Journal of Statistical Software
September 2010, Volume 36, Book Review 5. http://www.jstatsoft.org/

Reviewer: Christian Derquenne
Electricité de France

Regression Modeling: Methods, Theory, and Computation with SAS

Michael Panik
Chapman & Hall/CRC, Boca Raton, FL, 2009.
ISBN 978-1-4200-9197-7. 814 pp. USD 99.95.
http://www.crcpress.com/product/isbn/9781420091977

In his book entitled *Regression Modeling: Methods, Theory, and Computation with SAS*, Michael Panik gives a large field of modeling methods starting from simple linear regression to times series or spatial data, but also including robust regression or semi-parametric regression. The author's goal is also to provide, for most methods, some applications on comprehensive data using SAS software, in offering SAS code, output and detailed remarks. Generally, the framework of a chapter is the following: introducing the problem, one or several proposed solutions drawing from the literature, application on examples, implementation with SAS and appendices containing methodological support and exercises. This book has 19 chapters. The first one is specific and therefore analyzed separately: it reviews the fundamentals of probability and statistics. We have chosen a nonlinear discussion for the remaining chapters by defining three themes: the usual Gaussian linear model, methods in the case of misspecified disturbances and other regression models.

The first chapter gives elementary knowledge in probability and integration: random variables, probability distributions (expectation, variance, cumulative distribution function, etc.), the normal probability distribution and derived distributions, and bivariate composition of random variables (conditional independence, covariance, correlation, etc.). This chapter provides statistical inference (sampling, estimation, bias, minimum variance, large sample properties of point estimators, maximum likelihood, statistical hypothesis testing, etc.).

The study of the usual Gaussian linear model begins by the simple linear regression and correlation (Chapter 2). The least squares estimator is introduced, confidence intervals and hypothesis testing concerning slope and intercept, correlation coefficient are discussed. Confidence bands on the precision of the least squares regression equation and prediction of a particular value of response given the predictor are also developed. Lastly, the comparison between two linear regression equations is studied. Chapter 12, on multiple linear regression, is probably the best of this book, notably on the standardized coefficients, the decomposition of sum of squares regression into incremental sums of squares, stepwise regression and multicollinearity. A lot of detailed examples and corresponding SAS code illustrate the different stages. The following chapter, on correlation model, continues the previous one about the

link between correlation and regression models in a natural way, by means of partial correlation and test on individual regression parameters. Chapter 15, about regression on indicator variables, introduces the covariance analysis. Then, the author returns to the comparison between regression equations with interaction between indicator variable and numeric variable. The studied data are modeled with the `REG` procedure.

The second field tackled by Michael Panik is linked to the methods in case of misspecified disturbance of the usual Gaussian linear model. Chapter 3 introduces the main problems, such as: errors are not normally distributed, heterogeneity of variance of residuals, their autocorrelation, dependence between errors and predictors, stochastic explanatory variables and outliers. For each problem, one or several tests are proposed, with a proposition of a suitable model, if the test is rejected. The associated appendices offer the properties and detailed calculus, with aid of pedagogically process. Chapter 9 is about random coefficients regression and gives two aspects of this problem. The first one corresponds to a pure random coefficients model (Hildreth and Houck 1968) which is estimated by a weighted regression to take into account heteroskedasticity. The second aspect is linked to a linear random coefficients mixed model containing fixed effects and random effects. This model is applied on an example with the `MIXED` procedure. The appendices detail the generalized least squares estimator and the generalized linear mixed model. Chapter 14 returns to the multicollinearity problem with the ridge regression. The author offers an interesting and detailed analysis on bias of regression coefficients in the presence of multicollinearity. The `REG` with `RIDGE` option allows to put this regression in application. The appendices provide the properties of the ridge estimator. The example of Malinvaud's data could be more detailed. The robust regression, used when the presence of outliers is suspected, is developed in Chapter 7, with two estimators: Huber's M estimator (Huber 1996) with weighting function and the Rousseeuw's LMS estimator (Rousseeuw 1984) with breakdown point. Although these methods can perform very well, their implementation can be difficult. The author uses a SAS macro to apply the LMS regression, whereas he could use the `ROBUSTREG` procedure which is better in terms of quality of using. Unfortunately, the application is limited to simple regression.

The third field covers a large spectrum of other regression models associated to different type of data and problems. Chapter 4 touches upon nonparametric simple regression under aspect of rank statistic proposed by Theil (1950a,b,c). The slope and intercept coefficients are estimated by the ordinary median or by a weighted median and the associated tests are given. Even if the explanations are illustrated by small examples, neither application nor implementation in SAS are provided. The modeling of categorical response variables is studied by means of binary, ordinal or nominal logit regression in Chapter 5. The author explains the logit dichotomous model very well, notably with odds ratio, interpretation of the slope coefficient and on top of that SAS implementation. The proportional odds ratio is pedagogically introduced with formula and graphics. However, the associated test is not developed enough. The generalized logit model is briefly discussed. `CATMOD` is used to apply this model, whereas the `LOGISTIC` procedure is more suitable when the explanatory variable is numeric. Lastly, it is a pity that the multiple logistic regression is not discussed. The link between frequentist and Bayesian approaches is very well explained in the following chapter, notably with Bayesian regression and a proposal for a Bayesian estimator of slope and intercept parameters, confidence intervals and associated tests. Unfortunately, examples are not developed enough and no SAS code is provided. The fuzzy regression constitutes Chapter 8. Although the review of methods is detailed in a pedagogical way, fuzzy regression,

has little interest because it is not much used in real applications. Moreover as indicated by the author, this approach is not used in presence of outliers and when there is a large variability in data. In addition, no implementation in SAS is given. Chapter 10 is dedicated to robust modeling with L_1 regression and quantile regression. A link could be made with Chapter 7 on robust regression. In addition, the author uses the SAS `LAV` macro written in SAS/IML, whereas again, `ROBUSTREG` could be applied. Although the study of spatial data is complex, Michael Panik explains it elegantly in Chapter 11. He has chosen the geographically weighted regression developed by Hastie and Tibshirani (1990). He applies this method on a simple example and proposes some SAS/IML code. In Chapter 16, the spline transformation is pedagogically introduced by the author, with an orthogonal polynomials method. Again interaction is discussed, but this time between numeric variables. The stepwise regression is proposed to select the more convenient polynomial model. The SAS code is very clear. The following chapter is on a non-trivial problem: mix of parametric and nonparametric aspects in a model. Firstly, an approach by differencing is proposed, but it is limited because the data must be structurally ordered. Then, the LOWESS (locally weighted scatter plot smoothing) is described and applied with the `LOWESS` procedure. Nonlinear regression is studied in Chapter 18, in giving difference between purely nonlinear and intrinsically linear. Tests based on maximum likelihood, Wald's statistic and Lagrange's multiplier are used for testing the nonlinear parameter restrictions. `NLIN` and `MODEL` procedures are applied on several examples. The last chapter is devoted to time series modeling. The author provides the different steps of evaluation and modeling: unit root testing of nonstationarity, Dicker-Fuller statistics, cointegration and ARMA. Although examples and results are fully discussed, the problem of forecast validation is not introduced.

In his book, Michael Panik takes up many aspects of modeling with a pedagogical approach helping the reader to understand the process of a problem and the proposed methods. The appendices are useful to the readers that may want to broaden their knowledge. However, some parts are more achieved than others, maybe varying with author's degree of expertise, as illustrated by the different examples and the proposed SAS codes, procedures and macros. For instance, multiple linear regression is fully detailed, whereas the chapters on nonparametric, fuzzy and Bayesian regression do not offer implementation in SAS. Lastly, this book is a very good tool for students and teachers in statistics, but also for researchers wishing to improve their knowledge in statistical modeling.

References

Hastie T, Tibshirani R (1990). *Generalized Additive Models*. Chapman and Hall, New York.

Hildreth C, Houck JP (1968). "Some Estimates for a Linear Model with Random Coefficients." *Journal of the American Statistical Association*, **63**, 584–595.

Huber PJ (1996). *Robust Statistical Procedures*. 2nd edition. SIAM, Philadelphia.

Rousseeuw PJ (1984). "Least Median of Squares Regression." *Journal of the American Statistical Association*, **79**, 871–880.

Theil H (1950a). "A Rank-Invariant Method of Linear and Polynomial Regression Analysis, I." *Koninklijke Nederlandse Akademie van Wetenschappen, Proceedings A*, **53**, 386–392.

Theil H (1950b). "A Rank-Invariant Method of Linear and Polynomial Regression Analysis, II." *Koninklijke Nederlandse Akademie van Wetenschappen, Proceedings A*, **53**, 521–525.

Theil H (1950c). "A Rank-Invariant Method of Linear and Polynomial Regression Analysis, III." *Koninklijke Nederlandse Akademie van Wetenschappen, Proceedings A*, **53**, 1397–1412.

Reviewer:

Christian Derquenne
Electricité de France
Research and Development
1, avenue du Général de Gaulle
92141 Clamart Cedex, France
E-mail: christian.derquenne@edf.fr

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 36, Book Review 5 *Published:* 2010-09-21
September 2010