Multilevel and Longitudinal Modeling Using Stata

Volume II: Categorical Responses, Counts, and Survival

Third Edition

SOPHIA RABE-HESKETH
University of California, Berkeley
Institute of Education, University of London

ANDERS SKRONDAL
Norwegian Institute of Public Health

«

A Stata Press Publication
StataCorp LP
College Station, Texas

Contents

List of Tables xvii

List of Figures xix

V Models for categorical responses 499

10 Dichotomous or binary responses 501

10.1 Introduction 501

10.2 Single-level logit and probit regression models for dichotomous responses 501

10.2.1 Generalized linear model formulation 502

10.2.2 Latent-response formulation 510

Logistic regression 512

Probit regression 512

10.3 Which treatment is best for toenail infection? 515

10.4 Longitudinal data structure 515

10.5 Proportions and fitted population-averaged or marginal probabilities 517

10.6 Random-intercept logistic regression 520

10.6.1 Model specification 520

Reduced-form specification 520

Two-stage formulation 522

10.7 Estimation of random-intercept logistic models 523

10.7.1 Using xtlogit 523

10.7.2 Using xtmelogit 527

10.7.3 Using gllamm 527

10.8 Subject-specific or conditional vs. population-averaged or marginal relationships 529

10.9 Measures of dependence and heterogeneity 532
10.9.1 Conditional or residual intraclass correlation of the latent responses 532
10.9.2 Median odds ratio 533
10.9.3 Measures of association for observed responses at median fixed part of the model 533
10.10 Inference for random-intercept logistic models 535
10.10.1 Tests and confidence intervals for odds ratios 535
10.10.2 Tests of variance components 536
10.11 Maximum likelihood estimation 537
10.11.1 ••• Adaptive quadrature 537
10.11.2 Some speed and accuracy considerations 540
Advice for speeding up estimation in gllamm 542
10.12 Assigning values to random effects 543
10.12.1 Maximum "likelihood" estimation 544
10.12.2 Empirical Bayes prediction 545
10.12.3 Empirical Bayes modal prediction 546
10.13 Different kinds of predicted probabilities 548
10.13.1 Predicted population-averaged or marginal probabilities . . 548
10.13.2 Predicted subject-specific probabilities 549
Predictions for hypothetical subjects: Conditional probabilities 549
Predictions for the subjects in the sample: Posterior mean probabilities 551
10.14 Other approaches to clustered dichotomous data 557
10.14.1 Conditional logistic regression 557
10.14.2 Generalized estimating equations (GEE) 559
10.15 Summary and further reading 562
10.16 Exercises 563
11 Ordinal responses 575
11.1 Introduction 575

11.2 Single-level cumulative models for ordinal responses 575
11.2.1 Generalized linear model formulation 575
11.2.2 Latent-response formulation 576
11.2.3 Proportional odds 580
11.2.4 ••• Identification 582
11.3 Are antipsychotic drugs effective for patients with schizophrenia? . 585
11.4 Longitudinal data structure and graphs 585
11.4.1 Longitudinal data structure 586
11.4.2 Plotting cumulative proportions 587
11.4.3 Plotting cumulative sample logits and transforming the time scale 588
11.5 A single-level proportional odds model 590
11.5.1 Model specification 590
11.5.2 Estimation using Stata 591
11.6 A random-intercept proportional odds model 594
11.6.1 Model specification 594
11.6.2 Estimation using Stata 594
11.6.3 Measures of dependence and heterogeneity 595
Residual intraclass correlation of latent responses 595
Median odds ratio 596
11.7 A random-coefficient proportional odds model 596
11.7.1 Model specification 596
11.7.2 Estimation using gllamm 596
11.8 Different kinds of predicted probabilities 599
11.8.1 Predicted population-averaged or marginal probabilities . . 599
11.8.2 Predicted subject-specific probabilities: Posterior mean . . 602
11.9 Do experts differ in their grading of student essays? 606
11.10 A random-intercept probit model with grader bias 606
11.10.1 Model specification 606
11.10.2 Estimation using gllamm 607

11.11 Including grader-specific measurement error variances 608
11.11.1 Model specification 608
11.11.2 Estimation using gllamm 609
11.12 ••• Including grader-specific thresholds 611
11.12.1 Model specification 611
11.12.2 Estimation using gllamm 611
11.13 ••• Other link functions 616
Cumulative complementary log-log model 616
Continuation-ratio logit model 616
Adjacent-category logit model 618
Baseline-category logit and stereotype models 618
11.14 Summary and further reading 619
11.15 Exercises 620
12 Nominal responses and discrete choice 629
12.1 Introduction 629
12.2 Single-level models for nominal responses 630
12.2.1 Multinomial logit models 630
12.2.2 Conditional logit models 638
Classical conditional logit models 639
Conditional logit models also including covariates that vary only over units 645
12.3 Independence from irrelevant alternatives 648
12.4 Utility-maximization formulation 649
12.5 Does marketing affect choice of yogurt? 651
12.6 Single-level conditional logit models 653
12.6.1 Conditional logit models with alternative-specific intercepts 654
12.7 Multilevel conditional logit models 659
12.7.1 Preference heterogeneity: Brand-specific random intercepts 659

12.7.2 Response heterogeneity: Marketing variables with random coefficients 663
12.7.3 *•* Preference and response heterogeneity 666
Estimation using gllamm 667
Estimation using mixlogit 669
12.8 Prediction of random effects and response probabilities 672
12.9 Summary and further reading 676
12.10 Exercises 677

VI Models for counts 685

13 Counts 687

13.1 Introduction 687
13.2 What are counts? 687
13.2.1 Counts versus proportions 687
13.2.2 Counts as aggregated event-history data 688
13.3 Single-level Poisson models for counts 689
13.4 Did the German health-care reform reduce the number of doctor visits? 691
13.5 Longitudinal data structure 691
13.6 Single-level Poisson regression 692
13.6.1 Model specification 692
13.6.2 Estimation using Stata 693
13.7 Random-intercept Poisson regression 696
13.7.1 Model specification 696
13.7.2 Measures of dependence and heterogeneity 697
13.7.3 Estimation using Stata 697
Using xtpoisson 697
Using xtmepoisson 699
Using gllamm 700
13.8 Random-coefficient Poisson regression 701
13.8.1 Model specification 701

13.8.2 Estimation using Stata 702
Using xtmepoisson 702
Using gllamm 704
13.8.3 Interpretation of estimates 705
13.9 Overdispersion in single-level models 706
13.9.1 Normally distributed random intercept 706
13.9.2 Negative binomial models 707
Mean dispersion or NB2 708
Constant dispersion or NB1 709
13.9.3 Quasilikelihood 709
13.10 Level-1 overdispersion in two-level models 711
13.11 Other approaches to two-level count data 713
13.11.1 Conditional Poisson regression 713
13.11.2 Conditional negative binomial regression 715
13.11.3 Generalized estimating equations 715
13.12 Marginal and conditional effects when responses are MAR 716
••• Simulation 717
13.13 Which Scottish counties have a high risk of lip cancer? 720
13.14 Standardized mortality ratios 721
13.15 Random-intercept Poisson regression 723
13.15.1 Model specification 723
13.15.2 Estimation using gllamm 724
13.15.3 Prediction of standardized mortality ratios 725
13.16 ••• Nonparametric maximum likelihood estimation 727
13.16.1 Specification 727
13.16.2 Estimation using gllamm 727
13.16.3 Prediction 732
13.17 Summary and further reading 732
13.18 Exercises 733

VII Models for survival or duration data 741

Introduction to models for survival or duration data (part VII) 743

14 Discrete-time survival 749

14.1 Introduction 749

14.2 Single-level models for discrete-time survival data 749

14.2.1 Discrete-time hazard and discrete-time survival 749

14.2.2 Data expansion for discrete-time survival analysis 752

14.2.3 Estimation via regression models for dichotomous responses 754

14.2.4 Including covariates 758

Time-constant covariates 758

Time-varying covariates 762

14.2.5 Multiple absorbing events and competing risks 767

14.2.6 Handling left-truncated data 772

14.3 How does birth history affect child mortality? 773

14.4 Data expansion 774

14.5 ••• Proportional hazards and interval-censoring 776

14.6 Complementary log-log models 777

14.7 A random-intercept complementary log-log model 781

14.7.1 Model specification 781

14.7.2 Estimation using Stata 782

14.8 ••• Population-averaged or marginal vs. subject-specific or conditional survival probabilities 784

14.9 Summary and further reading 788

14.10 Exercises 789

15 Continuous-time survival 797

15.1 Introduction 797

15.2 What makes marriages fail? 797

15.3 Hazards and survival 799

15.4 Proportional hazards models 805

15.4.1 Piecewise exponential model 807

15.4.2 Cox regression model 815
15.4.3 Poisson regression with smooth baseline hazard 819
15.5 Accelerated failure-time models 823
15.5.1 Log-normal model 824
15.6 Time-varying covariates 829
15.7 Does nitrate reduce the risk of angina pectoris? 832
15.8 Marginal modeling 835
15.8.1 Cox regression 835
15.8.2 Poisson regression with smooth baseline hazard 838
15.9 Multilevel proportional hazards models 841
15.9.1 Cox regression with gamma shared frailty 841
15.9.2 Poisson regression with normal random intercepts 845
15.9.3 Poisson regression with normal random intercept and random coefficient 847
15.10 Multilevel accelerated failure-time models 849
15.10.1 Log-normal model with gamma shared frailty 849
15.10.2 Log-normal model with log-normal shared frailty 850
15.11 A fixed-effects approach 851
15.11.1 Cox regression with subject-specific baseline hazards 851
15.12 Different approaches to recurrent-event data 853
15.12.1 Total time 854
15.12.2 Counting process 858
15.12.3 Gap time 859
15.13 Summary and further reading 861
15.14 Exercises 862

VIII Models with nested and crossed random effects 871

16 Models with nested and crossed random effects 873

16.1 Introduction 873
16.2 Did the Guatemalan immunization campaign work? 873
16.3 A three-level random-intercept logistic regression model 875

16.3.1 Model specification 876
16.3.2 Measures of dependence and heterogeneity 876
Types of residual intraclass correlations of the latent responses 876
Types of median odds ratios 877
16.3.3 Three-stage formulation 877
16.4 Estimation of three-level random-intercept logistic regression models 878
16.4.1 Using gllamm 878
16.4.2 Using xtmelogit 883
16.5 A three-level random-coefficient logistic regression model 886
16.6 Estimation of three-level random-coefficient logistic regression models 887
16.6.1 Using gllamm 887
16.6.2 Using xtmelogit 890
16.7 Prediction of random effects 892
16.7.1 Empirical Bayes prediction 892
16.7.2 Empirical Bayes modal prediction 893
16.8 Different kinds of predicted probabilities 894
16.8.1 Predicted population-averaged or marginal probabilities: New clusters 894
16.8.2 Predicted median or conditional probabilities 895
16.8.3 Predicted posterior mean probabilities: Existing clusters . 896
16.9 Do salamanders from different populations mate successfully? . . . 897
16.10 Crossed random-effects logistic regression 900
16.11 Summary and further reading 907
16.12 Exercises 908
A Syntax for gllamm, eq, and gllapred: The bare essentials 915
B Syntax for gllamm 921
C Syntax for gllapred 933
D Syntax for gllasim 937

References 941

Author index 955

Subject index 963