

Index

- 95% confidence interval 27
- 100 α % critical value 30
- Acute Physiology and Chronic Health
 - Evaluation (APACHE) score, definition 3
- algebraic notation 1–2
- alternative hypothesis 26–7
- alternative models, hazard regression analysis 330
- analysis of covariance, fixed-effects analysis of variance 447
- analysis of variance 429 *see also* fixed-effects analysis of variance; repeated-measures analysis of variance
- area-under-the-curve response feature 468–9
- automatic methods of model selection 119–24
- bandwidth (lowess regression) 64
- Bernoulli distribution 163
- binomial distribution 162–3, 376
- box (and whiskers) plot 5–6
 - definition of outer bars 6
 - interquartile range 6
 - outliers 6
 - quartiles 6
 - skewed distributions 6
- Stata display 20, 21, 22
- breast cancer and estrogen receptor genotype, fixed-effects analysis of variance examples 435–8, 439–46
- Breslow–Day–Tarone test for homogeneity 204–6
- case–control studies
 - 95% confidence interval for odds ratio (Woolf’s method) 189
 - analysis of case–control data using Stata 195–6
 - case–control theory 188–9
 - frequency matched 258
 - Ille-et-Vilaine study of esophageal cancer risk 187–96
 - observed odds 188
 - observed odds ratio 188
 - odds 188
 - odds ratio 188
 - regressing disease against exposure 197–8
 - simple logistic regression 2×2 study example 187–90
 - test of null hypothesis that the odds ratio equals one 190
 - test of null hypothesis that two proportions are equal 190
- categorical hazard regression model 323–4
- categorical response variables, regression models 40–1, 278–81
- censoring and bias, survival analysis 296
- chi-squared distribution 38–9
- collinearity problem 124–5
- conditional expectation 48–9
- conditional logistic regression 258
- confounding variables
 - adjusting for 327–9
 - multiple linear regression 98
 - multiple logistic regression 201, 240
 - multiple Poisson regression 410–11
- continuous response variables, regression models 40, 41
- Cook’s distance 127–8
- correlated variables (population), definition 48
- correlation coefficient (population) 48
- correlation coefficient (sample) 47
- correlation matrix 471
- correlation structures 470–1
- covariance (population) 47
- covariance (sample) 45–7
- covariate (independent variable) 39
- covariates (explanatory variables) 97
- cumulative morbidity curve 289, 291, 295
- cumulative mortality function 287–8, 295
- data imputation methods 259–60
- data with missing values *see* missing data
- degrees of freedom 29–39
- density-distribution sunflower plots 87–92
- dependent (response) variable (multiple linear regression) 97
- dependent variables *see* response variables 39–40
- descriptive statistics
 - box (and whiskers) plot 5–6, 20, 21, 22
 - definition 1
 - dot plot 3–4, 20, 21, 22
 - example (ethylene glycol poisoning study) 6–7
 - example (Ibuprofen in Sepsis Study) 3–6, 7
 - histogram 6, 7, 9–12, 16–20

Cambridge University Press

978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Second Edition

William D. Dupont

Index

[More information](#)

- descriptive statistics (*cont.*)
 - median 5, 20–2
 - percentiles 5
 - residual 4
 - sample mean ($\bar{x}$) 4, 20–2
 - sample standard deviation (s) 5, 20–2
 - sample variance (s^2) 4–5, 20–2
 - scatter plot 6–7
 - Stata display 20–2
- deviance residual 423–4
- dichotomous response variables, regression
 - models 40, 41
- disease-free survival curve 288–9, 291
- dot plot 3–4, 20, 21, 22
- effect modifiers (multiple logistic regression) 240
- error (simple linear regression model) 49–50
- error messages (Stata) 14
- error standard deviation 49–50
- error variance 49
- esophageal cancer risk study *see* Ille-et-Vilaine Study of esophageal cancer risk
- estimated variance–covariance matrix 223–4
- ethylene glycol poisoning study (Brent et al., 1999)
 - descriptive statistics 6–7
 - simple linear regression examples 45–7, 54–66, 69–70
- exchangeable correlation structure 471
- expected response 99–100
- expected value of a statistic 25
- explanatory variables (covariates) 97
- F* distribution 430–1
- F* statistic 431
- Fisher's protected LSD (least significant difference) procedure 431–3
- fixed-effects analysis of variance 429–47
 - analysis of covariance 447
 - definition 429, 430
 - examples (breast cancer and estrogen receptor genotype) 435–8, 439–46
- F* distribution 430–1
- F* statistic 431
- Fisher's protected LSD (least significant difference) procedure 431–3
- Kruskal–Wallis test 435, 438
- multiple comparisons 431–3
- non-parametric methods 434–5
- one-way analysis of variance 429–31, 435–46
- P*-value adjustment for multiple comparisons 431–3
- parametric methods 434
- reformulating as a linear regression model 433–4
- two-way analysis of variance 446–7
- variety of models 446–7
- Wilcoxon–Mann–Whitney rank-sum test 435, 438
- Wilcoxon signed-rank test 435
- fixed-effects models 40–1
- Framingham Heart Study
 - hazard regression analysis examples 317–48, 350–70
 - multiple linear regression examples 102–24, 125–6, 128–33
 - multiple Poisson regression examples 404–27
 - Poisson regression examples 373–6, 379–81, 386–97
 - simple linear regression examples 66–9, 81–92
- frequency matched case–control studies 258
- generalized estimating equations (GEE)
 - analysis
 - nature of response variables 40, 41
 - repeated-measures analysis of variance 470, 471, 472–81
 - with logistic models 481
 - with Poisson models 481
- generalized linear model 163–4, 378–9
- genetic risk of recurrent intracerebral hemorrhage, survival analysis examples 291–5, 298–305, 310–11
- goodness-of-fit tests 240–4, 248–57
- graphics capabilities of Stata 14–15
- graphics editors 15
- grouped response data, logistic regression 176, 181–7
- hazard functions, survival analysis 306–7
- hazard ratios 307–9, 315–16
- hazard regression analysis 315–70
 - 95% confidence intervals 317
 - adjusting for confounding variables 327–9
- age-specific relative risks of CHD in men and women 359–70
- alternative models 330
- categorical hazard regression model (CHD risk and DBP) 323–4
- examples (Framingham Heart Study) 317–48, 350–70
- hazard ratios 307–9, 315–16
- hypothesis tests 317
- interaction of age and sex on CHD 350–1
- interpretation of models 329–30
- Kaplan–Meier survival curves 317–18
- Kaplan–Meier survival curves with ragged entry 350
- log–log plots and the proportional hazards assumption 354–9
- log-rank test with ragged entry 350
- model deviance 317
- multiplicative model (DBP and gender on risk of CHD) 325–6
- nested models 317
- predicted survival and the proportional hazards assumption 354–9
- proportional hazards assumption 354–61

Cambridge University Press

978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Second Edition

William D. Dupont

Index

[More information](#)

- proportional hazards assumption testing 361–70
- proportional hazards model 309–11, 315
- proportional hazards regression analysis 331–48, 351
- ragged study entry 349–54
- relative risks 315–16
- restricted cubic spline model (CHD risk and DBP) 320–3
- simple hazard regression model (CHD risk and DBP) 318–20
- simple hazard regression model (CHD risk and gender) 324–5
- step-functions 359–61
- stratified proportional hazards models 348–9
- survival analysis with ragged study entry 349–54
- time-dependent covariates 359–70
- using interaction terms to model effects of gender and DBP on CHD 326–7
- hazard regression models, nature of response variables 40, 41
- heteroscedastic error terms 50
- histograms 6, 7, 9–12, 16–20
- homoscedastic error terms 50
- Hosmer–Lemeshow goodness-of-fit test 242–4, 249, 255
- Huber–White sandwich estimator 472–5
- hypothesis tests
 - for weighted sums of coefficients 223
 - hazard regression analysis 317
 - multiple linear regression 101
- Ibuprofen in Sepsis Study (Bernard et al., 1997)
 - analyzing data with missing values 260–5
 - descriptive statistics 3–6, 7
 - descriptive statistics in Stata 9–12, 20–2
 - downloading data sets 8
 - independent *t* tests 34–5, 36–8
 - paired *t* test 30–4
 - simple logistic regression examples 159, 160, 170–6, 178–87
 - survival analysis example 288–9
- Ille-et-Vilaine Study of esophageal cancer risk
 - multiple logistic regression examples 201–9, 212–40, 242–4, 246–57
 - simple logistic regression examples 187–96
- imputed value 259
- incidence 373–4
- independent observations 23
- independent *t* tests
 - using a pooled standard error estimate 34–5
 - using separate standard error estimates 35–6
 - using Stata 36–8
- independent variable (covariate) 39
- independent variables (population), definition 48
- inferential statistics 22–39
 - 95% confidence interval 27
 - 100 α % critical value 30
- alternative hypothesis 26–7
- chi-squared distribution 38–9
- definition 1
- expected value 25
- independent observations 23
- independent *t* test using a pooled standard error estimate 34–5
- independent *t* test using separate standard error estimates 35–6
- independent *t* tests using Stata 36–8
- mean (μ) 24–5
- mutually independent observations 23
- normal distribution 24–5
- null hypothesis 26–39
- P*-value 26–7, 29–39
- paired *t* test 30–4
- parameters of the target population 24–5
- power curves 27–8
- power of a test 27–8
- probability density functions 23–7, 29–30, 38–9
- random variables 23
- sample size 28
- sampling bias 23
- significance level 26–30
- standard deviation (σ) 24–5
- standard error 25
- standardized normal distribution 29–30
- statistical power 27–8
- statistics of a sample 24
- Student's *t* distribution 29–30
- t* distributions 29–30
- target population 22–39
- Type I error 27
- Type II error 27
- unbiased estimate of a parameter 25
- unbiased sample 23
- variance (σ^2) 24
- z* distribution 29–30
- influence of individual observations
 - multiple linear regression 126–9
 - simple linear regression 66–7
- interaction terms, use in hazard regression analysis 326–7
- Internet sources of Stata programs 13–14
- interpretation of hazard regression models 329–30
- isoproterenol and blood flow study, repeated-measures analysis of variance examples 451–9, 460–8, 473–81
- jackknife residual 68, 125
- Kaplan–Meier cumulative mortality function 290–1
- Kaplan–Meier survival curves
 - hazard regression analysis 317–18, 350
 - survival analysis 290–6, 298–305
 - with ragged entry 350
- Kaplan–Meier survival function 290–1
- Kruskall–Wallis test 435, 438

Cambridge University Press

978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Second Edition

William D. Dupont

Index

[More information](#)

- least squares estimate 51, 99
- leverage
 - of an independent variable 67–8
 - of specific patient responses 101–2
- life tables 291
- likelihood function 164–6
- likelihood ratio tests 166–7
- linear predictor of the generalized linear model 163
- linear regression
 - contrast with logistic regression 164
 - origin of the term 52
 - see also* multiple linear regression; simple linear regression
- linear regression line 51
- linear regression models, nature of response variables 40
- link function of the generalized linear model 164
- log–log plots and the proportional hazards assumption 354–9
- log odds of death under a logistic probability function 161
- log-rank test
 - hazard regression analysis 350
 - survival analysis 296–8, 299, 302–5
 - with ragged entry 350
- logarithmic transformation of variables 71
- logistic probability function 159–61
- logistic regression models
 - generalized estimating equations (GEE) analyses 481
 - nature of response variables 40–1
 - see also* multiple logistic regression; simple logistic regression
- logit function 161
- longitudinal data 41, 451
- lost to follow-up, patients 289–90
- lowess regression 64–6
 - bandwidth 64
- Mantel–Haenszel chi squared statistic for multiple 2×2 tables 203–4
- Mantel–Haenszel estimate of an age-adjusted odds ratio 201–3, 206–9
- Mantel–Haenszel test for survival data (log-rank test) 296–8, 299, 302–5
- maximum likelihood estimation 164–6
- mean (μ) 24–5
 - sample mean ($\bar{x}$) 4
- mean squared error (MSE) (s^2) 53–5, 100
- median 5, 20–2
- missing data 258–65
 - analyzing data with missing values 258–65
 - classification 259–60
 - data imputation methods 259–60
 - imputed value 259
 - informatively missing 259
 - missing at random 259
 - missing completely at random 259
 - modeling missing values with Stata 263–5
 - multiple imputation 260
 - single imputation 259
 - Stata data imputation program 259–60
- model deviance
 - hazard regression analysis 317
 - multiple logistic regression 238–40
- model fitting (multiple logistic regression) 238–40
- model selection 40–1
- model selection methods
 - automatic methods with Stata 119–24
 - backward selection 122–3
 - backward stepwise selection 123–4
 - forward selection 120–2
 - forward stepwise selection 123
- model sum of squares (MSS) 54, 99
- multinomial logistic regression *see* polytomous logistic regression
- multiple 2×2 contingency tables, logistic regression 212–16
- multiple comparisons problem (multiple linear regression) 124
- multiple imputation 260
- multiple linear regression 97–154
 - 95% confidence interval 114
 - 95% confidence interval for $\hat{y}_i$ 102
 - 95% prediction interval 102, 114
 - accuracy of parameter estimates 100
 - automatic methods of model selection 119–24
 - collinearity problem 124–5
 - confounding variables 98
 - Cook's distance 127–8
 - covariates (explanatory variables) 97
 - dependent (response) variable 97
 - estimating the parameters 99–100
 - examples (Framingham Heart Study) 102–24, 125–6, 128–33
 - examples (SUPPORT study of hospitalized patients) 138–54
 - expected response 99–100
 - explanatory variables (covariates) 97
 - exploratory graphics 102–7
 - hypothesis tests 101
 - influence analysis using Stata 129–33
 - influence of individual observations 126–9
 - $\Delta\beta$ influence statistic 127
 - intuitive understanding of a multiple regression model 110–14
 - jackknifed residual 125
 - least squares estimates of parameters 99
 - leverage of specific patient responses 101–2
 - mean squared error (MSE) (s^2) 100
 - model 97
 - model selection methods 119–24
 - model sum of squares (MSS) 99
 - modeling interaction 107–9
 - multiple comparisons problem 124
 - non-linear models 133–4
 - non-linear models with restricted cubic splines 134–54

Cambridge University Press

978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Second Edition

William D. Dupont

Index

[More information](#)

- R^2 statistic 99
- regression coefficients 97
- residual analyses 125–6, 129–33
- residual values 99
- response (dependent) variable 97
- restricted cubic spline models 134–54
- root MSE (s) 100
- scatterplot matrix graphs (exploratory graphics) 105–7
- standardized residual 125
- modeling using Stata 114–19
- studentized residual 125–6
- total sum of squares (TSS) 99
- multiple logistic regression 201–81
 - 95% confidence interval for adjusted odds ratio 204, 211
 - analyzing data with missing values 258–65
 - analyzing multiple 2×2 tables with Stata 214–16
 - Breslow–Day–Tarone test for homogeneity 204–6
 - categorical response variables 278–81
 - categorical variables in Stata 216–17
 - conditional logistic regression 258
 - confidence intervals for weighted sums of coefficients 222–3
 - confounding variables 201, 240
 - data imputation methods 259–60
 - effect modifiers 240
 - estimated variance–covariance matrix 223–4
 - examples (hospital mortality in the SUPPORT study) 267–78
 - examples (Ibuprofen in Sepsis Study) 260–5
 - examples (Ille-et-Vilaine Study of esophageal cancer risk) 201–9, 212–40, 242–4, 246–57
 - fitting a model with interaction using Stata 234–8
 - fitting a multiplicative model with Stata 227–31
 - frequency matched case–control studies 258
 - goodness-of-fit tests 240–4
 - Hosmer–Lemeshow goodness-of-fit test 242–4
 - hypothesis tests for weighted sums of coefficients 223
 - influence analysis 244–8
 - $\Delta\hat{\beta}_j$ influence statistic 245–8
 - leverage 244–8
 - likelihood ratio test of influence of covariates on response variable 211
 - Mantel–Haenszel chi squared statistic for multiple 2×2 tables 203–4
 - Mantel–Haenszel estimate of an age-adjusted odds ratio 201–3, 206–9
 - missing data 258–65
 - model 210–11
 - model deviance 238–40
 - model fitting 227–31, 234–8, 238–40
 - model of alcohol, tobacco and esophageal cancer with interaction terms 233–4
 - model of two risk factors with interaction 231–2
 - modeling missing data with Stata 263–5
 - multinomial logistic regression *see* polytomous logistic regression
 - multiple 2×2 contingency tables 212–16
 - multiplicative model of smoking, alcohol and esophageal cancer 225–7
 - multiplicative models of two risk factors 224–31
 - nested models 238–40
 - odds ratios from multiple parameters 221
 - ordered logistic regression *see* proportional odds logistic regression
 - Pearson chi-squared goodness-of-fit statistic 241–2, 243
 - Pearson residual 241
 - polytomous logistic regression 278, 279–81
 - proportional odds logistic regression 278–9
 - residual analysis 244–8, 249–57
 - restricted cubic spline models 265–78
 - standard error of a weighted sum of regression coefficients 222
 - standardized Pearson residual 245
 - multiple Poisson regression 401–27
 - confounding variables 410–11
 - deviance residual 423–4
 - examples (Framingham Heart Study) 404–27
 - model 401–4
 - multiplicative model 405–7
 - multiplicative model with interaction terms 408–10
 - nuisance parameters 402, 403–4
 - residual analyses 423–7
 - standardized deviance residual 424–7
 - using Stata 411–23
 - see also* Poisson regression
 - multiplicative model (hazard regression analysis) 325–6
 - multiplicative models of two risk factors (multiple logistic regression) 224–31
 - mutually independent observations 23
 - nested models 238–40, 317
 - non-linear models (multiple linear regression) 133–4
 - with restricted cubic splines 134–54
 - non-linearity, correcting for 71–2
 - non-parametric methods, fixed-effects analysis of variance 434–5
 - normal distribution 24–5
 - nuisance parameters
 - multiple Poisson regression 402, 403–4
 - Poisson regression 378
 - simple logistic regression 191
 - null hypothesis 26–39
 - observed incidence 373–4
 - odds of death under a logistic probability function 161

Cambridge University Press

978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Second Edition

William D. Dupont

Index

[More information](#)

- odds ratios and the logistic regression model 174
- odds ratios from multiple parameters 221
- offset, Poisson regression 378
- one-way analysis of variance 429–31, 435–46
- ordered logistic regression *see* proportional odds logistic regression
- P*-value 26–7, 29–30
 - adjustment for multiple comparisons 431–3
 - critical region of values 26
 - two-sided (two-tailed) 26
- paired *t* test 30–4
- panel data 451
- parameters of a target population 24–5, 99–100
- parametric methods, fixed-effects analysis of variance 434
- patients lost to follow-up 289–90
- Pearson chi-squared goodness-of-fit test 241–2, 243, 249, 255
- Pearson residual 241
- percentiles 5
- person–year unit of observation *see* Poisson regression
- Poisson distribution 376
- Poisson model, generalized estimating equations (GEE) analyses 481
- Poisson regression 373–97
 - binomial distribution 376
 - classification variable 393–5
 - contrast with simple linear and logistic regression 379
 - converting survival data to person–time format 381–92
 - elementary statistics involving rates 373–4
 - example of a generalized linear model 378–9
 - examples (Framingham Heart Study) 373–6, 379–81, 386–97
 - for 2×2 tables 376–8
 - incidence 373–4
 - model 378
 - multiple data records 392–3
 - nature of response variables 40, 41
 - nuisance parameters 378
 - observed incidence 373–4
 - offset 378
 - person–year unit of observation 373, 381–92
 - Poisson distribution 376
 - relative risk 374–6
 - survival analysis 381–92
 - using Stata 379–81, 395–7
 - see also* multiple Poisson regression
- polytomous logistic regression 278, 279–81
 - nature of response variables 40–1
- population correlation coefficient 48
- population covariance 47
- power calculations 27–8
- power curves 27–8
- power of a test 27–8
- predicted survival, proportional hazards assumption 354–9
- probability density functions 23–7, 29–39
 - degrees of freedom 29–39
- probability distribution 162–3
- product limit survival function 290–1
- proportional hazards 306–7
- proportional hazards assumption 354–61
 - testing (time-dependent covariates) 361–70
- proportional hazards model 309–11, 315
- proportional hazards regression analysis 309–11, 331–48, 351
- proportional odds logistic regression 278–9
 - nature of response variables 40–1
- PS program 28
- quadratic approximations to the log likelihood ratio function 167
- quasi-likelihood technique 472
- R^2 statistic (multiple linear regression) 99
- ragged study entry 349–54
- random component of the generalized linear model 163
- random variables 23
- rates (response variables), regression models 40, 41
- regression, origin of the term 52
- regression coefficients 97
- regression models
 - multiple responses per patient 40, 41
 - one response per patient 40–1
 - selection of 40–1
- relative risk 307–9, 315–16, 374–6
- repeated-measures analysis of variance 451–81
 - area-under-the-curve response feature 468–9
 - common correlation structures 470–1
 - correlation matrix 471
 - definition 451
 - examples (isoproterenol and blood flow study) 451–9, 460–8, 473–81
 - exchangeable correlation structure 471
 - exploratory analysis of repeated-measures data 453–9
 - generalized estimating equations (GEE) analysis 470, 471, 472–81
 - Huber–White sandwich estimator 472–5
 - longitudinal data 451
 - panel data 451
 - quasi-likelihood technique 472
 - response feature analysis 459–69, 475–6
 - two-stage analysis *see* response feature analysis
 - unstructured correlation matrix 471
- repeated-measures models 41
- research funding and morbidity study, transforming *x* and *y* variables 72–9
- residual 4
 - definitions of 50–2
- residual analyses
 - multiple linear regression 125–6
 - multiple Poisson regression 423–7
 - simple linear regression 66–70

Cambridge University Press

978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Second Edition

William D. Dupont

Index

[More information](#)

- residual values, multiple linear regression 99
- response (dependent) variables 39–40, 97
- response feature analysis (two-stage analysis)
 - nature of response variables 40, 41
 - repeated-measures analysis of variance 459–69, 475–6
- response variables (multiple responses per patient)
 - continuous 40, 41
 - dichotomous 40, 41
 - selection of regression model 40, 41
- response variables (one response per patient)
 - categorical 40–1
 - continuous 40
 - dichotomous 40
 - fixed-effects models 40–1
 - rates 40, 41
 - selection of regression model 40–1
 - survival data 40, 41
- restricted cubic spline models 71
 - hazard regression analysis 320–3
 - multiple linear regression 134–54
 - use in logistic regression 265–78
- right censored data, survival analysis 289–90
- root MSE (s) 53–5, 100
- sample correlation coefficient 47
- sample covariance 45–7
- sample mean ($\bar{x}$) 4, 20–2
- sample size 28
- sample standard deviation (s) 5, 20–2
- sample statistics 24
- sample variance (s^2) 4–5
- sampling bias 23
- scatter plot 6–7
- score tests 166, 167–8
- semi-nonparametric model 309
- sigmoidal family of regression curves 159–61
- significance level 26–30
- simple hazard regression models 318–20, 324–5 *see also* hazard regression analysis
- simple linear regression 45–92
 - 95% confidence interval for $y[x] = \alpha + \beta x$ evaluated at x 55–6
 - 95% prediction interval for response of a new patient 56–7
 - accuracy of linear regression estimates 53–5
 - analysis of transformed data 74–9
 - bandwidth (lowess regression) 64
 - comparison of slope estimates 82–7
 - conditional expectation 48–9
 - contrast with Poisson and logistic regression 379
 - correcting for non-linearity 71–2
 - definitions of residual 50–2
 - density-distribution sunflower plots 87–92
 - error 49–50
 - error standard deviation 49–50
 - error variance 49
 - examples (ethylene glycol poisoning) 45–7, 54–66, 69–70
 - examples (Framingham Heart Study) 66–9, 81–92
 - examples (research funding and morbidity study) 72–9
 - fitting the linear regression model 50–2
 - heteroscedastic error terms 50
 - homoscedastic error terms 50
 - influence of specific observations 66–7
 - jackknife residual 68
 - least squares estimate 51
 - leverage of an independent variable 67–8
 - linear regression line 51
 - logarithmic transformation of variables 71
 - lowess regression 64–6
 - mean squared error (MSE) (s^2) 53–5
 - model 49–50
 - model sum of squares (MSS) 54
 - origin of the term *linear regression* 52
 - population correlation coefficient 48
 - population covariance 47
 - residual analyses 66–70
 - restricted cubic spline regression model 71
 - root MSE (s) 53–5
 - sample correlation coefficient 47
 - sample covariance 45–7
 - square root transformation of variables 71
 - stabilizing the variance 70–1
 - standardized residual 68, 69
 - studentized residual 68, 69–70
 - testing the equality of regression slopes 79–87
 - total sum of squares (TSS) 54
 - transforming the x and y variables 70–4
 - using Stata 57–64
- simple logistic regression 159–98
 - 95% confidence interval for odds ratio 191–2
 - 95% confidence interval for odds ratio associated with a unit rise in x 175–6
 - 95% confidence interval for $\pi[x]$ 176–7, 178–81
 - analysis of case-control data 195–6
 - APACHE score and mortality in patients with sepsis 159, 160
 - Bernoulli distribution 163
 - binomial distribution 162–3
 - contrast between logistic and linear regression 164
 - contrast with simple linear and Poisson regression 379
 - exact $100(1 - \alpha)\%$ confidence intervals for proportions 177–8
 - examples (Ibuprofen in Sepsis Study) 159, 160, 170–6, 178–87
 - examples (Ille-et-Vilaine study of esophageal cancer risk) 187–90
 - generalized linear model 163–4
 - likelihood function 164–6
 - likelihood ratio tests 166–7
 - linear predictor of the model 163
 - link function of the model 164

Cambridge University Press

978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Second Edition

William D. Dupont

Index

[More information](#)

- simple logistic regression (*cont.*)
 - log odds of death under a logistic probability function 161
 - logistic probability function 159–61
 - logistic regression with grouped response data 176, 181–7
 - logit function 161
 - maximum likelihood estimation 164–6
 - model 163
 - models for 2×2 contingency tables 191–2
 - nuisance parameters 191
 - odds of death under a logistic probability function 161
 - odds ratios and the logistic regression model 174
 - probability distribution 162–3
 - quadratic approximations to the log likelihood ratio function 167
 - random component of the model 163
 - regressing disease against exposure 197–8
 - score tests 166, 167–8
 - sigmoidal family of regression curves 159–61
 - simple 2×2 case–control study example 187–90
 - Stata data file creation 192–4
 - statistical tests and confidence intervals 166–70
 - using Stata 171–4
 - variance of maximum likelihood parameter estimates 165–6
 - Wald confidence interval 169
 - Wald tests 166, 167, 168
- simple Poisson regression *see* Poisson regression
- simple proportional hazards model 309
- single imputation 259
- square root transformation of variables 71
- standard deviation (σ) 20–2, 24–5
- standard error 25
- standardized deviance residual, multiple Poisson regression 424–7
- standardized normal distribution 29–30
- standardized residual 68, 69, 125
- Stata statistical software package 7–22
 - abbreviations of commands 13
 - analysis of case–control data 195–6
 - analysis of dose of alcohol and esophageal cancer risk 219–20
 - analysis of multiple 2×2 tables 214–16
 - analyzing transformed data 74–9
 - area under the response curve 469
 - automatic methods of model selection 119–24
 - box plot 20, 21, 22
 - calculating relative risk from incidence data 374–6
 - capitalization in variables and commands 13
 - case–sensitivity of commands and variables 13
 - categorical variables 216–17
 - collinearity problem 125
 - command abbreviations 13
 - command execution 15–20
 - command prefix 13
 - command punctuation 12
 - command qualifiers 12–13
 - command syntax 12–13
 - Command window 8
 - commands list *see* Appendix B
 - comparing regression slope estimates 82–7
 - converting survival data to person–time format 383–92
 - creating histograms 9–12, 16–20
 - customizing graphs 14–15
 - data file creation 192–4
 - data imputation program 259–60
 - defaults and schemes 14–15
 - density–distribution sunflower plots 88–92
 - descriptive statistics display 20–2
 - do files 15, 16
 - dot plot 20, 21, 22
 - downloading data sets 8
 - error messages 14
 - executing a command 8
 - exploratory analysis of repeated–measures data 453–9
 - fitting a model with interaction 234–8
 - fitting a multiplicative model 227–31
 - generalized estimating equations (GEE) analysis 476–81
 - goodness-of-fit tests (multiple logistic regression) 248–57
 - Graph window 8
 - graphics editor 15
 - help facility 13–14
 - histogram generation examples 9–12, 16–20
 - Hosmer–Lemeshow test of goodness-of-fit 249, 255
 - independent t tests 36–8
 - influence analysis 129–33
 - interactive help facility 13–14
 - Internet sources of Stata programs 13–14
 - Kaplan–Meier survival functions 298–305
 - log files 14, 15
 - logistic regression with grouped data 181–7
 - logistic regression with restricted cubic splines 271–8
 - log-rank test 302–5
 - lowess regression curve 64–6
 - Mantel–Haenszel odds ratio 206–9
 - median 20–2
 - modeling missing data 263–5
 - modeling time-dependent covariates 362–70
 - multiple linear regression modeling 114–19
 - multiple Poisson regression 411–23
 - non-linear models with restricted cubic splines 142–54

Cambridge University Press

978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data, Second Edition

William D. Dupont

Index

[More information](#)

- notation for point-and-click commands 16–20
- odds ratio associated with a unit rise in x 175–6
- one-way analysis of variance 439–46
- paired t test 31–4
- Pearson chi-squared test of goodness-of-fit 249, 255
- point-and-click commands 15–20
- power calculations 28
- proportional hazards assumption evaluation 357–9
- proportional hazards regression analysis 310–11, 331–48
- pull-down menus 15–20
- residual analysis (multiple linear regression) 129–33
- residual analysis (multiple logistic regression) 248, 249–57
- residual analysis (Poisson regression) 424–7
- response feature analysis 463–8, 469
- Results window 8
- Review window 8
- sample mean 20–2
- sample size calculations 28
- scatterplot matrix graphs 105–7
- schemes 14–15
- simple linear regression 57–64
- simple logistic regression 171–4
- simple Poisson regression 379–81, 395–7
- standard deviation 20–2
- stratified proportional hazards models 349
- studentized residual analysis 69–70
- survival analysis with ragged study entry 351–4
- user contributed software 438–9
- Variables window 8
- variance 20–2
- statistical power 27–8
- statistics of a sample 24
 - expected value 25
 - unbiased estimate of a parameter 25
- step-functions, hazard regression analysis 359–61
- stratified proportional hazards models 348–9
- Student's t distribution 29–30
- studentized residual 68, 69–70, 125–6
- study design
 - power calculations 27–8
 - sample size 28
- sunflower plots (density-distribution) 87–92
- SUPPORT study of hospitalized patients
 - multiple linear regression examples 138–54
 - multiple logistic regression examples 267–78
- survival analysis 287–312
 - 95% confidence intervals for survival functions 293–4
 - censoring and bias 296
 - cumulative morbidity curve 289, 291, 295
 - cumulative mortality function 287–8, 295
 - disease-free survival curve 288–9, 291
 - examples (genetic risk of recurrent intracerebral hemorrhage) 291–5, 298–305, 310–11
 - examples (Ibuprofen in Sepsis Study) 288–9
 - hazard functions 306–7
 - hazard ratios 307–9
 - Kaplan–Meier cumulative mortality function 290–1
 - Kaplan–Meier survival curves 290–6, 298–305
 - Kaplan–Meier survival function 290–1
 - life tables 291
 - log-rank test 296–8, 299, 302–5
 - log-rank test for multiple patient groups 305
 - Mantel–Haenszel test for survival data (log-rank test) 296–8, 299, 302–5
 - patients lost to follow-up 289–90
 - Poisson regression 381–92
 - product limit survival function 290–1
 - proportional hazards 306–7
 - proportional hazards regression analysis 309–11
 - relative risks 307–9
 - right censored data 289–90
 - simple proportional hazards model 309
 - survival function 287–8
 - tied failure times 311–12
 - with ragged study entry 349–54
- survival data, regression models 40, 41
- t distributions 29–30
- t tests
 - independent t test using a pooled standard error estimate 34–5
 - independent t test using separate standard error estimates 35–6
 - independent t tests using Stata 36–8
 - paired t test 30–4
- target population 22–39
- target population parameters 24–5
- tied failure times, survival analysis 311–12
- time-dependent covariates, hazard regression analysis 359–70
- total sum of squares (TSS) 54, 99
- transforming the x and y variables (simple linear regression) 70–4
- two-sided (two-tailed) P -value 26
- two-stage analysis *see* response feature analysis
- two-way analysis of variance 446–7
- Type I error 27
- Type II error 27
- unbiased estimate of a parameter 25
- unbiased sample 23
- unstructured correlation matrix 471

Cambridge University Press
978-0-521-61480-1 - Statistical Modeling for Biomedical Researchers: A Simple Introduction to the
Analysis of Complex Data, Second Edition
William D. Dupont
Index
[More information](#)

- variables
 - independent variable (covariate) 39
 - response (dependent) variables 39–40
- variance (σ^2) 24
- Wald confidence interval 169, 310
- Wald tests 166, 167, 168, 310
- Wilcoxon–Mann–Whitney rank-sum test 435, 438
- Wilcoxon signed-rank test 435
- Woolf’s method (95% confidence interval for odds ratio) 189
- z distribution 29–30