

Profiling Consumer Safety Violations and Violators*

Richard Berk

Department of Statistics

University of Pennsylvania

May 14, 2009

Introduction

Responding to reports from China of infant formula contaminated with melamine, the FDA issued in September of 2008 a Health Information Advisory stating that there was no known threat of contamination in infant formula manufactured by companies complying with U.S. regulations. The FDA also stated that no companies selling infant formula in the United State were using milk-based ingredients from China but that nevertheless, there would be ongoing testing of food items imported from China that could contain a significant amounts of milk or milk proteins.

*Thanks go to Cary Coglianese and Adam Finkel for very helpful comments on an earlier draft of this paper.

The melamine incident is one illustration of the systematic oversight of consumer products that can be the responsibility of any number of federal, state and local agencies. Such oversight requires resources, and with an almost limitless number of venues and actors, as noted in chapter 1, those resources are guaranteed to be insufficient. A sensible response is to be smart about how oversight resources are allocated. And being smart can depend substantially on the collection and analysis of data that allow administrators to anticipate where the impact of their oversight activities will be most cost-effective. This can be very demanding because to anticipate impact, one must look into the future.

Forecasts has long played an important role in a variety of decision-making, especially in criminal justice. They have been used to inform the need to build prisons, to allocate police personnel to certain neighborhoods, and to anticipate the amount and mix of crime. But perhaps the most common use of forecasts has been for making decisions about individuals convicted of crimes. Forecasts of future behavior can at least implicitly affect sentencing decisions by judges (U.S. Sentencing Commission, 2006), housing decisions made by prison officials (Berk et al., 2006), release decisions made by parole boards (Glaser, 1987), and supervisory decisions made by probation or parole officers (Berk et al., 2009). The overall success of these forecasting exercises is certainly open to debate, but there is no doubt that some recent statistical developments have begun to produce predictions of useful accuracy and that still better forecasting skill is on the horizon (Berk, 2008a).

The goal of this paper is to provide a translation from forecasting a wide range of criminal behavior to forecasting economic behavior, whether financially motivated or the results of

innocent error, that puts consumers at risk. The justification for both applications is much the same. Forecasts can help in the prevention of misconduct, in determining appropriate punishment for misconduct, in the allocation of scarce enforcement resources, and in the selection of effective remedies for the consequences of misconduct. Forecasts can also be used in a forensic manner to identify malfeasance that has already occurred. How these applications play out for consumer safety oversight will be considered in the pages ahead. We will see that many important lessons from forecasting criminal behavior carry over well.

Examples of Forecasts and Their Roles in Law Enforcement

The forecasting addressed in the pages ahead is related to other empirical activities. In particular, the kinds of forecasting to be discussed are sometimes characterized as “profiling.” From a set of attributes that some might call a profile, projections are made. For example, advertisers construct consumer profiles to forecast what certain individuals might be convinced to buy. Colleges construct applicant profiles to forecast who will do well academically. Medical researchers construct profiles of patients and their symptoms to forecast treatments likely to work well. More immediately relevant here, profiling is common in law enforcement. Parole boards, for instance, construct profiles of prison inmates using information about their past crimes and behavior behind bars to forecast how they will fare under supervision in their communities. Perhaps the best known instances in law enforcement involve profiles of

crimes and criminals used determine which pedestrians and drivers to stop for questioning. In theory at least, police officers are trying to apprehend individuals who have committed crimes or who may soon do so.

In short, profiling implies forecasting. Behavior is imputed that has not been observed, either because it occurred elsewhere or because it has not yet occurred. One important implication is that the construction and use of profiles should conform to the same statistical principles as conventional forecasting. This emphasis on forecasting also helps to distinguish the procedures discussed below from “risk assessments” sometimes undertaken in criminal justice settings or when the safety of consumer products is a concern. For example, risk assessments are often undertaken for households in which domestic violence is suspected (Dutton and Kropp, 2000) and also for a wide variety of food product contaminants, as described in the Bier and Zach chapter in this volume. The essential goal of a risk assessment is to establish cause and effect relationships between causal variables and some outcome of interest. One must bring to bear, therefore, all of the usual concerns about causal inference (Holland, 1986; Rubin, 1986). For example, careful thought must be given to defining the appropriate counterfactual conditions for each causal variable. Thus, for foods stored at a temperature that risks bacterial contamination, what is the appropriate temperature to which the suspect temperature should be compared?

When accurate forecasts are the goal, cause and effect is formally unnecessary. Consequently, the conceptual and statistical tools can differ dramatically (Granger, 1989). In particular, the primary criterion for evaluating forecasts is their forecasting accuracy. Whether

the variables used to make the forecasts have any sensible causal interpretation is at best a secondary concern. It follows that addressing uncertainty centers on the values forecasted, not the role of individual predictors. In practice, however, one can find forecasts made implicitly from risk assessments and causal interpretations given to forecasting predictors. Usually, crossing such boundaries is a bad idea. There is a mismatch between the tools being used and the goals of the research. Misleading conclusions can follow.

We can now turn to some criminal justice applications. They provide useful lessons for forecasting consumer product safety because the issues are often very similar. Other criminal justice examples could have been used to convey much the same material, but the ones discussed were chosen because they could be written from first hand and detailed knowledge.

Domestic Violence

One of the key decisions that police officers make in domestic violence incidents is whether to arrest the alleged perpetrator or just try to restore order. An important factor in that decision is a forecast of whether in the near future police will again be dispatched to that household. If the chances they will have to return are good, police might quite properly decide to employ a more intrusive intervention. Police officers on domestic violence calls routinely make such forecasts in an informal manner using their experience with similar cases in the past. But it is possible to do better.

With a data set of several hundred domestic violence cases, one can in principle link

a number of readily available predictors to the chances of another incident involving the same individuals and to the seriousness of that repeat incident. For example, past police dispatches to the same household are, not surprisingly, a good predictor of more dispatches in the future. Likewise, injuries that in the past have required medical attention anticipate subsequent life threatening violence. When police later confront domestic violence cases, they can exploit such information to make formal forecasts that in turn can be used to inform decisions about what should be done. Just such a procedure was developed for the Los Angeles County Sheriff's Department (Berk et al., 2005). The strong predictors were organized into a check list. The more checks, the greater the probability of a repeat call. This information helped to inform the actions of police officers.

This is an example of forecasts used to prevent unwanted behavior in the future. When the the probability of unwanted behavior is sufficiently high, police can choose to intervene in a more intrusive manner. They can try to prevent undesirable outcomes by incapacitating the likely perpetrator through an arrest and incarceration, by trying to aggressively mediate the dispute, ordering the perpetrator from the premises for several hours as a "cooling down period," or by threatening unpleasant sanctions in the future. And some of these options (and others) can be usefully combined. The general point is that law enforcement resources can be more effectively allocated if police officers have the ability to make more reliable predictions of future behavior.

Probation and Parole

Approximately 50,000 individuals a year are supervised by the Philadelphia Adult Probation and Parole Department (APPD). The supervision is costly and most would argue not effectively deployed. About a quarter of the probationers or parolees fail; they are sent to prison or returned to prison respectively. One possible explanation for the high rate of failure is that supervisory resources are not being allocated efficiently. For example, individual probationers and parolees can differ dramatically in the probability they will re-offend and in how serious the new offenses are likely to be. If one could forecast at a useful level of accuracy the risk to public safety an individual poses, it might then be possible to reallocate resources away from low risk individuals toward high risk individuals.

Just such an exercise is well under way. Recent research has shown that it is possible using information routinely available at intake to forecast with considerable skill which individuals under supervision by the APPD will commit a homicide or attempted homicide within two years after supervision begins (Berk et al., 2009a). That work has now been expanded. Using similar statistical procedure and data, forecasts of at least equal accuracy have been constructed for individuals at the other extreme: parolees and probationers who pose no serious risk to public safety (Berk et al., 2009b). This has, in turn, led to a third forecasting exercise in which there are three outcome classes: (1) a new arrest for a serious crime, (2) a new arrest for any other crime, and (3) no new arrest. Once again, forecasting accuracy has been satisfactory and supervisory resources are now being reorganized in response to need those forecasts imply.

Incarceration

Yet another forecasting study within the same spirit was conducted for the California Department of Corrections (Berk et al., 2006). The outcome to be forecasted was serious misconduct in prison. “Serious misconduct” was defined as a violation that would be considered a felony if committed on the outside: homicide, attempted homicide, assault, rape, drug trafficking and such. For prisons, a particularly dear resource is the availability of prison beds in high security facilities. High security facilities have per inmate yearly costs comparable to Ivy League tuition and each new cell built can easily cost six figures. High security placements, therefore, should only be made for inmates who really need them. Sufficiently accurate forecasts were produced, using information regularly available at admission, that would have routinely made more informed placement decisions possible. Unfortunately, several systemwide crises intervened and, to date, the forecasting procedures have not been used.

Environmental Crimes

The international purse-seine fishery for tuna in the Eastern Pacific is responsible for about a quarter of the total catch of yellowfin tuna. By international treaty, boat captains are required to abide by a number of fishing regulations and in particular, employ fishing practices that do not kill dolphins. Dolphin mortality can lead to a variety of sanctions, including the loss of all revenues from tuna caught when the dolphins were killed. And tuna caught when dolphins are killed cannot be sold with the “dolphin safe” label, which effectively eliminates

many markets in Europe and North America.

To enforce the fishing regulations, the Inter-American Tropical Tuna Commission, which supervises the fishery, places observers on a many of the fishing boats. The observer's job is report on each tuna set in substantial detail. A set is when a large net to encircle a school of tuna is deployed and then retrieved.

Among the items they are required to record are whether any dolphin were killed in the fishing process, and if so, how many there were. A problem with this approach is that captains have strong incentives to bribe observers because with a stroke of an observer's a pen, tens of thousands of dollars can be saved. Indeed, there have been allegations that some observers have falsified data and in a few such cases, observers have been caught and fired.

The statistical challenge was to identify from patterns in the data which observers were likely to be corrupt. The solution, in brief, was to develop a model of the purse-seine process that provided figures for expected dolphin mortality under different fishing conditions. For example, when a net is hauled in, dolphins are likely to become entangled when wind, currents, and the speed of the retrieval lead to a folding over of some parts of the net. Problems with the net were routinely recorded and could be used as predictors.

If an observer frequently reported no mortality when the model indicated there should have been some, that observer's credibility was undermined. A follow-up investigation was then initiated. The approach was validated with earlier data in which corrupt observers were identified by other means. The statistical procedures correctly found most of those individuals (after the fact) with very few false positives.

In short, the statistical forecasts of dolphin mortality, coupled with what was actually reported, allowed Inter-American Tropical Tuna Commission to develop profiles of corrupt observers. When these profiles were applied to observers reports, they could trigger investigations that usually proved to be productive. In addition to several other benefits, the statistical procedures aided in the allocation of investigative resources. Rather than trying to investigate all observers, only the “suspicious” observers were investigated (Lennert-Cody and Berk, 2007).

Making the Translation

The tasks undertaken for the kinds of criminal justice applications just described are much the same, at least broadly characterized, as the tasks that could be undertaken for consumer product safety forecasting. Just as with police interventions in cases of domestic violence, regulatory officials charged with consumer safety depend on cost-effective prevention initiatives. Probation and parole authorities employ oversight that may be especially intrusive once past offenses have been proved and that varies as individual profiles change over time, especially if undesirable behavior persists. Regulatory officials could proceed in a similar fashion. The tuna fishing illustration raises the question of when governments use third party auditors, who will the audit be auditors? We turn now to a summary of the tasks inherent in criminal justice forecasting that carry over into oversight of consumer product safety.

Design Decisions

The first and perhaps most important step is to develop a research design, beginning with the processes that are to be examined and the observational units. For example, if there are concerns about imported pharmaceuticals, one could try to forecast defective products at any of several different steps in the supply chain from the raw materials used, to the products coming off the assembly line, to shipments before leaving the factory, to shipments arriving at their destinations, to wholesale distributors, to retail establishments. The units could be individuals doses of the medication, single packages of the medications, or entire shipments from certain places and at certain times.

Consider an example involving food safety. Figure 1, obtained from the U.S. FDA website, is a simplified flow chart of how peanut products from a given processor are distributed. The flow chart was constructed after recent Salmonella contamination of peanut products by the Peanut Corporation of America (PCA). The public health consequences materialize after the events on the far right side of the flow chart. Salmonella contamination occurs on the far left hand side of the flow chart. Flow charts much like Figure 1 apply to virtually all consumer products whether from the United States or abroad.

For illustrative purposes, suppose Figure 1 applied to a large number of peanut product manufacturers, and the goal was to construct procedures from which to forecast Salmonella contamination. Also suppose for simplicity that Salmonella contamination gets carried along largely undisturbed. For example, there are no production processes in the distribution chain that heat the peanut products to very high temperatures. Then, the link in the distribution

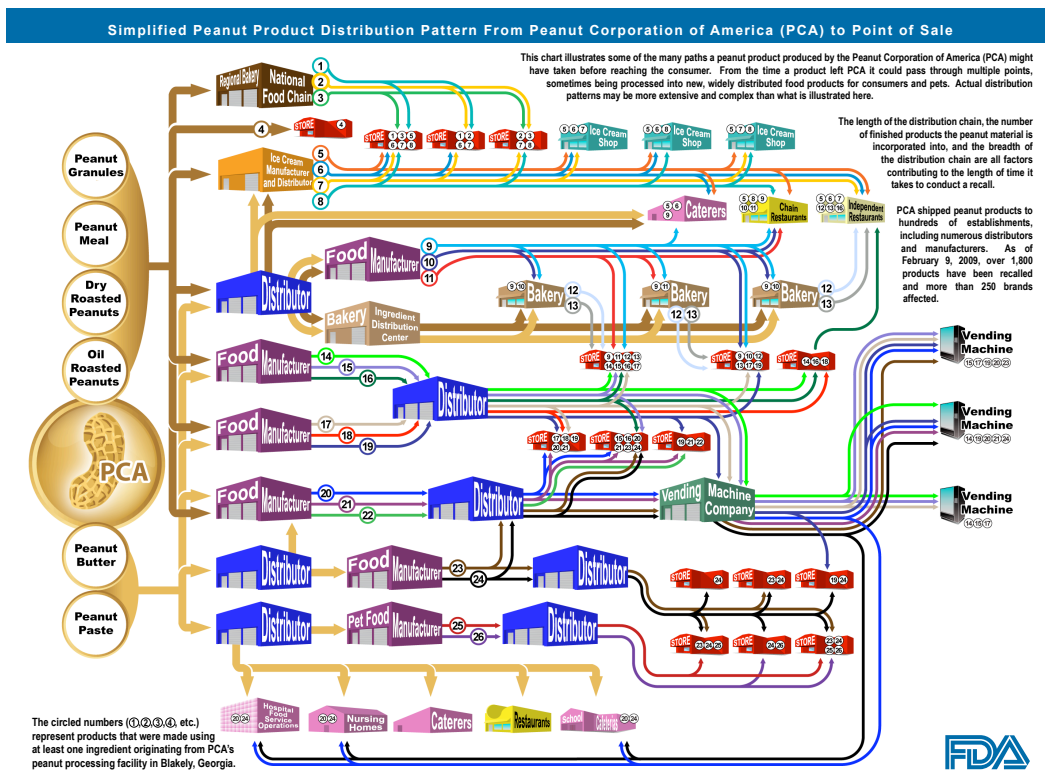


Figure 1: Distribution of Peanut Products

chain for which the forecasts of Salmonella concentrations were made would not matter much.

Developing any forecasting procedure for processes like those represented in Figure 1 would require a large sample of peanut products with measures of Salmonella concentrations for each. Predictors would need to be derived from information that would ordinarily be available. Otherwise, they could not be used when forecasts were desired. The precise nature of those predictors would depend in part on the stage in the distribution process. Some predictors might come from the size and design of the production or distribution facility, from past inspections of the entities in question or even from their recent financial circumstances. For example, it might turn out that smaller facilities, using older technology, and under significant economic pressure were more likely to have had Salmonella problems in the past. And if so, one might anticipate such problems for such establishments in the future.

The quality of the data is very important. Key predictors and outcomes must be well measured, and the data must on their own be a population of interest or a proper sample from such a population. For example, before any forecasts can be used, one needs to determine to what population of observational units they should apply. Would forecasts derived from a distribution chain such as Figure 1 apply to Salmonella contamination for all firms making peanut products or ten years into the future when the production technology may have changed? At the same time, these kinds of technical concerns do not stand alone. They must respond to the goals of stakeholders. Stakeholder intentions motivate the research. As

the famous statistician John Tukey purportedly said, “If it’s not worth doing, it’s not worth doing well.”

To illustrate, the work undertaken for the Inter-American Tropical Tuna Commission had three goals: (1) allocating investigatory resources, (2) preventing record falsification, and (3) punishing offenders. The analysis rested on a population of administrative records that were themselves of interest. The Commission was pursuing its three goals with respect to the actions represented in those administrative data. The focus was on the production step in which tuna were netted. The unit of observation was the set. The outcome of dolphin deaths was apparent, given the policy concerns. Extensive administrative data were readily available, which included a number of predictors with a least strong face validity.

Different policy concerns could have led to a dramatically different research design. For example, had the concern been mercury concentrations in tuna, the observational unit could have been individual tuna and the meat that people consume. The outcome variables and predictors would also have been entirely different, and there would have been no need to collect data on fishing boats. Data collected as the fish entered processing plants would probably have been of better quality and easier to obtain.

One key lesson is that the design decisions must respond to the needs, opportunities and constraints of each situation. There can be no design, or even small set of designs, that will properly apply to all settings. For example, an approach that was effective in an earlier setting may subsequently stumble because some key information is proprietary. Likewise, sometimes designs can rely on data that are already available, and sometimes they will need

to anticipate the collection of data that otherwise would not exist or the establishment of procedures by which the requisite data will be collected in an ongoing manner.

Another lesson is that any design should be driven by technical matters and substantive matters. Both sorts of expertise must be represented among those who are planning the work. And for both the technical and the substantive matters too, there will be the need for specialized expertise. In the pharmaceutical illustration, one would need on the team, for example, expertise on the production processes, legal matters, and the most relevant local and international politics. Such expertise would inform how much effort should go into the data collection. For some purposes “quick and dirty” will suffice. For other purposes, one would need a “full court press.”

Taking Forecasting Errors into Account

One must appreciate from the start that forecasting errors will be made. Perfect foresight is the stuff of science fiction novels and movies (e.g., *Minority Report*). There are no costs from such errors in themselves. There can be enormous costs when those errors are acted upon.

For categorical outcomes such as whether an observer has been taking bribes, the forecasting errors will be of two kinds: false positives and false negatives. For the dolphin analysis, a false positive would falsely identify an observer as a suspect. A false negative would be failing to identify an observer as a suspect who was in fact taking bribes. In product safety, a categorical outcomes might be whether the presence of health hazards exceeded some med-

ical threshold or not. A false positive would be incorrectly asserting that the threshold had been exceeded. A false negative would be failing to assert that the threshold was exceeded when it actually had been. A possible cost of a false positive would be a product recall in error. A possible cost of a false negative would be the deaths of several infants that could have been prevented. Sometimes there will be tradeoffs between false positives and false negatives, in the sense that research strategies to eliminate one of these types of errors might increase the possibility of the other type. How to make such tradeoffs is a policy decision.

Categorical variables are sometimes called “nominal” variables. Observations are placed into categories that are exhaustive and mutually exclusive. Categorical variables with two categories are most common, but more than two is certainly permitted. For example, baby formula can be made unsafe by several different kinds of contaminants. Then, false positives and false negatives should be considered for each type of contaminant.

In practice, one only needs to determine the *relative* costs of false positives and false negatives. For example, in the study done for the California Department of Corrections, prison officials determined that it was far more costly to have a very high risk inmate not be accurately identified as high risk than to have a low risk inmate incorrectly identified as high risk. Ten “over-incarcerations” were treated as if they had the same costs as one “under-incarceration.” The cost ratio of false negatives to false positive was 10 to 1.

There are often several sets of relative costs. Different stakeholders will value the relative costs of false positives and false negatives differently. There are then at least two options. Some overarching body or procedure can force a compromise ratio. Or, different sets of

forecasts can be provided for the different cost ratios. Ideally, the cost ratios are similar enough so that the forecasts are as well. Alternatively, that overarching body or procedure will determine which forecasts to use.

If the relative costs of false negatives and false positives are not explicitly introduced, equal costs are the usual default. This is likely to be a mistake. For the Department of Corrections, equal costs would have produced far less useful and very different forecasts from those that were being sought. In the baby formula example, if the costs of several avoidable infant deaths are much higher than a false-alarm product recall, the forecasts will by design increase the false positive rate relative to the false native rate. That is, the procedure will accept weaker evidence that a product recall is needed than it would if the relative costs were not so extreme, let alone if infant deaths were valued less than a formula producer's economic well-being.

When the response being forecasted is quantitative, constructing useful forecasts can be more technically demanding. For example, one may be interested in forecasting the concentration of a given contaminant in baby formula. Then, the forecasting errors will be over-estimates or under-estimates of those concentrations. The direction and the size of the forecasting error matters, and asymmetric costs are common. For example, overestimates or underestimates of contaminant concentrations with no medical or economic implications may have zero costs because no actions are required. But the costs of larger underestimates may increase in an increasing fashion, while the costs of larger overestimates may increase in a constant fashion. Building these kinds of complexities into a forecasting procedure can

be challenging and is beyond the scope of this paper. A useful discussion and application can be found in the recent paper by Kriegler and Berk (2009).

Analysis of the Data

The data available are analyzed so that associations between predictors and the outcomes to be forecasted are established. “Predictors” are sometimes called regressors, independent variables, or exogenous variables. Outcome variables are sometimes called response variables, dependent variables, or endogenous variables.

The data used for such purposes are often called “training data” because algorithms linking predictors to the response are “trained.” For example, in the study done for the Philadelphia Adult Department of Probation and Parole, the training data were a random sample of 30,000 individuals whose supervision began in 2006 and whose behavior on probation or parole was followed for two years. Among the many predictors were each individual’s past criminal record. The key response was a homicide or an attempted homicide.

Mad cows disease provides an illustration for consumer product safety. On January 9, 2007, the U.S. Food and Drug Administration announced the following informational requirement:

The Food and Drug Administration (FDA) is requiring that manufacturers and processors of human food and cosmetics that are manufactured from, processed with, or otherwise contain, material from cattle establish and maintain records sufficient to demonstrate that the human food or cosmetic is not manufactured

from, processed with, or does not otherwise contain, prohibited cattle materials. These recordkeeping requirements provide documentation for the provisions in FDA's interim final rule entitled "Use of Materials Derived From Cattle in Human Food and Cosmetics." FDA is requiring recordkeeping because manufacturers and processors of human food and cosmetics need records to ensure that their products do not contain prohibited cattle materials, and records are necessary to help FDA ensure compliance with the requirements of the interim final rule (Food and Drug Administration, 2007).

These requirements were imposed so that there would be a data base with which to monitor beef production and other products from cattle. The information sought was determined from earlier epidemiological studies using training data that linked the content of certain cattle feed to bovine spongiform encephalopathy. Particular feed constitutes (e.g., the brain, skull, eyes, trigeminal ganglia, spinal cord from uninspected cattle) would then serve as predictors to identify suspect cattle that were not to be imported or were to be destroyed. Other predictors could include the temperatures used in any rendering process and whether as young calves the cattle were fed infected protein supplements.

One must always evaluate the performance of the forecasting procedures. This requires "test data" that are a probability sample from the same population as the training data. The test data must contain the same predictors and outcomes as employed with the training data. Then, the procedures developed from the training data are applied to the test data. Ideally, the forecasts using the test data are sufficiently accurate for the purposes at hand.

For the Philadelphia study 26,000 (or so) individuals not included at random in the training data became the sample for the test data. For all practical purposes, the forecasting model built with the training data performed equally well in the training data and the test data.

It is important to understand that the common practice of evaluating forecasting performance using the training data is usually a mistake. One is not assessing forecasting skill. One is assessing goodness-of-fit. Goodness of fit refers to how well the statistical model accounts for the data on hand. Forecasting performance refers to how accurately the procedure forecasts with data not used to construct the forecasts (i.e. new data). Goodness of fit statistics will generally provide an inappropriately optimistic sense of forecasting skill because real forecasts are not undertaken. The development of a powerful forecasting algorithm will capitalize not just on systematic patterns in the data, but patterns resulting from noise. The result is “overfitting.” Because in a new data set the noise patterns will likely be quite different, forecasting algorithms generally perform better on the data used to build the algorithm than on new data. So, honest evaluations of forecasting performance require test data.

Honest forecasting assessments also should provide information on uncertainty. The basic problem is this: if the study were done again, the results would almost certainly differ, at least a bit. Some of the difference would result from unsystematic variation in the data that for purposes of this paper can be considered “noise.” It follows that a summary of the impact of that noise should be attached to any forecasts. “Error bands” representing the “margin of error” is one illustration.

Within the usual modeling paradigm used in forecasting, uncertainty is built into the model. That is, one imposes theory about the sources of uncertainty. It is then possible to construct overall measures of consequences of that uncertainty. The algorithmic methods favored in this discussion (described shortly), which can perform better than conventional models, provide special challenges when uncertainty is considered. The issues are quite technical, but substantial progress is being made. An important insight from that work is the uncertainty assessments from conventional forecasting approaches are often on less solid ground than many practitioners realize (Leeb and Pötscher 2005; 2006; 2008). All existing approaches to uncertainty calculations can be significantly compromised (Berk et al., 2009).

Going Operational

Finally, the forecasting procedures must be ported to the setting in which they will be used. This step may be trivial if the forecasting procedures were developed within the system in which they will be used in practice. Commonly, however, this is not the case and the challenges can be daunting. For example, for both the Department of Corrections Study and the Probation/Parole study, forecasting had to be available rapidly in real time as intakes occurred. This meant having forecasting capacity at a large number of desktop computers in many field offices. One approach is to link each desktop to a single server. An alternative approach is to install the forecasting tools necessary on each desktop. In both settings, the server option was chosen.

Forecasting Algorithms

In the past decade, there have been dramatic advances in the statistical tools from which one can construct forecasts. Conventionally, a “model” is formulated from past research or extant theory. The model is essentially a simplified description of how the data came to be. That model is then applied to the data and some very specific statistical features of the model are computed from the data. The main weakness of this approach is that the forecasts are highly model dependent. If the model is not a good one, the accuracy of the forecasts can suffer dramatically. In recent work, however, much less about the model need be specified in advance. Many of the model’s key features are arrived at inductively from the data. It now seems increasingly clear that there are a large number of situations in which this is a far better way to proceed. We turn to a brief summary.

As just noted, a model is a quantitative theory of the processes responsible for generating the data. That is, a model is a quantitative explanation of how the data on hand came to be. To take the linear regression example, one is interested in understanding how nature produced an outcome, often denoted by y_i , from a set of predictors, often denoted by $x_{i,p}$. One can think of $x_{i,p}$ as input and y_i as output. Then, $y_i = \beta_0 + \sum_{p=1}^P \beta_p x_{i,p} + \varepsilon_i$, where $\varepsilon \sim \text{NIID}(0, \sigma^2)$, there are P predictors $x_{i,p}$, and $i = 1, 2, 3, \dots, N$ cases. For each of the N cases, nature multiplies the values of each of the P predictor variables x_{ip} by its constant β_p , adds the value of constant β_0 and then adds an independent random perturbation ε , which is a random realization of a normal distribution having a mean of 0.0 and a variance of σ^2 .

The primary goal of this model is to represent how nature works — how nature generated

the outcome. It may turn out that the model also produces usefully accurate forecasts, but that is usually a secondary concern if the intent is to capture the data generation process. The same holds for more general models such as $y_i = f(X_i) + \varepsilon_i$, where X_i is a set of predictors, and as before, $\varepsilon_i \sim \text{NIID}(0, \sigma^2)$. We are no longer limited to a linear combination of the predictors. The function of X_i is to be determined by the data.

In this canonical context, it is easy to forget that the goal of forecasting is to forecast accurately. The goal is not to identify important causal variables, let alone develop an explanatory model, although such byproducts can be desirable on other grounds. In conventional econometric language, an “astructural” model will suffice, and forecasting accuracy should not be sacrificed so that “structural” models can be developed. Thus, if there is a predictor that adds forecasting skill even if it is not in any way responsible for generating the values of y_i , it should be allowed to contribute. If the shoe size of a peanut factory’s manager helps to forecast the level of salmonella contamination, it should be included among the predictors.

Focussing on forecasting skill alone can be quite liberating. In a single minded fashion, one can bring any technology to bear that produces good forecasts. It will also be clear which approach to use; forecasting accuracy will separate the winners from the losers.

Over that past decade, a wide variety of new data analysis procedures have been developed by statisticians and computer scientists that can be called “model free.” They can be used without any concerns about whether the data generating mechanisms are being properly represented. Called “statistical learning” by most statisticians and “machine learning”

by most computer scientists, their intent is to search extensively through the data to find how a set of predictors is associated with a response. An earlier generation of statisticians would call the approach “exploratory data analysis.” Consider the following example.

1. Fit a regression model to the data. If the outcome is binary, that regression model might be logistic regression.
2. Determine which outcome observations are accurately identified and which are not. In an epidemiological study of bovine spongiform encephalopathy, are the infected cows identified as such? Likewise are the uninfected cows identified as such?
3. Reweight the data. Give more weight to observations that were classified incorrectly.
4. Repeat steps 1-3 a large number of times (e.g., 1000).
5. For each case, compute a weighted average of the class designated by the model. The weights come from the models. The better a given model fits the data, the more weight given to its output when the average is computed.

The steps just outlined have much in common with a machine learning procedure called boosting. Boosting is one of the new kinds of inductive procedures that can perform much better than conventional parametric models (Hastie et al., 2001; Berk, 2008b). The $f(X)$ is determined as part of the fitting process. It is not specified in advance. The danger is that the algorithm will be too responsive and build on idiosyncratic features of the data. Insofar as this happens, overfitting is the result. Forecasting performance can degrade substantially

when applied to new data. The remedy is to have a test data set from which honest measures of forecasting accuracy can be obtained.

Boosting is one especially popular technique (actually a set of procedures) within a machine/statistical learning framework. Two popular competitors with at least comparable forecasting performance are random forests and support vector machines. For the problems at hand, all of these procedures can be seen as ways to arrive at the $f(X)$ inductively. Looked at in this manner (and there are other ways to look), machine/statistical learning has much in common with the older tradition of smoothing. The generalized additive model (Hastie and Tibshirani, 1990) is probably the closest analog. The usual regression equation of $y_i = \beta_0 + \sum_{p=1}^P \beta_p x_{ip} + \varepsilon_i$ is replaced by $y = \beta_0 + \sum_{p=1}^P f_p(x_{ip}) + \varepsilon_i$. Each predictor has its own inductively produced functional relationship with the response, which are then combined in an additive fashion.

Although the main goal in profiling is to forecast accurately, many machine/statistical learning procedures can provide measures of the forecasting contribution of each predictor. The basic idea is to compute how much less accurate a forecast is when any given predictor is not allowed to contribute to the forecasting exercise. Then, predictors may be ordered by their forecasting “importance.”

It is also possible to show the inductively-generated functional form linking each predictor to the response. For example, Figure 2 shows how the age of an individual on parole or probation is related to the likelihood that he or she will commit a homicide or attempted homicide, other predictors held constant. The technical details need not concern us. When

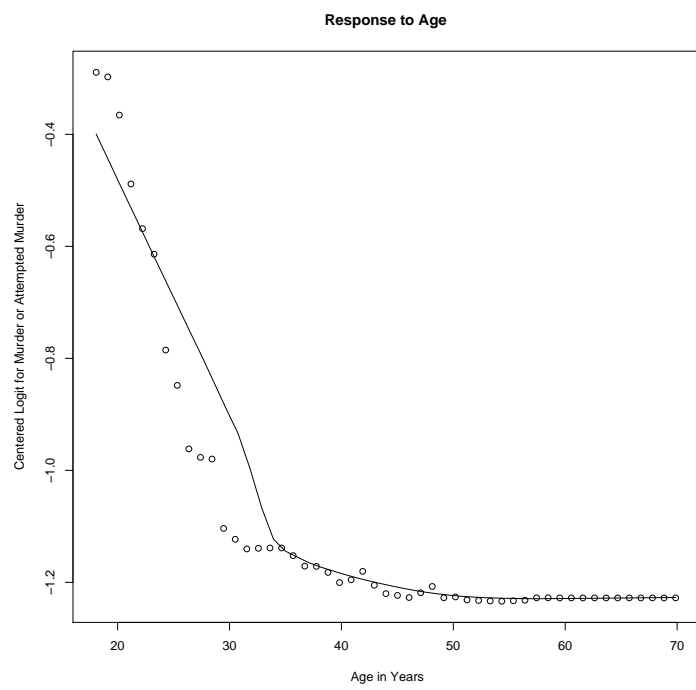


Figure 2: Partial Dependence Plot for Age

the parolee or probationer is under 20 years of age the risks are the highest. The risk drops precipitously to about age 30, after which age seems to have virtually no relationship to the outcome. Thus, although a difference in age between an individual of 18 and 25 can be very important, the difference in age between an individual at 28 and 35 hardly matters at all.

Some might worry whether it is “fair” to use age. The same holds for gender, which is also a very important predictor. Prospective murders have no control over either variable. And if there are causal links to homicide perpetration, they are not well understood. But, this is not a statistical decision. On ethical, legal, or political grounds one has to determine if the loss in accuracy is worth it. For example, in a city like Philadelphia, if age and gender are not used to forecast homicides, there could easily be 50 homicides a year that would have otherwise been prevented. There will be tradeoffs because most homicides are committed by young men. Moreover, background variables are already used informally in probation and parole decisions. At least in a forecasting model, no one is hiding the ball.

It is likely that similar issues will arise in profiling of consumer safety violators. What sorts of predictors are legitimate? For example, can one use as a predictor the country in which a food processor is located? In the instance of bovine spongiform encephalopathy, should all processed beef from Canada have been prohibited? Such decisions would depend on the forecasting importance of a given predictor and tradeoffs that would follow were it not used.

As noted briefly earlier, forecasts should be combined with assessments of uncertainty. For algorithmic methods, the usual approaches to uncertainty do not apply. The basic prob-

lem is that inductive methods violate a key assumption of conventional statistical inference (Barnett,1982). For such inference, the model must be known before the data are examined. However, relatively simple resampling procedures can be used to, in effect, simulate what would happen if the study were repeated. The result is a useful assessment of how stable one's forecasts really are. How different would the forecasts likely be if the study were repeated?

Conclusions

In principle, the motivations driving forecasting in criminal justice settings can apply to consumer product safety concerns, and the same kinds of benefits can follow. Looking forward, forecasts can help determine how to allocate scarce oversight resources. Product safety “hot spots” can be identified and subjected to especially close scrutiny. Concentrating oversight where it is most needed may help identify problems in a more cost-effective manner. One consequence can be a more efficient search for defective products and those responsible for them.

Better still is the prospect of prevention. Prevention implies an intervention before the product safety is compromised. Clearly, prevention cannot be undertaken without a forecast of what would happen if nothing were done. However, prevention too can be costly. One of the benefits of usefully accurate forecasts is that prevention resources can be concentrated where they are most needed.

There can be important long run benefits. Knowledge that more effective oversight is in place can serve as a deterrent. It is widely believed that deterrence is a function of the probability of apprehension and severity of punishments that follow. It is also widely believed that the probability of apprehension is the more important of the two. Effective product safety oversight — made possible through carefully developed statistical forecasts — speaks directly to that probability.

Finally, it is important to emphasize that forecasts of the sort considered in this chapter are meant to help inform decisions. They are not intended to wholly determine the decisions made. Even very accurate forecasts should be ignored if acting on them leads undesirable outcomes. There needs to be a separation between the information that the forecasts can provide and what is done with that information.

References

- Barnett, V. (1982) *Comparative Statistical Inference*. New York: John Wiley & Sons.
- Berk, R.A., (2008a) “Forecasting Methods in Crime and Justice.” *Annual Review of Law and Social Science*, J. Hagan, K.L. Schepple, and T.R. Tyler (eds.), Palo Alto: Annual reviews.
- Berk, R.A., (2008b) *Statistical Learning from a Regression Perspective*. New York: Springer.
- Berk, R.A., Brown, L., and L. Zhao (2009) “Statistical Inference After Model Selection.” Department of Statistics, University of Pennsylvania, working paper under review.
- Berk, R.A., Sorenson, S., and Y. He (2005) “Developing a Practical Forecasting Screener for Domestic Violence Incidents.” *Evaluation Review* 29(4): 358-382.
- Berk, R.A., Kriegler, B. and J-H Baek (2006) ”Forecasting Dangerous Inmate Misconduct: An Application of Ensemble Statistical Procedures.” *Journal of Quantitative Criminology* 22(2): 131-145.
- Berk, R.A., Sherman, L., Barnes, G., Kurtz, E., and L. Ahlman, (2009a) “Forecasting Murder within a Population of Probationers and Parolees: A High Stakes Application of Statistical Learning.” *Journal of the Royal Statistical Society* (Series A) 172, part 1: 191-211.
- Berk, R.A., Ahlman, L., Barnes, G., Kurtz, E., and L. Ahlman, (2009b) “When Second

- Best Is Good Enough: A Comparison Between A True Experiment and a Regression Discontinuity Quasi-Experiment.” Working Paper.
- Dutton, D.G., and R.R. Kropp (2000) “A Review of Domestic Violence Risk Instruments.” *Trauma, Violence, & Abuse* 1(2): 171-181.
- Food and Drug Administration, 21 CFR Parts 189 and 700, [Docket No. 2004N-0257], RIN 0910-AF48, January 9, 2007.
- Glaser, D. (1987) “Classification for Risk,” in D. M. Gottfredson and M. Tonry (eds.) *Prediction and Classification*, Chicago, University of Chicago Press.
- Granger, C.W. J. (1989) *Forecasting in Business and Economics*, second edition. New York: Academic Press.
- Hastie, T.J., and Tibshirani (1990) *Generalized Additive Models*. New York: Chapman and Hall.
- Hastie, T., Tibshirani, R., and J, Friedman (2001) *Elements of Statistical Learning*. New York: Springer.
- Holland, P.W. (1986) “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81: 945-960.
- Kriegler, B., and R.A. Berk (2009) “Estimating the Homeless Population in Los Angeles: An Application of Cost-Sensitive Stochastic Gradient Boosting.” UCLA Department of Statistics, Working paper and under review.

- Leeb, H., B.M. Pötscher (2005) “Model Selection and Inference: Facts and Fiction,” *Economic Theory* 21: 21–59.
- Leeb, H., B.M. Pötscher (2006) “Can one Estimate the Conditional Distribution of Post-Model-Selection Estimators?” *The Annals of Statistics* 34(5): 2554–2591.
- Leeb, H., B.M. Pötscher (2008) “Model Selection,” in T.G. Anderson, R.A. Davis, J.-P. Kreib, and T. Mikosch (eds.), *The Handbook of Financial Time Series*, New York, Springer: 785–821.
- Lennert-Cody, C.E. and R.A. Berk (2007) “Statistical Learning Procedures for Monitoring Regulatory Compliance: An Application to Fisheries Data.” *Journal of the Royal Statistical Society, Series A* 170, Part 3: 191-211.
- Rubin, D. B. (1986) “Which Ifs Have Causal Answers.” *Journal of the American Statistical Association* 81: 961-962.
- United States Sentencing Commission, (2006) *Final Report on the Impact of The United States v. Booker on Federal Sentencing*. Washington, D.C., United States Sentencing Commission.