

Index

- \doteq , nearly equal, 22
- $\| \|$, norm, 30
- \perp , orthogonal, 30
- $\perp\!\!\!\perp$, independent, 42
- \sim , distributed as, 68

- adj, classical adjoint, 32
- AIC, Akaike's information criterion, 210
- Aitken estimator; *see* feasible GLS estimator
- Alba and Logan, 106–107
- ambiguity in notation, 102
- American Cancer Society, 60
- Angrist, J., 208, 213
- Ansolabehere and Konisky, 204
- Arbuthnot, A., 94ff
- “as if by experiment”, “as if randomized”, 3, 6–9, 92, 96, 98, 180, 190–191, 217; *see also* natural experiments, randomization, causal inference, causation
- as the crow flies, 52
- assignment equation, selection equation, 134, 141, 143, 194
- association, 2ff; *see also* correlation
 - vs causation, 2ff, 12–13, 17, 53, 134, 207
- assumptions
 - for GLS and FGLS, 61–63
 - for IVLS, 181–182
 - for logit model, 128
 - for MLE, 118–119, 121, 149
 - for OLS, 42, 49, 61–62
 - for probit model, 121–124
- asymptotic covariance matrix
 - for FGLS, 64, 161–166, 167–172, 175
 - for IVLS, 183, 198
 - for MLE, 118–119, 123, 149, 150
- asymptotic mean
 - for IVLS, 198
 - for MLE, 118–119, 149
- asymptotic normality, 39, 59, 66, 70, 73, 118–119, 149, 198
- asymptotic SEs
 - compared with bootstrap SEs, 160–166, 167–172; *see also* asymptotic covariance matrix, plug-in SEs, SE as square root of diagonal element in covariance matrix

- asymptotic variance; *see* asymptotic covariance matrix, variance of random variable as diagonal element of covariance matrix
- asymptotics, 39, 59, 70, 73, 79, 118–119, 123, 149, 164–166, 175, 198, 211
 compared with bootstrap, 160–166, 167–173
- autoregression, 160–166, 174
 bootstrap principle for, 161
- average, 19
- average treatment effect, 132
- β , parameter vector, 41ff
- $\hat{\beta}$, $\hat{\beta}_{OLS}$, OLS estimator, 42ff
- $\hat{\beta}_{FGLS}$, feasible GLS estimator, 64, 161–166, 167–172, 174–175
- $\hat{\beta}_{GLS}$, GLS estimator, 64, 65
- $\hat{\beta}_{IVLS}$, two-stage least squares estimator, 176ff, 186
- $\hat{\beta}_{IVLS}$, instrumental-variables least squares estimator, 176ff, 183ff
- Bayesian methods, 210–211
- Beck, N., 80, 148, 175, 280
- Berkson, Joseph, 2
- bias, 5, 43, 53, 59, 64, 66, 68, 92, 112, 119, 124, 125–126, 130, 134–136, 139–140, 149, 160–166, 167–173, 174–175, 176ff, 184, 189–192, 195–196, 197, 270, 271
 in autoregression, 160–166, 167–172, 174
 due to endogeneity; *see* endogeneity
 due to failures in assumptions, 53, 59, 68, 124, 139–140, 189–192, 195–196
 in FGLS, 64, 66, 161–166, 167–172
 in IVLS, 184, 197–198
 in MLE, 119, 124, 149, 270, 271, 305
 omitted-variables, 59
 selection, 92, 130ff, 193–196
 due to simultaneity; *see* endogeneity
 small-sample, 184, 197–198
- bias-variance tradeoff, 150–151
- binary variable, 103
- binomial distribution, 116, 118, 119, 125–126, 269
- bivariate normal density, ii, 38, 137
- bivariate probit model, 134–138
- Blau, P. M., 81, 101
- Blau and Duncan, 81ff, 101, 302–303
- blinding, 14
- BLUE, best linear unbiased estimator, 61–63, 64, 78
- BMI, body mass index, 112, 152
- bootstrap, 155ff, 211
 compared with asymptotics, 166, 167–173
 compared with plug-in methods, 167–175
 to estimate bias, 160–166, 167–172, 174–175
 estimator, 156, 158
 parametric, 166–167
 replicates, 156–157
 sample, 156
 SE, 157, 161, 167–175
- bootstrap principle
 for autoregression, 161
 for FGLS, 165–166
 for regression, 159
 for sample mean, 157
- bootstrapping RDFOR, 167–175
- bootstrapping the bootstrap, 173
- box model, 26, 99, 156, 158, 169
- Braithwaite, B., 94ff
- breast cancer, 3, 4–5, 15
 and mammography, 4–5, 15, 200–201
 and telephones, 3
- butter market, 176ff
- χ_d^2 , chi-squared with d degrees of freedom, 68
- cancer and smoking, 2–3
- cancer and vitamins, 17
- cancer survivors, 200
- categorical variable, 104–105
 vs quantitative variable, 104–105
- Catholic schools, effectiveness of, 130ff
- causal inference, 1–17
 and constancy assumptions, 13, 91–102, 209ff
 from non-experimental data, 1ff, 6ff, 9ff, 81ff, 88ff, 91ff, 130ff, 176ff, 187ff, 193ff, 209ff
 from observational data; *see* causal inference from non-experimental data

- qualitative, quantitative, 11, 23, 89, 97, 101, 102
- and regression, 9ff, 81ff, 88ff, 91ff, 176ff, 187ff, 193ff, 209ff
- causation, 114, 209ff
 - vs association, 2ff, 12–13, 17, 53, 134, 207
 - as shown by controlled experiments, 1–5, 17, 22–23, 109–110, 144–145
 - as shown by logit models, 153–154
 - as shown by natural experiments, 6–9
 - as shown by path models, 81–86, 88–93, 94–102, 113–114, 209ff
 - as shown by probit models, 130–140
 - as shown by regression models, 1, 9–13, 81–86, 88–90, 91–102, 105–108, 176ff, 187ff, 193ff, 209ff
- causation, manipulationist and non-manipulationist views, 114, 214, 215
- centering residuals before bootstrapping, 174
- central limit theorem, 38–39, 59, 70, 154, 198, 211, 243–244, 251, 257
 - as inapplicable to GLS or FGLS, 66
- chi-squared distribution, 68
- cholera, 6–9, 16
- cigarettes; *see* smoking
- citations, model for determinants of, 107–108
- cloglog specification, 147
- cofactor in a matrix, 32
- Coleman, James S., 139–140, 151
- collinearity, 55, 59, 301–302
 - exact, 55
- column vector, 29
- computerized tomography, CT scans, 200
- concave function, 129, 177, 270, 273–274, 282; *see also* convex function
- concomitants, as non-manipulable variables, 192
- conditional probability, expectation, 28
- confounding, 2–4, 5, 11–12, 17, 42, 52, 92, 94, 130ff, 187ff, 193ff
 - in experiments, according to Victora et al, 112
- confounding variables
 - controlling for, 3–4, 11, 130ff, 187ff, 193ff; *see also* regression
- consistency, consistent estimator, 59, 198, 199–200
- constancy assumptions; *see* intervention
- constancy under intervention; *see* intervention
- Consumer Price Index, 60
- continuity correction, 244
- continuous variable, 104–105
 - vs discrete variable, 104–105
- control, 2–5
- control group, 2–5
 - vs treatment group, 2–5
- control variable; *see* covariate
- convex function, 28, 177, 243–274, 270, 282
 - defined, 28; *see also* concave function
- Cornfield, Jerome, 15, 153–154
- coronary heart disease, risk factors for, 153–154
- correlation, 3
 - coefficient, 19–21, 23
 - coefficient for random variables, 35
 - spurious, 3, 53, 56, 60
- cov, covariance, 27; *see also* covariance matrix
- covariance matrix
 - for FGLS estimates, 64, 161ff, 167ff
 - for GLS estimates, 64
 - as having variances on the diagonal, 46
 - for IVLS estimates, 183, 198
 - for MLE, 118–119, 149–150
 - for OLS estimates, 45ff
 - for random vectors, 35
- covariate, 42, 192
- critical value of a test, 70, 73, 299–300, 309
 - defined, 300
- cross-validation, 75
- cross-tabulation, 1, 3
 - vs modeling, 138
- crossover, 15
- Current Population Survey, 82, 83, 105, 113, 149, 305
- δ_i , disturbance, random error, 82, 89, 98, 179, 181, 189, 192, 194
- data snooping, 74–75, 79
- data variable, 18ff
 - mean of, 19
 - vs random variable, 24–25
 - standard deviation of, 19, 25
 - variance of, 19, 25
- death penalty, determinants of, 146–147

- degrees of freedom, 46–48, 68, 73, 85, 150
 demand curve, 176ff
 convexity of, 177
 as response schedule, 177ff
 demand, determinants of, 178ff
 density
 of jointly normal random variables, 38
 of a normal random variable, 38
 of a random variable, 36
 dependent variable; *see* response variable
 design matrix, 41ff
 det, determinant of a matrix, 31–32
 determinants of demand, 178ff
 determinants of supply, 178ff
 deviance, 150
 diagnostics, 210, 246, 297
 diagonal matrix, 36
 diet and cancer, 17
 DiNardo and Pischke, 208
 direct effects, 95ff
 discrete variable, 104–105
 vs continuous variable, 104–105
 discrimination against women, a statistical
 model for, 103–104
 disturbance term, disturbances, 22, 41ff
 assumptions on, 22, 41–42, 61–62,
 63–64, 91ff, 98ff, 182ff
 in autogression, 160
 vs residuals, 23–24, 44–45, 49, 53, 57
 Doll, Richard, 2
 dot product, 30
 dummy variable, 103–104, 113, 121,
 130–131
 and interactions, 138–139
 Duncan, Otis Dudley, 81, 114, 188
- E*, expectation, 24
 e_i , *see* residual
 ϵ_i , disturbance, random error, 22, 41, 61, 83,
 87, 93, 98, 179, 189, 192
 EC/IC Bypass Study, 14
 economic growth, in relation to left-wing and
 trade-union power, 147–148, 271
 Edgeworth, F. Y., 148
 education
 and fertility, 187ff, 306–308
 and PTA membership, 193ff
 educational level
 father and son, 87ff
 husband and wife, 105
- eigenvalue, eigenvector, 36
 email volume, determinants of, 205–206
 empirical covariance matrix, 65, 159, 164,
 171
 empirical distribution of a sample, 157–158
 endogeneity and exogeneity
 described, 59, 92, 96, 98, 134, 136,
 179–182, 198
 exogeneity assumptions behind causal
 inference, 92, 96–102, 136, 180,
 182, 188ff, 195ff
 mathematical examples, endogeneity
 bias, 54, 55–56, 59, 92, 112,
 198–199, 206, 207
 practical examples, endogeneity bias,
 139–140, 176ff, 189–192,
 195–196, 209ff
 statistical models for endogeneity,
 134–137, 176ff
 tests for exogeneity, 288
 endpoint maximum, 117
 epidemiology, 2ff, 15
 error function, 40
 error term; *see* disturbance term; *see also*
 residual
 error term compared with latent variable,
 123–124
 estimability, 125–127, 150
 vs identifiability, 125–127, 150
 estimate
 vs parameter, 23–24, 49, 57, 111
 estimators
 bias in, 53, 64, 68, 119–120, 160–166,
 167ff, 184, 197–200; *see also* bias
 due to failures in assumptions
 consistent, 59, 198, 200, 211
 FGLS, 64ff, 161ff, 167ff
 GLS, 63ff
 IISLS, 2SLS, 186
 inconsistent, 72, 199–200
 IVLS, 181ff, 306–308
 MLE, 115–124, 128ff, 303–306
 OLS, 9–13, 22–23, 34, 41ff, 295ff
 unbiased, 43, 46, 61–63, 63–64, 92,
 102–103, 109–110, 125–126, 269,
 270, 271
 Evans and Schwab, 130ff, 141–142, 151, 217
 data issues, 151
 modeling issues, 138–140
 exchangeable variables, 109

- exclusion restrictions; *see* identifying restrictions
- exogeneity, exogenous variable; *see* endogeneity and exogeneity
- expectation, expected value, 24
 compared with mean, 24
 conditional, 28
- experiments, 1–5, 14–17
 flawed, 14
 gedanken, hypothetical; *see* thought experiments
 unnecessary with large effects, 17
- experiments vs observational studies, 2–5, 14, 17
- explained variance, 51–53
 as related to F , 74
- explanatory variable, 41ff
- exponential families, 119, 149, 269
- exposed group; *see* treatment group
- eyewitness testimony, 201
- ϕ , standard normal density, 38
- Φ , standard normal distribution function, 121ff
- F -test, statistic, distribution, 59, 72–74, 300–301
- fertility and education, 187ff, 306–308
- FGLS, bootstrap principle for, 165–166
- FGLS, feasible GLS, 64–67, 161–166, 167–173, 174–175
 assumptions, 63ff
 estimator, 64–67
 one-step, 66, 161–166, 167–173
 two-step, 66
- Fisher, R. A., 2, 15, 72–73, 118–119, 123, 150, 257
- Fisher information, 118–119, 123, 150
- fitted value, 47
- fixed-effects models, 67, 78
 compared with random-effects models, 263
- flag, 103
- foreign investment, effects on political oppression, 105–106
- fraction of variance explained by regression, 50–53
- Framingham heart study, 153–154
- free arrow, 83–86, 88, 94, 102
- Frisch-Waugh theorem, 241
- full rank, 32, 41, 126–127, 129, 182, 184, 240
 and identifiability, 126–127, 129, 184
 vs rank deficient, 32
- Garrett, G., 147–148, 152, 280
- Gauss, Carl Friedrich, 9, 21, 34, 62
- Gauss' theorem
 for multiple regression, 34
 for the regression line, 21
- Gauss-Markov theorem, 62, 269
- Gibson, J. L., 88ff, 101, 105, 303
- Giffen, Robert, 148–149
- Gilens, 205
- GLS, generalized least squares, 63ff, 182
 assumptions, 63
 estimator, 63–64
 estimator is conditionally unbiased, 64
 model, GLS model; *see* GLS
 assumptions
- Goldberger, Joseph, 16
- Goldthorpe, J., 214
- goodness of fit, 53
 vs model validity, 53, 111, 207, 268, 292–293
- Gram-Schmidt process, 69
- graph of averages, 20–21
- H , hat matrix, 47
- happiness, determinants of, 203
- hat notation for estimators, 22
- Heckman, J. J., 151, 213
- heights of fathers and sons, 18–21, 23
- Hendry, D., 60, 212, 213
- Henschke et al, 200
- heteroscedasticity, 66, 78, 146–147, 279–280
- Hill, Bradford, 2
- HIP, Health Insurance Plan of New York, 4–5, 13, 15
- homoscedasticity, 60, 66
- Hooke's law, 22–23, 28, 43–44, 87, 91–93
- HRT, hormone replacement therapy, 17, 144, 152
- HS&B, High School and Beyond, 130, 136, 139–140, 151
- Huber-White correction for heteroscedasticity, 78
- hypothesis testing, 68–74, 299–300
 critical value, 68, 70, 73, 300
 deviance, 150
 F -test, 72ff, 300–301
 level, 300

- hypothesis testing (*cont.*)
 power, 300
 score test, 150
 significance levels, 70, 299–300
 size, 300
 statistical significance, 70, 300
 t -test, 68ff, 298–299, 308–309
 z -test, 68
- hypothetical experiments; *see* thought experiments
- I_θ ; *see* Fisher information
- $I_{n \times n}$, the $n \times n$ identity matrix, 30
- idempotent projection matrices, 47–48, 186
- identifiability, 125–127, 135–136, 150–151, 182, 213
 vs estimability, 125–127, 150
- identifying restrictions, 135–136, 140, 189–192, 196, 209ff
- IEEE arithmetic, 295
- IID, independent and identically distributed, 22, 24, 39, 42, 50, 68ff, 91ff, 118, 123ff, 155ff, 181ff, 209ff
- IISLS, 2SLS, two-stage least squares, 176, 181, 186
 relation to OLS, 181, 186; *see also* IVLS
- inconsistent estimator, 200
- independent and identically distributed; *see* IID
- independence assumptions as basis for computing SEs, 45–46, 54, 57, 59–60, 77
- independent effects, 70
- independent variable, 41ff
- independence vs orthogonality, 42, 244
- indicator variable, 103
- indirect effects, 95ff
- Indonesia, 202–203
- information, information matrix; *see* Fisher information
- information, observed, 119, 123, 131, 303–306
- inner product, 30
- instrument, instrumental variable, 135, 176, 181–185, 191, 195, 197–200
- instrumental-variables least squares; *see* IVLS
- intention-to-treat, 5, 15
 vs per protocol, 235
 vs treatment received, 5, 252
- interactions, 138–139, 143–144, 147–148
 of left-wing and trade-union power, 147–148, 280
 of TV ads with other variables in election model, 143–144
- intercept of regression line, 20, 50
- intermediate variables, 95ff
- International Agency for Research on Cancer, 15
- intervention, 13, 86, 87, 91–102, 114, 187, 190–191, 196, 209ff
 vs observation, 2, 13, 101, 191
 vs selection, 101; *see also* manipulation
- invariance assumptions; *see* intervention
- invariance under intervention; *see* intervention
- invertible matrix, 32
- iteratively reweighted least squares, 66
- IVLS, instrumental-variables least squares, 181ff, 197–200, 306–308
 assumptions, 181–182
 asymptotic normality, 198
 asymptotic variance, 183, 198
 bias in, 184, 197–198
 consistency, 197–198
 model, IVLS model; *see* assumptions
 relationship with OLS, 185, 186, 197–198
 simulations for, 199–200; *see also* IISLS
- Jacobs and Carmichael, 146–147
- Jensen's inequality, 270
- jointly normal random variables, 38–39
- just-identified system, 182
- Kahneman, D., 213–214
- Keefe et al, 14
- Keynes, John Maynard, 212
- King, Keohane, and Verba, 204, 216
- Koch, Robert, 6, 16
- Krueger, A., 185, 208, 213
- Λ , logistic distribution function, 128
 logit model defined by, 128
- L_n , log likelihood function, 116ff
- Labrie et al, 58–59, 252

- lag, lag term, in autoregression, 160–161, 161–166, 167–172, 174–175
- latent variable, 123–124, 128, 132–137, 139–140, 195
 vs error term, 124
- law of
 diminishing marginal utility, 177
 error, normal, 148
 large numbers, 14, 198, 236, 251, 252, 267, 268
 supply and demand, 177ff, 191
- lead time bias, 288
- least squares, 9–13, 22–23
 iteratively reweighted, 66
 weighted, 65, 67; *see also* regression, OLS
- left-wing political power, 147–148, 280
- Legendre, Adrien Marie, 9
- Lehmann, Erich, 73, 149, 217, 258
- level of a test, 300
- levels of measurement, 114
- Liebersohn, S., 214–215
- likelihood function, 116ff
- linear probability model, 123, 193–196
- log likelihood function, 116ff
- log odds ratio, 128
- logistic curve, history of, 153–154
- logistic distribution function, 128
 monotonicity of, 128
 symmetry of, 128
- logistic regression, 128
 history of, 153–154
- logit, 128
- logit model, 128–129, 149, 153–154, 305–306
- lung cancer death rate, 53, 56, 60, 252
- lung cancer screening, 200
- Lu-Yao and Yao, 201, 289
- malaria, 146, 279
- Malthus, Thomas, 153, 189
- Malthusian population theory, 153
- Mamaros and Sacerdote, 205–206
- mammography, 4–5, 15, 200–201
- manipulation, 86, 91ff, 94ff, 114, 209ff
 vs observation, 2, 13, 101–102; *see also* intervention
- manipulationist and non-manipulationist views of causation, 114, 214–215
- marathons, 201
- marginal effects, 131–132
- matrix, 29
 addition, 29
 classical adjoint, 32
 covariance, 35
 design, 41ff
 determinant of, 31–32
 diagonal, 36
 fixed, 42
 identity, 30
 inverse, 31–32
 invertible, 32
 multiplication, 30
 non-negative definite, 36–37
 orthogonal, 36
 positive definite, 36–37
 positive semi-definite, 36–37
 random, 42
 rank, 32
 symmetric, 30
 trace, 30, 33
 transpose, 30
 unitary, 36
 zero, 30
- maximum likelihood estimator; *see* MLE
- McCarthyism, 88ff, 101
- mean
 of data variable, 19, 25
 of random variable, 24–25
 of sample, 24
- measurement error, 112–113, 204, 290
- median, 28
- mediating variables; *see* indirect effects
- Meehl, P., 17, 215
- Megawati, 202–203
- microbiology, 16
- Mill, John Stuart, 17
- misspecification, 147, 149, 175, 209ff, 279–280
- MLE, maximum likelihood estimator, 115ff, 128ff, 130ff, 149–150, 303–306
 assumptions, 118–119, 121–124, 149
 assumptions, consequences of failures in, 124
 asymptotic normality, 118–119, 123, 149
 behavior with small samples, 119–120, 305
 bias in, 119, 120, 269–271, 305; *see also* bias due to failures in assumptions

- MLE, maximum likelihood estimator (*cont.*)
 in binomial model, 116–117, 118, 119
 compared with OLS in normal model,
 120, 271
 consistency, 118–119, 123, 149
 in logit model, 128, 291–292
 in normal model, 115–116, 119–120
 in Poisson model, 117
 in probit model, 121ff, 129, 130ff
 model selection, 204–205, 209ff
 MSE, mean square error, 21
 multicollinearity; *see* collinearity
 multiple comparisons; *see* data snooping
 multiple regression, 9–13, 26, 34, 41ff
 multivariate normal, 38–39
- $N(\mu, \sigma^2)$, normal distribution, 38
 natural experiments, 1, 6–9, 213
 NELS, National Educational Longitudinal
 Surveys, 151
 Newton-Cotes method, 152–153
 Newton-Raphson method, 152–153
 Neyman, Jerzy, 150, 217
 Neyman-Pearson statistic, 150
 nominal SEs, nominal variances, 80, 175; *see*
also asymptotic SEs, plug-in SEs
 non-experimental data; *see* observational data
 non-negative definite matrix, 36–37
 non-parametric methods, 258
 non-singular matrix; *see* invertible matrix
 norm, 30
 normal
 central limit theorem and the, 39
 density of the, 38, 124, 129, 137
 distribution function of the, 121ff, 129,
 130ff
 as exponential family, 119, 269
 joint normality, 38–39
 MLE in, 115–116
 moments of the, 80
 parametric bootstrap for the, 166–167
 probit model defined by the, 121–124
 random variables, 38–39
 regression models with errors that are,
 68ff
 standard, 38
 tail bounds for the, 129
 null hypothesis, 68, 72–73, 300
 mistakes formulating hypotheses, 71,
 111–112
 mistakes interpreting significance levels,
 71
 Nurses' Health Study, 144, 152
- observation vs manipulation, 2, 13, 101–102;
see also intervention
 observational data, 1ff
 causal inference from, 1–4, 6–9, 9–13,
 16–17, 81ff, 88ff, 91ff, 94ff, 130ff,
 176ff, 187ff, 193ff, 209ff
 observational studies, 1ff; *see also*
 observational data
 vs experiments, 1–5, 13, 17
 observed information, 119, 123, 131
 observed significance levels, 70, 300
 defined, 300
 mistaken interpretations, 71–72,
 111–112
 observed value, 25, 42
 vs random variable, 25
 occupational status, 81–86, 187–192
 odds ratio, 128
 OLS, ordinary least squares, 34, 41ff
 assumptions, 41–42, 61–62
 assumptions, consequences of failures
 in, 53, 59, 80
 asymptotic normality, 59, 299
 compared to IVLS, 186, 197–198; *see*
also regression, GLS, IVLS, path
 diagrams
 OLS estimator, 34, 42ff
 bias in; *see* OLS assumptions,
 consequences of failures in
 conditional variance, 45–46
 conditionally unbiased, 43
 consistency, 59
 OLS model, regression model; *see* OLS
 assumptions
 omitted-variables bias, 59
 one-step GLS, 65–66, 163ff, 167ff
 as problematic according to a social
 scientist, 77–78
 orthogonal, 30
 orthogonal matrix, 36
 orthogonality vs independence, 42, 244
 orthonormal, 39
 outer product, 31
 out-relief, 9–13, 45, 57, 148–149, 203,
 297–301
 over-identified system, 182

- P*-values, 70, 300; *see also* hypothesis testing, observed significance levels
- Pacini, F., 16
- pain control, 14
- parameter, 22
 vs estimate, 23–24, 49, 57, 111
- parametric bootstrap, 166–167
- Pasteur, Louis, 6, 16
- path diagram, 81ff
 as a causal model, 81–86, 88–90, 91–93, 94–102, 107–108, 113
 complete, 113
 represented as a box model, 99–100
- path model; *see* path diagram
- pauperism, 9–13, 45, 57, 148–149, 203–204, 297–301
- Pearson, Karl, 19, 23, 154
- Pearson and Lee, 19, 23
- per protocol analysis, 235; *see also* intention-to-treat, treatment received
- permeability of social structure, 81, 85–86
- Pettenkofer, Max von, 16
- philosophers' stones, 211
- Pisano et al, 200–201
- plug-in
 SEs, 64, 160ff, 167ff, 175
 SEs, compared with bootstrap SEs, 161ff, 167ff
 variance estimators, 46, 64, 119, 164, 171ff, 183, 198; *see also* asymptotic SEs, nominal SEs
- Podunk University, 266
- Poisson distribution, 117, 118, 119, 120, 269
- policy preferences, as related to political knowledge, 205
- political oppression, as related to foreign investment, 105–106
- population forecasting, modeling, 153
- positive definite matrix, 36–37
- positive semi-definite matrix, 36
- potential outcomes, 99, 213
- poverty, causes of, 9–13, 45, 57, 148–149, 297–301
- power of a test, 300
- Powers and Rock, 142–143, 151
- predicted value; *see* fitted value
- prediction, 1, 13
 vs description, 1, 13
- presidential elections and TV ads, 143–144
- probit model, 121–124, 129, 130ff
 bivariate, 134ff
- product
 inner, 30
 outer, 31
- projection theorem, 241
- prostate cancer, PSA, 58–59, 201–202, 252, 289
- PTA membership, determinants of, 193–196
- publication productivity, determinants of, 107–108
- Pythagoras' theorem, 33, 51–52
- quadrature, 137, 152–153
- qualitative vs quantitative causal inference, 11, 23, 89, 97, 101, 102
- qualitative vs quantitative variable, 104–105
- Quetelet, Adolphe, 11, 16
- r*, correlation coefficient, 19–21
- R, statistical package, 309
- R^2 , fraction of variance explained, 50–53
 for equations without an intercept, 75
 as measuring association rather than model validity, 53, 56–57, 111, 207, 292–293
 as related to *F*, 74
 as related to *r*, 53
- random-coefficients models, 209; *see also* random-effects models
- random-effects models, 78, 263
 compared with fixed-effects models, 263
- random error; *see* disturbance term
 vs residual; *see* disturbance term vs residuals
- random matrix, 42
- random number generators, 298
- random variables, 14, 24–27, 28
 correlation, 35
 covariance, 35
 density, 36
 expectation, expected value, 24
 independent, 24
 jointly normal, 38–39
 mean; *see* expectation
 normally distributed, 38–39
 observed values of, 25, 26, 42
 realizations of; *see* observed values
 relationship to data, 24–25, 42, 44
 relationship to samples, 24

- random variables (*cont.*)
 standard error, 25
 variance, 24, 25, 35–36
- randomization, 1–5, 14, 17, 109–110, 144–146
 as basis for causal inference, 1–5, 109–110, 144–146; *see also* “as if randomized”
- randomized controlled experiments; *see* randomization
- rank, 32
 deficient, 32
 full, 32, 41ff, 182ff, 240
 as related to identifiability, 41ff, 125–127, 182ff, 271–272
- rational choice theory, 213–215
- raw; *see* unstandardized regression coefficients, unstandardized variables; *see also* standardization, standardized regression coefficients
- RDFOR, Regional Demand Forecasting Model, 167ff
- reading, determinants of, 121–124, 149
 realization; *see* observed value
- Redelmeier and Greenwald, 201
- regression, 1ff, 9ff, 18ff, 41ff
 to control for confounding variables, 9–13, 187–192, 193–196
 diagnostics; *see* diagnostics
 effect, 236
 line, 18–23
 as a model for causation, 1, 9–13, 22–23, 81–86, 87–90, 91–93, 94–102, 105–111, 113–114, 176–181, 187–192, 193–196, 209ff
 multiple vs simple, 26
 uses of, 1, 13
- regression, bootstrap principle for, 159
- regression, uses of, 1, 13
 description, 13
 prediction, 13, 17
 to infer causation; *see* regression as a model for causation
- regression coefficients
 standardized, 86, 87
 unstandardized, 86, 87
- regression diagnostics; *see* diagnostics
- regression equation, 9–13, 22–23, 73, 81ff, 91ff, 94ff
 vs structural equation, 101–102
- regression line
 as flatter than SD line, 20–21
 as linear approximation to graph of averages, 20–21
- regression model, 9–13, 22–23, 41ff, 61–68
 for causation; *see* regression as a model for causation
 vs fitted model, 23–24, 44–45, 49, 53–54, 57; *see also* OLS model, OLS assumptions
- religious coping, 14
- rent control, 176–178
- replication, 79
- repression in the McCarthy era, 88–90, 303
- residential integration, determinants of, 106–107
- residual, 21, 23–24, 42–43
 vs random error; *see* disturbance term vs residuals
- residual variance; *see* unexplained variance
- response schedule, 13, 91–102, 113, 133–136, 176–181, 187, 189–192, 193–196, 217
- response variable, 41ff
- Rindfuss et al, 187ff, 217–218, 306–308
- RMS, root mean square, 21
- robust to misspecification, 146–147
- Rodgers and Maranto, 107–108, 266
- root mean square error, 21
- s_x , sample standard deviation, 19
- sample
 mean, 19
 mean, as a random variable, 24
 variance, 19
 variance, as a random variable, 24
- sample mean, bootstrap principle for, 157
- SAT, effects of coaching on, 93–94, 142–143, 151
- scatter diagram, 18ff
- Schneider et al, 193–196, 217–218
- school choice, effects of, 130–140, 193–196
- score test, 150
- screening; *see* lung cancer screening, mammography, prostate cancer
- SD, standard deviation, 18–19, 25
- SE, standard error, 25
 for slope and intercept of regression line, 50

- as square root of diagonal element in covariance matrix, 46; *see also* covariance matrix
- SE vs SD, 25
- selection bias, 92, 130, 134ff, 193–196
- selection equation; *see* assignment equation
- selection vs intervention, 101–102
- self-selection, models for, 130, 134ff, 193–196
- Semmelweis, Ignaz, 16
- Sen, A. K., 214
- Shaw, D. R., 143–144, 277–278
- significance, statistical; *see* significance levels
- significance levels, 68ff, 300
 - barely significant, statistically
 - significant, highly significant, 70
 - mistakes interpreting, 71–72, 111–112; *see also* hypothesis testing, observed significance levels, *P*-values
- simple vs multiple regression, 26
- Simpson's rule, 152
- simultaneity bias; *see* endogeneity
- simultaneous-equation models, 176ff, 209ff, 306–308
- size of a test, 300
- slope of regression line, 20, 50
- small-sample bias, 184, 197–198
- smoking, health effects of, 2–3, 15
- Snow, John, 6–9, 16
- social capital, 193–196
- social physics, 11, 16, 86, 89, 194
- social status, social stratification, 81ff, 101, 188
- sparse cross-tabs, handled by modeling, 138
- specification, specifying a model, 128, 149, 178–179, 193, 209ff
- specification error, 147, 149, 175, 308–309
- specification tests, 204–205, 210ff; *see also* diagnostics
- spectral theorem for matrices, 36–37
- stability under intervention; *see* intervention
- standard deviation
 - data variable, 19, 25
 - random variable, 25
- standard error; *see* SE
- standard normal density, 38
- standard units, 20
- standardization, 20, 82ff, 86, 87, 89ff, 90–91, 113, 263
- standardized regression coefficients, 86, 87
- statistical modeling, issues in, 209ff
- statistical packages, 309
 - as replacing statistical tables, 309
- Stouffer survey, 88–90
- stratification (cross-tabulation), 1–4, 138; *see also* social stratification
- structural equations, 101–102, 187, 190–192, 213–215
 - vs regression equations, 101–102
- structural zeros, 136
- student's *t*-test, statistic, distribution, 68–70, 298–299, 308–309
- supply, determinants of, 178ff
- supply curve, 176ff
 - as a response schedule, 178ff
 - concavity of, 177
- symmetry
 - of logistic distribution, 128
 - of matrices, 30
 - of normal distribution, 120
 - of projection matrices, 47, 186
- 2SLS, IISL, two-stage least squares, 176, 186; *see also* IVLS
- t*-test, statistic, distribution, 68–70, 298–299, 308–309
- telephones and breast cancer, 3, 15
- thought experiments, 95ff, 114, 190–192, 209ff
- Timberlake and Williams, 105–106
- Tinbergen, Jan, 212
- tolerance of dissent, 88ff, 303
- trace of a matrix, 30
- trade-union power, 147–148, 280
- traffic fatalities, 201
- transpose of a matrix, 30
- trapezoid rule, 152
- treatment group, 2–5
 - vs control group, 2–5
- treatment received, 5
 - vs intention-to-treat, per protocol, 5, 235, 252
- TV ads and presidential elections, 143–144
- Tversky, A., 213–214
- Two-County Study on mammography, 5
- two-equation model for effects of Catholic schools, 134ff

- two-stage least squares; *see* IISLS
 two-step GLS, 66
- unbiased estimators, 43, 46–48, 58, 61–64,
 78, 92, 102–103, 109–110,
 119–120, 125–127, 269–270
- unconditional vs conditional expectation, 59
- under-identification, 125–127, 136, 150–151,
 182, 209ff
- unexplained variance, 51–52
- unitary matrix, 36
- University, Podunk; *see* Podunk
 University
- unmodeled heterogeneity, 204
- unstandardized regression coefficients,
 variables, 86, 87
- var, variance, 19, 24–25, 27, 35
 of data, 19, 25
 of estimates in a simple regression
 model, 50
 explained, unexplained, 51–52
 of FGLS estimates, 64, 65–66, 161ff,
 167ff, 174–175
 of GLS estimates, 64
 of IVLS estimates, 183, 197–198
 of MLE, 118–119, 123, 149–150
 of OLS estimates, 45–48, 61–62
 of random variables, 24–25, 35
 of random variables as diagonal
 elements of the covariance matrix,
 35
 of samples, 24
- variable
 binary, 103
 categorical, 104–105
 continuous, 104–105
 dependent, 41ff
 discrete, 104–105
- dummy, 103–105, 113
- endogenous; *see* endogeneity and
 exogeneity
- exogenous; *see* endogeneity and
 exogeneity
- explanatory, 41ff
- independent, 41ff
- indicator, 103
- instrumental, 176, 181–186, 191–192,
 193–196
- latent, 123–124, 126–127, 128,
 132–133, 134–137
- non-manipulable, 114, 192, 196
- qualitative, 104–105
- quantitative, 104–105
- random, 14, 24–27
- response, 41ff
- vectorizing code, 301
- Verhulst, P. F., 153
- vitamins and cancer, 17
- voter turnout, determinants of, 204
- weighing designs, 108–109, 112
- weighted least squares, 65
- White, H., 78
- White's correction for heteroscedasticity, 78,
 146–147, 175
- Wilks' statistic, 150
- \bar{x} , sample mean, 19
- X , generally the design matrix, 41ff;
 sometimes, a random variable or
 vector, 14, 94, 102
- Yule, G. U., 9–13, 16–17, 60, 148–149,
 203–204, 297–301
- $0_{m \times n}$, an $m \times n$ matrix of zeros, 30
- z-test, 68