Journal Of Investment Management, Vol. 20, No. 1, (2022), pp. 37–47

www.joim.com

RELEVANCE

Megan Czasonis[a], *Mark Kritzman*[b] *and David Turkington*[c]

The authors describe a new statistical method for improving forecasting called relevance. They describe their new method from both a conceptual and mathematical perspective, and they show how relevance links regressions to event studies and machine learning algorithms.

1 Introduction

People rely on experience to shape their view of the future, but the way they naturally process experiences differs from the method that classical statistics prescribes. The natural process is to record experiences as narratives, to focus on past experiences that are like current conditions, and to focus on unusual experiences. Classical statistics advises us to record experiences as data, to include observations irrespective of their similarity to current circumstances, and to treat unusual observations with skepticism. The purpose of our research is to reconcile classical statistics with the way people naturally process experiences based on a notion called relevance.

We proceed as follows. First, we describe relevance conceptually, including how it is used to improve forecasting. We then explain relevance mathematically, and we show how relevance relates regressions to event studies and machine learning algorithms. We conclude with a summary.

2 Relevance Conceptually

Not all observations we use to form predictions are equally relevant. We therefore propose a way to filter observations based on a statistical measure of relevance that rests on a key insight from information theory and follows from a mathematical equivalence to linear regression analysis. We determine relevance from a set of independent variables and use it to predict a dependent variable as a relevance-weighted average of the prior values of the dependent variable. This approach offers unique insights about regression analysis and often leads to better predictions.

Relevance has two components: similarity and informativeness. First, let us consider similarity.

[a]mczasonis@statestreet.com
[b]kritzman@mit.edu
[c]dturkington@statestreet.com

Although classical statistics advises us to use as many observations as we can obtain, we often focus on more recent observations when compiling data samples to use in forecasting models. This practice of censoring or de-emphasizing older observations may be quite helpful, especially if the system that produced the observations underwent structural change. The implicit assumption is that recent data is more like current conditions; therefore, it is more relevant. But recency is not the only way to judge similarity, nor is it necessarily the best way. We should also look to history for experiences that we judge to be like current circumstances and use those similar experiences to provide guidance about how the future will unfold, irrespective of the chronological position of those past experiences. This perspective does not necessarily exclude recent experiences. Rather, it offers a more general framework by which to determine the similarity of an observation, and hence its relevance. Although it makes sense to extrapolate from past experiences that resemble current conditions, doing so subjectively could be quite challenging. We therefore propose a statistical measure of the similarity of past observations with current conditions.

Now let us consider the other component of relevance, informativeness. Classical statistics tells us that if we wish to exclude observations, we should exclude those that are most extreme because they might reflect errors or arise from unusual circumstances that are unlikely to reappear. While it certainly makes sense to exclude incorrect data, we should not exclude or de-emphasize correct, outlying observations. To the contrary, we should emphasize them because they are more likely to reflect useful information as opposed to noise. Indeed, people intuitively apply this notion of informativeness when they focus on notable experiences more than mundane ones. Unlike similarity, the statistical informativeness of an observation does not depend on current conditions. It instead depends on how different an observation is from average conditions, which is to say, its improbability. This view of informativeness is consistent with information theory, which posits that the information contained in an event is inversely related to its probability.[1]

Here is how to think about the connection between probability and informativeness. Suppose we are interested in forecasting the returns of a financial asset based on some related factor. And suppose that, in addition to this factor, there are other unobservable forces that influence the asset's return. First, consider an observation in which the factor behaves in a typical way. It is not very informative, because we cannot tell if the return was caused by the factor or by the multitude of other forces we do not observe. There are many plausible explanations for the return, and we should not expect that the factor's behavior is its key determinant. Now consider an observation of the factor that is very unusual. If there is a relationship between the factor and the return, then it will have a significant impact that stands out from other influences. Therefore, whatever return we observe in this rare instance reveals a lot about the underlying relationship between the factor and the return. In general, unusual conditions are more informative because it is easier to discern the true nature of a relationship.

If we consider that informativeness does not depend on current conditions whereas similarity does, it becomes clear that the sum of both quantities across all the observations in a prediction will vary according to current conditions. To make the average relevance zero, we must add the informativeness of current conditions to the sum of similarity and informativeness for each historical observation. This result is a mathematical outcome, which we show in our mathematical description of relevance. But it has an intuitive interpretation. By including the informativeness of the current observation, we rescale relevance from a relative quantity to an absolute quantity. It is analogous to shifting the line of least squares in a regression analysis of the full sample to fit just the subsample of relevant observations. If we did not include the informativeness of the current observation, relevance could sum to an arbitrarily large positive or negative value, with the consequence that we would struggle to distinguish clearly between relevant and non-relevant observations.

To summarize, the relevance of a given observation to a prediction is the sum of its similarity to current conditions, its dissimilarity from average conditions, and the dissimilarity of current conditions from average conditions.

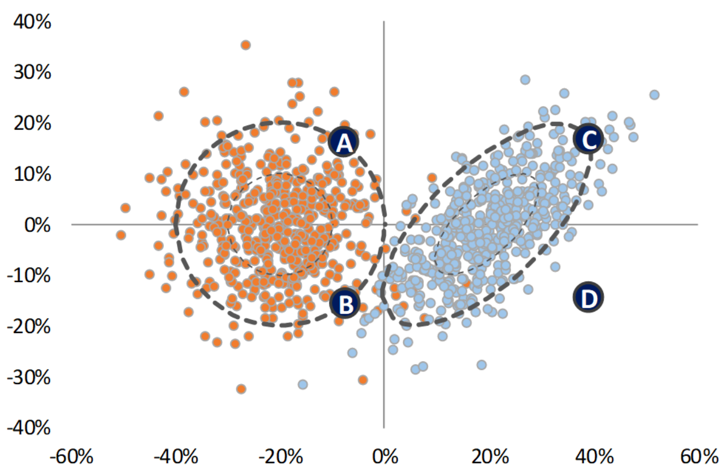
Similarity and informativeness are multivariate concepts. When we speak of observations and conditions, we have in mind a multivariate description of circumstances, specifically a vector of values for a set of independent variables. When we measure the similarity of a past observation with the current observation, we would like to consider not only the similarity of the values of each variable in isolation, but also the similarity of their co-occurrence. And when we measure the informativeness of an observation, we would like to consider both the dissimilarity of the values of each variable from average and the dissimilarity of their co-occurrence from their average co-occurrence. Put plainly, we would like to consider how variables behave independently as well as how they interact with each other when measuring similarity with current conditions or dissimilarity from average conditions.

We use a statistic called the Mahalanobis distance to measure these features of data precisely.[2] Unlike the standard Euclidean distance, the Mahalanobis distance accounts for the variances and correlations of variables. All else equal, two observations are more distant (less similar) if the spread between their values is large compared to the typical variance of those values. And all else equal, two observations are more distant if the pattern of differences between their values diverges from the typical pattern of differences in values. The Mahalanobis distance neatly summarizes these effects in a single number.

Exhibit 1 helps to visualize the Mahalanobis distance.

The scatter plot on the left is of two variables that are uncorrelated and have equal variances. The circles represent distances from the center of the

Exhibit 1 Scatter plot of two hypothetical variables.



data: that is, their average values. All points on a given circle, such as points A and B, have the same Mahalanobis distance from the center. In this case they also have the same Euclidean distance. The scatter plot on the right is of two variables that are positively correlated and have unequal variances. In this case, ellipses that are centered on the average values have the same Mahalanobis distances. However, not all observations that have the same Euclidean distance will have the same Mahalanobis distance. Consider, for example, points C and D. They both have the same Euclidean distance from the center, but C is closer in Mahalanobis distance than D. C is statistically closer because it is consistent with a positive correlation, whereas D represents an interaction that is inconsistent with a positive correlation. This exhibit illustrates how the Mahalanobis distance considers the interaction of the variables.

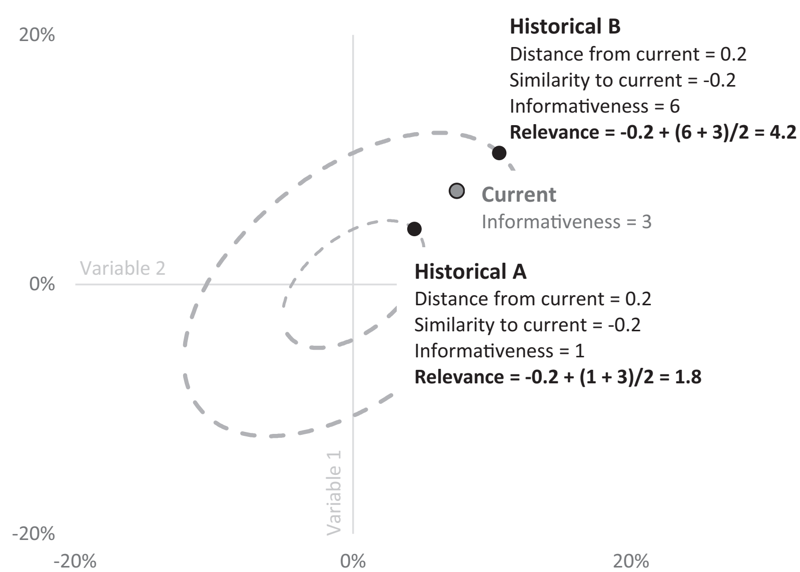
Exhibit 2 offers a visualization of the components of relevance, similarity and informativeness.

Exhibit 2 shows three observations of two hypothetical variables: two historical observations and the current observation. Historical observations A and B are equally distant from the current observation of the independent variables. They therefore have the same degree of similarity. However, historical observation B is further from the average of all the observations which are not shown but whose center is the intersection of the two axes. It is therefore more informative than historical observation A and more relevant.

Why should we care about relevance? We should care because it allows us to use observations more effectively in forecasts. To understand how relevance improves forecasting, we first need to understand how it is related to regression analysis.[3] The prediction from a linear regression equation is mathematically equivalent to a weighted average of the historical values of the dependent variable in which the weights are the relevance of the observations of the independent variables.

This equivalence reveals an intriguing feature of regression analysis. Owing to the symmetry of the observations around a fitted regression line, regression analysis places as much importance on non-relevant observations as it does on relevant

Exhibit 2 Similarity, informativeness, and relevance.



observations. It just flips the sign of the effect of the non-relevant observation on the dependent variable. This feature of regression analysis invites a fundamental question about forecasting: Are non-relevant observations as useful in forming a prediction as relevant ones? In some cases, they may be, but not always, and perhaps not usually. Suppose, for example, we wish to forecast the economic outcomes of a recession. Should we place as much importance on past conditions of robust growth as on past recessions? This is an empirical question, but we suspect that intuition is often right to suggest that relevant observations are more useful to a forecast than non-relevant observations.

This insight about how regression analysis treats relevant and non-relevant observations leads to the key innovation we propose for forecasting. Researchers should consider a two-step approach to forecasting. First, create a subsample of relevant observations. And second, form the prediction as a relevance-weighted average of the past values of the dependent variable in the subsample. This two-step approach to forecasting is called partial sample regression.[4]

One might ask why we should not simply apply regression analysis to the subsample of relevant observations. Why do we instead take a weighted average of the past values of the dependent variable? The answer is that the weights preserve valuable information about relevance in the context of the full sample. If we were to apply regression analysis to the relevant subsample, it would consider some of the relevant observations as not relevant and interpret them opposite to the way they should be used to inform the prediction.

Perhaps at this point it would be useful to summarize our concept of relevance.

(1) The relevance of an observation is determined by a set of independent variables for the purpose of forecasting a dependent variable. It equals the sum of an observation's similarity with current conditions and its dissimilarity from average conditions. Dissimilarity from average conditions measures an observation's informativeness.
(2) By including similarity in our definition of relevance, we are simply following intuition, which often directs us to consider past events that are like current conditions to help us think about the path forward.
(3) Observations that are dissimilar from their average values are more informative than observations that are like their average values, because it is easier to discern causality from unusual observations than from common observations.
(4) We should include both the informativeness of past observations as well as the informativeness of current observations because by including both, the relevance of all the observations sums to zero, which establishes zero as a natural threshold for relevant and non-relevant observations.
(5) To measure an observation's similarity with current conditions we should consider the isolated similarity of the variables' values with current values, as well as the similarity of their co-occurrence with the co-occurrence of current values. The same is true for how we measure dissimilarity from the average values to determine informativeness. We should consider the values of the variables in isolation as well as how they interact with each other.
(6) We use a statistic called the Mahalanobis distance to measure similarity and informativeness. The Mahalanobis distance considers variables independently as well as how they interact with each other. It also converts all values into common units.
(7) The prediction from a linear regression model is mathematically equivalent to a

relevance-weighted average of the past values of the dependent variable if it is averaged over the full sample.

(8) This equivalence reveals that regression analysis places as much importance on non-relevant observations as it does on relevant observations, which is often counter-productive.

(9) We should therefore consider forming our prediction as a relevance-weighted average of the dependent variable from a subsample of observations that have positive relevance.

(10) We should not, however, apply regression analysis to a subset of relevant observations, because it will interpret some of the relevant observations in a way that is opposite to how they should inform the prediction.

3 Relevance Mathematically

In our conceptual discussion, we explained relevance within the context of a current observation and past observations. We now define it more generally between any pair of observations x_i and x_j, to stress that relevance is a symmetric measure. We assume that there are N historical observations of M independent variables, which we organize into an N-by-M matrix X. The observations x_i and x_j are both row vectors with M elements corresponding to the value of each independent variable. We define similarity and informativeness in terms of these two vectors along with the mean vector $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and the M-by-M covariance matrix $\Omega = \frac{1}{N-1}(X - \bar{x})'(X - \bar{x})$ of X. We use the notation $'$ to indicate matrix transpose, and we denote the inverse of the covariance matrix as Ω^{-1}.

$$\begin{aligned} sim_{ij} &= sim(x_i, x_j) \\ &= -\frac{1}{2}(x_i - x_j)\Omega^{-1}(x_i - x_j)' \end{aligned} \quad (1)$$

$$\begin{aligned} info_i &= info(x_i) \\ &= (x_i - \bar{x})\Omega^{-1}(x_i - \bar{x})' \end{aligned} \quad (2)$$

Similarity equals the Mahalanobis distance between x_i and x_j, in its squared form, multiplied by negative 1/2. It may be helpful to consider the purpose of each step in this calculation. The spread between the vectors measures the similarity of the values for each variable in isolation. Multiplying by the inverse of the covariance matrix converts the spreads for each variable into common units, effectively dividing each spread by the variance of the corresponding variable. It also captures the similarity of the co-occurrence of the variables compared to their typical patterns of co-occurrence. When we post multiply by the spreads between the vectors, we collapse the result into a single number. The negative sign converts the notion of distance into one of closeness (similarity). The factor of 1/2 offsets the double counting that occurs from the identical multiplication of x_i with x_j and x_j with x_i for pairwise comparisons.

We measure informativeness as the Mahalanobis distance between x_i and $\bar{x}$, the full sample mean of X. This time, however, we retain the positive value of the distance, as we are interested in how dissimilar or distant the observations are from the average values. Also, we do not need to multiply by 1/2 because the observation is compared to the average, rather than to another observation.

We define relevance as in Equation (3).

$$r_{ij} = r(x_i, x_j) = sim_{ij} + \frac{1}{2}(info_i + info_j) \quad (3)$$

Recall from our earlier conceptual description of relevance that we include the average informativeness of both x_i and x_j so that the relevance of all observations sums to zero. This result enables us to use a threshold of zero to separate relevant observations from non-relevant observations. The inclusion of both vectors' informativeness is also motivated by symmetry, whereby the relevance of observation x_i to predicting conditions that match x_j is equal to the relevance of observation x_j to predicting conditions that match x_i. In the discussion that follows, we define the conditions of our prediction as x_t; therefore, r_{it} gives the relevance of any historical observation x_i to these current conditions.

Relevance is independent of the object of our prediction, Y. In the absence of any information from the X variables, our best prediction $\hat{y}_t$ of an unknown y_t would be the simple average, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. But the utility of relevance is that we may enhance that estimate by adding a weighted average of the historical deviations of Y from their average, where the weights are the relevance of each x_i to x_t.

$$\hat{y}_t = \bar{y} + \frac{1}{N-1}\sum_{i=1}^{N} r_{it}(y_i - \bar{y}) \tag{4}$$

Let us now proceed by demonstrating the equivalence between Equation (4) and ordinary least squares linear regression. First, we rearrange and consolidate the expression for relevance from Equation (3).

$$\begin{aligned} r_{it} = & -\frac{1}{2}(x_i - x_t)\Omega^{-1}(x_i - x_t)' \\ & + \frac{1}{2}(x_i - \bar{x})\Omega^{-1}(x_i - \bar{x})' \\ & + \frac{1}{2}(x_t - \bar{x})\Omega^{-1}(x_t - \bar{x})' \end{aligned} \tag{5}$$

$$\begin{aligned} r_{it} = & \, x_t\Omega^{-1}x_i' - x_t\Omega^{-1}\bar{x}' - \bar{x}\Omega^{-1}x_i' \\ & + \bar{x}\Omega^{-1}\bar{x}' \end{aligned} \tag{6}$$

$$r_{it} = (x_t - \bar{x})\Omega^{-1}(x_i - \bar{x})' \tag{7}$$

We substitute Equation (7) into Equation (4).

$$\begin{aligned} \hat{y}_t - \bar{y} = & \frac{1}{N-1}\sum_{i=1}^{N}(x_t - \bar{x})\Omega^{-1} \\ & \times (x_i - \bar{x})'(y_i - \bar{y}) \end{aligned} \tag{8}$$

Equation (8) predicts the value of Y above its average based on observations of X above its average and Y above its average. The covariance matrix, by definition, is also a function of X above its average. Therefore, without loss of generality we may rewrite the prediction formula under the assumption that X and Y have means of zero. We pull $x_t\Omega^{-1}$ out of the sum because this term does not depend on i.

$$\hat{y}_t = x_t\Omega^{-1}\frac{1}{N-1}\sum_{i=1}^{N} x_i' y_i \tag{9}$$

Using matrix notation whereby X contains all N observations in rows and M variables in columns, and Y contains all N observations in rows with one column and noting that $\Omega^{-1} = (N-1)(X'X)^{-1}$, we obtain the standard formula for a linear regression prediction.

$$\hat{y}_t = x_t(X'X)^{-1}X'Y \tag{10}$$

$$\hat{y}_t = x_t\beta' \tag{11}$$

4 A Unified Perspective on Observation-Based Prediction

We now show how the framework of relevance-weighted prediction relates two common prediction methods that are typically considered to be distinct: regression analysis and event studies.

We begin with the recognition that linear regression, and its relevance-weighted equivalent, does not discriminate between highly relevant and highly non-relevant observations other than flipping the signs of their predictive contributions. In cases where relevant observations are more reliable than non-relevant ones, it may be better to remove the non-relevant observations and generalize Equation (4) as a sum over a chosen subsample of n observations, where $1 \geq n \geq N$:

$$\hat{y}_t = \bar{y} + \frac{\lambda^2}{n-1} \sum_{i=1}^{n} r_{it}(y_i - \bar{y}) \qquad (12)$$

The application of Equation (12) to a subsample of the most relevant observations is called partial sample regression. The λ^2 in this formula equals the variance of r_{it} in the full sample divided by the partial variance of r_{it} in the subsample: $\lambda^2 = 1/(N-1)\sum_{i=1}^{N} r_{it}^2 / 1/(n-1)\sum_{i=1}^{n} r_{it}^2$. This ratio equals exactly 1 for a full sample regression and is very close to 1 for a partial sample regression that includes half the observations. It removes a bias that would otherwise occur for a more selective partial sample regression.

If we apply Equation (12) to a single observation, $n = 1$, which we choose for any reason, we end up with the outcome of a single event. For example, consider an event study intended to give a prediction of the path of a chosen variable following an event that just occurred or is anticipated to occur. As a simple approach, we might identify a single past observation based on judgment, intuition, or a set of exogenous variables, and record the outcome of Y at a range of time intervals after the event. The single observation we choose could be the most relevant one, but it need not be. In either case, we may consider each time interval observation of Y around the historical event as an application of Equation (12) with one data point. When $n = 1$, our prediction for Y converges to its actual occurrence following the chosen event.

A composite event study that draws upon multiple events, $n < N$, is potentially more interesting and more statistically robust. To conduct a composite event study, we identify a sample of events and align their chronological position to $t = 0$. We then observe the value of Y at various times following the event, $t+1, t+2, t+3, \ldots$, and we compute the arithmetic means of these outcomes across the n events. We interpret these post-event means as predictions for the path Y will take from a recent or anticipated event.

Selecting the events for a composite event study is equivalent to censoring non-relevant observations as we do in a partial sample regression, but in a composite event study we are using criteria other than the relevance of X to determine the subsample. Now suppose that rather than predicting the path forward as the simple arithmetic mean of the observed paths for our defined events, we weight the observations by their statistically determined relevance. This relevance-weighted event study is equivalent to partial sample regression, with the exception that non-relevant observations are censored based on identification as non-events as opposed to the statistical relevance of X.

To summarize:

(1) The prediction from a linear regression equation is mathematically equivalent to a weighted average of past values of the dependent variable in which the weights are the relevance of the independent variables.
(2) This equivalence allows one to form a relevance-weighted prediction of the dependent variable by using only a subsample of relevant observations. This approach is called partial sample regression.
(3) Like partial sample regression, an event study separates relevant observations from non-relevant observations, but it does so by using criteria other than statistically determined relevance.
(4) As an alternative to predicting the path from a recent or current event as an arithmetic mean of past paths, one could use

> a relevance-weighted average of past paths to form a prediction. This approach would be equivalent to partial sample regression in which the relevant subsample is determined by a separate identification process rather than statistical relevance.

It is worth noting that the general notion of relevance can extend to more complex prediction models as well. As we have described, relevance allows us to weight the contribution of historical observations in a rigorous statistical way that converges to ordinary least squares regression when we include all the observations in a sample. It also allows us to censor some observations by effectively setting their weight to zero and rescaling the remaining observations. The resulting partial sample regression forecasts still comprise a linear combination of historical outcomes for the variable we aim to predict. However, the predictions are no longer linear with respect to the current conditions that inform the predictions. This occurs because different input conditions censor different sets of historical observations, thereby introducing a conditional nonlinear dependency on the prediction inputs. The data that fuels the prediction consists of observations that are sufficiently informative and similar to current conditions. Various machine learning algorithms implement related logic, though typically in a more complex fashion. For example, tree-based prediction models including random forest and boosted machines solve for thresholds that determine which observations to include and which to censor. In this sense, these models determine their own notions of relevance which differ from ours in practice, though not in principle.

The measure of statistical relevance we propose based on the Mahalanobis distance has a few comparative advantages. First, it accounts for informativeness in addition to similarity, which many other methods do not. Second, it equates to linear regression when applied to a full sample of observations, thereby aligning with the simplicity and efficiency of linear regression analysis. Third, it is transparent, intuitive, and may be calculated directly, without resorting to numerical iteration. And fourth, by virtue of the simplicity of censoring irrelevant data while retaining the linear structure of prediction, relevance offers a powerful and parsimonious way to introduce conditionality into prediction, which we call partial sample regression. The practical efficacy of one approach compared to another is an empirical question, but we suspect there are many instances in which the properties of statistical relevance confer an advantage. Further linking the notion of relevance to more complex nonlinear prediction models may offer a productive area for future research.

5 Conclusion

Although most of us think long and hard about which variables to use in our forecasts, we typically tend not to think as much about which observations of those variables to include or emphasize. To the extent we do consider observations, we are often inclined to place greater emphasis on more recent observations than on more distant observations. However, when we think intuitively about how to forecast into the future from present conditions, we often look to past episodes in history that are like present conditions. This intuition is sound and helpful. Observations that are like current conditions are more relevant to a forecast than dissimilar observations. But not all observations that are equally like current conditions are equally relevant. Observations that are unusual are more relevant than common observations, because it is easier to discern causality from unusual observations than from common observations. Thus, unusual observations are more informative.

The statistical relevance of an observation to a prediction is determined by its similarity to current conditions, its dissimilarity from average conditions (which captures its informativeness), and the informativeness of current conditions. We include the informativeness of current conditions to facilitate a natural interpretation of relevance in absolute terms. By including it, the relevance of all observations sums to zero, which enables us to use a threshold of zero to separate relevant observations from non-relevant observations.

When we measure relevance, not only must we measure the similarity of variable values with their current values or their dissimilarity from average values in isolation. We must also consider the similarity or dissimilarity of their co-occurrence. We therefore use a statistic called the Mahalanobis distance to measure similarity and informativeness. This statistic has two valuable features: it considers the interaction of the variables, and it converts their values into common units.

Our conception of relevance is not arbitrary. The prediction from ordinary least squares linear regression is mathematically equivalent to a weighted average of the past values of the dependent variable in which the weights are the relevance of the independent variables. This equivalence reveals a key insight about regression analysis, which is that owing to the symmetry of observations around a fitted regression line, regression analysis places as much importance on non-relevant observations as it does on relevant observations; it just flips the sign of the effect of the non-relevant observation on the dependent variable.

This insight about regression analysis invites a fundamental question. Is it possible to produce a better forecast from a subsample of relevant observations than from the full sample? The answer, of course, can only be determined empirically, but it is not hard to imagine settings in which our intuition would rightly suggest that we exclude non-relevant observations.

We should therefore consider a two-step approach to forecasting. First create a subsample of relevant observations. Then, form the forecast by taking a relevance-weighted average of the observations from the relevant subsample. This two-step approach to forecasting provides a unified perspective on relevance, regression analysis, event studies, and machine learning algorithms.

Notes

Notes

1 See, for example, Shannon (1948).

2 The Mahalanobis distance was introduced by an Indian statistician in 1927 and modified by him in 1936 to analyze resemblances in human skulls among people of different parentage in India. Mahalanobis compared a set of measurements for a chosen skull with the average of those measurements across skulls from different groups. He also compared the co-occurrence of those measurements for a chosen skull to their covariation within each group. He summarized these comparisons in a single number which he used to place a given skull into a particular group.

[3] When we use the terms regression or regression analysis, we have in mind ordinary least squares linear regression analysis.

[4] See Czasonis *et al.* (2020) for a thorough discussion of partial sample regression.

References

Czasonis, M., Kritzman, M., and Turkington, D. (2020). "Addition by Subtraction: A Better Way to Predict Factor Returns (and Everything Else)," *The Journal of Portfolio Management* **46**(8) (September).

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

Mahalanobis, P. C. (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**, pp. 379–423, 623–656, July, October.

Keywords: Boosted machines, euclidean distance, event study, information theory, informativeness, machine learning algorithm, Mahalanobis distance, partial sample regression, random forest, relevance, similarity.