How is Ethics Like Logistic Regression?

Ethics decisions, like statistical inferences, are informative only if they're not too easy or too hard

Andrew Gelman and David Madigan

Consider the analogy between ethical dilemmas and statistical classification problems. The mapping is not perfect—when we categorize a decision option as ethical or not (in its context), there is ultimately no "true answer," thus we can't really speak of correct and incorrect decisions, or of error rates. But there is a sense in which many ethical questions become clear in hindsight. And as with prediction, some ethical decisions are clear-cut (for most of us) while others are tough calls. Perhaps some insights from the statistics of classification can help us better understand the study of ethics.

Election Forecasting

Consider a binary prediction problem of some outcome that occurs roughly half the time, fit to data using a logistic regression. For the data points where the predicted probability is near 50%, you learn almost nothing from the data, as the outcome is essentially pure chance. And for the points where the predicted probability is near 0 or 1, you also learn almost nothing, because you knew what to expect from the model alone. You learn more from the intermediate cases.

For example, various historians and political scientists have come up with methods for forecasting presidential elections, and have tested their models on past data. It is natural to evaluate such a method by counting how often the winner was correctly predicted, but such an evaluation is statistically inefficient.

In the past 60 years, there have been four elections that have been essentially tied in the final vote: 1960, 1968, 1976, and 2000. (You could throw 2004 in there too.) It's meaningless to say that a forecasting method predicts the winner correctly (or incorrectly) in these cases. And from a statistical point of view, you don't want to adapt your model to fit these tossups—it's just an invitation to over-fitting.

It's fine to make predictions for these elections, but after the fact essentially no information is provided by learning that the point prediction was correct or incorrect. For example, suppose that a particular method mis-predicted 1960, 1968, and 2000. Would we think any less of such a method? No. A method that predicts vote share (such as used by political scientists) could get credit from these close elections by predicting the vote share with high accuracy, but you should not get credit for correctly predicting the outcome of a coin flip.

From the other direction, landslide elections such as 1964 and 1984 are so easy to predict that any reasonable model should forecast them directly, and thus these predictions have no great discrimination power.

The statistical point is that it is better to predict a continuous measurement (in this case, the vote differential) than a discretization (the winner of the election); but it is counter to a natural intuition that we should be modeling what we ultimately care about. If the objective is to learn about wins, maybe we should predict wins directly. To which I reply, sure, predict the winner, but it will be more statistically efficient to do this in a two-stage process: first predict vote differential, then predict the winner given vote differential (not quite a trivial process in America's Electoral College system, but it can be done). The key is that vote differential is available, and performing a logistic regression for wins alone is implicitly taking this differential as latent or missing data, thus throwing away information.

Similar problems arise in sports: when predicting basketball games, don't model the probability of wins, model the expected score differentials. Sure, what you really want to know is who wins. But the most efficient way to get there is to model the score differential, and then map that back to win probabilities. Similarly in baseball: as the great Bill James wrote, if you want to predict a pitcher's win-loss record, it's better to use last year's earned run average than last year's won-lost record. The won-lost record is a noisy discretization.

Educational Testing

In other settings, no underlying continuous measurement is available. Consider, for example, grading on a multiple-choice test. If a question is so hard that almost everyone is simply guessing at random, then it provides very little information about the test-takers' abilities. Even if a student happens to get this sort of question correct, it is likely that the correct answer is just a guess, and it is not statistically appropriate to give the student full credit for the answer. From the other direction, if a question is so easy that almost everyone gets it correct, then this item provides information only about the very few students who happen to get it wrong.

For the purpose of evaluation, then, the most useful test items are those that are neither too easy nor too hard. Or, to give a fuller picture, the ideal test will have a mix of items of varying difficulty so as to be able to discriminate among a wide range of abilities. But when considering items one at a time, the most generally informative are those of intermediate difficulty.

It is a saying in statistics that any good idea first appeared in psychometrics 50 years earlier. And the ideas of psychological measurement are also relevant to ethics. In particular, ethics is all about drawing the line between acceptable and unacceptable behavior. It seems universal that ethics is taught and understood via case studies, which are, perhaps, comparable to items on a test. To continue the analogy, the different people facing ethical choices correspond to different test-takers in different situations. And the attempts to come to general ethical principles based on the study of special cases can be analogized to item-response models or support-vector machines that separate ethical from unethical behaviors.

Ethical Dilemmas

One challenge in writing about or teaching ethics is a temptation to focus on difficult ethical dilemmas. As the lawyers say, though, hard cases make bad law. That saying means a lot of different things, but in this case our point is that if you focus on the toughest calls, you can end up implicitly sending the message that ethics are completely arbitrary, and that any decision can be ethically justified.

At the other extreme, there's no point in teaching ethics based on easy cases. It is not interesting to discuss, for example, the ethical dilemma of a comfortable middle class citizen who chooses to take up armed robbery just for fun, or a gossip who spreads false stories about his friends just to watch people's reactions.

When we do discuss very hard or very easy ethics examples, it is to make particular points. For example, one might start with an extreme case (that is, an easy one) just to establish a baseline, to flush out any hard-line moral relativists before proceeding with discussion. Or one might consider a tough call to set a boundary on the other end, to make the point that there is, necessarily, individual variation on where to draw the ethical line: ethics can never be an exact science.

In general, though, the most informative ethics vignettes are those in which the call is not so close as to seem arbitrary, but not so obvious that the decision can be made without thought. The purpose of discussing the intermediate cases is to explore the way in which careful assessment of goals, motivations, costs, and benefits can help us make better decisions, and also help us understand the decisions of others.

Ethical Dilemmas in Pharmaceutical Research

Pharmaceutical research represents a very live area of ethical discussion in statistics, science, and public policy, and there are compelling stories on all sides of the controversies involving the drug-approval process, patents, pricing, and the like. From a statistical perspective, some of the most interesting ethical concerns come as products of the so-called moral hazard involved in the high stakes of dollars, health, politics, and scientific prestige. In particular, double-blind randomized experiments are considered a gold standard, but in real life there exists evidence of persistent threats to the validity of such experiments—for example, via protocol modification after partial data become available, or by failure to report trials with negative findings. Further inherent difficulties include patients being able to guess which treatment they are getting, and non-adherence by patients for whom the treatment is ineffective. The placebo effect also presents a challenge in that the most honest description of a treatment may be expected to yield a less than effective outcome, and so what then is the most ethical choice?

Imperfections in data collection and analysis have been the subject of statistical research for many years, but the questions are not merely technical. They link up to ethical dilemmas in several ways, most directly when dealing with researchers who are willing to break the rules in order to get a desired positive result, or in a softer way when considering which approximate analysis method to use (recognizing that no exact approach is possible that will handle the complexities of real data). A detailed study-protocol and data-analysis plan crafted before the study begins can mitigate some of these concerns. Standard practice, however, allows for much of the data-processing analysis choices to be crafted long after the study has begun and data have accumulated, but before the un-blinding of the treatment assignments to the researchers. Since even blinded data can give hints about the outcome, this practice seems risky. For observational studies in healthcare, pre-specified protocols are the exception rather than the rule, and all kinds of unethical practices are certainly possible.

Well-documented evidence points to a strong connection between study funder and study outcome, presumably as a result of the kinds of concerns we just described. In the United States almost all randomized pharmaceutical experiments are paid for (and in many cases run by) the manufacturers. The conflict of interest is obvious and of grave concern, but given the extraordinary cost of such experiments, no practical solution to this problem is apparent.

Drug safety presents its own ethical dilemmas. For example, some years ago anecdotal evidence raised the possibility that long-term use of statins could be associated with adverse mental effects. Does a statin manufacturer have an obligation to communicate this low-quality evidence? Should physicians tell their patients? What if the patients stop taking statins? Have we empowered patients to make sensible risk-benefit tradeoffs in this kind of situation? The statistical tools of decision analysis should enable us to make useful recommendations here, but as statisticians it is hard for us to get to that point, given the challenges we have in addressing our own ethical issues.

Perhaps it makes sense to start by considering the principle introduced at the start of this article, that we gain the most information from cases that are neither too easy nor too hard. In the context of ethics in statistics, this might be settings where there is no outright fraud, but where financial and professional incentives can motivate statisticians and other quantitative researchers to present misleading results. From the standpoint of statistical theory, the issue is that theoretical evaluations typically assume that whatever method is being considered will be applied universally, whereas in practice a researcher typically has a choice of methods, and a choice of options within any method. So we offer no easy answers here, merely suggestions of directions for further study. ■

Further Reading

Bekelman, J. E., Y. Li, and C.P. Gross, 2003. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. *Journal of the American Medical Association* 289(4): 454-465.

Lexchin, J., L.A. Bero, B. Djulbegovic, and O. Clark, 2003. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *British Medical Journal* 326(7400): 1167-1170.

U.S. Food and Drug Administration. 2014. FDA expands advice on statin risks. *www.fda.gov/ForConsumers/ConsumerUpdates/ucm293330.htm*

Yaphe, J., R. Edman, B. Knishkowy, and J. Herman, 2001. The association between funding by commercial interests and study outcome in randomized controlled drug trials. *Family Practice* 18(6): 565-568.

About the Authors

Andrew Gelman is a professor of statistics and political science at Columbia University. His books include *Bayesian Data Analysis, Teaching Statistics: A Bag of Tricks,* and *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do.*

David Madigan is a professor of statistics and dean of arts and sciences at Columbia University. His research areas include Bayesian statistics, text mining, Monte Carlo methods, pharmacovigilance and probabilistic graphical models.