Statistics in Medicine

WILEY

Statistics in Medicine

COMMENTARY

Some Ways to Make Regression Modeling More Helpful Than Misleading

Sander Greenland

Department of Epidemiology and Department of Statistics, University of California, California, Los Angeles, USA

Correspondence: Sander Greenland (lesdomes@ucla.edu)

Received: 29 November 2024 | **Accepted:** 4 December 2024

Funding: The author received no specific funding for this work.

I congratulate Carlin and Moreno-Betancur [1] on their excellent call for reform. I would like to expand on their call with the following recommendations for improvements in the teaching and practice of regression analysis:

1. Regression analysis and modeling should begin as part of data description. In that role, fundamental concepts such as model forms, curve fitting, residuals, and influence can be introduced as data summaries, before the more abstract concepts of probability and inference.
2. Conventional regression analysis provides inferences about the data generator, not the assumed target population. We thus need to stop the implicit, automatic equating of the data generator and the target population by using descriptions and notation that makes the distinction explicit.
3. Models are not right or wrong, but are inevitably limited in what they can capture and where they are useful. Thus, to guard against misleading the users, model predictions need to be evaluated against contextual information, model residuals, and the data to see what they captured and what they obscured or missed entirely.
4. To minimize bias from model misspecification, models need to be flexible but not overparameterized. In particular, unnecessary categorizations of quantitative variables should be replaced with simple smooth curves between the extremes of straight lines and cubic splines.
5. Even if the analysis goal is only description of a target population (as in a survey), our analyses need to reflect the causal mechanisms that generated the data.
6. The usual potential-outcome notations for distinguishing regression models from causal models are elegant for the math and confusing for students. Thus, for teaching, more explicit notation is advisable.
7. Sample-size requirements need far more attention than provided by considerations of precision, power, or positivity. At a minimum, when using common default (Wald) large-sample methods we should check for potential breakdown of their approximations.

These recommendations are neither exhaustive nor new; the third especially has a vast historical literature behind it (e.g., [2–4]). Nonetheless, my impression is that all the above points need more coverage in teaching and more deployment in practice.

1 | "Descriptive Regression" Can Refer to Several Different Uses of Modeling

Paralleling the classification of research purposes in Carlin and Moreno-Betancur [1], regression analyses are often classified into "descriptive," "predictive," and "causal." Unfortunately, the most basic descriptive uses of regression seem neglected in teaching and practice, even though they can serve as a foundation for understanding the meaning and limitations of regression in data analysis. To do so we need to subclassify descriptive regression further; when that is done, descriptive uses of regression can be seen as a prerequisite to predictive and causal uses. More radically, descriptive uses can be seen as subsuming predictive uses, and predictive uses can be seen as subsuming causal uses.

Specifically, the term "descriptive regression" may refer to the following different concepts and goals, of which only the third seems to be emphasized in typical statistics discussions:

1. *Description of select data features*. This goal involves no inference, only degrees of detail and neglect in describing data features and their fuzziness. Explication of this goal can be found in the engineering and computer-science literature in which models are used in data-compression algorithms. This non-inferential version of description can illustrate many general points of regression without the subtleties of probability theory, including the geometry of least-squares fitting and its consequences. A simple teaching example of a U-shaped (e.g., parabolic) regression can show how a linear-regression coefficient represents an average difference in outcome per unit change in a regressor; it thus can summarize monotonic trends in data, but also can mask important data patterns when changes cancel each other out across the range of the regressor. The more general lesson is that a parametric regression can only display the patterns allowed by its model form, and that graphs from nonparametric regression have a role in data summarization.
2. *Inferred description of the behavior of the data-generating process* (DGP). This goal could be described as predicting what the data generator will produce next, without regard to how far the prediction may be from a corresponding target population. An example would be predicting the patterns a survey (the data generator) will produce in the face of low response rates and the errors in the survey measurements, as opposed to predicting the actual distributions in the surveyed population. The data generator thus incorporates without distinction all influences on the final data, such as selection bias and measurement errors—and it is the only object directly addressed by conventional interval estimates, p values, and other "superpopulation" inferences.
3. *Inferred description of a target population*. This goal can be described as predicting a specified target-population quantity from a sample, or as filtering out known sources of error (bias and noise) to decode sample information about the target population. This target population may be current, as in disease surveys (which have a descriptive purpose); or it may be a future population, as in epidemic forecasts (which have a predictive purpose); or it may include potential outcomes, as in causal modeling (which predicts outcomes under alternative actions [5]).

Most teaching of regression pretends as if inferential statistics apply directly to the target population. Unfortunately, this pretense is based on very strong and often implausible set of assumptions, typically that data generation (item 2) involved only probability sampling of known structure from the target (e.g., random conditional on a design matrix), with measurements of variables that are error-free or with known error structures, and with parameters that are homogeneous (constant) across what are clearly very different person, times, and places.

I thus argue that, in teaching and practice, regression analysis would best begin with description of the data (item 1), and then move to description of generator behavior (item 2) before claiming to provide a description of a target-population (item 3). This advice applies not only to observational studies, but also to randomized clinical trials due to their extreme selectivity in trial recruitment, treatment adherence, and retention. The key advantage conferred by treatment randomization is that it eases identification of treatment effects *within* the trial. Those within-trial effects are part of the data generator (item 2), and typically very far from effects in actual medical practice (item 3). Estimation of the latter effects requires methodology and data well beyond what is covered in most regression textbooks; for that we must turn to modern literature on transportability (e.g., [6, 7]).

In survey sampling, the sampler attempts to impose known selection probabilities over a sampling frame that covers the entire population. If however there is non-negligible post-sampling nonresponse, loss, or missing data, the selection probabilities are not identified by the data. The literature on dealing with such problems is too vast and sophisticated to cover in basic courses, but I believe that the core distinction between data-generating and target distributions should be spelled out early and clearly, with emphasis on the facts that (a) the only data we observe come from the generator, yet (b) conventional statistics (as seen in the vast majority of medical research articles) pretends as if the data come straight from the target, or at least that selection varies in a way that leads to no distortion of targeted parameters.

1.1 | Describing the Data With Fitted Models

The use of regression models for data description is apparent in the works of Mosteller, Tukey, and others in the 1950s–1970s (e.g., [8]). In this descriptive role, parametric model forms are seen as filters that pass along information that follows the form, and filter out information which does not so conform; the goal is to save only data patterns that fall within the model form. A fitted model is then a data summary that may be more or less adequate for the purpose at hand, and can be viewed as a type of very lossy data compression.

If the application requires extracting particular patterns from the data, a model form will be adequate to the extent it allows through those patterns. For example, linear regression extracts and summarizes linear components of data patterns and filters out the rest as residuals; it will be an adequate description if only linear components are needed, and inadequate to the extent higher-order components are needed. Thus, model adequacy is a contextual, application-specific property rather than an absolute property of a model relative to some reality [4].

More abstractly, a regression model provides a data summary or compression based on projecting the data onto the subspace of expected data sets following the model form [9, 10]. The model extracts specific data features for which it is tuned or sensitive to—namely, the features that can be reproduced by model—while the residuals display features that the model filters out. If, as usual, the model has not been pre-specified based on which patterns are contextually important and which are not, its summarization adequacy can be examined directly by seeing if clear patterns were left behind in the residuals, and if the patterns the model captures are general or instead driven by a few influential data records.

1.2 | Describing the Data Generator With Fitted Models

The task of data description can be translated into a parallel task of describing the behavior of the data generator. Doing so, the model is a filter that will capture features of the data generator for which it is tuned or allows in as a signal, and will filter out the rest as residual. Residuals now display what the model is missing about the generator, and *p* values for model fit based on these residuals provide one index (of many) for evaluating the adequacy of the data-generating model in light of the data [4, 10].

The difference between data and generator description is that generator description involves uncertainty: We usually observe just one data set from the generator, and can see directly how close fitted values are to observed values. If however there are unobserved sources of variation in the generator, we must invoke assumptions about the distribution of residuals to infer behavior of the generator. For example, assuming a parametric regression model with independent identically distributed residuals allows each observation to be treated as a replicate observation from the generator, with an estimable systematic offset that follows the model form. The results can capture patterns in generator output even when the model form does not correctly capture generator behavior (i.e., is misspecified) [11]; again, residual analysis [12] can aid in detection of assumption violations and what the model missed.

1.3 | Describing the Target Population With Fitted Models

Description of the data or its generator is not the goal of most studies, and so descriptive regression is usually framed only as a tool for inference about target-population features, that is, for describing the target population which fed the generator. As mentioned under item 3, the convention is to use very strong and often questionable assumptions that equate patterns or parameters in the data generator with patterns or parameters in the target. That is, the data distribution produced by the actual data-generating process is treated as the one that would be produced by perfect random sampling of the target population.

In this fashion, inferences about the data generator become identical to and thus confused with inferences about the target population. Most statistical textbooks and presentations make this identification automatically, with some narrative asides in the discussion about possible discrepancies. Yet the assumptions needed for the identification are uncertain and often implausible, which makes conventional statistical inferences at best overconfident if not misleading. In particular, parameter estimates, *p* values and other inferential outputs from regression methods take no account of gaps between the data-generating distribution and the distribution of the target population, and thus understate warranted total uncertainty about the target.

1.4 | Nonparametric Regression and Related Methods

Nonparametric regression and machine learning (algorithmic modeling) do not pre-specify data or generator structure, apart from what may be unobjectionable constraints (e.g., smoothness); they can thus adapt to the data in fine detail, leaving unpatterned residuals [13]. The resulting data reduction is a complex response surface, and intelligible presentation requires graphs of fitted outcomes plotted against covariates, rather than a table of transformed coefficient estimates. Addition of uncertainty bounds to these graphs apply to data-generator behavior, but the sample size required for the validity of these bounds is beyond that of most health and medical studies. Finally, interpreting those bounds as applying to the target population requires the same strong assumption as in the parametric case, that is, equating the generator distribution to the distribution of samples from the target population under a known sampling design.

The need for graphical presentation of nonparametric regressions does not lend itself well to publication constraints. Hence in clinical and epidemiologic reports nonparametric methods remain mostly in intermediate roles, as seen in estimation of average treatment (marginal) effects [14] or weight construction for parametric model fitting. Nonetheless, these methods deserve to be integrated into basic regression education, and may even displace much of traditional methodology for data summarization.

2 | Models Need to be Flexible but not Too Flexible: Trend Models as a Case Study

Beyond the basic assumption that the data identify some features of the target that we are after, parametric models add more assumptions in the form of constraints imposed by the chosen model, such as additivity and linearity on some scale. Where does that model come from? As Box wrote [4], it *should* be derived from contextual information, such as constraints imposed by the study design and conduct, and assumptions uncontroversial in the context (which usually includes uniform smoothness). Nonetheless, most applications use off-the-shelf parametric models which impose constraints well beyond contextual information (such as additivity) and may even conflict with it (as monotonicity often does). We thus should ask what fitted misspecified models are telling us about the generator distribution. A large literature on this topic emerged in the 1960s–1990s (e.g., [3, 4, 11, 15]). Among the many recommendations there, general practice has adopted "specification-robust" methods such as sandwich covariances and bootstrapping of residuals [11, 12]; however, those methods capture only uncertainties about the patterns allowed by the fitted model, and do not account for inflexibilities of the model.

For quantitative regressors such as age or drug dose, this problem is partially addressed by using more than one model term for the outcome trend across the regressor. This expansion increases model flexibility and reduces estimation bias, at a cost of increased estimation variance and attendant power loss. Unfortunately, conventional analyses continue to use wastefully inefficient, biased, and unrealistic model expansions. One example is breaking the regressor into categories determined by covariate quantiles. At the other extreme, cubic splines involve more terms than needed for optimal estimation; their popularity is based on an extreme flexibility which is useful when complex trends are expected (as in engineering) but which leads to overfitting in typical medical studies. Between these extremes are simple quadratic splines and fractional polynomials, which are easy to fit, provide more realism and efficiency than categorization, and avoid the overfitting inherent in typical cubic splines [16, 17]. These models provide a much-needed compromise between parsimony and flexibility, and thus deserve coverage in teaching and use in real analyses.

A deeper problem is that the model used in presentations is often the output of a preliminary selection algorithm, yet it is used to create "statistical inferences" which assume that final model was wholly pre-specified, leading to severe miscalibration of the results [9, 12, 18–20]. More sophisticated methods either account for the preliminary model selection or average results over model candidates [18, 19], which are in effect fitting a more flexible model so that selection can be avoided. These are useful technical advances for predicting generator behavior, but still fail capture uncertainty about discrepancies between the data-generator and target-population patterns, which may show up as dramatic failures in clinical prediction [17]. Uncertainties about those discrepancies are a central topic in bias analysis [21, 22], which focuses on model expansion to encompass distortions of target-population patterns by the data generator.

3 | We Need Causal Thinking in the Foundations, Teaching, and Practice of Statistics

The current causality literature focuses on forecasting what would happen to a population or individual under different intervention strategies. Nonetheless, paraphrasing Hill [23], no contextually sensible probability model can be constructed without asking "what caused this data set rather than some other to appear before us?" I thus use the term "data generator" for all the causal processes (physical mechanisms) leading to the analyzed data set, and regard causal thinking as fundamental for all applied probability and statistics [24].

In causal modeling, extrapolations are made from observed treatment groups to counterfactual allocations, and thus model misspecification can be disastrous for health practices and medical care. For example, almost all medication and nutrient effects on health change direction across dosing, and thus present a dose-optimization problem. But supplement overdoses often occur due to naïve monotonic ("more is better") extrapolation from health benefits observed at lower doses, raising again the need for models capable of capturing effect reversal.

To formally represent causal extrapolations, we need to expand beyond ordinary regression models to potential-outcome models that forecast outcomes of each observational unit under different interventions [25–27]. Conventional regression modeling remains a useful component in fitting causal functions, for example when fitting outcome regressions using inverse probability-of-treatment weights (IPTW) obtained from treatment-assignment regression to obtain "doubly robust" estimates of causal effects [26]. But the target of these combined modeling efforts are causal functions, which go beyond conventional regression functions in complexity and interpretation. Thus, the distinction between these functions is essential for teaching and practice.

There are several ways to symbolically display the distinction between regression and causal functions. This is usually done by denoting the regressand (regression-function output) by Y, distinguishing the structural output (potential outcome) for intervention $X = x$ by Y_x, $Y(x)$, or Y^x. These notations make the math compact and facilitate a useful mapping between missing-data and causal-modeling methods [26, 28]. Nonetheless, I have found them problematic for teaching, for example, Y_x gets confused with Y, Y^x gets confused with exponentiation, and both give the incorrect impression that potential outcomes must have a joint distribution or be deterministic [29].

Thus, for teaching I much prefer the more explicit notation in which the probabilities for a discrete outcome variable Y when treatment X is set to level x are denoted by $\Pr[Y = y \mid \text{set}(X = x)]$ or $\Pr[Y = y \mid \text{do}(X = x)]$ [23, 25]. This notation shows the stochastic potential outcome of treating the entire population with $X = x$, regardless of actual X values, and represents treatment as an observable function $\text{set}(\cdot)$ or $\text{do}(\cdot)$ rather than as selection among hypothetical pre-existing variables. It should be contrasted to the conditional probability function for regression, $\Pr[Y = y \mid X = x]$, which shows the probability of $Y = y$ in the subpopulation for which $X = x$. The two concepts can be combined, for example, $\Pr[Y = y \mid \text{set}(X = x), Z = z]$ provides the potential outcome of applying the treatment $X = x$ to the subpopulation for which $Z = z$, and randomization of X stratified on Z is the primary basis for the simplification

$$\Pr[Y = y \mid \text{set}(X = x), Z = z] = \Pr[Y = y \mid X = x, Z = z]$$

which enables estimation of the causal effects of treatment X using regression models [25–27].

Whatever notation is chosen, causal models provide a basis for deriving and extending covariate adjustments, as well as for more encompassing strategies such as target-trial emulation [26, 30] to bring regression outputs closer to targeted causal effects.

4 | Sample-Size Requirements for Large-Sample (Asymptotic) Methods

A purely technical problem is that few analyses or programs check for approximation failures, even though their effects can be profound [31, 32]. As a holdover from the computing limits of the past century, most software will default to presenting Wald statistics for a coefficient β, which combine a point estimate b of β with its estimated standard deviation (standard error) s to form a Z-statistic b/s for the hypothesis that $\beta = 0$ and a $b \pm 1.96\,s$ 95% interval estimate (which contains all values for β that have a two-sided Wald $p > 0.05$). Unfortunately, for common models such as logistic regression, these statistics converge more slowly to asymptotic behavior than do likelihood-ratio and score statistics. Alternatives statistics based on likelihood ratios have long been available in major software, but must be known to and chosen by the user.

The current causal-inference literature tends to focus on "positivity" conditions (e.g., positive probability of receiving each treatment being compared, or, more strongly, positive numbers at each treatment level), but these conditions are far too weak to ensure sample-size adequacy for estimating effects of qualitative treatments. Instead, adequacy requires at least multiple cases and noncases at each treatment level; precise requirements in multivariable applications will further depend on a host of parameters that vary with setting.

In particular, there is no universal numeric rule analogous to the classic "at least 5 observations expected per cell" for contingency tables, which implicitly assume models loglinear in the expected counts (e.g., logistic, Poisson, and proportional-hazards models): Requirements vary dramatically with the form of the model fit, with very slow convergence to asymptotic behavior for models linear on a bounded scale (such as additive risk and rate models) and for nonparametric regressions. I have thus suggested that, at the very least, software should routinely compute and report diagnostics that signal approximation breakdowns, for example by providing the score or likelihood-ratio p values for the limits of Wald intervals [33].

There have been many simulation studies and numeric examples of approximation accuracy; most however assume that the fitted model is correct and that the goal is estimation of a model coefficient or an average effect, for which model errors may cancel out. But prediction of patient-specific effects are what clinical practitioners most need for choosing treatments. Outside of regions of highly influential data, individual predictions can be enormously biased by model misspecification, and highly unstable [12, 17]. This aspect of the misspecification problem would favor more flexible, highly parameterized models, or nonparametric regression and machine-learning algorithms to avoid rigid (and thus likely erroneous) model specification [12]; again however they require sample sizes well above that of most health and medical studies to reach their nominal large-sample behavior and to provide stable individual predictions.

5 | Closing Cautions

The effort needed to absorb the regression extensions advised above may tax students and researchers if their understanding of math is weak. They may also tax mathematics and statistics majors lacking experience with actual study operations (the real data generators) and scientific controversies. Both types of student present educational challenges that need to be met by narrative and graphical explanations to illustrate abstract theory.

Unfortunately, statistical training has often fostered the misimpression that adopting theories of inference and understanding their mathematical intricacies is sufficient for competent data analysis. This "math delusion" has long been criticized for leading to analyses based on implausible assumptions, to faulty interpretations of statistical outputs, and for neglecting the more basic judgments needed to employ mathematical results in formulating scientific inferences [1–4, 15, 18–25, 34–39]. Teaching about this problem is aided by analyses of real, messy data sets with full narrative description of the actual processes that generated them, pointing out how those cannot be fully described by tractable models.

Proper interpretation will be further aided by requiring the purpose of modeling to be clearly defined before models are specified and fitted [1], and by switching to more modest, accurate terminology to describe program outputs, such as replacing overconfident claims of "significance," "power," and "confidence" with more modest observations of compatibility [10, 36, 38, 40]. Such exercises illustrate how all models are wrong in detail, but some can be useful to a degree *if* their assumptions are listed in full with verbal translation into the application context, then checked against both background causal narratives and the analysis data [2, 4, 12, 15, 24, 26, 28, 34].

Acknowledgments

I would like to thank John Carlin, Margarita Moreno-Betancur, and Alessandro Rovetta for helpful comments on the initial draft of this article.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. J. B. Carlin and M. Moreno-Betancur, "On the Uses and Abuses of Regression Models: A Call for Reform of Statistical Practice and Teaching," *Statistics in Medicine* 44, no. 13-14 (2025): 1–16, https://doi.org/10.1002/sim.10244.

2. J. G. Skellam, "Models, Inference, and Strategy," *Biometrics* 25 (1969): 457–475.

3. E. E. Leamer, *Specification Searches* (New York: Wiley, 1978), https://www.anderson.ucla.edu/faculty/edward.leamer/books/specification_searches/SpecificationSearches.pdf.

4. G. E. P. Box, "Sampling and Bayes Inference in Scientific Modeling and Robustness (With Discussion)," *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 143 (1980): 383–430.

5. S. Greenland, "Chapter 5: Causal Inference as a Prediction Problem: Assumptions, Identification, and Evidence Synthesis," in *Causality: Statistical Perspectives and Applications*, eds. C. Berzuini, A. P. Dawid, and L. Bernardinelli (Chichester, UK: John Wiley and Sons, 2012), 43–58.

6. P. N. Zivich, J. K. Edwards, E. T. Lofgren, S. R. Cole, B. E. Shook-Sa, and J. Lessler, "Transportability Without Positivity: A Synthesis of Statistical and Simulation Modeling," *Epidemiology* 35 (2024): 23–31, https://doi.org/10.1097/EDE.0000000000001677.

7. I. J. Dahabreh, S. E. Robertson, A. Jon, and J. A. Steingrimsson, "Learning About Treatment Effects in a New Target Population Under Transportability Assumptions for Relative Effect Measures," *European Journal of Epidemiology* 39 (2024): 957–965, https://doi.org/10.1007/s10654-023-01067-4.

8. F. Mosteller and J. W. Tukey, *Data Analysis and Regression* (Menlo Park: Addison Wesley, 1977).

9. S. Vansteelandt and O. Dukes, "Assumption-Lean Inference for Generalized Linear Model Parameters," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 84, no. 3 (2022): 657–685.

10. S. Greenland, "Divergence Versus Decision P-Values: A Distinction Worth Making in Theory and Keeping in Practice," *Scandinavian Journal of Statistics* 50 (2023): 54–88, https://doi.org/10.1111/sjos.12625.

11. H. White, *Estimation, Inference, and Specification Analysis* (New York: Cambridge University Press, 1994).

12. F. Harrell, *Regression Modeling Strategies* (New York: Springer, 2015).

13. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. (New York: Springer, 2009).

14. M. S. Schuler and S. Rose, "Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies," *American Journal of Epidemiology* 185 (2017): 65–73, https://doi.org/10.1093/aje/kww165.

15. D. A. Freedman, "Statistical Models and Shoe Leather," *Sociological Methodology* 21 (1991): 291–313, https://www.jstor.org/stable/i212010.

16. S. Greenland, "Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis," *Epidemiology* 6 (1995): 356–365.

17. E. W. Steyerberg, *Clinical Prediction Models*, 2nd ed. (New York: Springer, 2019).

18. D. Draper, "Assessment and Propagation of Model Uncertainty (With Discussion)," *Journal of the Royal Statistical Society, Series B* 57 (1995): 54–97.

19. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. (New York: Chapman and Hall, 2013).

20. A. Gelman and E. Loken, "The Statistical Crisis in Science," *American Scientist* 102 (2014): 460–465.

21. M. P. Fox, R. F. MacLehose, and T. L. Lash, *Applying Quantitative Bias Analysis to Epidemiological Data*, 2nd ed. (New York: Springer, 2021).

22. S. Greenland, "Chapter 19: Sensitivity Analysis and Bias Analysis," in *Handbook of Epidemiology*, 3rd ed., eds. W. Ahrens and I. Pigeot (New York: Springer, 2024).

23. A. B. Hill, "The Environment and Disease: Association or Causation?," *Proceedings of the Royal Society of Medicine* 58 (1965): 295–300, https://doi.org/10.1177/003591576505800503.

24. S. Greenland, "Chapter 31: The Causal Foundations of Applied Probability and Statistics," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, vol. 36, eds. R. Dechter, J. Halpern, and H. Geffner (New York: ACM Books, 2022), 605–624, https://dl.acm.org/doi/10.1145/3501714.3501747; corrected version at, https://arxiv.org/abs/2011.02677.

25. J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. (New York: Cambridge University Press, 2009).

26. M. A. Hernán and J. M. Robins, *Causal Inference: What if?* (New York: Chapman & Hall, 2020), https://www.hsph.harvard.edu/miguel-hernan/wp-content/uploads/sites/1268/2024/04/hernanrobins_WhatIf_26apr24.pdf.

27. S. Greenland, "Chapter 30: Regression Methods for Epidemiological Analysis," in *Handbook of Epidemiology*, 3rd ed., eds. W. Ahrens and I. Pigeot (New York: Springer, 2024).

28. D. B. Rubin, "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association* 100 (2005): 322–331, https://doi.org/10.1198/016214504000001880.

29. T. J. VanderWeele and J. M. Robins, "Stochastic Counterfactuals and Stochastic Sufficient Causes," *Statistica Sinica* 22 (2012): 379–392, https://doi.org/10.5705/ss.2008.186.

30. M. A. Hernán and J. M. Robins, "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available," *American Journal of Epidemiology* 183 (2016): 758–764, https://doi.org/10.1093/aje/kwv254.

31. S. Greenland, M. A. Mansournia, and D. G. Altman, "Sparse-Data Bias: A Problem Hiding in Plain Sight," *British Medical Journal* 353, no. i1981 (2016): 1–6, https://www.bmj.com/content/352/bmj.i1981.

32. M. Gosho, T. Ohigashi, K. Nagashima, Y. Ito, and K. Maruo, "Bias in Odds Ratios From Logistic Regression Methods With Sparse Data Sets," *Journal of Epidemiology* 33 (2023): 265–275, https://doi.org/10.2188/jea.JE20210089.

33. S. Greenland, "Likelihood-Ratio Testing as a Diagnostic Method for Small-Sample Regressions," *Annals of Epidemiology* 2 (1992): 311–316, https://doi.org/10.1016/1047-2797(92)90063-V.

34. G. E. P. Box, "Science and Statistics," *Journal of the American Statistical Association* 71 (1976): 791–799, https://doi.org/10.1080/01621459.1976.10480949.

35. G. E. P. Box, "Comment on Shafer," *Statistical Science* 5 (1990): 448–449.

36. S. Greenland, "Invited Commentary: The Need for Cognitive Science in Methodology," *American Journal of Epidemiology* 186 (2017): 639–645, https://doi.org/10.1093/aje/kwx259.

37. M. A. Hernán, "Spherical Cows in a Vacuum: Data Analysis Competitions for Causal Inference," *Statistical Science* 34 (2019): 69–71, https://doi.org/10.1214/18-STS684.

38. S. Greenland, "Connecting Simple and Precise p Values to Complex and Ambiguous Realities (Includes Rejoinder to Comments on "Divergence vs. Decision P-Values")," *Scandinavian Journal of Statistics* 50, no. 3 (2023): 899–914, https://arxiv.org/abs/2304.01392, https://onlinelibrary.wiley.com/doi/10.1111/sjos.12645.

39. D. Draper, J. S. Hodges, C. L. Mallows, and D. Pregibon, "Exchangeability and Data Analysis (With Discussion)," *Journal of the Royal Statistical Society, Series A* 156 (1993): 9–37, https://doi.org/10.2307/2982858.

40. V. Amrhein, S. Greenland, and B. McShane, "Retire Statistical Significance," *Nature* 567 (2019): 305–307, https://media.nature.com/original/magazine-assets/d41586-019-00857-9/d41586-019-00857-9.pdf.