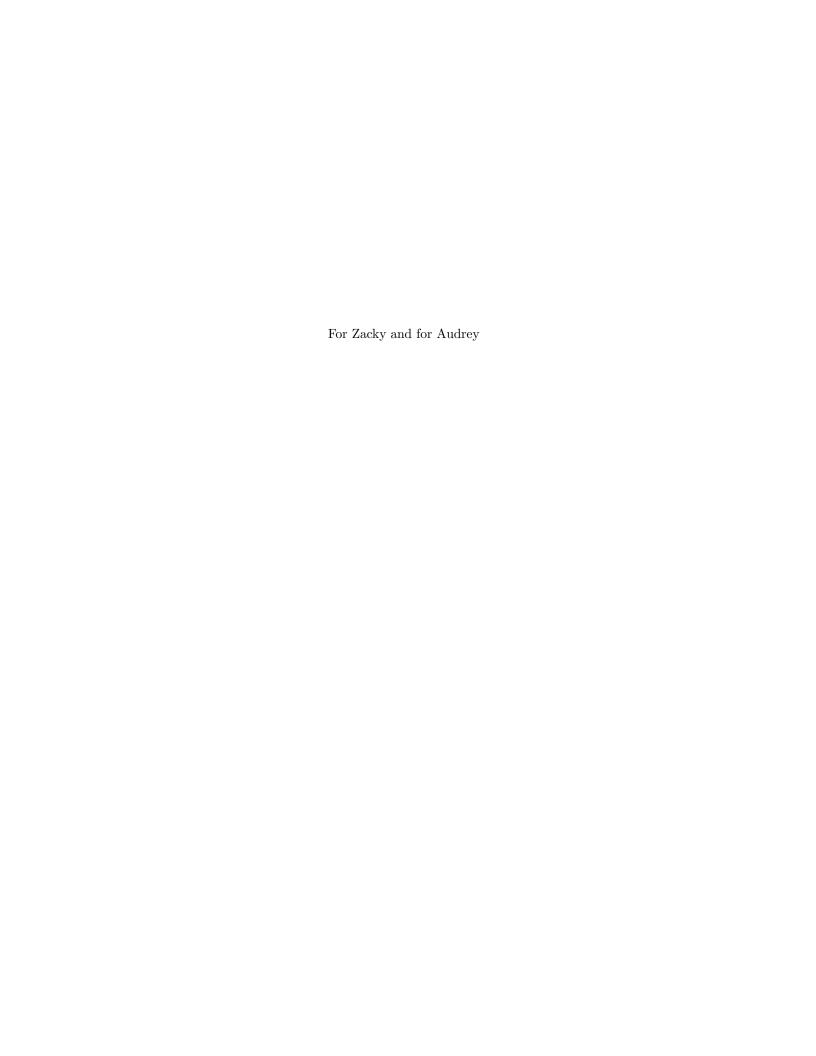
Data Analysis Using Regression and Multilevel/Hierarchical Models (Final version: 5 July 2006) Please do not reproduce in any form without permission

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University, New York

Jennifer Hill School of International and Public Affairs Columbia University, New York

©2002, 2003, 2004, 2005, 2006 by Andrew Gelman and Jennifer Hill To be published in October, 2006 by Cambridge University Press



Contents

Li	List of examples		xvii
P	refac	e	xix
1 Why?		y?	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	Cor	cepts 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
Pa	art 1.	A: Single-level regression	29
3	Line	ear 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	Line	ear 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
	47		73

X	CONTENTS

	4.8	Bibliographic note	74	
	4.9	Exercises	74	
5	Logi	stic regression	7 9	
	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	
c	Com	eralized linear models	100	
6			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	Multinomial 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 Exercises	131	
	0.10	Exercises	132	
Pa	rt 1E	3: Working with regression inferences	135	
7	Sim	lation 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	Sim	ulation 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	
•	C		105	
9		sal 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	

CC	NTE	NTS	xi
	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
		Exercises	194
10	Cau	sal inference using more advanced models	199
		Imbalance and lack of complete overlap	199
		Subclassification: effects and estimates for different subpopulations	204
		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:	200
	10.4	regression discontinuity	212
	10.5	Estimating causal effects indirectly using instrumental variables	215
		Instrumental variables in a regression framework	$\frac{210}{220}$
			220
	10.7	Identification strategies that make use of variation within or between	226
	10.0	groups Diblic markis note	229
		Bibliographic note Exercises	
	10.9	Exercises	231
Pa	rt 2/	A: Multilevel regression	235
11	Mul	tilevel 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	Mul	tilevel linear models: the basics	251
		Notation	251
		Partial pooling with no predictors	252
		Partial pooling with predictors	254
		Quickly fitting multilevel models in R	259
		Five ways to write the same model	262
		Group-level predictors	265
		Model building and statistical significance	270
		Predictions for new observations and new groups	272
		How many groups and how many observations per group are	
		needed to fit a multilevel model?	275
	12.10	Bibliographic note	276
		Exercises	277
13	Mul	tilevel linear models: varying slopes, non-nested models, and	
		r complexities	279
		Varying intercepts and slopes	279
		Varying slopes without varying intercepts	283
		· · · · · · · · · · · · · · · · · · ·	

xii CONTENTS

	13.3	Modeling multiple varying coefficients using the scaled inverse- Wishart distribution	284
	13 4	Understanding correlations between group-level intercepts and	204
	10.1	slopes	287
	13.5	Non-nested models	289
		Selecting, transforming, and combining regression inputs	293
		More complex multilevel models	297
		Bibliographic note	297
		Exercises	298
14	Mul	tilevel 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
		Bibliographic note	321
	14.6	Exercises	322
15	Mul	tilevel generalized linear models	325
	15.1	Overdispersed Poisson regression: police stops and ethnicity	325
		Ordered categorical regression: storable votes	331
		Non-nested negative-binomial model of structure in social networks	332
	15.4	Bibliographic note	342
	15.5	Exercises	342
Pa	rt 2E	3: Fitting multilevel models	343
16	Mul	tilevel modeling in Bugs and R: the basics	345
		Why you should learn Bugs	345
		Bayesian inference and prior distributions	345
		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	Fitti	ng multilevel linear and generalized linear models in Bugs	
	and	R	375
		Varying-intercept, varying-slope models	375
		Varying intercepts and slopes with group-level predictors	379
		Non-nested models	380
		Multilevel logistic regression	381
		Multilevel Poisson regression	382
		Multilevel ordered categorical regression	383
	1 7 7	Latent-data parameterizations of generalized linear models	384

CC	CONTENTS xiii			
	17.8	Bibliographic note	385	
		Exercises	385	
	11.0	Zhorobou -	900	
18	Like	lihood and Bayesian inference and computation	387	
		Least squares and maximum likelihood estimation	387	
		Uncertainty estimates using the likelihood surface	390	
		Bayesian inference for classical and multilevel regression	392	
		Gibbs sampler for multilevel linear models	397	
		Likelihood inference, Bayesian inference, and the Gibbs sampler:	001	
	10.0	the case of censored data	402	
	18.6	Metropolis algorithm for more general Bayesian computation	408	
		Specifying a log posterior density, Gibbs sampler, and Metropolis	-00	
		algorithm in R	409	
	18.8	Bibliographic note	413	
		Exercises	413	
19	Deb	ugging and speeding convergence	415	
		Debugging and confidence building	415	
		General methods for reducing computational requirements	418	
		Simple linear transformations	419	
		Redundant parameters and intentionally nonidentifiable models	419	
		Parameter expansion: multiplicative redundant parameters	424	
		Using redundant parameters to create an informative prior		
		distribution for multilevel variance parameters	427	
	19.7	Bibliographic note	434	
	19.8	Exercises	434	
Pa		From data collection to model understanding to model		
	chec	king	435	
	~			
20		ple size and power calculations	437	
		Choices in the design of data collection	437	
	20.2	Classical power calculations: general principles, as illustrated by	400	
	20.0	estimates of proportions	439	
		Classical power calculations for continuous outcomes	443	
		Multilevel power calculation for cluster sampling	447	
		Multilevel power calculation using fake-data simulation	449	
		Bibliographic note	454	
	20.7	Exercises	454	
91	Und	erstanding and summarizing the fitted models	457	
41		Uncertainty and variability	457	
		Superpopulation and finite-population variances	459	
		Contrasts and comparisons of multilevel coefficients	462	
		Average predictive comparisons	466	
		Average predictive comparisons R^2 and explained variance	473	
		Summarizing the amount of partial pooling	$473 \\ 477$	
		Adding a predictor can <i>increase</i> the residual variance!	480	
		Multiple comparisons and statistical significance	481	
		Bibliographic note	484	
		Exercises	485	

xiv	CONTENTS
xiv	CONTE

22	Ana	lysis of variance	487
		Classical analysis of variance	487
		ANOVA and multilevel linear and generalized linear models	490
		Summarizing multilevel models using ANOVA	492
		Doing ANOVA using multilevel models	494
		Adding predictors: analysis of covariance and contrast analysis	496
		Modeling the variance parameters: a split-plot latin square	498
		Bibliographic note	501
		Exercises	501
23	Cau	sal inference using multilevel models	503
		Multilevel aspects of data collection	503
		Estimating treatment effects in a multilevel observational study	506
		Treatments applied at different levels	507
		Instrumental variables and multilevel modeling	509
		Bibliographic note	512
		Exercises	512
24	Mod	lel checking and comparison	513
		Principles of predictive checking	513
		Example: a behavioral learning experiment	515
		Model comparison and deviance	524
		Bibliographic note	526
		Exercises	527
25	Miss	sing-data imputation	529
_0		Missing-data mechanisms	530
		Missing-data methods that discard data	531
		Simple missing-data approaches that retain all the data	532
		Random imputation of a single variable	533
		Imputation of several missing variables	539
		Model-based imputation	540
		Combining inferences from multiple imputations	542
		Bibliographic note	542
		Exercises	543
ΑĮ	peno	lixes	545
\mathbf{A}	Six	quick tips to improve your regression modeling	547
		Fit many models	547
	A.2	Do a little work to make your computations faster and more reliable	547
		Graphing the relevant and not the irrelevant	548
	A.4		548
	A.5	Consider all coefficients as potentially varying	549
		Estimate causal inferences in a targeted way, not as a byproduct	
	-	of a large regression	549
В	Stat	istical graphics for research and presentation	551
	B.1	Reformulating a graph by focusing on comparisons	552
	B.2	Scatterplots	553
		Miscellaneous tips	559

CC	CONTENTS		xv	
	B.4	Bibliographic note	562	
	B.5	Exercises	563	
\mathbf{C}	Soft	ware	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	
Re	eferen	aces	575	
Αι	Author index 60			
Su	Subject index 60			