

Data Analysis Using Regression and Multilevel/Hierarchical Models

ANDREW GELMAN
Columbia University

JENNIFER HILL
Columbia University



CAMBRIDGE
UNIVERSITY PRESS

Contents

List of examples	page xvii
Preface	xix
1 Why?	1
1.1 What is multilevel regression modeling?	1
1.2 Some examples from our own research	3
1.3 Motivations for multilevel modeling	6
1.4 Distinctive features of this book	8
1.5 Computing	9
2 Concepts and methods from basic probability and statistics	13
2.1 Probability distributions	13
2.2 Statistical inference	16
2.3 Classical confidence intervals	18
2.4 Classical hypothesis testing	20
2.5 Problems with statistical significance	22
2.6 55,000 residents desperately need your help!	23
2.7 Bibliographic note	26
2.8 Exercises	26
Part 1A: Single-level regression	29
3 Linear regression: the basics	31
3.1 One predictor	31
3.2 Multiple predictors	32
3.3 Interactions	34
3.4 Statistical inference	37
3.5 Graphical displays of data and fitted model	42
3.6 Assumptions and diagnostics	45
3.7 Prediction and validation	47
3.8 Bibliographic note	49
3.9 Exercises	49
4 Linear regression: before and after fitting the model	53
4.1 Linear transformations	53
4.2 Centering and standardizing, especially for models with interactions	55
4.3 Correlation and “regression to the mean”	57
4.4 Logarithmic transformations	59
4.5 Other transformations	65
4.6 Building regression models for prediction	68
4.7 Fitting a series of regressions	73

4.8	Bibliographic note	74
4.9	Exercises	74
5	Logistic regression	79
5.1	Logistic regression with a single predictor	79
5.2	Interpreting the logistic regression coefficients	81
5.3	Latent-data formulation	85
5.4	Building a logistic regression model: wells in Bangladesh	86
5.5	Logistic regression with interactions	92
5.6	Evaluating, checking, and comparing fitted logistic regressions	97
5.7	Average predictive comparisons on the probability scale	101
5.8	Identifiability and separation	104
5.9	Bibliographic note	105
5.10	Exercises	105
6	Generalized linear models	109
6.1	Introduction	109
6.2	Poisson regression, exposure, and overdispersion	110
6.3	Logistic-binomial model	116
6.4	Probit regression: normally distributed latent data	118
6.5	Ordered and unordered categorical regression	119
6.6	Robust regression using the t model	124
6.7	Building more complex generalized linear models	125
6.8	Constructive choice models	127
6.9	Bibliographic note	131
6.10	Exercises	132
Part 1B: Working with regression inferences		135
7	Simulation of probability models and statistical inferences	137
7.1	Simulation of probability models	137
7.2	Summarizing linear regressions using simulation: an informal Bayesian approach	140
7.3	Simulation for nonlinear predictions: congressional elections	144
7.4	Predictive simulation for generalized linear models	148
7.5	Bibliographic note	151
7.6	Exercises	152
8	Simulation for checking statistical procedures and model fits	155
8.1	Fake-data simulation	155
8.2	Example: using fake-data simulation to understand residual plots	157
8.3	Simulating from the fitted model and comparing to actual data	158
8.4	Using predictive simulation to check the fit of a time-series model	163
8.5	Bibliographic note	165
8.6	Exercises	165
9	Causal inference using regression on the treatment variable	167
9.1	Causal inference and predictive comparisons	167
9.2	The fundamental problem of causal inference	170
9.3	Randomized experiments	172
9.4	Treatment interactions and poststratification	178

9.5	Observational studies	181
9.6	Understanding causal inference in observational studies	186
9.7	Do not control for post-treatment variables	188
9.8	Intermediate outcomes and causal paths	190
9.9	Bibliographic note	194
9.10	Exercises	194
10	Causal inference using more advanced models	199
10.1	Imbalance and lack of complete overlap	199
10.2	Subclassification: effects and estimates for different subpopulations	204
10.3	Matching: subsetting the data to get overlapping and balanced treatment and control groups	206
10.4	Lack of overlap when the assignment mechanism is known: regression discontinuity	212
10.5	Estimating causal effects indirectly using instrumental variables	215
10.6	Instrumental variables in a regression framework	220
10.7	Identification strategies that make use of variation within or between groups	226
10.8	Bibliographic note	229
10.9	Exercises	231
Part 2A:	Multilevel regression	235
11	Multilevel structures	237
11.1	Varying-intercept and varying-slope models	237
11.2	Clustered data: child support enforcement in cities	237
11.3	Repeated measurements, time-series cross sections, and other non-nested structures	241
11.4	Indicator variables and fixed or random effects	244
11.5	Costs and benefits of multilevel modeling	246
11.6	Bibliographic note	247
11.7	Exercises	248
12	Multilevel linear models: the basics	251
12.1	Notation	251
12.2	Partial pooling with no predictors	252
12.3	Partial pooling with predictors	254
12.4	Quickly fitting multilevel models in R	259
12.5	Five ways to write the same model	262
12.6	Group-level predictors	265
12.7	Model building and statistical significance	270
12.8	Predictions for new observations and new groups	272
12.9	How many groups and how many observations per group are needed to fit a multilevel model?	275
12.10	Bibliographic note	276
12.11	Exercises	277
13	Multilevel linear models: varying slopes, non-nested models, and other complexities	279
13.1	Varying intercepts and slopes	279
13.2	Varying slopes without varying intercepts	283

13.3	Modeling multiple varying coefficients using the scaled inverse-Wishart distribution	284
13.4	Understanding correlations between group-level intercepts and slopes	287
13.5	Non-nested models	289
13.6	Selecting, transforming, and combining regression inputs	293
13.7	More complex multilevel models	297
13.8	Bibliographic note	297
13.9	Exercises	298
14	Multilevel logistic regression	301
14.1	State-level opinions from national polls	301
14.2	Red states and blue states: what's the matter with Connecticut?	310
14.3	Item-response and ideal-point models	314
14.4	Non-nested overdispersed model for death sentence reversals	320
14.5	Bibliographic note	321
14.6	Exercises	322
15	Multilevel generalized linear models	325
15.1	Overdispersed Poisson regression: police stops and ethnicity	325
15.2	Ordered categorical regression: storable votes	331
15.3	Non-nested negative-binomial model of structure in social networks	332
15.4	Bibliographic note	342
15.5	Exercises	342
Part 2B:	Fitting multilevel models	343
16	Multilevel modeling in Bugs and R: the basics	345
16.1	Why you should learn Bugs	345
16.2	Bayesian inference and prior distributions	345
16.3	Fitting and understanding a varying-intercept multilevel model using R and Bugs	348
16.4	Step by step through a Bugs model, as called from R	353
16.5	Adding individual- and group-level predictors	359
16.6	Predictions for new observations and new groups	361
16.7	Fake-data simulation	363
16.8	The principles of modeling in Bugs	366
16.9	Practical issues of implementation	369
16.10	Open-ended modeling in Bugs	370
16.11	Bibliographic note	373
16.12	Exercises	373
17	Fitting multilevel linear and generalized linear models in Bugs and R	375
17.1	Varying-intercept, varying-slope models	375
17.2	Varying intercepts and slopes with group-level predictors	379
17.3	Non-nested models	380
17.4	Multilevel logistic regression	381
17.5	Multilevel Poisson regression	382
17.6	Multilevel ordered categorical regression	383
17.7	Latent-data parameterizations of generalized linear models	384

CONTENTS	xiii
17.8 Bibliographic note	385
17.9 Exercises	385
18 Likelihood and Bayesian inference and computation	387
18.1 Least squares and maximum likelihood estimation	387
18.2 Uncertainty estimates using the likelihood surface	390
18.3 Bayesian inference for classical and multilevel regression	392
18.4 Gibbs sampler for multilevel linear models	397
18.5 Likelihood inference, Bayesian inference, and the Gibbs sampler: the case of censored data	402
18.6 Metropolis algorithm for more general Bayesian computation	408
18.7 Specifying a log posterior density, Gibbs sampler, and Metropolis algorithm in R	409
18.8 Bibliographic note	413
18.9 Exercises	413
19 Debugging and speeding convergence	415
19.1 Debugging and confidence building	415
19.2 General methods for reducing computational requirements	418
19.3 Simple linear transformations	419
19.4 Redundant parameters and intentionally nonidentifiable models	419
19.5 Parameter expansion: multiplicative redundant parameters	424
19.6 Using redundant parameters to create an informative prior distribution for multilevel variance parameters	427
19.7 Bibliographic note	434
19.8 Exercises	434
Part 3: From data collection to model understanding to model checking	435
20 Sample size and power calculations	437
20.1 Choices in the design of data collection	437
20.2 Classical power calculations: general principles, as illustrated by estimates of proportions	439
20.3 Classical power calculations for continuous outcomes	443
20.4 Multilevel power calculation for cluster sampling	447
20.5 Multilevel power calculation using fake-data simulation	449
20.6 Bibliographic note	454
20.7 Exercises	454
21 Understanding and summarizing the fitted models	457
21.1 Uncertainty and variability	457
21.2 Superpopulation and finite-population variances	459
21.3 Contrasts and comparisons of multilevel coefficients	462
21.4 Average predictive comparisons	466
21.5 R^2 and explained variance	473
21.6 Summarizing the amount of partial pooling	477
21.7 Adding a predictor can <i>increase</i> the residual variance!	480
21.8 Multiple comparisons and statistical significance	481
21.9 Bibliographic note	484
21.10 Exercises	485

22 Analysis of variance	487
22.1 Classical analysis of variance	487
22.2 ANOVA and multilevel linear and generalized linear models	490
22.3 Summarizing multilevel models using ANOVA	492
22.4 Doing ANOVA using multilevel models	494
22.5 Adding predictors: analysis of covariance and contrast analysis	496
22.6 Modeling the variance parameters: a split-plot latin square	498
22.7 Bibliographic note	501
22.8 Exercises	501
23 Causal inference using multilevel models	503
23.1 Multilevel aspects of data collection	503
23.2 Estimating treatment effects in a multilevel observational study	506
23.3 Treatments applied at different levels	507
23.4 Instrumental variables and multilevel modeling	509
23.5 Bibliographic note	512
23.6 Exercises	512
24 Model checking and comparison	513
24.1 Principles of predictive checking	513
24.2 Example: a behavioral learning experiment	515
24.3 Model comparison and deviance	524
24.4 Bibliographic note	526
24.5 Exercises	527
25 Missing-data imputation	529
25.1 Missing-data mechanisms	530
25.2 Missing-data methods that discard data	531
25.3 Simple missing-data approaches that retain all the data	532
25.4 Random imputation of a single variable	533
25.5 Imputation of several missing variables	539
25.6 Model-based imputation	540
25.7 Combining inferences from multiple imputations	542
25.8 Bibliographic note	542
25.9 Exercises	543
Appendixes	545
A Six quick tips to improve your regression modeling	547
A.1 Fit many models	547
A.2 Do a little work to make your computations faster and more reliable	547
A.3 Graphing the relevant and not the irrelevant	548
A.4 Transformations	548
A.5 Consider all coefficients as potentially varying	549
A.6 Estimate causal inferences in a targeted way, not as a byproduct of a large regression	549
B Statistical graphics for research and presentation	551
B.1 Reformulating a graph by focusing on comparisons	552
B.2 Scatterplots	553
B.3 Miscellaneous tips	559

B.4	Bibliographic note	562
B.5	Exercises	563
C	Software	565
C.1	Getting started with R, Bugs, and a text editor	565
C.2	Fitting classical and multilevel regressions in R	565
C.3	Fitting models in Bugs and R	567
C.4	Fitting multilevel models using R, Stata, SAS, and other software	568
C.5	Bibliographic note	573
	References	575
	Author index	601
	Subject index	607