Prediction

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 14, 2013

Beginning Quotations

Its tough to make predictions—especially about the future.
— Yogi Berra

I wold not say that the future is necessarily less predictable than the past. I think the past was not predictable when it started.
— Donald Rumsfeld

Week 5: Prediction

— Henry A. Wallace and the modeling of expert judgment ("What Is In the Corn Judge's Mind?"); the distinction between actuarial and clinical prediction, and the Dawes notion of the "robust beauty of improper linear models"

— Barefoot v. Estelle (1983): There is no merit to petitioner's argument that psychiatrists, individually and as a group, are incompetent to predict with an acceptable degree of reliability that a particular criminal will commit other crimes in the future, and so represent a danger to the community.

Required Reading:
SGEP (141–173) —
Regression Toward the Mean
Actuarial Versus Clinical Prediction
Incorporating Reliability Corrections in Prediction
Differential Prediction Effects in Selection
Interpreting and Making Inferences From Regression Weights
The (Un)reliability of Clinical Prediction

Suggested Reading:
Appendix: Continuation of the American Psychiatric Association, Amicus Curiae Brief: Barefoot v. Estelle
Appendix: Opinion and Dissent in the U.S. Supreme Court, Barefoot v. Estelle (Decided, July 6, 1983)

Film: *The Thin Blue Line* (102 minutes)

Simple and Multiple Regression

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

The most common combination rule is a linear function of the independent variables obtained by least squares; that is, 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 relation, the possible effects of outliers on the slope of the least-squares line, and the influence of individual observations 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.

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 used by Galton in his well-known 1886 article, "Regression Towards Mediocrity in Hereditary Stature."

Galton showed that heights of children from very tall or short parents regress toward mediocrity (that is, toward the mean) and exceptional scores on one variable (parental height) are not matched with such exceptionality on the second (child height).

This observation is purely due to the fallibility for the various measures and the concomitant lack of a perfect correlation between the heights of parents and their children.

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 (for example, a crackdown on speeding or drunk driving as fatalities spike, treatment groups formed from individuals who are seriously depressed, or individuals selected because of extreme good or bad behaviors).

In all such instances, whatever remediation is carried out will be followed by some lessened value on a response variable. Whether the remediation was itself causative is problematic to assess given the universality of regression toward the mean.

A variety of phrases seem to get attached whenever regression toward the mean is probably operative:

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

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

Or, when a change of diagnostic classification results upon repeat testing for an individual given subsequent one-on-one tutoring (after being placed, for example, in a remedial context).

More personally, there is "editorial burn-out" when someone is chosen to manage a prestigious journal at the apex of a career, and things go quickly downhill from that point.

Actuarial Versus Clinical Prediction

Paul Meehl in his iconic 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, for example, 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 a more mechanical manner will generally do better than the person choosing such information in the first place.

A 2005 article by Robyn Dawes in the *Journal of Clinical Psychology* (*61*, 1245–1255) has the intriguing title "The Ethical Implications of Paul Meehl's Work on Comparing Clinical Versus Actuarial Prediction Methods."

Dawes' main point is that given the overwhelming evidence we now have, it is unethical to use clinical judgment in preference to the use of statistical prediction rules. We quote from the abstract:

Whenever statistical prediction rules . . . are available for making a relevant prediction, they should be used in preference to intuition. . . . Providing service that assumes that clinicians "can do better" simply based on self-confidence or plausibility in the absence of evidence that they can actually do so is simply unethical.

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 referred to as "paramorphic."

Improper Linear Models

In an influential review article, Dawes (1979) discussed proper and improper linear models and argued for the “robust beauty of improper linear models.”

A proper linear model is one obtained by an 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, for example, by 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.

A Good Fit Does Not Mean a Good Model

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 particularly germane: “With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk.

Henry A. Wallace and the Modeling of Expert Judgments

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 an article (1923) in the *Journal of the American Society of Agronomy* (*15*, 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, and 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 judgments 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

In prediction, two aspects of variable unreliability have consequences for ethical reasoning.

One is in estimating a person's true score on a variable;

the second is in how regression might be handled when there is measurement error in the independent and/or dependent variables.

In both of these instances, there is an implicit underlying 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.

Estimation of True Score

The estimation, $\hat{T}_X$, of a true score from an observed score, X, was derived using the regression model by Kelley in the 1920s, 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 the person is in relation to the group—upward if below the mean, downward if above.

This equation has been labeled "Kelley's Paradox."

In addition to obtaining a true score estimate from an obtained score, Kelly's regression model also provides a standard error of estimation (which in this case is now referred to as the standard error of measurement).

An approximate 95% confidence interval on an examinee's true score is given by

$$\hat{T}_X \pm 2\hat{\sigma}_X((\sqrt{1-\hat{\rho}})\sqrt{\hat{\rho}}) \ ,$$

where $\hat{\sigma}_X$ is the (estimated) standard deviation of the observed scores.

By itself, the term $\hat{\sigma}_X((\sqrt{1-\hat{\rho}})\sqrt{\hat{\rho}})$ is the standard error of measurement, and is generated from the usual regression formula for the standard error of estimation but applied to Kelly's model predicting true scores.

The standard error of measurement most commonly used in the literature is not Kelly's but rather $\hat{\sigma}_X\sqrt{1-\hat{\rho}}$, and a 95% confidence interval taken as the observed score plus or minus twice this standard error.

An argument can be made that this latter procedure leads to "reasonable limits" (after Gulliksen, 1950) whenever $\hat{\rho}$ is reasonably high, and the obtained score is not extremely deviant from the reference group mean.

Why we should assume these latter preconditions and not use the more appropriate procedure to begin with, reminds us of a Bertrand Russell quotation: "The method of postulating what we want has many advantages; they are the same as the advantages of theft over honest toil."

There are several remarkable connections between Kelley's work in the first third of the twentieth century and the modern theory of statistical estimation developed in the last half of the century.

In considering the model for an observed score, X, to be a sum of a true score, T, and an error score, E, plot the observed test scores on the x-axis and their true scores on the y-axis.

As noted by Galton in the 1880s, any such scatterplot suggests two regression lines:

One is of true score regressed on observed score (generating Kelley's true score estimation equation given in the text);

the second is the regression of observed score being regressed on true score (generating the use of an observed score to directly estimate the observed score).

Kelley clearly knew the importance for measurement theory of this distinction between two possible regression lines in a true-score versus observed-score scatterplot.

The quotation given below is from his 1927 text, *Interpretation of Educational Measurements*.

The reference to the "last section" is where the true score was estimated directly by the observed score; the "present section" refers to his true score regression estimator:

This tendency of the estimated true score to lie closer to the mean than the obtained score is the principle of regression. It was first discovered by Francis Galton and is a universal phenomenon in correlated data. We may now characterize the procedure of the last and present sections by saying that in the last section regression was not allowed for and in the present it is. If the reliability is very high, then there is little difference between [the two methods], so that this second technique, which is slightly the more laborious, is not demanded, but if the reliability is low, there is much difference in individual outcome, and the refined procedure is always to be used in making individual diagnoses.

Kelley's preference for the refined procedure when reliability is low (that is, for the regression estimate of true score) is due to the standard error of measurement being smaller (unless reliability is perfect); this is observable directly from the formulas given earlier.

There is a trade-off in moving to the regression estimator of the true score in that a smaller error in estimation is paid for by using an estimator that is now biased.

Such trade-offs are common in modern statistics in the use of "shrinkage" estimators (for example, ridge regression, empirical Bayes methods, James–Stein estimators).

Other psychometricians, however, apparently just don't buy the trade-off; for example, see Gulliksen (*Theory of Mental Tests*; 1950);

Gulliksen wrote that "no practical advantage is gained from using the regression equation to estimate true scores."

We disagree—who really cares about bias when a generally more accurate prediction strategy can be defined?

Stein's Paradox

What may be most remarkable about Kelley's regression estimate of true score is that it predates the work in the 1950s on "Stein's Paradox" that shook the foundations of mathematical statistics.

A readable general introduction to this whole statistical kerfuffle is the 1977 *Scientific American* article by Bradley Efron and Carl Morris, "Stein's Paradox in Statistics."

Keep in mind that the class referred to as James–Stein estimators (where bias is traded off for lower estimation error) includes Kelley's regression estimate of the true score.

We give an excerpt below from Stephen Stigler's 1988 Neyman Memorial Lecture, "A Galtonian Perspective on Shrinkage Estimators" that makes this historical connection explicit:

The use of least squares estimators for the adjustment of data of course goes back well into the previous century, as does Galton's more subtle idea that there are two regression lines. . . . Earlier in this century, regression was employed in educational psychology in a setting quite like that considered here. Truman Kelley developed models for ability which hypothesized that individuals had true scores . . . measured by fallible testing instruments to give observed scores . . . ; the observed scores could be improved as estimates of the true scores by allowing for the regression effect and shrinking toward the average, by a procedure quite similar to the Efron–Morris estimator.

Errors-in-Variables Modeling

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.

Such biasing may be particularly troubling when discussing econometric models where unit changes in observed variables are supposedly related to predicted changes in the dependent measure; possibly the unit changes are more desired at the level of the true scores.

Milton Friedman's 1992 article entitled "Do Old Fallacies Ever Die?" gives a downbeat conclusion regarding errors-in-variables modeling:

Similarly, in academic studies, the common practice is to regress a variable Y on a vector of variables X and then accept the regression coefficients as supposedly unbiased estimates of structural parameters, without recognizing that all variables are only proxies for the variables of real interest, if only because of measurement error, though generally also because of transitory factors that are peripheral to the subject under consideration. I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data, alleviated only occasionally by consideration of the bias introduced when "all variables are subject to error."

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.

Most discussions about fairness of selection are best phrased as 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, or age).

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

Generally, an understanding of how a regression/selection model works with this kind of variation is necessary for a numerically literate discussion of its intended or unintended consequences.

Interpreting and Making Inferences From Regression Weights

An all-too-common error in multivariable systems is to overinterpret the meaning of the obtained regression weights.

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.

For example, in applying regression models to argue for employment discrimination (such as in pay, promotion, or hiring), the multivariable system present could be problematic in arriving at a “correct” analysis.

Depending on the variables included, some variables may “act” for others (as “proxies”) or be used to hide (or at least, mitigate) various effects.

If a case for discrimination rests on the size of a coefficient for some polychotomous 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 relation to the polychotomous variable.

In short, based on how the regressions are performed and one's own (un)ethical predilections, different conclusions could be produced from what is essentially the same dataset.

In considering regression in econometric contexts, our interest is typically not in obtaining any deep understanding of the interrelations among the independent variables, or in the story that might be told.

The goal is usually more pragmatic and phrased in terms of predicting a variable reflecting value and characterized in some numerical way (for example, as in money or performance statistics).

The specific predictor variables used are of secondary importance; what is central is that they "do the job."

One recent example of success for quantitative modeling is documented by Michael Lewis in *Moneyball* (2003), with its focus on data-driven decision making in baseball.

Instead of relying on finding major league ball players using the hordes of fallible scouts visiting interminable high-school and college games, one adopts quantitative measures of performance, some developed by the quantitative guru of baseball, Bill James.

Moneyball relates the story of the Oakland Athletics and their general manager, Billy Beane, and how a successful team, even with a limited budget, could be built on the basis of statistical analysis and insight, and not on intuitive judgments from other baseball personnel (such as from coaches, scouts, or baseball writers).

Expert Judgment Versus Formula

A contentious aspect of using regression and other types of models to drive decision making arises when "experts" are overridden (or their assessments second-guessed and discounted, or their livelihoods threatened) by replacing their judgments with those provided by an equation.

One particularly entertaining example is in the prediction of wine quality in the Bordeaux or elsewhere.

Here, we have wine experts such as Robert Parker (of the *Wine Advocate*), pitted against econometricians such as Orley Ashenfelter (of Princeton).

One good place to start is with the *Chance* article by Ashenfelter, Ashmore, and LaLonde, “Bordeaux Wine Vintage Quality and the Weather.”.

As the article teaser states: “Statistical prediction of wine prices based on vintage growing-season characteristics produces consternation among wine ‘experts’.”

We also note an earlier article from the *New York Times* by Peter Passell (March 4, 1990), with the cute double-entendre title “Wine Equation Puts Some Noses Out of Joint.”

The (Un)reliability of Clinical Prediction

This last section on prediction concerns the (un)reliability of clinical (behavioral) prediction, particularly for violence, and notes two extensive redactions in the Appendix Supplements:

one is the majority opinion in the Supreme Court case of *Barefoot v. Estelle* (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.

Both of these documents are detailed, self-explanatory, and highly informative about our current lack of ability to make clinical assessments that lead to accurate and reliable predictions of future behavior.

As noted in the various opinions and *amicus* brief given in *Barefoot v. Estelle*, the jury in considering whether the death penalty should be imposed, has to answer affirmatively one question:

whether there was a probability that the defendant would commit criminal acts of violence that would constitute a continuing threat to society.

The use of the word "probability" without specifying any further size seems odd to say the least, but Texas courts have steadfastly refused to delimit it any further.

So, presumably a very small probability of future violence would be sufficient for execution if this small probability could be proved "beyond a reasonable doubt."

The point of much of this section has been to emphasize that actuarial evidence about future violence involving variables such as age, race, or sex, is all there really is in making such predictions.

More pointedly, the assignment of a clinical label, such as "sociopath," adds nothing to an ability to predict, and to suggest that it does is to use the worst "junk science," even though it may be routinely assumed true in the larger society.

All we have to rely on is the usual psychological adage that the best predictor of future behavior is past behavior.

Thus, the best predictor of criminal recidivism is a history of such behavior, and past violence suggests future violence.

The greater the amount of past criminal behavior or violence, the more likely that such future behavior or violence will occur (a behavioral form of a “dose-response” relationship).

At its basis, this is statistical evidence of such a likely occurrence and no medical or psychological diagnosis is needed or useful.

The Goldwater Rule

The offering of a professional psychiatric opinion about an individual without direct examination is an ethical violation of the Goldwater Rule, named for the Arizona Senator who ran for President in 1964 as a Republican.

Promulgated by the American Psychiatric Association in 1971, it delineated a set of requirements for communication with the media about the state of mind of individuals.

The Goldwater Rule was the result of a special September/October 1964 issue of *Fact:* magazine, published by the highly provocative Ralph Ginzburg.

The issue title was “The Unconscious of a Conservative: Special Issue on the Mind of Barry Goldwater,” and reported on a mail survey of 12,356 psychiatrists, of whom 2,417 responded: 24% said they did not know enough about Goldwater to answer the question; 27% said he was mentally fit; 49% said he was not.

Much was made of Goldwater’s “two nervous breakdowns,” because such a person should obviously never be President because of a risk of recurrence under stress that might then lead to pressing the nuclear button.

Goldwater brought a $2 million libel suit against *Fact:* and its publisher, Ginzburg.

In 1970 the United States Supreme Court decided in Goldwater's favor giving him $1 in compensatory damages and $75,000 in punitive damages.

More importantly, it set a legal precedent that changed medical ethics forever.

For an updated discussion of the Goldwater Rule, this time because of the many psychiatrists commenting on the psychological makeup of the former chief of the International Monetary Fund, Dominique Strauss-Kahn, after his arrest on sexual assault charges in New York, see Richard A. Friedman's article, "How a Telescopic Lens Muddles Psychiatric Insights" (*New York Times*, May 23, 2011).