**PREDICTING PAROLE VIOLATORS**

In many criminal justice systems around the world, inmates deemed not to be a threat to society are released from prison under the parole system prior to completing their sentence. They are still considered to be serving their sentence while on parole, and they can be returned to prison if they violate the terms of their parole.

Parole boards are charged with identifying which inmates are good candidates for release on parole. They seek to release inmates who will not commit additional crimes after release. In this problem, we will build and validate a model that predicts if an inmate will violate the terms of his or her parole. Such a model could be useful to a parole board when deciding to approve or deny an application for parole.

For this prediction task, we will use data from the United States 2004 National Corrections Reporting Program (http://www.icpsr.umich.edu/icpsrweb/NACJD/series/38/studies/26521?archive=NACJD&sortBy=7), a nationwide census of parole releases that occurred during 2004. We limited our focus to parolees who served no more than 6 months in prison and whose maximum sentence for all charges did not exceed 18 months. The dataset contains all such parolees who either successfully completed the term of parole during 2004 or those who violated the terms of their parole during that year. The dataset contains the following variables:

* *male*: 1 if the parolee is male, 0 if female
* *race*: 1 if the parolee is white, 2 otherwise
* *age*: the parolee's age in years at release from prison
* *state*: a code for the parolee's state. 2 is Kentucky, 3 is Louisiana, 4 is Virginia, and 1 is any other state. The three states were selected due to having a high representation in the dataset.
* *time.served*: the number of months the parolee served in prison (limited by the inclusion criteria to not exceed 6 months).
* max.sentence: the maximum sentence length for all charges, in months (limited by the inclusion criteria to not exceed 18 months).
* *multiple.offenses*: 1 if the parolee was incarcerated for multiple offenses, 0 otherwise.
* *crime*: a code for the parolee's main crime leading to incarceration. 2 is larceny, 3 is drug-related crime, 4 is driving-related crime, and 1 is any other crime.
* *violator*: 1 if the parolee violated the parole, and 0 if the parolee completed the parole without violation.

UNDERSTANDING AND PREPARING THE DATASET

* How many parolees are contained in the dataset?
* How many of the parolees in the dataset violated the terms of their parole?
* Which variables in this dataset are unordered factors with at least three levels?
* Are there any variables that should be converted into factors before further analysis?
* Why are we taking this step of preparing the variables before splitting the data into a training and testing set?
* Now split the data into training and testing set

BUILDING THE MODEL

* Run the Logistics Regression Model and interpret the output.
* What can we say based on the coefficient of the multiple.offenses variable?
* Consider a parolee who is male, of white race, aged 50 years at prison release, from the state of Maryland, served 3 months, had a maximum sentence of 12 months, did not commit multiple offenses, and committed a larceny. Answer the following questions based on the model's predictions for this individual. (HINT: You should use the coefficients of your model, the Logistic Response Function, and the Odds equation to solve this problem.)
* Report the Classification Matrix. What is the sensitivity, specificity and accuracy of the Model?
* Our goal has been to predict the outcome of a parole decision, and we used a publicly available dataset of parole releases for predictions. In this final problem, we'll evaluate a potential source of bias associated with our analysis. It is always important to evaluate a dataset for possible sources of bias.
* The dataset contains all individuals released from parole in 2004, either due to completing their parole term or violating the terms of their parole. However, it does not contain parolees who neither violated their parole nor completed their term in 2004, causing non-violators to be underrepresented. This is called "selection bias" or "selecting on the dependent variable," because only a subset of all relevant parolees were included in our analysis, based on our dependent variable in this analysis (parole violation). How could we improve our dataset to best address selection bias?
  + There is no way to address this form of biasing.
  + We should use the current dataset, expanded to include the missing parolees. Each added parolee should be labeled with violator=0, because they have not yet had a violation.
  + We should use the current dataset, expanded to include the missing parolees. Each added parolee should be labeled with violator=NA, because the true outcome has not been observed for these individuals.
  + We should use a dataset tracking a group of parolees from the start of their parole until either they violated
  + parole or they completed their term.