CONFIDENCE REVISITED: THE DISTRIBUTION OF INFORMATION

THIS VERSION: October 17, 2025

Megan Czasonis, Mark Kritzman, Fangzhong Liu, and David Turkington

Megan Czasonis is a managing director at State Street Associates in Cambridge, MA.
mczasonis@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

Mark Kritzman is the chief executive officer of Windham Capital Management in Cambridge, MA, and a senior lecturer at the MIT Sloan School of Management in Cambridge, MA.
kritzman@mit.edu
100 Main Street, Cambridge MA, 02142

Fangzhong Liu is a vice president at State Street Associates in Cambridge, MA.
fliu6@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

David Turkington is senior managing director and head of State Street Associates in Cambridge, MA.
dturkington@statestreet.com
140 Mt Auburn Street, Cambridge MA, 02138

Key Takeaways

Classical statistics assesses the confidence of a prediction by measuring error variance, estimation variance, or the sum of both. More complex machine learning models assess confidence using empirical methods such as resampling or conformal prediction.

Model-based assessments of confidence, however, are likely to be unrealistically tight because they assume that the model is specified correctly and the predictive variables are measured without error.

Relevance-based prediction, a model-free prediction technique, assesses the reliability of a prediction from the distribution of information that is used to form a prediction. It gives a more realistic assessment of an individual prediction's reliability because it does not rely on model correctness nor the assumption that the predictive variables are error free.

Abstract

Classical statistics assesses the confidence of a prediction by measuring error variance, estimation variance, or the sum of both. More complex machine learning models assess confidence using empirical techniques such as resampling or conformal prediction. These methods implicitly assume that the model is correctly specified and the predictive variables are measured without error. The authors describe a model-free approach to prediction called relevance-based prediction that does not rely on a model's correctness, nor does it assume that the predictive variables are error free. It measures the uncertainty of each individual prediction based on the unique distribution of information that is used to form the prediction.

CONFIDENCE REVISITED: THE DISTRIBUTION OF INFORMATION

Data-driven predictions are usually formed by models such as a linear regression model or a neural network. In classical statistics, there are three distinct concepts that relate to a prediction's confidence: error variance reflects irreducible randomness in outcomes, estimation variance reflects noise in the prediction value that results from error-prone parameter estimates, and predictive variance is the sum of error variance and estimation variance which determines the width of a prediction interval. These concepts can be derived analytically for a linear regression model. For more complex machine learning models such as a neural network, it is usually impossible to determine these quantities theoretically, but one may approximate them using empirical techniques such as resampling or conformal prediction. In all these cases, for both classical statistics and machine learning, the estimated measures of confidence may be unrealistically tight because they assume that the model is correctly specified and the predictive variables are measured without error.

We propose a new framework for measuring prediction uncertainty based on a model-free prediction technique called relevance-based prediction (RBP). RBP gives a measure of prediction uncertainty that does not depend on the correctness of a parameterized model because there is no model, nor does it assume that the predictive variables are error free. RBP forms a prediction as a weighted average of observed outcomes, which allows us to see precisely how each observation informs the prediction. This transparency allows us to observe the distribution of information that goes into each prediction. RBP associates prediction uncertainty with the alignment of the information that goes into a prediction. If the information is mutually supportive, which would show up as a tight distribution, we should be more confident in the prediction. If it is mutually contradictory, as evidenced by a wide distribution, we should be less confident.

We proceed as follows. First, we describe partial sample regression, which is a building block of RBP. It forms a prediction as a weighted average of observed outcomes in which the weights come from a precisely defined and theoretically justified statistic called relevance. Linear regression predictions emerge as a special case of partial sample regression. We show how to construct distributions of the information that is used to form individual predictions, and we provide simulations to support our analysis. We then introduce the notion of fit which quantifies the prevalence of useful patterns in a dataset pertaining to one specific prediction. Fit enables us to compute a composite prediction from a prediction grid that comprises many combinations of observations and predictive variables. Each cell in the prediction grid gives a partial sample regression prediction and an associated measure of reliability. The grid's composite prediction is formed as a reliability-weighted average of each cell's prediction. We describe how to form distributions of information aggregated from all the cells that form the grid's composite prediction. We conclude with a summary.[1]

Partial Sample Regression

Partial sample regression forms a prediction as a weighted average of observed outcomes in which the weights are based on a statistic called relevance. Relevance is composed of similarity and informativeness which are both measured as Mahalanobis distances, as shown in equations 1 through 4.

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

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

$$info(x_i, \bar{x}) = (x_i - \bar{x})\Omega^{-1}(x_i - \bar{x})' \tag{3}$$

$$info(x_t, \bar{x}) = (x_t - \bar{x})\Omega^{-1}(x_t - \bar{x})' \tag{4}$$

In equations 1 through 4, x_i is a vector of the values of K predictive variables for a prior observation, x_t is a vector of the values of the predictive variables for a specific prediction task, $\bar{x} = 1_N 1_N' X N^{-1}$ is the average of the predictive variables across all observations, and Ω^{-1} is the inverse covariance matrix of all the observations of the variables. The vector $(x_i - x_t)$ measures how distant each variable's observed value is from its corresponding value in the prediction task, when measured in isolation. By multiplying this vector by the inverse covariance matrix, we capture the interaction of the predictive variables, and at the same time we standardize the distances by dividing by variance. By multiplying this product by the transpose of the vector $(x_i - x_t)$ we consolidate the outcome into a single number. All else being equal, observations that are like current circumstances but different from average circumstances are more relevant than those that are not.

This definition of relevance is not arbitrary. We know from information theory that the information contained in an observation is the negative logarithm of its likelihood.[2] We also know from the Central Limit Theorem that the relative likelihood of an observation from a multivariate normal distribution is proportional to the exponential of a negative Mahalanobis distance. Therefore, the information contained in a point on a multivariate normal distribution is proportional to a Mahalanobis distance.

If we define weights in terms of relevance as shown by equation 5, which admits the relevance-weighted average of every prior outcome in the full sample of observations, the result is precisely equivalent to the prediction given by linear regression analysis.[3]

$$w_{it,linear} = \frac{1}{N} + \frac{1}{N-1} r_{it} \tag{5}$$

Partial sample regression assumes that we can produce a more reliable prediction by censoring observations that are less relevant than a chosen threshold, which leads to the following definition of prediction weights.

$$w_{it,retained} = \frac{1}{N} + \frac{\lambda^2}{n-1}\left(\delta(r_{it}) r_{it} - \varphi \bar{r}_{sub}\right) \tag{6}$$

$$\delta(r_{it}) = \begin{cases} 1 & if\ r_{it} \geq r^* \\ 0 & if\ r_{it} < r^* \end{cases} \tag{7}$$

$$\lambda^2 = \frac{\sigma^2_{r,full}}{\sigma^2_{r,retained}} = \frac{\frac{1}{N-1}\sum_{i=1}^{N} r_{it}^2}{\frac{1}{n-1}\sum_{i=1}^{N} \delta(r_{it}) r_{it}^2} \tag{8}$$

In equations 5 through 8, $n = \sum_{i=1}^{N} \delta(r_{it})$ is the number of observations that are fully retained, $\varphi = n/N$ is the fraction of observations in the retained sample, and $\bar{r}_{sub} = \frac{1}{n}\sum_{i=1}^{N} \delta(r_{it}) r_{it}$ is the average relevance value of the observations in the retained sample. It is important to note that $w_{it,retained}$ depends crucially on the prediction circumstances x_t. Relevance is reassessed for each prediction circumstance which further affects the identification of the retained subsample and introduces nonlinear conditional dependence of the prediction $\hat{y}_t$ on the prediction circumstances x_t. The scaling factor λ^2 compensates for a bias that would otherwise result from relying on a small subsample of highly relevant observations. In the case of linear regression analysis $n = N$ and $\lambda^2 = 1$. Lastly, note that the regression weights always sum to 1.[4]

The prediction given by partial sample regression is shown by equation 9.

$$\hat{y}_t = \sum_i \left(\frac{1}{N} + \frac{\lambda^2}{n-1} (\delta(r_{it}) r_{it} - \varphi \bar{r}_{sub}) \right) y_i \tag{9}$$

We can equivalently express the prediction in terms of deviations from the full sample $\bar{y}$ as shown by equation 10.

$$\hat{y}_t - \bar{y} = \frac{\lambda^2}{n-1} \sum_i \delta(r_{it}) r_{it} (y_i - \bar{y}) \tag{10}$$

Even though equation 10 uses $\bar{y}$ based on the full sample, only non-censored retained observations cause the prediction to tilt away from the full sample average. The prediction's tilts emanate only from the subsample where $\delta(r_{it}) = 1$. This feature allows us to construct a predictive distribution because we ignore censored observations. We cannot ignore censored observations in equation 6 because the term $\varphi \bar{r}_{sub}$ implicitly manufactures the effect of the full sample $\bar{y}$.

As we have shown previously,[5] we can express λ^2 equivalently as follows.

$$\lambda^2 = \frac{\sigma^2_{r,full}}{\sigma^2_{r,part}} = \frac{info(x_t)}{\sigma^2_{r,part}} = \frac{info(x_t)}{\frac{1}{n-1} \sum_j \delta(r_{jt}) r^2_{jt}} = \frac{(n-1) info(x_t)}{\sum_j \delta(r_{it}) r^2_{jt}} \tag{11}$$

These equivalences allow us to cancel out $n - 1$ in equation 10, giving us a new equation.

$$\hat{y}_t - \bar{y} = \frac{info(x_t)}{\sum_j \delta(r_{jt})r_{jt}^2} \sum_i \delta(r_{it}) r_{it}(y_i - \bar{y}) \tag{12}$$

Solo prediction

By removing $n - 1$ from equation 10, we are now able to make a prediction from a single observation, $solo(i)$, which has deviations from $\bar{y}$ as shown in equations 13 and 14.

$$\hat{y}_{t,solo(i)} - \bar{y} = \frac{info(x_t)}{r_{it}^2} r_{it}(y_i - \bar{y}) \tag{13}$$

$$\hat{y}_{t,solo(i)} - \bar{y} = \frac{info(x_t)}{r_{it}} (y_i - \bar{y}) \tag{14}$$

The solo prediction scales the outcome deviation $(y_i - \bar{y})$ by a simple factor, which is intuitive. If the observation is equal to x_t, the scaling factor will equal 1 and the prediction will be formed from the outcome that occurred in the identical circumstance. All else being equal, as the informativeness of the current prediction circumstance x_t increases, the prediction amplifies the impact of the deviation for observation i.[6] If observation i is highly relevant, the deviation will be scaled back. If the observation is not very relevant, it will be scaled up. This scaling is necessary to express the observed outcome in the same expected scale as the prediction task. The scaling factor also contains a positive or negative sign, so it has the potential to flip the direction of the deviation. The solo prediction is technically undefined when relevance equals exactly zero because the observation is orthogonal to the task and cannot be used to inform it.

As we show next, we can combine solo predictions to form the actual prediction thereby accounting for multiple observations. Observations that we censor using $\delta(r_{it}) = 0$ receive zero weight in the composite sum, thus nullifying the impact of their extreme solo predictions. Solo predictions that are undefined because relevance equals exactly zero also receive zero weight in the composite sum, and it is appropriate to treat these instances as zero contributions to the composite prediction.

Contribution of a solo prediction

Consider the contribution, $contr(i)$, of a single observation to a multi-observation prediction's deviation from the unconditional average:

$$\hat{y}_{t,contr(i)} - \bar{y} = \frac{info(x_t)}{\sum_j \delta(r_{jt}) r_{jt}^2} r_{it}(y_i - \bar{y}) \tag{15}$$

If we express contribution as a weight ξ_i times that observation's solo prediction:

$$\hat{y}_{t,contr(i)} - \bar{y} = \xi_i\left(\hat{y}_{t,solo(i)} - \bar{y}\right) \tag{16}$$

Then:

$$\xi_i = \frac{\hat{y}_{t,contr(i)} - \bar{y}}{\hat{y}_{t,solo(i)} - \bar{y}} \tag{17}$$

This substitution cancels out many terms, leading to:

$$\xi_i = \frac{r_{it}^2}{\sum_j \delta(r_{jt}) r_{jt}^2} \tag{18}$$

The weight placed on the solo prediction of a retained observation ($\delta(r_{it}) = 1$) is equal to its squared relevance as a fraction of the sum of the squared relevance of all the other retained observations. Clearly these weights sum to 1, they are all nonnegative, and the weights of the censored observations in this context are all 0.

Building a distribution of solo predictions

Rewriting the weight formula so that the weights are all positive allows us to calculate a distribution of solo predictions. Each solo prediction is scaled and signed as a conceptually viable prediction which receives a weight ξ_i in the actual partial sample regression prediction. Together, these solo prediction deviations and weights define a histogram or distribution of relative likelihood.

Censored observations receive zero weight, so they are censored from the distribution as they are from the prediction. They have no effect irrespective of their solo prediction values. Observations with very low relevance may result in unrealistic solo predictions, but their probability weights in the histogram will be extremely small. Rather than showing unrealistic outliers in a distribution that extrapolates noise to extremes, a graphical solution for this issue is to show cumulative tail bars on both sides of the distribution, limiting the display to a reasonable range. This distribution will only reflect retained observations, just as the prediction itself only reflects retained observations (setting aside the computation of $\bar{y}$).

It is important to distinguish this distribution from the distributions implied by model-based estimates of error variance and estimation variance. The distribution of solo predictions is a distribution of the information that is used to form each individual prediction $\hat{y}_{t,solo(i)}$. It, therefore, allows us to measure prediction uncertainty in a way that captures nuanced details about how each prediction is formed which would be obscured by a distribution that is derived from summary statistics or empirically resampled predictions. Moreover, unlike classical measures of confidence which are based on the same theoretical distribution for all predictions, the distributions of solo predictions are specific to each individual prediction task.

Simulations of the distribution of solo predictions

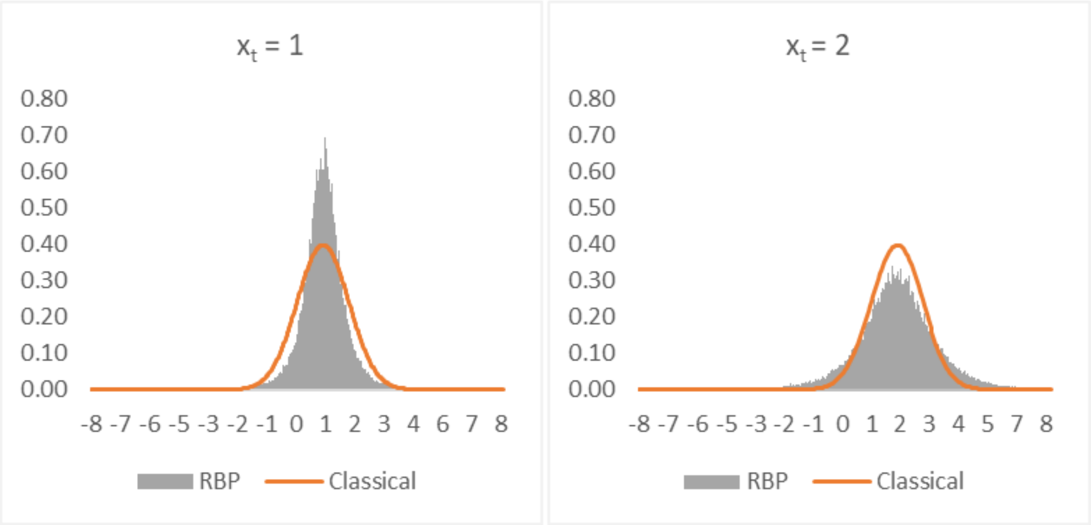
Linear Regression Analysis

We now simulate a synthetic data sample of 100,000 observations to illustrate the distributions of $\hat{y}_{t,solo(i)}$. First, we randomly draw values of a single predictive X variable from a standard normal distribution. Next, we form outcomes Y as the values of the predictive variable plus the values of a simulated error term which also comes from a standard normal distribution.

To provide a point of reference with classical statistics, we show the distribution of $\hat{y}_{t,solo(i)}$ for predictions given by linear regression analysis, which is the only model-based prediction for which we observe the distribution of $\hat{y}_{t,solo(i)}$. We can observe this distribution owing to the equivalence of linear regression analysis with RBP in the special case in which partial sample regression is applied to all the observations.

The left panel of Exhibit 1 shows the distribution of the solo predictions for the prediction task of $x_t = 1$ and the right panel shows the distribution for the task $x_t = 2$. The gray bars show the distribution of the solo predictions given by each observation, and the orange lines show the classical distribution of potential outcomes implied by the error variance estimated from the regression residuals. Both are probability density functions that integrate to 1.

Exhibit 1: Solo Distributions and Classical Error Variance Distributions



This example highlights the differences between RBP's distribution of solo predictions, classical error variance, and classical estimation variance. In this experiment, the classical estimation variance of the prediction is extremely small because the noise in 100,000 observations mostly cancels out in the prediction estimates. In other words, the predictions $\hat{y}_t$ are nearly certain to be close to their theoretical values given this large sample. We do not show the distributions implied by the estimation variance in Exhibit 1 because they would be extremely pointed narrow distributions near the average prediction. By contrast, the classical error variance, which creates the distributions shown in orange, do not become so narrow because no matter how reliable the overall prediction value becomes, actual outcomes include the influence of the error term which has an irreducible variance. Likewise, the distribution of solo predictions does not converge to a narrow outcome because the information contributed by each observation contains noise from the randomness in the data sample.

There are several interesting takeaways from Exhibit 1. First, the distribution of the solo predictions when the value of the predictive variable equals 1 and 2 are both reasonably close to the classical error distribution which confirms that they are sensible. Second, the distributions of solo predictions are different from each other which highlights the fact that the solo predictions depend crucially on the prediction circumstances. The classical error distribution, by contrast, has the same variance irrespective of the value of the predictive variable. It is based on the average squared residuals across all predictions and does not recognize that some predictions can be more reliable than others. It is worth noting that if we average the variance of the solo prediction distributions across all prediction tasks x_t in the sample, the average of all these distributions will converge to the classical error variance for the special case in which there is a single predictive variable. See the Appendix for an explanation of this result.

Partial Sample Regression

We now illustrate distributions of $\hat{y}_{t,solo(i)}$ for predictions given by partial sample regression based on simulated data from a more complex data generating process. Partial sample regression is based on the premise that we can form more reliable predictions from subsamples of relevant observations than the full sample of observations. This is so because in many prediction circumstances non-relevant observations contradict or obscure useful patterns that would otherwise allow us to form a more reliable prediction.

To illustrate the information distributions that come from partial sample regression, we draw simulated observations from two distinct regimes whose observations for X collectively form a bimodal distribution. One of the regimes, which we refer to as the majority regime, occurs 75% of the time. Its predicted outcomes are equal to the values of X drawn from a normal distribution with a mean of 5 and a standard deviation of 3 plus an error term that is normally distributed with a mean of 0 and a standard deviation of 3. The second regime, which we refer to as the minority regime, occurs 25% of the time. Its predicted outcomes are equal to 2 times the values of X drawn from a normal distribution with a mean of -5 and a standard deviation of 2 plus an error term that is normally distributed with a mean of 0 and a standard deviation of 3.

Exhibit 2: Boxplots of Predictions of Outcomes for Different Censoring Thresholds

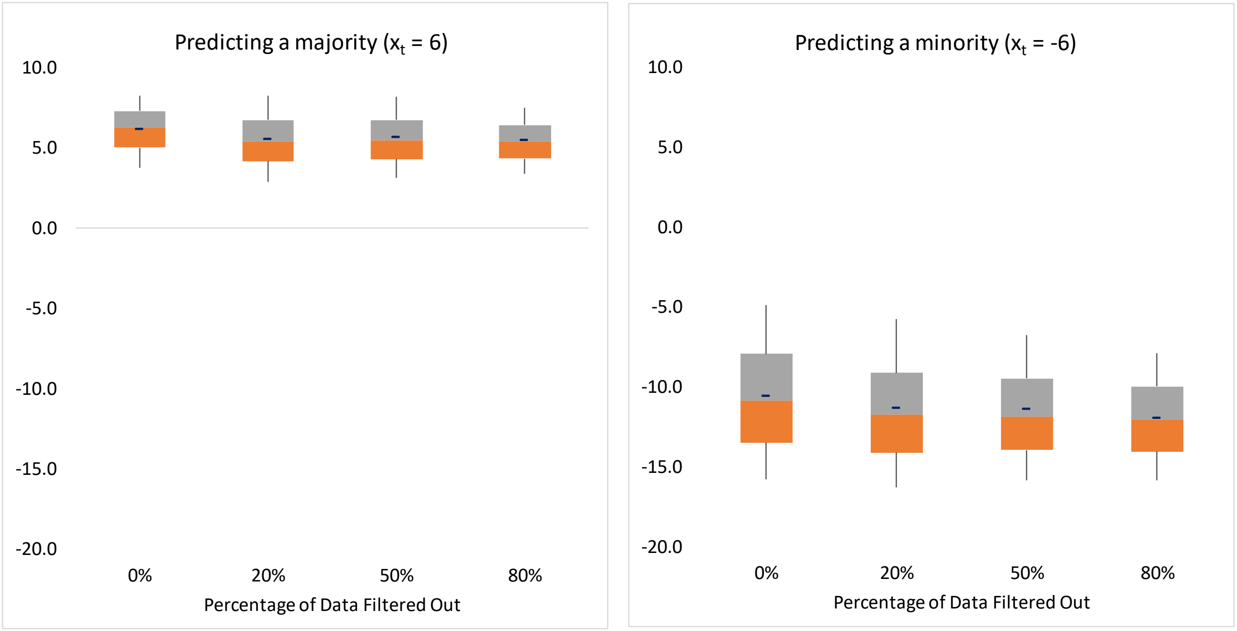


Exhibit 2 shows box plots demarcating the median, 25th, 75th, 10th, and 90th percentiles, along with the actual prediction (average), for predictions of a majority regime circumstance and a minority regime circumstance given several censoring thresholds. This exhibit reveals that the distributions as summarized by these boxplots are not far apart from each other for the majority regime shown in the left panel, which makes sense because the majority regime is highly prevalent and therefore well represented by most subsamples of observations.

The distributions for the minority regime in the right panel are wider because uncertainty in X is magnified due to its greater contribution to the Y outcomes. In addition, the more focused subsamples of relevant observations give notably tighter distributions. This occurs because partial sample regression carefully curates the observations to predict uncommon outcomes, censoring the irrelevant information from the majority regime.

This insight raises the question of how to determine the optimal subsample of relevant observations for a given prediction task, which leads us to the broader formulation of relevance-based prediction.

Relevance-Based Prediction

To focus attention on the distribution of $\hat{y}_{t,solo(i)}$ we illustrated it using partial sample regression, which is a simplified and less refined version of relevance-based prediction (RBP). RBP extends partial sample regression in two important ways: it applies a principled way for determining the optimal subsample of relevant observations for each prediction task based on a statistic called fit, and it incorporates a prediction grid to account for the codependence of observations and predictive variables on the specific circumstances of each individual prediction task.

Fit

Fit quantifies the prevalence of useful patterns in a dataset, which provides a principled way to evaluate the relative efficacy of alternative calibrations for each prediction task. Additionally, fit reveals how much confidence we should have in a specific prediction task, separately from the confidence we have in the overall prediction routine.

Consider a pair of observations that are used to form a prediction. Each observation has a weight and an outcome. We are interested in the alignment of the weights of the two observations with their outcomes. We first standardize them by subtracting the average value and dividing this difference by standard deviation – in essence, converting them to z-scores. We then measure their alignment by taking the product of these standardized values. If the product is positive, their relevance is aligned with their outcomes, and the larger the product, the stronger the alignment. We perform this calculation for every pair of observations in our sample. We should also note that all the formulas we have thus far considered for weights rely only on relevance, which in turn relies only on the x_is, the x_t, and the $\bar{x}$. They do not use any of the information from outcomes. To determine fit, however, we must consider outcomes (the y_is).

$$fit_t = \frac{1}{(N-1)^2} \sum_i \sum_j z_{w_{it}} z_{w_{jt}} z_{y_i} z_{y_j} \tag{19}$$

Equation 20 intuitively describes fit as the squared correlation of relevance weights and outcomes, which conceptually matches the notion of the conventional R-squared statistic.

$$fit_t = \rho(w_t, y)^2 \tag{20}$$

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 observations that are more relevant, we should expect the fit of the subsample to increase, but we should also expect more noise as we shrink the number of observations. The fit across pairs of all observations in the full sample implicitly captures this tradeoff between subsample fit and noise by overweighting observations that are more relevant and underweighting observations that are less relevant.

Like relevance, fit is not arbitrary. In the case of linear regression analysis with $n = N$, the informativeness-weighted average fit across all prediction tasks in the observed sample equals R-squared.[7]

$$R^2 = \frac{1}{N-1}\sum_{t=1}^{N} info(x_t, \bar{x})fit_t \tag{21}$$

Censoring observations that fall below a relevance threshold is more effective to the extent there is asymmetry between the fit of the weights formed from the retained subsample of observations and the fit of the weights formed from the complementary set of censored observations. We measure asymmetry between the fit of the retained and censored subsamples as shown by equation 22. The $(+)$ superscript designates weights formed from the retained observations while the $(-)$ superscript designates weights formed from the censored observations. Asymmetry recognizes the benefit of censoring non-relevant observations that contradict the predictive relationships that exist among the relevant observations. This assessment also inherently considers the relative sample sizes of the two subsamples.

$$asymmetry_t = \frac{1}{2}\left(\rho\left(w_t^{(+)}, y\right) - \rho\left(w_t^{(-)}, y\right)\right)^2 \tag{22}$$

To calculate adjusted fit, we add asymmetry to fit and multiply this sum by K, the number of predictive variables included in the prediction, as shown by Equation 23. Multiplication by the number of predictive variables allows us to compare predictions based on different numbers of predictive variables. Adjusted fit recognizes that we are more likely to observe a spurious relationship from prediction weights based on just one or a few variables than we are based on a collection of many variables.

$$adjusted\ fit_t = K(fit_t + asymmetry_t) \tag{23}$$

Grid Prediction

Grid prediction employs a grid in which the columns represent different combinations of predictive variables, and the rows represent subsamples of observations determined by different relevance thresholds. Each cell contains a prediction and an associated adjusted fit. The assessment of reliability using adjusted fit occurs before the prediction is rendered and the subsequent outcome is known. Grid prediction forms a composite prediction as a reliability-weighted average of the predictions from all possible calibrations. Equation 24 defines reliability weights, ψ_θ, as the adjusted fit for a parameter calibration, θ, divided by the sum of all adjusted fits across all parameter calibrations.

$$\psi_\theta = \frac{adjusted\ fit_\theta}{\sum_{\tilde{\theta}} adjusted\ fit_{\tilde{\theta}}} \tag{24}$$

Equation 25 describes the composite prediction.

$$\hat{y}_{t,grid} = \sum_\theta \psi_\theta \hat{y}_{t,\theta} \tag{25}$$

Exhibits 3 and 4 illustrate how RBP forms a prediction. Exhibit 3 shows how we compute the prediction for a single cell in the prediction grid. It includes hypothetical values for the X and Y variables. The panel on the right gives values for the similarity and informativeness of prior observations and the informativeness of the observations for the current prediction task. It also shows the relevance of each prior observation and the observation's relevance weight.

Exhibit 3: Single Cell Prediction

Variables	Y	X_1	X_2	X_3	X_4	X_5	X_6	Similarity	Info$_i$	Info$_t$	Relevance	Weight
Prediction Task t	?	2.78	8.75	0.28	0.61	0.31	0.58					
Observation 1	20.67	3.13	10.21	0.29	0.00	0.47	0.53	-4.30	12.13	11.96	7.75	4.9%
Observation 2	6.30	4.14	12.24	0.21	0.60	0.29	0.48	-4.06	2.99	11.96	3.41	2.0%
Observation 3	5.19	1.99	9.78	0.16	0.52	0.10	0.48	-7.36	2.43	11.96	-0.17	-0.4%
Observation 4	10.41	3.21	13.47	0.26	0.34	0.48	0.54	-3.41	3.94	11.96	4.54	2.7%
•	•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•	•	•
Observation n	4.49	4.14	3.14	0.23	0.31	0.22	0.37	-7.36	2.75	11.96	-0.01	-0.4%
			Prediction						19.40			
			Adjusted Fit:						2.32			

Exhibit 4 gives a visual representation of grid prediction. The columns represent different subsets of variables, and the rows represent different subsamples of observations as determined by different relevance thresholds. Each cell represents a calibration θ; that is, a unique combination of predictive variables and observations. In practice, we would consider all 63 combinations of six variables, but for illustrative purposes we show only six columns in Exhibit 4. The first values shown in the cells are the calibration-specific predictions $\hat{y}_t$ for a given prediction task t. The second values are the weights ψ_θ we apply to the calibration-specific predictions to form the composite prediction. The values in the grid are specific to each prediction task. This illustration gives a composite prediction of 16.30 (15.7 x 1.72% + 15.7 x 1.15% + 10.1 x 0.24% + . . . + 9.3 x 0.04%).

Exhibit 4: Grid Prediction – Illustrative Example

Variable Combinations

Observation Censoring Threshold	$X_1 X_2 X_3 X_4 X_5 X_6$		$X_1 X_2 X_3 X_4$		$X_1 X_3 X_4$		$X_2 X_5 X_6$		$X_3 X_6$		X_2		X_6	
0.0	15.7	1.72%	15.7	1.15%	10.1	0.24%	15.3	1.37%	10.9	0.54%	15.3	0.47%	7.4	0.06%
0.1	16.4	2.02%	16.7	1.39%	10.4	0.23%	15.4	1.88%	12.5	0.73%	15.5	0.50%	7.7	0.04%
0.2	17.5	2.20%	17.4	1.43%	10.3	0.18%	15.4	1.91%	12.6	0.64%	15.5	0.44%	7.9	0.05%
0.3	17.8	2.17%	17.7	1.43%	10.5	0.20%	15.5	2.24%	12.6	0.62%	15.5	0.42%	7.9	0.05%
0.4	18.2	2.29%	18.0	1.50%	10.6	0.22%	15.4	2.18%	12.7	0.65%	15.5	0.41%	8.1	0.07%
0.5	18.6	2.50%	18.2	1.58%	10.7	0.25%	14.3	2.50%	12.8	0.70%	15.3	0.41%	8.1	0.06%
0.6	18.7	2.47%	18.4	1.61%	10.7	0.23%	15.4	1.21%	13.1	0.73%	15.4	0.42%	8.8	0.10%
0.7	19.0	2.47%	18.8	1.63%	10.7	0.19%	15.4	2.20%	12.9	0.62%	15.4	0.41%	8.7	0.07%
0.8	**19.4**	**2.32%**	19.1	1.50%	11.5	0.20%	15.3	2.04%	13.7	0.57%	15.5	0.37%	8.6	0.04%
0.9	19.5	1.26%	18.8	0.81%	12.9	0.22%	15.5	1.73%	14.0	0.32%	15.3	0.25%	9.3	0.04%

Composite Prediction : 16.30

Note that each cell's prediction is a linear function of observations, and the grid prediction is a linear function of each cell's prediction. Therefore, we can express the grid prediction in terms of composite weights applied to each observation, as shown by Equation 26. Composite weights are important because they preserve the transparency of each observation's contribution to the current prediction task, and they allow us to calculate fit from composite weights as a final gauge of the grid prediction's reliability.

$$w_{it,grid} = \sum_{\theta} \psi_{\theta} w_{it,\theta} \qquad (26)$$

Illustration of distributions of solo predictions from the prediction grid

As an illustration, we now apply grid prediction to forecast the future one-year change in the effective federal funds rate. We use two predictive variables:

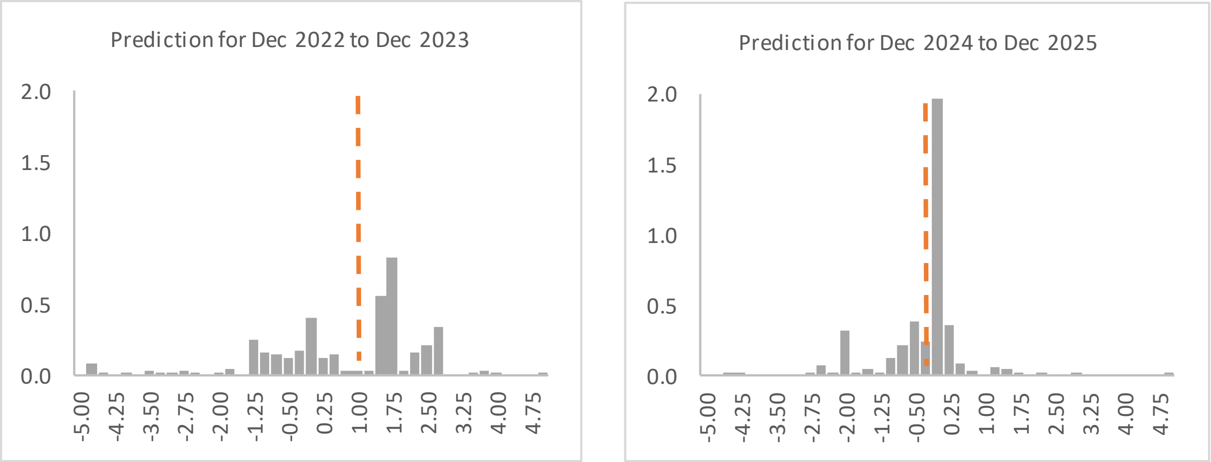
- X1: Employment, measured as the annual change in non-farm payrolls
- X2: Inflation, measured as the annual change in personal consumption expenditures excluding food and energy

We consider all three combinations of the predictive variables: X1 and X2, X1 by itself, and X2 by itself. We consider four subsamples of observations in the grid based on censoring thresholds of 0% (full sample), 20%, 50% and 80%. For each prediction, we filter on both relevance and similarity.[8] Therefore, the prediction grid has 24 cells (three variable combinations times four subsamples times two censoring thresholds). For our training data, we use non-overlapping calendar year observations from 1961 to 2023.

We highlight two predictions: the one-year rate change for 2023 and the one-year rate change for 2025. For each prediction, our inputs reflect the employment and inflation values for the prior year. For 2023, RBP predicted that the effective federal funds rate would increase 90 basis points compared to an actual increase of 120 basis points. For the calendar year that will end in December 2025, RBP predicted the effective federal funds rate would decrease 30 basis points.

The composite predictions that are given by the grid are fit-weighted averages of the individual cell predictions. And each cell's prediction is a weighted average of the solo predictions from the individual observations that are used by each cell. We can therefore construct a distribution of the information given by pooling the solo predictions across all the cells in the grid, which is what we show in Exhibit 5. The dotted lines in Exhibit 5 show the aggregate predictions which are the means of each distribution.

Exhibit 5: Distribution of Solo Predictions from Grid Prediction


Prediction for Dec 2022 to Dec 2023

Prediction for Dec 2024 to Dec 2025

Exhibit 5 demonstrates how the grid preserves and assembles the information from each observation across all the grid cells to form composite distributions of all the information used to form each prediction. It also reveals that these distributions differ from prediction to prediction and that they are distinctly different from classical distributions based on a model's summary statistics. From first principles, we should expect the distributions of information given by the grid to have greater resolution than distributions composed from a single combination of variables and observations, because the grid distributions capture a much broader set of distinct solo predictions based on different combinations of variables.

Summary

We first introduced the notion of partial sample regression, which is a model-free prediction technique that forms a prediction as a weighted average of observed outcomes in which the weights are based on a precise and theoretically justified statistic called relevance. We explained that the premise of partial sample regression is to form a prediction from a subsample of relevant observations, which in most cases gives a more reliable prediction than the full sample of observations.

We then showed how the model-free nature of partial sample regression enables us to form single-observation predictions called solo predictions. These solo predictions are the building blocks that aggregate to the prediction given by all the observations that are used by partial sample regression to form the prediction.

We evaluated the distribution of solo predictions on simulated distributions, assuming that both the observations and errors come from a standard normal distribution, for a special case of partial sample regression that uses the full sample of observations. This special case gives the same prediction as linear regression analysis, which enabled us to compare our information distributions with the classical distribution that comes from the stylized assumptions of linear regression analysis. This comparison showed how the distributions of information differ from each other and from the classical distribution of possible values.

We then evaluated distributions of solo predictions in which the observations were drawn from a bimodal distribution composed of a majority regime and a minority regime. We summarized these distributions with boxplots. We showed that the boxplots were relatively similar for predictions of outcomes associated with the majority regime but different for outcomes associated with the minority regime. This distinction highlighted the fact that partial sample regression carefully curates the observations to predict unusual outcomes.

We next introduced RBP, which depends crucially on fit. Fit quantifies the prevalence of useful patterns in a dataset for forming a prediction. We showed how fit gives advance guidance of a prediction's reliability, which enables us to identify the best subsample of observations for a given prediction task. We then extended this concept to the choice of predictive variables. We discussed how the choice of both observations and predictive variables is codependent on the unique circumstances of each prediction task. We described the prediction grid which considers a vast number of combinations of observations and predictive variables. The columns of the grid represent different combinations of predictive variables, and the rows represent different subsamples of observations based on different censoring thresholds. Each cell in the grid has a prediction and an associated fit. The grid forms a composite prediction as a fit weighted average of the predictions across all the grid cells. It, therefore, produces a composite prediction that is diversified across many calibrations but in a way that bends toward those combinations of observations and predictive variables that give more reliable predictions.

We then presented a simple illustration of grid prediction in which we predicted annual changes in the effective federal funds rate. We demonstrated that because each cell's prediction is a weighted average of solo predictions and because the grid's composite prediction is a weighted average of all the cell predictions, we can construct distributions of all the information in the grid that is used to form a composite prediction.

Finally, we discussed how our information-based measure of prediction uncertainty relates to model-based assessments of confidence. Model-derived distributions tell us about the range of prediction errors we should expect and the range of estimated predictions we should expect if we repeated the prediction with different data. We should have less confidence in a prediction if these distributions are wide than if they are tight. These model-based methods, however, assume that the model is correctly specified and that the predictive variables are measured without error. By contrast, RBP assesses a prediction's reliability based on the consistency of the information that is used to form each individual prediction. RBP tells us that we should have more trust in a prediction if it is formed from information that is mutually supportive than if it is formed from information that is mutually contradictory.

In conclusion, we do not assert that RBP's measure of prediction uncertainty is necessarily better than a model-based assessment of confidence. It does, however, give a completely novel perspective about a prediction's reliability, and it does so in a way that does not assume perfect accuracy of the model and the predictive variable observations. We therefore recommend it as a valuable complement to conventional measures of confidence.

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.

References

Czasonis, Megan, Mark Kritzman, and David Turkington. 2022a. "Relevance." *The Journal of Investment Management*, 20 (1).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2022b. *Prediction Revisited: The Importance of Observation*. Hoboken, New Jersey: John S. Wiley & Sons.

Czasonis, Megan, Mark Kritzman, and David Turkington. 2023. "Relevance-Based Prediction: A Transparent and Adaptive Alternative to Machine Learning." *The Journal of Financial Data Science*, 5 (1).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2025a. "The Virtue of Transparency: How to Maximize the Utility of Data Without Overfitting." *The Journal of Financial Data Science*, 7 (2).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2025b. "A Transparent Alternative to Neural Networks with an Application to Predicting Volatility." *Journal of Investment Management*, 23 (3).

Czasonis, Megan, Mark Kritzman, and David Turkington. 2025c. "Prediction with Incomplete Information." *The Journal of Financial Data Science*, 7 (3).

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

Appendix: A Note about the Distribution of Solo Predictions

In the following analysis, we derive an expression for the average variance of the RBP solo prediction distributions across all prediction tasks in the case of a linear regression and we relate it to the classical variance of the error term from a linear model.

Let us assume without loss of generality that $\bar{y} = 0$. The following expressions give the solo predictions and their associated weights.

$$\hat{y}_{t,solo(i)} = \frac{info(x_t)}{r_{it}} y_i \tag{A1}$$

$$\xi_i = \delta(r_{it}) \frac{r_{it}^2}{\sum_j \delta(r_{jt}) r_{jt}^2} \tag{A2}$$

Let us calculate the mean of the distribution as a probability-weighted sum:

$$\mu_{t,solo} = \sum_i \xi_i \hat{y}_{t,solo(i)} \tag{A3}$$

$$\mu_{t,solo} = \sum_i \delta(r_{it}) \frac{r_{it}^2}{\sum_j \delta(r_{jt}) r_{jt}^2} \frac{info(x_t)}{r_{it}} y_i \tag{A4}$$

$$\mu_{t,solo} = \frac{info(x_t)}{\sum_j \delta(r_{jt}) r_{jt}^2} \sum_i \delta(r_{it}) r_{it} y_i \tag{A5}$$

$$\mu_{t,solo} = \hat{y}_t \tag{A6}$$

Now, calculate the variance of the distribution:

$$\sigma_{t,solo}^2 = \sum_i \xi_i \left(\hat{y}_{t,solo(i)} - \hat{y}_t\right)^2 \tag{A7}$$

$$\sigma_{t,solo}^2 = \sum_i \xi_i \left(\frac{info(x_t) y_i}{r_{it}} - \hat{y}_t\right)^2 \tag{A8}$$

$$\sigma^2_{t,solo} = \sum_i \delta(r_{it}) \frac{r_{it}^2}{\sum_j \delta(r_{jt}) r_{jt}^2} \left(\frac{info(x_t)^2 y_i^2}{r_{it}^2} + \hat{y}_t^2 - \frac{2info(x_t)\hat{y}_t y_i}{r_{it}} \right) \tag{A9}$$

$$\sigma^2_{t,solo} = \frac{1}{\sum_j \delta(r_{jt}) r_{jt}^2} \sum_i \delta(r_{it})(info(x_t)^2 y_i^2 + r_{it}^2 \hat{y}_t^2 - 2info(x_t) r_{it} \hat{y}_t y_i) \tag{A10}$$

Consider the case of a linear regression, where $\delta(r_{it}) = 1$ for all i. Now, we have:

$$\sigma^2_{t,solo} = \frac{1}{(N-1)info(x_t)} \sum_i \delta(r_{it})(info(x_t)^2 y_i^2 + r_{it}^2 \hat{y}_t^2 - 2info(x_t) r_{it} \hat{y}_t y_i) \tag{A11}$$

$$\sigma^2_{t,solo} = \frac{1}{N-1} \sum_i (info(x_t) y_i^2 + r_{it}^2 \hat{y}_t^2 - 2 r_{it} \hat{y}_t y_i) \tag{A12}$$

For linear regression the definition of $\hat{y}_t$ is:

$$\hat{y}_{t,linear} = \frac{1}{N-1} \sum_i r_{it} y_i \tag{A13}$$

Substituting this in the previous equation gives:

$$\sigma^2_{t,solo,linear} = \frac{1}{N-1} \sum_i \left(info(x_t) y_i^2 + r_{it}^2 \left(\frac{1}{N-1} \sum_j r_{jt} y_j \right)^2 - \frac{2r_{it}}{N-1} \sum_i r_{it} y_j y_i \right) \tag{A14}$$

$$\sigma^2_{t,solo,linear} = \frac{1}{N-1} \sum_i \left(info(x_t) y_i^2 + \frac{r_{it}^2}{(N-1)^2} \sum_j \sum_k r_{jt} r_{kt} y_j y_k - \frac{2r_{it}}{N-1} \sum_j r_{jt} y_j y_i \right) \tag{A15}$$

$$\sigma^2_{t,solo,linear} = \frac{info(x_t)}{N-1} \sum_i y_i^2 + \frac{1}{(N-1)^3} \sum_i r_{it}^2 \sum_j \sum_k r_{jt} r_{kt} y_j y_k - \frac{2}{(N-1)^2} \sum_i \sum_j r_{jt} r_{it} y_j y_i \tag{A16}$$

$$\sigma^2_{t,solo,linear} = \frac{info(x_t)}{N-1} \sum_i y_i^2 + \frac{info(x_t)}{(N-1)^2} \sum_j \sum_k r_{jt} r_{kt} y_j y_k - \frac{2}{(N-1)^2} \sum_i \sum_j r_{jt} r_{it} y_j y_i \tag{A17}$$

Now, let us take the average of $\sigma^2_{t,solo,linear}$ across all prediction tasks t in the observed sample of N (dividing by $N - 1$ rather than N, which makes little difference):

$$\frac{1}{N-1} \sum_t \sigma^2_{t,solo,linear} = \frac{K}{N} \sum_i y_i^2 + \frac{K}{(N-1)^2} \sum_i \sum_j r_{ij} y_i y_j - \frac{2}{N(N-1)^2} \sum_i \sum_j r_{ij} y_i y_j \tag{A18}$$

$$\frac{1}{N-1}\sum_t \sigma_{t,solo,linear}^2 = \frac{K}{N}\sum_i y_i^2 + \frac{K-2}{(N-1)^2}\sum_i \sum_j r_{ij} y_i y_j \qquad \text{(A19)}$$

Equation A19 characterizes the average variance of the solo distribution. Notably, it increases with the number of predictive variables, K. When there are more predictive variables, the noise in those variables creates more dispersion in the solo predictions, reflecting greater uncertainty in the information that underlies a linear regression prediction. For a sufficiently large sample, adding a pure noise variable to linear regression will have little impact on prediction values because the noise cancels out. But this noise variable will still increase the variance of the information that informs the prediction. For RBP grid predictions, the grid assigns less weight to cells that do not exhibit useful patterns for a prediction. As a result of this prediction logic, noise variables may contribute less to information variance than they would in a linear regression context.

Let us now consider how equation A19 relates to the classical range of outcomes for a prediction. For the special case of just one predictive variable ($K = 1$), we have:

$$\frac{1}{N-1}\sum_t \sigma_{t,solo,linear}^2 = \frac{1}{N-1}\sum_i y_i^2 - \frac{1}{(N-1)^2}\sum_i \sum_j r_{ij} y_i y_j \qquad \text{(A20)}$$

$$\frac{1}{N-1}\sum_t \sigma_{t,solo,linear}^2 = \sigma_y^2(1 - R^2) = \sigma_\epsilon^2 \qquad \text{(A21)}$$

Equation A21 reveals why the solo distributions from Exhibit 1 were close to the classical distributions with a variance of σ_ϵ^2, and why one of the distributions was narrower than the classical distribution while the other was wider. On average across all empirical prediction tasks, the solo distributions will converge to a variance of σ_ϵ^2 for the special case where $K = 1$.

[1] The descriptions of partial sample regression and relevance-based prediction throughout this article follow closely language used from Czasonis, Kritzman, and Turkington (2022a, 2022b, 2023, 2025a, 2025b, and 2025c), but they are modified to fit the context of the current discussion.

[2] Shannon showed that information is an inverse logarithmic function of probability, which is a key insight from his comprehensive theory of communication. See Shannon (1948).

[3] See Czasonis, Kritzman, and Turkington (2023) for proof of this result.

[4] See Czasonis, Kritzman, and Turkington (2023) for proof of this result.

[5] See Czasonis, Kritzman, and Turkington (2022b).

[6] Each prediction contains both a signal and noise. As the impact is scaled up so too is the noise, which leads to more uncertainty about the prediction.

[7] See Czasonis, Kritzman, and Turkington (2022b) for proof of this result.

[8] It is often helpful to censor the observations on just similarity as well as relevance, though we always form the prediction as a relevance-based weighted average. Moreover, we need not choose which censoring criterion to use. We let adjusted fit determine the best censoring criterion.