CAMBRIDGE

Cambridge University Press
978-1-108-48301-8 — Data Analysis for Business, Economics, and Policy
Gábor Békés , Gábor Kézdi
Index
More Information

INDEX

A/B test, 557, 559
absolute frequency, 60
accuracy, 467
adjusted R-squared, 379
administrative data, 13
aggregate trends, 661, 671
AIC, 379
algorithm, 387, 402, 407, 432, 686
  greedy, 432
alternative hypothesis, 148
annotation, 70
API, 13
approximate normality, 126
AR models, 500
AR(1), 500
area under the ROC curve, 469
arm of experiment, 560
ARMA (ARIMA) models, 503, 504
Ashenfelter’s dip, 694
assignment rule, 561, 566
assignment to treatment, 523
asymptotic normality, 126
attenuation bias, 225
attrition, 578
auto-arima algorithm, 502
autocorrelation, 500
autoregressive models, 500
average intent-to-treat effect, 568, 570, 613
average marginal effects, 310
average partial effects, 310
average treatment effect, 525, 568, 570, 602, 605
average treatment effect on the treated, 527, 602, 605

bad conditioning variables, 287, 544, 545, 572, 634
bag of words, 416
balanced panel data, 5, 677
base rate, 117
Bayesian statistics, 120, 130
benchmarking, 19, 45
Beta of an asset, 343
Bernoulli, 90, 94
between-subject comparison, 534
bias of prediction, 314, 376
BIC, 380
Big Data, 22, 393, 412, 494, 565, 626, 645
binary causal variable, 522
binary (dummy) variables, 279, 283, 521, 660
binomial, 84, 90
bin scatter, 104, 174
binomial distribution, 91
black box models, 453
boosting, 440, 449
bootstrap, 130, 273, 276, 440
bootstrap estimate of the standard error, 131
bootstrap sample, 130
box plot, 77, 104, 108
Brier score, 313, 317, 463

calibration, 315, 463
calibration curve, 315
CART, 419, 439, 475
categorical variable, 31
causal interpretation, 668
causal map, 531, 536
causal mechanism, 287
causal variable, 97, 520, 521
causality, 190
censoring of data, 322
Central Limit Theorem, 140
ceteris paribus, 287, 530
CI of the predicted value, 243
CI of the regression line, 243
classical measurement error, 223
classification, 300, 368, 459
  actual values, 467
  class imbalance, 480
  class rebalancing, 481
  classes, 459
  loss function, 471
  mean squared error, 463
  negatives, 467
  positives, 467
  root mean squared error, 463
  SMOTE class rebalancing , 481
  threshold, 466
  true classifications, 467
classification random forest, 476
  optimal classification threshold, 472
classification tree, 475
click through rate, 571
clustered standard errors, 665
coarsened exact matching, 603
collider variable, 544
common cause confounder, 541, 544, 547
common consequence bad conditioner, 544
common consequence variable, 634
common effect bad conditioner, 544
common support, 603, 608
comparative case study, 682
comparison group, 560
compliance (perfect/imperfect), 567, 570
conditional comparison, 100
conditional difference, 272
conditional distribution, 103
conditional event, 101
conditional expectation, 104, 172
conditional mean, 104, 172
conditional probability, 101, 458
conditioning, 100, 572, 633

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48301-8 — Data Analysis for Business, Economics, and Policy
Gábor Békés , Gábor Kézdi
Index
More Information

conditioning event, 101
conditioning variables, 100, 276
confidence Interval (CI), 128
  of the predicted value, 243
  of the regression coefficient, 238
  of the regression line, 243
  of the regression coefficient, 238, 541, 634
confounder variables, 272, 276, 541, 566, 572, 592, 636, 660
confusion table (matrix), 467
constant, 254
constructive skepticism, 551
continuous variable, 31
control group, 560, 682
controlled experiment, 539, 550
controlled variation, 113
controlling variable, 276
conversion rate, 571
correlation coefficient, 108, 398
counterfactual, 534, 683, 688
counterfactual outcome, 522
covariance, 108
covariates, 272, 564, 572
covariate balance, 564
complexity parameter (cp), 421, 426
cross-section time series data, 5, 34
cross-sectional data, 5, 33, 534
cross-validation (CV), 382, 398, 403, 426, 489
csv file, 4
cumulative association, 356
cumulative coefficient, 657
cumulative distribution function, 90
cumulative effect, 653, 668

data cleaning, 46, 397
data cleaning process, 48
data table, 4, 35
dataset, 4, 35
dataviz
  heatmap, 496
data wrangling, 48
degrees of freedom, 265, 273
density plot, 62
dependence of variables, 108
descriptive analysis, 59
df, 265
difference in differences (diff-in-diffs), 622–624, 629, 633, 650, 688
  estimator, 623
  regression, 623, 641
  analysis, 622
  with pooled cross-sections, 640
direct effect, 538
directed acyclic graph, 533
disambiguation, 43
discrete variable, 31
discrimination, 317
distribution, 61, 123
documentation table, 80
domain knowledge, 385, 400, 494, 536, 655
donor pool, 683
downsampling, 481
duplicate observations, 42
dummy variable, 32, 279
duration variables, 322

effect of the intervention, 523, 526
elasticity, 206
endogeneity, 605, 633
endogenous sources of variation, 536, 593, 595, 621
engineering, 460
ensemble methods, 439, 441
entity resolution, 42
error term, 596
error term of the regression, 173
estimate, 123
estimate of regression coefficient, 181
estimated value of a statistic, 121
estimation error, 370, 376, 413
event study method, 688, 694
event study regressions, 688, 689, 695
event time, 688
exact matching, 601
exogeneity of the instrument, 612
exogenous in the regression, 595
exogenous source bad conditioner, 544
exogenous sources of variation, 536, 593
experiment, 144, 287
experimental data, 113, 191, 539
experimental design, 561
explanatory variable, 173
exploratory data analysis (EDA), 59, 398
exponential trend, 332
external validity, 120, 135, 145, 237, 241, 384, 403, 410, 509, 550, 581
extraneous effects, 576
extrapolation, 185
extreme values, 62, 65, 221, 331, 432, 461

F-test, 275
factor variable, 31
false negative, 151, 154, 158, 250, 469, 473, 478, 483, 562, 589
false positive, 151, 152, 154, 158, 163, 469, 474, 478, 483, 564, 589
first difference (FD) regressions, 666, 671, 675, 689
fixed effects (FE) regressions, 659, 665, 671
feature engineering, 393, 397, 402, 461
features, 397
field experiment, 557, 559
first difference of time series variable, 338
first differences, 688
first stage of IV, 611
first-order serial correlation, 349
fit of regression, 188
fixed effects, 660, 688
flag variable, 46
flow variable, 33, 331

© in this web service Cambridge University Press www.cambridge.org

fold, 382
forecast horizon, 488
forecasting, 368, 488
formula, 387
frequency of the time series, 34, 330, 490
functional form, 174, 398

gaps in time series data, 330
general pattern, 120, 122, 145, 290, 378, 431, 549–550
genuine error, 370
geometric object, 69
Gini impurity, 476
goodness of fit, 188, 312
gradient boosting, 440
gradient boosting machine (GBM), 450

heterogeneous treatment effects, 524
heteroskedasticity, 239, 273, 666
histograms, 61, 103
holdout, 489
holdout set, 403
homoskedasticity, 239, 273
hypothesis testing, 144, 275, 562

identical and independently distributed (i.i.d.) variable, 140
identifier (ID) variable, 5
idiosyncratic error, 370
independence of variables, 108
independent events, 101
indicator variable, 32
indirect effects, 538
individual treatment effect, 523, 531
inference, 119, 144
influential observations, 221
instrument, 611
instrumental variables (IV), 610
inter-quartile range, 73
interaction term in regression, 282
intercept of linear regression, 179, 338
internal validity, 522, 542, 549, 577
interpolation, 185
interval prediction, 368
interval variable, 32
intervention, 520, 568, 682, 683
inverse probabilities, 101
irreducible error, 370, 376
ISO 8601 standard for dates, 57

joining data tables, 39
joint distribution, 104
joint probability, 101

k-fold cross-validation, 382
kernel density estimate, 62

lab experiments, 557
label engineering, 393, 394
lagged association, 356
lagged dependent variable, 351
lagged value (lag), 349, 522, 542
LASSO, 407, 413, 418
  penalty term, 407
  tuning parameter, 408
latent variable, 110, 223, 543
Law of Large Numbers, 140
lead terms (leads), 653, 654, 657
left-hand-side variable, 173
level, 202
level of power, 562
level of significance, 154, 562
linear probability model (LPM), 298
linear trend, 332
link function, 307
linking data tables, 39
live data, 366, 377, 378, 489, 509
ln scale, 212
local average treatment effect, 568
logarithm (log), 85, 201, 202
log approximation, 203, 339
log change, 339
log correction, 395, 396
log point, 203
logit, 307, 462, 604
lognormal distribution, 85
long format for xt data, 37
long-horizon (many periods ahead) forecasts, 492
long-run association, 356
long-run effects, 653, 668
longitudinal data, 5, 34, 534
loss function, 374
  asymmetric, 374
  convex, 374
  linear, 374
  log-loss, 313
  squared, 374
  symmetric, 374
lowess, 175, 396

machine learning, 387
marginal differences, 310
marginal effects, 310
matching model, 601
matching data tables, 39
matching on the propensity score, 604, 608
maximum likelihood estimation, 326
mean, 72
mean reversion, 190, 691
mean squared error (MSE), 375, 465, 486
mean-dependence, 108
measurement error in variables, 222
mechanism, 521
mechanism bad conditioner, 544
mechanism of reverse causality, 541
mechanism variable, 634
mechanism variables, 521, 532
median, 72
mediator variables, 521, 532

merging data tables, 39
minimum number of subjects, 563, 571
misclassification, 467
missing at random, 45, 578, 677
missing outcome values, 578
missing values, 44, 330, 397, 678
mode, 62, 72
model complexity, 379
model error, 370, 376
model-agnostic tool, 445
moderator variable, 282
moments, 91
multi-arm experiment, 560
multicollinearity, 274
multinomial variables, 321

natural experiment, 540
natural logarithm, 85, 106, 202
nearest neighbor matching on the propensity score, 604
negative spillovers, 577
negatively correlated, 108
Newey–West standard error, 351, 355, 652, 655
noise-to-signal ratio, 224
nominal variables, 32, 321
non-binary causal variables, 522
non-compliance, 568
non-parametric regression, 174, 426
non-treatment group, 560
nonlinear patterns, 179, 201
nonlinear relationships, 398
nonresponse, 578
normal distribution, 84
np-hard, 402, 432
null hypothesis, 148, 249

observable outcome, 522
observational data, 113, 191, 287, 539, 540, 595
observations, 4
omitted variable bias, 269, 543, 595, 660
omitted variables, 276
one-sided alternative, 148, 158
optimal classification threshold, 475
ordered categorical variables, 322
ordinal variables, 32, 321, 322
ordinary least squares (OLS), 181, 182, 185, 194, 197, 234, 263, 326, 379, 389, 407, 408, 435, 614, 619
original data, 366, 370, 377
outcome variable, 97, 100, 366, 520, 521
outlier, 65
overfitting, 290, 378, 403, 421, 440

p-hacking, 163
p-value, 154, 250, 277, 564
panel data, 5, 34, 534, 640, 665, 688
panel regression in first differences, 666
parallel trends, 641
parallel trends assumption, 629, 633, 652, 682, 684, 694
parameters, 84, 91
Pareto distribution, 92
partial dependence plot (PDP), 445
path diagram, 533
pathway, 521
penalty for model complexity, 379
per capita measures, 210
percentage change in variables, 338
percentage difference, 202
percentiles, 72
perfect collinearity, 274
Phillips–Perron test, 362
piecewise linear spline, 215, 277
placebo effect, 579
point prediction, 368
polynomial, 216, 277
pooled cross-sectional data, 640
pooled time series, 655, 659, 671
population, 18, 119, 122, 145, 378, 431, 549
positive spillovers, 576
positively correlated, 108
post-prediction diagnostics, 410
potential outcomes, 522, 535, 540, 560, 621
potential outcomes framework, 522, 531
potential treated/untreated outcome, 522
power calculation, 563
power-law distribution, 92
pre-intervention, 631
pre-intervention dip, 694
pre-intervention outcome, 622
pre-intervention trends, 630, 654, 668
precision of regression coefficient estimate, 238
predicted probability, 299
predicted value of linear regression, 184
prediction, 290
prediction error, 369, 373, 413
prediction interval (PI), 243, 244, 368, 399, 403, 410
prediction intervals, 397
predictive analytics, 366, 387
predictive data analysis, 366
predictive model, 366
predictor variables, 366
principal component analysis, 110
probability, 60, 458
probability distribution function, 90
probability prediction, 368, 458
probit, 307, 462
proof beyond reasonable doubt, 250
proof of concept, 250
propensity score, 604
propensity score matching, 604
Prophet, 497
proxy variable, 110, 223
pseudo R-squared, 313
pseudo-intervention, 694
publication bias, 164

quadratic function, 216
qualitative variables, 31, 279, 398

Cambridge University Press
978-1-108-48301-8 — Data Analysis for Business, Economics, and Policy
Gábor Békés , Gábor Kézdi
Index
More Information

quantiles, 72
quantitative causal variables, 527
quantitative prediction, 368
quantitative variable, 31, 521

R-squared, 188, 276, 290, 312, 375, 377, 660
random assignment, 535, 539, 556, 560, 564, 630
random forest, 439, 440, 445, 475
  bagging, 440
  bootstrap aggregation, 440
  decorrelated trees, 441
  probability random forest, 476
  pruning, 440
  variable importance, 444
random numbers generated by computers, 561
random sampling, 19, 561
random walk, 333, 345, 491
random walk with drift, 334
range, 73
ratio variable, 32
reduced form of IV, 611
reference category, 279
refinement, 317
regression, 172, 402
  analysis, 172, 203
  dependent variable, 173
  estimator of regression coefficient, 181
  left-hand-side variable, 173
  linear, 237, 445
  level–log, 206
  log-level, 206
  log-log, 206
  multiple, 172, 267, 291
  parametric, 174
  residual, 184
  right-hand-side variable, 173
  reverse, 189
  simple, 172, 267
  slope of linear regression, 179
  smoothing non-parametric, 174
  spurious, 344
  tables, 253
  time series, 651, 653
  weights, 228, 625
  x–x, 269
regression toward the mean, 190
regression tree, 419, 430, 439
  cost complexity pruning, 426
  pruning, 426
  automatic pattern detection, 432
  bins, 419
  branch, 419
  bucket, 420
  building, 419
  complexity parameter, 421
  cutoff point, 419
  growing, 419
  level, 420
  node, 419
  relative error, 420
  splitting, 419
  step function, 420
  stopping rule, 420, 421
  terminal node, 420
  top node, 419
  variable importance, 431
regression-discontinuity design (RDD), 613
relational data, 39
relative change in variables, 338
relative differences, 202
relative frequency, 60, 102
repeated samples, 122
representative sample, 18
residual, 351
results table, 80
reverse causality, 539, 654
reverse causality mechanism, 542
root mean squared error (RMSE), 375, 403, 420, 490
robust SE formula, 239
robust standard error (SE), 241, 254, 273, 351
robustness checks, 49, 219
ROC curve, 468
root-n convergence, 126
running variable, 613

sample, 18
sample design, 393, 460
sample size calculation, 563
sampling, 18
sampling distribution, 126
saturated models, 325
scaffolding, 69
scale variables, 32
scale-free distribution, 92
scatterplot, 104, 291
search algorithm, 327, 408, 419
season dummies, 344
seasonal binary variables, 492
seasonality, 332, 333, 344, 393, 491, 652
selection, 541, 642
selection bias, 45
self-selection, 541
sensitivity, 467
serial correlation, 349, 351, 500, 652, 666
short-horizon forecasts, 488, 500
shrinkage, 407
sign the omitted variable bias, 596
simple SE formula for slope coefficient of linear
    regression, 238
simulation exercise, 124
skewed distributions, 62, 73
skewness, 73, 107
slope coefficient, 179, 343
sources of variation, 113, 536
specificity, 467
spillovers, 576
splines, 277
spurious correlation, 344

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-108-48301-8 — Data Analysis for Business, Economics, and Policy
Gábor Békés , Gábor Kézdi
Index
More Information

stability, 385
stacked bar chart, 102, 103
standard deviation (SD), 73, 565
standard error (SE), 123, 276
standard error formula, 134
standard error of the predicted value, 243
standard normal distribution, 84
standard prediction error, 244
standardized difference, 73
standardized value of a variable, 73
stationarity, 345, 349, 385
stationary time series variable, 333
statistic, 72, 120, 123
statistical dependence, 108
statistical inference, 119, 135, 148, 237, 241
statistically significant, 250
step function, 174
stock variables, 33, 330
string variable, 31
subjective probability, 61
subjects, 520
summary statistics, 72
survey experiment, 557
survival time variables, 322
synthetic control, 683, 684, 695

t-statistic, 149, 249
t-test, 149, 283
target observations, 366
target variable, 366
terminal node, 426
test sets, 489
test statistic, 149
testing, 149
  a model, 381
  against a one-sided alternative, 251
  ambiguous identification, 43
  hypothesis tests, 564
  joint hypotheses, 251, 252, 275
  more than one coefficient, 251
  multiple hypotheses, 160
  power of test, 154
  size of test, 154
  two-stage least squares, 612
  Type-I error, 151
  Type-II error, 151
text parsing, 416
thought experiment, 589, 590
tidy data, 35
tidy data tables, 35, 691
time fixed effects (dummies), 661, 671
time index, 338
time series data, 5, 34, 330, 488
time series frequency, 34
total effect of the intervention, 526
training a model, 380
training–test method, 381
treated group, 560
treated subjects (units), 521
treatment, 540
treatment effect, 523
treatment group, 560
treatment variable, 521
trend, 332, 491, 333, 344, 652
trend line, 493
true value of a statistic, 120, 333, 344, 652
tuning parameter, 442
TwoSLS (2SLS), 612
two-sided alternative, 148
type of graph, 69
Type-I error, 151
Type-II error, 151

unbalanced panel, 5
unbalanced panel data, 677
unbiased probability prediction, 314
unbiasedness, 126
uncorrelated, 108
underfitting, 378
uniform distribution, 92
unit root, 334
unit root test, 362
unobserved heterogeneity, 616
untreated group, 560
untreated subjects (units), 521
unwanted mechanism confounder, 542, 642
usage of a graph, 68

value labeling, 48
variable labels, 48
variable selection, 597
variables, 4, 427
variance, 73
variance of prediction, 376
vector autoregression, 505
violin plots, 78, 104, 108

web scraping, 13
wide format for xt data, 37
winsorization, 461
wisdom of crowds, 439
within R-squared, 660
within-subject comparison, 534, 651, 660
work set, 403
workfile, 35, 460

xsec data, 5, 33
xt data, 5, 34
xt panel data, 660, 661
xt panel regression, 688

y–y plot, 291, 312, 410
year-over-year differences, 344
Youden index, 472, 475, 485

z-score, 73, 110
Zipf’s law, 92

© in this web service Cambridge University Press www.cambridge.org