

## Index

- acceleration and braking, 291
  - accelerometers, 291
  - accident period, 61
  - actual versus forecast (AvF), 68
    - by calendar period, 72*f*
    - heat map, 72*f*
  - agglomerative nesting (AGNES), 165–166, 173
  - AIB. *See* Automobile Insurers Bureau of Massachusetts (AIB)
  - allocated loss adjustment expense (ALAE), 233
  - ankle contusions, 239
  - area under the ROC curve (AUC), 192, 298, 299*f*, 303, 304*t*, 306*t*
  - auto insurance, usage-based (UBI), 290–308
    - classification trees, 301–305
      - applied to DDVs from iterative stepwise approach, 304–305
      - applied to larger sets of DDVs, 302–303
      - Poisson regression applied to DDVs by, 303–304
    - data collectible by telematics, 291–292
    - data preparation, 292–294
    - future research, 306–307
    - machine learning for, 301
    - models
      - implementation with traditional rating plan, 305–306
      - objectives, 292–294
      - Poisson model, 294–301
      - predictive modeling, 290–308
  - overview, 290–291
  - Poisson model for, 294–301
    - data and model form, 294–295
    - holdout driving period, 297–301
    - loss ratio charts, 299–301
    - ROC curves, 298–299
    - validation, 297
    - variable selection, 295–296
  - predictive modeling for, 290–308
- Automobile Insurers Bureau of Massachusetts (AIB), 182–184, 201–202
  - bagging (bootstrap aggregation), 194
  - Bailey-Simon approach, 100
  - balanced iterative reducing and clustering using hierarchies (BIRTH), 165–166
  - Bayesian MCMC model, 208–209
  - bike age, motor collision insurance, 44
  - billing data, 269–271
  - BIRTH (balanced iterative reducing and clustering using hierarchies), 165–166
  - boosting (data mining), 193
  - bootstrap aggregation, 194
  - bumps, 291
  - Burr-gamma distribution, 224
  - calibration, 66
  - capital management, 216–223
  - CAR. *See* Commonwealth Automobile Reinsurers (CAR)
  - CART. *See* classification and regression trees (CART)
  - CAS Loss Reserve Database, 210
  - CCL. *See* correlated chain ladder (CCL)
  - chain ladder
    - defined, 61
    - disadvantages of, 73
  - claim adjuster notes, 267–269
  - claim escalation, 261–289
    - factor selection, 271–274
    - loss development models, 262–271
    - modeling method, 274–276
      - data format, 274, 275*t*
      - logistic regression, 275
      - practical considerations, 275–276
  - penalized regression, 280–289
    - elastic net, 285–288
    - extension to GLMs, 288

- claim escalation (*cont.*)
  - fitting the elastic net in R, 288–289
  - lasso, 284–285
  - ridge regression, 281–284
- research opportunities in, 277–280
  - elastic net, 279–280
  - individual claims to ultimate, 278
  - text mining, 278–279
- stages of modeling, 262
- triage models, 267–271
  - medical and billing data, 269–271
  - research opportunities, 279
  - text mining claim adjuster notes, 267–269
- claims
  - analytics, 261
  - count, 5*t*
  - frequency, 74–75, 103*t*, 304*t*, 306*t*
  - frequency modeling, 40
  - questionable, 185, 187
  - severity, 74–75
  - severity modeling, 40–41
    - advantages of, 43
    - conventional triangular, 76
  - triangle, 61
- classification and regression trees (CART), 193
  - applied to DDVs from iterative stepwise approach, 304–305
  - applied to larger sets of DDVs, 302–303
  - machine learning and, 301
  - Poisson regression applied to DDVs by, 303–304
- Clayton copula, 118
- CLIQUE method, 168
- clustering, 159–170
  - datasets, 161–163
  - dendrogram, 176*f*
  - density-based methods, 167
    - density-based clustering (DENCLUE), 167
    - density-based clustering of application with noise (DBSCAN), 167
    - ordering points to identify the clustering structure (OPTICS), 167
  - exposure-adjusted hybrid (EAH), 159, 168–171, 173–177
  - grid-based methods, 167–168
  - hierarchical methods, 165–166
    - agglomerative nesting (AGNES), 165–166, 173
    - balanced iterative reducing and clustering using hierarchies (BIRTH), 165–166
  - CHAMELEON, 166
  - clustering using representatives (CURE), 166
  - divisia analysis (DIANA), 165–166
  - kernel-based, 168
  - methods, 163–171
  - overview, 159
  - partitioning methods, 163–165
  - expectation maximization, 165
    - k*-means method, 164, 171–173, 174*f*
    - k*-medoids method, 164
  - purpose in insurance, 160–161
  - spectral, 168
- clustering using representatives (CURE), 166
- co-insurance, 139
- Commonwealth Automobile Reinsurers (CAR), 102
- compound symmetry-correlation metric form, 143
- conditional distribution, 120
- contusion injuries, 239
- copulas, 116, 118
- corporate governance, 291
- correlated chain ladder (CCL), 208, 212–213
- correlation analysis, 244–246
- Cost and Claim Counts model, 107
- Cost Only model, 107
- cost-sharing agreement, 137–139. *See also* group health insurance
  - co-insurance, 139
  - limit on amount, 137–138
  - limit on number, 138
  - types of, 137–139
- coverage per policy, 131
- CPT (Current Procedural Terminology) codes, 270–271
- cross-classified form, 62
- cross-validation for frequency, 22–23
- CURE (clustering using representatives), 166
- Current Procedural Terminology (CPT) codes, 270–271
- data mining, 192
- data visualization
  - Kohonen neural networks, 204–205
  - multidimensional scaling, 202–204
  - Random Forest, 202–204
- DBSCAN (density-based clustering of application with noise), 167
- DDVs. *See* driving data variables (DDVs)
- decision trees, 192–194
  - advantages of, 193
  - applied to DDVs from iterative stepwise approach, 304–305
  - applied to larger sets of DDVs, 302–303
  - machine learning and, 301
  - Poisson regression applied to DDVs by, 303–304
- degree of service, 135
- DENCLUE (density-based clustering). *See also* generalized linear models (GLM)
- density-based methods, 167
  - density-based clustering (DENCLUE), 167
  - density-based clustering of application with noise (DBSCAN), 167
  - ordering points to identify the clustering structure (OPTICS), 167
- dependence ratio, 115

- development period, 61  
 development year, 61  
 deviance residual, 65  
 DFA. *See* dynamic financial analysis (DFA)  
 DGLM. *See* double generalized linear model (DGLM)  
 DIANA (divisia analysis), 165–166  
 dissimilarity, 195  
 distribution analysis, 234–239  
 divisia analysis (DIANA), 165–166  
 double generalized linear model (DGLM), 230–231.  
   *See also* generalized linear models (GLM)  
   actuarial applications, 255–257  
   defined, 42  
   parameter estimation, 249*t*  
   projected large claim counts versus actual  
     observations, 252*t*  
   projected means of selected segmentations by, 251*t*  
   residual statistics by injury codes, 253*t*  
   residual statistics by predicted value for bias, 253*t*  
   Tweedie, 42–43  
   workers' compensation, 249–254  
 driver characteristic variables, 4  
 driving data variables (DDVs), 293, 294, 297, 300,  
   305–306  
   Poisson regression applied by tree, 303–304  
   tree applied to large set of, 302–303  
 dynamic financial analysis (DFA), 225, 229
- EAH clustering. *See* exposure-adjusted hybrid (EAH)  
 clustering  
 EDF (exponential dispersion family), 64–65  
 Egyptian Financial Supervisory Authority (EFSA), 130  
 Egyptian insurance market, 129–130  
 elastic net, 279–280, 285–288  
   fitting in R, 288–289  
 ensemble model, 193  
 enterprise risk management (ERM), 225  
 Euclidean clustering  
   application to suspicious data claims, 196–199  
   ranking, 200*t*  
 excess of loss (XOL) reinsurance, 230  
 expectation maximization, 165  
 explanatory variables, 103*t*, 245*t*, 263  
 exploratory data analysis, 101  
   for frequency, 6–11  
 exponential dispersion family (EDF), 64–65  
 exposure, 5*t*  
 exposure-adjusted hybrid (EAH) clustering, 159,  
   168–171, 173–177
- false positive rate (FPR), 298  
 fat-tailed distributions, 227–230  
 feature space, 168  
 finite mixture model (FMM), 231–232  
   actuarial applications, 255–257  
   projected large claim counts versus actual  
     observations, 252*t*  
   projected means of selected segmentations by, 251*t*  
   workers' compensation, 249–254  
 first notice of loss (FNOL), 232–233, 274, 276  
 FMM. *See* finite mixture model (FMM)  
 forecast error, 67  
 FPR (false positive rate), 298  
 Frank copula, 116, 118  
 fraud, 182–184  
 frequency, 5*t*  
   cross-validation for, 22–23  
   exploratory data analysis, 6–11  
   modeling, 12–23  
   multivariate, 115–117  
   multiway frequency models, 18–22  
   one-way frequency models, 13–17  
 frequency modeling, 40  
   advantages of, 43  
   motor collision insurance, 47–51  
   versus pure premium model, 54–55  
   splitting offset in, 45–46  
 frequency-severity model, 107–113
- gamma distribution, 39–40, 41, 51–52, 111, 234–239  
 gamma error structure, 101  
 generalized additive model (GAM), 79  
 generalized linear models (GLM), 273  
   actuarial applications, 255–257  
   double, 42–43, 230–231  
   elastic net, 279–280  
   extension to, 288  
   insurance data, 39–59  
   motor collision insurance, 39–59  
   as predictive claim models, 57–59  
   projected means of selected segmentations by,  
     251*t*  
   property and casualty (P&C) insurance, 64–68  
   in pure premium modeling, 39–40  
   ratemaking and, 101  
   regression analysis, 226  
   Tweedie distribution, 105–106  
   workers' compensation, 246–248, 249–254  
 Gini index, 35–36, 37*f*, 50, 53–54, 57–59  
 GLM. *See* generalized linear models (GLM)  
 Global Positioning system (GPS), 293  
 graph Laplacian matrices, 168  
 grid-based clustering, 167–168  
 group health insurance  
   data, 130–141  
   Egyptian market, 129–130  
   key variables, 133–141  
     benefit package/coverage, 134–135  
     cost-sharing agreement, 137–139  
     degree of service, 135–136

- group health insurance (*cont.*)
  - industrial classification, 139–140
  - type of service, 136–137
- model building strategy, 141–142
- models
  - comparing, 142–143
  - multiple-level models fitted to dataset, 151–157
  - single-level models fitted to dataset, 148–150
- multilevel modeling for, 128–129
- policies, 127–128
  - experience rating, 128
  - multidimensional nonstandardized benefit packages, 128
  - multilevel data, 127–128
  - panel/longitudinal aspects, 128
  - unbalanced records, 128
- ratemaking, 127
- Gumbel copula, 118, 120, 121*t*
- hard fraud, 182
- hierarchical insurance claims model, 118–120
- hierarchical methods, 165–166
  - agglomerative nesting (AGNES), 165–166
  - balanced iterative reducing and clustering using hierarchies (BIRTH), 165–166
  - CHAMELEON, 166
  - clustering using representatives (CURE), 166
  - divisia analysis (DIANA), 165–166
- High-level SNA/ISIC aggregation, 140
- holdout driving period, 297–301, 304*t*, 306*t*
- ICD (International Classification of Diseases) codes, 269–270
- incurred losses, motor collision insurance, 44
- industrial classification, 139–140
- inflation, 82–86
- injury codes, 232–233
  - correlation analysis, 244–246
  - generalized linear models (GLM), 246–248
  - loss statistics, 240*t*, 241*t*
  - univariate analysis, 239–244
- injury severity modeling, 75–76
- in-sample, 143, 144
- Insurance Fraud Bureau of Massachusetts, 183, 201
- insurance products, pricing of, 1–2
- International Classification of Diseases (ICD) codes, 269–270
- International Standard Industrial Classification of All Economic Activities (ISIC), 140
- inverse Gaussian distribution, 41, 52, 111, 234–239
- iterative stepwise approach, 304–305
- kernel-based clustering, 168
- k-fold cross-validation, 6
- k-means method, 164, 171–173, 174*f*
- k-medoids method, 164
- Kohonen neural networks, 204–205
- lagging indicators, 261
- large loss distribution, 224–225
- lasso, 284–285
- leading indicators, 261
- limit of insurance (LOI), 225
- linear predictor, 65
- link function, 65
- logistic regression, 49, 275
- lognormal distribution, 234–239
- Lorenz curve, 35–36, 37*f*
- loss ratio charts, 299–301
- loss reserve, 60
  - CAS Loss Reserve Database, 210
  - correlated chain ladder (CCL), 212–213
  - implications for capital management, 216–223
  - loss development patterns, 211*f*
  - predictive distribution of estimates, 213–216
- loss reserving
  - defined, 60
  - forecasting, 62–63
  - modeling, 62–63
  - notation, 61–62
- lower back contusions, 239
- machine learning, 301
- Mack model, 62, 212
- manufacturer suggested retail price, motor collision insurance, 44
- Markov chain Monte Carlo (MCMC) models, 208–209
- Markov chain of transitions, 75
- Markov transition matrix, 278
- MDSplot function, 202
- mean absolute error (MAE), 145
- medical and billing data, 269–271
  - Current Procedural Terminology (CPT) codes, 270–271
  - International Classification of Diseases (ICD) codes, 269–270
  - prescription drugs, 271
- merging points, 173
- model building strategy, 141–142
- motor collision insurance, 39–59
  - dataset, 44–47
  - parameter estimates, 46
  - splitting offset in frequency/severity approach, 45–46
- frequency models, 47–51
- frequency/severity versus pure premium, 54–55
- generalized linear models, 39–59
- pure premium models, 52–54
- severity models, 51–52

- multidimensional scaling, 202–204
- multilevel modeling, 141*t*
- multinomial logit model, 121*t*
- multivariate frequency, 115–117
- multivariate ratemaking, 113–120
  - hierarchical insurance claims model, 118–120
  - model comparisons, 122–123
  - multivariate frequency, 115–117
  - multivariate severity, 117–118, 119*t*
  - two-part model, 114–118
- multivariate severity, 117–118, 119*t*
- multiway frequency models, 18–22
- multiway severity models, 29–30
  
- National Association of Insurance Commissioners (NAIC) annual, 208
- National Council on Compensation Insurance (NCCI), 233
- negative binomial distribution, 40, 48
- negative binomial regression, 110
- nodes, 301
- nonlinearity, 67
- nonnormal observations, 68
  
- odds ratio, 115
- OLS. *See* ordinary least squares (OLS)
- one-way frequency models, 13–17
- one-way severity models, 25–29
- operational time, 77–82
  - warped, 92–97
- OPTICS (ordering points to identify the clustering structure), 167
- ordinary least squares (OLS)
  - linear model, 111
  - penalized regression, 280–281
- out-of-sample, 143, 144–146
- overdispersed Poisson, 62–63
  
- Pareto distribution, 228–229, 234–239
- partitioning methods, clustering, 163–165
  - expectation maximization, 165
  - k*-means method, 164, 171–173, 174*f*
  - k*-medoids method, 164
- payments per claim incurred (PPCI), 77
- PCA. *See* principal component analysis (PCA)
- Pearson correlation, 244–245
- Pearson residual, 65
- penalized regression, 280–289
  - elastic net, 285–288
  - extension to GLMs, 288
  - fitting the elastic net in R, 288–289
  - lasso, 284–285
  - ridge regression, 281–284
- personal injury protection (PIP), 102, 104, 182–185
  
- Poisson distribution
  - frequency models and, 47–48, 109–110
  - in loss development models, 265
  - Tweedie distribution, compared with, 39–40
- Poisson error structure, 101, 294
- Poisson model, 13, 294–295
- Poisson regression
  - applied to DDVs by tree, 303–304
  - frequency modeling and, 109
  - pure premium ratemaking and, 101
  - in usage-based auto insurance, 294–295
- policy year method, 131
- prediction error, 67
- predictive analytics, 261
- predictive distribution of estimates, 213–216
- predictive modeling in claim escalation, 261–289
  - factor selection, 271–274
  - loss development models, 262–271
  - modeling method, 274–276
    - data format, 274, 275*t*
    - logistic regression, 275
    - practical considerations, 275–276
  - penalized regression, 280–289
    - elastic net, 285–288
    - extension to GLMs, 288
    - fitting the elastic net in R, 288–289
    - lasso, 284–285
    - ridge regression, 281–284
  - stages of modeling, 262
  - triage models, 267–271
    - medical and billing data, 269–271
    - research opportunities, 279
    - text mining claim adjuster notes, 267–269
- prescription drugs, 271
- PRIDIT (Principal Components of RIDITS), 185–186
  - computing, 187–188
  - overview, 180–181
  - processing questionable claims for, 187–188
  - questionable claims data processing, 187
  - ranking, 200*t*
  - score, 189–192, 201–202
  - scree plot, 188*f*
- principal component analysis (PCA), 301
- prior weights, 265–266
- probability density function, 213
- property and casualty (P&C) insurance, 60–99
  - claim severity modeling, 76
  - datasets, 73–74
    - claim frequency and severity, 74–75
    - legislative change, 86–92
    - operational time, 77–82
    - row and diagonal effects, 82–92
    - superimposed inflation, 82–86
    - warped operational time, 92–97

- property and casualty (P&C) insurance (*cont.*)
- diagnostics, 68–72
    - distributional assumptions, 68–69
    - goodness-of-fit, 71–72
    - residual, 69–70
  - generalized linear model, 64–68
    - advantages of, 66–68
    - calibration, 66
    - nonlinearity, 67
    - prediction error, 67
  - generalized linear models
    - emerging claims experience, 68
    - nonnormality, 67–68
  - injury severity modeling, 75–76
  - loss reserving, 60–63
    - data and notation, 61–62
    - forecasting, 62–63
    - modeling, 62–63
    - notation, 63–64
  - Protection Class, 225–226
  - proximity, 195
  - “pruning” algorithms, 301
  - pure premium modeling, 41–43
    - advantages of, 43–44
    - dependent/outcome variable, 133
    - exploratory data analysis, 6–12
    - with exposures as weight, 56
    - frequency modeling, 12–23
      - cross-validation for frequency, 22–23
      - versus frequency/severity models, 54–55
    - multiway, 18–22
    - one-way, 13–17
  - generalized linear models in, 39–40
  - model dollars of loss with exposures<sup>(*p*–1)</sup> as weight, 57
  - motor collision insurance, 53–54
  - proof of equivalence, 55–57
  - pure premium formula, 30–34
  - severity modeling, 23–30
    - multiway, 29–30
    - one-way, 25–29
  - specifications, 41
  - Tweedie distribution, 41–43
  - validation, 34–36
- p*-values, 297
- questionable claims, 185, 187
- R code, 187
- Random Forest
  - application to suspicious data claims, 195–199
  - decision tree and, 192–194
  - overview, 180–181
  - ranking, 200<sup>t</sup>
  - software, 195–199
  - unsupervised learning with, 194–195
  - visualization via multidimensional scaling, 202–204
- ratemaking, 100–125
  - Bailey-Simon approach, 100
  - correlation analysis, 123
  - defined, 100
  - generalized linear models, 101
  - group health insurance, 127
  - multivariate, 113–120
  - percentiles of claim size by type and, 103<sup>t</sup>
  - scatterplot of predicted scores, 122–123
  - univariate, 104–113
    - frequency-severity model, 107–113
    - Tweedie model, 105–107
- Receiver Operating Characteristics (ROC) curve, 192
- regression analysis
  - with fat-tailed distributions, 227–230
  - on severity, 225–226
  - workers’ compensation, 246–257
    - double generalized linear model (DGLM), 249–254
    - with fat-tailed distributions, 227–230
    - finite mixture model (FMM), 249–254
    - generalized linear models (GLM), 246–248
    - on severity, 225–226
- regularization term, 280
- residual diagnostics, 69–70
- restricted maximum likelihood (REML), 142
- ridge regression, 281–284
- RIDIT, 185–186
- ROC curves, 298–299
- SAS Enterprise Miner, 302
- saturated model, 65
- scree plot, 188
- self-organizing feature maps, 204–205
- service
  - degree of, 135–136
  - type of, 136–137, 138<sup>f</sup>
- severity, 5<sup>t</sup>
  - models, 23–30, 40–41
  - multivariate, 117–118
  - regression analysis on, 225–226
- severity, exploratory data analysis, 11–12
- shrinkage, 280
- similarity matrix, 168
- similarity measure, 160
- SIU (special investigation unit), 185
- soft fraud, 182
- Solvency II, 209–210
- special investigation unit (SIU), 185
- spectral clustering, 168
- sprain injuries, 239
- spread plot, 70
- STING method, 168

- superimposed inflation, 82–86
- suspicious claims, 195–199
- tail forecasting, 267
- telematics, 291–292, 306
- territory loss cost, motor collision insurance, 44
- text kernels, 168
- text mining, 267–269, 278–279
- third party liability, 102
- traditional rating plans, 305–306
- trees, 193
  - applied to DDVs from iterative stepwise approach, 304–305
  - applied to larger sets of DDVs, 302–303
  - machine learning and, 301
  - Poisson regression applied to DDVs by, 303–304
- true positive rate (TPR), 298
- Tweedie distribution
  - defined, 39–40
  - estimation, 108
  - generalized linear model, 105–106
  - in loss development models, 265
  - pure premium models and, 41–42, 52–53
  - univariate ratemaking and, 105–107
- Tweedie double GLM, 53
- underwriting year method, 131
- univariate ratemaking, 104–113
  - frequency, 109–111, 112*t*
  - frequency-severity model, 107–113
  - severity, 111–113, 113*t*
  - Tweedie model, 105–107
- unsupervised learning, 181–182
  - with Random Forest, 194–195
- usage-based auto insurance (UBI), 290–308
  - classification trees, 301–305
    - applied to DDVs from iterative stepwise approach, 304–305
    - applied to larger sets of DDVs, 302–303
    - Poisson regression applied to DDVs by, 303–304
  - data collectible by telematics, 291–292
    - complexity, 292
    - depth, 292
    - dimensionality, 292
    - overlap, 292
  - data preparation, 292–294
  - future research, 306–307
  - machine learning for, 301
  - models
    - implementation with traditional rating plan, 305–306
    - objectives, 292–294
    - Poisson model, 294–301
    - predictive modeling, 290–308
  - overview, 290–291
  - “pay as you drive”, 307
  - Poisson model for, 294–301
    - data and model form, 294–295
    - holdout driving period, 297–301
    - loss ratio charts, 299–301
    - ROC curves, 298–299
    - validation, 297
    - variable selection, 295–296
  - predictive modeling for, 290–308
- vehicle characteristic variables, 4
- visualization, data, 202–204
  - Kohonen neural networks, 204–205
  - multidimensional scaling, 202–204
  - Random Forest, 202–204
- warped operational time, 92–97
- WaveCluster method, 168
- Weibull distribution, 234–239
- winter temperature, motor collision insurance, 44
- workers’ compensation, 224–225
  - actuarial applications, 254–257
  - correlation analysis, 244–246
  - data, 232–234
  - double generalized linear model (DGLM), 230–231
  - finite mixture model (FMM), 231–232
  - large loss distribution analysis, 224–225
  - regression analysis, 246–257
    - double generalized linear model (DGLM), 249–254
    - with fat-tailed distributions, 227–230
    - finite mixture model (FMM), 249–254
    - generalized linear models (GLM), 246–248
    - on severity, 225–226
  - traditional distribution analysis, 234–239
  - univariate analysis, 239–244
- World Health Organization (WHO), 129
- XOL (excess of loss) insurance, 230
- zero-inflated negative binomial model (ZINB), 110
- zero-inflated Poisson model (ZIP), 110
- z-score, 297