

Regression Analysis of Count Data

Second Edition

A. Colin Cameron
Department of Economics
University of California
Davis, CA 95616, U.S.A.
Telephone: 530-752-8396
Fax: 530-752-9382
E-mail: accameron@ucdavis.edu

Pravin K. Trivedi
Department of Economics
Indiana University
Bloomington, IN 47405, U.S.A.
Telephone: 812-855-3567
Fax: 812-855-3736
E-mail: trivedi@indiana.edu

April 2012

Copyright © 2012 by A. Colin Cameron and Pravin K. Trivedi.
All rights reserved.

Please do not copy without permission from the authors.

Contents

List of Figures	ix
List of Tables	xii
Preface	xvii
1 Introduction	1
1.1 Poisson Distribution and its Characterizations	3
1.2 Poisson Regression	8
1.3 Examples	10
1.4 Overview of Major Issues	17
1.5 Bibliographic Notes	19
2 Model Specification and Estimation	21
2.1 Introduction	21
2.2 Example and Definitions	22
2.3 Likelihood-Based Models	24
2.4 Generalized Linear Models	29
2.5 Moment-Based Models	39
2.6 Testing	47
2.7 Robust Inference	57
2.8 Derivation of Results	59
2.9 Bibliographic Notes	65
2.10 Exercises	65
3 Basic Count Regression	67
3.1 Introduction	67
3.2 Poisson MLE, QMLE, and GLM	69
3.3 Negative Binomial MLE and QGPMLE	78
3.4 Overdispersion Tests	86
3.5 Use of Regression Results	89

3.6	Ordered and Other Discrete-Outcome Models	95
3.7	Other Models	99
3.8	Iteratively Reweighted Least Squares	104
3.9	Bibliographic Notes	105
3.10	Exercises	106
4	Generalized Count Regression	107
4.1	Introduction	107
4.2	Mixture Models	108
4.3	Truncated Counts	123
4.4	Censored Counts	128
4.5	Hurdle Models	130
4.6	Zero-Inflated Count Models	134
4.7	Hierarchical Models	136
4.8	Finite Mixtures and Latent Class Analysis	139
4.9	Count Models with Cross-sectional Dependence	150
4.10	Models Based on Waiting Time Distributions	155
4.11	Katz, Double Poisson and Generalized Poisson	160
4.12	Derivations	164
4.13	Bibliographic Notes	167
4.14	Exercises	168
5	Model Evaluation and Testing	171
5.1	Introduction	171
5.2	Residual Analysis	172
5.3	Goodness of Fit	182
5.4	Discriminating among Nonnested Models	189
5.5	Tests for Overdispersion	193
5.6	Conditional Moment Specification Tests	199
5.7	Derivations	212
5.8	Bibliographic Notes	213
5.9	Exercises	214
6	Empirical illustrations	217
6.1	Introduction	217
6.2	Background	218
6.3	Analysis of Demand for Health Care	220
6.4	Analysis of Recreational Trips	236
6.5	Analysis of Fertility Data	245
6.6	Model Selection Criteria: A Digression	248

6.7	Concluding Remarks	250
6.8	Bibliographic Notes	251
6.9	Exercises	252
7	Time Series Data	253
7.1	Introduction	253
7.2	Models for Time Series Data	254
7.3	Static Count Regression	258
7.4	Serially Correlated Heterogeneity Models	265
7.5	Autoregressive Models	270
7.6	Integer-valued ARMA models	274
7.7	State Space Models	278
7.8	Hidden Markov Models	280
7.9	Dynamic Ordered Probit Model	282
7.10	Discrete ARMA Models	283
7.11	Applications	284
7.12	Derivations	289
7.13	Bibliographic Notes	290
7.14	Exercises	291
8	Multivariate Data	293
8.1	Introduction	293
8.2	Characterizing and Generating Dependence	294
8.3	Sources of Dependence	299
8.4	Multivariate Count Models	299
8.5	Copula-based Models	305
8.6	Moment-based Estimation	313
8.7	Testing for Dependence	315
8.8	Mixed Multivariate Models	320
8.9	Empirical Example	323
8.10	Derivations	324
8.11	Bibliographic Notes	325
9	Longitudinal Data	327
9.1	Introduction	327
9.2	Models for Longitudinal Data	328
9.3	Population Averaged Models	334
9.4	Fixed Effects Models	337
9.5	Random Effects Models	345
9.6	Discussion	349
9.7	Specification Tests	351

9.8	Dynamic Longitudinal Models	353
9.9	Endogenous Regressors	360
9.10	More Flexible Functional Forms for Longitudinal Data	361
9.11	Derivations	363
9.12	Bibliographic Notes	365
9.13	Exercises	365
10	Endogenous Regressors and Selection	369
10.1	Introduction	369
10.2	Endogeneity in Recursive Models	370
10.3	Selection Models for Counts	372
10.4	Moment-based Methods for Endogenous Regressors	379
10.5	Example: Doctor Visits and Health Insurance	385
10.6	Selection and Endogeneity in Two-Part Models	388
10.7	Alternative Sampling Frames	389
10.8	Bibliographic Notes	393
11	Flexible Methods for Counts	395
11.1	Introduction	395
11.2	Flexible Distributions using Series Expansions	396
11.3	Flexible Models of the Conditional Mean	403
11.4	Flexible Models of the Conditional Variance	407
11.5	Quantile Regression for Counts	413
11.6	Nonparametric Methods	416
11.7	Efficient Moment-Based Estimation	418
11.8	Analysis of Patent Counts	423
11.9	Derivations	426
11.10	Bibliographic Notes	427
12	Bayesian Methods for Counts	429
12.1	Introduction	429
12.2	Bayesian Approach	429
12.3	Poisson Regression	432
12.4	Markov chain Monte Carlo methods	434
12.5	Count models	440
12.6	Roy Model for Counts	443
12.7	Bibliographic Notes	446
13	Measurement Errors	447
13.1	Introduction	447
13.2	Measurement Errors in Regressors	448

13.3 Measurement Errors in Exposure	458
13.4 Measurement Errors in Counts	463
13.5 Underreported Counts	466
13.6 Underreported and Overreported Counts	471
13.7 Simulation Example: Poisson with Mismeasured Regressor	473
13.8 Derivations	474
13.9 Bibliographic Notes	476
13.10 Exercises	476
A Notation and acronyms	479
B Functions, distributions and moments	483
B.1 Gamma function	483
B.2 Some distributions	484
B.3 Moments of truncated Poisson	486
C Software	487
References	489

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, the current version reflects an orientation towards 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 while the more marginal topics have disappeared altogether. For similar reasons bootstrap inference has moved from Chapter 5 to Chapter 2. 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 which 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 25 per cent. 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 material on the bootstrap and generalized estimating equations now appears in this chapter. 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 considerably expanded and updated. As previously mentioned, the new Chapter 12 covers Bayesian analysis of count model, 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 a 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. Jeff Racine provided valuable advice for Chapter 11. We thank them all.

A. Colin Cameron

Davis, CA

Pravin K. Trivedi

Bloomington, IN

April 2012