### Index

| Academic Technology Services (ATS) web site, xv–xvi |
| Acceptable discrimination, 177. See also Receiver Operating Characteristic (ROC) curve |
| Adaptive quadrature, 351 |
| estimation of, 322–323, 326–327 |
| Adaptive rejection sampling, 413 |
| Additive difference, 79 |
| Additive interaction, 448 |
| estimating and testing, 451–456 |
| Additive scale, multiplicative scale vs., 448–451 |
| Adjacent-category logistic model, 290–291, 294–296 |
| Adjusted logistic regression coefficient, interpretation of, 443–444 |
| Adolescent Placement Study (APS) data set, 26–27 |
| multinomial assessment of fit and interpretation, 284–289 |
| multinomial modeling, 279–283 |
| ordinal logistic modeling, diagnostics, proportional odds assumption, 305–310 |
| Aftercare placement study, 272–278 |
| Aggregated data sets, 165 |
| Akaike Information Criterion (AIC), 120–121, 134, 136–339 |
| Algorithm performance checks, 414 |
| Alpha level, choosing, 126–127 |
| Alternate coding, 55. See also Coding |
| Alternative link functions, roles of, 436 |
| Alternative ordinal score, 304–305 |
| Analysis of covariance, 65 |
| Analysis of covariance model, 228 |
| Analysis of variance table, 11 |
| Approximate methods, 387 |
Bayesian inference Using Gibbs Sampling (BUGS) software package, 409, 413. See also OpenBUGS statistical package
Bayesian logistic regression models, 409, 410–411
Bayesian methods for logistic regression, 408–433
Bayesian perspective, on multiple imputation, 397
Bayesian residuals, 430–433
Bayes’ theorem, 229–230, 245, 411
Best Linear Unbiased Predictions (BLUPs), 330
Best model, choosing a, 89
Best subsets linear regression, 136
Best subsets logistic regression, 133–139
advantage of, 139
applying (weighted least squares) best subsets linear regression software, 134, 139
Best subsets selection, 94
using Score test method, 137
Between-chain variability ($B$), in MCMC simulations, 418
Between-cluster correlation, 316
Bias. See also Best Linear Unbiased Predictions (BLUPs); Estimated Best Linear Unbiased Predictions (EBLUPs); Median unbiased estimator (MUE)
in discriminant function estimators, 45–46
in maximum likelihood, 387
in maximum likelihood estimators, 391
Binary data, correlated, 314–315
Binary models, fitting separate, 282–283
unconstrained continuation-ratio model, 295–296
when proportional odds assumption is not satisfied, 309
Binary outcome models, 273
cluster-specific, 315
Binary outcome variable, 1, 229, 270, 278, 283
Binary regression models, link functions for, 434–441
linear link, 451–453
Binary variable coding, for the conditional likelihood function construction, 271
Binomial errors, 186
Biological interaction, 448
Bootstrapping methods, 82, 380, 456
Boxplots, of standardized residuals, 370–371
Brant’s Wald test, 302, 306
Breslow–Day test, 85
Brooks–Gelman–Rubin (BGR) statistic, 417–418
Burn in iterations, 414–416, 419
Burn Injury Study, 27
BURN1000 data, 27
BURN_EVAL_1 data set, 205–207
BURN_EVAL_2 data, 207–211
classification table for, 173
diagnostic statistics and, 201–202
discrimination for the model fit to, 181
evaluating model fit to, 161–162
for fitting link functions, 437–441
fitting multivariable model to, 116
main effects model for, 124
1–3 matched data set from, 260–267
plots related to, 220–222
using purposeful selection in, 115–124
results from MFP cycle fits applied to, 143–144
Calibration, of models, 186
Case-control data, analysis of stratified, 232
likelihood function, 229
pairs, uninformative, 246
studies, 229–233
Case-wise diagnostic measures, computing, 253
diagnostic statistics, 308
diagnostic tools, 354
Categorical independent variables, included or excluded from models, 41
Categorical variables, examining scale of continuous covariate, 95–96
Caterpillar plot, 331–332
Cell coding, reference, 55, 57–59. See also Coding
Cell counts, 145–146
Chi-square ($\chi^2$) distribution, 13–14, 158
Chi-square goodness of fit tests, 232
Chi-square random variable, 14
Chi-square ($\chi^2$) statistic, Pearson, 135–136, 155–157, 163
Chi-square ($\chi^2$) tests, 157
likelihood ratio, 90
Pearson, 90, 157
Classification, 18. See also Cross-classification sensitivity of, 171–172
Classification tables, 169–173
for GLOW Study, 171–175
sensitivity/specificity for, 175
Closed test method, 343
Closed test procedure, 98, 140
Clumping, in MCMC simulations, 415–416
Cluster effects, 330
Cluster estimates, 330–331
Cluster influence, measures of, 362–363
Cluster-level covariates, 313, 317, 329, 354
Cluster-level leverage, 362
Cluster-level variables, 330
Clusters, 241
  influential, 371–374
  outlying, 372–374
Cluster-specific binary outcome model, 315
  coefficients, 335–336
  covariates, 313
  estimates, 327
  fitted values, 365–366
Cluster-specific models, 315–317, 320–321, 326–333, 374
  alternative estimation methods for, 333–334
  assessment of fit of, 354–365
  with correlated data, 344–350
  fitting, 351
  logistic normal, 315
  population average model vs., 334–337
  random effects in, 367
Cluster-specific odds ratio, estimate of, 328
Cluster variance estimate, 329
Coded design variables, 58
methods for, 55
Coding. See also Cell coding
  alternate, 55
  binary variable, 271
  deviation from means, 55–56, 59–62
  dichotomous, 279–280
  effect of, 54–56
  of outcome variables, 293
Coefficient of discrimination, 185
Coefficients. See also Intracluster correlation coefficient (ICC)
  adjusted, 210–211
  changes in the estimated values of, 191
  changing signs of, 301
  cluster-specific, 335–336
  estimated interaction, 454–455
  estimated slope, 349
  estimates of model, 212
  intercept, 208
  interpretation of adjusted logistic regression, 443–444
  interpreting for correlated-data analysis, 323–337
  logistic regression, 403–404
  population average, 336
  regression, 241
  significance of, 237
  significance of estimated, 272–278
  univariable, 71
  univariable estimated, 72
  vector of, 244–245
  vector of estimated, 237
  Coefficient significance
    likelihood ratio test for, 276
    testing for, 10–15
  Cohort studies, 227–229
  Collaborative Longitudinal Evaluation of Ethnicity and Refractive Error (CLEERE) Study, 31
  Collinearities, among independent variables, 149
  Common odds ratio assumption, 85
  Comparative residuals, 368
  Comparative standard errors, 368–369
  Complementary log–log model, 435–436, 438
  Complete separation problem, 147–149
  Complex sample survey data, fitting logistic regression models to, 233–242
  Complex survey data, 236–237
  Concordance correlation, 359, 367
  Concordant pairs, 246
  Conditional distributions
    in Bayesian logistic regression models, 411
    in MCMC simulations, 413
    of outcome variables, 7
  Conditional exact maximum likelihood estimate (CMLE), 390–391, 394
  Conditional likelihood, 244–245, 388
  full, 247
  Conditional likelihood analysis, 244–245
  Conditional likelihood function, 247
  Conditional log-likelihood, maximizing, 247
  Conditional maximum likelihood estimates, 247.
    See also Conditional exact maximum likelihood estimate (CMLE)
  Conditional maximum likelihood estimators, 22
  Conditional maximum likelihood point estimates, exact, 390–391
  Conditional mean, 2–7
  dichotomous outcome variable and, 5–6
  estimates of, 5–6
  Conditional model, 315
  Conditional probability, 260, 270–271
  estimated, 249
  Confidence bands, 75, 82, 220–222
  Confidence interval (CI) endpoints, 63–64
  calculating, 56
  Confidence interval estimates, 215, 267, 440
Confidence interval (CI) estimation, 15–20, 53–54
for the multivariable model, 42–45
Confidence interval (CI) estimators, 16
likelihood-based, 19
logit, 43–45
for the multivariable model, 42–45
profile likelihood, 43, 54
Wald-based, 380
Confidence intervals (CIs), 80, 209, 212–213, 258, 269, 392
endpoints of, 16–17, 276
for intraclass correlation, 345
log-likelihood function–based, 18
for odds ratios, 62, 274
Confidence limits, 75–76
for odds ratios, 59
Confounders, 64, 131, 237. See also Controlling for confounding
Confounding, 90, 377–379
controlled, 447
controlling for, 379
residual, 384
uncontrolled, 447
Confounding variables, 456
Conservative effective sample size, in MCMC simulations, 418
Constant covariates, 261
Constant odds ratios, 82
Constant term, as estimator, 16
Constrained baseline logistic model, 291, 294–295
Constrained multinomial logistic regression model, 310
Constrained ordinal models, 302
Contingency table(s), 90
approach, xiii
frequency of zero in, 145–146
Contingency table analysis methods of, 161–162
stratified, 50
Continuation-ratio logistic model, 290–291, 295–297
Continuous covariates, 69–71, 78, 139–140, 253, 278, 281, 324
checking the scale of, 338, 342, 347
dichotomizing, 112
Continuous covariate scale, methods to examine, 94–107
Continuous independent variables, 62–64
Continuous outcome model, 301
Continuous outcomes, regression based on, 298
Continuous response variable, 297–298
Continuous variables, 106–107
univariable analysis of, 91
Contribution to the likelihood function, 8
Controlled confounding, 447
Controlling for confounding, 64, 67
Convergence, of MCMC chains, 417–418. See also Chain convergence
Convergence issues, 351–352
Cook’s Distance diagnostic, 191–192, 197, 371–372. See also Delta-beta-hat percent asymptotic distribution of, 193
plotting of, 196, 255
Coronary heart disease (CHD)
frequency by AGE group, 6
table, 3–5
Correlated binary data
analysis of, 314–315
logistic regression models for, 375
modeling, 374
Correlated categorical response data, xiv
Correlated data, 313–315
cluster-specific model with, 344–350
logistic regression modeling with, 337–353
population average model with, 339–344
Correlated-data analysis
choosing model for, 338–339
goals of, 313–314
interpreting coefficients for, 323–337
logistic regression models for, 313–375
Correlated-data logistic regression models, estimation methods for, 318–323
Correlated-data modeling software, 314–315
Correlated data models, Hosmer–Lemeshow test and, 354
Correlation(s)
concordance, 359, 367
ignoring, 325
intracluster, 327, 335–336, 351, 354
within- and between-cluster, 316
Correlation estimates, 357
Correlation structures
autoregressive, 340
checking, 358
choosing/selecting, 318–320, 339, 344, 359
exchangeable, 325
Covariance(s)
analysis of, 65
estimators of, 37–38
significance of, 286–287
within-cluster, 319, 320
Covariance matrix estimator, 46
Covariance matrix/matrices, 319
estimated, 17, 275
measures for comparing estimated, 359
Covariance parameter, 350
Covariate(s), 1
    adjusted probability, 82
categorizing, 103
checking the scale of continuous, 342, 374
cluster-level, 313, 317, 329, 354
cluster-specific, 313
collapsing categories for, 341
constant, 261
continuous, 69, 139–140, 253, 278, 281, 324, 338
dichotomous, 198–199, 218, 261, 341
effects, estimating and interpreting, 374
estimated odds ratio for, 258–259
events per, 402, 407–408
identifying dependencies among, 149–150
interactions among, 262
necessity of, 76–77
overlapping distributions of, 147–148
parameterization of, 96
partitioning into g regions, 163
probability distribution of, 230
purposeful selection of, 70, 89–124
in regression sampling, 227–228
scale, methods to examine continuous, 94–107
selecting/checking scale using multivariable fractional polynomials, 139–144
subject-specific, 317
time-invariant, 313
time-varying, 313
Covariate selection
alternative methods of, 124–144
methods, purposeful, 344
stepwise, 125–133
Coverage, of an interval estimator, 18
Credible interval
Bayesian, 421
equal-tailed, 421
Cross-classification, 273, 277, 293, 389, 392
of DEATH by FLAME, 83
by vital status, 83
Crude odds ratio, 82, 86
\( \hat{C} \) statistic. See Hosmer–Lemeshow goodness of fit statistic (\( \hat{C} \))
Cubic spline covariates, restricted, 101
Cubic splines, restricted, 105–106
Cubic splines model, restricted, 118–119
Cubic spline variables, 101
Cumulative distributions, 6
Cumulative distribution function, 298
Cumulative sums of residuals, tests based on, 164
Cutpoints
defining, 170
optimal, 174–176
Data
correlated, 313–315
ignorable, 396
missing, 314, 395–401
unavailable, 235–236
Data analysis, choosing model for correlated, 338–339
Data collection, retroactive, 201
Data sets
aggregated, 165
developmental, 168
imputed, 398
modeling of, 10
used in examples and exercises, 22–32
validation, 168
Data vectors, 245
DBETAC\(_i\) statistic, 364
dCLS\(_i\) statistic, 362–363
Decile of risk goodness of fit test, extension of, 283–284
Decile of risk group strategy, 160–162
disadvantage of, 162–163
likelihood ratio test using, 163
Decile of risk statistic, 440
Decile of risk test, 168, 205, 239–240
Decile of risk type tests, grouped, 167
Decile size, imbalance in, 161
Degree of freedom statistic, 166
Degrees of freedom, 41, 139
for assessing model performance, 154
inferences and, 397–398
for multinomial goodness of fit test, 304–305
between variables, 125
Delete/refit procedure, 285
Deletion, of variables, 127
Delta-beta-hat percent, 67, 349, 445
Dependent variables, values of, 135
Design-based methods, 240–242
Design matrix, 187. See also X matrix
Design variable method, 109, 110
Design variable(s), 35, 94–96, 398–399
coding of, 56, 58–62
collections of, 36
for multiple logistic regression model, 35–36
for polychotomous independent variables, 57–58
quartile, 103–104, 110, 112–113, 117
quartile-based, 121
Design variables method, 338, 347
Developmental data set, 168
Deviance (D), 12–13, 155–157
with and without independent variable, 13
Deviance Information Criteria (DIC), in
Bayesian analysis, 428–429
Deviance residual, 156
Deviation from means coding, 55–56, 59–60, 62
estimated coefficients obtained using, 61
Diagnostic(s)
evaluation, 240
influence, 197, 250
interpreting the value of, 192
logistic regression, 186–292
regression, 186
residuals, 368
standard errors, 368–369
stratum specific totals of, 250
Burn Injury Study data and, 201
calculating, 285–286
case-wise, 308
data and values of, 264–265
estimating matched set effect on, 255
for multinomial logistic regression model, 283–289
statistical package calculation of, 188
subject-specific, 359–360
Diagonal matrix, 319
Dichotomous coding, 279–280
Dichotomous–continuous covariate model, 70–71
Dichotomous covariates, 198–199, 261, 341
estimating odds ratios for, 218
Dichotomous independent variables, 21–22, 50–56
Dichotomous outcome variable, 1, 5–8
regression analysis with, 7–8
Dichotomous variable(s), 69–71, 170
odds ratio and, 56
Difference data approach, to 1–1 matched
design, 250–251
Diffuse prior distribution, 423
Direct effect, 443
Discriminant function estimate, univariable
linear, 91
Discriminant function estimators
bias in, 45–46
maximum likelihood estimators vs., 21–22
in the multivariable case, 45–46
Discriminant function models, normal theory,
231
Discrimination. See also Coefficient of
discrimination
levels of, 176, 178–181
visual methods for assessing, 178
Distribution functions
cumulative, 298
for use in dichotomous outcome variable
analysis, 6–7
Distributions
in Bayesian logistic regression models, 410–411
of maximum likelihood estimators, 18
mixture, 345
Distribution theory, 192
relevant, 157
D matrix, 234
Due regression sum-of-squares (SSR), 11–12.
See also Sum-of-squares (S)
Dummy variables, 36. See also Design variables
Effect
direct, 443
estimates of, 440–441
of independent variables, 444
indirect, 443, 445
total, 443
Effective number of parameters, in Bayesian
analysis, 428
Effective sample size, in MCMC simulations, 418
Effect modification, 64, 455–456, 448, 450
statistically testing for, 451
Effect modifier, 68
Empirical residuals, 320
Endpoints
of confidence intervals, 16, 56, 63–64, 276
exponentiating, 54
of likelihood intervals, 19
of Wald-based confidence intervals, 19, 42–43
Equality, test for, 296
Equal-tailed credible interval, 421
Error (e), 7
hypothesis testing, 167
Errors, binomial, 186
Estimated Best Linear Unbiased Predictions
(EBLUPs), 330
adjusted coefficients vs., 210–211
changes in the values of, 191
drawing inferences from, 49
obtained using deviation from means coding, 61
significance of, 272–278
vector of, 237
Estimated conditional probability, 249
Estimated covariance matrix/matrices, 17, 275
measures for comparing, 359
Estimated expected risk frequencies, 160–161
Estimated expected value, 85
Estimated frequency, 85
Estimated interaction coefficient, 454–455
Estimated logistic probability, 17, 44
Estimated logistic regression coefficients, 86
exponentiation of, 278
Estimated logit
estimating the variance of, 43–44
95 percent confidence interval for, 45
Estimated odds ratio(s), 56–57, 86, 214, 216–219, 267, 286, 325, 383–384
for covariates, 258–259
tabulation of, 84
Estimated odds ratio interpretation, 327–328
for a continuous variable, 64
Estimated population average odds ratio, 326
of death, 80–81, 116, 162
distributions of, 176–181
histograms of, 174–176
importance of, 77
lack of fit diagnostic vs., 263–264
leverage values vs., 262–263
Estimated propensity score, 379–380, 382
Estimated slope coefficients, 275, 349
Estimated standard error(s), 17, 59, 62, 149, 231–232, 274, 278, 325, 327
of pooled log-odds ratio estimator, 380
Estimated stratum-specific probability, 262
Estimates. See also Estimation; Estimation methods; Estimator(s)
cluster, 330–331
cluster-specific, 327
conditional maximum likelihood, 247
certainty interval, 215, 440
correlation, 357
exact conditional maximum likelihood point, 390–391
fixed, 334
linear, 241
model-based, 339
odds ratio, 440
parameter, 302
random-effects, 333–334
sandwich, 320, 325, 339, 358
shrinkage in, 183–184
Estimates of effect, 440–441
Estimating equations, 319
Estimation
adaptive quadrature, 322–323, 326–327
of covariant effects, 374
Markov Chain Monte Carlo, 353
maximum likelihood, 228
quadrature, 323
of treatment effect, 377–378
Estimation methods
additional, 20–22, 45–46
choice of, 322–323
classes of, 321–322
for cluster-specific models, 333–334
for correlated-data logistic regression models, 318–323
numerical issues in, 353
quasilikelihood, 321, 352–353
Estimator(s)
covariance matrix, 46
discriminant function, 45–46
information matrix, 272
information sandwich, 320, 325
of the logit, 17
logit-based, 83, 86
Mantel–Haenszel, 83–86
maximum likelihood, 46, 231, 244, 248, 271
pooled log-odds ratio, 380
robust, 320, 325, 339, 358–359
stratified odd ratio, 86
Wald-based confidence interval, 380
Events per covariate, 402, 407–408
Events per parameter, 407–408
Exact conditional maximum likelihood point
estimate, 390–391
Exact distribution, of p sufficient statistics, 388–393
Exact logistic models, results of fitting, 394
Exact logistic regression, 377
Exact methods, 393
for logistic regression models, 387–395
in statistical software packages, 388
Examination process, iterative, 284
Excellent discrimination, 177. See also Receiver Operating Characteristic (ROC) curve
Exchangeable correlation, 318, 357–358
assumption, 343–344
structure, 318–320, 325
Expected frequencies, 207, 284
Exponentiation
of endpoints, 54
of estimated logistic regression coefficients, 278
of logit difference, 51–52
External validation, assessment of fit via, 202–211
Extrabinomial variation, 201
Fagerland–Hosmer goodness of fit test (statistic), 304–305
APS application, results of, 306
F-corrected test statistic, 240
F distribution, 234–235
Fears–Brown model, 232
Final model, 93
preliminary, 282
Finite population correction factor, 234
Firth estimates, 392–393
Firth’s modified likelihood function, 391–392
Fisher’s exact test, 388
Fit assessment, in Bayesian analysis, 429–430
Fit-assessment methods
for multinomial logistic regression model, 283–289
in 1–M matched study, 248–251
Fitted logistic regression model, 58, 60, 85–86, 104
interpretation of, 49–88
results from, 212–223
Fitted logit functions, 79
Fitted logit values, plotting, 103
Fitted models, 8–14. See also Measure of fit;
Model fit
assessing, 162
to Burn Injury Study data, 181
estimated logits for, 17
interpretation of, 49, 77–82
logistic regression, 18
logits for, 80
log-likelihood of, 19
multiple logistic regression, 37–39, 40–45
plot of, 105–106
Fitted multiple logistic regression model, 77–82
Fitted restricted cubic spline model, 106
Fitted values, 18, 80, 153
graphical presentations of, 77
of multiple logistic regression model, 37, 39
presentation and interpretation of, 77–82
Wald-based confidence intervals for, 17–18
Fitting. See also Goodness of fit; Maximum
likelihood fit; Summed measure of fit adjusted, 209–210
assessing, 153–225
of cluster-specific models, 351
of exact logistic models, 394
for multiple logistic regression model, 37–39
numerical problems related to, 145–150
reduced model, 40
of separate binary models, 282–283
of separate logistic models, 282–283
of unconstrained continuation-ratio logit model, 296
of univariable models, 260
Fixed estimates, 334
Forward selection, 127–129
Four-category outcome variable, 272–273
Four-level categorical variable, for examining
scale of continuous covariate, 95–96
Fractional polynomial analysis, 241, 347–348
results of, 117–118
weighted, 238
Fractional polynomial model
one-term, 98, 117–120, 143, 262
two-term, 98, 104–106, 113, 117–119, 141–143
Fractional polynomial procedure, multivariable, 139–144
Fractional polynomials, 113, 121, 342–343
multivariable models and, 99
results of using, 104–105
selecting/checking scale of covariates using
multivariable, 139–144
Fractional polynomials method, 94, 96–99, 109, 111, 382
STATA software package and, 99
Frequency
estimated, 85
estimated expected risk, 160–161
expected, 207, 284
Full conditional likelihood, 247
Full log-likelihood, for cluster-specific model, 321
Fully conditional specifications (FCS), 397
Furnival–Wilson algorithm, 133
Gauss–Hermite quadrature, 321–322
Generalized estimating equations (GEE), 318
estimation, 343–344
INDEX

Generalized logistic model, parameters of, 166–167
Generalized Score statistics, 320
Geometric mean odds, 60–61
Geweke test, 418–419
Gibbs sampler, 397, 413
Gibbs sampling, 409
Global Longitudinal Study of Osteoporosis in Women (GLOW) Study
Bayesian logistic regression models using, 414
GLOW/500 data set, 24–26, 38–39
ALR3_GLOW_BONEMED data set, 382
classification table for, 171–175
classifying the observations of, 170–173
code sheet for variables in, 25
decile of risk strategy and, 160–162
dichotomous independent variable in, 52–53
1-1 matched data set from, 251–259
mediation testing and, 445–448
model, estimated probabilities from, 176
multiple imputation and, 398
plots related to, 194–196
polychotomous independent variable in, 57–58
using purposeful selection in, 107–115
results from MFP cycle fits applied to, 140–143
ROC Curve for, 176–178
“rule of 10” and, 408
sample size with, 402–408
stepwise variable selection procedure applied to, 129–132
two-level models and, 323–337
Goodness of fit, 11
assessing, 153–154
summary measures of, 154–186
Goodness of fit statistics
advantage of, 162
Hosmer–Lemeshow, 157–158
Pearson chi-square, modifications of, 163–164
for population average models, 355–356
Goodness of fit tests. See also Goodness of fit statistics.
chi-square, 232
in 1-1 Matched studies, 259
for the multinomial logistic model, 283–284, 304–305
for proportional odds model, 303
use of, 169
Graphical approach, to diagnostics, 192
Grizzle, Starmer, and Koch (GSK) method, 20–21
Grouped decile of risk type tests, power of, 167
Grouping strategies/methods, for goodness of fit, 157–158, 160, 163
Group mean, 59
Groups
in assessment of population average model fit, 354–355
specifying the number of, 168–169
G statistic, 13, 15, 39–41
Hat matrix (H), 187–188, 249–250, 360, 362
Heidelberger–Welsh stationarity test, 419
Hierarchical models, 316
Highest Density Interval (HDI), 421
Highest Posterior Density (HPD) interval, 421
Histograms, of estimated probabilities, 174–176
Homogeneity, assessment of odds ratio, 86
Homogeneity assumption, 84–85
Hosmer, David W., Jr., xvi
Hosmer–Lemeshow goodness of fit statistic (Ĉ), 158–164, 204, 354
calculation of, 161
Hosmer–Lemeshow tests, 157–169, 204
with the cluster-specific model, 365–366
in correlated data setting, 354
extension of, for multinomial model, 303–304
Hsieh’s correction factor, 406–407
Hypothesis testing error, 167
ICU (intensive care unit) study data set, 22–23
Identity function, 50, 435–436
Identity link models, 436–437, 439–440
Ignorable data, 396
Important variables, including, 92
Imputation chain equations (ICE), 397
Imputations, number of, 400–401
Imputed data sets, fitting a model to, 398
Independence assumption, 313
Independent correlation structure, 318
Independent variables, 1, 13, 36
categorical, 41
collinearities among, 149
dichotomous, 21–22, 50–56
estimated coefficients for, 49
estimating effect of, 444
included in models, 89
outcome variable vs., 442–443
polychotomous, 56–62
relationship with outcome variables, 2
univariable analysis of, 90
Indicator variables
in assessment of population average model fit, 354–355
in Bayesian analysis, 426–427
Indirect effect, 443, 445
Inferences, degrees of freedom and, 397–398
Influence diagnostic, 197, 250
Influence diagnostic statistic, values of, 264
Influential clusters, 371–374
Information matrix, estimator of, 272
Information sandwich estimator, 320, 325. See also Sandwich estimates
Interaction model, 72–73, 114–115
Interaction(s). See also Statistical interaction
among covariates, 262
assessments, 448
coefficients, estimated, 454–455
contrast, 453
estimating and testing additive, 451–456
among main effects, 281–282
of matching variables with model covariates, 262
multiplicative, 451
numerical problems with, interaction terms, 146
purposeful selection of, 259
selecting, 124, 343
statistical significance of, 93
stepwise selection of, 132–133
submultiplicative, 455
terms, 92–93, 132
among variables, 253, 348
variables, 92–93
Intercept coefficient, 208
Intercept only model, 126, 251
Intercepts
predicted, 331
random, 316–317, 347, 348–349
Interval estimators, 15–20
Interval Odds Ratio (IOR), 328–330
Intracluster correlation coefficient (ICC), 241
327, 334–336, 351, 354
confidence intervals for, 345
Iterations, “burn in” period of, 414–415, 419
Iterative examination process, 284
Iterative methods, 9
Jittered values, 178
John Wiley web site, data sets available at, xiv
Joint hypotheses, in population average models, 320
Just Another Gibbs Sampler (JAGS) software package, 409, 413
Knot placement, distribution percentiles defining, 102
Knots, spline functions and, 99–106
Kuo–Mallick (KM) approach, 426–428
Lag, 415–416
Lawless–Singhal method, 133–134
Least squares estimators, 20
Least squares method, 8
Lemeshow, Stanley, xvi. See also Hosmer–Lemeshow entries
Leverage(s), 360–362
cluster-level, 362
magnitude of, 189–190
Leverage values, 187, 249, 253–254
estimated probability vs., 262–263
Likelihood, 13
Likelihood-based confidence interval estimator, 19
Likelihood equations, 9, 231, 271, 436
of multiple logistic regression model, 37
Likelihood function(s), 8–9
case-control, 229–230
contribution to, 8
extension of, 228
Firth’s modified, 391–392
modification of, 231–232
of multiple logistic regression model, 37
in regression sampling, 228
stratum-specific, 228–229
Likelihood intervals, 19
endpoints of, 19
Likelihood ratio, 12
Likelihood ratio chi-square test, 90
Likelihood ratio test(s), 12, 14–15, 18, 39–41, 86, 111, 114–115, 125, 231, 261–262, 276, 280, 295, 345, 350, 353
approximate, 302
using deciles of risk, 163
partial, 97–98, 140–143
Linear discriminant function, 91
Linear equations, 9
Linear estimates, 241
Linearity, in the logit, 63, 94, 103
Linearized models, 321
Linear link binomial model, 451–453
Linear link function, 435–436, 450
Linear mixed effects models, random effects
models vs., 315
Linear models, best models vs., 97–98. See also Linear regression model
Linear regression, 8, 11–12, 249
best subsets, 136, 139
logistic regression vs., 125
INDEX

stepwise, 125
weighted, 164–165, 249
Linear regression model, logistic regression model vs., 1–2, 7
Linear regression software, 134–135
best subsets, 134
weighted least squares best subsets, 139
Linear splines, 99–106
fitting, 103–104
Linear spline variables, 100
Link function(s), 49–50, 203
for binary regression models, 434–441
Burn Injury Study data for fitting, 437–441
linear, 450
maximum likelihood and, 436
roles of alternative, 436
test for choice of, 367
Lipsitz test, 303–305, 309
Log, of odds ratio, 57
Logistic coefficients, 21
Logistic distribution, choosing, 7
Logistic function, model form and, 164
Logistic model(s), 200, 201
binary, 309
fitting separate, 282–283
fitting to sample survey data, 236
parameters of generalized, 166–167
Logistic normal cluster-specific model, 315
Logistic probability, estimated, 44
Logistic regression
advantage of using, 53
for assessing mediation, 445–448
Bayesian methods for, 408–433
best subsets, 133–139
binary, 295–296
exact, 377
guiding principle of, 12
linear regression vs., 125
for matched case-control studies, 243–268
model-building strategies/methods for, 89–151
sampling models for, 227–242
stratified analysis vs., 82–86
underlying theory of, xiii
univariable, 246
model fitting, sample size and, 401–408
Logistic regression analysis
plots in, 193–197
for 2 × 2 tables, 82–86
Logistic regression coefficients, 403–404
estimated, 86
exponentiation of estimated, 278
interpretation of, 50–51
interpretation of adjusted, 443–444
Logistic regression diagnostics, 186–292
extensions of, 284
Logistic regression modeling. See also Logistic regression models
with correlated data, 337–353
propensity score methods in, 377–387
Logistic regression model(s), 1–33, 127
Bayesian, 409–411
for correlated binary data, 375
for correlated-data analysis, 313–375
developing, xiii
exact methods for, 387–395
fitted multiple, 77–82
fitting, 8–10, 58, 60, 85–86
fitting to complex sample survey data, 233–242
fitting to the CHDAGE data, 10
fitting univariable, 107–108
flexibility of, 200
goal of analysis using, 1
interpretation of coefficients for univariable, 50
interpretation of fitted, 49–88
linear regression model vs., 1–2, 7
in 1–1 matched studies, 251–260
in 1–M matched study, 260–267
maximum likelihood estimate (MLE) of, 390–391
multinomial, 269–289
for multinomial and ordinal outcomes, 269–311
multiple, 35–47
numerical problems when fitting, 145–150
ordinal, 289–310
results of fitting, 104, 212–223
slope coefficient in, 50
specific form of, 7
statistical aspects of, xiii
stratum-specific, 244–245
strength of, 35
univariable, 405
values of, 52
wide use of, 229
Logistic regression software packages. See
Software; Software packages
Logit(s)
baseline, 290–291
calculating adjusted, 81
confidence interval estimator for, 43–45
continuous covariate scale in, 94–107
estimated, 17
estimating the variance of estimated, 43–44
estimator of, 16
first estimated adjacent-category, 294
in fractional polynomials method, 96–97
linearity in, 103
Logit(s) (Continued)
lowess smoothed, 119, 122–123
modified, 80, 82
of the multiple logistic regression model, 35–36
95 percent confidence interval for estimated, 45
with population average model, 317
second estimated adjacent-category, 294
third estimated adjacent-category, 295
variance estimators of, 79
Logit assumptions, linear, 63. See also Linearity, in the logit
Logit-based estimators, 83, 86
Logit difference, 66
exponentiating, 51–52
Logit difference estimator, 63
Logit equation, 449
Logit functions, 61, 270, 273, 282
plotting, 74–75
Logit link model, 434
Logit model(s), 438–440
unconstrained continuation-ratio, 296
Logit transformation \[ g(x) \], 7, 35–37, 50
Logit values, fitted, 103
Log-likelihood, 9–10, 13, 321
for fitted baseline model, 295
profile, 19, 20
Log-likelihood-based \( R^2 \), 184
Log-likelihood function, 233–234, 271, 292
Log-likelihood function–based confidence interval, 18
Log-likelihood value, 40
Log link function, 435
Log–log models, 435–436, 438
Log model, 439–440
Log-odds, 300
expression for, 55
estimation of, 64
modification of, 64
Log-odds ratio estimator, estimated standard error of pooled, 380
Log-odds ratio plot, 307
Log-odds ratios, 288–289
equations for, 287
plots for, 287–288
standard error of, 308
Log transformation, 348
Low Birth Weight Study (LOWBWT) data, 24
ordinal logistic regression application, 293–303
Lower confidence limit, 76, 79
Lowess (locally weighted scatterplot smoothing)
method, 102–103, 109–110
Lowess smoothed logit, 119, 122–123
Lowess smoothed plots, 112, 342–343
Main effect coefficient, change in, 73
Main effects, interactions among, 281–282
Main effects model(s), 92, 109, 114, 132, 261–262
for burn injury data, 124
“locking,” 93
preliminary, 281, 341, 347
preliminary final, 349
refining, 94
Mallow’s \( C_q \), 136
Score test approximation of, 137–138
Mann–Whitney U statistic, 178
Mantel–Haenszel estimator, 82–86
Marginal model, 317
Marginal pseudolikelihood (MPL), 322
Marginal quasilikelihood (MQL), 322
Markov chain, 411
Markov Chain Monte Carlo (MCMC) estimation, 353
Markov Chain Monte Carlo (MCMC) simulations, 396–397, 409–419
in Bayesian analysis, 419–433
m-asymptotics, 155, 160
Matched case-control studies, logistic regression for, 243–268
Matched data
methods designed for, 385–386
model building methods for, 247
Matched designs, 1–1 (one to one), 243–244, 250–251
Matched pairs, breaking, 259
Matched sample creation, 381
Matched set effect, estimating on diagnostic statistics, 255
Matched studies 1–1, 251–260
1-M, 260–267
Matching variables, interaction with model covariates, 262
Matrix of second partial derivatives, 271–272
Maximum (M) likelihood (ML), 322
bias in, 387
fit, 135
link functions and, 436
method, 8, 20, 243–244
in multiple logistic regression model, 37
point estimates, exact conditional, 390–391
principle, 9
uses of, 22
INDEX
Maximum likelihood estimates (MLEs), 9–10, 13, 134, 393
conditional, 247
Maximum likelihood estimation, 228
with a misspecified model, 200
Maximum likelihood estimation theory, 37
Maximum likelihood estimators, 8, 19–20, 231, 244, 248, 271
bias in, 391
conditional, 22
discriminant function estimators vs., 21–22
distribution of, 18
under the multivariate normal model, 46
MCLS_i statistic, 363
MCMC algorithms, 412, 414. See also Markov Chain Monte Carlo (MCMC) entries
MCMC chains, convergence of, 417
Mean, estimate of, 18
Measure of fit, 192. See also Fitted entries; Fitting
Median Odds Ratio (MOR), 328–329
Median unbiased estimator (MUE), 393
Mediation, 441–448
assessing, 445–448
Mediational hypothesis, 446
Mediators
adapting for, 444–445
confounders vs., 441–443
in multivariable model, 444
Method of least squares. See Least squares method
Method of maximum likelihood. See Maximum likelihood method
Metropolis Algorithm, 411–412
variations of, 413
Metropolis–Hastings (M–H) algorithm, 413
MFP cycle fits. See Multivariable fractional polynomial procedure
Missing at random (MAR) assumption, 396
Missing completely at random (MCAR) assumption, 395–396
Missing data, 91, 314, 395–401
in longitudinal studies, 395
Missing not at random assumption. See Not missing at random assumption (NMAR)
Misspecified models, maximum likelihood estimation with, 200
Mittlböck–Schemper criteria, 182–184
Mixture distribution, 345
MLWin software program, 332
Model assessment
of the multiple logistic regression model, 39
in validation samples, 205
Model-based approach, 378–379
Model-based estimates, 339
Model-based estimators, 16
Model-based inferences, 82
Model-based methods, 240–242
Model building, 222–223, 337–338
multiple imputations and, 401
with polypharmacy data, 338
purposeful, 340
traditional approach to, 89–90
Model-building methods/strategies/techniques, xiv
for logistic regression, 89–151
for matched data, 247
for multinomial logistic regression, 278–283
for ordinal logistic regression models, 305–310
Model building process, 154
Model checking, missing data and, 401
Model coefficients, estimates of, 212
Model covariates, interaction with model variables, 262
Model fit. See also Fitted models
assessment of, 354–375
assessment via external validation, 202–211
in GLOW Study data, 212
informed decisions about, 169
summary tests of, 167–169
of within-quintile models, 384
Model fit statistics, 200
Model fitting
to imputed data sets, 398
sample size and logistic regression, 401–408
Model form, logistic function and, 164
Modeling
within Bayesian framework, 425–426
of correlated binary data, 374
Model misspecification, 233
Model parameters, inferences about, 234
Models. See also Binary regression models; Data sets, modeling of; Dichotomous–continuous covariate model; Fitted logistic regression model; Linear regression model; Logistic regression models; Multiple logistic regression model; Multivariable models;
Sampling models
adjacent-category logistic, 290–291, 294–296
adjusted, 209–211
alternative, 267
assessing the fit of, 153–225
baseline logit, 290, 294
Bayesian, 409, 414
binary logistic, 309
binary outcome, 273
Models (Continued)
categorical independent variables included or excluded from, 41
cluster-specific binary outcome, 315
complementary log–log, 435–436, 438
conditional, 315
constrained baseline logistic, 291, 294–295
constrained ordinal, 302
continuation-ratio logistic, 290–291, 295–297
continuous outcome, 301
correlated-data, 314–315, 354
discriminant analysis, 170
discovariables included in, 89
discovariables in multivariable, 65
interaction, 72–73, 114–115
intercept only, 126
linearized, 321
linear link binomial, 451–453
linear mixed effects, 315
log, 439–440
logistic, 200–201
logistic normal cluster-specific, 315
logit, 438–440
logit link, 434
log–log, 435–436, 438
main effects, 92–94, 109, 114, 124, 132, 261–262
marginal, 317
maximum likelihood estimation with misspecified, 200
mediators in multivariable, 444
multilevel, 316
multivariate normal, 46
normal theory discriminant function, 231
one-term fractional polynomial, 120
parameters of generalized logistic, 166–167
parsimonious, 116
polypharmacy, 358–359, 363–365
preliminary final, 92–93, 115, 124, 282
preliminary final main effects, 349
preliminary main effects, 92, 109, 116, 281, 341, 347
Probit, 434–436, 438
propensity score, 382–383, 387
proportional odds, 290–292, 297–302, 303, 305
purposeful selection, 131
quadratic, 97–98, 382
random effects, 315–316, 323, 348, 367–368, 413–414
regression sampling, 227–228
restricted cubic splines, 118–119
risk, 386
saturated, 12–13, 184
saturation/calibration of, 186
simple, 386
single-dichotomous-covariate, 273
stratum-specific, 380
stratum-specific logistic regression, 244–245
transitional, 315
two-level, 323–337
unconstrained continuation-ratio logit, 296
univariable logistic regression, 50
well established, 172
Model significance, testing for, 39–42
Model simplification, in multinominal logistic regression, 280
Model validation, 202, 211
Modified logit, 80, 82
Modified Wald statistics, 234–235, 240
Monte Carlo Standard Error (MCSE), 418, 421
Multilevel models, 316
Multinomial likelihood, adaptation of, 292
Multinomial logistic regression model, 269–289
assessment of fit and diagnostic statistics for, 283–289
constrained, 310
goodness of fit test, degrees of freedom for, 304–305
model building strategies for, 278–283
satisfying proportional odds assumption via, 309
Multinomial outcome setting, odds ratios in, 273–278
Multilevel models, 316
Multiple chain production, in MCMC simulations, 416–417
Multiple imputation method, 395–397
Bayesian perspective on, 397
GLOW 500 data and, 398–400
INDEX

model building and, 401
steps in, 396
software packages and, 398
Multiple logistic regression model, 35–47
fitting, 37–39
formulation of, 38
Multiple odds ratios, in multinomial models, 289
Multiplicative interaction, 451, 448
Multiplicative scale, additive scale vs., 448–451
Multiplicity, perfect, 450
Multivariable fractional polynomial procedure,
139–144
applied to Burn Injury Study data, 143–144
applied to GLOW 500 data, 140–143
Multivariable modeling, using fractional
polynomials, 99
Multivariable models, 64–77, 91, 139–140
fitting, 108, 116, 252
independent variables in, 65
mediators in, 444
Multivariable Score test, 42, 340
Multivariable Wald tests, 42, 236–237, 320, 340,
342
Multivariate normal (MVN) distribution, 396
Multivariate normality assumption, 45
Multivariate normal model, maximum likelihood
estimators under, 46
Multivariate test, 15
Myopia study (MYOPIA)
data, 28–31
statistical adjustment illustration, 70–71
n-asymptotics, 155–156
National Burn Repository research data set, 27.
See also Burn Injury Study
National Health and Nutrition Examination
Survey (NHANES) study
complex survey application, 235–242
data, 29, 31
Noise variables, 129
Nominal scale variables, 36
Noniterative weighted least squares method,
20–21
Nonlinearity in the logit, checking for, 238
Normal distribution,
assumption for random effects, 321
standard, 14
Normalized Pearson chi-square, normalized
sum-of-squares vs., 166
Normal probability (PP) plots, 367–369
Normal quantile (QQ) plots, 367–369
Normal theory analysis of covariance model, 228
Normal theory discriminant function model, 231
Not missing at random (NMAR) assumption, 396
Nuisance parameters, 244
Null hypothesis
analogue for Bayesian methods, 421–422
for goodness of fit, 165–166
Numerical integration techniques, 321–323
Numerical problems, in logistic regression model
fitting, 145–150
pooling strategies for, 147
Observed information matrix, 37–38
Observed values, 11–12
Odds, 51. See also Log-odds entries
general mean, 60–61
Odds ratio(s) (OR), 51–56, 212–213
adjusted, 82, 229
for baseline logit model, 294
cluster-specific, 328
confidence intervals for, 62, 274
confidence limits for, 59
constant, 82
correction of, 213
crude (unadjusted), 82, 86
dichotomous variables and, 56
estimated population average, 326
expanding the number of, 276–277
interpretation of, 325–326
log of, 57
as a measure of association, 52, 54
in multinomial outcome setting, 273–278
multiple, 289
for prior fracture, 73–74
relationship of regression coefficient to, 51–52
risk difference vs., 448–451
“Odds ratio approximates relative risk”
argument, 213
Odds ratio constancy assumption, 84
Odds ratio estimates, 56–57, 74–76, 107, 214,
216–219, 258–259, 286, 288–289,
300–302, 325, 327–328, 383–384, 440
Odds ratio estimation, 54, 90, 300–302, 307,
288–289
Odds ratio estimator, 54–55, 61–63
stratified, 86
Odds ratio homogeneity, assessment of, 86
1–1 matched data set, from GLOW Study data,
251–259
1–1 matched design, 243–244
difference data approach to, 250–251
1–1 matched studies, logistic regression model
in, 251–260
1–3 matched data set, from Burn Injury Study
data, 260–267
INDEX

1–M matched study
- fit-assessment methods in, 248–251
- logistic regression model in, 260–267
1-specificity, 174–177, 179–181
Open BUGS statistical package, xiv. See also Bayesian inference Using Gibbs Sampling (BUGS) software package
Optimal cutpoints, 174–176
Optimality properties, of maximum likelihood method, 243–244
Ordinal logistic regression models, 289–310
model-building strategies for, 305–310
Ordinal (scale) outcomes, 289–290, 292–293, 299–300, 302
modeling, 310
Ordinal score, alternative, 304
Orinda Longitudinal Study of Myopia (OLSM) data set, 31. See also Myopia entries
Outcome(s), 179–181
- jittered, 178
- logistic regression models for multinomial and ordinal, 269–311
- logit of, 66
- ordinal (scale), 289–290, 292–293, 299
- predicting, 174
- reference, 273, 293
- regression based on continuous, 298
Outcome categories, pooling, 276, 289
Outcome probabilities, computing in Bayesian analysis, 422
Outcomes Research. Center for, web site, 25
Outcome variable(s), 1–2
- binary, 229, 270, 278, 283
coding, 293
- conditional distribution of, 7
- independent variables vs., 2, 442–443
- nominal, 269, 270
time to event, 228
Outlying clusters, 372–374
Outstanding discrimination, 177. See also Receiver Operating Characteristic (ROC) curve
Overall mean, 59, 66
Overestimation, relative risk, 213
Pairs
deletion of, 258
- fit sensitivity to, 257–258
Parameter distributions, in Bayesian logistic regression models, 410–411
Parameter estimates, 302
- in Bayesian analysis, 424
- computation of, 145
Parameterization. See also Events per parameter of covariates, 96
unconstrained, 291
Parsimonious model, 116
Partial likelihood ratio test, 97–98, 140–143
Partial proportional odds models, 297, 309–310
Pearson chi-square residuals, 166, 188–193, 249, 250, 356, 360–361, 371
- standardized, 191, 250
- summary statistics based on, 186
- variance estimator of, 190
Pearson chi-square ($X^2$) statistic, 135–136, 155–157, 249
- computing the significance of, 166
- decrease in the value of, 191
- goodness of fit testing with, 163–164
- as a measure of lack of fit, 254–255
- value of, 206–207
Pearson chi-square ($X^2$) test, 90, 157, 355
- for the cluster-specific model, 366
Pearson correlation coefficient ($r^2$), squared, 182–184
Penalized quasilikelihood (PQL), 322
Perfect multiplicity, 450
Plots. See also Boxplots; Scatterplots
- advantage of, 76
caterpillar, 331–332
- of estimated logistic regression coefficients, 113
of fitted logit values, 103
of fitted models, 105–106
- lack of fit diagnostic, estimated probability vs., 263–264
- in logistic regression analyses, 193–197
- of logit functions, 74–75
- for log-odds ratio, 287–288, 307
- lowess smoothed, 112, 342–343
- normal probability, 367–369
- normal quantile, 367–369
- of posterior distribution residuals, 432–433
- of profile log-likelihood, 20
- related to Burn Injury Study data, 220–222
- related to GLOW Study, 194–196
- of residuals, 371–373
- sensitivity/specificity, 175
- smoothed, 347
- of squared deviance residuals, 373
- Plotted confidence bands, 78
Point estimates, exact conditional maximum likelihood, 390–391
Points, removing, 362
Polychotomous independent variables, 56–62
design variables for, 57–58
Polychotomous logistic regression model, 269
Polynomials, fractional. See Fractional polynomial entries
Polypharmacy study (POLYPHARM) data, 30–32
model building with, 338, 358–365
Polytomous logistic regression model, 269
Pooled log-odds ratio estimator, estimated standard error of, 380
Poor discrimination, 177. See also Receiver Operating Characteristic (ROC) curve
Population average coefficients, Wald statistics for, 336
assessments of fit of, 354–365
cluster-specific model vs., 334–337
with correlated data, 339–344
weakness of, 328
Population average odds ratio, 326
estimated, 326
Posterior distributions, 330
in Bayesian analysis, 411, 419–420, 424–425, 432
Posterior mean, in Bayesian analysis, 425
Posterior predictive checking, in Bayesian analysis, 429–430
Posterior probabilities
in Bayesian analysis, 426
in MCMC simulations, 412
Posterior simulated values, in Bayesian analysis, 430
Power function, 97
Precision parameters, in Bayesian logistic regression models, 410
Predicted intercept, 331
Predicted probabilities, 332–333
Predicted random effects, 331–332
standard error of, 368
Predicted values, 11–12
missing data and, 401
Predicting outcomes, 174
Predictive squared error, measure of, 136
Pregibon linear regression—like approximation, 190
Preliminary final main effects model, 349
Preliminary final model, 92–93, 115, 124, 282
Preliminary main effects model, 92, 109, 116, 281, 341, 347
Primary sampling units, 233–235
Principle of maximum likelihood. See Maximum likelihood principle
Prior distributions
in Bayesian analysis, 429
in Bayesian logistic regression models, 410
changing, 424
choice of, 423
Prior information weight, tolerance and, 423–424
Prior mean, in Bayesian analysis, 425
Prior probability, in Bayesian analysis, 426–427
Probability. See also Estimated probabilities conditional, 260, 270–271
covariate adjusted, 82
estimated, 303, 333
estimated stratum-specific, 262
leverage values vs. estimated, 262–263
meaning of, 171
population average model and, 317
predicted, 332–333
propensity score and, 378
Probability distributions, of covariates, 230
Probability of miscalculation (PMC), 170–171
Probability of moving, in MCMC simulations, 412
Probit model, 434–438
Profile likelihood confidence interval (CI), 54
estimator of, 43
Profile log-likelihood, 19–20
plot of, 20
Propensity score, 378–380
estimated, 379–380, 382
purpose and properties of, 379
Propensity score methods
advantages and disadvantages of, 387
in logistic regression modeling, 377–387
Propensity score model, 382–383
approaches to using, 387
Proportional odds assumption, 306
not supported by data, 308–310
options for satisfying, 309–310
testing, 302
Proportional odds models, 290–292, 297–302, 305
goodness of fit tests for, 303
partial, 297
Proposal distribution, in MCMC simulations, 412–413
Pseudolikelihood (PL) estimation, 321, 333–334
methods for, 322–323, 352–353
Pseudo-studies, constructing, 407
Public health interaction, 448
for cluster-specific models, 344
of covariates, 70
examples of, 107–124
for population average models, 340
p-value removal, 130–131
p-values, 14, 40–41, 85, 91–92, 127, 240
in Bayesian analysis, 421–422
in stepwise selection procedures, 128
two-tailed, 14–15, 165–166, 203–204, 356, 390
Wald statistic, 261
Quadratic models, 97–98, 382
Quadrature, 321–322
adaptive, 351
Quadrature check, 351–352
Quadrature estimation, 323
adaptive, 322–323, 326–327
Quadrature points, 352
Quartile-based design variables, 121
Quartile design variable analyses, 117
results of, 121
Quartile design variables, 103–104, 110, 112–113
Quasicomplete separation, 148
Quasilikelihood (QL) estimation method, 321, 352–353
Quasilikelihood function, 339
Quasilikelihood information criteria (QIC), 339–340, 343–344, 356–358
Quasilikelihood information criteria approximation (QIC_u), 339–340, 343
Quintiles, analysis using, 381
R^2 measures, 182–186, 356–357, 406
Raftery–Lewis tests, 419
predicted, 331–332
standard error of predicted, 368
Random-effects estimates, 333–334
Random effects models, 315–316, 323, 348, 367–368
linear mixed effects models vs., 315
MCMC simulations and, 413–414
Random effect standard deviation, 345–346
Random intercepts, 316–317, 347–349
Random intercept values, 336
Randomized trials, 228
Random slopes, 349–350
Random variable assumption, in Bayesian logistic regression models, 410
Random variables, chi-square, 14
Ranges of values, 77
Rare disease assumption, 52
Receiver Operating Characteristic (ROC) curve, area under, 173–182, 206. See also ROC analysis
Reduced model fitting, 40, 237–238
Reference cell coding, 55, 57–59
Reference covariate value, 277
Reference levels, 212
Reference outcome, 273, 293
Regression analysis, with dichotomous outcome variable, 7–8
Regression coefficients, 241
relationship to odds ratio, 51–52
Regression diagnostics, 186
Regression methods, 1
Regression sampling model, 227–228
Relative difference, 79
Relative Excess Risk due to Interaction (RERI), 455–456
Relative risk, 52
overestimation of, 213
Relevant distribution theory, 157
Replacement, sampling with, 381
Residual confounding, 384
Residuals
Bayesian, 430–433
empirical, 320
likelihood methods using, 322
plots of, 371–373
posterior distribution plots of, 432–433
tests based on cumulative sums of, 164
Residual sum-of-squares (SSE, RSS), 11–12, 164–165, 186
Response variable
possible predictors of, 126–127
values of, 11
Restricted cubic spline analysis, 121–123
Restricted cubic spline covariate, 101
Restricted cubic spline(s) model, 118–119
results of fitting, 106
Restricted cubic splines, 105–106, 118–120
fit modeling TBSA with, 123
Retroactive data collection, 201
Ridge regression methods, 150
Risk. See also Relative Excess Risk due to Interaction (RERI)
decile of, 160–163, 167–168, 205
relative, 52
Risk difference, odds ratios vs., 448–451
Risk factors, 68
adding, 385
modeling, 389
Risk overestimation, 213
Risk ratio, 456
R matrix, 319
Robust estimator, 320, 325, 339, 358–359
ROC analysis, 289. See also Receiver Operating Characteristic (ROC) curve
INDEX

R R Development Core Team statistical package, xiv
“Rule of 10,” 407–408
Rule of thumb, in Bayesian analysis, 428–429
R values, in MCMC simulations, 418
Sampled clusters, 241
Sample distribution, of Wald statistic, 403
Sample size(s), 168–169
  logistic regression models and, 401–408
  in MCMC simulations, 418
Sample size questions, 401–402
Sample survey data
  complex, 233–242
  fitting logistic models to, 236
  regression modeling of, xiv
Sampling, adaptive rejection, 413
Sampling distribution, 54
Sampling models, for logistic regression, 227–242
Sampling rates, stratum-specific, 232
Sampling units, primary, 233–235
Sampling with replacement, 381
Sandwich estimates, 325, 339, 358. See also Information sandwich estimator
SAS procedures
  GLIMMIX procedure, 331, 334, 352, 367
  logistic regression (PROC Logistic), 19, 41, 137
  PROC Logistic output, 301
SAS statistical package, xiv, 129, 132, 249, 353, 413. See also Software packages/programs
Bayesian methods software in, 409
diagnostics in, 188
missing data and, 396–400
PL estimation in, 333–334
score test for proportional odds assumption in, 302
Saturated models, 12–13, 184, 186
Scale variables, discrete nominal, 62
Scatterplots, 2
  of presence/absence of coronary heart disease, 5
  smoothed, 94–95
Score test, 14–15, 86, 129, 137, 163, 167. See also Generalized Score statistics
  approximation of Mallow’s $C_p$, 137–139
  multivariable, 340
  multivariable analog of, 42
Second partial derivatives, matrix of, 271–272
Sensitivity/specificity
  for classification tables, 175
  plots of, 174–176
Spline covariates, restricted cubic, 101. See also Cubic spline entries
Spline functions, 94, 99–102
  knots and, 99–106
  functions method, 109
Sequential regression multivariate imputation (SRMI), 397
Sequential test procedure, 98
Shapiro–Wilk test, 369
Shrinkage, 183–184
Significance levels, 91–92, 140, 309
Single-dichotomous-covariate model, 273
Single independent variables, 14
Single prior distribution, in Bayesian analysis, 429
Slope coefficients, 39, 50–51, 53
  estimates of, 275, 394
Slope parameter, 421
  in Bayesian analysis, 423
Slopes, random, 349–350
Smoothed plots, 347. See also Lowess entries
Smoothed scatterplots, 94–95
Software packages/programs. See also SAS entries; STATA entries; SPSS software package; MLWin software program; SUDAAN software; OpenBUGS statistical package; Just Another Gibbs Sampler (JAGS) software package
  for Bayesian methods, 409
capabilities of, xiii
  complex sample surveys in, 233
  conditional logistic regression in, 247
  correlated-data modeling, 314–315
design variables in, 55, 57
differences among, xiv
deviance vs. log-likelihood in, 12
exact methods in, 388
handling of weights in, 249
modified Wald statistic in, 234–235
multinomial logistic regression model diagnostics in, 284
  for multivariable fractional polynomial methods, 139
  point and confidence interval estimates in, 54
  score test in, 43
  weighted least squares best subsets linear regression, 139
  weighted ordinary logistic regression programs in, 239
  zero cells in, 90
Specificity. See also 1-specificity
  for classification tables, 175
  plots of, 174–176
Spline functions, 94, 99–102
  knots and, 99–106
  functions method, 109
Splines, restricted cubic, 105–106
Spline variables
  cubic, 101
  linear, 100
SPSS software package, 57
Squared deviance residuals, plots of, 373
Squared Pearson correlation coefficient ($r^2$), 182–184
S-shaped curve, 6
Standard deviation, random effect, 345–346
Standard error(s), 45
  estimated, 17, 59, 62, 149, 231–232, 274, 278, 325, 327
  estimation of, 37
  estimators of, 16, 63
  of log-odds ratio, 308
  of pooled log-odds ratio estimator, 380
  of predicted random effects, 368
Standardized comparative residuals, 368–369
Standardized Pearson residual, 191
Standardized residuals, boxplots of, 370–371
Standardized Pearson chi-square statistic, 203
Standard normal distribution, 14
STATA commands/procedures/programs
clogit command, 251, 260
GLLAMM procedure, 370–371, 373
  for Pearson chi-square statistic, 166
psmatch2 program, 385
test/lincom commands, 276
xlogit procedure, 331
xtmelogit procedure, 331
STATA log option, for fractional polynomial analysis, 118
conditional logistic regression in, 247
cubic spline variables and, 101
diagnostics in, 188, 248
fractional polynomial method and, 99
  lowest smooth via, 102–103
  missing data and, 396–400
Stationarity test, Heidelberger-Welch, 419
  Statistical adjustment, 64, 66–67, 69, 70–72
  mediation and, 441
  Statistical analyses, of survey data, 240
Statistical considerations, for fractional polynomial models, 106
Statistical evidence, for variables, 14
Statistical hypothesis, formulating and testing, 10
Statistical interaction, 64, 69–73, 448–456
  presence and absence of, 68–69
  Statistically important variables, 131
  Statistically significant interaction, 77
Statistical model building, traditional approach to, 89–90
Statistical packages, xiv. See also SAS entries;
  STATA entries; Software packages/programs
Statistical significance, of interactions, 93
Statistical software packages. See Software packages/programs
Statistics. See also Diagnostic statistics; Model fit
  statistics; Pearson chi-square ($X^2$) statistic
  goodness of fit, 355–356
  influence, 255–256, 360–363
  standardized, 203
Stepwise backward elimination, 134. See also
  Backward elimination
Stepwise covariate/variable selection, 125–133
  of interactions, 132–133
  method for, 93–94
  for multinomial models, 279
  results of applying, 130
Stepwise linear regression, 125
Stepwise selection procedure. See also Four-step
  process
  applied to GLOW data, 129–132
  modification of, 129
  $p$-values in, 128
  for variables, 90–93, 128
Stepwise variable selection, 279
  results of applying, 130
Strata
  accessing, 250
  deleting, 265–266
  uninformative, 260
Stratification variables, 147, 228, 243–244
Stratified analysis, 385
  of case-control data, 232
  logistic regression vs., 82–86
  for 2 $\times$ 2 tables, 82–86
Stratified contingency table analysis, 50
  estimates, 86
  odd ratio estimator, 86
Stratum number, stratum sum vs., 265
Stratum-specific
  likelihood functions, 228–229
  logistic regression model, 244–245
  mean, weighted, 248
  models, 380
  probability, estimated, 262
  sampling rates, 232
  totals, of diagnostics, 250
Structural zero, 308–309
Stukel test, 166–167, 436, 438
Sturdivant, Rodney X., xvi
Subject-specific covariates, 317
diagnostic statistics, 359–360
pseudolikelihood (SPL), 322
Submultiplicative interaction, 455
SUDAAN software, 233–234
Sufficient statistics, exact distribution of $p$, 388–393
Sum, variance of, 16–17
Summary statistics, 154–155
in Bayesian analysis, 420–421
Summed measure of fit, 255–256
Sum-of-squares $(S)$, 155–156, 183–184, 204
residual, 11–12, 186
total, 11
value of, 206–207
weighted residual, 135
Superadditivity, 450–451, 455
Survey data
complex, 233–242
statistical analyses of, 240
Tarone test, 85
Taylor expansion, 322
t-distribution
in Bayesian analysis, 429
in multiple imputation, 397–398
Test statistics
for likelihood ratio test, 12
for score test, 15
for univariable Wald test, 40
Thin data, 260
Time-invariant covariates, 313
Time-to-event data, 228–229
Time-varying covariates, 313
Tolerance, prior information weight and, 423–425
Tolerance parameters, in Bayesian logistic regression models, 410
Total effect, 443
Total sum-of-squares, 11
Trace plot, 414–417
Transitional model, 315
Treatment effect estimation, 377–387
t-tests
in correlated data, 353
two-sample, 91
univariable analysis based on, 91
Two degree of freedom likelihood ratio test, in multinomial logistic modeling, 280
Two-level models, GLOW data and, 323–337
Two-sample t-test, 91. See also t-tests.
Two-tailed $p$-value, 14–15, 165–166, 203–204, 356, 390
$2 \times 2$ classification tables, 171–173
$2 \times 2$ tables, logistic regression vs. stratified analysis for, 82–86
$U$ matrix, in 1-M matched study diagnostics, 248–249
Unadjusted difference, adjusted difference vs., 67
Unadjusted odds ratio, 82
Unavailable data problem, 235–236
Unconstrained continuation-ratio logit model, fitting, 296. See also Continuation-ratio logistic model.
Unconstrained parameterization, 291
Uncontrolled confounding, 447
Uninformative case-control pairs, 246
Uninformative stratum, 260
Univariable analyses, 65, 90–91, 340–341, 344–346
of continuous variables, 91
of independent variables, 90
Univariable (model) coefficient, 70–72
Univariable linear discriminant function, 91
Univariable logistic regression, 246
Univariable logistic regression model, 405
fitting, 107–108
interpretation of coefficients for, 50
Univariable models
fitting, to assess thin data, 260
results of fitting, in 1-1 matched study, 251–252
Unstructured correlation structure, 318
Upper confidence limit, 76, 79
$U$ statistic, Mann-Whitney, 178
Validation data, 168, 202–203. See also External validation, assessment of fit via model assessment in, 205
Variable deletion, 127
Variables. See also Design variables; Response variable
Adolescent Placement, 305–310
binary outcome, 283
categorical, 95
cluster-level, 330
confounding, 456
continuous, 106–107
continuous response, 297–298
cubic spline, 101
dichotomous, 69, 170
grouping, 303
importance of, 125–126
including important, 92
Variables (Continued)
indicator, 354–355
interaction, 92–93
interactions among, 253, 348
linear spline, 100
minimizing the number of, 90
for multiple logistic regression model, 35–36
ordinal outcome, 290, 300
outcome, 452
quartile design, 103–104, 110, 112–113
removal of, 252
significance of, 39, 279–281
single independent, 14
statistical evidence for, 14
statistically important, 131
stepwise selection of, 279
stratification, 228, 243–244

Variable selection
approaches to, 93–94
criteria for, 136
methods, 128
pitfalls of, 94
steps in, 90–93, 128
tests used in, 353

Variable significance, assessment of, 10–15

Variance
assumed, in cluster-specific model, 327
estimation of, 62
of a sum, 16–17

Variance estimators, 37–38, 207, 232
of logits, 79
of residuals, 190

Variation, extrabinomial, 201

Vector notation, for logit confidence interval estimator, 43
Vector of coefficients (β), in matched case-control studies, 244–245
Venzon–Moolgavkar method, for likelihood-based confidence intervals, 18–19

Visual assessment, of diagnostics, 192–193
Vittinghoff–McCulloch simulations, for sample size determination, 408

V matrix, 38, 134–135, 187, 319

Wald-based confidence interval (CI), 16–17
asymmetry of, 19–20
for coefficients, 16
estimator of odds ratio, 380
for fitted values, 17–18
for logit, 17

Wald statistic, see Wald test statistic
Wald (W) test(s), 14–16, 70, 234, 353
adjusted, 237
Brant’s, 302, 306
equivalence to Score test, 14–15
multivariable, 42, 236–237, 320, 340, 342
Wald (W) test statistic(s), 40–42, 69–70, 72, 237
adjusted, 235–237
approximation, 137
modified, 234–235, 240
for population average coefficients, 336
p-values, 91–92, 261
sample distribution of, 403

Weighted fractional polynomial analysis, 238

Weighted least squares best subsets linear regression software, 139

Weighted linear regression
used in model fit assessment, 164–165
used in 1-M matched study fit assessment, 249

Weighted ordinary logistic regression program, 239

Weighted residual sum-of-squares, 135

Weighted stratum-specific mean, 248

Weighting, in statistical packages, 249

Whitemore formula for sample size, 403
modifications of, 406

Wiley web site, data sets available at, xiv

Within-chain variability (W), in MCMC simulations, 417–418

Within-cluster correlation, 316
Within-cluster covariance, 319–320
Within-quintile models, fit of, 384
W (weight) matrix, 234

Working correlation, 318

X matrix (design matrix), 38, 134–135, 187, 234, 248–249

Zero, structural, 308–309
Zero (frequency) cell, in contingency tables, 90, 145–147