Semiparametric Regression with R, Jaroslaw Harezlak, David Ruppert, and Matt P. Wand, New York, NY: Springer, 2018, xi+331 pp., $149.99(P), $109.00(e-Book), ISBN 978-1-4939-8851-8(P)/978-1-4939-8853-2(e-Book).

Semiparametric Regression with R by Harezlak, Ruppert, and Wand is a truly heroic book! To many of us, statistics can be roughly dichotomized into parametric and nonparametric. Semiparametric statistics aim to bring together the best of these two worlds by modeling the components of scientific interest parametrically for the ease of interpretation and the "nuisance" components nonparametrically to avoid unnecessary model misspecification. Yet its adoption into applied data analysis has been slower than expected compared to some other areas (e.g., high-dimensional linear models or machine learning). This could be attributed to the lack of an easy-to-follow textbook on this topic. Fortunately, HRW fill this huge void, a much overdue for practitioners awaiting guidance from the experts. A related book with more emphasis on the theoretical aspects can be found in Ruppert et al. (2003).

1. The Learning-by-Doing Philosophy, the R package, Datasets, and Seamless Bayesian-Semiparametric Methods with `stan`

First and foremost, the authors should be applauded for espousing the "learning-by-doing" philosophy, writing the R package `HRW` and compiling more than 20 datasets for the book[1]. All these together close the gap between methodology and application, and drastically ease the transition from reading the book to gaining hands-on experience. In particular, we highly recommend readers spend some time on Sec. 1.3 that encapsulates a condensed version of the entire book by previewing one data analysis example per chapter. For instance, in Sec. 1.3.5, by simple visualization, HRW explained vividly why one should prefer modeling the conditional quantile instead of the conditional mean function in the presence of outliers, thereby motivating readers to learn semiparametric quantile regression techniques in Chapter 6. Fast forwarding to later chapters, the authors did not simply stop at teaching how to implement and interpret a given semiparametric method in R—instead, they went above and beyond by exhibiting R codes for informative and esthetic visualization. The way that HRW analyzed data and presented the analysis results in this book set a high standard for practitioners to follow in their own applied works.

Before embarking on the main content, we feel obliged to highlight another unique feature of the book—the seamless inclusion of the Bayesian paradigm, which could be more convenient to fit complex statistical models than frequentist paradigm in certain cases. However, the learning curve of Bayesian methods is usually steeper for students and practitioners, due to the lack of off-the-shelf R packages. Though Bayesian engines like `stan` have been made as automatic as possible, it is still not as simple as pressing a single button. HRW greatly accelerate the learning process by presenting the entire `stan` workflow for each example involving Bayesian inference. Following HRW's recipe, students and practitioners should easily adapt the R codes from this book to their own data analysis. That being said, it is worth emphasizing to practitioners that the correctness of a Bayesian-semiparametric procedure shall not be taken for granted (Ghosal and van der Vaart 2017), in case this is a less familiar topic to some.

2. An Odyssey of Classical Topics: Splines, Additive Models, and Models Beyond iid

Additive models have been the workhorse of semiparametric regressions in practice. Not surprisingly, a large portion of the book is devoted to additive models and their immediate extensions.

Chapter 2 of the book first introduces penalized splines, because in additive models, the nonlinear components are often approximated by linear combinations of splines. Many facets of penalized splines are covered, including the pros and cons of different types of splines (Sec. 2.2), and how to choose tuning parameters: for example, the smoothing parameters for penalty with generalized cross-validation (Sec. 2.3) and the number of basis functions (Sec. 2.4). Regarding the number of basis functions, HRW mentioned on page 23 that "For signals that typically arise in applications $K \approx 35$ will often be more than enough." This is presumably an empirical conclusion based on the authors' long experience in data analysis. But we do want the readers to take this point with some caution—it is not entirely impossible that this rule of thumb may not hold for a particular application. This chapter then proceeds to the common model checking practice of visualizing the residual plots (Sec. 2.5), followed by Sec. 2.6, defining the effective degree of freedom (EDF). EDF is an important concept, often combined with the log residual sum-of-squares to form a metric of model goodness-of-fit (GoF) such as AIC; see Sec. 2.9 on how to perform additive model GoF test in R. Sec. 2.7 connects additive models to mixed models (also see later paragraphs on Chapter 4). This interesting connection motivates alternative model fitting strategies developed in the mixed-model literature, for example, the celebrated *restricted maximum likelihood* (REML). The connection also hints at the use of the Bayesian paradigm (Gelman et al. 2013) for fitting additive models, covered later in Sec. 2.10.

[1] Also see the accompanying website *http://semiparametric-regression-with-r.net/* of the book.

Statistical analysis usually does not stop at simply fitting the model. It is our statisticians' second nature to also quantify the uncertainty of the model fit. In the context of additive models, HRW explain how to build variability bands in Sec. 2.9. An advantage of the Bayesian paradigm (Sec. 2.10) is that we can directly use the posterior samples to build variability bands (aka credible bands). Sec. 2.10 includes an in-depth tutorial on using `stan` for Bayesian inference of additive models, including common practice of using MCMC to generate the posterior samples for statistical inference. This section is user-friendly for those who are unfamiliar with Bayesian inference. Finally, Chapter 2 ends with a prelude for more complex settings than additive models (Sec. 2.11) and an extension of additive models by allowing for spline-factor interactions (Sec. 2.12). The spline-factor interactions are useful when some covariates are categorical and naturally stratify the sample into several groups (e.g., districts in a city, gender,...). In such cases, one may want to model the nonlinear components via different linear combinations of splines, illustrated by analyzing the *Warsaw Apartments* datasets.

With the foundations built in Chapter 2, Chapter 3 moves onto generalized additive models (GAMs) that can handle non-continuous response variables, such as labels or counts. Such an extension is analogous to the extension from linear to generalized linear models. Hence, the materials in Chapter 3 are conceptually straightforward. Apart from introducing GAMs, Chapter 3 also considers the important but tricky problem of model selection. Commonly-used model selection procedures such as stepwise selection and penalty-based strategies are included. While illustrating model selection in real data analysis, one point that is not covered but nonetheless worth emphasizing is the interpretation of the fitted model after selection. It is well known that statistical inference after model selection is generally invalid without further tweaks. Similar concern persists for GAMs when building variability bands covered in Sec. 2.8. Thus, care needs to be taken in practice (Rügamer et al. 2022).

Chapter 4 is concerned with analyzing grouped data, including longitudinal, multilevel, panel, and small area data. The journey starts from the additive mixed models, which model the within-group correlations by using random effects (so parameters are random variables, as in the Bayesian paradigm). Sec. 4.2 begins with using frequentist methods such as the R package `gamm` to fit additive mixed models. As mentioned, mixed-models are intimately connected with the Bayesian paradigm, which nowadays is often used to fit complicated hierarchical models. R codes are also provided for fitting additive mixed models with `stan`, with extensions to non-Gaussian response variable in Sec. 4.5.

An alternative approach to mixed models is to fit marginal models, that do not model within-group covariance structure using random effects. Instead, they directly model the covariances between and within group marginally. Parameters in marginal parametric models can be estimated via generalized estimating equations (GEE) together with sandwiched robust variance estimators for valid inference. HRW take a different path for estimating non- or semiparametric models—they first explain how marginal models and mixed models are related and then use estimation methods from additive mixed models to indirectly estimate the marginal parameters.

Chapter 5 moves one step beyond the realm of additive models. It considers bivariate nonparametric regression that can model complex pairwise interactions between covariates. The bivariate model is especially suitable for modeling geo-spatial data, because the geographic locations are inherently two-dimensional. Other applications of bivariate models include varying-coefficient models with coefficients depending on time (Sec. 5.4), and covariance function (of two time points) estimation for (sparse) functional data (Secs. 5.5 and 5.6). Two types of splines are introduced for modeling bivariate functions: tensor product of univariate splines and thin plate splines. Readers should be able to grasp the essence of bivariate nonparametric regression after reading the fascinating case study, Ozone Levels in Midwest USA in Sec. 5.2.1. Both tensor product splines and thin plate splines are used to analyze the same data and arrive at similar conclusions. In Sec. 5.3, Geoadditive models are introduced, which is nothing but extending the above "bivariate splines" to GAMs, explained in conjunction with a very comprehensive data analysis example. Varying-coefficient models and covariance function estimation for (sparse) functional data are also treated in-depth under the bivariate function framework. Finally, this chapter ends with a brief digression to the so-called fast "sandwich smoother" for modeling imaging data. Overall, this chapter is highly pedagogical as the authors discuss many details on (spatial) data analysis. As an example, in Sec. 5.3.1, the distribution of residuals of Geoadditive model fits are analyzed in-depth, including several directions an analyst could try in practice, for example, adding more covariates into the model or modeling quantiles instead of expectations.

3. A Glance at Some Additional Topics

Chapter 6 (the final chapter) of the book quickly walks readers through snippets of a variety of miscellaneous topics, including robust methods for heavy-tailed data, functional data, and missing data and measurement errors, kernel machines, and more complicated Bayesian methods.

Regarding heavy-tailed data or data with outliers, Sec. 6.2 mainly covers GAMs with outcomes modeled by t-distribution using the `VGAM` package, Bayesian t-regression, and semiparametric quantile regression. As explained by HRW, one disadvantage of VGAM is the lack of data-driven selection for smoothing parameters. This is then resolved by using the Bayesian t-regression, with hyperpriors specified for the smoothing parameters and letting the data speak for itself. The above two approaches still try to fit the conditional mean function of the response. An alternative is to consider semiparametric quantile regression, which is also robust against the heavy-tailedness or outliers but bears a different interpretation.

Chapter 6 then proceeds to more complex data structures—functional covariates (Secs. 6.3–6.5) and functional response and covariates (Sec. 6.6). This part of the book is intimately related to the development of functional data analysis (`fda`) in recent decades (Wang et al. 2016). `fda` garners more and more attention throughout the years because of our increasing ability to collect high-frequency high-dimensional data in fields like brain sciences and mobile health. HRW start with linear models for handling functional data using three datasets with different types of functional covariates (e.g., the conditional mean depends on the integral or derivative of a function). With the preparation under functional linear models in Sec. 6.3, it becomes easier for readers to combine what they have learnt on GAMs in previous chapters with functional response/covariates. Secs. 6.4 and 6.5 explain two different approaches of GAMs with functional covariates, one additive in the functions (Sec. 6.4) and the other additive in the principal components (or eigenfunctions) (Sec. 6.5). To illustrate the difference between these two approaches, HRW use the same dataset (*Fat Content of Meat Samples*) to compare the model fit using AIC. In Sec. 6.6, methods for also handling functional responses are covered, but under simpler linear models instead of additive models. Such methods are not trivial to use and in Sec. 6.6.1 the authors illustrate the entire R workflow in great detail.

Kernel machines are a powerful tool for function approximation and were developed fairly recently. As pointed out at the beginning of Sec. 6.7, many semiparametric methods, including the ones for grouped data, can be interpreted as kernel machines. The utility of kernel machines is demonstrated via the standard support vector machine (svm) for classification tasks, together with the *penalized svm* that has more semiparametric flavor by combining additive models, in order to improve the interpretability of the classification algorithm. We would like to add that idea similar to penalized svm has been extended to modern deep neural nets (DNNs) setting (Agarwal et al. 2021). Considering the resemblance between kernels and DNNs (Jacot et al. 2018), we expect that kernel-machine-based semiparametric methods become even more prevalent in practice in the near future.

All previous sections assume there is neither missing data nor measurement error, which, however, is the exception rather than the rule in reality. Sec. 6.8 undertakes a Directed Acyclic Graphical (DAG) approach for Bayesian semiparametric regression with missing data or measurement error and explains how DAGs can represent the hierarchical structure of the Bayesian paradigm, in which parameters are random variables and thus also represented as vertices together with the data on a DAG. We do find that this section is more challenging for students and practitioners, considering that graphical models are not yet a part of the core curriculum in most of the (bio)statistics or data science graduate programs.

The authors quickly switch gear to a more concrete method termed as *nonparametric regression with a partially observed Gaussian predictor* in Sec. 6.8.2 and its applications. One advantage of Bayesian paradigm is its ease of incorporating sensitivity analysis, a critical step when missing-not-at-random (MNAR) is deemed more plausible than missing-at-random (MAR) by the analyst. Unfortunately, model checking on possible violation of MAR is only skimmed through in Sec. 6.8.3. For readers who want a more rigorous treatment on missing data or measurement error, we recommend that they study this part concurrently with books like Little and Rubin (2019). Finally, Sec. 6.9 covers more complex semiparametric models such as heteroscedastic additive models that could benefit from the power of `stan`. But as acknowledged by the authors, the computation cost of MCMC sampling is now a major bottleneck in applications.

4. Concluding Remark

As students and researchers in the field of semiparametric statistics, we sincerely thank Professors Harezlak, Ruppert, and Wand, for their time and effort dedicated to writing this book, and their tremendous service and contribution to the entire field of (semiparametric) statistics. One can tell from reading the book that the authors have done a lot of trouble shooting for their readers: for example, they realized that it was difficult to install the `cosso` package and provided a concrete solution (though at the time of this writing, `cosso` has already been removed from CRAN). This is just one of many such examples.

For the foreseeable future, this book will surely remain *the* introductory textbook on semiparametric statistics for students in their first or second year of graduate schools and for almost all applied statisticians and data scientists. The book has covered many important aspects of semiparametric regression. To end our review, we take the liberty to make a short wish-list on topics not covered in HRW but nonetheless we would hope to see either in future edition(s) of HRW or somewhere else:

- *Semiparametric methods for high- or ultrahigh-dimensional data.* One missing piece is semiparametric methods for handling high- or even ultrahigh- dimensional datasets, in which the sample size is much smaller than the number of covariates. On page 101, HRW briefly mentioned the R package `gamsel` could handle high-dimensional data. But we hope future edition(s) of the book or others take on the challenge of covering this topic in detail.
- *Computational issues.* In the age of big data, computational efficiency can no longer be swept under the rug. Sec. 6.9.3 briefly comments on the computational issues of Bayesian paradigm for complicated semiparametric models. It could be an important next step to develop computational efficient Bayesian-semiparametric methods that are ready to be used and presented in a textbook like this remarkable piece.
- *More treatment on missing data.* In most real-life data analysis, missing data is unavoidable. Developing semiparametric methods under MNAR is also an active research area in recent years. Thus, a dedicated book with R implementation for semiparametric methods with missing data, in a similar style to HRW, will be at the top of the wish-list of many practitioners.

Funding

The authors gratefully acknowledge Shanghai NSF grant no.21ZR1431000, Shanghai Science and Technology Commission grant no.21JC1402900 and Major project no.2021SHZDZX0102, NSFC grant no.12090024 and no.12101397 for funding support.

Zixiao Wang
Department of Biostatistics,
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD

Yi Feng
School of Mathematical Sciences,
Shanghai Jiao Tong University
Shanghai, China

Lin Liu
Institute of Natural Sciences, MOE-LSC,
School of Mathematical Sciences, CMA-Shanghai,
SJTU-Yale Joint Center for Biostatistics and Data Science
Shanghai Jiao Tong University
Shanghai, China
linliu@sjtu.edu.cn

References

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. (2021), "Neural Additive Models: Interpretable Machine Learning with Neural Nets," in *Advances in Neural Information Processing Systems* (Vol. 34), pp. 4699–4711. [2286]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, Boca Raton, FL: Chapman and Hall/CRC. [2284]

Ghosal, S., and van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference* (Vol. 44), Cambridge: Cambridge University Press. [2284]

Jacot, A., Gabriel, F., and Hongler, C. (2018), "Neural Tangent Kernel: Convergence and Generalization in Neural Networks," in *Advances in Neural Information Processing Systems* (Vol. 31). [2286]

Little, R. J., and Rubin, D. B. (2019), *Statistical Analysis with Missing Data* (Vol. 793), New York: Wiley. [2286]

Rügamer, D., Baumann, P. F., and Greven, S. (2022), "Selective Inference for Additive and Linear Mixed Models," *Computational Statistics & Data Analysis*, 167, 107350. [2285]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression* (Vol. 12), Cambridge: Cambridge University Press. [2284]

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295. [2285]

Sampling: Design and Analysis, 3rd ed., Sharon L. Lohr, Boca Raton, FL: Chapman & Hall/CRC Press, 2022, xxiii+650 pp., $79.95(H), ISBN 978-0367279509.

A generation of students have been trained on the two previous editions of this sampling text, and the third edition will ensure that this continues. This book is remarkable for its utility for both learners and practitioners, much as Cochran's classic text (1977) was for decades. For teaching, I have been using the text only for a one-semester upper-level graduate course for statistics students. It provides sufficient technical detail, challenging exercises, and up-to-date references to prepare students to conduct research in the field. The preface also outlines subsections of the text that could be used for upper-level undergraduates majoring in technical fields and graduate students outside statistics who need to design and analyze surveys for their own research. I do believe the text could be effectively used for these audiences because of the range of topics covered and types of exercises provided. However, the level of technical detail embedded in even the recommended sections could be intimidating and would require the instructor to explain the salient points carefully.

Sampling practice has undergone profound changes since publication of the second edition of this text in 2009. This edition reflects those changes beautifully. A new chapter on nonprobability sampling provides an overview of the technical details for the newest methods of adjustment. I especially like the presentation, popularized by Meng (2018), of the partitioning of bias of the sample mean in a nonprobability sample. The subsequent discussion and examples inform intuition and should make the chapter's main points accessible to all the text's target audiences. In fact, I believe this chapter, along with its enlightening exercises, would make an excellent curriculum for a short course on nonprobability sampling.

The chapter on nonresponse is the most extensively revised of the existing chapters and expands both the practical and technical discussions of nonresponse methods. It provides a more complete description of response propensities, step-by-step instructions for constructing weighting class and post-stratification weights, more detail on methods of imputation (such as Multiple Imputation by Chained Equations, or MICE), and a section on the various American Association for Public Opinion Research (AAPOR) nonresponse measures. Other updates include replacing some discussion of nonresponse in telephone and mail surveys with discussion of internet surveys, including differences in causes and mitigation of nonresponse, and its effect on estimator bias. The remaining chapters have been thoroughly updated to reflect more recent citations and applications. Most exercises and examples based on datasets from the last century have been replaced with recent ones. An estimate of the percentage of citations published after the last edition (i.e., 2010 or later) in the 48 pages of references was 48%, based on a systematic sample of 100 citations. The American Community Survey has replaced the National Crime Victimization Survey in this edition to illustrate methods used in large scale surveys. Except for minor cases, notation is unchanged, making updating class notes painless for the instructor. The chapters have been signposted better, with additional headings allowing easier navigation.

A new feature of this edition is the inclusion of companion software guides. The two guides for SAS and R provide code showing how to obtain results for most of the examples in the text using the survey procedures (SAS) and packages/functions (R). The guides also provide thorough explanations of the use of options and workarounds that make them better than available software documentation for instruction. This provides great flexibility for instructors whose classes include students with diverse computer training. The best thing is that soft copies of these guides are freely available from the author's website, but reasonably priced hard copies are also available.

My favorite part of the previous editions of this text was its exercises. They are even better in this edition. The chapter exercises are still divided into four types: Introduction, Working with Survey Data, Working with Theory, and Projects and Activities. A substantial number of exercises in previous editions have been removed, especially those based on datasets prior to 2000, so instructors may lose some of their favorites. Some have been updated with newer data sources, but mostly, new problems have been added, so most chapters contain more exercises than before of every type. The exercises have something to interest everyone; they are contemporary (Census household pulse sur-