

Subject Index

- Adaptive rounding, *see* rounding
- Adherence, 348–350
- Adverse event, 264
- Ante-dependence, 373
- Assumptions, *see* MAR, MCAR, MNAR, Model
- Asymptotic, 63, 64, 204
- coverage, 69
 - linear estimator, 67
 - normality for Rubin's rules, 45
 - upward variance bias with weights, 320
- Auxiliary variables, 65, 141, 145, 149, 150, 156, 162, 163, 357
- imputing survival times, 187, 189–190
 - in survey settings, 319
 - interaction & non-linear, 152
- Bahadur model, 116
- Bayes theorem, 384–385
- Bayesian, 10
- analysis, 34, 49, 183
 - approach, 72, 85
 - Beta prior, 130
 - bootstrap, 113, 168, 188
 - conditional predictive distribution, 46, 48, 49, 50
 - conjugate prior, 59
 - Dirichlet prior, 130
 - draw, 40, 44, 46, 48
 - framework, 48, 73
 - improper prior, 90
 - joint model, 130
 - justification for MI, 48–49
 - pattern mixture model, 248
 - perspective, 130
 - posterior, 43, 48–49, 53, 57, 90, 113, 384
 - multivariate normal, 49, 391–393
 - prior, 43, 59, 85, 92, 384
 - censored survival time, 179–180
 - elicitation, 255, 265
 - for missing data, 85, 254
 - improper, 85, 105
 - selection model, 248
- Bernstein–von-Mises theorem, 49
- Beta distribution, 113, 130
- Binary data, *see* Data, binary
- Bootstrap
- as alternative to Rubin's variance, 362
 - Bayesian, 188
 - data, 190
 - full mechanism, 68
 - regression calibration, 304
 - to assess model performance, 209
 - use with MI, 69
- Calibration, 205–206
- Case-cohort studies, 192–197

- Censoring, 176
 - at random with auxiliary variables, 189
 - informative, 260
 - mechanisms, 186
- Chained equations, *see* Full conditional specification (FCS)
- Coarsening, 296
- Coin flip rounding, 98–100
- Compatibility, 135, 136, 152, 153, 158, 160, 165
- Complete
 - data quantities, 46, 63
 - records analysis, 28, 40, 356, 360, 361, 362, 365, 381
 - interactions, 152
 - just another variable, 144
- Compliance, 250, 264
- Composite variables, 140–141
- Conditional
 - mean, variance, 48
 - predictive distribution, 49
- Confounding assumptions, 335
- Congenial yet mis-specified, 72
- Congeniality, 48, 49, 50, 62, 63, 64, 66, 72, 73, 131, 136, 162, 164, 315
 - auxiliary variables, 65
 - constructing, 72
 - log-linear model, 126, 129
 - with interaction, 150
 - with multilevel data, 217
 - with non-linear relationship, 141, 142
- Conjugate prior, 59
- Contingency table, 126, 129
- Copy increments in reference, 269, 273, 274, 275
- Copy reference, 269, 273, 274, 275
- Covariance
 - ante-dependence, 373
 - differs by treatment arm, 150
 - matrix, 85, 97, 126
 - as random effect, 151
 - common, 150
 - constrained, 107
 - ill-conditioned, 371
 - prior distribution, 85, 392, 393
 - shrinkage, 151
 - unstructured, 84, 85, 91, 150
 - unstructured vs. structured, 371
 - updating elementwise, 103
- Coverage error, 69–71
- Cox model, 176, 178, 181, 184, 261
- Cross-classified data, 231–232
- Cubic splines, 142
- Cycles, 91, 109
 - within and across times, 374
- Data
 - binary and binomial, 96–118
 - higher order associations, 116
 - categorical, 115, 116, 119–132
 - continuous, 81–94, 115
 - cross-sectional, 5
 - longitudinal, *see* Longitudinal data
 - mix of types, 97, 128
 - multilevel, *see* Multilevel
 - ordinal, 98, 100, 105
 - survey, *see* Survey
- Data augmentation, 114, 120, 130
 - algorithm, 387
 - for perfect prediction, 112–114
- Derived variables
 - no missing data in, 141–142
- Descriptive analyses, 23, 364–366
- Design weights, 317
- Diagnostics
 - interactions, 152
 - MCMC sampler, 87, 229, 386
- Dirichlet distribution, 130
- Discriminant analysis, 126
- Discrimination, 22, 206
- Donor pools, 189
- Double resampling, 188
- Double robustness, 6, 191
- Dropout pattern, 21
- E–M algorithm, 3, 35, 87, 93, 104, 114
- Efficiency, 10, 70, 71
 - relative, 55

- Estimand, 250–251
 hypothetical, 266–267, 271–273
 intercurrent event, 263
 policy, 271–273
 intercurrent event, 317
- Estimating equations, 46, 49, 317
- Estimator
 asymptotic linear, 67
 self-efficient, 67
- Exponential distribution, 178
- External validation, 208
- FCS, 81, 92, 109, 112, 113, 115, 116, 130
 cycles, 109
 end effect, 109
 equivalence to log-linear model, 115
 just another variable (JAV), 143
 justification
 binary data, 128–129
 categorical data, 128–129
 continuous data, 91–92
 perfect prediction, 114
 relationship to joint modelling, 129
 shrinkage, 373
 speed vs. joint modelling, 115
 survival data, 181, 184–186
 two-fold MI, 373
 with skips, 170
- Fisher–Behrens distribution, 52
- Fraction of missing information (FMI), 42, 55, 380, 383
 equal across vector components, 54
- Fractional polynomials, 166–168
- Frequentist, 40
 MI inference, 49
 MI properties by simulation, 377–381
 properties of Rubin’s variance
 formula, 58
- Full conditional specification, *see* FCS
- G-formula, 336
 equivalence to MI, 209, 210
- Gaussian quadrature, 123
- General location model, 96, 108, 119, 125–126, 130, 131
 relationship to latent normal model, 125
- Generalized estimating equations (GEE), 66
- Genetic sub-types, 150
- Gibbs sampler, 85, 91, 106
 as special case of Metropolis
 Hastings, 385
 binary data, 101
 for covariates, 89–90
 joint distribution, 92
 multivariate normal, 85–88
 poor man’s, 92
- Hastings ratio, 103, 153
- Hazard ratio, 45, 136, 158, 177, 178, 180, 260, 298
 log transform, 49
- Heckman selection model, 19
- Heteroscedasticity, 240
- Hierarchical, *see* Multilevel
- Ignorable missingness mechanism, 21, 33, 85, 187
- Importance sampling, 279
- Improper imputation, 67
- Imputation model, 47, 358
 richer, poorer, 71
 auxiliary variables, 358
 FMI indicates problems, 359
 joint, *see* Joint modelling
 practical considerations for, 355–360
- Impute then transform, 143
 simulation study, 156
- Incomplete data at model
 implementation, 208–209
- Inference, *see also* Frequentist, Bayesian
 framework, 10–11
 robustness with missing data, 5, 9, 245
- Information, 42, 46, 49, 63
- Information anchoring, 275–278

- Informative censoring hazard ratio, 260
 Informative missingness odds ratio, 256
 Instrumental variables, 347–349
 assumptions, 347
 estimand, 348
 exclusion restriction, 347
 monotonicity assumption, 348
 via MI, 349–351
 Intention to treat, 251, 363
 Interaction, 129, 137
 auxiliary variables, 152
 binary and continuous, 149
 borrow across categories, 151
 categorical variable, 149
 complete records, 152
 continuous and categorical variable, 151
 group indicator missing, 152
 impute separately in groups defining, 148–152
 proportion of missing data small, 152
 skips, 170
 survival data, 187
 two-way, 130
 variables fully observed, 148–151
 Intercurrent event
 change of slope post, 264–267
 definition, 249, 263–264
 distribution post, 267–275
 Internal validation, 207
 Inverse probability weighting, 4, 289, 316
 Item, 5–6, 329
 Joint modelling, 85
 constraining partial associations, 373
 vs. FCS, 92, 129
 large datasets, 372–374
 longitudinal and survival data, 176
 multilevel imputation model, 216
 perfect prediction, 97
 R package jomo, 241
 relationship to FCS, 129
 robustness, 115
 Jump to reference, 268, 273, 275
 relationship to copy reference, 274
 technical details, 269
 Just another variable (JAV), 144, 145
 bias under MAR, 145
 joint imputation model, 144
 logistic regression performance, 145
 Rubin’s rules, 144
 simulation study, 144, 156
 validity conditions, 149
 K nearest neighbours, 189
 Kaplan–Meier, 190, 261–262
 Kronecker product, 89, 222
 Large datasets, 56, 189, 371–376
 two-fold MI, 374
 Last mean carried forward, 268, 273
 Last observation carried forward, 269
 Latent normal model, 113, 116
 binary data, 100–104
 binomial data, 104
 categorical data, 119, 121–125
 covariates in variance model, 131
 imputation of mix of data types, 127, 128
 MCMC sampler, 101–104
 ordinal data, 105–108
 overfitting, 116
 relationship to general location model, 125–126
 software, 130
 Likelihood, *see also* ML, MLE
 analysis, 28
 dominating prior, 49
 methods, 24
 partial, 34, 183
 pattern mixture model, 247
 penalised, 113
 ratio test, 54
 selection model, 248
 Linear regression, *see* Regression, linear
 Local residual draws, 148
 Log odds ratio, *see* Odds ratio

- Log-linear model, 126, 130
 Gibbs sampler, 129
 partial association, 129
 saturated, 129, 130
 two-way interactions, 129
- Logistic regression, *see* Regression, logistic
- Longitudinal data, 81–88
 change of slope post intercurrent event, 264–266
 pattern mixture model, 263–275
 post-intercurrent event distributions from other groups, 267–268
- Loss to follow-up, 263–267
- Louis formula, 46
- Malahanobis distance, 189
- MAR, 61, 82, 83, 84, 85, 359
 definition of, 13–17 discussion and examples, 35
 connection to confounding, 336
 departure point for sensitivity analysis, 21, 247, 264
 pattern mixture assumption, 246
 plausibility, 376
 randomised arm, 268
 Rubin's rules, 45–47
 selection assumption, 246
 subtlety of, 15–17
 untestable, 289
- Marginalising over Y_M , 49
- Markov Chain Monte-Carlo, *see* MCMC
- Maximum indicant model, 121–125
 MCMC algorithm for, 123
 Monte-Carlo estimate of probabilities, 130
 probabilities implied by, 123
 software, 130
- Maximum Likelihood Estimate, *see* ML & MLE
- MCAR, 15, 17, 23, 57, 68, 84
 definition of, 12
 for validity of JAV, 143
- MCMC, 85–88, 102–106, 109 115
 convergence, 87, 229, 386–387
 efficiency of, 103
 pattern mixture model, 253–254, 267–268
 survival data, 260–262
 sampler, 386–387
 Cox model, 183
 for covariates, 127
 Gibbs, 387
 maximum indicant model, 100
 mixing, 386
 multilevel data, 224–228
 non-linear relationship, 153–156
 transformation to improve mixing, 387
- Mean imputation, 209
- Mean Square Error (MSE), 67, 360
- Measurement error, 294
 as a missing data problem, 295
 effects of, 295
 external information, 307–308
 fraction of missing information, 309
 non-differential, 296, 299
 regression calibration, 298
 replication data, 301–307
 validation data, 296–301
- Meta-analysis
 individual participant data, 237
- Meta-regression for hazard ratio, 263
- Metropolis–Hastings, 85, 102, 104, 105, 113
 proposal distribution, 103
 sampler, 386, 387
 step for missing covariates, 153
- MI
 approximations, 49
 combining
 likelihood ratio statistics, 51–55
 p-values, 55
 confidence interval, 51
 connections to causal inference, 335, 351
 cross-sectional for multilevel data, 238

MI (*Continued*)

descriptive analyses, 364–366
 equivalence to G-formula, 336–338
 frequentist inference, 50–55
 general justification, 64–71
 generic procedure, 45–47
 hot deck, 188–191
 hypothesis test, 51
 information recovered, 115
 instrumental variables, 349–351
 complier average causal effect, 348
 intuitive outline, 40–45
 max. like. MI, 381, 382
 mis-specified, 72
 model building, 202–204
 non-parametric, 114, 188–189
 MNAR, 189
 no of imputations, 191
 propensity matching, 189
 survival data, 189–191
 variance underestimation, 189
 not making up data, 56–59, 362
 objections to, 355, 360–363
 principal stratification, 346–347
 propensity scores, 340–344
 proper, 116
 public use data, 328
 record linkage, 376–378
 reference *t*-distribution, 42, 51–53
 replication data, 301–307
 sequential regression imputation, 129
 small or large missingness, 363
 small sample inference, 53
 two-fold for large datasets, 374
 validation data, 296–301
 variance estimators
 compared, 69–71
 exactly unbiased, 60
 limiting distribution, 63
 Robins and Wang *vs.* Rubin, 67
 vector inference, 53–54
 Wald tests, 55
 when to pool, 205

with weights, 317–320
 conditions consistent, 318
 within imputation variance, 47
 Misclassified covariates, 297–299
 Missing At Random, *see* MAR
 Missing Completely At Random, *see* MCAR
 Missing data
 bias, 10
 consequences, 10
 loss of efficiency, 10
 Missing indicator method, 210
 Missing Not At Random, *see* MNAR
 Missingness mechanism, 5, 10, 11,
 see also Ignorable; MCAR;
 MAR; MNAR
 assumption untestable, 15
 implic. odds ratio, 30–32
 implic. regression, 24–33
 plausibility, 358
 Missingness pattern, 7–10, 16, 17, 253,
 357
 as covariate, 341
 causal view, 351
 ML, 67, 88, 97
 MLE, 25, 49, 54, 57, 112
 MNAR, 17–21, 22, 23, 359
 bias extent, 360
 composite endpoints, 140
 depends on estimand, 251
 implic. lin. reg., 32–33
 implic. odds ratio, 30–32
 missing covariates, 256
 modelling, 246–249
 non-parametric MI, 189
 on covariates, 28
 on response, 26, 27
 on response and covariate, 29
 pattern mixture model, 18
 selection model, 18
 sensitivity analysis, 360
 survival data, 260
 untestable, 15, 248
 via MI, 253–259
 via re-weighting, 279

- Model
 assessment, 206–208
 diagnostics, 366–367
 performance, 204–206
 saturated repeated measures, 265
 validation, 207–208
 external, 208
 internal, 207
- Moments, 49, 116, 123, 235
- Monotone, 7–8, 11, 82, 85, 90, 96, 108, 112, 116, 119, 129
 consistent with MAR, 83
- Monte-Carlo error, 65, 309, 310, 359, 363, 367, 370
- Multilevel data, 131, 213–214, 216, 320, 371
- Multivariate imputation by chained equations, *see* Full conditional specification (FCS)
- Multivariate normal, 85, 91, 93, 100, 115, 150, 181, 268, 386
 and general location model, 125–126
 density, 388
 for binary variables, 98–100
 for categorical data, 121
 measurement error, 297
 multilevel imputation, 222
 posterior density, 392
 robustness, 115
- Nearest neighbours, 190
- No. of imputations, 55–56, 367–369, 383
 large, 50–51
 simulation studies, 379
 small, 51–54
- Non-linear relationships, 67, 72, 185, 223, 236, 360
 general, 142
 measurement error, 300
 splines, 142, 165
 survival data, 187
 with weights, 320
- Non-monotone, 85, 93, 97
- Non-parametric, *see* MI, non-parametric
- Numerical integration, 123, 248, 289
- Odds ratio, 30–32, 45, 99, 112, 114, 119
 heterogeneity in, 129
 informative missingness, 256
 log transform, 49
- Ordinary least squares (OLS), *see* Regression, linear
- p-values, MI combination rule, 55
- Parametric survival model, 261
- Partially observed
 covariates and response, 15–33
 covariates, 27–30
 response, 24–27
- Passive imputation, 145–146
 simulation study, 144
- Pattern mixture model, 18, 20, 247
 avoiding formal priors, 267–268
 examples, 248, 255
 expressed pictorially, 20
 implies selection model, 18, 246, 289
 likelihood, 247
 longitudinal data, 263–275
 missing covariates, 256–258
 more accessible than selection model, 289
 prior, 255, 265
 survival data, 260–263
 via MI, 251–263
- Pattern submodels, 211–212
- Pivot, 53
- Potential outcomes, 334
- Precision matrix prior, 303
- Prediction, perfect, 112–116
 categorical data, 121, 130
- Predictive mean matching, 146–148
 simulation study, 144
- Principal stratification, 344
 estimation by MI, 346
- Prior, *see* Bayesian
- Prior informed MI, 376
- Probability density functions, 387–392

- Prognostic modelling, 201
 discrimination, 206
 MI, 202–204
 validation, 206–207
- Propensity scores, 189, 339
 definition, 339
 matching on, 340
 MI for, 341–344
 modelling, 340
 weighting by, 339–340
- Proper imputation, 67
- Proportional
 hazards, 190
 odds model, 97
 probit model, 104
- Proposal distribution, 103, 104, 164,
 184, 227, 385, 387
- QRISK, 201
- Random coefficient model, 142, 214
- Randomised arm MAR, 268, 271, 275
- Ratios, 136, 138, 165
- Receiver operating characteristic
 (ROC), 22, 205
- Record linkage, 376–378
- Regression
 imputation, 81–83
 linear, 24, 27, 29, 32, 64, 67, 72, 81,
 92, 108, 116, 126, 256
 logistic, 31, 32, 45, 96, 99, 108, 111,
 113, 116
 difference from probit, 100
 propensity, 189
 multinomial, 119, 129, 181
 perfect prediction, 120
 non-linear relationship, 154
 ordinal, 97, 113
 probit, 32, 100, 112
 survey setting, 318
 switching, *see* Full conditional
 specification (FCS)
- Regulatory agencies, 251
- Reiter's rules, 307–308
- Rejection sampling
 binary data, 102
 bound Cox regression, 185
 categorical data, 123–124
 missing covariates, 153–155
 perfect prediction, 130
 rate, 115
 simulation study, 156
 with FCS, 184–185
- Relative survival, 180, 181
- Reporting, 364–365
- Response probability ratio, 256
- Restricted maximum likelihood
 (REML), 84, 88, 89
- Retrieved dropout analysis, 274
- Ridge
 parameter, 116
 regression, 372, 373
vs. constrained partial association,
 375
- Robust variance estimate, 68, 69, 320
- Robustness, 5
- Rounding, 115
 adaptive, categorical data, 121
 bias due to, 99
 methods compared, 99
 simple, 98–100, 113
 software, 116
- Rubin's rules, 41–43, 45–47, 396
 application order, 239
 derivation, 51–54
 domains, 329
 frequentist properties, 58–64
 likelihood ratio statistics, 54
 linear regression theoretical
 examples, 56–64
 non-parametric MI, 189
 situations biased, 66–71, 172,
 314–317, 328
 survey setting, 71, 312, 319
 survival data, 187, 261
 validity condition with weights, 314
 with 'just another variable', 144
 within strata, 314
- Rubin's typology, 5

- Sandwich variance, 68, 320
- Saturated repeated measurements, 271
- Score equations, 46, 62, 284, 318
- Seed
 - for random numbers, 359, 377, 378
- Selection mechanism, 19, 26
- Selection model, 18, 246
 - examples, 248
 - implies pattern mixture, 18, 247, 289
 - likelihood, 248
 - longitudinal data, 284
 - probit, 18
 - sensitivity parameter, 281
 - untestable assumptions, 248
 - via re-weighting, 279–289
- Self-efficiency, 67, 315
- Semi-parametric, 4, 189
- Sensitivity analysis
 - from MAR, 247
- Sensitivity analysis, 10, 21, 128, 360
 - change of slope, 264–267
 - estimands, 249–251
 - framing, 249–251
 - hypothetical estimand, 266–267, 271–275
 - information anchoring, 275–278
 - longitudinal data, 263–275
 - missing covariates, 256–258
 - motivation, 245–246
 - pattern mixture, 251–254
 - policy estimand, 271–275
 - selection model via re-weighting, 279–289
 - reliability of, 281–284
 - survival data, 187, 260–263
- Sequential regression (multivariate)
 - imputation, *see* Full conditional specification (FCS)
- Shared parameter model, 246
- Skewed continuous variables, 115
- Skips, 169–172
 - interactions, 170
 - Rubin’s variance formula
 - conservative, 172
 - variable fully observed, 171
 - variable partially observed, 171
 - vary size of imputed datasets, 172
- Software, 4
 - REALCOM-impute, 93, 103, 116, 127, 130, 387
 - CAT for categorical data, 130
 - MIX for mix of data types, 130
 - smcfc (R, Stata), 161
 - winBUGS, 287
 - binary data, 116
 - categorical data, 130
 - general location model, 123
 - latent normal model, 130
 - multivariate normal data, 93
 - R, 93, 394
 - SAS, 93
 - Stata, 93, 396
- Squares, 137–138
- Statistical analysis plan, 363
- Substantive model, 40, 41, 44, 45, 47, 50, 63, 67
 - estimating equations, 317
 - interaction & non-linear, 148, 149
- Substantive model compatible
 - imputation, 152–165
 - case-cohort data, 193, 195–197
 - competing risks data, 186
 - multilevel data, 222
 - survey data, 325
 - survival data, 181
 - time-varying covariate effects, 186
- Sum scores, 139, 141
- Survey, 4, 5, 39, 71
 - auxiliary variables, 319
 - condition to eliminate bias in Rubin’s variance formula, 316
 - design and estimated weights, 318
 - domains, 71, 328
 - impute separately in strata, 313–314
 - linear estimators, 314, 319
 - logistic regression, 322
 - missing covariates, 320
 - missing values in weight model, 329
 - model based predictions, 313
 - parameters vary across strata, 313

- Survey (*Continued*)
two stage MI, 329
weights form strata for multilevel MI,
320–322
- Survival data, 72, 175, 176
as categorical, 180, 181
baseline hazard, 183
censoring mechanisms, 175
competing risks, 176
cumulative hazard as covariate,
176–181
grouped, 180
imputing censored, 179, 186–191
misclassified covariates, 297
missing covariates, 176–181
Nelson–Aalen estimator, 179, 180
normalising transformation, 179
simulation study, 178–179
time-varying covariates, 179
- Survivor function, 187
- Tobit model, 179
- Transform then impute, 143–145
- Transformation
normalising, 49, 93, 179
- Unblinding, 264
- Unit, 3, 5–6
- Wald tests, 55
- Weibull distribution, 178
- Wishart distribution, 85, 102, 226, 232,
325

