THE FUTURE OF PREDICTION

Megan Czasonis
mczasonis@statestreet.com

Mark Kritzman
kritzman@mit.edu

David Turkington
dturkington@statestreet.com

THIS VERSION: SEPTEMBER 17, 2022

Abstract

The authors describe a new prediction system based on the concept of statistical relevance, which defines the most useful observations for forming predictions. Their prediction system also encompasses the notion of fit, which enables one to assess the unique reliability of each individual prediction task. Fit also enables one to identify the uniquely optimal subsample of relevant observations, and it may be used to facilitate other aspects of situational learning. Like linear regression analysis, their relevance-based approach is grounded in theory; yet it addresses complexities that are beyond the capacity of linear regression analysis. They also compare their approach to machine learning and show that their relevance-based system is more transparent, more flexible, and less arbitrary than widely used machine learning algorithms.

THE FUTURE OF PREDICTION

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 uniquely optimal subsample of relevant observations for each prediction task.

This relevance-based approach to prediction addresses complexities that are beyond the capacity of linear regression analysis, and it does so 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 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 view of the future of prediction in 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 between the predictive variables and the outcomes. We next proceed to the present. We first discuss several model-based machine learning algorithms that are best characterized as

enhancements to linear regression analysis. Then we describe two model-free machine learning algorithms that serve as a bridge to our relevance-based prediction system. Finally, we describe our relevance-based approach to prediction, which we advocate as the future of prediction, and we compare it to both 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 estimate.

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

$$y_t = \alpha + x_t\beta + \epsilon_t \quad (1)$$

In Equation 1, ϵ_t is an error term that is assumed to be normally distributed and centered on zero. The method of least squares solves for the parameters α and β 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 β to be a column vector containing an equal number of coefficients, one can extend this linear regression model to include multiple predictive variables. And by including transformations of the original variables as new variables in the list, it can be used to linearize non-linear 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 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. And we consider two model-free algorithms: near neighbors and Gaussian kernels.

Model-Based Machine Learning Algorithms

Lasso regression focuses on the selection of predictive variables. Introduced by Tibshirani (1996), its name stands for "least absolute shrinkage and selection operator." Its practical goal is to remove, or minimize, the influence of variables that contribute the least predictive efficacy within the context of a multivariate linear regression. Whereas traditional linear regression seeks coefficients that minimize the sum of squared prediction errors, Lasso regression seeks coefficients that minimize the sum of squared errors plus an additional penalty term 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 it will be deselected – or the penalty may only justify a smaller coefficient than linear regression – in which case its value shrinks. The extent of the penalty, and thus the amount of shrinkage and selection, must be specified or 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 specified or 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. 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 must be specified or 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, and it is extremely difficult to interpret the 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.[1] 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. Generally speaking, this approach to prediction is referred to as kernel regression. One additional distinction regarding terminology is worth noting. Model-based algorithms focus on the selection of predictive variables. Model-free algorithms, instead, focus on the selection of observations. In deference to this shift in focus, we refer to predictive variables as attributes that are used to describe observations, when discussing model-free algorithms.

The near neighbors approach, for example, forms its predictions from localized sets of observations with respect to the current values of the attributes (think predictive variables). These algorithms simply exclude those observations that fall outside a chosen range around the current values of the attributes 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 attributes 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. It should be flexible, by which we mean it should be responsive to changing circumstances.
3. And it should be non-arbitrary, by which we mean it should be theoretically justified and mathematically unified. Non-arbitrariness 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, non-arbitrariness, 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.

$$\hat{y}_t = \sum_i w_{it} y_i \qquad (2)$$

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 non-arbitrary solution for computing the weights to be applied to prior outcomes, which is statistical 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.

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

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

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

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

Here, $\bar{x}$ is the average of X, and Ω^{-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 a set of attributes X.

As we emphasized earlier, this definition of statistical 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 non-normal processes can 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:

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

Information theory 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:

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

We can also justify the non-arbitrariness 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 statistically relevant observations, as determined by X. Forming a prediction from a subsample of the most

statistically relevant observations is called partial sample regression. As we show in the Appendix, these weights always sum to one.

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

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 $n = \sum_i \delta(r_{it})$ 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 $\bar{r}_{sub} = \frac{1}{n}\sum_i \delta(r_{it})r_{it}$. Finally, we include a term $\lambda^2 = \sigma_{r,full}/\sigma_{r,partial} = \frac{1}{N-1}\sum_i r_{it}^2 \,/\, \frac{1}{n-1}\sum_i \delta(r_{it})r_{it}^2$ 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 statistical relevance.
- An observation's relevance equals its similarity to current circumstances measured as the negative of half the Mahalanobis distance from current circumstances, and 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. From the Central Limit Theorem, 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 are, therefore, 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 them from 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, 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.

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

In Equation 9:

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

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

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

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

Or as follows in Equation 13:

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

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

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

or as:

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

or as:

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

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. As 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 tradeoff 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) show 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.

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

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 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 tradeoff between decreasing the number of observations included in the prediction, φ, and increasing subsample fit among the n non-censored 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 tradeoff between subsample fit $fit_{t,sub}$ and noise, but it may be useful to consider this tradeoff in more detail.

Recall from Equation 8 that we write the weights for partial sample regression as $w_{it,partial} = \frac{1}{N} + \frac{\lambda^2}{n-1}(\delta(r_{it})r_{it} - \varphi\bar{r}_{sub})$. 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, φ, λ^2, and c^2 all decrease, which unambiguously penalizes the adjusted 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.

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

The term c^2 is defined as follows:

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

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

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 test different values for r^* to maximize fit, which reflects the penalty associated with the decrease in sample size from N to n.

Thoughts on Machine Learning and Relevance-Based Prediction

Connection to Lasso Regression

As we discussed earlier, Lasso regression is a model-based machine learning algorithm for selecting predictive variables. And as we have described in detail, relevance-based prediction is a theoretically grounded system for weighting prior observations. But we can also use relevance, and, in particular, fit to select predictive variables. Rather than limit our choice of attributes (remember, our term for predictive variables) to a single collection, we can propose a variety of collections of attributes. Then we can maximize fit as a joint function of the selection of attributes and the selection of weights. Fit is specific to each prediction task, so it may recommend some collections of attributes only in rare circumstances. Lasso, by contrast, must decide whether to always or never use a given variable. This alternative approach to Lasso regression has the dual virtues of theoretical justification and flexibility to adapt to changing circumstances. This general concept may apply as well to other model-based machine learning algorithms.

Connection to Kernel Regression

Kernel regression and relevance-based prediction both form predictions as a weighted average of prior observations. They differ only in how the weights are selected. Therefore, it is quite 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. If we substitute similarity for simple proximity by using the Mahalanobis distance instead of the Euclidean distance to measure differences from current circumstances, and 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, a prediction from a kernel regression converges to a relevance-based prediction. Though these modifications may seem minor, these changes are profound. They convert an arbitrary weighting scheme to one that is theoretically justified and mathematically unified, as we discussed before.

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 non-arbitrariness.

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 works well only if the influence of the predictive variables on the prediction is static across all observations. Linear regression analysis is, therefore, incapable of addressing conditional relationships.

Non-arbitrariness: 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 estimate 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 pre-specified, 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 pre-specified by a fixed set of calibrated parameters. They do not adapt their approach to prediction circumstances.

Non-arbitrariness: Machine learning contains many disparate approaches, and their use is guided by empirical efficacy rather than by a core set of theoretical principles. This, therefore, creates a risk that its solutions result from historical happenstance rather than underlying relationships.

Relevance-Based Prediction

Transparency: Relevance-based prediction is remarkably 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 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.

Non-arbitrariness: 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 of a model's prediction tasks equals the classical R-squared statistic. Finally, fit may be used to rigorously determine the uniquely optimal subsample of relevant observations for each prediction task.

To summarize:

	Transparency	Flexibility	Non-Arbitrariness
Linear Regression Analysis	No	No	Yes
Machine Learning	No*	No*	No
Relevance-Based Prediction	Yes	Yes	Yes

* Apart from the model-free algorithms

Conclusion

We propose a new approach to prediction based on the notion of statistical relevance. 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 leaning. 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:

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

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

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

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

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

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

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

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

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

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

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

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

$$\frac{1}{N}\sum_i \varphi \bar{r}_{sub} = \varphi \bar{r}_{sub} \quad \text{(A8)}$$

This result 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:

$$w_{it,linear} = \frac{1}{N} + \frac{1}{N-1} r_{it} \quad \text{(A9)}$$

The prediction equation corresponding to full sample linear regression equals:

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

Expanding out the expression for relevance gives:

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

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

$$X_d = (X - 1_N \bar{x}) \quad \text{(A12)}$$

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

Where:

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

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:

$$\hat{y}_t = (\bar{y} - \bar{x}\beta) + x_t\beta \quad \text{(A15)}$$

$$\alpha = (\bar{y} - \bar{x}\beta) \quad \text{(A16)}$$

$$\hat{y}_t = \alpha + x_t\beta \quad \text{(A17)}$$

Result 3: Why fit is normalized by the square of N minus 1

The definition of fit in terms of weights and outcomes equals:

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

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:

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

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.

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

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

$$Sum = \frac{1}{N(N-1)}\sum_i \sum_a (w_i - w_a)(y_i - \bar{y}) \quad \text{(A21)}$$

Likewise, for $\bar{y}$:

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

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:

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

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

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

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

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

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

$$\sigma^2_{w,psr} = \left(\frac{\lambda^2}{n-1}\right)^2 \frac{1}{N-1}\sum_i (\delta(r_{it})r_{it} - \varphi\bar{r}_{sub})^2 \quad \text{(A27)}$$

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

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

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:

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

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.

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

(A30)

We break apart the sums as follows:

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

(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:

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

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

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

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 φ^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:

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

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

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

Where:

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

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.

Notes

This material is for informational purposes only. The views expressed in this material are the views of the authors, are provided "as-is" at the time of first publication, are not intended for distribution to any person or entity in any jurisdiction where such distribution or use would be contrary to applicable law and are not an offer or solicitation to buy or sell securities or any product. The views expressed do not necessarily represent the views of Windham Capital Management, State Street Global Markets®, or State Street Corporation® and its affiliates.

We thank Maryam Farboodi, Robin Greenwood, and Andrew Yimou Li for helpful comments.

References

E. T. Bell. 1986. "Men of Mathematics." Simon and Schuster. pp. 218 - 269.

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

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

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

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

S. M. Stigler. 1986. "The History of Statistics." The Belknap Press of Harvard University Press. pp. 139 - 158.

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

[1] 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.