

Regression Analysis of Count Data

Second Edition

A. Colin Cameron
University of California, Davis

Pravin K. Trivedi
Indiana University, Bloomington



CAMBRIDGE
UNIVERSITY PRESS

Contents

<i>List of Figures</i>	<i>page</i> xiii
<i>List of Tables</i>	xv
<i>Preface</i>	xix
<i>Preface to the First Edition</i>	xxiii
1 Introduction	1
1.1 Poisson Distribution and Its Characterizations	3
1.2 Poisson Regression	9
1.3 Examples	11
1.4 Overview of Major Issues	17
1.5 Bibliographic Notes	20
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	40
2.6 Testing	48
2.7 Robust Inference	58
2.8 Derivation of Results	61
2.9 Bibliographic Notes	67
2.10 Exercises	67
3 Basic Count Regression	69
3.1 Introduction	69
3.2 Poisson MLE, QMLE, and GLM	71
3.3 Negative Binomial MLE and QGPMLE	80
3.4 Overdispersion Tests	89
3.5 Use of Regression Results	92

3.6	Ordered and Other Discrete-Outcome Models	98
3.7	Other Models	102
3.8	Iteratively Reweighted Least Squares	108
3.9	Bibliographic Notes	108
3.10	Exercises	109
4	Generalized Count Regression	111
4.1	Introduction	111
4.2	Mixture Models	112
4.3	Truncated Counts	128
4.4	Censored Counts	133
4.5	Hurdle Models	136
4.6	Zero-Inflated Count Models	139
4.7	Hierarchical Models	142
4.8	Finite Mixtures and Latent Class Analysis	144
4.9	Count Models with Cross-Sectional Dependence	156
4.10	Models Based on Waiting Time Distributions	161
4.11	Katz, Double Poisson, and Generalized Poisson	167
4.12	Derivations	171
4.13	Bibliographic Notes	174
4.14	Exercises	175
5	Model Evaluation and Testing	177
5.1	Introduction	177
5.2	Residual Analysis	178
5.3	Goodness of Fit	188
5.4	Discriminating among Nonnested Models	196
5.5	Tests for Overdispersion	200
5.6	Conditional Moment Specification Tests	207
5.7	Derivations	220
5.8	Bibliographic Notes	222
5.9	Exercises	223
6	Empirical Illustrations	225
6.1	Introduction	225
6.2	Background	226
6.3	Analysis of Demand for Health Care	228
6.4	Analysis of Recreational Trips	245
6.5	Analysis of Fertility Data	253
6.6	Model Selection Criteria: A Digression	257
6.7	Concluding Remarks	260
6.8	Bibliographic Notes	260
6.9	Exercises	261

7	Time Series Data	263
7.1	Introduction	263
7.2	Models for Time Series Data	264
7.3	Static Count Regression	268
7.4	Serially Correlated Heterogeneity Models	276
7.5	Autoregressive Models	281
7.6	Integer-Valued ARMA Models	284
7.7	State Space Models	289
7.8	Hidden Markov Models	291
7.9	Dynamic Ordered Probit Model	293
7.10	Discrete ARMA Models	294
7.11	Applications	295
7.12	Derivations	301
7.13	Bibliographic Notes	302
7.14	Exercises	302
8	Multivariate Data	304
8.1	Introduction	304
8.2	Characterizing and Generating Dependence	305
8.3	Sources of Dependence	310
8.4	Multivariate Count Models	311
8.5	Copula-Based Models	317
8.6	Moment-Based Estimation	325
8.7	Testing for Dependence	327
8.8	Mixed Multivariate Models	333
8.9	Empirical Example	336
8.10	Derivations	338
8.11	Bibliographic Notes	339
9	Longitudinal Data	341
9.1	Introduction	341
9.2	Models for Longitudinal Data	342
9.3	Population Averaged Models	349
9.4	Fixed Effects Models	351
9.5	Random Effects Models	360
9.6	Discussion	364
9.7	Specification Tests	366
9.8	Dynamic Longitudinal Models	368
9.9	Endogenous Regressors	378
9.10	More Flexible Functional Forms for Longitudinal Data	379
9.11	Derivations	381
9.12	Bibliographic Notes	383
9.13	Exercises	384

10	Endogenous Regressors and Selection	385
10.1	Introduction	385
10.2	Endogeneity in Recursive Models	386
10.3	Selection Models for Counts	388
10.4	Moment-Based Methods for Endogenous Regressors	397
10.5	Example: Doctor Visits and Health Insurance	402
10.6	Selection and Endogeneity in Two-Part Models	406
10.7	Alternative Sampling Frames	407
10.8	Bibliographic Notes	412
11	Flexible Methods for Counts	413
11.1	Introduction	413
11.2	Flexible Distributions Using Series Expansions	414
11.3	Flexible Models of the Conditional Mean	421
11.4	Flexible Models of the Conditional Variance	425
11.5	Quantile Regression for Counts	432
11.6	Nonparametric Methods	435
11.7	Efficient Moment-Based Estimation	438
11.8	Analysis of Patent Counts	442
11.9	Derivations	446
11.10	Bibliographic Notes	447
12	Bayesian Methods for Counts	449
12.1	Introduction	449
12.2	Bayesian Approach	449
12.3	Poisson Regression	453
12.4	Markov Chain Monte Carlo Methods	454
12.5	Count Models	460
12.6	Roy Model for Counts	464
12.7	Bibliographic Notes	467
13	Measurement Errors	468
13.1	Introduction	468
13.2	Measurement Errors in Regressors	469
13.3	Measurement Errors in Exposure	479
13.4	Measurement Errors in Counts	485
13.5	Underreported Counts	488
13.6	Underreported and Overreported Counts	494
13.7	Simulation Example: Poisson with Mismeasured Regressor	496
13.8	Derivations	498
13.9	Bibliographic Notes	499
13.10	Exercises	499

Contents

	xi
A Notation and Acronyms	501
B Functions, Distributions, and Moments	505
B.1 Gamma Function	505
B.2 Some Distributions	506
B.3 Moments of Truncated Poisson	507
C Software	509
<i>References</i>	511
<i>Index</i>	543

PROOF

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 25%. 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 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. 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 Pariss of the 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

PROOF