Regression Analysis of Count Data, Second Edition

Students in both social and natural sciences often seek regression methods to explain the frequency of events, such as visits to a doctor, auto accidents, or new patents awarded. This book provides the most comprehensive and up-to-date account of models and methods to interpret such data. The authors have conducted research in the field for more than 25 years. In this book, they combine theory and practice to make sophisticated methods of analysis accessible to researchers and practitioners working with widely different types of data and software in areas such as applied statistics, econometrics, marketing, operations research, actuarial studies, demography, biostatistics, and quantitative social sciences. The book may be used as a reference work on count models or by students seeking an authoritative overview. Complementary material in the form of data sets, template programs, and bibliographic resources can be accessed on the Internet through the authors' homepages. This second edition is an expanded and updated version of the first, with new empirical examples and more than two hundred new references added. The new material includes new theoretical topics, an updated and expanded treatment of cross-section models, coverage of bootstrap-based and simulation-based inference, expanded treatment of time series, multivariate and panel data, expanded treatment of endogenous regressors, coverage of quantile count regression, and a new chapter on Bayesian methods.

A. Colin Cameron is Professor of Economics at the University of California, Davis. His research and teaching interests span a range of topics in microeconometrics. He is a past director of the Center on Quantitative Social Science at UC Davis and is currently an associate editor of the *Stata Journal* and *Journal of Econometric Methods.*

Pravin K. Trivedi is Professor of Economics, University of Queensland, and Distinguished Professor Emeritus and J. H. Rudy Professor Emeritus of Economics at Indiana University, Bloomington. During his academic career, he has taught undergraduate and graduate-level econometrics in the United States, Europe, and Australia. His research interests are in microeconometrics and health economics. He served as coeditor of the *Econometrics Journal* from 2000 to 2007 and has been on the board of *Journal of Applied Econometrics* since 1988.

Professors Cameron and Trivedi are coauthors of the first edition of *Regression Analysis of Count Data* (Cambridge University Press, 1998), *Microeconometrics: Methods and Applications* (Cambridge University Press, 2005), and *Microeconomics Using Stata Revised Edition* (2010).

Econometric Society Monographs

Editors:

Rosa L. Matzkin, University of California, Los Angeles
George J. Mailath, University of Pennsylvania

The Econometric Society is an international society for the advancement of economic theory in relation to statistics and mathematics. The Econometric Society Monograph series is designed to promote the publication of original research contributions of high quality in mathematical economics and theoretical and applied econometrics.

Other titles in the series:

G. S. Maddala, *Limited dependent and qualitative variables in econometrics*, 9780521241434, 9780521338257
Gerard Debreu, *Mathematical economics: Twenty papers of Gerard Debreu*, 9780521237369, 9780521335614
Jean-Michel Grandmont, *Money and value: A reconsideration of classical and neoclassical monetary economics*, 9780521251419, 9780521313643
Franklin M. Fisher, *Disequilibrium foundations of equilibrium economics*, 9780521378567
Andreu Mas-Colell, *The theory of general equilibrium: A differentiable approach*, 9780521265140, 9780521388702
Truman F. Bewley, Editor, *Advances in econometrics – Fifth World Congress (Volume I)*, 9780521467261
Truman F. Bewley, Editor, *Advances in econometrics – Fifth World Congress (Volume II)*, 9780521467254
Hervé Moulin, *Axioms of cooperative decision making*, 9780521360555, 9780521424585
L. G. Godfrey, *Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches*, 9780521424592
Tony Lancaster, *The econometric analysis of transition data*, 9780521437899
Alvin E. Roth and Marilda A. Oliviera Sotomayor, Editors, *Two-sided matching: A study in game-theoretic modeling and analysis*, 9780521437882
Wolfgang Härdle, *Applied nonparametric regression*, 9780521429504
Jean-Jacques Laffont, Editor, *Advances in economic theory – Sixth World Congress (Volume I)*, 9780521484596
Jean-Jacques Laffont, Editor, *Advances in economic theory – Sixth World Congress (Volume II)*, 9780521484602
Halbert White, *Estimation, inference and specification*, 9780521252805, 9780521574464
Christopher Sims, Editor, *Advances in econometrics – Sixth World Congress (Volume I)*, 9780521444590, 9780521566100
Christopher Sims, Editor, *Advances in econometrics – Sixth World Congress (Volume II)*, 9780521444606, 9780521566094
Roger Guesnerie, *A contribution to the pure theory of taxation*, 9780521629560
David M. Kreps and Kenneth F. Wallis, Editors, *Advances in economics and econometrics – Seventh World Congress (Volume I)*, 9780521589833
David M. Kreps and Kenneth F. Wallis, Editors, *Advances in economics and econometrics – Seventh World Congress (Volume II)*, 9780521589826
David M. Kreps and Kenneth F. Wallis, Editors, *Advances in economics and econometrics – Seventh World Congress (Volume III)*, 9780521580137, 9780521589819
Donald P. Jacobs, Ehud Kalai, and Morton I. Kamien, Editors, *Frontiers of research in economic theory: The Nancy L. Schwartz Memorial Lectures, 1983–1997*, 9780521632225, 9780521635387
Steinar Strom, Editor, *Econometrics and economic theory in the 20th century: The Ragnar Frisch Centennial Symposium*, 9780521633239, 9780521633659

Continued on page following the index

Regression Analysis of Count Data

Second Edition

A. Colin Cameron
University of California, Davis

Pravin K. Trivedi
University of Queensland
Indiana University, Bloomington, Emeritus

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107667273

First edition published 1998
Second edition published 2013
Reprinted 2013

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication data
Cameron, Adrian Colin.
Regression analysis of count data / A. Colin Cameron, University of California, Davis, Pravin K. Trivedi, Indiana University, Bloomington. – Second edition.
pages cm. – (Econometric society monographs)
Includes bibliographical references and index.
ISBN 978-1-107-01416-9 (hardback) – ISBN 978-1-107-66727-3 (paperback)
1. Regression analysis. 2. Econometrics. I. Trivedi, P. K. II. Title.
QA278.2.C36 2013
519.5′36–dc23 2012043350

ISBN 978-1-107-01416-9 Hardback
ISBN 978-1-107-66727-3 Paperback

To Michelle and Bhavna

Contents

List of Figures

List of Tables

Preface

Since *Regression Analysis of Count Data* was published in 1998, significant new research has contributed to the range and scope of count data models. This growth is reflected in many new journal articles, fuller coverage in textbooks, and wide interest in and availability of software for handling count data models. These developments (to which we have also contributed) have motivated us to revise and expand the first edition. Like the first edition, this volume reflects an orientation toward practical data analysis.

The revisions in this edition have affected all chapters. First, we have corrected the typographical and other errors in the first edition, improved the graphics throughout, and where appropriate we have provided a cleaner and simpler exposition. Second, we have revised and relocated material that seemed better placed in a different location, mostly within the same chapter though occasionally in a different chapter. For example, material in Chapter 4 (generalized count models), Chapter 8 (multivariate counts), and Chapter 13 (measurement errors) has been pruned and rearranged so the more mainstream topics appear earlier and the more marginal topics have disappeared altogether. For similar reasons bootstrap inference has moved to Chapter 2 from Chapter 5. Our goal here has been to improve quality of synthesis and accessibility of material to the reader. Third, the final few chapters have been reordered. Chapter 10 (endogeneity and selection) has moved up from Chapter 11. It replaces the measurement error chapter that now appears as Chapter 13. Chapter 11 now covers flexible parametric models (previously Chapter 12). And the current Chapter 12, which covers Bayesian methods, is a new addition. Fourth, we have removed material that was of marginal interest and replaced it with material of potentially greater interest, especially to practitioners. For example, as barriers to implementation of more computer-intensive methods have come down, we have liberally sprinkled illustrations of simulation-based methods throughout the book. Fifth, bibliographic notes at the end of every chapter have been refreshed to include newer references and topics. Sixth, we have developed an almost complete set of computer code for the examples in this book.

The first edition has been expanded by about 35%. This expansion reflects the addition of a new Chapter 12 on Bayesian methods as well as significant

additions to most other chapters. Chapter 2 has new sections on robust inference and empirical likelihood and includes material on the bootstrap and generalized estimating equations. In Chapter 3 and throughout the book, the term "pseudo-ML" has been changed to "quasi-ML" and robust standard errors are computed using the robust sandwich form. Chapter 4 improves the coverage and discussion of how many alternative count models relate to each other. Censored, truncated, hurdle, zero-inflated, and especially finite mixture models are now covered in greater depth, with a more uniform notation, and hierarchical count models and models with cross-sectional and spatial dependence have been newly added. Chapter 5 moves up presentation of methods for discrimination among nonnested models. Chapter 6 adds a new empirical example of fertility data that poses a fresh challenge to count data modelers. The time series coverage in Chapter 7 has been expanded to include more recently developed models, and there is some rearrangement so that the most often used models appear first. The coverage of multivariate count models in Chapter 8 uses a broader and more modern range of dependence concepts and provides a lengthy treatment of parametric copula-based models. The survey of count data panel models in Chapter 9 gives greater emphasis to moment-based approaches and has a more comprehensive coverage of dynamic panels, the role of initial conditions, conditionally correlated random effects, flexible functional forms, and specification tests. Chapter 10 provides an improved exposition of models with endogeneity and selection, including consideration of latent factor and two-part models as well as simulation-based inference and control function estimators. A major new topic in Chapter 11 is quantile regression models for count data, and the coverage of semiparametric and nonparametric methods has been expanded and updated. As previously mentioned, the new Chapter 12 covers Bayesian analysis of count models, providing an entry to the world of Markov chain Monte Carlo analysis of count models. Finally, Chapter 13 provides a comprehensive survey of measurement error models for count data. As a result of the expanded coverage of old topics and appearance of new ones, the bibliography is now significantly larger and includes more than two hundred additional new references.

To emphasize its empirical orientation, the book has added many new examples based on real data. These examples are scattered throughout the book, especially in Chapters 6–12. In addition, we have a number of examples based on simulated data. Researchers, instructors, and students interested in replicating our results can obtain all the data and computer programs used to produce the results given in this book via Internet from our respective personal web sites.

This revised and expanded second edition draws extensively from our jointly authored research undertaken with Partha Deb, Jie Qun Guo, Judex Hyppolite, Tong Li, Doug Miller, Murat Munkin, and David Zimmer. We thank them all. We also thank Joao Santos Silva for detailed comments on Chapter 10 and Jeff Racine for detailed comments on Chapter 11. The series editor Rosa Matzkin and an anonymous reviewer provided helpful guidance and suggestions for

improvements for which we are grateful. As for the first edition, it is a pleasure to acknowledge the overall editorial direction and encouragement of Scott Parris of Cambridge University Press throughout the multiyear process of bringing the project to completion.

A. Colin Cameron
Davis, CA

Pravin K. Trivedi
Bloomington, IN
August 2012

Preface to the First Edition

This book describes regression methods for count data, where the response variable is a non-negative integer. The methods are relevant for analysis of counts that arise in both social and natural sciences.

Despite their relatively recent origin, count data regression methods build on an impressive body of statistical research on univariate discrete distributions. Many of these methods have now found their way into major statistical packages, which has encouraged their application in a variety of contexts. Such widespread use has itself thrown up numerous interesting research issues and themes, which we explore in this book.

The objective of the book is threefold. First, we wish to provide a synthesis and integrative survey of the literature on count data regressions, covering both the statistical and econometric strands. The former has emphasized the framework of generalized linear models, exponential families of distributions, and generalized estimating equations, while the latter has emphasized nonlinear regression and generalized method of moment frameworks. Yet between them there are numerous points of contact which can be fruitfully exploited. Our second objective is to make sophisticated methods of data analysis more accessible to practitioners with different interests and backgrounds. To this end we consider models and methods suitable for cross-section, time series, and longitudinal data. Detailed analyses of several data sets as well as shorter illustrations, implemented from a variety of viewpoints, are scattered throughout the book to put empirical flesh on theoretical or methodological discussion. We draw on examples from, and give references to, works in many applied areas. Our third objective is to highlight the potential for further research by discussion of issues and problems that need more analysis. We do so by embedding count data models in a larger body of econometric and statistical work on discrete variables and, more generally, on nonlinear regression.

The book can be divided into four parts. The first two chapters contain introductory material on count data and a comprehensive review of statistical methods for nonlinear regression models. Chapters 3, 4, 5, and 6 present models and applications for cross-section count data. Chapters 7, 8, and 9 present methods for data other than cross-section data, namely time series, multivariate

and longitudinal or panel data. Chapters 10, 11, and 12 present methods for common complications, including measurement error, sample selection and simultaneity, and semiparametric methods. Thus the coverage of the book is qualitatively similar to that in a complete single book on linear regression models.

The book is directed toward researchers, graduate students, and other practitioners in a wide range of fields. Because of our own background in econometrics, the book emphasizes issues arising in econometric applications. Our training and background also influence the organizational structure of the book. But areas outside econometrics are also considered. The essential prerequisite for this book is familiarity with the linear regression model using matrix algebra. The material in the book should be accessible to people with a background in regression and statistical methods up to the level of a standard first-year graduate econometrics text such as Greene's *Econometric Analysis*. While basic count data methods are included in major statistical packages, more advanced analysis can require programming in languages such as Splus, Gauss, or Matlab.

Our own entry into the field of count data models dates back to the early 1980s when we embarked on an empirical study of the demand for health insurance and health care services at the Australian National University. Since then we have been involved in many empirical investigations that have influenced our perceptions of this field. We have included numerous data analytic discussions in this volume to reflect our own interest and those of readers interested in real data applications. The data sets, computer programs, and related materials used in this book will be available through Internet access to our individual web sites. These materials supplement and complement this book and will help new entrants to the field, especially graduate students, to make a relatively easy start.

We have learned much on modeling count data through collaborations with coauthors, notably Partha Deb, Shiferaw Gurmu, Per Johansson, Kajal Mukhopadhyay, and Frank Windmeijer. The burden of writing this book has been eased by help from many colleagues, coauthors, and graduate students. In particular, we thank the following for their generous attention, encouragement, help, and comments on earlier drafts of various chapters: Kurt Brännäs, David Hendry, Primula Kennedy, Tony Lancaster, Scott Long, Grayham Mizon, Neil Shephard, and Bob Shumway, in addition to the coauthors already mentioned. We especially thank David Hendry and Scott Long for their detailed advice on manuscript preparation using Latex software and Scientific Workplace. The manuscript has also benefited from the comments of a referee and the series editor, Alberto Holly, and from the guidance of Scott Parris of Cambridge University Press.

Work on the book was facilitated by periods spent at various institutions. The first author thanks the Department of Statistics and the Research School of Social Sciences at the Australian National University, the Department of Economics at Indiana University in Bloomington, and the University of California in Davis for support during extended leaves to these institutions in 1995 and

1996. The second author thanks Indiana University and European University Institute, Florence, for support during his tenure as Jean Monnet Fellow in 1996, which permitted a period away from regular duties. For shorter periods of stay that allowed us to work jointly, we thank the Department of Economics at Indiana University, SELAPO at University of Munich, and the European University Institute.

Finally we would both like to thank our families for their patience and forbearance, especially during the periods of intensive work on the book. This work would not have been possible at all without their constant support.

A. Colin Cameron
Davis, CA

Pravin K. Trivedi
Bloomington, IN