|  |  |  |  |
| --- | --- | --- | --- |
| Method | Categorical Response | Quantitative Response | Notes |
| Linear regression | No | Yes | Simple lin reg—one predictor  Multiple lin reg—multiple predictors (use y ~ . or y ~ x1 + x2) |
| Logistic regression | Yes (binary only, unless you use multinomial logistic regression or ordinal logistic regression) | No |  |
| Robust regression | No (in this class) | Yes | Downweights outliers  Tukey’s bisquare, Huber, Hampel  rlm |
| Other Iteratively Weighted Least Squares | No (in this class) | Yes | Dealing with heteroskedasticity: Weight = 1/predicted(y), others |
| Regression with Autocorrelation | No | Yes | Good for time series  Plot data versus lagged data  acf( ) plot |
| Regression with compound symmetry correlation | No | Yes | Good for grouped data, such as different states |
| KNN | Yes (knn) | Yes (knn.reg) | Good for communicating with non-data scientists  Good for modeling interactions, non-linear relationships, but can be hard to see what those relationships are  Typically best for p << n (p = number of predictors, n = number of rows) |
| Penalized regression: Ridge regression | No (in this class) | Yes | Doesn’t set coefficients = 0  Reduce variance/overfitting by encouraging small coefficients  Outliers can make overfitting more of a problem  Multicollinearity is a source of model variance |
| Penalized regression: LASSO | No (in this class) | Yes | Sets coefficients = 0  Reduce variance/overfitting by encouraging small coefficients  Outliers can make overfitting more of a problem  Multicollinearity is a source of model variance |
| Penalized regression: Elastic net | No (in this class) | Yes | Ranges from Ridge (alpha = 0) to LASSO (alpha = 1)  Ridge and LASSO are special cases of Elastic Net  Reduce variance/overfitting by encouraging small coefficients  Outliers can make overfitting more of a problem  Multicollinearity is a source of model variance |
| Single Decision tree | Yes | Yes | Good for communicating with non-data scientists  Good for modeling interactions, non-linear relationships |
| Random Forests | Yes | Yes | Mtry = p (all predictors) gives classic bagging  Good for modeling interactions, non-linear relationships |
| Boosting (gbm or XGBoost) | Yes | Yes | Good for modeling interactions, non-linear relationships  XGBoost is often faster than other methods (random forest, ANNs, SVM) for a large dataset |
| ANN | Yes | Yes | Good for modeling complexities of large datasets, but implementation in R is not especially fast |
| LDA | Yes | No | Extension of Naïve Bayes  2 assumptions: Predictors are multivariate normal within each group (formed by categories of the response), and Covariance matrices of predictors (1 matrix per category of response) are equal |
| QDA | Yes | No | 1 assumption: Predictors are multivariate normal within each group (formed by categories of the response). Covariance matrices can be equal or not equal.  Expect to do better than LDA if there are a lot of rows of data, relative to the number of parameters (which depends on number of predictors and categories) |
| SVM | Yes | No (in this class) | Curved separator  svmRadial or svmPoly(?)  Designed to work when p is large, possibly > n  Implementation in R is slow |
| SVC | Yes | No (in this class) | Linear separator  svmLinear |