Prediction

Lawrence Hubert

July 25, 2011

Starting Quotes

The race is not always to the swift nor the battle to the strong — but that's the way to lay your bets.
– Damon Runyon

The only function of economic forecasting is to make astrology look good.
– John Kenneth Galbraith

If all else fails, immortality can always be assured by spectacular error.
– John Kenneth Galbraith

I like also the men who study the Great Pyramid, with a view to deciphering its mystical lore. Many great books have been written on this subject, some of which have been presented to me by their authors. It is a singular fact that the Great Pyramid always predicts the history of the world accurately up to the date of publication of the book in question, but after that date it becomes less reliable.
– Bertrand Russell

The list of studies in which the regression factor has been neglected grows monotonous, as well as distressing.
– Philip Rulon (1941)

What you already know

The attempt to predict the values on some (dependent) variable by a function of (independent) variables is typically approached by simple or multiple regression, for one and more than one predictor, respectively.

The most common combination rule is a linear function of the independent variables obtained by least-squares, i.e., the linear combination that minimizes the sum of the squared residuals between the actual values on the dependent variable and those predicted from the linear combination.

In the case of simple regression, scatterplots again play a major role in assessing linearity of the relationship, the possible effects of outliers on the slope of the least-squares line, and the influence of individual objects in its calculation.

Regression slopes, in contrast to the correlation, are neither scale invariant nor symmetric in the dependent and independent variables.

One usually interprets the least-squares line as one of expecting, for each unit change in the independent variable, a regression slope change in the dependent variable.

There are several topics involving prediction that do not (necessarily) concern linear regression, and because of this, no extended discussion of these is given. One area important for the legal system is Sex Offender Risk Assessment, and the prediction of recidivism for committing another offense.

The Rapid Risk Assessment for Sexual Offender Recidivism (or the more common acronym, RRASOR, and pronounced "razor"). It is based on four items: Prior Sex Offense Convictions – 0, 1, 2, or 3 points for 0, 1, 2, or 3+ prior convictions, respectively; Victim Gender: only female victims (0 points); only male victims (1 point); Relationship to Victim: only related victims (0 points); any unrelated victim (1 point); Age at Release: 25 or more (0 points); 18 up to 25 (1 point).

Another approach to prediction that we do not develop, is in the theory behind chaotic systems, such as the weather. A hallmark of such dynamic prediction problems is an extreme sensitivity to initial conditions, and a general inaccuracy in prediction even over a relatively short time frame.

The person best known for chaos theory is Edward Lorenz and his "butterfly effect" – very small differences in the initial conditions for a dynamical system (e.g., a butterfly flapping its wings somewhere in Latin America), may produce large variations in the long term behavior of the system.

Topics you may not know as well

regression toward the mean;

methods involved in using regression for prediction that incorporate corrections for unreliability;

differential prediction effects in selection based on tests;

interpreting and making inferences from regression weights;

the distinction between actuarial (statistical) and clinical prediction.

Regression toward the mean

Regression toward the mean is a phenomenon that will occur whenever dealing with (fallible) measures with a less-than-perfect correlation.

The word "regression" was first used by Sir Francis Galton in his 1886 paper, *Regression Toward Mediocrity in Hereditary Stature*, where he showed that heights of children from very tall or short parents would regress toward mediocrity (i.e., toward the mean) — exceptional scores on one variable (parental height) would not be matched with such exceptionality on the second (child height).

Regression toward the mean is a ubiquitous phenomenon, and given the name "regressive fallacy" whenever cause is ascribed where none exists.

Generally, interventions are undertaken if processes are at an extreme, e.g., a crackdown on speeding or drunk driving as fatalities spike; treatment groups formed from individuals who are seriously depressed; individuals selected because of extreme behaviors, both good or bad; and so on.

There are many common instances where regression may lead to invalid reasoning: I went to my doctor and my pain has now lessened; I instituted corporal punishment and behavior has improved; he was jinxed by a *Sports Illustrated* cover because subsequent performance was poorer (i.e., the "sophomore jinx"); although he hadn't had a hit in some time, he was "due," and the coach played him; and on and on.

More generally, any time one optimizes with respect to a given sample of data by constructing prediction functions of some kind, there is an implicit use and reliance on data extremities. In other words, the various measures of goodness-of-fit or prediction we might calculate need to be cross-validated either on new data or by a clever sample reuse strategy such as the well-known jackknife or bootstrap procedures.

The degree of "shrinkage" we see in our measures based on this cross-validation, is an indication of the fallibility of our measures and the (in)adequacy of the given sample sizes.

We have the "winner's curse," where someone is chosen from a large pool (e.g., of job candidates), who then doesn't live up to expectation; or when we attribute some observed change to the operation of "spontaneous remission."

As Campbell and Kenny note: "many a quack has made a good living from regression toward the mean."

Actuarial Versus Clinical Prediction

Paul Meehl in his classic 1954 monograph, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, created quite a stir with his convincing demonstration that mechanical methods of data combination, such as multiple regression, outperform (expert) clinical prediction.

The enormous amount of literature produced since the appearance of this seminal contribution, has uniformly supported this general observation; similarly, so have the extensions suggested for combining data in ways other than by multiple regression, e.g., by much simpler unit weighting schemes, or those using other prior weights.

It appears that individuals who are conversant in a field are better at selecting and coding information than they are at actually integrating it.

Combining such selected information in some more mechanical manner will generally do better than the person choosing such information in the first place.

This conclusion can be pushed further: if we formally model the predictions of experts using the same chosen information, we can generally do better than the experts themselves. Such formal representations of what a judge does, are called "paramorphic."

In an influential review paper, Dawes (1979) discussed what he called proper and improper linear models, and argued for the "robust beauty of improper linear models."

A proper linear model is one obtained by some optimization process, usually least-squares.

Improper linear models are not "optimal" in this latter sense, and typically have their weighting structures chosen by a simple mechanism, e.g., random or unit weighting.

Again, improper linear models generally outperform clinical prediction, but even more surprisingly, improper models typically outperform proper models in cross-validation.

What seems to be the reason, is the notorious instability of regression weights with correlated predictor variables, even if sample sizes are very large.

Generally, we know that simple averages are more reliable than individual observations, so it may not be so surprising that simple unit weights are likely to do better on cross-validation than those found by squeezing "optimality" out of a sample.

Given that the *sine qua non* of any prediction system is its ability to cross-validate, the lesson may be obvious — statistical optimality with respect to a given sample may not be the best answer when we wish to predict well.

The idea that statistical optimality may not lead to the best predictions, seems counterintuitive, but as argued by Roberts and Pashler (2000), just the achievement of a good fit to observations does not necessarily mean we have found a good model.

In fact, because of the overfitting of observations, choosing the model with the absolute best fit is apt to result in poorer predictions.

The more flexible the model, the more likely it is to capture not only the underlying pattern but unsystematic patterns such as noise.

A single general purpose tool with many adjustable parameters is prone to instability and greater prediction error as a result of high error variability.

An observation by John von Neumann is particulary germane: "With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk."

More generally, this notion that "less-is-more" is difficult to get one's head around, but as Gigerenzer and others have argued (e.g., see Gigerenzer and Brighton, 2009), it is clear that simple heuristics can at times be more accurate than complex procedures.

All of the work emanating from the idea of the "robust beauty of improper linear models" *et sequelae* may force some reassessment of what the normative ideals of rationality might be.

Most reduce to simple cautions about overfitting one's observations, and then hoping for better predictions because an emphasis has been placed on immediate optimality instead of the longer-run goal of cross-validation.

Henry A. Wallace and the modeling of expert judgements

There are several historical connections between Henry A. Wallace, one of Franklin D. Roosevelt's Vice-Presidents (1940–1944), and the formal (paramorphic) modeling of the prediction of experts, and applied statistics more generally.

Wallace wrote a paper (1923) in the *Journal of the American Society of Agronomy* (*13*, 300–304), entitled: *What Is In the Corn Judge's Mind*?

The data used in this study were ratings of possible yield for some 500 ears of corn from a number of experienced corn judges.

In addition to the ratings, measurements were taken on each ear of corn over six variables: length of ear; circumference of ear; weight of kernel; filling of the kernel at the tip (of the kernel); blistering of kernel; starchiness.

Also, because all the ears were planted in 1916, one ear to a row, the actual yields for the ears were available as well. The method of analysis for modeling both the expert judgements of yield and actual yield was through the new method of path coefficients just developed by Sewall Wright in 1921 (*Correlation and Causation*, *Journal of Agricultural Research*, *20*, 557–585).

The results were final "scorecards" for how the judges and the actual yield values could be assessed by the six factors (each was normalized to a total of 100 "points").

JUDGES' SCORE CARD:
Length – 42.0
Circumference – 13.6
Weight of kernel – 18.3
Filling of kernel at tip – 13.3
Blistering of kernel – 6.4
Absence of starchiness – 6.4
Total – 100.00

ACTUAL YIELD SCORE CARD:
Length – 7.7
Circumference – 10.0
Weight of kernel – 50.0
Filling of kernel at tip – 18.0
Blistering of kernel – 9.0
Absence of starchiness – 5.3
Total – 100.00

Incorporating reliability corrections in prediction

The model for how any observed score, X, might be constructed additively from a true score, T_X, and an error score, E_X, where E_X is typically assumed uncorrelated with T_X: $X = T_X + E_X$.

When we consider the distribution of an observed variable over, say, a population of individuals, there are two sources of variability present in the true and the error scores.

If we are interested primarily in structural models among true scores, then some correction must be made because the common regression models implicitly assume that variables are measured without error.

The estimation, $\hat{T}_X$, of a true score from an observed score, X, was derived using the regression model by Kelley in the 1920's , with a reliance on the algebraic equivalence that the squared correlation between observed and true score is the reliability.

If we let $\hat{\rho}$ be the estimated reliability, Kelley's equation can be written as

$$\hat{T}_X = \hat{\rho}X + (1 - \hat{\rho})\bar{X} \ ,$$

where $\bar{X}$ is the mean of the group to which the individual belongs.

In other words, depending on the size of $\hat{\rho}$, a person's estimate is partly due to where they are in relation to the group — upwards if below the mean; downwards if above.

The application of this statistical tautology in the examination of group differences provides such a surprising result to the statistically naive, that this equation has been called "Kelley's Paradox".

We might note that this notion of being somewhat punitive of performances better than the group to which one supposedly belongs, was not original with Kelley, but was known at least 400 years earlier. In the words of Miguel de Cervantes (1547–1616): "Tell me what company you keep and I'll tell you what you are."

In the topic of errors-in-variables regression, we try to compensate for the tacit assumption in regression that all variables are measured without error.

Measurement error in a response variable does not bias the regression coefficients per se, but it does increase standard errors, and thereby reduces power. This is generally a common effect: unreliability attenuates correlations and reduces power even in standard ANOVA paradigms.

Measurement error in the predictor variables biases the regression coefficients. For example, for a single predictor, the observed regression coefficient is the "true" value multiplied by the reliability coefficient.

Thus, without taking account of measurement error in the predictors, regression coefficients will generally be underestimated, producing a biasing of the structural relationship among the true variables.

Differential prediction effects in selection

One area in which prediction is socially relevant is in selection based on test scores, whether for accreditation, certification, job placement, licensure, educational admission, or other high-stakes endeavors.

We note that most of these discussions about fairness of selection need to be phrased in terms of regression models relating a performance measure to a selection test; and whether the regressions are the same over all identified groups of relevance, e.g., ethnic, gender, age, and so on.

Specifically, are slopes and intercepts the same? If so or if not, how does this affect the selection mechanism being implemented, and whether it can be considered fair?

Interpreting and making inferences from regression weights

Mathematics has given economics rigor, but alas, also mortis.
– Robert Heilbroner

Statistics are the triumph of the quantitative method, and the quantitative method is the victory of sterility and death.
– Hillaire Belloc (*The Silence of the Sea*)

Years ago a statistician might have claimed that statistics deals with the processing of data ... today's statistician will be more likely to say that statistics is concerned with decision making in the face of uncertainty.
– H. Chernoff and L. E. Moses, *Elementary Decision Theory* (1959)

Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex.
– Fritz Machlup

When a true genius appears in this world, you may know him by this sign, that the dunces are all in confederacy against him.
– Jonathan Swift

Although multiple regression can be an invaluable tool in many arenas, the interpretive difficulties that result from the interrelated nature of the independent variables must always be kept in mind.

As in the World War II example in the reading, depending on what variables are (or are not) included, the structural relationship among the variables can change dramatically. At times, this malleability can be put to either good or ill usage.

For example, in applying regression models to argue for employment discrimination (e.g., in pay, promotion, hiring, and so on), the multivariable system present could be problematic in arriving at a "correct" analysis.

Depending on what variables are included, some variables may "act" for others (as "proxies"), or be used to hide (or at least, to mitigate) various effects. If a case for discrimination rests on the size of a coefficient for some "dummy" variable that indicates group membership (according to race, sex, age, and so on), it may be possible to change its size depending on what variables are included or excluded from the model, and their relationship to the dummy variable.

The (Un)reliability of Clinical Predictions (of Violence)

If your mother says she loves you, check it out.
Adage from the Chicago City News Bureau

Prosecutors in Dallas have said for years – any prosecutor can convict a guilty man. It takes a great prosecutor to convict an innocent man.
Melvyn Bruder (The Thin Blue Line)

The last section in this chapter concerns the (un)reliability of clinical (behavioral) prediction, particularly for violence, and will include two extensive redactions at the end of the section:

one is the majority opinion in the Supreme Court case of Barefoot v. Estelle (Decided, July 6, 1983) and an eloquent Justice Blackmun dissent; the second is an Amicus Curiae brief in this same case from the American Psychiatric Association on the accuracy of clinical prediction of future violence.

The Psychiatrist, James Grigson, featured so prominently in the opinions for Barefoot v. Estelle and the corresponding American Psychiatric Association Amicus brief, played the same role repeatedly in the Texas legal system.

For over three decades before his retirement in 2003, he would testify when requested at death sentence hearings to a high certainty as to "whether there is a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society."

An affirmative answer by the sentencing jury imposed the death penalty automatically, as it was on Thomas Barefoot; he was executed on October 30, 1984.

The (Questionable) Use of Statistical Models

The form of statistical practice most commonly carried out by those with a mathematical bent (and in contrast to those more concerned with simple manifest forms of data analysis and visualization), is through the adoption of a stochastic model commonly containing (unobserved) latent variables.

Here, some data generating mechanism is postulated, characterized by a collection of parameters and strong distributional assumptions.

Based on a given data set, the parameters are estimated, and usually, the goodness-of-fit of the model assessed by some statistic.

We might even go through a ritual of hoping for non-significance in testing a null hypothesis that the model is true (generally through some modified chi-squared statistic heavily dependent on sample size).

The cautionary comments of Roberts and Pashler (2000) should be kept in mind that the presence of a good fit does not imply a good (or true) model.

Moreover, models with (many) parameters are open to the problems engendered by over-fitting and of a subsequent failure to cross-validate.

The handout provides the abstract of the Roberts and Pashler (2000) article, *How Persuasive Is a Good Fit? A Comment on Theory Testing* (*Psychological Review*, *107*, 358–367).

A model-based approach is assiduously avoided throughout this monograph.

It seems ethically questionable to base interpretations about some given data set and the story that the data may be telling, through a model that is inevitably incorrect, at least at the periphery if not at its core.

As one highly cherished example in the behavioral sciences, it is now common to frame questions of causality through structural equation (or path) models, and to perform most data analysis tasks through the fitting of various highly parameterized latent variable models.

In a devastating critique of this type of practice, David Freedman in a *Journal of Educational Statistics* article (*As Others See Us: A Case Study in Path Analysis*; 1987, *12*, 101–128) ends with this paragraph:

My opinion is that investigators need to think more about the underlying social processes, and look more closely at the data, without the distorting prism of conventional (and largely irrelevant) stochastic models.

Estimating nonexistent parameters cannot be very fruitful. And it must be equally a waste of time to test theories on the basis of statistical hypotheses that are rooted neither in prior theory nor in fact, even if the algorithms are recited in every statistics text without caveat.

The late Leo Breiman took on the issue directly of relying on stochastic models (or, as he might have said, "hiding behind"), in (almost) all of contemporary statistics.

What Breiman advocates is the adoption of optimization in place of parameter estimation, and of methods that fall under the larger rubric of (supervised or unsupervised) statistical learning theory.

Currently, this approach is best exemplified by the comprehensive text, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second Edition, 2009) (T. Hastie, R. Tibshirani, and J. Friedman).

We give two quotes from Breiman. The first is from an invited discussion of a paper from *Statistical Science* by Jim Ramsay (*Monotone Regression Splines in Action* (1988, *3*, 425–441).

It reacts to Ramsay's preference for maximum likelihood estimation or a Bayesian approach to the fitting of splines, both of which require distributional assumptions:

In describing the fitting of the yarn data, the author states that "The fitting criterion could be least squares, but this is not desirable when the dependent variable is being transformed," and he opts for maximum likelihood or Bayesian approaches; in this instance for maximum likelihood.

The reason why least squares is not desirable is not stated. Least squares is an old and reliable friend. Maximum likelihood or Bayesian approaches always impose distributional assumptions on the data which are usually difficult to verify.

If you clap when Tinkerbell asks "do you believe in fairies" then fine. If you are in doubt, as I often am with real data, then use an earthy friend

The second quote is the abstract from Leo Breiman's *Statistical Science* piece, *Statistical Modeling: The Two Cultures* (2001, *16*, 199–215):

There are two cultures in the use of statistical modeling to reach conclusions from data.

One assumes that the data are generated by a given stochastic data model.

The other uses algorithmic models and treats the data mechanism as unknown.

The statistical community has been committed to the almost exclusive use of data models.

This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.

Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics.

It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets.

If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

The view of statistics to be followed in this monograph is, to use a Breiman term, "earthy."

We might go so far as to consider what (linear) regression models can (or cannot) do, or the implications of a basic sampling model, but we go no further than least-squares treated as an algorithmic optimization process, and a suggestion to adopt various sample reuse methods to gauge stability and assess cross-validation.

Remembering the definition of a *deus ex machina* – a plot device in Greek drama whereby a seemingly insoluble problem is suddenly and abruptly solved with the contrived and unexpected intervention of some new character or god –

we will not invoke any statistical *deus ex machina* analogues.

A (slightly) amusing story told in some of our beginning statistics sequences reflects this practice of postulating a *deus ex machina* to carry out statistical interpretations.

Three academics – a philosopher, an engineer, and a statistician – are walking in the woods toward a rather large river that needs to be crossed

The pensive philosopher stops, and opines about whether they really need to cross the river;

the engineer pays no attention to the philosopher and proceeds immediately to chop down all the trees in sight to build a raft;

the statistician yells to the other two: "stop, assume a boat."

Stochastic data models do have a place but not when that is only as far as it goes.

When we work solely within the confines of a closed system given by the model, and base all inferences and conclusions under that rubric alone (e.g., we claim a causal link because some path coefficient is positive and significant), the ethicality of such a practice is highly questionable.

George Box has famously said that "all models are wrong, but some are useful" (or Henri Theil's similar quip: "models are to be used, but not to be believed").

Box was referring to the adoption of a model heuristically to guide a process of fitting data; the point being that we only "tentatively entertain a model," with that model then subjected to diagnostic testing and reformulation, and so on iteratively.

The ultimate endpoint of such a process is to see how well the fitted model works, for example, on data collected in the future.

Once again, some type of (cross-)validation is essential, which should be the *sine qua non* of any statistical undertaking.