Who's likely to get into trouble?

A statistical modelling of parole violation

Preet Khowaja

12/7/2021

Summary

In this study, I analyzed the effect of various demographic factors on the likelihood to violate parole. I used a logistic regression model to see what factors affects the likelihood of an inmate violating parole terms. My model deems state and race of the parolee to be significant predictors and deems that the effect of age varies by type of crime committed. Age is not a significant factor in affecting likelihood of violating parole according to my model.

Introduction

The United States incarceration system was introduced to the concept of parole in the late 1800s, with the main purpose being avoiding overcrowding in prisons. Over time, the advantages of the parole system were realized. It assisted those who were capable of integration into society with a chance to rectify their mistakes. The parole system in the U.S. decrees that those prisoners who demonstrate good behaviour are allowed to be released before their jail time is over, under supervision of the parole board. There are restrictions on them during this time and if they do commit a crime, they are imprisoned or the terms of their parole are tightened.

There are many factors which could explain why parolees violate their parole, on an individual case level. If there are general factors that do seem to impact the likelihood of an inmate to commit a crime again, such as state, age or the type of crime, taking a closer look might reveal dysfunctionalities in the system of parole, indicating that changes need to be made so that the primary motivator for prisoners to violate is not caused by the way parole is structured to advantage some groups over others. For this reason, modelling the violation of parole is a worthwhile and interesting endeavour.

In this study, I am interested in three main questions of analysis about parole violation:

- Does inmate age affect whether or not they will violate parole?
- Does the state the prisoner is in affect how likely a parolee is to violate the terms of their parole?
- Does the effect of age vary by crime?

Data

The data I obtained is from the National Archive of Criminal Justice Data (NACJD) website and includes data on prison inmates who either violated or did not violate their parole in 2004. The dataset is limited to prisoners who have spent less than 6 months in prison already and were not sentenced to more than 18 months in prison for their crime.

The predictor variables in this dataset are sex, race, age, state, time.served, max.sentence. multiple.offenses, and crime. The response variable, violator, is binary and encodes whether or not a prisoner violated their parole in 2004. The data dictionary is as follows:

Variable	Description
violator	1 if prisoner violated parole; 0 otherwise
sex	1 if the parolee is male, 0 if female
age	the parolee's age (in years) when he or she was released from prison
race	1 if the parolee is white, 2 otherwise
state	a code for the parolee's state. 2 is Kentucky, 3 is Louisiana, 4 is Virginia, and 1 is any other state
time.served	the number of months the parolee served in prison
max.sentence	the maximum sentence length for all charges, in months (less than 18 for all)
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

Exploratory Data Analysis

The dataset has information on 675 parolees who were observed in 2004. The states they mainly belonged to are Kentucky, Louisiana and Virginia, which is why the state variable focuses on them. The mean time served in prison by these 675 individuals is 4.19 months and the mean age of these prisoners is 35 years. I also looked at the spread of the data by the predictor variables and found that 18.4% of the inmates were female and 81.6% were male. Moreover, 42.4% of them were white and 57.6% were not. The distribution of age can be seen in the histogram below, which is interestingly fairly normal.

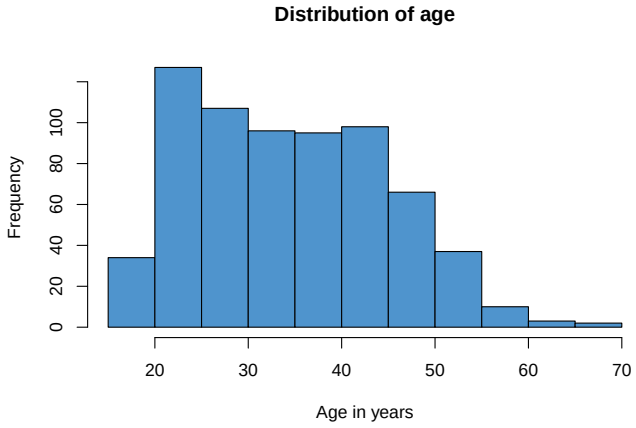

Figure 1: Distribution of parolee's ages in data

While performing EDA, I explored the graphic relationship between continuous predictors (age, time.served and max.sentence) and the response variable (violator). Along with this, I used Pearson's Chi-squared test to check for independence between the categorical variables (race, sex, state, multiple.offenses, crime) and the response variable.

Generating boxplots for the continuous predictors revealed no clear relationship between age and violator. The median age for those who violated and those who did not is approximately the same. In comparison, the median time served in prison for violators is slightly lower than for non-violators. This could indicate a statistically significant relationship and so I decided to include time served in any model I use. Since age is also a predictor I am interested in, I didn't completely write it off and held onto it for model selection. The boxplots for maximum sentence show that non-violators have a higher maximum sentence on average than

violators. These boxplots are included in Appendix A of the report, for reference. I also used binned plots for continuous predictors but did not find any noticeable trends in them that would indicate a relationship between these predictors and violator.

For the categorical predictors, I performed chi-squared tests. Race, sex and type of crime had large p-values indicating that there is little evidence for them being related to violator. Multiple offenses was significant at 99% level. State had a very small p-value indicating that there is evidence it is related to violator. Hence, this is also a variable I want to explore in future models.

After exploring interactions between predictors, my most interesting finding was the interaction between age and the type of crime committed. For those arrested for crimes of larceny (theft), older inmates were more likely to violate parole than younger inmates. In comparison, for drug and driving related crimes, older inmates were **less** likely to violate parole than older inmates. This is interesting because the relationship between age and violator seems to significantly change by the type of crime (see Figure below). I decided to keep an eye out for this interaction in my model to see whether it was statistically significant. One possible insight into this interaction could be that theft, by nature, is a different type of crime. There is perhaps something unique to the nature of this crime that causes even older parolees to violate the terms of their parole.

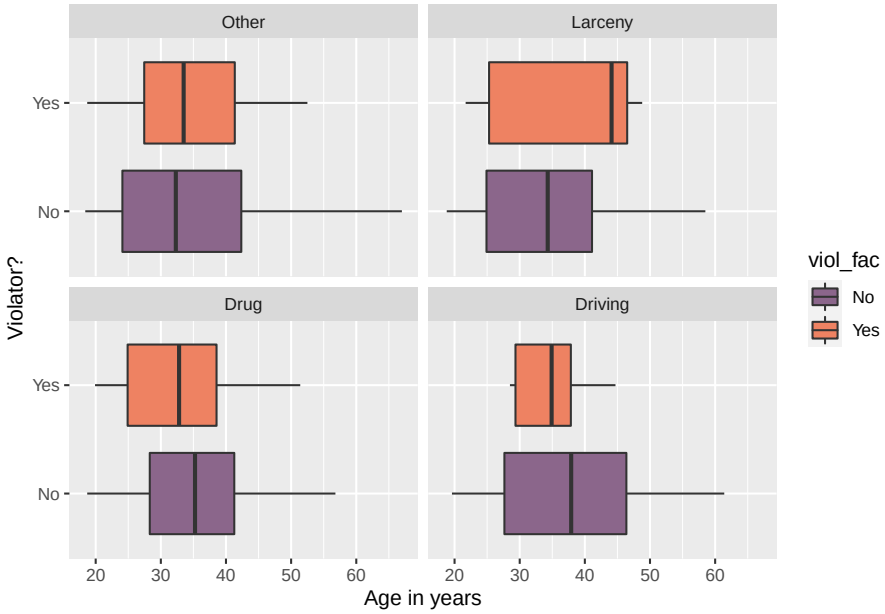

Figure 2: Interaction plot between age and type of crime

Other interactions were explored but none were particularly interesting. The interactions between gender and other predictors were also not reliable since most of the inmates in our dataset are men. At the conclusion of my EDA, I have state, age and the interaction between age and crime as some predictor terms to look out for as I move onto modelling.

Modelling

Final Model

The final model I use is a logistic regression with violator as a binary response as follows:

$$violator_i | x_i \sim \text{Bin}(n_i, \pi_i);$$

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 age_i + \beta_2 state_i + \beta_3 race_i + \beta_4 crime_i +$$
$$\beta_5 time.served_i + \beta_6 age_i : crime_i$$

Model Selection Process

Before any kind of modelling, I centered the continuous predictors in order to aid in the interpretation of coefficients that were output from models I created. I then created a null model which included only the predictors age, type of crime and the time served in prison before parole was granted since these are variables I'm interested in and necessarily want them in my model. My full model included all the variables and interactions possible. I then used AIC and BIC stepwise selection methods on these null and full models. AIC Stepwise gave me the final model I have mentioned in the previous sub-section whereas BIC stepwise gave me the same model without the predictors race and the interaction between age and type of crime. I used a chi-squared test to see if these two predictors were significant and they were. Hence, my final model is the AIC stepwise model.

Model Evaluation

In order to assess how my model stands on the assumptions of linear regression I looked at the binned residual plots to ensure there was random-ness and the points were captured within the confidence bands. I noticed that there were quite a few points outside the confidence bands and so I tried two transformations on the response variable including log() and square root. However, neither fixed the problem so I chose to work with the model without transformations. The binned residual plot is below.

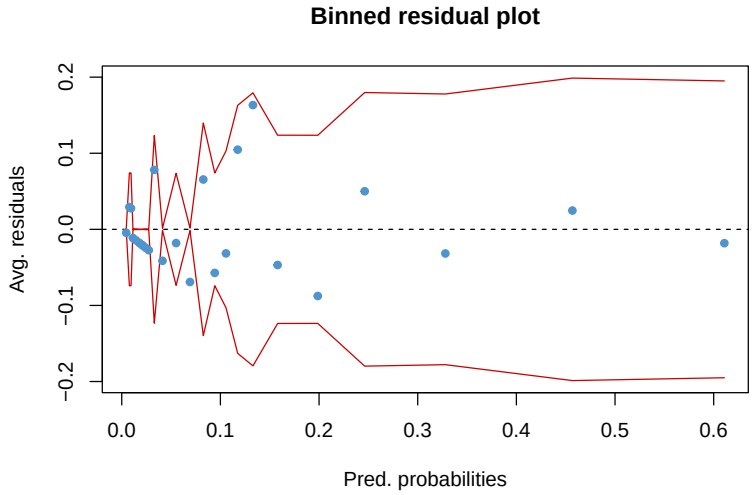

Figure 3: Binned Residual Plot of Final Model

I also plotted each binned centered continuous predictor in our model against average residuals to ensure there were no visible trends. The binned residual plots for age and time served in prison had randomness and displayed no visible trends. Furthermore, the points were mostly inside the confidence bands. These plots are in Appendix B.

With my current model, I am able to achieve a specificity of 82% and sensitivity of 76% and plotting the ROC Curves gives us an AUC of 0.85 which means our model does quite well on predicting whether or not a parolee will violate. The ROC curve is included in Appendix C.

Results

The table below shows the results of our fitted final model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8702	0.2951	-6.34	0.0000
ageC	0.0169	0.0177	0.95	0.3397
crime2	0.2826	0.4127	0.68	0.4934
crime3	-0.3749	0.3443	-1.09	0.2762
crime4	0.0257	0.4816	0.05	0.9574
time.servedC	-0.1522	0.0950	-1.60	0.1092
state2	-0.0723	0.3978	-0.18	0.8557
state3	1.1859	0.3818	3.11	0.0019
state4	-2.6801	0.5054	-5.30	0.0000
race2	0.7937	0.3119	2.54	0.0109
ageC:crime2	0.0578	0.0380	1.52	0.1281
ageC:crime3	-0.0466	0.0353	-1.32	0.1875
ageC:crime4	-0.0901	0.0456	-1.98	0.0482

The significant coefficients in my model are those for state and race as well as the interaction term between age and driving related crimes. The intercept of my model deems that at average age and time served, a non-white parolee in an 'other' state who has not committed a drug, driving or theft related crime has odds of 0.15 of violating their parole.

According to this model, a parolee with the baseline characteristics described above, is 3.27 times more likely to violate parole if they are in Louisiana as compared to being in the 'other' state category and 0.07 times as likely to violate if they are in Virginia. The coefficient for Kentucky is not significant in my model. Moreover, a parolee with baseline characteristics is less likely to violate parole as age increases if they have committed driving crimes as compared to other crimes. White parolees at the baseline are twice as likely to violate as non-white parolees. With this, we've looked at all the coefficients that are significant in our model.

Conclusions

Conclusively, we look at what my model says about the questions of interest. Age in my model, is not a significant predictor and the coefficient suggests that the odds of violating do not change with age. In comparison the state of the parolee is very significant. this suggests that rather than age demographics it is the location of the prisoner that influences their likelihood to violate parole. The details of parole law vary by state and could explain this significance. Finally, there is evidence that while the effect of age on its own is not significant, it varies by the type of crime committed and driving related crimes reduce the effect age have on likelihood to violate.

It is important to note that there are limitations in this study, primarily due to the limitedness of data. The study is also limited to 2004 and a particular subset of prisoners (those who have served less than 6 months in prison already and not been sentenced to more than 18 months) which makes the conclusions difficult to generalize. Furthermore, parole is a complex legal process and there are many factors which are not included in this dataset such as the terms of parole and employments status of parolees.

Appendix A

Figure 4: Age vs Violator

Figure 5: Time Served vs Violator

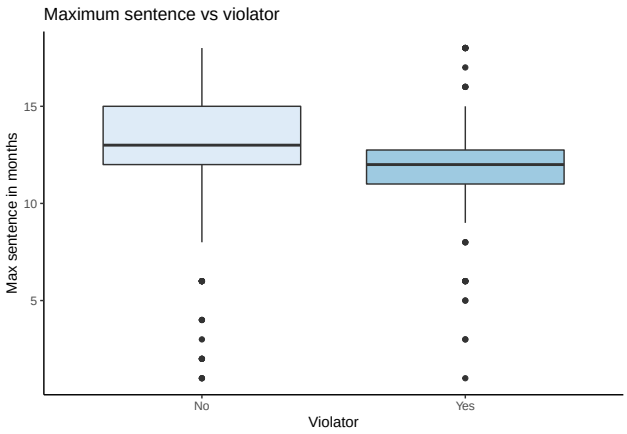

Figure 6: Maximum Sentence vs Violator

Appendix B

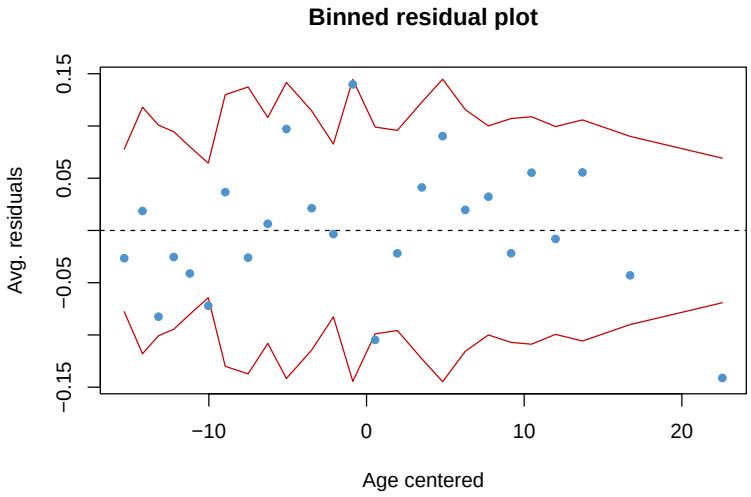

Figure 7: Binned Residual Plot of Predictors

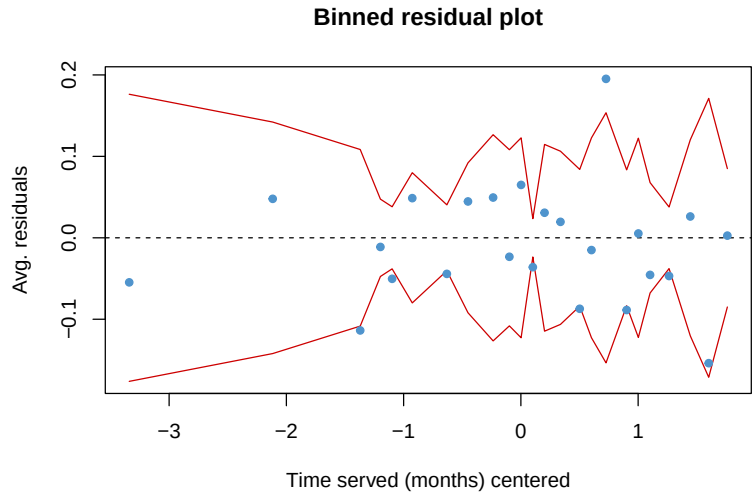

Figure 8: Binned Residual Plot of Predictors

Appendix C

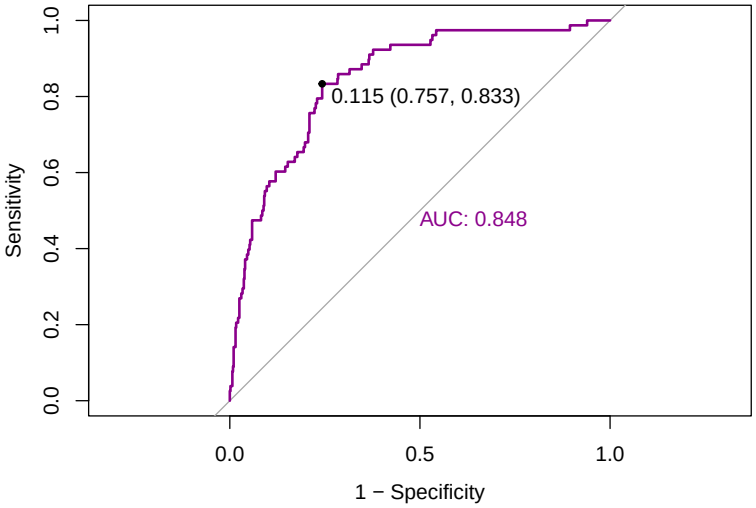

Figure 9: ROC for Final Model