Teaching Bayes to Graduate Students in Political Science, Sociology, Public Health, Education, Economics, . . .

Andrew GELMAN

1. INTRODUCTION

I was trying to draw Bert and Ernie the other day, and it was really difficult. I had pictures of them right next to me, but my drawings were just incredibly crude, more "linguistic" than "visual" in the sense that I was portraying key aspects of Bert and Ernie but in pictures that didn't look anything like them. I know that drawing is difficult—every once in awhile, I sit for an hour to draw a scene, and it's always a lot of work to get it to look anything like what I'm seeing—but I didn't realize it would be so hard to draw *cartoon characters*!

This got me to thinking about the students in my statistics classes. When I ask them to sketch a scatterplot of data, or to plot some function, they can never draw a realistic-looking picture. Their density functions don't go to zero in the tails, the scatter in their scatterplots does not match their standard deviations, $E(y|x)$ does not equal their regression line, and so forth. For example, when asked to draw a potential scatterplot of earnings versus height, they have difficulty with the x-axis (most people are between 60 and 75 inches in height) and having the data consistent with the regression line, while having all earnings be nonnegative. (Yes, it's better to model on the log scale or whatever, but that's not the point of this exercise.)

Anyway, the students just can't make these graphs look right, which has always frustrated me. But my Bert and Ernie experience suggests that I'm thinking of it the wrong way. Maybe they need lots and lots of practice before they can draw realistic functions and scatterplots. They'll certainly need lots of practice to learn Bayesian methods.

2. TEACHING BAYESIAN STATISTICS TO SOCIAL SCIENTISTS, INCLUDING A DISCUSSION OF WHAT IS BAYESIAN ABOUT MAKING GRAPHS TO GET A BETTER UNDERSTANDING OF THE DETERMINISTIC PART OF A MODEL

I teach Bayesian statistics to nonstatisticians in two settings.

Sometimes I teach a course in Bayesian statistics and computation. This class attracts all sorts of graduate students and postdocs on campus who have heard about Bayesian methods or are using Bayesian methods but don't have a foundation in the topic. This semester I had a couple of biology postdocs studying spatial patterns of animals and trees; they were already using Bugs and R, but they wanted to know more about the actual models they were fitting. These sorts of students often come in with very specific questions about convergence of Markov chain simulations but sometimes get the most out of learning some of the basics, such as where do likelihoods and priors come from and what is random simulation. (When I teach computing, I introduce simulation in several steps: first simulation of stochastic processes such as simple queues, then simple simulation of regression inferences using the multivariate normal distribution defined by the estimate and covariance matrix, then finally simulating from posterior distributions.)

The other way I teach Bayesian statistics is in an introductory graduate course on applied regression and multilevel models. Here we get a lot of students from all departments, especially the social sciences, who want to fit and understand their models beyond what they'll get from Stata output. These students, and also those who take the more advanced Bayesian class, tend to be highly motivated because they are trying to solve real problems in their applied fields. Because I'm 50% in the political science department, I get some of the top Ph.D. students in political science taking my class. Unfortunately, I'm afraid that the top quantitative students in sociology, economics, and public health do not take this sort of applied class, instead following the fallacy that the most mathematical courses are the best for them. (I know about this fallacy—it was the attitude that I had as an undergraduate, until I stumbled into an applied statistics class as a senior and realized that this stuff was interesting and difficult.)

My applied regression and multilevel modeling class has no derivatives and no integrals—it actually has less math than a standard regression class, since I also avoid matrix algebra as much as possible! What it does have is programming, and this is an area where many of the students need lots of practice. The course is Bayesian in that all inference is implicitly about the posterior distribution. There are no null hypotheses and alternative hypotheses, no Type 1 and Type 2 errors, no rejection regions and confidence coverage. (To expand upon this a bit: we do talk about statistical significance, but we frame it in terms of uncertainty about particular inferences, rather than in terms of error rates.)

Instead, the course is all about models, understanding the models, estimating parameters in the models, and making predictions. Prediction is a great topic because (a) it's clearly useful in many contexts, from business to medicine to politics; (b) it can be directly implemented using simulation and without the complexities of Gibbs sampling etc.; and (c) it's a great way to operationalize multilevel models. You can make predictions for new observations in existing groups or new observations in new groups, and you can see the different variance components pop out.

Andrew Gelman is Professor, Department of Statistics and Department of Political Science, Columbia University (E-mail: *gelman@stat.columbia.edu*). The author thanks the National Science Foundation, the National Institutes of Health, and the Columbia University Applied Statistics Center for financial support, Peter Westfall for inviting this article and supplying helpful comments, and many students over the years for their questions and suggestions.

  DOI: 10.1198/000313008X330829

Beyond programming and simulation, probably the Number 1 message I send in my applied statistics class is to focus on the deterministic part of the model rather than the error term. The error term is important—no doubt about it—and we do lots of simulations and interval estimates to explain that the point estimate is not the whole story. And I have examples of where people make hasty conclusions from small sample sizes. But, but, ... I think that understanding the model comes first. Even a simple model such as $y = a + bx + error$ is not so simple if x is not centered near zero. And then there are interaction models—these are incredibly important and so hard to understand until you've drawn some lines on paper. We draw lots of these lines, by hand and on the computer. I think of this as Bayesian as well: Bayesian inference is conditional on the model, so you have to understand what the model is saying. Arguably this is less of a concern in classical statistics and econometrics, where there is an emphasis on getting estimates with good statistical properties even if model assumptions are violated. I have no problem with students learning this in other courses, but in my class I want them to know what their models are doing. The students seem to like it, but my observations are surely subject to selection bias, and probably the students who take other sorts of statistics classes get a lot out of those other approaches in their own way. One positive thing I've noticed is that more and more of our political science students are presenting their regression results graphically. This is standard practice in medical and public health statistics, and it's good to see it moving into political science. Maybe the economists and statisticians will be next.

3. OTHER THOUGHTS ON TEACHING STATISTICS TO NONSTATISTICIANS

Different instructors have different styles. After 20 years of teaching, I've come to the conclusion that teaching skills works better than teaching concepts (or, should I say, trying to teach concepts). This is related to the fundamental insight that you can't "cover" material in a course; students ultimately have to teach themselves how to do things. No easy answers here, but I can certainly believe there are better and worse ways to proceed. I've found that the teaching tricks that work well with undergraduates (in particular, frequently stopping the class and having students work together in pairs) work with graduate students as well. In general, I think teaching works best when you have a good script to follow—not just a good textbook, but an hour-by-hour lesson plan. Unfortunately, these are hard to come by in statistics. I guess I should prepare such things based on my own textbooks.

The more general principle is that just about any teaching method (or, for that matter, research method) can work well, as long as (a) you put in the effort to do it right, and (b) keep in mind the ultimate goal, which is for the students to have certain skills and certain experiences by the time the class is over. Related to these is (c) the method should be appropriate for your own teaching style. Even old-fashioned blackboard lecturing is fine—if you can pull it off in a way that keeps the students' brains engaged while you're doing it. I developed a more active teaching style for myself (Gelman and Nolan 2002) because that was the only way I could keep the students thinking, and I give this advice to our instructors in training (Gelman 2005).

Let's also not forget the benefit of the occasional dumb but fun example. For example, I came across the following passage in a *New York Times* article: "By the early 2000s, Whitestone was again filling up with young families eager to make homes for themselves on its quiet, leafy streets. But prices had soared. In October 2005, the Sheas sold the house, for which they had paid $28,000 nearly 40 years ago, for more than $600,000." They forgot to divide by the Consumer Price Index! Silly but, yes, these things happen, and it's good to remind social science students that if they know about these simple operations, they're already ahead of the game. The next step is to discuss more difficult problems such as adjusting the CPI for quality improvements. (For example, if the average home today is larger than the average home 40 years ago, should the CPI adjustment be per home or per square foot?) I also like to mention points such as, "The difference between 'significant' and 'nonsignificant' is not itself statistically significant" (Gelman and Stern 2006). But I haven't surmounted the challenge of how to fit this sort of "good advice" into a coherent course so that students have a sense of how to apply these ideas in new problems.

4. A CASE STUDY: THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

The hardest thing to teach in any introductory statistics course is the sampling distribution of the sample mean, a topic that is at the center of the typical intro-stat-class-for-nonmajors. All of probability theory builds up to it, and then this sample mean is used over and over again for inferences for averages, paired and unpaired differences, and regression. This is the standard sequence, as in the books by Moore and McCabe (1993), and De Veaux, Velleman, and Bock (2004). The trouble is, most students don't understand it. I'm not talking about proving the law of large numbers or central limit theorem—these classes barely use algebra and certainly don't attempt rigorous proofs. No, I'm talking about the derivations that lead to the sample mean of an average of independent, identical measurements having a distribution with mean equal to the population mean, and sd equal to the sd of an individual measurement, divided by the square root of n.

This is key, but students typically don't understand the derivation, don't see the point of the result, and can't understand it when it gets applied to examples.

What to do about this? I've tried teaching it really carefully, devoting more time to it, and so on—nothing works. So here's my proposed solution: deemphasize it. I'll still teach the sampling distribution of the sample mean, but now just as one of many topics, rather than the central topic of the course. In particular, I will not treat statistical inference for averages, differences, and so on, as special cases or applications of the general idea of the sampling distribution of the sample mean. Instead, I'll teach each inferential topic on its own, with its own formula and derivation. They still mostly won't follow the derivations, but then at least if they're stuck on one of them, it won't muck up their understanding of everything else.

5. STARTING AN (IMPLICITY) BAYESIAN APPLIED REGRESSION COURSE: TWO WEEKS OF CLASSROOM ACTIVITIES

I conclude with a discussion of how I begin my applied regression course—the first two weeks of activities in the classroom. The class is based on Gelman and Hill (2007). In the first few weeks, we give the class a chance to get familiar with R, which they'll have to use more of when working with more complicated models—especially when trying to use inferences beyond simply looking at parameter estimates and standard errors. The first two homework assignments involve fitting simple regressions in R, graphing the data and the fitted regression lines, and building a multiple regression to fit the beauty and teaching evaluations data of Daniel Hamermesh (2005). The T.A. has to spend a lot of time helping students get started with R. The main mathematical difficulties are learning and understanding linear and logarithmic transformations.

Now let's move to the classroom. I'll mention lots of examples without giving all the details here, just to give a sense of how the class feels.

Lecture 1 starts with some motivating examples, including roaches, rodents, and red and blue states. I stop and give the students a few minutes to work in pairs to come up with explanations for the patterns of income and voting within and between states. I describe two studies I worked on with the New York City Department of Health—one study was about roaches and one was about rodents—and then give the students a minute to discuss in pairs to see if they can figure out the key difference between the two studies. (The difference is that the roach study has the goal of estimating a treatment effect—integrated pest management compared to usual practice—and the rodent study is descriptive—to understand the differences between rodent levels in apartments occupied by whites, blacks, hispanics, and others. We return to causal inference later in the semester.) I yammer on a bit about the skills they'll learn by the time the course is over, and how I expect them to teach themselves these skills. Analogies between statistics and child care, sports, and policy analysis. The beauty and teaching evaluations example. I give the equation of the regression line, the students have to work in pairs to draw it. Use the computer to fit some regressions in R and plot the data and fitted regression lines. (No residual plot for now, no q-q plot: we're focusing on the important things first.)

Lecture 2 starts with the cancer-rate example. I hand out Figure 2.7 from our Bayesian book (Gelman et al. 2003) and give the students a few minutes to work in pairs to come up with explanations for why the 10% of counties with highest kidney-cancer deaths are mostly in the middle of the country. I write various explanations on the blackboard and then hand out Figure 2.8. We discuss: this is a motivator for multilevel models. I then give them some regression lines and scatterplots to draw—see Section 1 of this article for a discussion of how students have difficulties with this sort of activity. We talk transformations for a bit—some more activities in pairs (e.g., what's the equation of the regression line if we first normalize x by subtracting its mean and dividing by its sd). Discussion of appropriate scale of the measurements and how much to round off. Comparisons of men to women: adding sex into the regression model. In pairs: What's the difference in earnings between the average man and the average woman (it's not the coefficient for sex, since the two sexes differ in height)? Why it's better to create a variable called "male" than one called "sex."

Lecture 3 starts with answering some questions brought in by students. What are outliers and should we care about them? (My answer: outliers are overrated as a topic of interest.) Why is it helpful to standardize input variables before including interactions? A long discussion ensues using the earnings, height, and sex example. Standardize earnings by subtracting mean and dividing by 2*sd. Standardize sex by recoding as male $= 1/2$, female $= -1/2$. There is lots of working in pairs, drawing regression lines and figuring out regression slopes. Understanding coefficients of main effects and interactions. Categorized predictors—for example, modeling age as continuous, with quadratic term, using discrete categories. Start talking about the logarithm. The amoebas example—at time 1, there is 1 amoeba; at time 2, 2 amoebas, at time 3, 4 amoebas; and so on. In pairs: give the equation of number of amoebas as a function of time. Then give the linear relation on the log scale. (I should have had this example starting at time 0. Having to subtract time $=$ 1 is a distraction that the students didn't need.) Graph of world population versus time since year 1, graph on log scale. Interpreting exponential growth as a certain percentage per year, per 100 years (in pairs again).

Lecture 4 is all about logarithms. On the blackboard I give the equation for a cube's volume V as a function of its length L. Then also $\log V = 3 \log L$. Then, in pairs, they have to figure out the corresponding formulas for surface area S as a function of volume. It's not so easy for students who haven't used the log in awhile. Then we discuss the example of metabolic rate and body mass of animals. We then go to interpreting log regression models. Log earnings versus height. Log earnings versus log height. Interpreting log-regression coefficients as multiplicative factors (if the coefficient is 0.20, then a one-unit difference in x corresponds to an approximate 20% difference in y). Interpreting log-log coefficients as elasticities (if the coefficient is 0.6, then a 1% increase in x corresponds to an approximate 0.6% increase in y). All these are special cases of transformations. We also discuss indicator variables, combinations of inputs, and model building. How to interpret statistical significance of regression coefficients.

We haven't got to Bayes yet. In our applied regression course, we introduce Bayes by stealth, first simulating from the posterior distribution obtained by point estimates and standard errors, then showing what happens when we throw in a prior distribution. In our Bayesian data analysis course, we present the material more formally and then loop back and show how it reduces to classical estimation as a special case.

6. HOW IS THERE TIME, IN A COURSE WITH CLASS PARTICIPATION, TO COVER ALL THE MATERIAL?

People have often told me that they'd like to do group activities but they can't spare the class time. I disagree with that line of thinking. My impression is that students learn by prac-

ticing. A lecture can be good because it gives students a template for their own analyses, or because it motivates students to learn the material (e.g., by demonstrating interesting applications or counterintuitive results), or by giving students tips on how to navigate the material (e.g., telling them what sections in the book are important and what they can skip, helping them prepare for homework and exams, etc.). The lecture room also can be a great way to answer questions, since when one student has a question, others often have similar questions, and the feedback is helpful as the class continues.

But I don't see the gain in "covering" material. I don't need to do everything in lecture. It's in the book, and they're only going to learn it if it's in the homework and exams anyway. The class-participation activities allow the students to confront their problem-solving difficulties in an open setting, where I can give them immediate feedback and help them develop their skills. And having them work in pairs keeps all of them (well, most of them) focused during the class period.

REFERENCES

De Veaux, R. D., Velleman, P. F., and Bock, D. E. (2004), *Stats: Data and Models*, Boston: Addison-Wesley.

Gelman, A. (2005), "A Course on Teaching Statistics at the University Level," *The American Statistician*, 59, 4–7.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.) London: CRC Press.

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press.

Gelman, A., and Nolan, D. (2002), *Teaching Statistics: A Bag of Tricks*, Cambridge, MA: Oxford University Press.

Gelman, A., and Stern, H. S. (2006), "The Difference Between 'Significant' and 'Nonsignificant' is not Itself Statistically Significant," *The American Statistician*, 60, 328–331.

Hamermesh, D. (2005), "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity," *Economics of Education Review*, 24, 369–376.

Moore, D. S., and McCabe, G. P. (1993), *An Introduction to the Practice of Statistics*, New York: W. H. Freeman.