Index

• A •

additive error, linearity in parameters and, 94–95
aggregation of data, 10
aggregation level used in measuring variables, 11
alternative hypothesis, 54
American Community Survey (ACS), 69
AR(q) process, detecting autocorrelation of (Breusch-Godfrey test), 221–223
AR(1) process, detecting autocorrelation of, 219–222
assumptions
  ceteris paribus, 8, 9, 64, 318
  CLRM, 10, 93–94, 97
    no autocorrelation, 209–210
  linear in parameters, 94–95
  normality, 111–114
  relying on sensible, 8–9
  stationarity, 213–214
  of this book, 2
asymptotic property, consistency as, 44
autocorrelation (serial correlation)
  AR(1) process, 219–222
  AR(q) process (Breusch-Godfrey test), 221–223
  autoregressive errors, 212–214
  CLRM assumption of no autocorrelation, 209–210
  d statistic, 220–221
  described, 99–101
  Durbin-Watson test, 219–221
  in dynamic models, 270
  examining patterns of, 209–212
  misspecification and, 212
  negative, 100, 210–211
  positive, 100, 210–211
  remedying harmful, 223–228
  residual analysis, 214–215
  stationarity assumption, 213–214
autoregression, first-order, 213
autoregressive errors, 212–214
autoregressive models, 268–270
auxiliary regressions, variance inflation factor (VIF) and, 183–185

• B •

base group, 154
bell-shaped probability density function, 44
beta coefficients, 87, 88
bias
  heterogeneity, 292
  omitted variable, 148–149
binomial probability distribution for the dependent variable, 232–233
bivariate probability density, 28–29
bizarre results, looking the other way when you see, 319
BLUE (best linear unbiased estimators), 101, 108
Break button, in STATA, 17
Breusch-Godfrey (BG) test, 221–223, 270
Breusch-Pagan (BP) test, 196–197
Bureau of Labor Statistics, 71

• C •

causation, statistical results cannot prove, 63
CDF (cumulative density function)
  for continuous random variables, 27
  for discrete random variables, 26–27
  probit model, 241–245
  shape of, 27
censored dependent variables, 254–255
census, 69
Census Bureau, 71
central limit theorem (CLT), 48–49, 112
ceteris paribus assumption, 8, 9, 64, 318

COPYRIGHTED MATERIAL

chi-squared distribution
  deriving from the random error, 115–119
  description and uses, 49–51
  table, 323–325
Chow test
  to determine structural stability, 151–152
  dummy variable approach to, 169–172
chronological order, time-series data, 70
classical linear regression model (CLRM)
  assumptions, 10, 93–94, 97
  autocorrelation (serial correlation), 99–101, 209–210
  error term with constant variance, 98
  error term with a zero conditional mean, 97–98
  Gauss-Markov theorem, 101, 103, 105, 106, 108, 109
  random sampling and variability, 95
  stationarity assumption, 213–214
  zero correlation of error observations, 99–101
Cobb-Douglas production functions, 141
Cochrane-Orcutt (CO) transformation, 223, 226
coefficient estimates, sensitivity to changes in specification, 179
coefficients
  logit, 251–252
  OLS
    best linear unbiased estimators (BLUE), 108
    deriving a chi-squared distribution from the random error, 115–119
    error term and sampling distribution of, 113–114
    expected value of, 102–105
    variance of, 105–108
    WLS coefficients and, 205
  probit, 249–251
  regression. *See* regression coefficients
  slope
    regression analysis, 84, 85
    relationship between *t* and F, 127
    t-test of significance for, 127
  standard errors of the, chi-squared distribution, 117–118
  variables with statistically insignificant, 313
collinearity
  high, 97
  perfect, 96–97
common sense, failing to use your, 315
composite error, defined, 295
composite error term, 298–299
composite index variable, 188
compounded growth rate, 144–145
conditional mean operator, 64
conditional probabilities, 29–30, 233
confidence interval approach, hypotheses testing, 54, 56
confidence intervals
  prediction, 133–134
  regression coefficients, 121–122
consistent estimators, 44, 108
constant
  expected value of, 32
  population regression function (PRF), 64
  variance of, 34
constant variance (homoskedasticity)
  Chow test
    to determine structural stability, 151–152
    dummy variable approach to, 169–172
  defined, 192
  error term in classical linear regression model (CLRM), 98, 106
  Goldfeld-Quandt (GQ) test, 199
  heteroskedasticity distinguished from, 191–194
  normality assumption, 112
  variance of the estimated slope coefficient, 193
  weighted least squares (WLS) technique, 203–205
constant-elasticity demand curves, 141
consumer price index (CPI), as composite index variable, 188
consumption spending, linear-log model, 145–146
contemporaneous variables, 267
continuous probability density function, 44, 47
continuous random variables
  cumulative density function (CDF), 27
  defined, 22
  expected value, 31
  probability density function (PDF), 24–25
  variance, 33

core variables, 152
correlation
  described, 42
  introduction to, 35, 37, 38
  random variables, 37
  serial. *See* autocorrelation
  spurious, 63
  spurious time-series data, 273–274
correlation coefficients
  pairwise, 181–182
  zero (or very close to zero), 37–38
correlation matrix, 181–182
covariance
  calculating, 34, 35
  described, 41
  positive or negative, 36–38
CPI (consumer price index), as composite index variable, 188
critical *t* value, 122
cross-section analysis, pooled, 281–290
cross-sectional data
  defined, 10
  pooled, 72–73
  regression analysis, 68–69
cubic functions, 139–140
cumulative density function (CDF)
  for continuous random variables, 27
  for discrete random variables, 26–27
  probit model, 241–245
  shape of, 27
Current Population Survey (CPS), 69

• D •

d statistic, 220–221, 271
data
  collecting and organizing, for regression analysis, 68–73
  cross-sectional, 68–69
  detailed description of, in research project, 311
  failing to familiarize yourself with the, 316–317
  panel (longitudinal), 71–72
  pooled cross-sectional, 72–73, 281–290
  time-series, 70–71
  trend-adjusted, 274–276
  types of, 10–11, 68
  values that appear legitimate but are actually censored, 317
data mining, 12
decomposing variance, 88–89
decomposition of differences to find policy effect, 288
degrees of freedom
  adjusting the goodness of fit, 91
  chi-squared distribution, 50, 116
  described, 41
  F distribution, 52–53
Department of Commerce, 71
Department of Energy, 71
dependent variables
  censored, 254–255
  decomposing variance, 88–89
  impact of omitted variable on, 148–149
  inverse function, 140
  limited
    censored, 254–255
    modifying regression analysis for, 256–264
    normality assumption, 113
    overview, 253
    Tobit model, 256–258
    truncated regression, 255–256, 258–262
  linear-log model, 145–147
  ordinary least squares (OLS) technique, 77
  PRF model, 65
descriptive statistics
  calculating parameters and estimators, 40–43
  determining whether an estimator is good, 43–44
  STATA software, 42–43
deseasonalizing time-series data, 278–280
deterministic linear relationship, perfect multicollinearity, 175
detrending time-series data, 274–276
difference-in-difference, 288–290
discrete dependent variables, 254
discrete random variables
  cumulative density function (CDF) for, 26–27
  defined, 22
  expected value, 31

discrete random variables *(continued)*
probability density function (PDF) for, 22–24
variance, 33
dispersion, measures of, 33
distributed lag model, 268–270
distribution of sample means, 48. *See also* chi-squared distribution; normal distribution; sampling distributions; *t*-distribution
dummy dependent variable models, 231. *See also* qualitative dependent variables
dummy variable (DV) regression, 294–295
dummy variables
defined, 12
finding average differences by using, 156–158
interacting quantitative variables and, 162–165
juggling multiple characteristics with, 155–156
need to create, 317
with only two possible characteristics, 154
pooled cross-sectional data, 283–285
Durbin's alternative statistic, 221–223
Durbin-Watson *d*-Statistic, table, 325
Durbin-Watson test, 219–221, 271
dynamic models
distributed lag model, 268
problems with, 268–270
testing and correcting for autocorrelation in, 270–271

• E •

EconLit website, 309
econometric models. *See* model(s)
The Econometric Society, 7
econometrics research project
conceptual or theoretical framework, 309
detailed description of data, 311
discussing the relevance and importance of the topic, 308
estimation method(s), 310–311
explaining your econometric model, 310
interpreting the reported results, 312–313
introducing the topic and posing the research question, 307–308
literature review, 308–309
summarizing the results, 313
tables and graphs, 312
economic journals, on the web, 309
economic theory
describing outcomes and making predictions with, 8
econometrics and, 8
failing to use your knowledge of, 315
Economics For Dummies (Flynn), 61, 138
economics journals, websites of, 309
education, FE model, 297–298
efficient estimators, 43
elasticity, constant, log-log model, 141
Engel curve, 145
error
composite, 295
forecast, applying to OLS predictions, 130
idiosyncratic, 295
linearity in parameters, and additive error, 94
random, regression analysis, 65, 67
type I, 58, 123
type II, 58, 123
error covariance matrix, 227
error messages, STATA, 16
error term
composite, 298–299
normality assumption, 111–114
zero conditional mean, 97–98
estimation
discussing in research project, 310–311
point, 54
in STATA, 19
estimators
calculating, 40–43
consistency, 44

defined, 40
desirable properties, 43–44
linearity, 44
OLS
heteroskedasticity and, 238
standard errors and the *t*-distribution, 119
standard normal distribution, 114–115
variance of, 107
expected value (or mean), 31–32
experiments
cumulative density function table, two-coin-toss experiment, 26–27
to estimate policy effects with pooled cross sections, 285–286
natural (quasi), 287–290
probability density function, 3-coin-toss experiment, 23–24
true, random assignment, 286
explained sum of squares (ESS), 90
exponential growth, log-linear model, 144
exponential time trend, 271

• F •

FD (first difference) transformation, 294
F-distribution (*F*-statistic)
Chow test, 151–152
derivation and basic properties, 52–53
Goldfeld-Quandt (GQ) test, 199–200
for joint significance, 168–169
joint significance for subsets of variables, 128–130
normality assumption, 124–126
table, 325–326
White test, 198
feasible generalized least squares (FGLS)
Cochrane-Orcutt and Prais-Winsten transformations, 223
random effects (RE) estimator, 299
weighted least squares (WLS) technique, 204, 205
Federal Reserve, 71
FGLS. *See* feasible generalized least squares
first difference (FD) transformation, 294
first-differenced equation, 187
first-differenced estimators, 187
first-differencing, 187–188
first-order autoregression, 213
fixed effects, individual, 292, 293, 295, 296, 302
fixed effects estimator (within estimator), 295–298
fixed effects (FE) model, 298–299
Flynn, Sean Masaki (*Economics For Dummies*), 61, 138
forecast error, applying to OLS predictions, 130, 131
forecasting. *See* prediction(s)
frequency of data
explained, 11
time-series data, 70
functions
cubic, 139–140
inverse, 140
quadratic, 138–139

• G •

Gauss-Markov theorem
BLUE (best linear unbiased estimators), 101, 108
CLRM assumptions, 101, 103, 105, 106, 108, 109
expected value of OLS coefficients, 102–105
linearity of OLS, 101–102
omitted variable bias, 148
proving, 101–108
summary of assumptions, 109
Geary test (run test), 215–218
general mathematical specification of your model, 64
generalized least squares (GLS) estimator, 204

generalized method of moments (GMM) technique, 76
GLS (generalized least squares) estimator, 204
Goldfeld-Quandt (GQ) test, 199–201
goodness of fit
  adjusting, in multiple regression, 90–91
  measuring, 88–90
Google Scholar, 309
graphical examination of residuals, 194–195
graphs, research project, 312
growth rate
  compounded, 144–145
  exponential time trend, 271
  instantaneous, 144

• H •

HAC (heteroskedasticity-autocorrelation-corrected) standard error. *See* serial correlation robust standard errors
Hausman test, 302–304
Heckman selection model, 260–264
heterogeneity bias, 292
heteroskedasticity
  Breusch-Pagan (BP) test, 196–197
  consequences of, 193–194
  correcting regression model for the presence of, 203–208
  detecting, with residual analysis, 194–195
  distinguishing between homoskedastic and heteroskedastic disturbances, 191–194
  Goldfeld-Quandt (GQ) test, 199–201
  Hausman test, 302
  linear probability model (LPM), 238–239
  Park test, 201–202
  robust standard errors (White-corrected standard errors), 205–208
  White test, 197–199
heteroskedasticity-autocorrelation-corrected (HAC) standard error. *See* serial correlation robust standard errors
heteroskedasticity-corrected standard errors (robust standard errors), 205–208
high collinearity, 97
high multicollinearity
  described, 175, 176
  expelling the problem variable(s), 188–190
  gathering additional data, 186
  resolving, 185–188
  zeroing in on, 178–180
homoskedasticity (constant variance). *See also* heteroskedasticity
  Chow test
    to determine structural stability, 151–152
    dummy variable approach to, 169–172
  defined, 192
  error term in classical linear regression model (CLRM), 98, 106
  Goldfeld-Quandt (GQ) test, 199
  heteroskedasticity distinguished from, 191–194
  normality assumption, 112
  variance of the estimated slope coefficient, 193
  weighted least squares (WLS) technique, 203–205
human capital functions, 143
hypotheses testing
  explained, 53
  performing a hypothesis test, 54–55
  probability distributions, 53–55

• I •

icons used in this book, 3
idiosyncratic error, 295
impact multiplier, 268
independence, statistical, checking for, 30
independent variables
  common time trend component, 178
  decomposing variance, 89
  described, 62
  imperfect linear relationships among, 96–97
  including irrelevant variables, 149

multicollinearity issues
first-differencing technique, 187–188
high multicollinearity, 178–180
pairwise correlation coefficient, 181–182
resolving multicollinearity issues, 185–186
variance inflation factor (VIF), 183–185
selecting, 147–149
that capture similar phenomena, 178–179
individual fixed effects, 292, 293, 295, 296, 302
inferences
foundational concepts, 39
making, with probability distributions, 53
inflexion point, 139
instantaneous rate of growth, 144
interacted econometric model, 162–166
interacting
quantitative and qualitative variables (interacted econometric model), 162–166
two (or more) qualitative characteristics, 165–168
interaction term, 162–166
intercept(s)
ordinary least squares (OLS) technique, 78–80
pooled cross-sectional data, 282–283
intercept term
Gauss-Markov theorem, 104
ignoring, in real-world scenarios, 85
OLS coefficients, 114
inverse function, 140
investments, calculating rates of return, 143–145
irrelevant variables, including, 149

• J •

joint probability
density, 28–29
likelihood function, 244–245
mathematical notation, 28–29
joint significance
F-test, 168–169
subsets of variables, 128–130
tests of, 123–124
justifying your model, 62

• L •

lagged variables
high multicollinearity, 178
in regression analysis, 267–268
least squares principle, defining and justifying, 76
likelihood function, 244–245
limited dependent variables
censored, 254–255
modifying regression analysis for, 256–264
normality assumption, 113
overview, 253
Tobit model, 256–258
truncated regression, 255–256, 258–262
linear estimators (coefficients), 108. *See also* coefficients
linear in parameters assumption, 94–95
linear probability model (LPM)
estimating with OLS, 232–237
heteroskedasticity, 238–239
problems, 237–240
unbounded predicted probabilities, 239–240
linear time trend, 271
linearity
of estimators as desirable property, 44
Gauss-Markov theorem, 101–102
in parameters, and additive error, 94
linear-log model, 145–147
literature review, 308–309
log transformation, OLS technique, 95
logistic units (logits), 246
logit coefficients, 251–252
logit model, 240, 242–244, 246
log-linear model, 143–145
log-log model, 141–143
longitudinal data (panel data)
defined, 11
dummy variable (DV) regression, 294–295
first difference (FD) transformation, 294
fixed effects (FE) model, 298–299
Hausman test, 302–304
increasing the efficiency of estimation with random effects, 298–301

longitudinal data (panel data) *(continued)*
pooled cross-sectional data compared to, 71–72
random effects (RE) estimator, 299–301
uniqueness of each individual unit, 291–294
long-run propensity, distributed lag model, 268
LPM (linear probability model)
estimating with OLS, 232–237
heteroskedasticity, 238–239
problems, 237–240
unbounded predicted probabilities, 239–240

• M •

macroeconomic theory, 8
marginal effects
Heckman selection model, 262
probit model, 249–252
Tobit model, 257
truncated regression, 259
marginal probabilities, 29
maximum likelihood (ML) estimation, 76, 95, 241, 244–246
mean (average), defined, 40
mean prediction
and forecast error, 130–131
variance of, 131–133
mean square error (MSE)
chi-squared distribution, 116, 118, 119
dropping a variable, 189
microeconomic theory, 8
misspecification. *See also* overspecification; specification issues
autocorrelation and, 212
checking for, 147–149
described, 97
mistakes in econometric analysis, ten common, 315–320
asking the wrong questions first, 316
assuming your results are robust, 320
being inflexible to real-world complications, 318
failing to familiarize yourself with the data, 316–317
failing to use common sense and knowledge of economic theory, 315
forgetting about economic significance, 320
ignoring the work and contributions of others, 316
looking the other way when you see bizarre results, 319
making it too complicated, 317–318
obsessing over measures of fit and statistical significance, 319
model(s). *See also specific models*
dummy dependent variable, 231. *See also* qualitative dependent variables
dynamic, 268–271
fixed effects (FE), 298–299
general mathematical specification of your, 64
Heckman selection, 260–264
justifying your, 62
linear-log, 145–147
logit, 240, 242–244, 246
log-linear, 143
log-log, 141–143
nonlinear, giving linearity to, 141–147
probit, 240–242, 246–252
specification, 62
Tobit, 256–258
MSE (mean square error)
chi-squared distribution, 116, 118, 119
dropping a variable, 189
multicollinearity
auxiliary regression and the variance inflation factor (VIF), 183–185
described, 175
distinguishing between types of, 175–180
high
described, 175, 176
expelling the problem variable(s), 188–190
gathering additional data, 186
resolving, 185–188
zeroing in on, 178–180

imperfect linear relationships among independent variables, 96
pairwise correlation coefficient, 181–182
rules of thumb for identifying, 180–181
variance of OLS estimators, 107
multiple regression
adjusting the goodness of fit in, 90–91
ordinary least squares (OLS) technique, 82

• N •

National Longitudinal Surveys (NLS), 72, 291
natural experiment (quasi experiment), 287–290
natural log transformation of the likelihood function, 245–246
negative autocorrelation, 210–211
Newey-West (NW) standard errors (serial correlation robust standard errors), 223, 225–228
NLS (National Longitudinal Surveys), 72, 291
nonlinear models, giving linearity to, 141–147
nonlinearity
identifying, 38
in parameters, 240
non-normality of the error term, 237–238
nonrandom sample selection, truncation, 255
nonsensical coefficient signs and magnitudes, 179, 185
normal distribution. *See also* normality assumption
density function of, 44–45
random variables with, 45
shorthand way of indicating, 45
standard
defined and explained, 46–47
OLS estimators, 114–115
table, 321
normality assumption
describing role of, 111–113
error term, 111–113
F-distribution, 124–126
homoskedasticity, 112
linear probability model (LPM), 237
mathematical notation, 112
sampling distribution of OLS coefficients, 113–114
null hypothesis, 54

• O •

observable variables, 292
OLS (ordinary least squares) assumptions, 93–94
OLS coefficients
best linear unbiased estimators (BLUE), 108
deriving a chi-squared distribution from the random error, 115–119
error term and sampling distribution of, 113–114
expected value of, 102–105
variance of, 105–108
WLS coefficients and, 205
OLS estimation
heterogeneity bias, 292–293
heteroskedasticity and, 193
OLS estimators
heteroskedasticity and, 238
standard errors and the *t*-distribution, 119
standard normal distribution, 114–115
variance of, 107
OLS standard errors, *t*-distribution and, 119–121
OLS (ordinary least squares) technique. *See also* OLS coefficients; OLS estimators
applying forecast error to OLS predictions, 130
desirable characteristics of, 76
as easier than alternative techniques, 76
estimates of the regression parameters, 78
estimating LPM with, 232–237
estimating the regression function and the residuals, 77
formulas necessary to produce optimal coefficient values, 79–80

OLS (ordinary least squares) technique *(continued)*
Gauss-Markov theorem, 101
heteroskedasticity, 99
interpreting regression coefficients, 84–88
least squares principle, 76
linearity in parameters, 95
measuring goodness of fit, 88–89
measuring proportion of variance with R-squared, 89–90
as most common technique, 10
normality assumption, 111–113
overview, 75
prediction confidence interval, 133–134
projecting time trends with, 271–273
reported *F*-statistic from, 124–126
for seasonal adjustments, 276–278
as sensible, 76
standardizing regression coefficients (beta coefficients), 86–88
omitted variable bias, 148–149
one-tailed hypothesis tests, 54
overall significance, tests of, 123–125
overspecification, 317–318. *See also* misspecification; specification issues
overspecified regression model, 149, 188

• *p* •

pairwise correlation coefficients, 181–182
panel data (longitudinal data)
defined, 11
dummy variable (DV) regression, 294–295
first difference (FD) transformation, 294
fixed effects (FE) model, 298–299
Hausman test, 302–304
increasing the efficiency of estimation with random effects, 298–301
pooled cross-sectional data compared to, 71–72
random effects (RE) estimator, 299–301
uniqueness of each individual unit, 291–294
Panel Study of Income Dynamics (PSID), 72, 291
parameters
calculating, 40–43
defined, 40
Park test, 201–202
partial-slope coefficients, log-log model, 143
perfect collinearity, 96–97
perfect multicollinearity, 175, 176, 178
point estimate
defined, 40
hypotheses testing, 54
policy effects, experiments to estimate, with pooled cross sections, 285–286
polynomials, 138
pooled cross-section analysis, 281–290
pooled cross-sectional data
described, 72–73
estimating policy effects with, 285–286
intercepts and/or slopes, 282–283
time dummy variables, 283–285
population covariance, 41, 42
population mean, 40–41
population parameters
calculating, 40
sampling distributions, 47, 53, 54
testing hypotheses, 53, 54
population regression function (PRF)
an example, 66–67
identifying, 63–67
linearity in parameters, 94
mean prediction and forecast error, 130–131
ordinary least squares technique, 77
stochastic, 65
population standard deviation, 41
population variance, 41
positive autocorrelation, 210–211
Prais-Winsten (PW) transformation, 223, 226
predicted probabilities, unbounded, 239–240
prediction(s)
applying forecast error to OLS predictions, 130
econometrics used in making, 8
mean and forecast error, 130–131
variance of mean, 131–133

prediction confidence interval, 133–134
probability
  conditional, 29–30, 233
  linear probability model (LPM)
    estimating with OLS, 232–237
    heteroskedasticity, 238–239
    problems, 237–240
    unbounded predicted probabilities, 239–240
  random variables and, 21–22
  unconditional, 29, 30
probability density function (PDF)
  for continuous random variables, 24–25
  defined, 22–25
  for discrete random variables, 22–24
  probit model, 242
probability distribution (density), sampling distribution, 47–48
probability units (probits), 246
probit coefficients, 249–251
probit model, 240–242, 246–252
proxies, 150, 318
pseudo R-squared measure of fit, 248–249
PSID (Panel Study of Income Dynamics), 72, 291
p-values, levels of significance and, 58, 123

• Q •

quadratic functions, 138–139
qualitative data
  defined, 10
  combining quantitative and, in regression model, 159
  quantifying, 153–156
qualitative dependent variables
  described, 231
  linear probability model (LPM). *See* linear probability model
  logit function. *See* logit coefficients; logit model
  maximum likelihood (ML) estimation, 241, 244–246
  probit function. *See* probit coefficients; probit model
qualitative variables
  incorporating quantitative information and, 12, 19
  interacting quantitative variables and, 162
  quantifying qualitative information, 153
quantifying qualitative information, 153–156
quantitative information, 12
quasi (natural) experiments, 287–290
quasi-differencing, 223

• R •

random assignment, true experiment, 286
random effects (RE) estimator, 299–301
random effects (RE) model, increasing the efficiency of estimation with, 298–301
random error
  deriving a chi-squared distribution from, 115–119
  regression analysis, 65, 67
random sampling, variability and, 95
random variables
  bivariate or joint probability density, 28–29
  continuous
    cumulative density function (CDF), 27
    defined, 22
    expected value, 31
    probability density function (PDF), 24–25
    variance, 33
  correlation between, 37
  covariance and correlation, 35–37
  discrete
    cumulative density function (CDF) for, 26–27
    defined, 22
    expected value, 31
    probability density function (PDF) for, 22–24
    variance, 33
  expected value, 31–32
  introduction to, 21–22
  nonlinear relationship, 38
  normal distribution, 44–46

random variables *(continued)*
OLS estimators, 114–115
variance and standard deviation, 33–34
rates of return, calculating, log-linear model, 143–145
reference group, 154
reference period, 283
regression analysis. *See also* classical linear regression model (CLRM); ordinary least squares (OLS) technique
causality and, 63
collecting and organizing data for, 68–73
combining quantitative and qualitative data, 159–161
contemporaneous and lagged variables in, 267–276
correcting for the presence of heteroskedasticity, 203–207
cross-sectional data, 68–69
decomposing variance, 88–89
modifying for limited dependent variables, 256–264
panel (longitudinal) data, 71–72
pooled cross-sectional data, 72–73
population regression function (PRF), 63–67
time-series data, 70–71
truncated regression, 258–262
regression coefficients
calculation of estimated, 80–81
interpreting, 84–88
linear-log model, 146–147
omitting relevant variables, 147–149
ordinary least squares (OLS) technique, 79–80
standardizing, 86–88
statistically significant, 119–123
confidence interval approach, 121–122
test of significance approach, 122–123
regression function, estimating, 77
regression line, ordinary least squares (OLS) technique, 78, 79
regression models. *See* regression analysis
regression parameters, obtaining estimates of the, 78
repeated sampling property, 179
reported *F*-statistic from OLS, 124–126
research project. *See* econometrics research project
RESET (regression specification error test), 150–151
residual sum of squares (RSS), 90, 91, 125, 151, 169, 199
residuals (residual analysis)
autocorrelation and, 214–215
detecting heteroskedasticity with, 194–195
run test (Geary test), 215–218
restricted dependent variables, 97
RMSE (root mean squared error), 116
robust standard errors
heteroskedasticity, 205–208
serial correlation, 223, 225–228
robustness/sensitivity analysis, 152
root mean squared error (RMSE), 116
R-squared (R^2)
adjusting the goodness of fit, 90–91
detrended data, 274
measuring proportion of variance with, 89–90
obsessing over, 319
reasons not to use as only measure of regression's quality, 92
reported *F*-statistic from OLS, 124–126
RSS (residual sum of squares), 90, 91, 125, 151, 169, 199
Rumsey, Deborah (*Statistics For Dummies*), 39
run test (Geary test), 215–218

• S •

sample correlation coefficient, 42
sample covariance, 41, 42
sample means
distribution of, 48
population mean, 40, 41
sample regression function (SRF)
calculation of estimated regression coefficients, 81–82
expected value of OLS coefficients, 102–105

mean prediction and forecast error, 131
ordinary least squares (OLS) technique, 76–78
sample regression line (SRL), 78
sample sizes, normality assumption, 113
sample standard deviation, 41
sample variance, 41
sampling distributions, 47–48
seasonal adjustments
deseasonalizing time-series data, 278–280
OLS technique for, 276–278
self-selection
Heckman selection model, 260–264
truncation, 255
sensitivity, 152
sequencing of positive and negative error values, 210, 211
serial correlation (autocorrelation)
AR(1) process, 219–222
AR(q) process (Breusch-Godfrey test), 221–223
autoregressive errors, 212–214
CLRM assumption of no autocorrelation, 209–210
d statistic, 220–221
described, 99–101
Durbin-Watson test, 219–221
in dynamic models, 270
examining patterns of, 209–212
misspecification and, 212
negative, 100, 210–211
positive, 100, 210–211
remedying harmful, 223–228
residual analysis, 214–215
stationarity assumption, 213–214
serial correlation robust standard errors, 223, 225–228
short-run propensity, distributed lag model, 268
significance, test of, 54, 57–58
simultaneous equations, 95
SIPP (Survey of Income and Program Participation), 72, 291
skewness, chi-squared distribution, 50
slope coefficients
regression analysis, 84, 85
relationship between *t* and F, 127
t-test of significance for, 127
slopes, pooled cross-sectional data, 282–283
Social Science Research Network, 309
software, econometric, 7. *See also* STATA
ordinary least squares (OLS) technique, 82–83
specification bias, 148
specification issues. *See also* overspecification
checking for misspecification, 147–149
Chow test, 151–152
described, 137
misspecification, 97
RESET (regression specification error test), 150
spurious correlation, 63, 273–274
squared residuals, ordinary least squares (OLS) technique, 76, 78–80
squared standard normal distribution, 50
standard deviation
described, 41
measuring variance and, 33–34
standard errors
of the coefficients, chi-squared distribution, 117–118
high multicollinearity, 179
prediction confidence interval, 133–134
of the regression, chi-squared distribution, 116
robust (White-corrected standard errors), 205–208
standard normal distribution
defined and explained, 46–47
OLS estimators, 114–115
table, 321
standard situation, defined, 93
STATA
analysis of variance (ANOVA), 126
autoregressive and distributed lag models, 269–270
Breusch-Godfrey (BG) test, 222

STATA *(continued)*
Breusch-Pagan (BP) test, 196–197
calculation of estimated regression coefficients, 82
Chow test and *F*-test for joint significance, 169–172
colors of text in the results area, 16
combining quantitative and qualitative data, 161
as command-driven program, 13
confidence interval for the OLS mean prediction, 133–134
creating new variables, 18–19
creating and saving datasets, 14
d statistic, 221
descriptive statistics, 42–43
deseasonalized model, 278–279
detrended data, 274
difference-in-difference model, 289–290
documentation, 13
error messages, 16
estimating, testing, and predicting in, 19–20
estimating a model with dummy variable, quantitative variable, and interacted variable, 164–165
FE model, 297–298
Goldfeld-Quandt (GQ) test, 200–201
Hausman test, 303–304
Heckman selection model, 262–264
importing data in non-STATA format, 14
logit results with estimated marginal effects, 252
Park test, 202
perfect multicollinearity, 176–177
as point-and-click software, 13
pooled cross section, 284
probit and logit results, 246–248, 250
random effects (RE) model, 300–301
"regress" command" for OLS results, 118
regression output and correlation matrix for independent variables, 182
regression output with dummy variable as the only independent variable, 158
regression output with VIF values for each independent variable, 189, 190
regression with a quantitative variable and two interacted dummy variables, 167
RESET (regression specification error test), 150
robust standard errors, 208
robustness/sensitivity analysis, 152
run test for auto-correlation, 218
saving your commands and output, 17–18
seasonal adjustments, 276, 277
serial correlation robust (Newey-West) standard errors, 228
significance of subsets of independent variables, 129–130
size of dataset, 15
stopping, 17
table-generating commands, 312
time trend model, 271–272
Tobit model, 257, 258
truncated regression, 259–261
viewing data, 14–16
weighted least squares (WLS) estimates, 205, 206
White test, 198
"xtset" command," 303
stationarity assumption, 213
statistical independence, checking for, 30
statistical insignificance, obsessing over findings of, 319
statistical significance
forgetting about, 320
obsessing over, 319
statistical tables, 321–326
statistically insignificant coefficients, ignoring variables with, 313
statistically insignificant test, 57
statistically significant regression coefficients, 120
statistically significant test, 57
statisticians, econometricians distinguished from, 10
Statistics For Dummies (Rumsey), 21, 39
stochastic population regression function, 65
strong relationship, 97
structural stability, Chow test, 151–152

summarizing the results of a research project, 313
summary measures (or moments), 31
Survey of Income and Program Participation (SIPP), 72, 291
surveys
  cross-sectional data collected through, 69
  panel data collected through, 72
switching of positive and negative error values, 211
symmetrical probability density function, 44, 46

• T •

tables, statistical, 312, 321–326
t-distribution (*t*-statistics or test)
  derivation and basic properties, 51–52
  high multicollinearity, 179
  OLS standard errors and, 119–121
  p-values associated with, 123
  slope coefficients, 127
  table, 322–323
  test of significance approach, 122–123
ten common mistakes in econometric analysis, 315–320
  asking the wrong questions first, 316
  assuming your results are robust, 320
  being inflexible to real-world complications, 318
  failing to familiarize yourself with the data, 316–317
  failing to use common sense and knowledge of economic theory, 315
  forgetting about economic significance, 320
  ignoring the work and contributions of others, 316
  looking the other way when you see bizarre results, 319
  making it too complicated, 317–318
  obsessing over measures of fit and statistical significance, 319
test of significance approach, 54, 57–58, 122–123
theoretical framework, description of, 309
theory, economic
  describing outcomes and making predictions with, 8
  econometrics and, 8
  failing to use your knowledge of, 315
time demeaning the data, FE estimation, 295–298
time effects, FE models, 296–297
time trends, 271–273
time-effect controls, 300
time-period dummy variables, pooled cross-sectional data, 283–285
time-series data
  autocorrelation, 101, 212–214, 270
  defined, 11
  deseasonalizing, 278–280
  detrending, 274–276
  first-differencing, 187–188
  gathering additional data, 186
  introduction to, 70–71
  spurious correlation problem, 273–274
Tobit model, 256–258
total sum of squares (TSS), 90, 91
transformation
  Cochrane-Orcutt (CO), 223, 226
  FD (first difference), 294
  first difference (FD), 294
  log, OLS technique, 95
  natural log, of the likelihood function, 245–246
  Prais-Winsten (PW), 223, 226
  within, 296
treatment (policy) effects, 285–286
trend-adjusted data, 274–276
true experiment, random assignment, 286
truncated dependent variables, 255–256
truncated regression, 258–262
two-tailed hypothesis tests, 54
type I error, 58, 123
type II error, 58, 123
types of data, 10–11

• U •

unbiased estimators (coefficients), 43, 108
unbounded predicted probabilities, 239–240

unconditional probabilities, 29, 30
unobservable variables, 292, 294–296

• V •

variability, random sampling and, 95
variables
  aggregation level in measuring, 11
  composite index, 188
  core, 152
  creating new, in STATA, 18–19
  dummy. *See* dummy variables
  observable, 292
  proxies, 150, 318
  qualitative. *See* qualitative variables
  random. *See* random variables
  statistically correlated but economically irrelevant, 10
  unobservable, 292, 294–296
variance
  chi-squared distribution, 50–51
  constant. *See* homoskedasticity
  decomposing, 88–89
  defined, 41
  determining overall or joint significance, 123–127
  of error term, 106
  F-distribution and normality assumption, 124
  of mean prediction, 131–133
  measuring, 33–34
  measuring proportion of, with R^2, 89–90
  OLS coefficients, 105
  of the regression, calculating, 116
  of *X,* 106
variance inflation factor (VIF)
  auxiliary regression and, 183–185
  dropping highly collinear independent variables, 189

• W •

websites, 309
weighted least squares (WLS) technique, 203–205
White test, 197–199
White-corrected standard errors (robust standard errors), 205–208
within estimator (fixed effects estimator), 296
within transformation, 296

• Z •

zero conditional mean error, 97–98