J Quant Criminol (2010) 26:481–487
DOI 10.1007/s10940-010-9116-4

ORIGINAL PAPER

What You Can and Can't Properly Do with Regression

Richard Berk

Published online: 29 September 2010

Introduction

Regression analysis, broadly construed, has over the past 60 years become the dominant statistical paradigm within the social sciences and criminology. In its most canonical and popular form, a regression analysis becomes a "structural equation model" from which "causal effects" can be estimated. Consider some examples from the two most recent issues of *Criminology*. Volume 47, Issue 4, has ten articles, seven of which employ some form of casual modeling. One exception (Mears and Bales 2009) estimates causal effects using matching, another exception (Schultz and Tabanico 2009) estimates causal effects using randomized experiments, and the final exception (Guerette and Bowers 2009) estimates causal effects using meta-analysis.[1] Volume 48, Issue 1, has nine articles of which eight employ some form of causal modeling. The one exception (Skardhamar 2010) is a useful critique of a causal modeling special case: "semiparametric group-based modeling."[2]

The many problems with causal modeling are well articulated in a very large literature (e.g., Box 1976; Leamer 1978; 1983; Lieberson 1985; Holland 1986; Rubin 1986; Freedman 1987, 2005; Kish, 1987; de Leeuw 1994; Manski 1995; Pearl 2000; Heckman 2000; Breiman 2001; Berk 2004; Imbens 2009). Perhaps too simply put, before the data are analyzed, one must have a causal model that is nearly right.[3] With the model specified, all that remains is to estimate the values of the regression parameters. In practice, one or two

[1] Of course, much the research on which Guerette and Bowers rely employs causal modeling. For a discussion of well-known, but typically ignored, problems with meta-analysis, see Berk (2007).

[2] See Berk et al. (2010) for a more formal and general discussion.

[3] A model is the "right" model when it accurately represents how the data on hand were generated. A detailed discussion can be found in many formal expositions of causal modeling (e.g., Freedman 2005).

R. Berk (✉)
Department of Statistics, University of Pennsylvania, Philadelphia, PA, USA
e-mail: berkr@sas.upenn.edu

R. Berk
Department of Criminology, University of Pennsylvania, Philadelphia, PA, USA

small flaws can sometimes be tolerated as long as they are easily identified and corrected with the data on hand. It is very difficult to find empirical research demonstrably based on nearly right models.

In the absence of a nearly right model, the many desirable statistical properties of a causal model can be badly compromised. Omitted variables, for instance, can badly bias parameter estimates even in very large samples. Alternatively, using the data to inductively arrive at a nearly right model assumes that all of the required regressors are included in the data set. Even if they are (and how would you know?), badly biased estimates can result because the model and the parameter estimates are extracted from the same data (Leeb and Pötscher 2005, 2006; Berk et al. 2010a).

There are five kinds of responses to the causal modeling critique. The first is rhetorical and is by far the least constructive. David Freedman compiled the following (but incomplete) list of ripostes (2005: 195).

> We all know that. Nothing is perfect. Linearity has to be a good first approximation. Log linearity has to be a good first approximation. The assumptions are reasonable. The assumptions don't matter. The assumptions are conservative. You can't prove the assumptions are wrong. The biases will cancel. We can model the biases. We're only doing what everyone else does. Now we will use more sophisticated techniques. If we don't do it, someone else will. What would you do? The decision-maker has to be better off with us than without us. We all have mental models, not using a model is still a model. The models aren't totally useless. You have to do the best you can with the data. You have to make assumptions in order to make progress. You have to give the models the benefit of the doubt. Where's the harm?

The obvious problem with these and the many other rhetorical devices is that they are little more than a reflexive defense of the problematic status quo.

The second response is to assert that the current assortment of regression diagnostics can find any serious problems in a regression model and that these problems can then be readily fixed. For example, model misspecification can be identified with "specification tests," and a subsequent instrumental variable estimator can save the day (e.g., Greene 2003, Sect. 5.5). In principle, this view has merit. In practice, it usually stumbles badly. For both the diagnostics and the remedies, new and untestable assumptions are required even before one gets to a number of thorny technical complications. On the matter of diagnostics, Freedman observes (2009: 839) that , "...skepticism about diagnostics is warranted. As shown by the theorems presented here, a model can pass diagnostics with flying colors yet be ill-suited for the task at hand." He is no more sanguine about about instrumental variable solutions (Freedman 2005: Sect. 8.5) and has lots of company (e.g., Berk 2004: Sect. 9.5; Kennedy 2003: Chap. 9; Leamer 2010).

The third response is what I have elsewhere called "Model-Lite Causal Analysis" (Berk 2009). The basic idea is apply the randomized experimental paradigm to observational data. One tries to make the case that conditional on a set of covariates, "nature" conducted a randomized experiment. This naturally leads to a variety of matching or post-stratification analysis methods that can be more robust than causal models (Rubin 2008). Regression models can creep in, however, in the construction of matching variables such propensity scores or in post-matching adjustments for sample imbalance (Rosenbaum 2002b, 2010).

Morgan and Winship (2007) provide an excellent overview of the issues. The primary literature can be found in the influential writings of William Cochran, Donald Rubin, Paul Rosenbaum, Judea Pearl and others. Unfortunately, the elephant in the room remains: there

is no apparent way to overcome the impact of omitted variables although in some cases, sensitivity analyses can suggest that the omitted variables may not be important (Rosenbaum 2002a, Chap. 4; 2010, Chap. 14).

The fourth response is to rely on research designs far better suited for causal inference than the usual observational approaches (Angrist and Pischke 2010; Campbell and Stanley 1963; Imbens 2004). Properly implemented randomized experiments are the reigning poster child. But strong quasi-experiments will often perform nearly as well (Cook et al. 2008). For example, recent advances in the regression discontinuity design (Imbens and Lemieux 2008) can make that approach very attractive when a randomized experiment is not feasible (Berk et al. 2010b). Capitalizing on stronger designs is clearly a useful suggestion when such designs are possible.

The fifth response is to go back to the formal definition of a regression analysis and reconsider what can be learned depending on the assumptions that one can credibly make. We turn to that approach now. There will be some good news.

What is Regression Analysis?

Cook and Weisberg (1999: 27) provide a definition of regression analysis that comports well with standard statistical perspectives: "[to understand] as far as possible with the available data how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors." The definition makes the entire conditional distribution of y fair game although in practice, the conditional mean and/or conditional variance are the primary focus. A key point is that there is no mention of estimation, hypothesis tests, or confidence intervals nor any reference to causal inference. One can do a proper regression analysis without any effort to address the role of chance or to make causal statements.

Level I Regression Analysis

We have seen that by definition, regression analysis is a procedure by which conditional relationships in data may be *described*. One might consider, for example, how the homicide rate in a city varies with unemployment, police practices fixed. And one might consider how the homicide rate in a city varies with police practices, unemployment fixed.

In effect, one is identifying interesting patterns in the data, which can be subtle, complicated and even rare. The patterns can be found over time, over space, and over observational units that can differ in complex ways. The analysis can be directed by existing theory or can be highly exploratory.

One is not limited to conventional linear regression. For example, the definition includes response variables that are categorical, ordinal or counts. Categorical regressors are can be included. Nonlinear relationships can be taken into account. One can consider several response variables at once such as a parolee's arrests for new crimes and the charges for those new crimes.

It follows that for description, one need not worry about all of the potentially problematic assumptions required if one is to undertake credible statistical or causal inference. One just characterizes associations in the data at hand. Thus, it does not matter, for instance, if the data are a probability sample or the product of well-defined stochastic process. Likewise, there is no concern about omitted variables. Indeed, it is not clear how to define an omitted variable in this context. It is the proverbial "what you see is what you

get." Consistent with a more expansive discussion elsewhere (Berk 2003: Sect. 11.5), one can call descriptive regression a "Level I" regression analysis. Level I regression analyses are *always* formally appropriate when a regression analysis could be useful and do not depend on any of the assumptions required for statistical inference or causal inference.

This is not a call to abandon advanced statistical procedures and complex quantitative reasoning. Description has long been at the center of state-of-the-art statistical practice. Among the more popular approaches are multivariate statistics (Anderson 1958; Gifi 1990), exploratory data analysis (Tukey 1977; Diaconis 1985), and more recently, dynamic graphics (Cook and Swayne 2007) and Machine/Statistical Learning (Hastie et al. 2009; Berk 2008). The growing use of GIS studies in criminology is one example of graphical methods (Groff and La Vigne 2002; Chainey and Ratcliffe 2005). It is then a small step to build visualizations that take both time and space into account (Berk and MacDonald 2009). Machine learning is also finding its way into the criminology journals (Berk et al. 2009).

Making descriptive regression analysis the central approach to data analysis in criminology is not a radical suggestion. Despite the statistical framing that one finds in the criminology journals, description is usually the actual enterprise. And the reason is apparent: in criminology, and for most social science applications more generally, there are rarely any widely accepted, nearly right models that can be used with real data. By default, the true enterprise is description. Most everything else is puffery.

Level II Regression Analysis

Level I regression analysis does not require any assumptions about how the data were generated. If one wants more from the data analysis, assumptions are required. For a Level II regression analysis, the added feature is statistical inference: estimation, hypothesis tests and confidence intervals. When the data are produced by probability sampling from a well-defined population, estimation, hypothesis tests and confidence intervals are on the table.

A random sample of inmates from the set of all inmates in a state's prison system might be properly used to estimate, for example, the number of gang members in state's overall prison system. Hypothesis tests and confidence intervals might also be usefully employed. In addition, one might estimate, for instance, the distribution of in-prison misconduct committed by men compared to the in-prison misconduct committed by women, holding age fixed. Hypothesis tests or confidence intervals could again follow naturally. The key assumption is that each inmate in the population has a known probability of selection. If the probability sampling is implemented largely as designed, statistical inference can rest on reasonably sound footing. Note that there is no talk of causal effects and no causal model. Description is combined with statistical inference.

In the absence of probability sampling, the case for Level II regression analysis is far more difficult to make. There are several options addressed in some depth elsewhere (Berk 2003: 39–58).

1. *Treating the data as a population*—In effect, this is a fallback to Level I regression analysis in which there is no statistical inference.
2. *Treating the data as if it were a probability sample from a well-defined population*—For example, a convenience sample of big-city police departments might be treated as if it were a simple random sample of police departments from cities of over 100,000 residents. One would have to argue convincingly that despite the lack of formal probability sampling, the way in which the data were generated is de facto a close approximation. And that, in turn, would lead to an in-depth discussion of why one

should believe that each department in the sample was, in effect, drawn independently of all others with a known probability of selection. If access was obtained by a sequence of referrals, the as-if approach would fail. It would also fail if the sample was limited to the subset of departments for which the requisite data were available. The as-if approach is rarely credible in practice.

3. *Treating the data as a random realization from an imaginary "superpopulation"*—Inferences to imaginary populations are imaginary. However, there is an important distinction between an imaginary population that could exist and an imaginary population that could not. Inferences to the former might not be imaginary, but are even more difficult to justify than inferences from the as-if strategy. The result can be a mind numbing exercise in circular reasoning. The population is defined as the population that would exist if the data were a random sample from that population.
4. *Model based sampling*—For this strategy, one needs a credible stochastic model of how the data were generated. For example, with conventional linear regression, one assumes

$$y_i = \boldsymbol{\beta}^T \mathbf{X}_i + \varepsilon_i, \tag{1}$$

where $\varepsilon_i \sim NIID(0,\sigma^2)$. If the model really captures how the data were generated, statistical inference can easily follow. But, inferences are made to the data generation process, not to a real population in the usual sense.[4]

Model-based sampling leads one back into much of the critical literature cited earlier. Why should one believe the model? An omitted variable, for instance, will mean that the assumed properties of the disturbance term do not hold, and statistical inference can be badly compromised. The same can follow when the wrong functional forms are assumed.

In short, a Level II regression analysis depends on how the data were generated. If the data are a well-implemented probability sample, Level I description can be supplemented by Level II statistical inference. Without probability sampling, Level II regression analysis can be difficult to justify.

Level III Regression Analysis

The goal in a Level III regression analysis is to supplement Level I description and Level II statistical inference with causal inference. In conventional regression, for instance, one needs a nearly right model like Eq. 1, but one must also be able to argue credibly that *manipulation* of one or more regressors alters the expected conditional distribution of the response. Moreover, any given causal variable can be manipulated independently of any other causal variable and independently of the disturbances. There is nothing in the data itself that can speak to these requirements. The case will rest on how the data were actually produced. For example, if there was a real intervention, a good argument for manipulability might well be made. Thus, an explicit change in police patrolling practices ordered by the local Chief will perhaps pass the manipulability sniff test. Changes in the demographic mix of a neighborhood will probably not. Here too, there is extensive discussion elsewhere (e.g., Berk 2003, Chap. 5). Suffice it to say, it is very difficult to find in criminology credible examples of a Level III regression analysis. Either the parallel to Eq. 1 is not credible, claims of manipulability are not credible, or both.

[4] Alternatively, one might define the population as all possible realizations of the given stochastic process. This population is imaginary.

Conclusions

Level I regression analysis is always formally appropriate and does not depend on any assumptions about how the data were generated. Truth be told, description is really the de facto product of most regression analyses in criminology. Description is not just a legitimate scientific activity, but corresponds well to the developmental stage in which criminology finds itself.

Level II regression analysis adds to description, estimation, hypothesis tests and confidence intervals. When the data are a probability sample from a well-defined population, statistical inference can follow easily. When the data are not, statistical inference can be difficult to justify.

Level III regression analysis adds to description and statistical inference, causal inference. One requires not just a nearly right model of how the data were generated, but good information justifying any claims that all causal variables are independently manipulable. In the absence of a nearly right model and one or more regressors whose values can be "set" independently of other regressors and the disturbances, causal inferences cannot make much sense.

The implications for practice in criminology are clear but somewhat daunting. With rare exceptions, regression analyses of observational data are best undertaken at Level I. With proper sampling, a Level II analysis can be helpful. The goal is to characterize associations in the data, perhaps taking uncertainty into account. The daunting part is getting the analysis past criminology gatekeepers. Reviewers and journal editors typically equate proper statistical practice with Level III.

Acknowledgments Thanks go to Alex Piquero and Jim Lynch for very help comments on earlier drafts of this paper.

References

Anderson T (1958) An introduction to multivariate statistical analysis, 1st edn. Wiley, New York
Angrist J, Pischke S (2010) The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J Econ Perspect 24(2)
Berk RA (2003) Regression analysis: a constructive critique. Sage Publications, Newbury Park
Berk RA (2004) Regression analysis: a constructive critique. Sage Publications, Newbury Park
Berk RA (2007) Meta-analysis and statistical inference (with commentary). J Exp Criminol 3(3):247–297
Berk RA (2008) Statistical learning from a regression perspective. Springer, New York
Berk RA (2009a) Cant tell: comments on "does the death penalty save lives?". Criminol Public Policy 8(4):843–849
Berk RA (2009b) The role of race in forecasts of violent crime. Race Social Problems 1:131–242
Berk RA, MacDonald J (2009) The dynamics of crime regimes. Criminology 47(3):971–1008
Berk RA, Sherman L, Barnes G, Kurtz E, Ahlman L (2009) Forecasting murder within a population of probationers and parolees: a high stakes application of statistical forecasting. J Royal Stat Soc (Series A) 172, Part 1:191–211
Berk RA, Brown L, Zhao L (2010a) Statistical inference after model selection. J Quant Criminol, forthcoming
Berk RA, Barnes G, Ahlman L, Kurtz E (2010b) When second best is good enough: a comparison between a true experiment and a regression discontinuity quasi-experiment. J Exp Criminol, forthcoming
Box GEP (1976) Science and statistics. J Am Stat Assoc 71:791–799
Breiman L (2001) Statistical moldeling: two cultures, (with discussion). Stat Sci 16:199–231
Campbell DT, Stanley J (1963) Experimental and quasi-experimental designs for research. Wadsworth Publishing, New York
Chainey S, Ratcliffe J (2005) GIS and crime mapping. Wiley, New York

Cook D, Swayne DF (2007) Interactive dynamic graphics and data analysis. Springer, New York
Cook RD, Weisberg S (1999) Applied regression including computing and graphics. Wiley, New York
Cook TD, Shadish WR, Wong VC (2008) Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. J Policy Anal Manag 727(4):724–750
de Leeuw J (1994) Statistics and the sciences. In: Borg I, Mohler PP (eds) Trends and perspectives in empirical social science. Walter de Gruyter, New York
Diaconis P (1985) Theories of data analysis: from magical thinking through classical statistics. In: Hoaglin DC, Mosteller F, Tukey J (eds) Exploring data tables, trends, and shapes. Wiley, New York
Freedman DA (1987) As others see us: a case study in path analysis (with discussion). J Educ Stat 12:101–223
Freedman DA (2005) Statistical models: theory and practice. Cambridge University Press, Cambridge
Freedman DA (2009) Diagnostics cannot have much power against general alternatives. Int J Forecast 25(4):833–839
Gifi A (1990) Nonlinear multivariate analysis. Wiley, New York
Groff E, La Vigne NG (2002) Forecasting the future of predictive crime mapping. In: Tilley N (ed) Analysis for crime prevention 13. Criminal Justice Press, Monsey, pp 29–58
Guerette RT, Bowers KJ (2009) Assessing the extent of crime displacement and diffusion of benefits: a review of situational crime prevention evaluations. Criminology 47(4):1331–1368
Greene WH (2003) Econometric analysis, 5th edn. Prentice Hall, New York
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer, New York
Heckman JJ (2000) Causal parameters and policy analysis in economics: a twentieth century retrospective. Q J Econ 45–97
Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81:945–960
Imbens G (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86(1):4–29
Imbens G (2009) Better late than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). Working Paper, Department of Economics, Harvard University
Imbens GW, Lemieux T (2008) Regression discontinuity designs: a guide to practice. J Econ 142(2): 615–635
Kennedy P (2003) A guide to econometrics. MIT Press, Boston
Kish L (1987) Statistical design for research. Wiley, New York
Leamer EE (1978) Specification searches: ad hoc inference with non-experimental Data. Wiley, New York
Leamer EE (1983) Let's take the con of econometrics. Am Econ Rev 73:31–43
Leamer EE (2010) Tantalus on the road to asymptopia. J Econ Perspect 24(2):1–16
Leeb H, Pötscher BM (2005) Model selection and inference: facts and fiction. Econometr Theory 21:21–59
Leeb H, Pötscher BM (2006) Can one estimate the conditional distribution of post-model-selection estimators?. Annals Stat 34(5):2554–2591
Lieberson S (1985) Making it count: the improvement of social research and theory. University of California Press, Berkeley
Manski C (1995) Identification problems in the social sciences. Cambridge University Press, Cambridge
Mears DP, Bales WD (2009) Supermax incarceration amd recodovism. Criminology 47(4):1131–1166
Morgan SL, Winship C (2007) Counterfactuals and causal inference: methods and principles for social research. Cambridge University Press, Cambridge
Pearl J (2000) Causality: models, reasoning and inference. Cambridge University Press, Cambridge
Rosenbaum P (2002a) Observational studies, 2nd edn. Springer, New York
Rosenbaum P (2002b) Covariance adjustments in randomized experiments and observational studies. Stat Sci 17(3):286–327
Rosenbaum P (2010) The design of observational studies. Springer, New York
Rubin DB (1986) Which ifs have causal answers. J Am Stat Assoc 81:961–962
Rubin DB (2008) For objective causal inference, design trumps analysis. Annals Appl Stat 2(3):808–840
Schultz TD, Tabanico JJ (2009) Criminal beware: a social norms perspective on posting public warning signs. Criminology 47(4):1201–1222
Skardhamar T (2010) Distinguishing facts from artifacts in group-based modeling. Criminology 48(1):295–320
Tukey J (1977) Exploratory data analysis. Addison-Wesley, Reading