PREFACE TO BERK

JAN DE LEEUW

It is pleasure to write a preface for the book ”Regression Analysis” of my fellow series editor Dick Berk. And it is a pleasure in particular because the book is about regression analysis, the most popular and the most fundamental technique in applied statistics. And because it is critical of the way regression analysis is used in the sciences, in particular in the social and behavioral sciences. Although the book can be read as an introduction to regression analysis, it can also be read as a thorough critique of this class of techniques, or at least of some of the methodological superstructure that has been build on top of it. The subtitle, a constructive critique, is appropriate enough if one interprets it as applying to regression analysis in general, but the book does some pretty destructive work in the area of regressing modeling and regression inference. That’s another reason why it is a pleasure to write this preface. I like debunking, and I wholeheartedly agree with Berk that there is plenty to debunk in this area.

Let me add my own perspective, which is not very different from that of Berk, but which perhaps emphasizes other aspects of the methodological situation. Regression analysis is by far the most frequently used data analysis technique. It dominates data analysis in the social, behavioral, educational, environmental, and biomedical sciences, and it features prominently in policy studies, court cases, and various types of interventions. It is a multi-functional technique, used in many different situations and for many different purposes. Let us try to disentangle some of its uses.

1. DESCRIPTION

In the first place, regression analysis is used for description. As explained extensively in this book, regression analysis is used to describe the distribution of a variable under a number of different conditions. These conditions

are usually defined as combinations of the values of a number of other variables, and we often say we are studying the conditional distribution of an outcome variable given the values of a number of regressors. Thus, for instance, we look at the distribution of school achievement for students with various combinations of gender, race, and income. Specifically, we look at the the eight distributions of SAT scores for students that are Male/Female, Black/White, and Rich/Poor, in all possible combinations. Regression analysis is a data analysis technique to present the differences in these conditional distributions in a clear and convincing way.

This definition is very general. In our example we could, for instance, present eight ”parallel” histograms or boxplots, and we have an example of regression analysis. But this only works in the example, because we do not have too many ”cells” in our ”design”. We only have eight possible combinations, and presumably enough observations in each of the eight cells to draw the histograms or boxplots. But what happens if we only have a small number of observations, for instance only fifty. On the average, there will be about six observations in each cell, not enough for a decent histogram. What we will tend to do, in that case, is to apply smoothing. Our regression analysis could take the form of plotting eight parallel normal densities, with different means and variances. This is a smoothed version of the parallel histograms, and it provides us with a cleaner picture. It also gives us the opportunity to summarize our data analysis in sixteen numbers (eight means, eight variances), which is a pretty concise data reduction summary.

The problem with this form of smoothing the histograms is clear. The device of using the different normals may or may not be appropriate in a particular empirical situation. It is never false, because it is just a graphical or analytical device, but it may be wasteful because it throws away much interesting information, or misleading because it distorts information. Maybe the conditional distributions are skewed, for instance, or maybe they have very heavy tails.

We get into more serious problems if the number of regressors, and thus the number of cells, increases. Ten variables, with five values each, produce about ten million cells, and we will not have enough observations to fill the

cells. Most of them will be empty, the other will have one or two observations. We cannot compute cell variances any more, and thus we have to resort to heavier smoothing. We use normal densities which all have the same variance, and are merely shifted along the real axis. And even that may not be enough, because we may not be able to compute cell means reliably. Additional smoothing is introduced by requiring our cell means to be linear combinations of our five variables, and thus we reduce the number of parameters in the graphical representation from ten million means (and a variance) to five regression coefficients (and a variance). Of course we are not actually going to graph the ten million normal densities, we just summarize the data analysis using the six parameters. A gigantic amount of data reduction, and plenty of possibilities to distort and/or to smooth away interesting aspects of the data.

As Berk explains in his book, using regression analysis for description cannot really be criticized in general methodological terms. It can definitely be criticized in any specific situation, because the actual device that is chosen may be misleading or wasteful, it may oversmooth and undersmooth. But comparing conditional or cell distributions is a basic technique in many of the sciences, and in the case of small samples or many cells we have do use data reduction or smoothing devices of some sort. What we criticize in this context is a lack of craftsmanship, or in some cases even a fraudulent use of the available tools.

2. PREDICTION

Regression analysis is also used for prediction. Imagine the following situation. Parents want to enroll their child in one of a number of high schools, and they want to do this in such a way that the child is most likely to be admitted to a particular university after six years. They go to a counselor, and they give the counselor all kinds of information about the child. The counselor plugs these data, together with data about the high schools, into a regression equation she keeps in a drawer, and the regression equation comes up with estimates of admission probabilities. The counselor then suggests to the parents to choose the high school with produces the highest

admission probability and collect the counseling fee. Similar scenarios can be constructed for stock-market brokers, economic macro-model predictors, clinical psychologists, and so on.

It seems clear that using regression analysis to construct prediction devices again cannot be criticized in general methodological terms. They either work, or they don't work. If they don't work, the counselor will go out of business, and a competitor with a better regression equation takes over. You can criticize a device by inventing a better one, but you cannot say the equation in the counselor's drawer is "false". That does not make sense.

There are more complicated forms of prediction, but basically the same methodological considerations apply. If we have ten million cells, and a smoothed description, then we can give values to the empty cells. Although we have no observations in the empty cells, we can still use our regression coefficients to make the linear combination of the variable values corresponding with the cell. We can predict the value of an observation in that cell, and then go out into the world, collect such an observation, and compare with out prediction. Or someone else may come up with such an observation. If our interpolation is wrong, we have to adjust our smoothing device, because otherwise we clearly are at a competitive disadvantage.

In linear regression models, prediction also occurs in other forms. If we have used gender, race, and income as our predictors, and we have computed regression coefficients, then we can say "increasing family income by $ 100 will result in so and so many additional points on the SAT". This refers to a possible experiment, and it predicts the outcome of that experiment. And it is useful in so far as such an experiment is feasible, and will actually be undertaken. If it is just a hypothetical experiment, then predicting its outcome is not very interesting. We could all sit at our desks and perform hypothetical experiments in our heads all day, and science would not advance one iota. Berk shows in this book that these types of predictions, often presented in the form of explanations, are very common in the social sciences, and are quite useless. The corresponding experiments can never be carried out, and thus the predictions are not really predictions. They are just another form of "idling of the machine", in this case the social

science regression machine, with all the fancy LISREL bells and whistles huffing and puffing.

3. INFERENCE

Regression analysis is also used as a statistical technique to make inferences from a sample to a population from which the sample is drawn. As Berk explains, this particular approach gets us into trouble right away in many situations. We are dealing with a sampling model and with a regression model on top of that. Often the sampling model cannot be plausibly defended. There is no sample, or the sample is not random, or the whole notion of a sample does not make sense. And even if the notion does make sense, the assumption that we are dealing with a simple random sample cannot be falsified in any conceivable way. We are specifying, at least if we are frequentists, what would happen if we repeated our experiment or our data gathering procedure a large number of times. But we have no way to actually replicate, so the sampling model, or the replication framework, is just a leap of faith.

And even if we would perform all these hypothetical replications, for instance in a thought experiment, it would be abundantly clear that the linear model that sits on top of the sampling model would be false. One could argue that maybe it would only be a little bit false, and still useful, but since we are talking about hypothetical replications in our head anyway, that sort of discussion is somewhat grotesque.

The standard statistical way of doing inference is to assume our observations are realizations of random variables. This means that we make up a framework of hypothetical replications, and all our subsequent statistical statements are about this hypothetical framework. To use Bishop Berkeley's apt phrase, we are talking continuously about "ghosts of departed quantities". As long as we are tossing coins, or drawing random samples from well-defined finite populations, we at least have a plausible replication framework. But these situations are rare, and it is far more common to have an implausible linear model on top of an equally implausible sampling model.

Berk quotes Box, who said that all models are false but some models are useful. Many statisticians are now familiar with this quote, but they still happily go about their business, which is to make various statements about random variables that are all conditional on the assumption that the model describing them is true. In the meantime they have thoroughly lost the connection with the analysis of the actual data. I think the only way out of this dilemma has to be a radical break with the idea that statistics is about models. Statistics is about techniques for describing data. In many cases it is useful to study the properties of techniques by applying them to models formulated in terms of hypothetical random variables, but this is just one way of validating the techniques, one form of quality control. Sometimes models are even useful to suggest techniques, by applying some general principle such as maximum likelihood or posterior mode, but such a technique still must be tested out and validated in actual data analysis situations.

We see that our discussing of regression analysis arrives at basically the same conclusion as this interesting book. Regression analysis is an eminently useful, and quite indispensable, statistical technique that can be used both for description of a large variety of data sets and for prediction of outcomes in many situations. As with all statistical techniques, its inferential aspects are problematical, even in simple situations. And, as Berk's book illustrates in great detail in the later chapters, these fundamental problems cannot be solved by using more complicated models that introduce hosts of additional parameters to "save the phenomena".

It was a pleasure for me to read this book. I see it as a critique of quantitative social science, which often takes the unholy route of forcing the data to serve an obviously silly model, but more generally as a critique of all those statistical methods and publications that concentrate on the statistical model and use the data merely as an afterthought. It is good to have this book in our series.