3. Methods

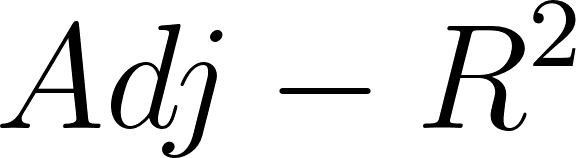
3.1 Data Description

The dataset we worked on contains detailed information on hate crime which happened in 51 states in the United States in 2016. This dataset was adapted from the one used by a FiveThirtyEight article to analyze the same topic. Our dataset contains 51 rows, corresponding to 51 states, and 9 columns. After dropping NAs, there are 45 rows in our dataset.

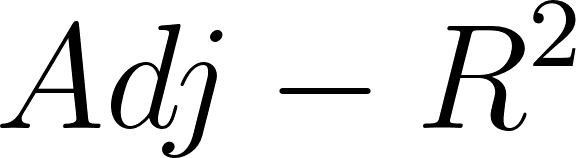
Continuous variables include: hate\_crimes\_per\_100k\_splc (hate crime rate per 100,000 population), median\_household\_income (median household income per state), perc\_population\_with\_high\_school\_degree (percentage of adults (>25 yrs.) with a high school degree), perc\_non\_citizen (percentage 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 include: 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 (Fig. 3 & Fig. 4). In this way, we decide to take the log of the hate crime rate as the outcome in our model.

Noticing that the hate crime rate is very high in the District of Columbia (1.522 per 100,000 population, while overall median = 0.226 per 100,000 population), we make plot of residuals vs. leverage to find out whether it’s an influential observation (Fig. 5). As the plot shows, case 9, which is the observation for the District of Columbia, is close to the dashed lines. Moreover, before removing case 9, there is supposed to be positive linear relationship between Gini index and hate crime rate (Fig. 6), while after removing it, there is slightly negative association between Gini index and hate crime rate (Fig. 7). Based on analyses above, we consider the observation for the District of Columbia as an influential outlier and omit it when conducting further analysis.

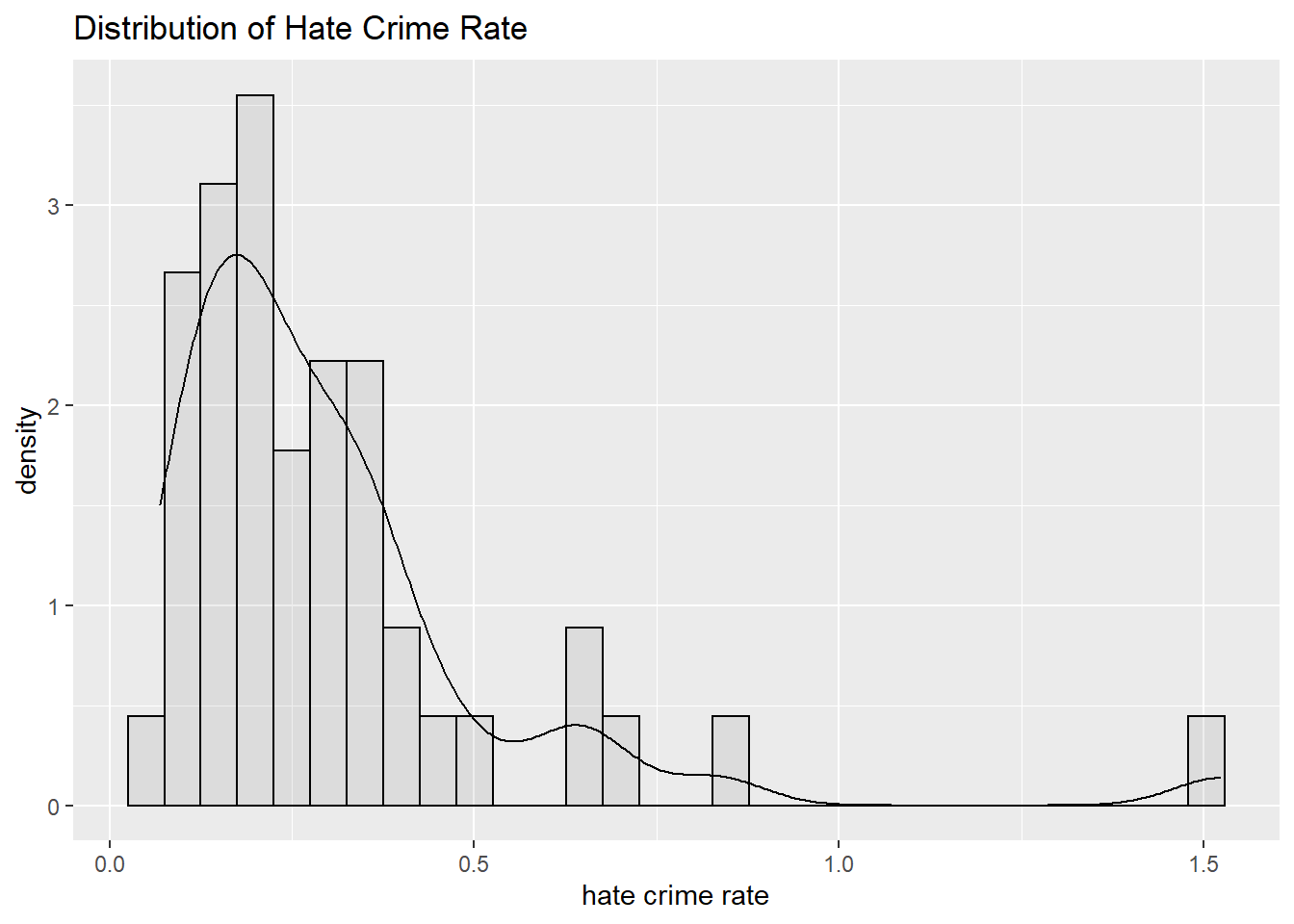
Another issue to be addressed is the intercorrelation between potential predictors. Based on the correlation matrix, there are strong correlation between perc\_non\_white and perc\_non\_citizen (*ρ* = .73) and moderate correlation between  urbanization and perc\_non\_citizen (*ρ* = .67), perc\_population\_with\_high\_school\_degree and median\_household\_income (*ρ* = .66), as well as gini\_index and perc\_population\_with\_high\_school\_degree (*ρ* = -.66). Actually, none of the [](https://www.codecogs.com/eqnedit.php?latex=VIF#0) of these variables exceeds 5. The results of the ANOVA test indicate that adding perc\_population\_with\_high\_school\_degree, which has the highest [](https://www.codecogs.com/eqnedit.php?latex=VIF#0) (4.41), is not redundant (P\_value = 0.148) under the criteria of α = 0.15 and improves [](https://www.codecogs.com/eqnedit.php?latex=Adj-R%5E2#0) from 0.027 to 0.057. However, the same procedure implies that perc\_non\_citizen, which is highly correlated with perc\_non\_white, should not be added into our model (*P\_value* = 0.772).

3.2 Predictor Selection

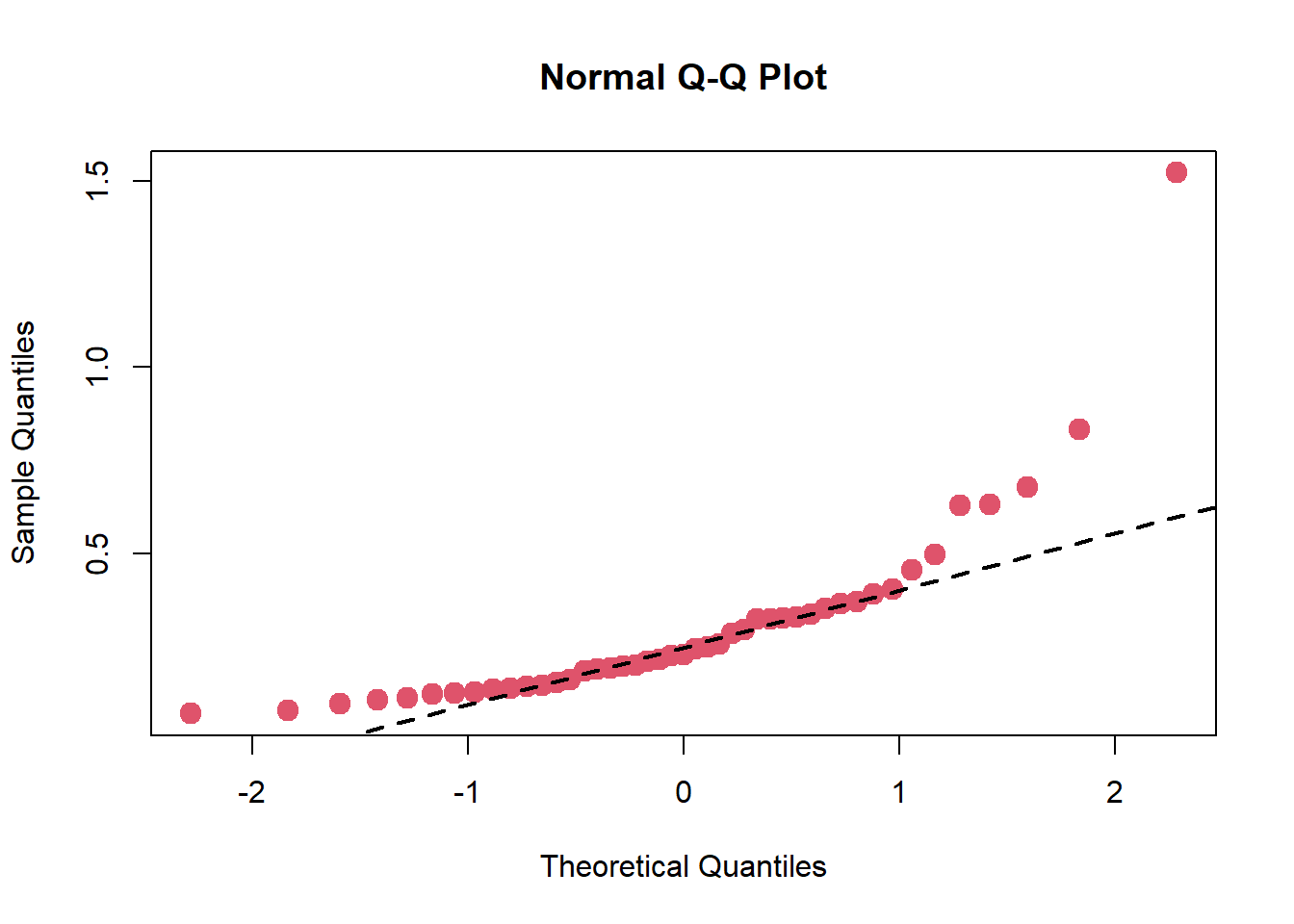
To begin with, we use automatic procedure to identify different best models of different sizes and evaluate them based on multiple criteria: [](https://www.codecogs.com/eqnedit.php?latex=C_p#0), [](https://www.codecogs.com/eqnedit.php?latex=Adj-R%5E2#0) and [](https://www.codecogs.com/eqnedit.php?latex=BIC#0) (Fig. 8). According to these plots, we think that the performance of models with two, three and four predictors are similar and decide to further evaluate them with cross validation (Fig. 9). In this plot, we can see that the RMSE of the model with two predictors is a little smaller than the other two candidates, which indicates better fitness. Combined the rule of parsimony, the model with 2 predictors :[](https://www.codecogs.com/eqnedit.php?latex=ln~hate~crimes~per~100k~splc%3D%5Chat%7B%5Cbeta_0%7D%20%2B%5Chat%7B%5Cbeta_1%7D%20gini~index%2B%5Chat%7B%5Cbeta_2%7D%20perc~population~with~high~school~degree%2B%5Cepsilon_i%2C~i%3D1%2C2%2C..%2C44#0) is selected as our basic model (Table 1).

3.3 Model Modification

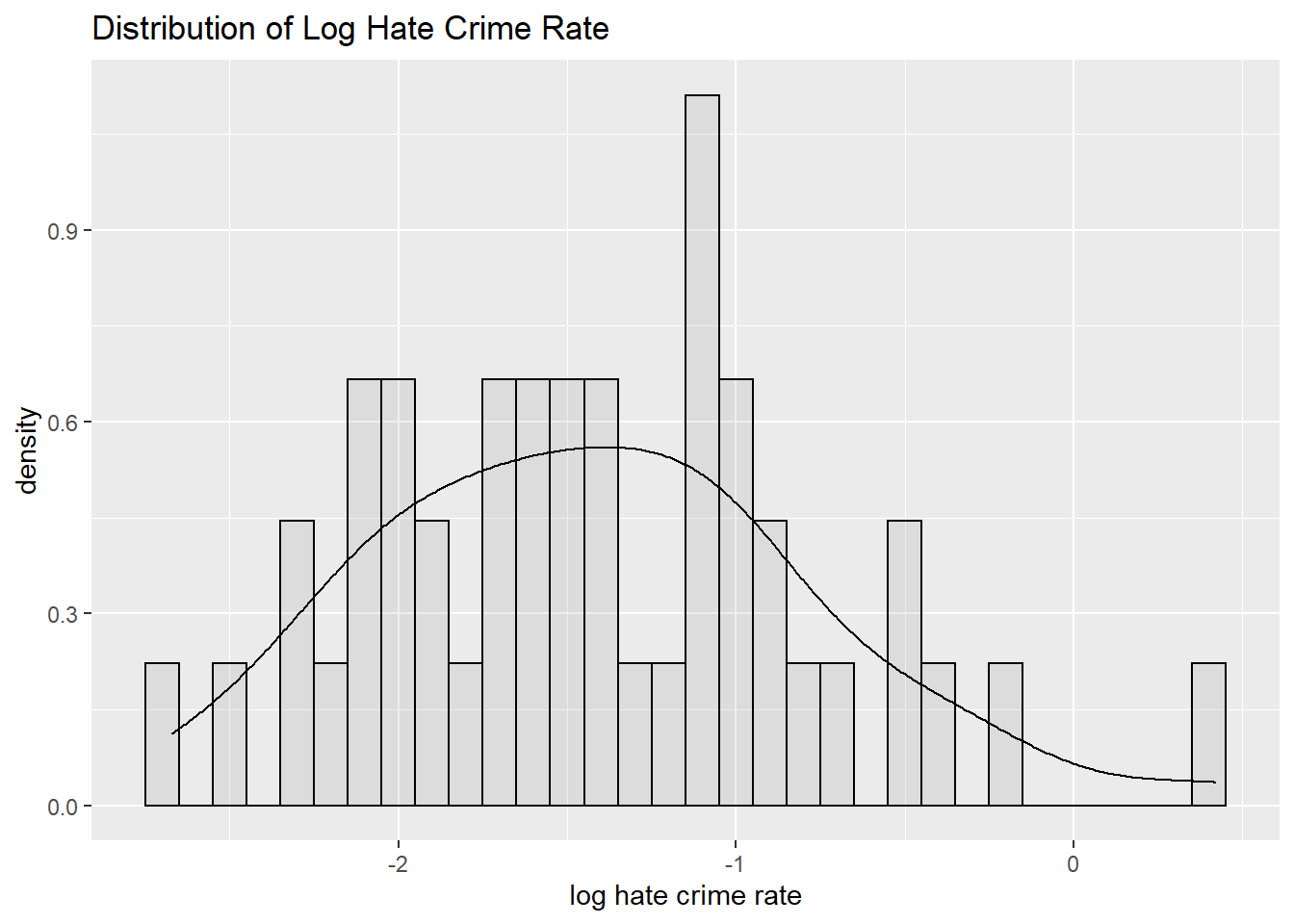
Appendix



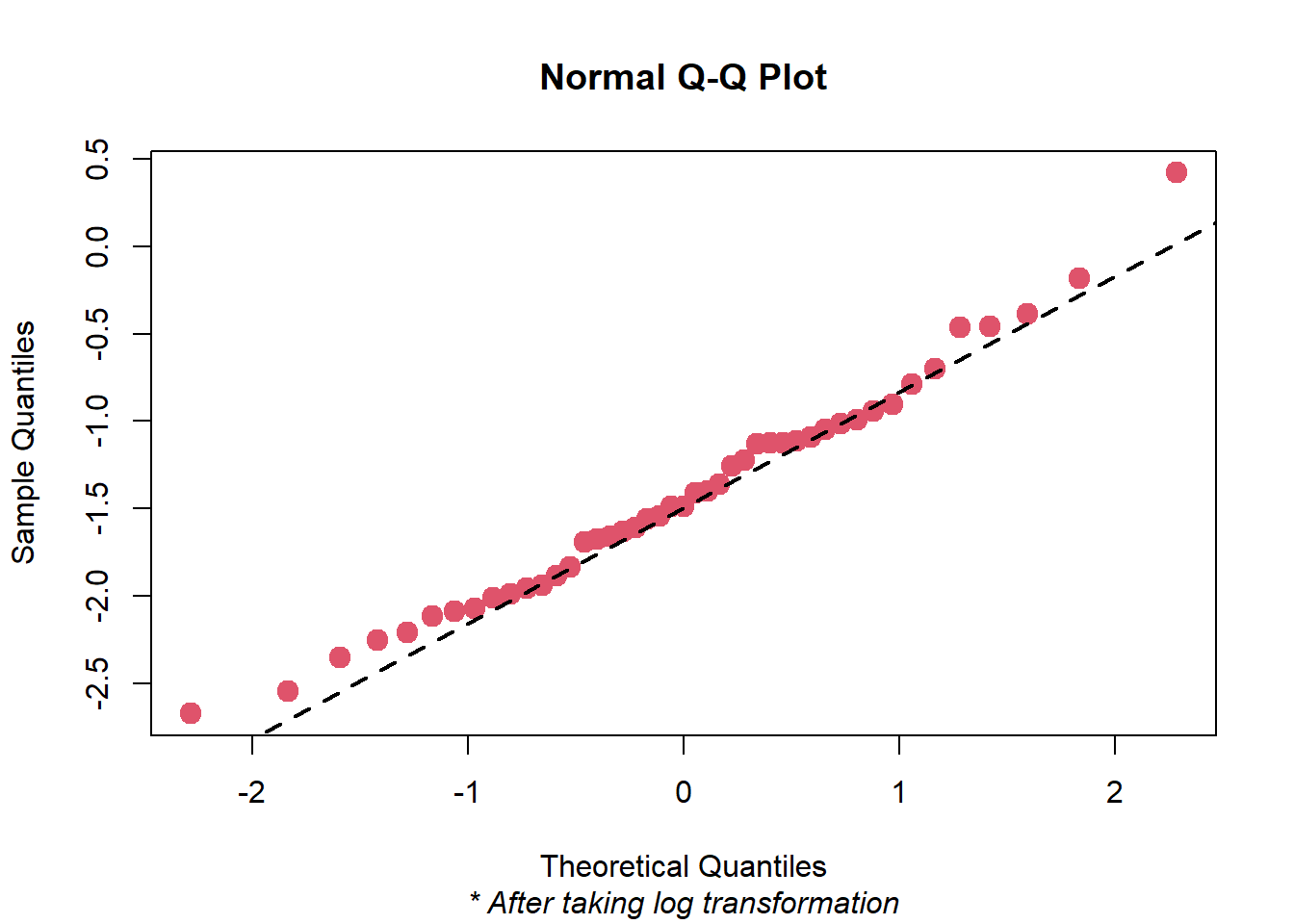
**Fig. 1**



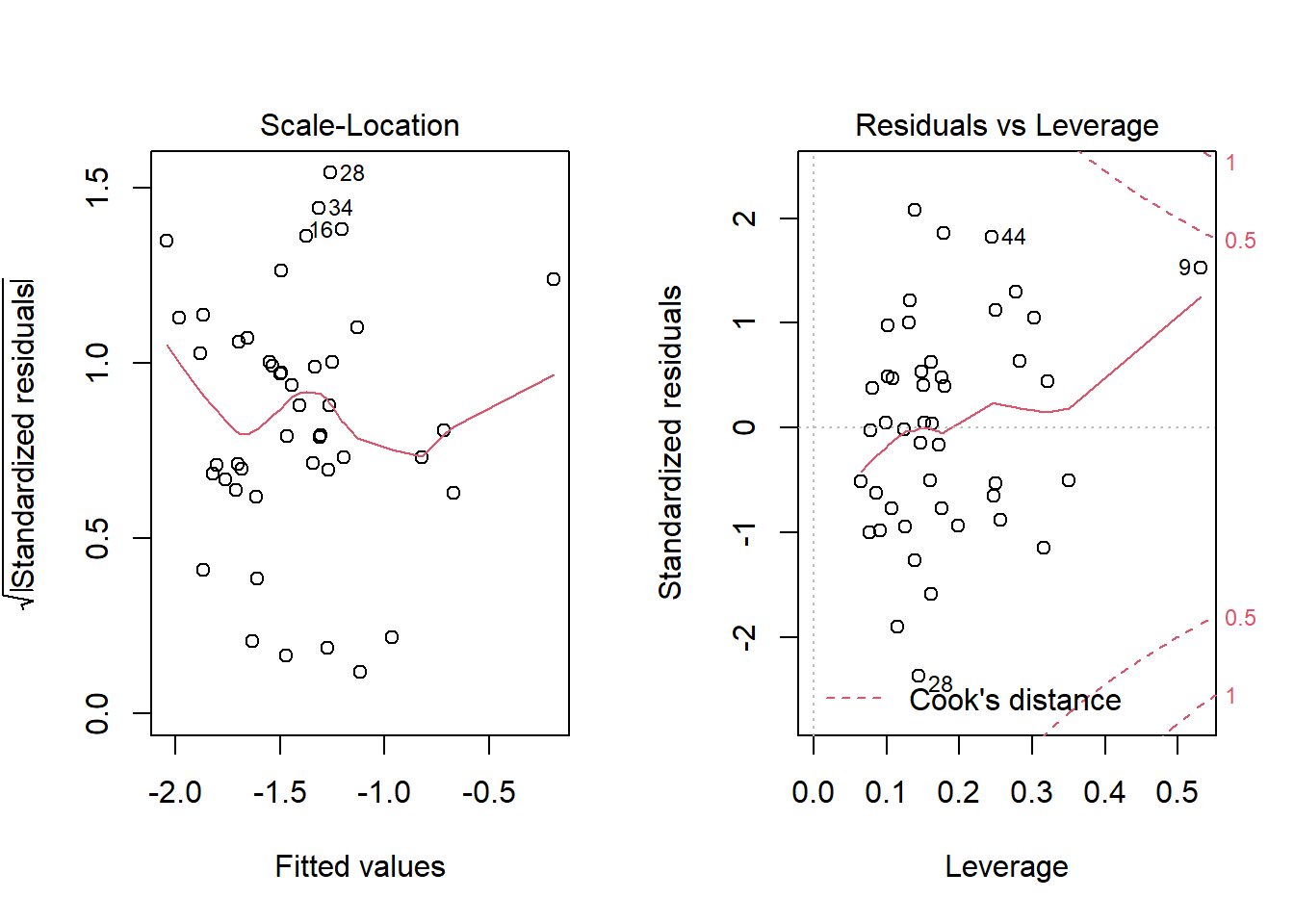
**Fig. 2**



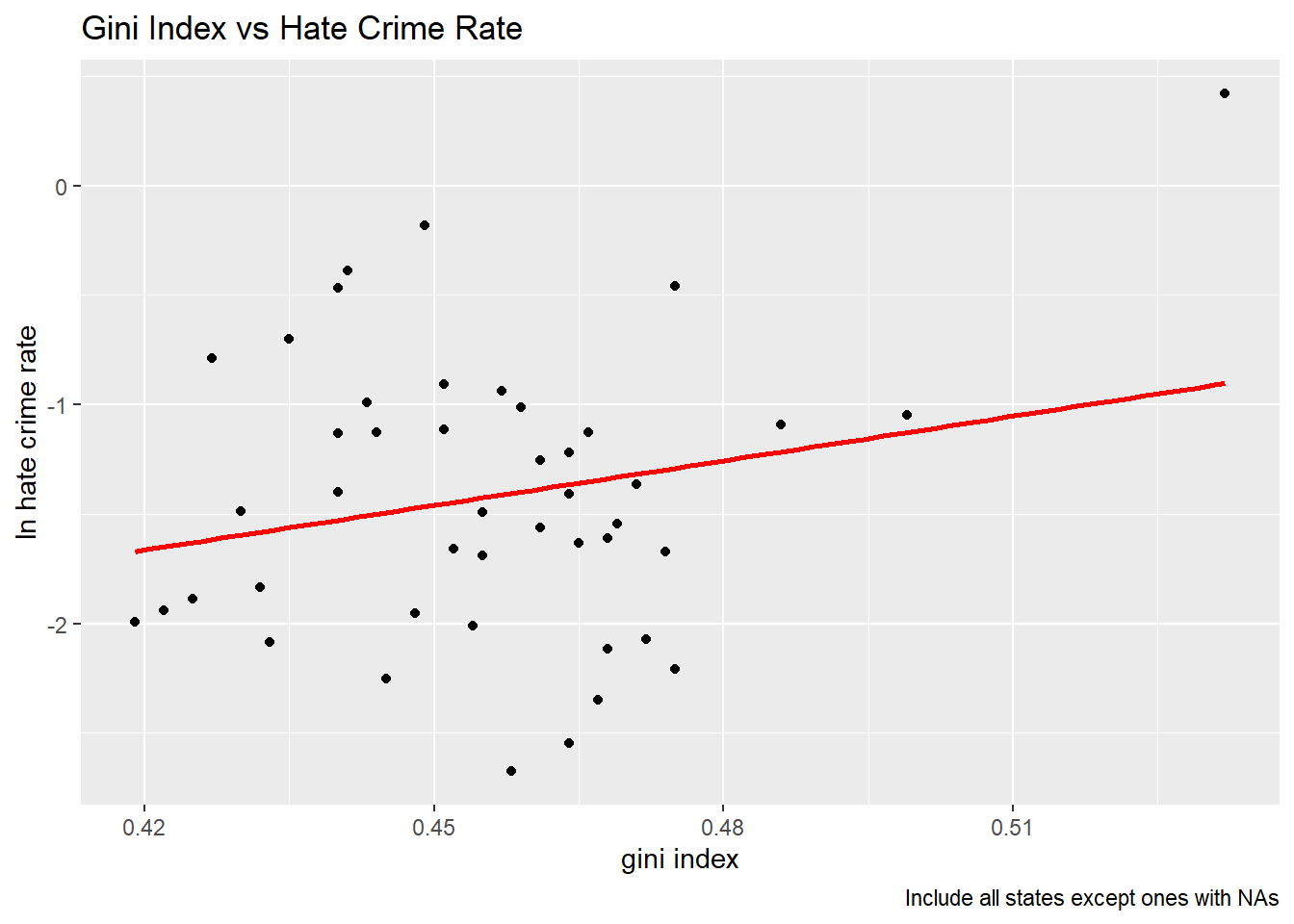
**Fig. 3**



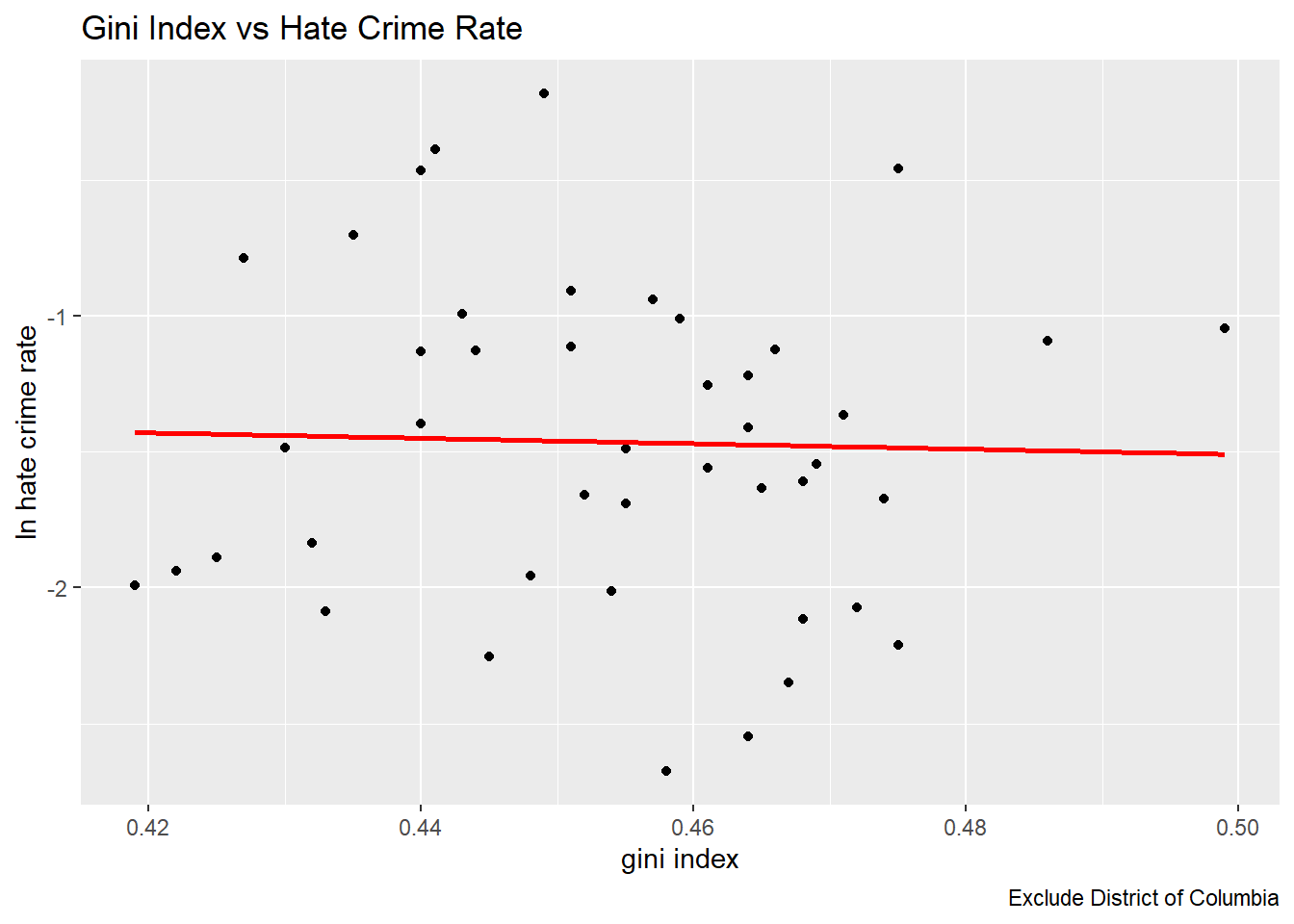
**Fig. 4**



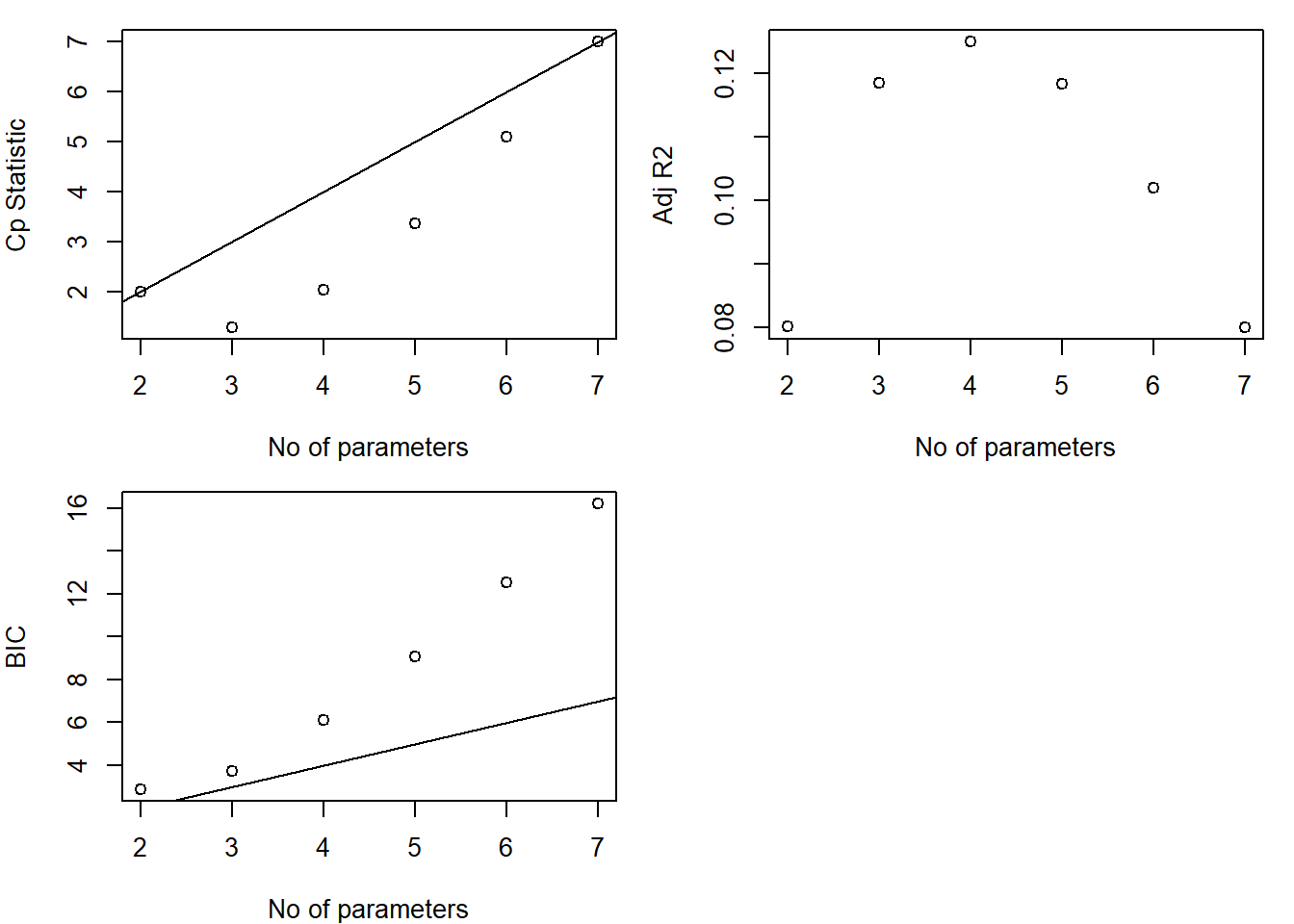
**Fig. 5**



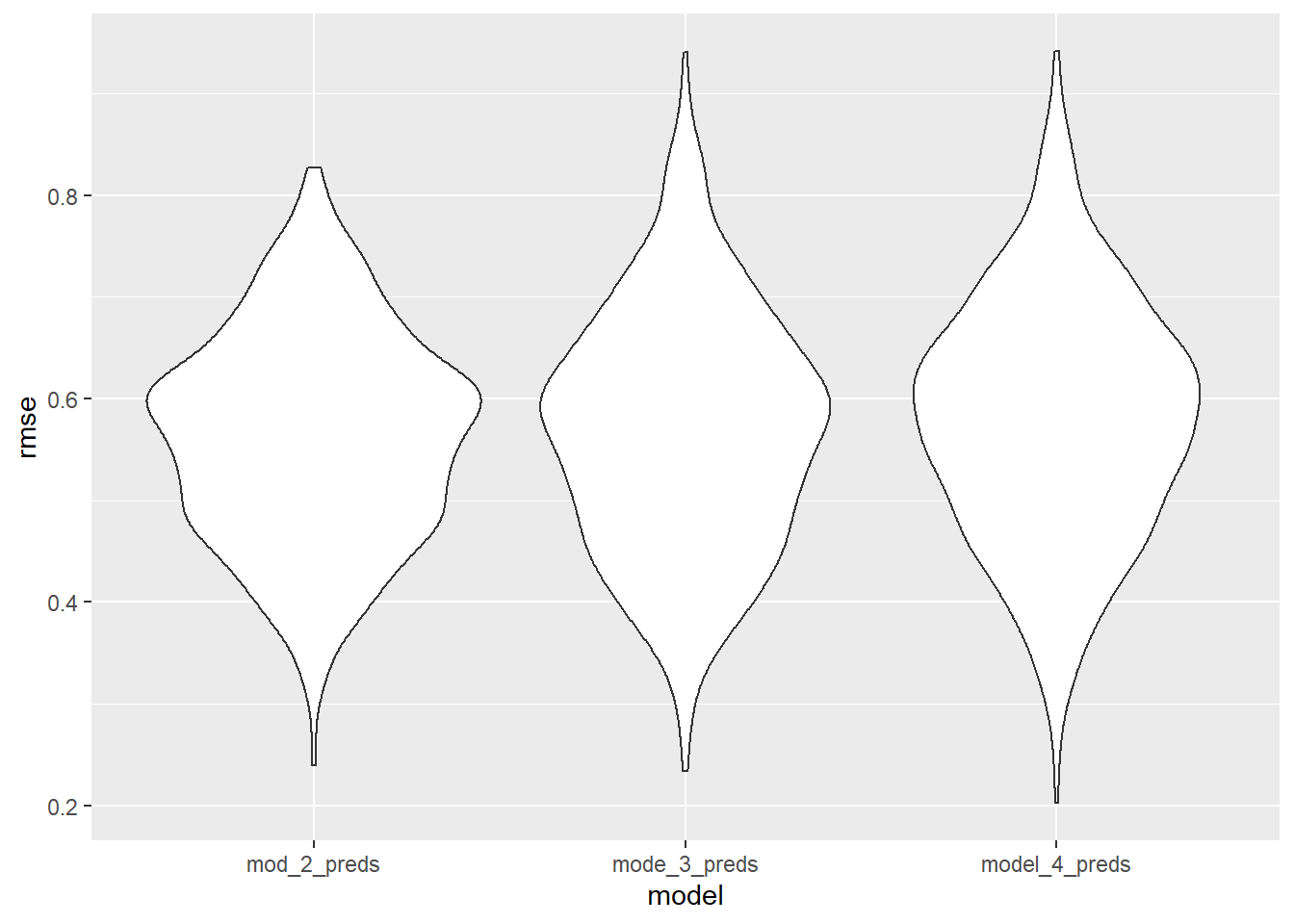
**Fig. 6**



**Fig. 7**

****

**Fig. 8**



**Fig. 9**

**Table 1: Regression Results for the Basic Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Estimate** | **Std.error** | **Statistic** | **P.value** |
| (Intercept) | -14.611 | 5.359 | -2.726 | 0.009 |
| perc\_population\_with\_high\_school\_degree | 9.509 | 3.419 | 2.781 | 0.008 |
| gini\_index | 10.811 | 6.429 | 1.682 | 0.100 |