
Index

Note: Locators followed by ‘*f*’ denote figure in the text

A

Acceptance region, 30
Active database, 2, 16, 19
Actuarial estimate of survivor function, 127
adjust command, 54, 180
Alternative hypothesis (H_a), 29
Analysis of variance (ANOVA), 4, 5*f*, 44
 data structure, 62
 with fixed effects, 64–65
 linear model, 63–64
 null hypotheses in, 64
 one-way, 78, 79*f*
 overview, 61
 planned comparisons
 comparison of expected values, 68–69
 linear contrast, 69–70
 programming for, 65–67
 random effects, 72–74
 example, 75–78
 measures related to, 74–75
 sample size, 78–79
 statistical power, 78–79
 unplanned comparisons, 70–72
Analytical epidemiology, 82
anova command, 66, 68, 70
Archives, Stata, 7
Asterisk (*) symbol, 14

B

Bar chart, 25–28, 27*f*
Bayesian approach, 29
Binary logistic regression model, 89
binreg command, 93
bmig command, 62–63

Body mass index (bmi), 14
 graph box for, 25*f*
 histogram for, 26*f*
 mean by sex, 27*f*
 mean with one standard deviation by sex, 28*f*
Bonferroni’s method, 70–71
Bootstrapping, Stata, 49
Box Plot option, 23
 window, 24*f*

C

capture command, 169
Case-control study, 84–86
Cases (diseased), study, 84
Censoring information, survival analysis,
 124–125
Centering, linear regression model, 47–49
Classical statistical methods, 43, 49
Classical statistical tests, 29
clear command, 7
Coefficient of determination (R^2), 44–45
Coef option, 93
Cohort study, 82–84
collapse command, 28
Combined RR, 82
Commands, Stata, 11–12. *See also specific commands*
 drop, 16
 generate, 14–15
 label, 16–17
 list, 12
 recode, 15
 replace, 16
 summarize, 17–18
 tabulate, 20–21

Complete model, 52
 Confidence intervals for testing null hypothesis, 35
 Confounding effect, 100–101
 Confounding variables, 82
 Controls (nondiseased), study, 84
 Correlated data
 in cross-sectional studies, 149
 mixed models, 154–156
 in sampling design, 163–166
 random intercept, 156–157
 linear mixed model with, 156–157
 mixed/gllamm commands with, 157–161
 regression models with, 149–154
 Covariance, 74–75
 Cox proportional hazards model, 143–145
 interaction terms in, 144
 stratified, 146–148
 survival function estimation, 146
 Crude HR, estimating, 144
 Crude OR, 95
 .csv file, 7
 Cumulative hazard function
 KM method, 135–136
 Nelson–Aalen method, 136, 136*f*

D

Database
 active, 2, 16, 19
 alternative, 94–95
 constructing user-friendly, 8
 in Stata, 7, 91, 94, 104
 statistics from, 63
 to summarize variables in, 17
 Data Editor, 7–10
 assigning labels to variable, 10*f*
 assigning value label, 9*f*
 icon, 2
 variable name change, 9*f*
 window, 8*f*
 Data file, reading, 6
 Data structure, 62
 Delimiters, 171–172
 Dependent variable, 41
describe command, 6
 Descriptive epidemiology, 81
detail command, 18
 Deviance (D), 96
 do-file Editor, 19
 database with, 176
 extension, 111

local command in, 177
 to open new, 167
 tool, 167–168
 window, 19*f*
drop command, 16, 169

E

Effect modification, 101–102
ef option, 95
 Epidemiology, 81
 analytical, 82
 descriptive, 81
 Equivalent process, 97
 Error sum of squares, 44
 Event occurrence in survival analysis, 125*f*
 Expected values
 planned comparisons
 comparison of, 68–69
 linear contrast, 69–70
 unplanned comparisons, 70–72
 Explanatory variable, 41
 Exposed group, 82

F

fam(bin) command, 92, 94
fam(bin total) command, 94
 Files, types of, 7
 Fisher transformation, 57
 Fixed effects
 ANOVA with, 64–65
 example for, 62–63
 linear model with, 63–64
foreach command, 176
forvalues command, 178
 Framingham Study, 115–116
Ftail command, 54

G

generate/gen command, 14–15
gllamm command, 157–161
glm command, 92, 94, 97–99, 179
 Glycohemoglobin
 groups, 106
 likelihood of, 108
 Graph box for bmi, 25*f*
 Graphics button, 3
 Graphics option, 23, 24*f*

H

- Hazard function ($h(t)$), survival analysis, 132–134, 135^f
 - cumulative, 135–136
 - by sex, 139^f
- Hazard ratio (HR), 143–144, 148
- Height in meters (heimt), 14
- Help ANOVA, 5^f
- Help, Stata, 3–4, 4^f
- Histogram, chart, 25
 - for bmi, 26^f
 - window, 26^f
- Hypothesis testing, 43–44
 - aim in, 30
 - for logistic regression model, 96
 - survival curves, 139

I

- Icons, 3^f
- Incomplete model, 52
- Independent variable, 41
- Indexing, 172–173
- insheet* procedure, 7
- Intercept of regression line, 41
- Intraclass correlation coefficient, 75, 157, 160
 - estimated, 78
 - with random intercept, 162

K

- Kaplan–Meier (KM) method, 127–128
 - cumulative hazard function, 135–136
 - survival function using, 133^f
- kwallis* command, 35

L

- Label box, 8
- label* command, 16–17
- label define* command, 16
- label value* command, 16
- Least squares method, 43
- Likelihood function, 90–92
 - for Poisson regression model, 115
 - relative comparison of, 96
- Likelihood ratio test, 96, 99, 109
 - Poisson regression model using, 117
- lincom* command, 99, 120
- Linear contrast, 69–70

- Linear mixed model with random intercept, 156–157
- Linear model
 - with fixed effects, 63–64
 - with random effects, 72–73
- Linear regression model, 41, 46
 - bootstrapping, 49
 - with categorical predictor, 65^f
 - centering, 47–49
 - coefficient of determination, 44–45
 - considerations for assumptions of, 59–60
 - hypothesis testing, 43–44
 - MLRM, 50–52
 - model assumptions, 42–43
 - parameter estimation, 43
 - partial hypothesis, 52–54
 - Pearson correlation coefficient, 45–46
 - polynomial, 55–57, 57^f
 - prediction, 54–55
 - running model, 47
 - sample size, 57–59
 - scatter plot, 46, 46^f
 - statistical power, 57–59
- link(log)* command, 102, 165
- list* command, 12
- local* command, 180
 - application for prevalence estimation, 179–182
- Local macros, 174
- Log-binomial regression model, 165
- Log files, 170
- logistic* command, 93
- Logistic regression model, 89–90
 - alternative database, 94–95
 - binary, 89
 - estimating OR, 95
 - extension of, 97–100
 - ordinal, 103, 105, 107–108
 - outcome, 103
 - overdispersion, 109
 - parameter estimation, 90–91
 - programming, 91–93
 - using binreg, 93
 - using glm, 92
 - using logistic, 93
 - using logit, 92
 - sample size, 109–111
 - significance tests, 96–97
 - likelihood ratio test, 96
 - Wald test, 96–97
 - statistical power, 109–111
- logit* command, 90, 92, 180

Log-rank test, 139–141
 evaluating survival curves, 140
 Stata command for, 141
loneway command, 77
 Longitudinal design in survival analysis,
 125, 126*f*
 Loops (foreach and forvalues), 176–179
ltable command, 131–133

M

Mann–Whitney test, 35
 Mantel–Haenszel method, 82, 84
mata command, 111
 Mata, matrix programming language, 111
 Mathematical and logical operators, 12–14
matrix command, 180
 application for prevalence estimation,
 179–182
 Maximum-likelihood estimation (MLE), 90–91
 Mean squares (MS), 44
 Median survival time and percentiles, 136–137
mixed command, 157–161, 159
 with random intercept and slope, 161–163
 Mixed models, correlated data, 154–156
 in sampling design, 163–166
 Multinomial logistic regression model, 103
 Multiple linear regression model (MLRM),
 50–52
 Multivariable logistic regression model, 97–99

N

Nelson–Aalen method, 136, 136*f*
noisily command, 168
 Nominal and ordinal outcomes, 103–109
 Nonexposed group, 82
 Nonparametric tests for unpaired groups,
 35–36
 Normality test, 31
 Normal option, 25
 Null hypotheses, 29, 53, 71
 in ANOVA, 64
 confidence intervals for testing, 35
 log-rank test, 140
 sample size for assessing, 86, 87*f*, 88*f*

O

Odds ratio (OR), 84, 86
 adjusted, 100–101
 crude, 95

 estimating, 95
 interpretation of, 104, 106, 108
 One-way ANOVA, 78, 79*f*
oneway command, 66–67, 70
 Open icon, 6
 Operators, mathematical and logical, 12–14
 Ordinal logistic regression model, 103,
 105, 107
 syntax in Stata to run, 108
 Overdispersion
 logistic regression model, 109
 Poisson regression model, 121

P

Partial hypothesis, 52–54
 Partial residuals, 145
 Pearson correlation coefficient, 45–46
 Poisson probability distribution, 113
 Poisson regression model
 advantage of, 113
 assessing interaction terms, 117–121
 defined, 113–114
 example, 115–116
 incidence
 age-specific, 118–119
 of cardiovascular disease, 121
 likelihood function, 115
 likelihood ratio test, 117
 overdispersion, 121
 parameter estimation, 115
 programming, 116–117
 RR, 114–116
 Polynomial linear regression model,
 55–57, 57*f*
 Polytomous logistic regression model, 103
 Power and sample size analysis, 36, 37*f*, 39*f*, 57,
 58*f*, 78–79, 86–88
 Predictor variable, 41
 Prevalence ratio (PR), 102–103
 Probability of survival, 126
program command, 168–170
 Programming
 for ANOVA, 65–67
 for logistic regression model, 91–93
 using *binreg*, 93
 using *glm*, 92
 using *logistic*, 93
 using *logit*, 92
 Poisson regression model, 116–117
 of *S(t)*, 128–132
 Stata, 137, 182

Proportional hazards (PH) assumption,
138–139
 assessment of, 139, 145–146
Proportional odds assumption, 106–107
proportion command, 179–180
Puerto Rico, sampling design in, 150*f*
P-value, statistical test, 30, 33–34

Q

Quadratic polynomial model, 55
quietly command, 168

R

Random effects model, 72–74
 example of, 75–78
 linear model with, 72–73
 measures related to, 74–75
 scheme, 73
Random intercept, 156–157
 intraclass correlation coefficients with, 162
 linear mixed model with, 156–157
 mixed/gllamm commands with, 157–161
recode command, 15
Recruitment time, survival analysis, 125
Regression model
 with correlated data, 149–154
 linear. *See* Linear regression model
 log-binomial, 165
 logistic. *See* Logistic regression model
 MLRM, 50–52
 Poisson. *See* Poisson regression model
 polynomial linear, 55–57
regress/reg command, 47, 51, 54, 65
Rejection region, 30
Relative risk (RR), 82, 84, 113
 calculating for age group, 119–120
 Poisson regression model, 114–116
replace command, 7, 16
reshape command, 76
Residual(s)
 distribution, 59*f*
 partial, 145
 Schoenfeld, 145
 sum of squares, 44, 52
Response variable, 42
return command, 175
return list command, 175
Right censoring, 125
rvpplot command, 60

S

Sampling design
 mixed models in, 163–166
 in Puerto Rico, 150*f*
Scalars, 175–176
Scatter plot, 46, 46*f*
Scheffé's method, 70–72
Schoenfeld residuals, 145
sdtest command, 34
Second-order polynomial model, 55
Shapiro–Wilk test, 31
Simple linear regression model (SLRM),
 41, 46
 bootstrapping, 49
 centering, 47–49
 coefficient of determination, 44–45
 hypothesis testing, 43–44
 MLRM, 50–52
 model assumptions, 42–43
 parameter estimation, 43
 partial hypothesis, 52–54
 Pearson correlation coefficient,
 45–46
 polynomial linear regression model,
 55–57, 57*f*
 prediction, 54–55
 running model, 47
 sample size, 57–59
 scatter plot, 46, 46*f*
 statistical power, 57–59
Single RR, 82
Start date, survival analysis, 124
Stata 14 window, 2, 2*f*
Stata, computer program, 1
 bootstrapping in, 49
 command for log-rank test, 141
 commands. *See* Commands, Stata
 database in, 91, 94, 104
 structure of, 107
Data Editor, 7–10
delimiters, 171–172
do-file Editor, 19, 19*f*
entering, 2
help, 3–4, 4*f*
indexing, 172–173
insheet procedure, 7
for log-rank test, 142
median survival time and percentiles,
 136–137
reading data file, 6
to run survival analysis in, 130

Stata, computer program (*Continued*)

- syntax in, 104
 - overdispersion, 109
 - to run ordinal logistic model, 108
- taskbar, 2–3
- working directories, 4–6

Statistical power, 36–38, 37*f*

Statistics and graphs, descriptive, 19–20, 20*f*

Statistics option, 3, 36, 36*f*

stcox command, 144–145

Stratified Cox proportional hazards model, 146–148

sts graph command, 137

Student's *t*-test statistics, 30, 175

- for independent samples, 33–34

summarizelsum command, 17–18

Summary statistics, 19

- creating table to display, 20*f*
- window for displaying table of, 20*f*

Sum of squares (SS), 44, 64

Survival analysis, 123

- actuarial estimate, 127
- applications of, 148
- censoring information, 124–125
- hazard function, 132–134
- longitudinal design, 125, 126*f*
- median survival time and percentiles, 136–137
- objective in, 123, 126
- occurrence of event in, 125*f*
- probability of survival, 126
- programming of $S(t)$, 128–132
- proportional hazards assumption, 138–139
 - assessment of, 145–146
- relationship between $S(t)$ and $h(t)$, 134–135
- significance assessment, 139–142
 - log-rank test, 140–141
 - Tarone–Ware test, 142
 - WGB test, 141–142
- start date, 124
- in Stata, 130
- study design components, 126–127
- study time in, 124
- survival curves
 - comparison, 137–138
 - by sex, 147*f*
- use of, 124

Survival function ($S(t)$), 126

- estimation, 146
- KM method, 127–128, 133*f*
- programming, 128–132
- relationship between $h(t)$ and, 134–135
- by sex, 138*f*

swilk command, 31

T

tabulate/tab command, 20–21

Tarone–Ware test, 142

Taskbar, Stata, 2–3

- and icons, 3*f*

test command, 70

Test statistics, 30, 68–69, 71

Time to event analysis, 123

Total sum of squares, 44

t-probability distribution, 33, 43

trace command, 170–171

ttest command, 33

twoway command, 28, 46

U

Underdispersion, 109

Unpaired groups, nonparametric tests for, 35–36

User-friendly database, constructing, 8

V

Value Label option, 8

Variables Manager window, 8

Variance

- ANOVA. *See* Analysis of variance (ANOVA)
- homogeneity, 31–33
- of residuals, 42

W

Wald test, 96–97, 99

Weight in kilograms (weikg), 14

Wide format, 76

Wilcoxon–Gehan–Breslow (WGB) test, 139, 141–142

Wilcoxon rank sum test, 35

Working directories, Stata, 4–6