Chandan Mukherjee, Howard White, and Marc Wuyts. *Econometrics and Data Analysis for Developing Countries.* London: Routledge, 1998. 496 pages. Paperback, £ 24.99.

The book describes, in detail, rigorous and accessible methods for modern data analysis. Traditionally, data analysis was not considered important in model specification and estimation. The theoretical models were estimated against given data. However, currently, the emphasis is on in-depth data analysis before utilising it for model estimation. Utilising the data from developing countries, the authors discuss suitable techniques in different situations.[1] After the introductory chapter, the book is divided into the following five parts: Foundation of Data Analysis; Regression and Data Analysis; Analysing Cross-section Data; Regression with Time-series Data; and Simultaneous Equation Models.

The first part discusses the problems in model estimation, using data infected with problems like skewness and outliers, etc. In general, data are assumed to be normally distributed and invalidity of the assumption may give unreliable/biased results. Therefore, initial model specification and results can not be taken at face value. The tools of Exploratory Data Analysis (EDA), i.e., graphical method, scatter plot, and statistical tests, are used to deal with the problems. Generally, least square estimators are assumed to be appropriate in all circumstances. The fact is that least square summaries (the first two moments) are sensitive to the presence of outliers in data. The authors give median and interquartile range, respectively, as alternatives to these moments. Diagnostic tools to test for different problems allow data to play an active part in model specification. To look at the shape of an empirical distribution, EDA starts with numerical summaries of the data, i.e., median, mode, Q_L, Q_u, and two extreme values, X_u and X_L. The box plot, the best way to check data, shows the basic structure of the data, i.e., its location, spread, skewness, tail length, and outliers. The authors recommend the use of skewness and Kurtosis statistics to judge whether the normality assumption holds in practice.

The second part of the book is devoted to simple regression, partial regression, sweeping out concept, and model selection. This part shows that inappropriate application of the least square principle leads to nonsensical results. Often, researchers arrive, through trial and error, at a regression which looks good on paper, and do not bother about its fragile foundations. A diagnostic tool, exploratory band regression, which is non-resistant to the pull of data points, is used to check non-linearity in the model. But its coefficients can not be used to interpret an economic relationship. The authors show that outliers, leverage, and influential point play a very important role in applied work, as the least square method does not work when assumptions are seriously violated. Special statistics; studentised residuals, hat statistics, and DFBETA, are designed to detect them. For multiple regression, partial regression and its plot are used

[1] To give an opportunity to the reader to practice, the authors provide the data on floppy disks.

to look deep into the structure of data. Partial regression plot allows us to look carefully at the scatter plot that underscores a particular regression coefficient in a multiple regression. This part also shows the importance of the concept of sweeping out as it teaches us how a coefficient in a multiple regression is arrived at by controlling, not keeping constant, for the linear influence of the other regression in the model. This part discusses the general-to-specific modelling, the fragility analysis known as extreme-bound analysis. General-to-specific modelling allows for data choice between rival models. Restriction test is an essential tool for the general-to-specific modelling. Fragility analysis compares regression coefficients across neighbouring specifications and conditional bounds on regression coefficients which can be of great help in guarding against making fragile inferences. Variable selection in model specification is a challenging task and applied research has serious consequences for the validity of the inferences. This part discusses the problems of the omitted relevant variable and the added irrelevant variables also. Griffin's well-known model is used to explain it. Hypothesis testing is very important for statistical inferences. In data analysis, many inferences drawn on the basis of testing these hypotheses may be proven worthless if normality assumption is not valid. So before drawing conclusions from a model, it is necessary to make sure that its foundations are sound. This part suggests examination of the residuals for clues to probable misspecification.

The third part mainly analyses cross-section data problems and their solutions. Cross-section data are mainly infected with heteroscedasticity. The authors apply diagnostic tools, visual and statistical tests, to residuals to detect heteroscedasticity. If data are heteroscedastic, then standard formulas for the errors will not be applicable. This part presents many alternatives to overcome this problem. Power transformation is very effective to eliminate heteroscedasticity. This part clearly shows that model respecification is the first course of action. In case of genuine heteoscedasticity, the method of weighted least squares (WLS) provides efficient estimates of regression and valid inferences. If WLS is not possible, then standard error may be calculated through White's heteroscedastic consistent standard errors (HCSEs).

In applied research, qualitative response data occupy an important place. This type of modelling is very important, and this part discusses such application to real data in empirical analysis. Contingency tables are used to see the interdependency of two variables. Then this analysis is extended to discuss logit modelling in the context of multiway contingency tables. Sometimes calculations involved in explaining the estimates are difficult (in logit model) but they are very useful to design the plots for conditional effect and poorness of fit, which are helpful in explaining logit regression and to show outliers in data. The book helps us by giving practical examples.

The fourth part is devoted to time-series data problems. Time series is assumed to behave as a set of data randomly sampled without any connection between successive observation. But, in fact, the ordering of the data matters a great deal as the first and second moments are time-dependent. Regression of one random walk on another is

likely to yield a significant result even if the two series are totally unrelated. On the other hand, non-stationarity may produce the autocorrelation problem. Formal tests are required to check non-stationarity. One possible way to avoid spurious regression is through transforming data so as to make the data stationary. Next, it discusses misspecification, incorrect functional form, and omitted variables as the main reason to produce correlated successive error terms even in cross-section data; although we do not expect it in cross-section data, as unlike time-series data, they have no natural ordering. Lastly, it discusses cointegration and the error-correction model and their importance in applied research. This part shows that the least squares principle works if two non-stationary series are cointegrated. For this purpose, we have to test for cointegration. The full dynamic model is estimated as an Error Correction Model (ECM) and provides estimates for the long-run relationship. The book presents ADF, DF, and CRDW statistics to test for non-stationarity and order of cointegration. This chapter also explores the model dynamics with some simulations. Interpretation of the results from ECM is very technical, and the book provides the facility to interpret results from real estimated models.

The final part discusses estimation of simultaneous equation models with limited and full information methods. It discusses the simultaneous equation bias problem, Granger's and Sims' test for causality, and the exogeneity concept. Granger causality is used for strong exogeneity test and the Hausman test is applied for weak exogeneity test. Rank and order condition are important to determine the appropriate means of estimation, as Indirect Least Squares (ILS) can not be used when the equation is over-identified. Instrumental Variables (IV) and Two Stage Least Squares (2SLS) should be used in that case. All these methods are limited information techniques. The inefficiency of the estimation arises since the method does not exploit two possible pieces of information: (a) Cross-correlation between error terms in different equations; (b) the pre-determined variable omitted from the equations being estimated may also be omitted from other equations. Full information techniques estimate all equations simultaneously. Seemingly Unrelated Regression equation takes into account the first problem and the Three Stage Least Squares (3SLS) method takes account of both.

The book is quite outstanding for the clarity and accuracy of its detailed arguments. It permits the reader to see the close connection between EDA and correct policy formulation through application of the correct model to data. The implication for statistical analysis is that we should not apply the regression model to the data straight away. The formulation of the models after exploratory data analysis increases the significance of the results and enhances their power from the economist's and the policy-maker's point of view.

The main message of the book is that exploratory data analysis before estimation and residual analysis after estimation are desirable features for good researchers. By analysing these two elements, we can construct the best models and their estimates may in turn yield good economic policies. The authors manage to reveal what the problems

are and how they can be solved by providing useful techniques, which require familiarity with mathematics and statistics plus computer programming tools. Another significant contribution of the book is its use of empirical work to refer to data problems. Topics such as the contingency table, cointegration analysis, and ECM bring the book up to date in several areas. I believe that the book will provide new insight to many readers who are already familiar with most of the material covered, it lacks mathematical proof of some of the formulas offered. It does give information about the statistical packages used to obtain such exploratory statistics. As the book has been written by economists, it explains more clearly the problems with results gotten through data mining.

Rizwana Siddiqui

Pakistan Institute of Development Economics,
Islamabad.