3. Methods

3.1 Data Description

The dataset we worked on contains detailed information on hate crimes happened in 51 states in United States in 2016. This dataset was adapted from the one used by a FiveThirtyEight article for analyzing the same topic. Our dataset containing 51 rows, corresponding to 51 states, and 9 columns. Continuous variables include: hate\_crimes\_per\_100k\_splc (hate crime rate per 100,000 population), median\_household\_income (median household income per state), perc\_population\_with\_high\_school\_degree (percentage of adults (>25 yrs.) with a high school degree), perc\_non\_citizen (percentage of population that are not US citizens), perc\_non\_white (percentage of population that are non-white) and gini\_index (index measuring income inequality). Categorical variables are: unemployment (level of state unemployment) and urbanization (level of state urbanization).

The distribution of hate crime rate is significantly right skewed (Fig. 1), which is also indicated by the Q-Q plot (Fig. 2). After taking the log transformation, the distribution is much more normal now (Fig. 3 & Fig. 4). In this way, we decided to take the log of the hate crime rate as our outcome in models.

Noticing that in District of Columbia, the hate crime rate very high compared to others, we made plot of residuals vs leverage to find out whether it’s an influential case (Fig. 5). As the plot shows, case 9, which represents record for District of Columbia, is close to the dashed lines. Moreover, before removing case 9, there was supposed to be positive linear relationship between Gini index and hate crime rate (Fig. 6), while there was slightly negative association between Gini index and hate crime rate (Fig. 7). Based on analyses above, we considered the case of District of Columbia as an influential outlier and omit it when conducting further study.

Another issues to be addressed is the intercorrelation between our potential predictors. Based on correlation matrix, there were strong correlation between perc\_non\_white and perc\_non\_citizen (*ρ* = .73) and moderate correlation between urbanization and perc\_non\_citizen (*ρ* = .67), perc\_population\_with\_high\_school\_degree and median\_household\_income (*ρ* = .66), as well as gini\_index and perc\_population\_with\_high\_school\_degree (*ρ* = -.66). Actually, none of VIF value of the predictors exceeded 5. The results of ANOVA test indicated that adding perc\_population\_with\_high\_school\_degree, which had the highest VIF (4.41) was not redundant (p\_value = 0.148) under the criteria of α = 0.15 and improved Adj-R^2 from 0.027 to 0.057. However, the same procedure implied that perc\_non\_citizen, which was highly correlated with perc\_non\_white, should not be considered in our models.

3.2 Predictor Selection

To begin with, we used automatic procedure to identify different best models of different sizes and evaluate them base on multiple criteria: C\_p, Adj R^2 and BIC (Fig. 8). According to these plots, we though that the performance of model with two, three and four predictors were similar and decided to further evaluate them with cross validation (Fig. 9). In this plot, we can see that RMSE of the model with two predictors was a little smaller than other two candidates, which indicated better fitness. With the rule of parsimony, we finally chose the model with 2 predictors( ln\_ hate\_crimes\_per\_100k\_splc = gini\_index + perc\_population\_with\_high\_school\_degree) as our basic model.

Appendix

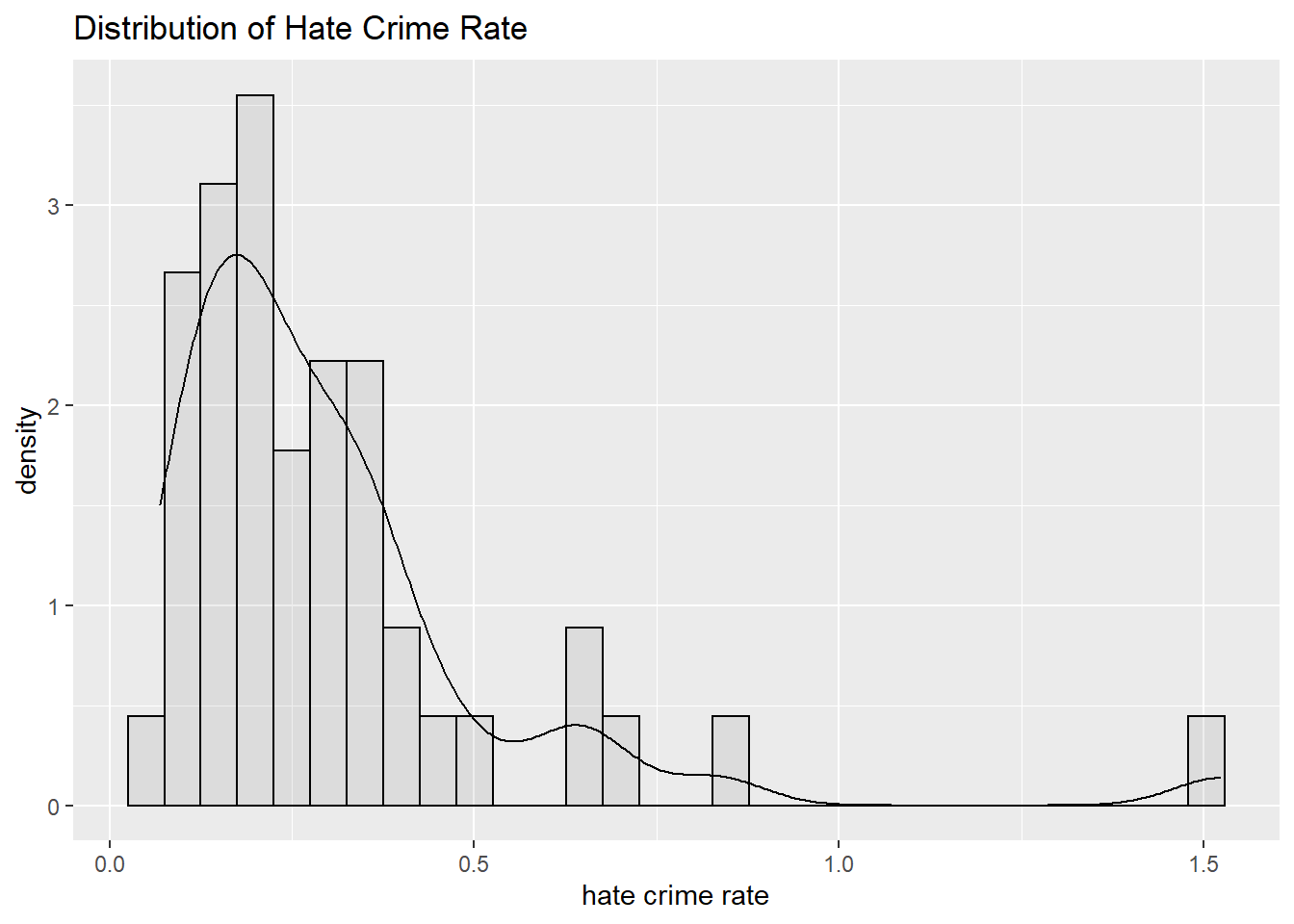


Fig. 1

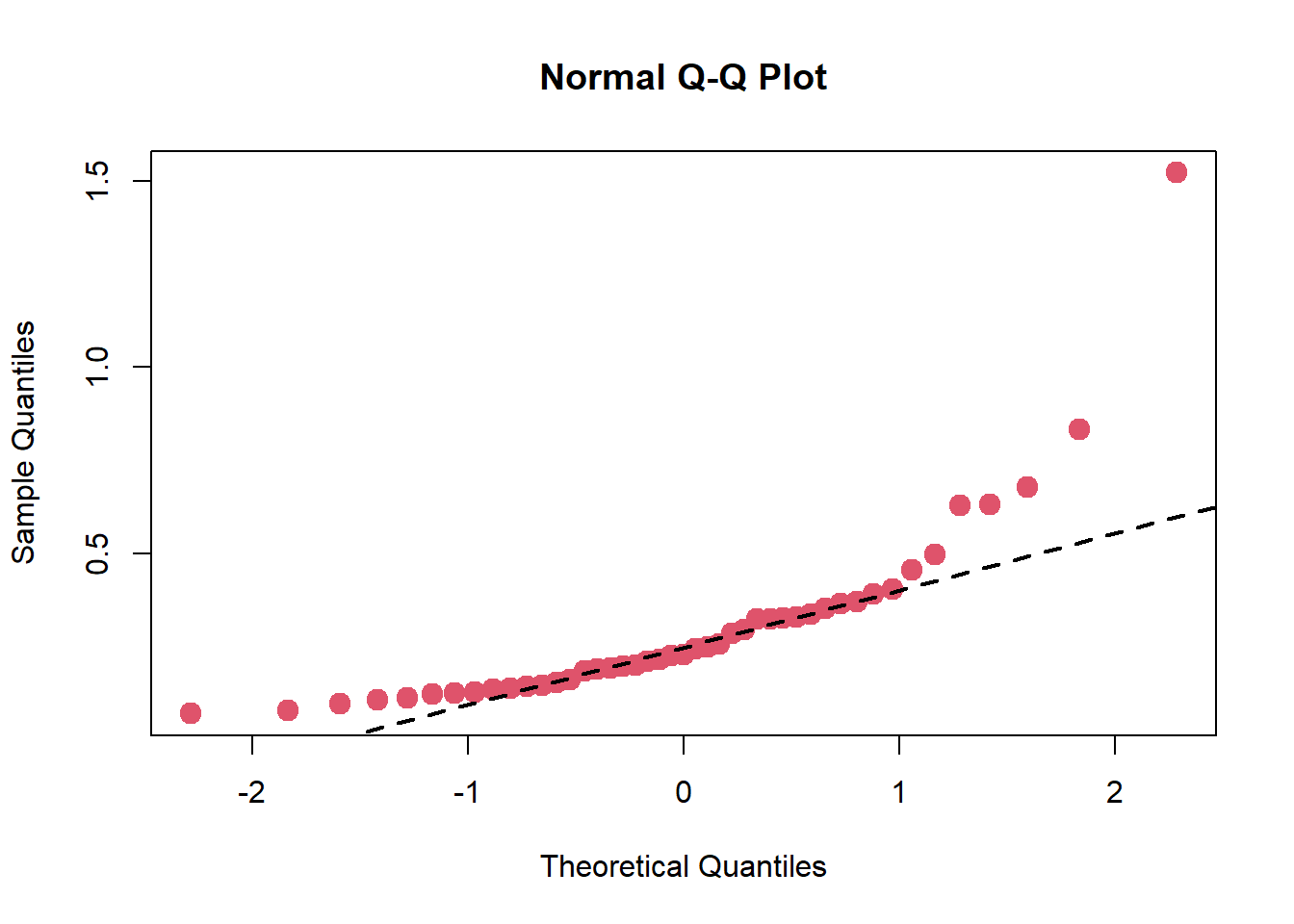


Fig. 2

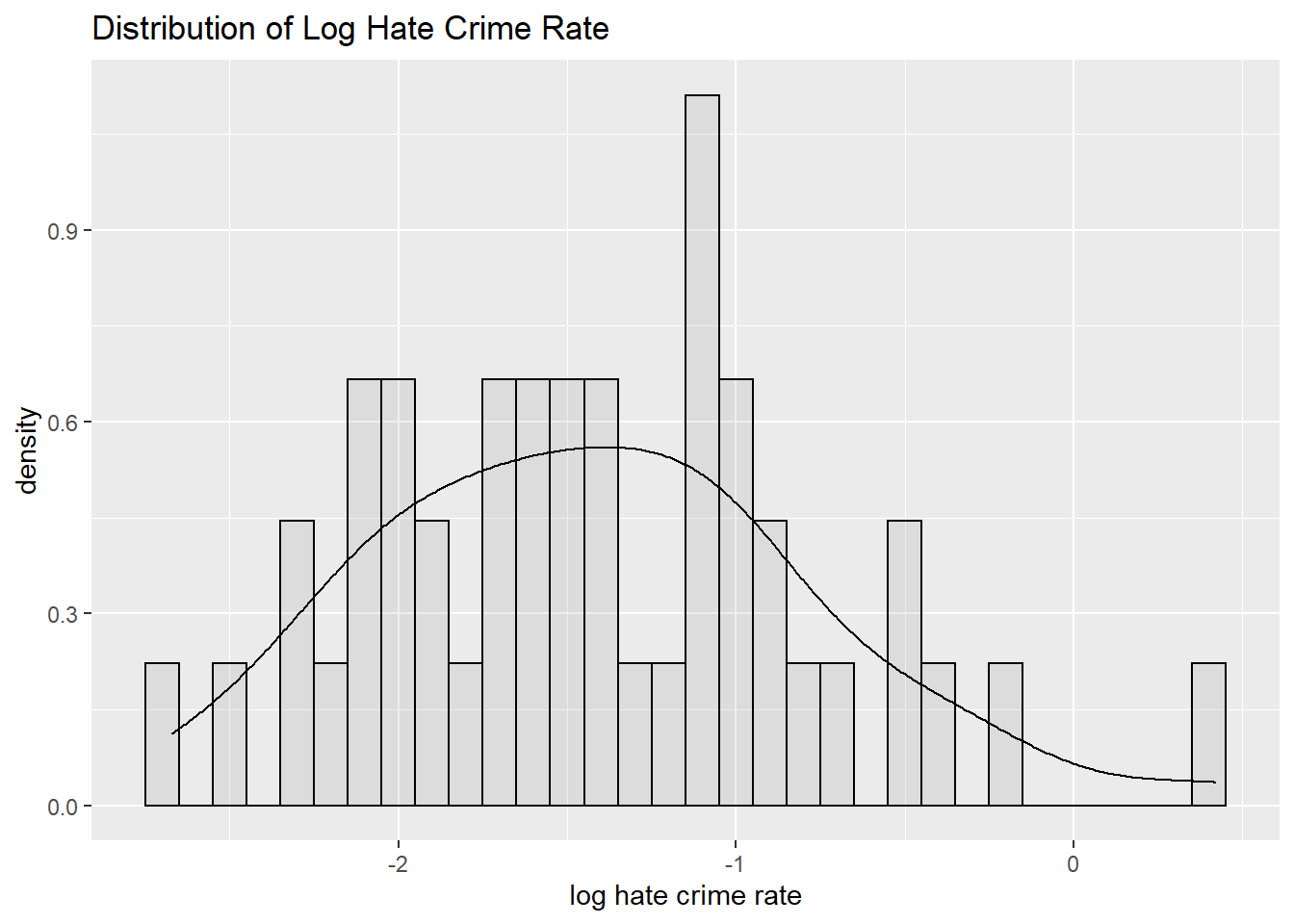


Fig. 3

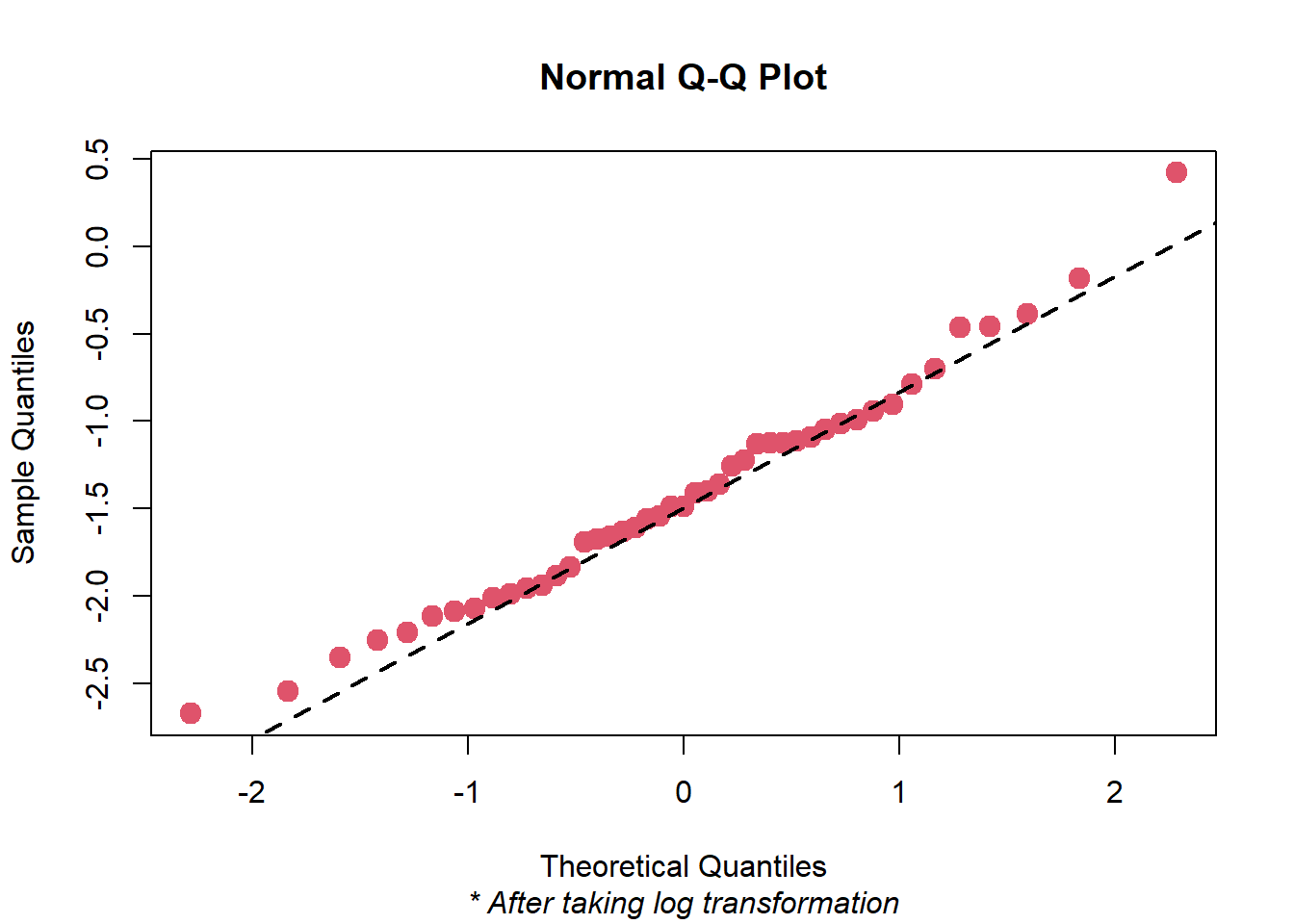


Fig. 4

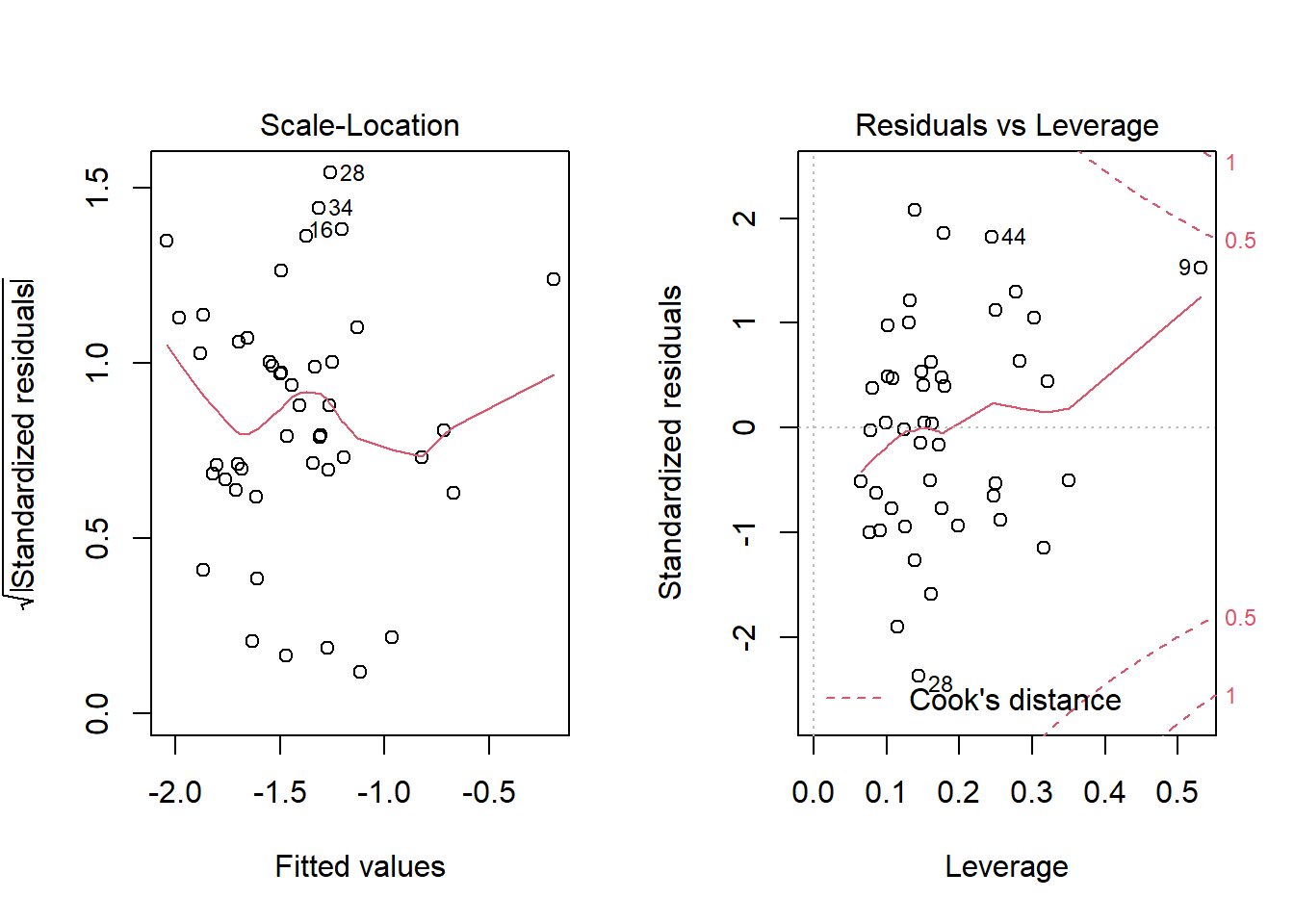


Fig. 5

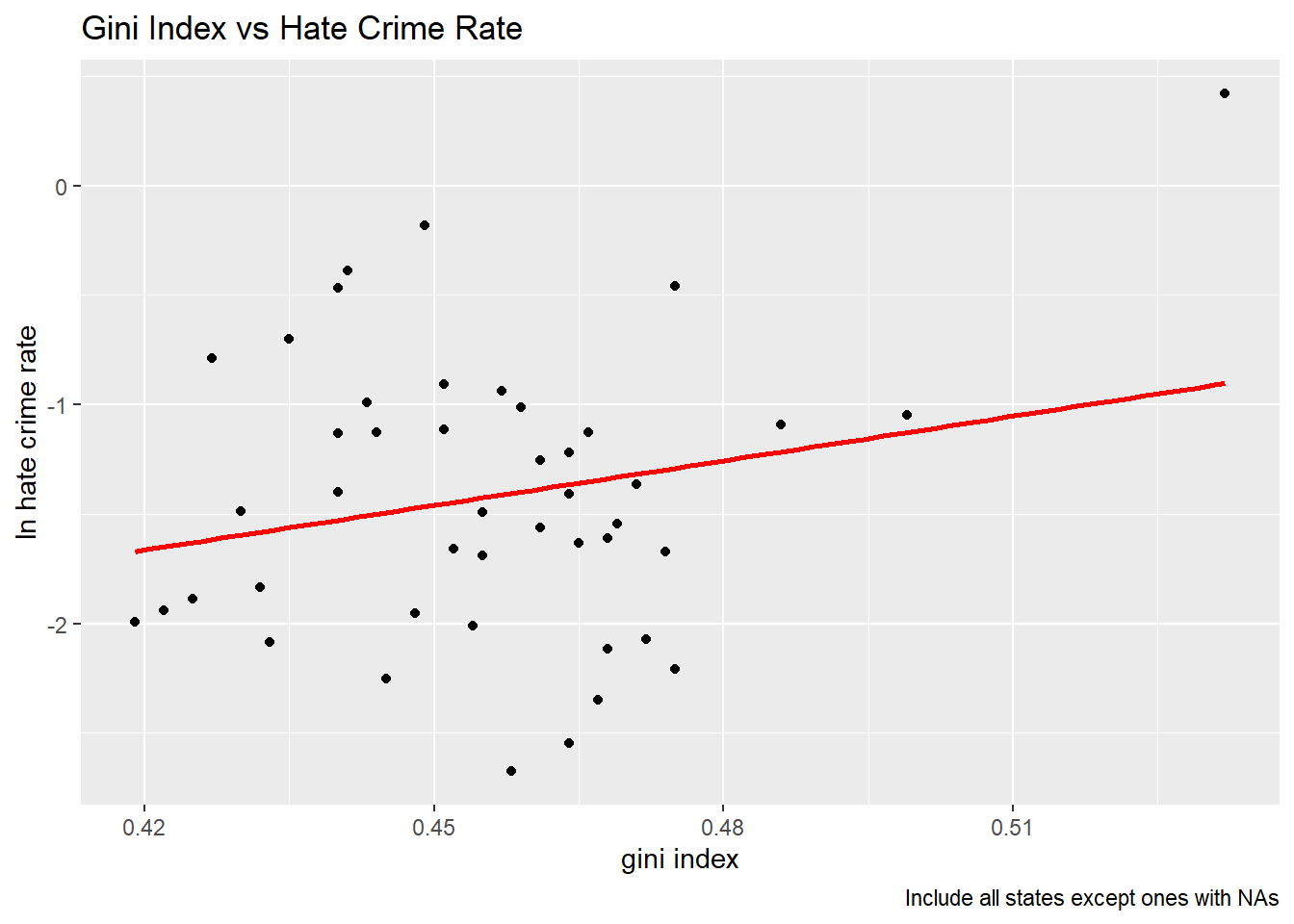


Fig. 6

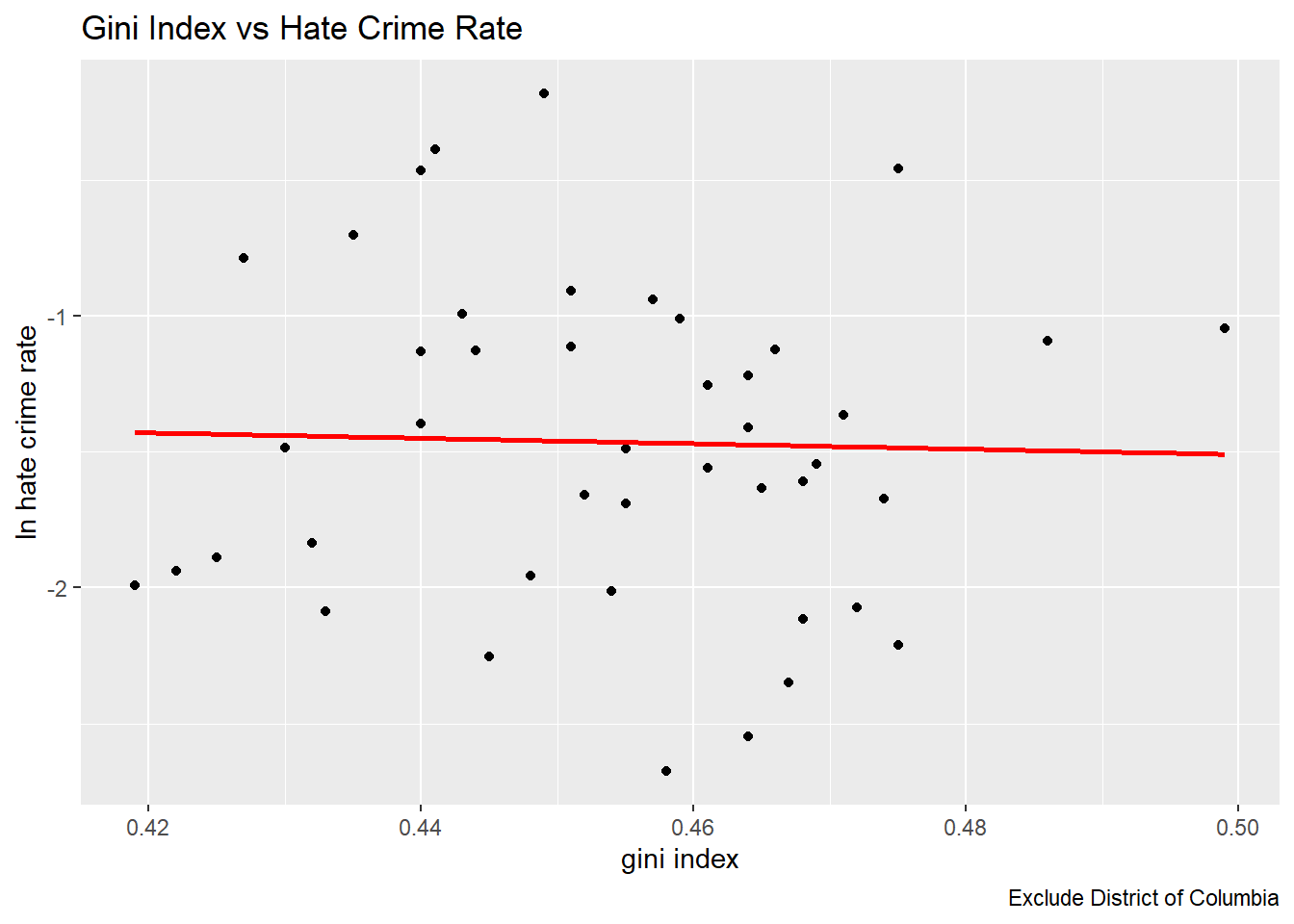


Fig. 7

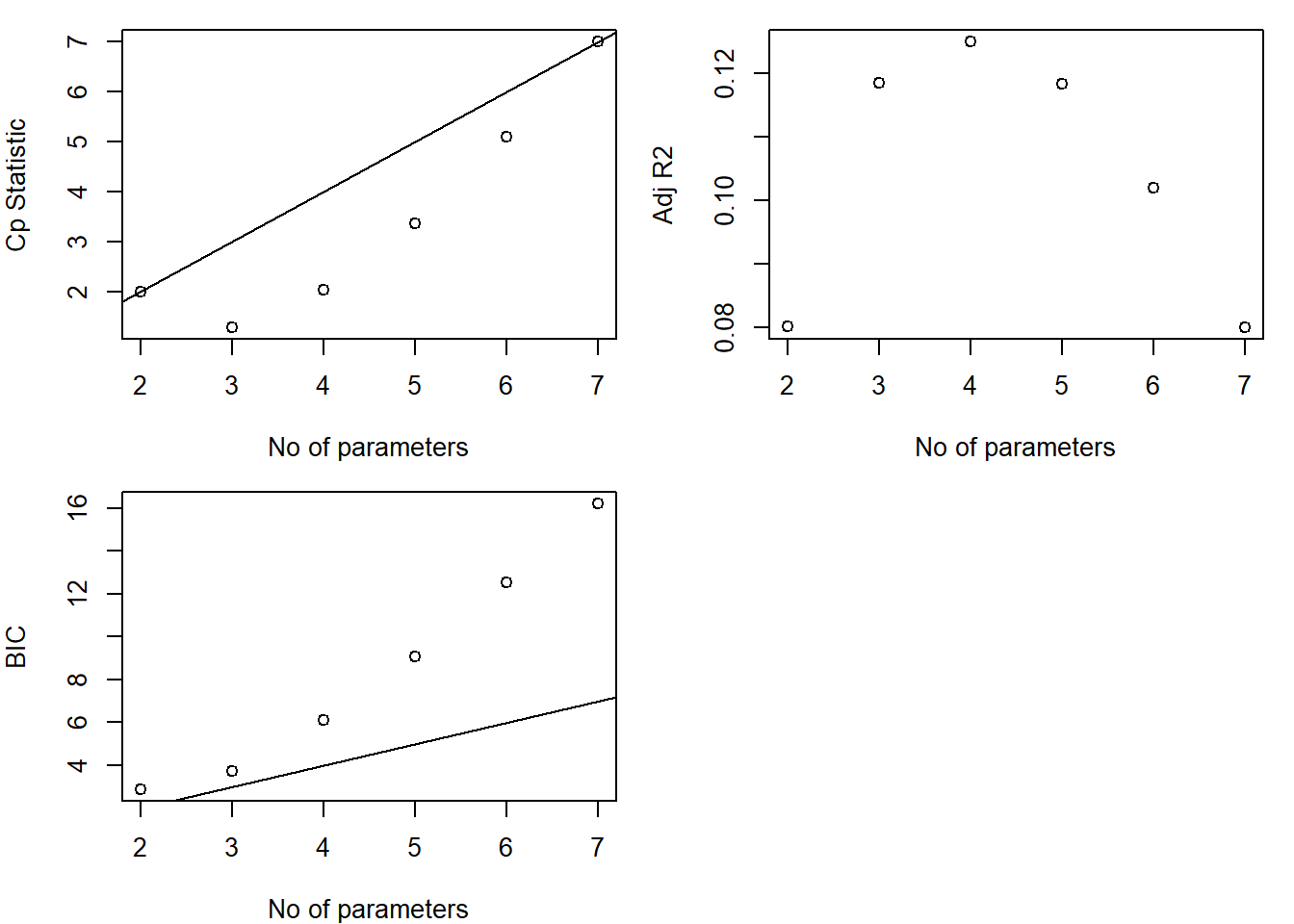


Fig. 8

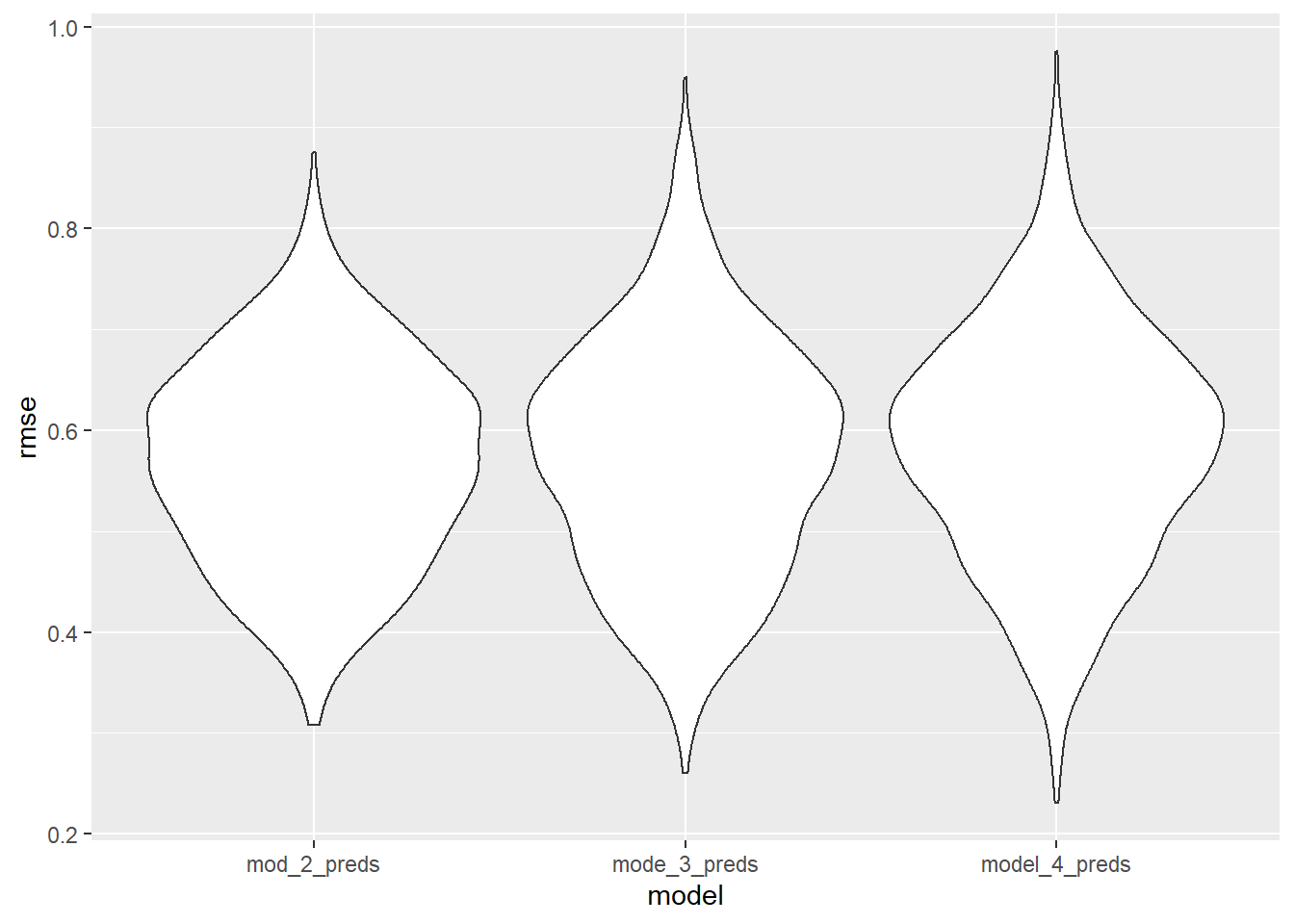


Fig. 9