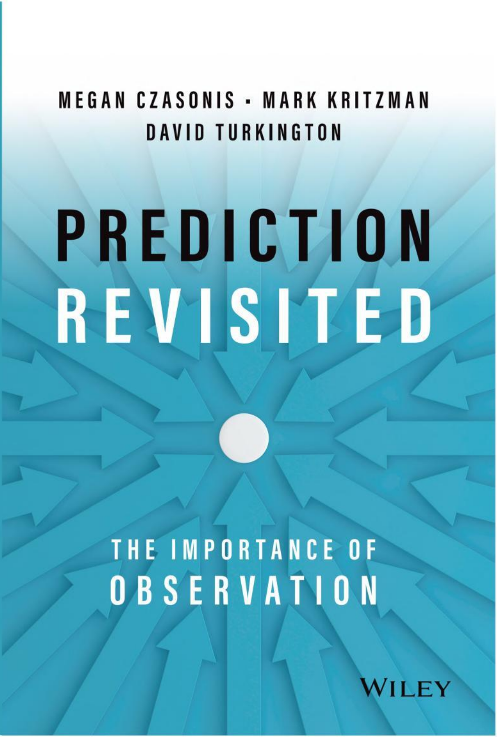
CZASONIS
KRITZMAN

PREDICTION REVISITED

LEY



MEGAN CZASONIS · MARK KRITZMAN
DAVID TURKINGTON

PREDICTION REVISITED

THE IMPORTANCE OF OBSERVATION

WILEY

Classical statistics originated more than two centuries ago to facilitate navigation by predicting the motion of heavenly bodies and to offer guidance for games of chance. These systems obey relatively simple rules. Today, we are challenged to predict outcomes that are driven by far more complex systems, such as the dynamics of social behavior. Researchers have come to recognize that classical statistics cannot accommodate the complexity of social dynamics; they have therefore turned to the emergent field of machine learning. But they also struggle with machine learning algorithms because these algorithms are often opaque and unintuitive, and they lack a theoretical core. This book offers another way forward—a way that is theoretically grounded, transparent, and intuitive.

This path forward requires a new perspective. We must view data as experiences and think of variables as attributes for describing those experiences. And we must recognize that some experiences are more relevant than others. Indeed, determining relevance is the essence of prediction. The authors provide a guided tour of this groundbreaking insight, from its foundations in information theory to its central role in forecasting. They reveal the specific components of relevance and show how to measure them, not only in concept but with mathematical precision.

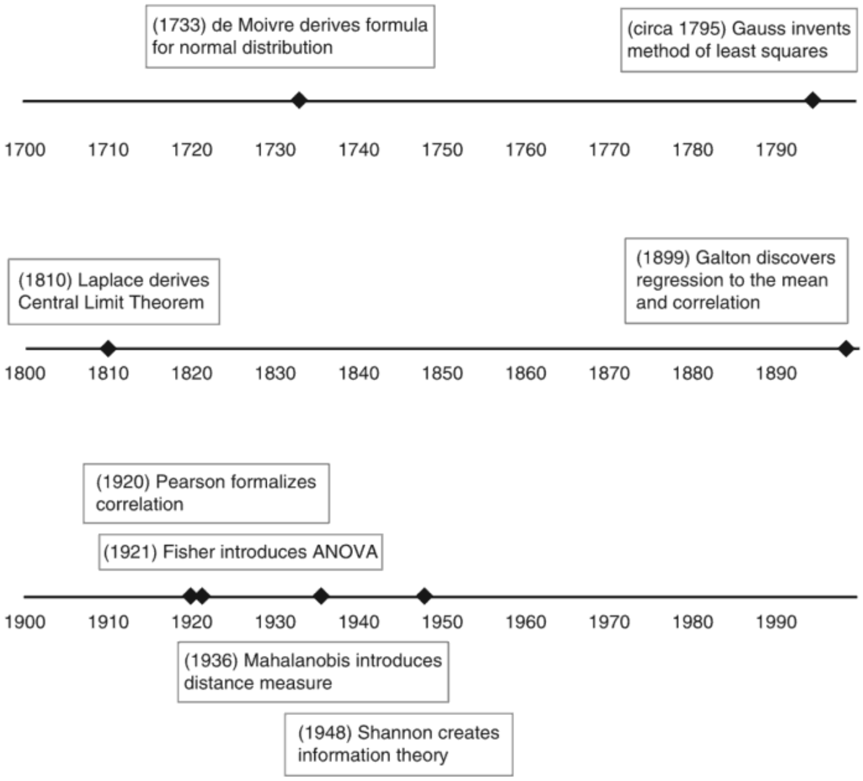
There is a practical reward to this journey. You will learn that the prediction from a linear regression equation is equivalent to a relevance-weighted average of past outcomes. This critical insight enables you to form more reliable predictions from a subset of the most relevant observations, using an approach called partial sample regression. And you will learn how to judge the unique reliability of an individual prediction separately from

(continued on back flap)

Contents

Timeline of Innovations

Relevance is the centerpiece of our approach to prediction. The key concepts that give rise to relevance were introduced over the past three centuries, as illustrated in this timeline. In Chapter 8, we offer more detail about the people who made these groundbreaking discoveries.


(1733) de Moivre derives formula for normal distribution
(circa 1795) Gauss invents method of least squares
1700
1710
1720
1730
1740
1750
1760
1770
1780
1790
(1810) Laplace derives Central Limit Theorem
(1899) Galton discovers regression to the mean and correlation
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
(1920) Pearson formalizes correlation
(1921) Fisher introduces ANOVA
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
(1936) Mahalanobis introduces distance measure
(1948) Shannon creates information theory

Essential Concepts

This book introduces a new approach to prediction, which requires a new vocabulary—not new words, but new interpretations of words that are commonly understood to have other meanings. Therefore, to facilitate a quicker understanding of what awaits you, we define some essential concepts as they are used throughout this book. And rather than follow the convention of presenting them alphabetically, we present them in a sequence that matches the progression of ideas as they unfold in the following pages.

Observation: One element among many that are described by a common set of attributes, distributed across time or space, and which collectively provide guidance about an outcome that has yet to be revealed. Classical statistics often refers to an observation as a multivariate data point.

Attribute: A recorded value that is used individually or alongside other attributes to describe an observation. In classical statistics, attributes are called independent variables.

Outcome: A measurement of interest that is usually observed alongside other attributes, and which one wishes to predict. In classical statistics, outcomes are called dependent variables.

Arithmetic average: A weighted summation of the values of attributes or outcomes that efficiently aggregates the information contained in a sample of observations. Depending on the context and the weights that are used, the result may be interpreted as a typical value or as a prediction of an unknown outcome.

Spread: The pairwise distance between observations of an attribute, measured in units of surprise. We compute this distance as the average of half the squared difference in values across every pair of observations. In classical statistics, the same quantity is usually computed as the average of squared deviations of observations from their mean and is referred to as variance. However, the equivalent evaluation of pairwise spreads reveals why we must divide by $N - 1$ rather than N to obtain an unbiased estimate of a sample's variance; it is because the zero distance of an observation with itself (the diagonal in a matrix of pairs) conveys no information.

Information theory: A unified mathematical theory of communication, created by Claude Shannon, which expresses messages as sequences of 0s and 1s and, based on the inverse relationship of information and probability, prescribes the optimal redundancy of symbols to manage the speed and accuracy of transmission.

Circumstance: A set of attribute values that collectively describes an observation.

Informativeness: A measure of the information conveyed by the circumstances of an observation, based on the inverse relationship of information and probability. For an observation of a single attribute, it is equal to the observed distance from the average, squared. For an observation of two or more uncorrelated attributes, it is equal to the sum of each individual attribute's informativeness. For an observation of two or more correlated attributes—the most general case—it is given by the Mahalanobis distance of the observation from the average of the observations. Informativeness is a component of relevance. It does not depend on the units of measurement.

Co-occurrence: The degree of alignment between two attributes for a single observation. It ranges between -1 and $+1$ and does not depend on the units of measurement.

Correlation: The average co-occurrence of a pair of attributes across all observations, weighted by the informativeness of each observation. In classical statistics, it is known as the Pearson correlation coefficient.

Covariance matrix: A symmetric square matrix of numbers that concisely summarizes the spreads of a set of attributes along with the signs and strengths of their correlation. Each element pertains to a pair of attributes and is equal to their correlation times their respective standard deviations (the square root of variance or spread).

Mahalanobis distance: A standardized measure of distance or surprise for a single observation across many attributes, which incorporates all the information from the covariance matrix. The Mahalanobis distance of a set of attribute values (a circumstance) from the average of the attribute values measures the informativeness of that observation.

Half of the negative of the Mahalanobis distance of one circumstance from another measures the similarity between them.

Similarity: A measure of the closeness between one circumstance and another, based on their attributes. It is equal to the opposite (negative) of half the Mahalanobis distance between the two circumstances. Similarity is a component of relevance.

Relevance: A measure of the importance of an observation to forming a prediction. Its components are the informativeness of past circumstances, the informativeness of current circumstances, and the similarity of past circumstances to current circumstances.

Partial sample regression: A two-step prediction process in which one first identifies a subset of observations that are relevant to the prediction task and, second, forms the prediction as a relevance-weighted average of the historical outcomes in the subset. When the subset from the first step equals the full-sample, this procedure converges to classical linear regression.

Asymmetry: A measure of the extent to which predictions differ when they are formed from a partial sample regression that includes the most relevant observations compared to one that includes the least relevant observations. It is computed as the average dissimilarity of the predictions from these two methods. Equivalently, it may be computed by comparing the respective fits of the most and least relevant subsets of observations to the cross-fit between them. The presence of asymmetry causes partial sample regression predictions to differ from those of classical linear regression. The minimum amount of asymmetry is zero, in which case the predictions from full-sample and partial-sample regression match.

Fit: The average alignment between relevance and outcomes across all observation pairs for a single prediction. It is normalized by the spreads of relevance and outcomes, and while the alignment for one pair of observations may be positive or negative, their average always falls between zero and one. A large value indicates that observations that are similarly relevant have similar outcomes, in which case one should have more confidence in the prediction. A small value indicates that relevance does not line up with the outcomes, in which case one should view the prediction more cautiously.

Bias: The artificial inflation of fit resulting from the inclusion of the alignment of each observation with itself. This bias is addressed by partitioning fit into two components—outlier influence, which is the fit of observations with themselves, and agreement, which is the fit of observations with their peers—and using agreement to give an unbiased measure of fit.

Outlier influence: The fit of observations with themselves. It is always greater than zero, owing to the inherent bias of comparing observations with themselves, and it is larger to the extent that unusual circumstances coincide with unusual outcomes.

Agreement: The fit of observations with their peers. It may be positive, negative, or zero, and is not systematically biased.

Precision: The inverse of the extent to which the randomness of historical observations (often referred to as noise) introduces uncertainty to a prediction.

Focus: The choice to form a prediction from a subset of relevant observations even though the smaller subset may be more sensitive to noise than the full sample of observations, because the consistency of the relevant subset improves confidence in the prediction more than noise undermines confidence.

Reliability: The average fit across a set of prediction tasks, weighted by the informativeness of each prediction circumstance. For a full sample of observations, it may be computed as the average alignment of pairwise relevance and outcomes and is equivalent to the classical R-squared statistic.

Complexity: The presence of nonlinearities or other conditional features that undermine the efficacy of linear prediction models. The conventional approach for addressing complexity is to apply machine learning algorithms, but one must counter the tendency of these algorithms to overfit the data. In addition, it can be difficult to interpret the inner workings of machine learning models. A simpler and more transparent approach to complexity is to filter observations by relevance. The two approaches can also be combined.

quest for intuition. But mostly we are motivated by a stubborn refusal to stop asking the question: Why?

Practitioners have difficult problems to solve and often too little time. Those on the front lines may struggle to absorb everything that technical training has to offer. And there are bound to be many useful ideas, often published in academic articles and books, that are widely available yet seldom used, perhaps because they are new, complex, or just hard to find.

Most of the ideas we present in this book are new to us, meaning that we have never encountered them in school courses or publications. Nor are we aware of their application in practice, even though investors clearly thrive on the quality of their predictions. But we are not so much concerned with precedence as we are with gaining and sharing a better understanding of the process of data-driven prediction. We would, therefore, be pleased to learn of others who have already come to the insights we present in this book, especially if they have advanced them further than we do in this book.

1

Introduction

We rely on experience to shape our view of the unknown, with the notable exception of religion. But for most practical purposes we lean on experience to guide us through an uncertain world. We process experiences both naturally and statistically; however, the way we naturally process experiences often diverges from the methods that classical statistics prescribes. Our purpose in writing this book is to reorient common statistical thinking to accord with our natural instincts.

Let us first consider how we naturally process experience. We record experiences as narratives, and we store these narratives in our memory or in written form. Then when we are called upon to decide under uncertainty, we recall past experiences that resemble present circumstances, and we predict that what will happen now will be like what happened following similar past experiences. Moreover, we instinctively focus more on past experiences that were exceptional rather than ordinary because they reside more prominently in our memory.

Now, consider how classical statistics advises us to process experience. It tells us to record experiences not as narratives, but as data. It suggests that we form decisions from as many observations as we can assemble or from a subset of recent observations, rather than focus on observations that are like current circumstances. And it advises us to view unusual observations with skepticism. To summarize:

Natural Process

- Records experiences as narratives.
- Focuses on experiences that are like current circumstances.
- Focuses on experiences that are unusual.

Classical Statistics

- Record experiences as data.
- Include observations irrespective of their similarity to current circumstances.
- Treat unusual observations with skepticism.

The advantage of the natural process is that it is intuitive and sensible. The advantage of classical statistics is that by recording experiences as data we can analyze experiences more rigorously and efficiently than would be allowed by narratives. Our purpose is to reconcile classical statistics with our natural process in a way that secures the advantages of both approaches.

We accomplish this reconciliation by shifting the focus of prediction away from the selection of variables to the selection of observations. As part of this shift in focus from variables to observations, we discard the term *variable*. Instead, we use the word *attribute* to refer to an independent variable (something we use to predict) and the word *outcome* to refer to a dependent variable (something we want to predict). Our purpose is to induce you to think foremost of experiences, which we refer to as observations, and less so of the attributes and outcomes we use to measure those experiences. This shift in focus from variables to observations does not mean we undervalue the importance of choosing the right variables. We accept its importance. We contend, however, that the choice of variables has commanded disproportionately more attention than the choice of observations. We hope to show that by choosing observations as carefully as we choose variables, we can use data to greater effect.

Relevance

The underlying premise of this book is that some observations are relevant, and some are not—a distinction that we argue receives far less attention than it deserves. Moreover, of those that are relevant, some observations are more relevant than others. By separating relevant observations from those that are not, and by measuring the comparative relevance of observations, we can use data more effectively to guide our decisions. As suggested by our discussion thus far, relevance has two components: similarity and unusualness. We formally refer to the latter as informativeness. This component of relevance is less intuitive than similarity but is perhaps more foundational to our notion of relevance; therefore, we tackle it first.

Informativeness

Informativeness is related to information theory, the creation of Claude Shannon, arguably the greatest genius of the twentieth century.[1] As we discuss in Chapter 2, information theory posits that information is inversely related to probability. In other words, observations that are unusual contain more information than those that are common. We could stop here and rest on Shannon's formidable reputation to validate our inclusion of informativeness as one of the two components of relevance. But it never hurts to appeal to intuition. Therefore, let us consider the following example.

Suppose we would like to measure the relationship between the performance of the stock market and a collection of economic attributes (think variables) such as inflation, interest rates, energy prices, and economic growth. Our initial thought might be to examine how stock returns covary with changes in these attributes. If these economic attributes behaved in an ordinary way, it would be difficult to tell which of the attributes were driving stock returns or even if the performance of the stock market was instead responding to hidden forces. However, if one of the attributes behaved in an unusual way, and the stock market return we observed was also notable, we might suspect that these two occurrences are linked by more than mere coincidence. It could be evidence of a fundamental relationship. We provide a more formal explanation of informativeness in Chapter 2, but for now let us move on to similarity.

[1] Some might prefer to assign this accolade to Albert Einstein, but why quibble? Both were pretty smart.

2

Observing Information

Our journey into data-driven prediction begins with some basic ideas. In this chapter, we set forth principles which may at first seem obvious, but which, upon deeper inspection, have profound implications. These ideas lay the foundation for everything that follows.

Observing Information Conceptually

Whenever we approach a new dataset the first order of business is to get our bearings. We have before us a series of observations, each of which is described by a set of attributes. The observations could be of people, described by attributes like age, health, education, salary, and place of residence. They could be times at-bat for a major league baseball player, with attributes of runs-batted-in, home runs, walks, strikeouts, weather conditions, and where the game took place. Or the observations could be periods of economic performance measured by attributes such as growth in output, inflation, interest rates, unemployment, stock market returns, and perhaps the political parties in power at the time. What matters is that we have a set of observations characterized by a consistent collection of attributes. A conventional statistics approach would have us focus on these attributes and refer to them as variables, but as we stated earlier, we ask that you indulge us as we focus mainly on how we observe these attributes.

of what he called a quincunx.[1] It generates a histogram of the normal probability curve by allowing pellets to cascade down a lattice of pins, falling randomly to the left or right of each pin. Few fall to the far left or to the far right; most cluster near the middle. Galton famously used this contraption to show, right before one's eyes, that the normal curve arises time and again from nothing more than the aggregation of the simplest random outcomes. With a bit more patience and even less technology, you can observe the same thing by tallying the number of heads you get from a sequence of coin flips. It is plausible that this is what the French mathematician Abraham de Moivre had in mind as he worked on his book *The Doctrine of Chances* in London, after having fled religious persecution and imprisonment in France. In 1733 de Moivre published his finding that for the sum of many binomial outcomes, such as coin flips, the occurrence of large deviations from average decays exponentially as a function of the distance squared. Though this discovery was a triumph, it was not well-known nor widely applied until much later.

In the 1770s, Pierre-Simon Laplace confronted similar ideas but in a more general context than just the coin flip equivalent. Though aware of de Moivre's work, Laplace appears to have been somewhat tormented for decades by the question of what curve best reflects the rarity of extreme events. After multiple false starts, he presented the essence of the Central Limit Theorem in 1810. It was a profound breakthrough.

Meanwhile, in 1805, another French mathematician named Adrien-Marie Legendre was actively promoting his method of least squares for solving the day's most pressing problems in astronomy. Departing from tradition, he blended noise-prone measurements together in what would later be seen as a form of linear regression analysis. The widespread attention he gained led to a conflict with Carl Friedrich Gauss, who argued he had invented the same method a decade earlier, although he did not publish his result at the time. Nonetheless, Gauss eventually outdid Legendre by connecting the method of least squares to the normal distribution in his 1809 book about the orbits of heavenly bodies around the sun.

Gauss's reference to the normal distribution was a minor side note at the end of his book. He was fond of using the arithmetic average of observations to mitigate measurement errors, and he asked himself:

[1] This curious term derives originally from the Roman word for five-twelfths, often depicted on currency as five dots. The term came to mean an arrangement of five dots in a lattice, such as on the side of a die, and eventually to describe such a lattice pattern in general.