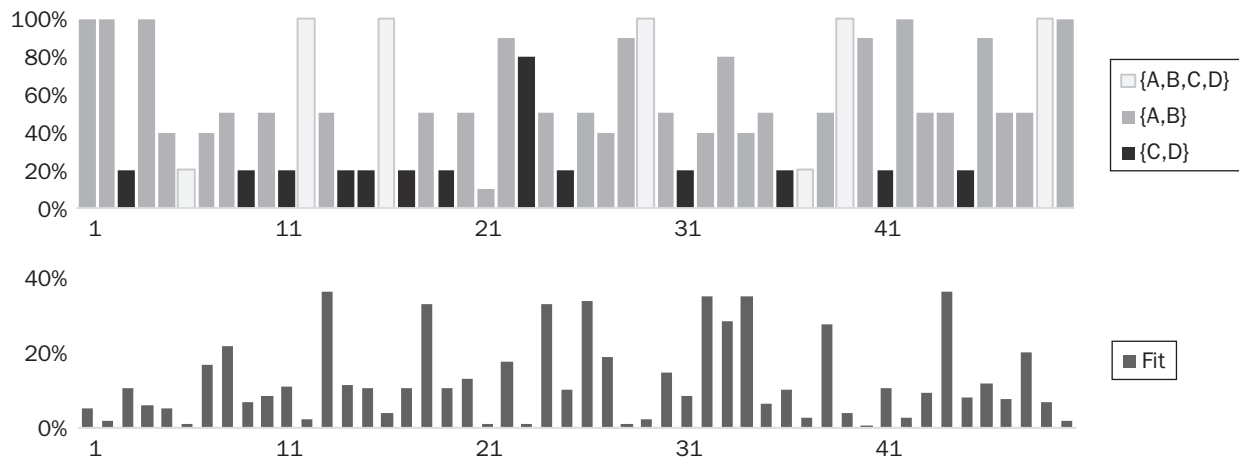


Nonetheless, we verify that fit provides useful information about the confidence of each prediction. Across the 50 prediction tasks, the absolute value prediction errors are −43% correlated with fit. In other words, predictions with higher fits resulted in lower realized prediction errors out-of-sample.

## COMPARATIVE SUMMARY

We now compare linear regression analysis, machine learning, and relevance-based prediction based on the three principles we proposed earlier: transparency, flexibility, and nonarbitrariness.

### Linear Regression Analysis

**Transparency.** Although linear regression analysis shows the influence of the predictive variables on the prediction, it is silent about how each observation informs the prediction. Moreover, the *R*-squared statistic only reveals the efficacy of predictions on average and does not provide information about the fit that underlies each individual prediction.

**Flexibility.** Linear regression analysis is inflexible. It assumes a fixed set of observations and variables for all prediction tasks. It is, therefore, incapable of addressing conditional relationships, and it ignores the codependence of observations and variables.

**Nonarbitrariness.** Linear regression analysis is theoretically justified. Carl Friedrich Gauss showed that the prediction from ordinary least squares is closer to the truth than any other unbiased prediction from a linear model.

### Machine Learning

**Transparency.** Although model-free machine learning algorithms are transparent in that they disclose the effect of each observation on the prediction, the most powerful model-based algorithms are opaque and hard to interpret. They may implement conditional reasoning, but they do not explain or justify it transparently. The logic is concealed in the black box that results from previous training and testing. It is, therefore, impractical to gauge the effect of the observations on the predictions.

**Flexibility.** Because model-based machine learning algorithms are prespecified, the only quantity that varies along with the prediction task is $x_t$. The model parameters are constant, and no new outcomes are observed prior to the prediction for $y_t$. Therefore, the flexibility in machine learning algorithms must be governed by their internal logic and prespecified by a fixed set of calibrated parameters. They do not adapt their approach to prediction circumstances, nor do they recognize the codependence of observations and variables.

**Nonarbitrariness.** Machine learning contains many disparate approaches, and their use is guided by empirical efficacy rather than by a core set of theoretical principles. It is therefore difficult to know whether the solutions result from historical happenstance or underlying relationships.

### Relevance-Based Prediction

**Transparency.** Relevance-based prediction is highly transparent. It reveals the comparative importance of each observation and shows precisely how it informs the prediction. Relevance-based prediction also quantifies the confidence one should attach to each unique prediction task beforehand, in contrast to linear regression analysis, which assigns the same confidence to all prediction tasks ex post based on the average quality of the regression model. Finally, relevance-based prediction shows explicitly how subsample fit and noise separately determine the relevance threshold that determines the optimal subsample for each prediction task.

**Flexibility.** Relevance-based prediction is specifically designed to adapt to asymmetry by imposing conditionality on the prediction process. Moreover, it automatically adapts to new prediction circumstances and explicitly considers the codependence of observations and variables.

**Nonarbitrariness.** Relevance-based prediction is theoretically justified by both the central limit theorem and information theory. Moreover, it is mathematically unified in the following sense. When applied to the full sample of observations, relevance-based prediction yields precisely the same prediction as linear regression analysis. And an informativeness-weighted average of the fit of all a model's prediction tasks equals the classical $R$-squared statistic. Finally, fit may be used to rigorously determine the optimal combination of observations and predictive variables.

## CONCLUSION

We propose a new approach to prediction based on relevance, which gives a measure of the importance of an observation to a prediction, and fit, which measures the reliability of a specific prediction task.

We show how relevance-based prediction addresses the codependence of observations and variables by using an algorithm called CKT regression. This algorithm identifies the optimal combination of observations and predictive variables for any given prediction task, thereby presenting a single alternative to both kernel regression and lasso regression.

We argue that relevance-based prediction compares favorably to linear regression analysis because it is more transparent and because, unlike linear regression analysis, it efficiently adapts to asymmetry between predictive variables and outcomes.

We also claim that relevance-based prediction has important advantages with respect to machine learning. It is more transparent, more flexible, and less arbitrary than commonly used machine learning algorithms.

# APPENDIX

## RESULT 1: PARTIAL SAMPLE REGRESSION WEIGHTS SUM TO ONE

Partial sample regression adds the average outcome to scaled deviations of outcomes from average:

$$\hat{y}_t = \overline{y} + \frac{1}{N}\sum_i \frac{N}{n-1}\delta(r_{it})\lambda^2 r_{it}(y_i - \overline{y}) \tag{A1}$$

Delta is a censorship function wherein $\delta(r_{it}) = 1$ if $r_{it} > 0$, otherwise $\delta(r_{it}) = 0$, and $n = \Sigma_i \delta(r_{it})$.

We want to determine weights of the form $\hat{y}_t = \Sigma_i w_i y_i$ to express these predictions. First, we express $\overline{y}$ as a sum over individual outcomes:

$$\hat{y}_t = \sum_i \frac{1}{N}\left(1 + \frac{N}{n-1}\delta(r_{it})\lambda^2 r_{it}\right)y_i - \frac{1}{N}\sum_i \sum_j \frac{N}{n-1}\frac{1}{N}\delta(r_{it})\lambda^2 r_{it}y_j \tag{A2}$$

To proceed, we flip the *i* and *j* index notations in the second term (the double sum) to get

$$\hat{y}_t = \sum_i \frac{1}{N}\left(1 + \frac{N}{n-1}\delta(r_{it})\lambda^2 r_{it}\right)y_i - \frac{1}{N}\sum_i y_i \sum_j \frac{1}{n-1}\delta(r_{jt})\lambda^2 r_{jt} \tag{A3}$$

$$\hat{y}_t = \sum_i \left(\frac{1}{N} + \frac{1}{n-1}\lambda^2 \frac{n}{n}\delta(r_{it})r_{it} - \frac{1}{n-1}\lambda^2 \frac{n}{N}\overline{r}_{sub}\right)y_i \tag{A4}$$

$$\hat{y}_t = \sum_i \left(\frac{1}{N} + \frac{\lambda^2}{n-1}(\delta(r_{it})r_{it} - \varphi\overline{r}_{sub})\right)y_i \tag{A5}$$

Thus, with $\varphi = n/N$ we have weights defined without reference to $y_i$:

$$w_{it,psr} = \frac{1}{N} + \frac{\lambda^2}{n-1}(\delta(r_{it})r_{it} - \varphi\overline{r}_{sub}) \tag{A6}$$

When we average across all weights, we once again find that $\Sigma_i w_{it} = 1$ because the two terms inside the parentheses average to the same amount:

$$\frac{1}{N}\sum_i \delta(r_{it})r_{it} = \frac{n}{N}\frac{1}{n}\sum_i \delta(r_{it})r_{it} = \varphi\overline{r}_{sub} \tag{A7}$$

$$\frac{1}{N}\sum_i \varphi\overline{r}_{sub} = \varphi\overline{r}_{sub} \tag{A8}$$

This result also applies to traditional full-sample linear regression, which is a special case of partial sample regression.

## RESULT 2: RELEVANCE-WEIGHTED AVERAGE OF PRIOR OUTCOMES EQUALS LINEAR REGRESSION PREDICTION

If we include every observation in a partial sample regression, the weights from Equation 8 may be written in a more simplified form:

$$w_{it,linear} = \frac{1}{N} + \frac{1}{N-1} r_{it} \tag{A9}$$

The prediction equation corresponding to full-sample linear regression equals

$$\hat{y}_t = \overline{y} + \frac{1}{N-1} \sum_{i=1}^{N} r_{it}(y_i - \overline{y}) \tag{A10}$$

Expanding out the expression for relevance gives

$$\hat{y}_t = \overline{y} + (x_t - \overline{x}) \frac{1}{N-1} \sum_{i=1}^{N} \Omega^{-1}(x_i - \overline{x})'(y_i - \overline{y}) \tag{A11}$$

To streamline the arithmetic, we recast this expression using matrix notation:

$$X_d = (X - 1_N \overline{x}) \tag{A12}$$

$$\hat{y}_t = \overline{y} - \overline{x}\beta + x_t\beta - (x_t - \overline{x})(X_d' X_d)^{-1} X_d' 1_N \overline{y} \tag{A13}$$

where

$$\beta = (X_d' X_d)^{-1} X_d' Y \tag{A14}$$

Noting that $X_d' 1_N$ equals a vector of zeros because $X_d$ represents attribute deviations from their own respective averages, we get the familiar linear regression prediction formula

$$\hat{y}_t = (\overline{y} - \overline{x}\beta) + x_t\beta \tag{A15}$$

$$\alpha = (\overline{y} - \overline{x}\beta) \tag{A16}$$

$$\hat{y}_t = \alpha + x_t\beta \tag{A17}$$

## RESULT 3: WHY FIT IS NORMALIZED BY THE SQUARE OF *N* MINUS 1

The definition of fit in terms of weights and outcomes equals

$$fit_t(w, y) = \frac{1}{(N-1)^2} \sum_i \sum_j r(w_{it}, w_{jt}) r(y_i, y_j) \tag{A18}$$

For notational convenience, we suppress the subscript *t* moving forward. Writing out the deviations from average inside the relevance functions and grouping the terms that depend on *i* versus *j*, we have

$$fit_t(w, y) = \frac{1}{\sigma_w^2 \sigma_y^2} \left( \frac{1}{N-1} \sum_i (w_i - \overline{w})(y_i - \overline{y}) \right) \left( \frac{1}{N-1} \sum_j (w_j - \overline{w})(y_j - \overline{y}) \right) \tag{A19}$$

It now suffices to show that either of these sums, in isolation, should be normalized by $\frac{1}{N-1}$. We proceed by writing the average weight as a sum.

$$Sum = \frac{1}{N-1} \sum_i \left( w_i - \frac{1}{N} \sum_a w_a \right)(y_i - \overline{y}) \tag{A20}$$

This expression is equivalent to summing everything over $a$ and diving everything by $N$:

$$Sum = \frac{1}{N(N-1)} \sum_i \sum_a (w_i - w_a)(y_i - \overline{y}) \tag{A21}$$

Likewise, for $\overline{y}$

$$Sum = \frac{1}{N^2(N-1)} \sum_i \sum_a \sum_b (w_i - w_a)(y_i - y_b) \tag{A22}$$

In the triple sum, there are $N^3$ terms for $(w_i - w_a)(y_i - y_b)$. However, some of the terms are trivially equal to zero and hence contain no information. The trivially zero terms occur when $a = i$ or $b = i$. There are $N$ ways for $a = i$ and $N$ ways for $b = i$ to occur, so there are $N^2$ trivially zero terms in total. We must omit the trivially zero terms from the normalizing factor, so we end up with $N^3 - N^2 = N^2(N-1)$ terms.

## RESULT 4: PARTIAL SAMPLE REGRESSION FIT

We want to evaluate the fit for partial sample regression weights:

$$fit_t(w_{psr}, y) = \frac{1}{(N-1)^2} \sum_i \sum_j r(w_{it,psr}, w_{jt,psr}) r(y_i, y_j) \tag{A23}$$

$$w_{it,psr} = \frac{1}{N} + \frac{\lambda^2}{n-1}(\delta(r_{it})r_{it} - \varphi \overline{r}_{sub}) \tag{A24}$$

We start by considering the relevance of a pair of weights:

$$r(w_{it,psr}, w_{jt,psr}) = \frac{(w_{it,psr} - \overline{w})(w_{jt,psr} - \overline{w})}{\sigma^2_{w_{psr}}} \tag{A25}$$

$$r(w_{it}, w_{jt}) = \frac{\left(\frac{\lambda^2}{n-1}\right)^2 (\delta(r_{it})r_{it} - \varphi \overline{r}_{sub})(\delta(r_{jt})r_{jt} - \varphi \overline{r}_{sub})}{\sigma^2_{w_{psr}}} \tag{A26}$$

The variance of these weights (across all $N$ observations) may be expressed as a sum over all the weights $i$ for current circumstances $t$:

$$\sigma^2_{w_{psr}} = \left(\frac{\lambda^2}{n-1}\right)^2 \frac{1}{N-1} \sum_i (\delta(r_{it})r_{it} - \varphi \overline{r}_{sub})^2 \tag{A27}$$

Substituting this term into the denominator from earlier and indexing to *k* for clarity gives

$$r(w_{it}, w_{jt}) = \frac{(\delta(r_{it})r_{it} - \varphi\bar{r}_{sub})(\delta(r_{jt})r_{jt} - \varphi\bar{r}_{sub})}{\frac{1}{N-1}\sum_k(\delta(r_{kt})r_{kt} - \varphi\bar{r}_{sub})^2} \tag{A28}$$

Note that this expression includes the censored observations. Their deviations are negative, and in the context of fit they are positively aligned with other censored observations, and they are negatively aligned with retained observations. Therefore, even the censored observations are important to the evaluation of fit.

A nice feature of this approach is that it inherently accounts for the smaller number of observations present in a highly focused, censored sample. A narrow sample is required not only to have a strong relationship among the retained observations, but it must also exhibit dispersion in outcomes for pairs of in versus out observations.

Using shorthand of *D* for the denominator, fit equals

$$fit_t(w_{psr}, y) = \frac{1}{D(N-1)^2}\sum_i\sum_j(\delta(r_{it})r_{it} - \varphi\bar{r}_{sub})(\delta(r_{jt})r_{jt} - \varphi\bar{r}_{sub})r(y_i, y_j) \tag{A29}$$

We expand out the product, noting that the two cross-terms are identical because they are specified symmetrically with respect to *i* and *j*. Therefore, we write them as two times one of the terms.

$$fit_t(w_{psr}, y) = \frac{1}{D(N-1)^2}\sum_i\sum_j(\delta(r_{it})r_{it}\delta(r_{jt})r_{jt} - 2\delta(r_{it})r_{it}\varphi\bar{r}_{sub} + (\varphi\bar{r}_{sub})^2)r(y_i, y_j) \tag{A30}$$

We break apart the sums as follows:

$$fit_t(w_{psr}, y) = \frac{1}{D(N-1)^2}\Big(\sum_i\sum_j\delta(r_{it})\delta(r_{jt})r_{it}r_{jt}r(y_i, y_j)$$
$$-\sum_i(2\delta(r_{it})r_{it}\varphi\bar{r}_{sub} + (\varphi\bar{r}_{sub})^2)z_{y_i}\sum_j z_{y_j}\Big) \tag{A31}$$

We know that $\sum_j z_{y_j} = 0$, so we are left only with the first term. Multiplying by $\frac{\sigma_r^2(n-1)^2}{\sigma_r^2(n-1)^2}$ allows us to write

$$fit_t(w_{psr}, y) = \frac{\sigma_r^2}{D}\frac{(n-1)^2}{(N-1)^2}fit_t(\delta(r_t)r_t, y) \tag{A32}$$

$$fit_t(w_{psr}, y) = \frac{\sigma_r^2}{D}\frac{N^2}{N^2}\frac{n^2}{n^2}\frac{(n-1)^2}{(N-1)^2}fit_t(\delta(r_t)r_t, y) \tag{A33}$$

$$fit_t(w_{psr}, y) = \varphi^2\frac{\sigma_r^2}{D}\left(\frac{N}{N-1}\right)^2\left(\frac{n-1}{n}\right)^2fit_t(\delta(r_t)r_t, y) \tag{A34}$$

Here, we denote the partial sample fit, which completely ignores censored observations and is normalized according to the partial sample *n*, as

$$fit_t(\delta(r)r, y) = \frac{1}{(n-1)^2}\sum_i\sum_j\delta(r_{it})\delta(r_{jt})r_{it}r_{jt}r(y_i, y_j) \tag{A35}$$

We see that the numerator of the overall fit includes $\varphi^2$. We now expand the denominator:

$$D = \frac{1}{N-1} \sum_k (\delta(r_{kt}) r_{kt} - \varphi \bar{r}_{sub})^2 \tag{A36}$$

$$D = \frac{1}{N-1} \sum_k (\delta(r_{kt}) r_{kt}^2 - 2\delta(r_{kt}) r_{kt} \varphi \bar{r}_{sub} + \varphi^2 (\bar{r}_{sub})^2) \tag{A37}$$

$$D = \frac{1}{N-1} \sum_k \delta(r_{kt}) r_{kt}^2 - 2\varphi \bar{r}_{sub} \frac{1}{N-1} \sum_k \delta(r_{kt}) r_{kt} + \frac{N}{N-1} \varphi^2 (\bar{r}_{sub})^2 \tag{A38}$$

Expressing in terms of $N$ and $n$ as desired gives us

$$D = \frac{1}{N-1} \frac{N}{N} \frac{n}{n} \frac{n-1}{n-1} \sum_k \delta(r_{kt}) r_{kt}^2 - 2\varphi \bar{r}_{sub} \frac{1}{N-1} \frac{N}{N} \frac{n}{n} \sum_k \delta(r_{kt}) r_{kt} + \frac{N}{N-1} \varphi^2 (\bar{r}_{sub})^2 \tag{A39}$$

$$D = \varphi \frac{N}{N-1} \frac{n-1}{n} \sigma_{r,partial}^2 - 2 \frac{N}{N-1} \varphi^2 (\bar{r}_{sub})^2 + \frac{N}{N-1} \varphi^2 (\bar{r}_{sub})^2 \tag{A40}$$

$$D = \varphi \frac{N}{N-1} \left( \frac{n-1}{n} \sigma_{r,partial}^2 - 2\varphi(\bar{r}_{sub})^2 + \varphi(\bar{r}_{sub})^2 \right) \tag{A41}$$

$$D = \varphi \frac{N}{N-1} \left( \frac{n-1}{n} \sigma_{r,partial}^2 - \varphi(\bar{r}_{sub})^2 \right) \tag{A42}$$

We substitute the expression for the denominator into the formula for fit and simplify to arrive at

$$fit_t(w_{psr}, y) = \varphi \left( \frac{\sigma_r^2 \frac{N}{N-1} \left( \frac{n-1}{n} \right)^2 fit_t\left( \delta(r_t) r_t, y \right)}{\frac{n-1}{n} \sigma_{r,partial}^2 - \varphi(\bar{r}_{sub})^2} \right) \tag{A43}$$

$$fit_t(w_{psr}, y) = \varphi \left( \frac{\lambda^2 \frac{N}{N-1} \frac{n-1}{n} fit_t(\delta(r_t) r_t, y)}{1 - \frac{\varphi(\bar{r}_{sub})^2}{\frac{n-1}{n} \sigma_{r,partial}^2}} \right) \tag{A44}$$

$$fit_t(w_{psr}, y) = \left( \frac{\varphi \lambda^2}{1 - \varphi c^2} \right) \frac{N}{N-1} \frac{n-1}{n} fit_t(\delta(r_t) r_t, y) \tag{A45}$$

where

$$c^2 = \frac{\bar{r}_{sub}^2}{\sigma_{r,partial}^2} \frac{n}{n-1} \tag{A46}$$

The adjusted fit of a partial sample regression is directly related to the fit associated with the subsample in isolation. However, the other terms unambiguously decrease the adjusted fit when the threshold for relevance is raised.

## ACKNOWLEDGMENTS

## REFERENCES

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Czasonis, M., M. Kritzman, and D. Turkington. *Prediction Revisited: The Importance of Observation*, pp. 155–162. Hoboken, New Jersey: John S. Wiley & Sons, 2022.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–1232.

Ho, T. K. "Random Decision Forests." In *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, August 14–16*, pp. 278–282. IEEE, 1995.

Mahalanobis, P. C. 1927. "Analysis of Race-Mixture in Bengal." *Journal of the Asiatic Society of Bengal* 23: 301–333.

——. 1936. "On the Generalised Distance in Statistics." *Proceedings of the National Institute of Sciences of India* 2 (1): 49–55.

Shannon, C. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27: 379–423, 623–656.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.