2. Introduction

Since 2014, the number of hate crimes has been increasing every year and reached a total of 7,314 [1] reported incidents and 59 deaths in 2019 [2]. Though the total number is not comparable to violent crimes and property crimes, hate crimes is under the attention of the Civil Rights Unit supported by the U.S. Attorney's Office, addressed as a violation of civil rights against the whole society [3]. By FBI’s definition, hate crime is the “criminal offense against a person or property motivated in whole or in part by an offender’s bias against a race, religion, disability, sexual orientation, ethnicity, gender or gender identity.” [4] It is highly related to the offender’s personal will based on his/her bias but violating the boundary between freedom of speech and illegitimacy.

According to the nature of hate crime, it is highly randomized and unpredictable and has brought difficulty to the FBI’s strategy to combat this problem. Unlike other crime types, like theft and robbery, primarily driven by poverty and economically inequality, the motivation of hate crimes is hard to be elaborated. Though defined to be in relation with one’s bias, the factors measuring bias are hard to be counted as statistics and used as prediction of future incidences. Does it mean hate crimes are not associated with any social measurements except bias? That is not the case because there has been analysis by FBI and Southern Poverty Law center data suggests that from the scope of the whole country, states with higher income inequality are likely to have higher hate crime incidents [5]. Such analysis could be a clue to the investigation of association between hate crimes and social measurements.

In this project, we are investigating the association between hate crimes rate per 100,000 population and social measurements including unemployment, urbanization, income inequality, education level, racial heterogeneity level etc. The data are describing the behavior of each states in the US aggregately, omitting the detailed performances of individuals. In the analysis, we are identifying the association of each variables with the incidences and the relation within the variables on this case.

3. Methods

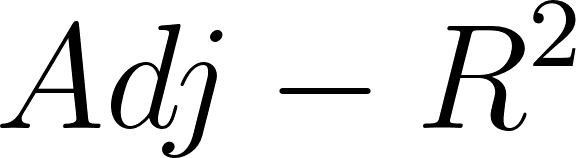
3.1 Data Description

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.

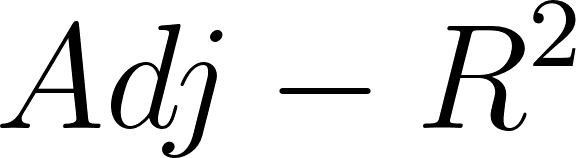
Continuous variables include: hate\_crimes\_per\_100k\_splc (hate crime rate per 100,000 population), median\_household\_income (median household income per state), perc\_population\_with\_high\_school\_degree (percentage of adults (>25 yrs.) with a high school degree), perc\_non\_citizen (percentage 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.

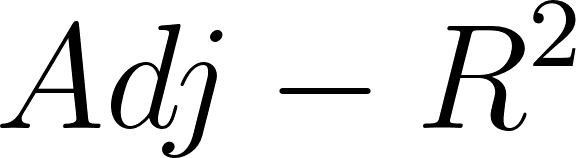
Noticing that the hate crime rate is very high in the District of Columbia (1.522 per 100,000 population, while overall median = 0.226 per 100,000 population), we make plot of residuals vs. leverage to find out whether it’s an influential observation (Fig. 5). As the plot shows, case 9, which is the 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

We want to further test if there exists interaction between Gini index and unemployment, we first draw a plot of two slopes of Gini index vs. high unemployment data and Gini index vs. low unemployment data, the result is in the (Fig. 10). The plot shows that there exists a cross between these two slopes, which is an indication of interaction between Gini index and unemployment. We then added the interaction term into our basic model, after fitting, we see the interaction term has p-value = 0.09434, which is significant under the criteria of α = 0.15. The [](https://www.codecogs.com/eqnedit.php?latex=Adj-R%5E2#0) increases from 0.1185 to 0.1654 (more than 6%), means adding this interaction term improved out model fitness.

Since there is an interaction, we did the stratified analysis based on different category of unemployment level, using low-unemployment data, we find a positive association (19.37) between Gini index and the hate crime rate, the *p-value* is 0.018, which is statistically significant. When using high-unemployment data, we find a negative association (-1.178) between Gini index and the hate crime rate, the *p-value* is 0.918, which is not statistically significant.

We then construct an ANOVA test to compare our new model (with the interaction term) to our basic model. We find the *p-value* is 0.1296, using a criterion of α = 0.15, we conclude the p-value is significant, and thus the larger model is superior.

Now we compare the RMSE value between the new model (with adding the interaction term between Gini index and unemployment) with the basic model, using 1000-fold cross validation. From figure 11, we find there are hard to see the difference between our new model and the basic model, when the RMSE value of the basic model is more concentrated in the middle (0.56) and the RMSE value of the new model is more spreading, with smaller minimum RMSE value and larger maximum RMSE value.

Furthermore, we also draw a plot using fitted value and model residuals of our new model, from figure 12**,** we can see most of the residuals are distributed evenly (constant variance) around the 0 line. Our normality assumption seems hold.

4. Results

After automatic stepwise selection, evaluating through multiple criteria, and further testing of interaction and analysis, we find our overall best fitted model would be:

Which contains three predictors, with p-value = 0.02515, and adjusted R-squared value of 0.1654. This modified model (see Table 2) has a slight improvement comparing with our basic model using two predictors.

Reference

[1] *Incidents and Offenses*. FBI. (2020). Retrieved 16 December 2020, from https://ucr.fbi.gov/hate-crime/2019/topic-pages/incidents-and-offenses.

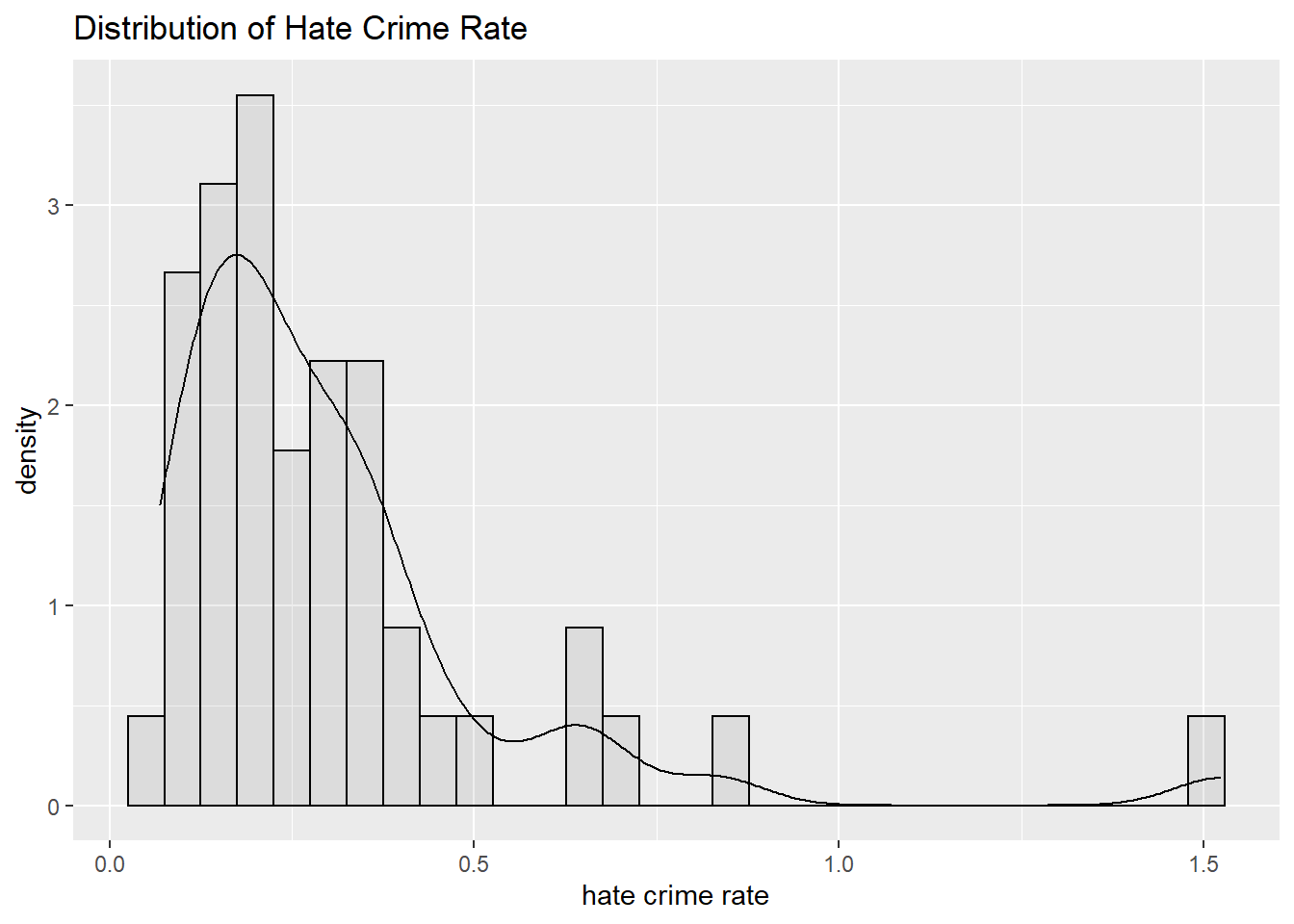
[2] *US hate crime highest in more than a decade - FBI*. BBC News. (2020). Retrieved 16 December 2020, from https://www.bbc.com/news/world-us-canada-54968498.

[3] *The Civil Rights Program*. Justice.gov. (2020). Retrieved 16 December 2020, from https://www.justice.gov/usao-wdtn/civil-rights-program.

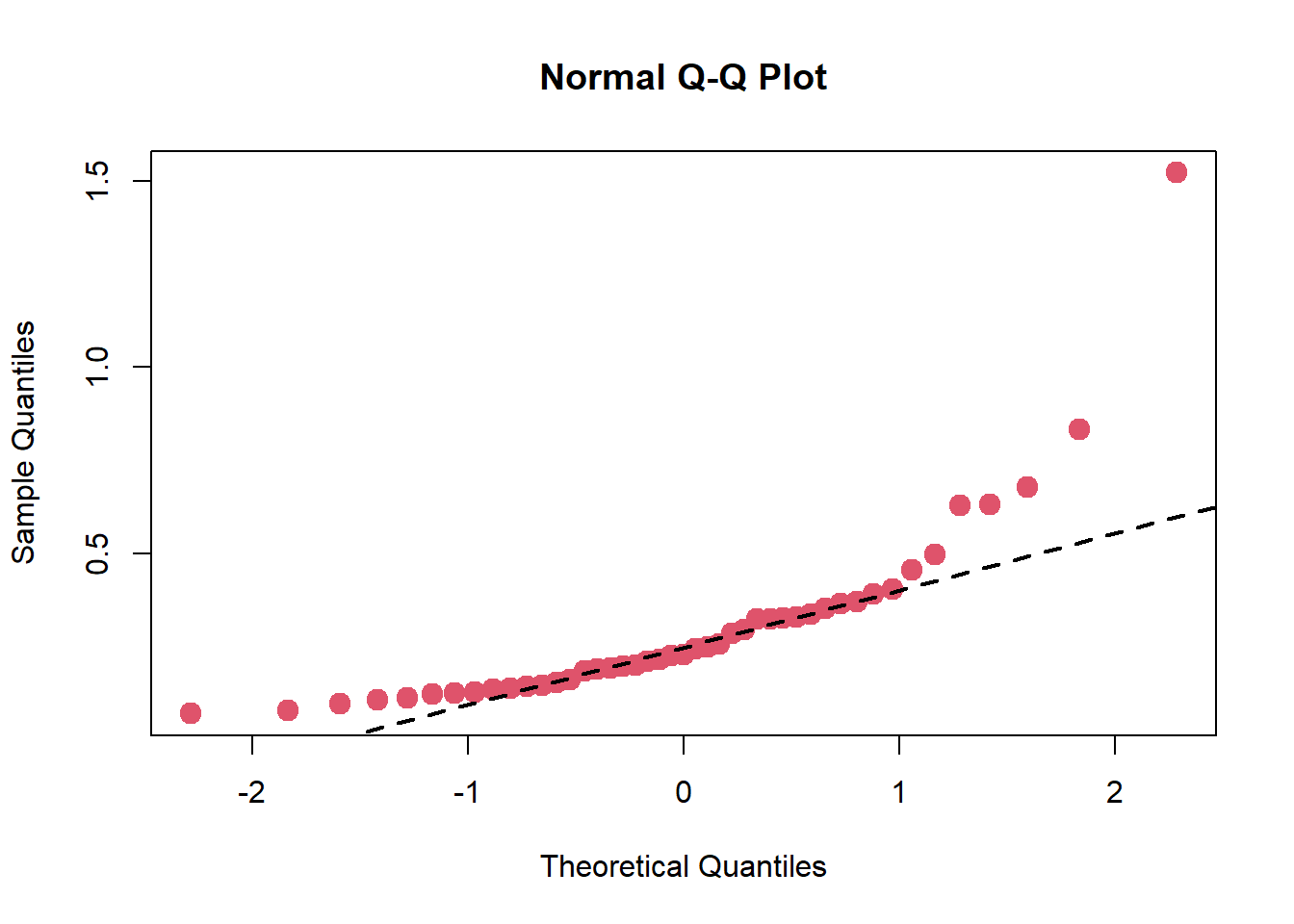
[4] *Hate Crimes | Federal Bureau of Investigation*. Federal Bureau of Investigation. (2020). Retrieved 16 December 2020, from https://www.fbi.gov/investigate/civil-rights/hate-crimes.

[5] Majumder, M. (2020). *Higher Rates Of Hate Crimes Are Tied To Income Inequality*. FiveThirtyEight. Retrieved 16 December 2020, from https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/.

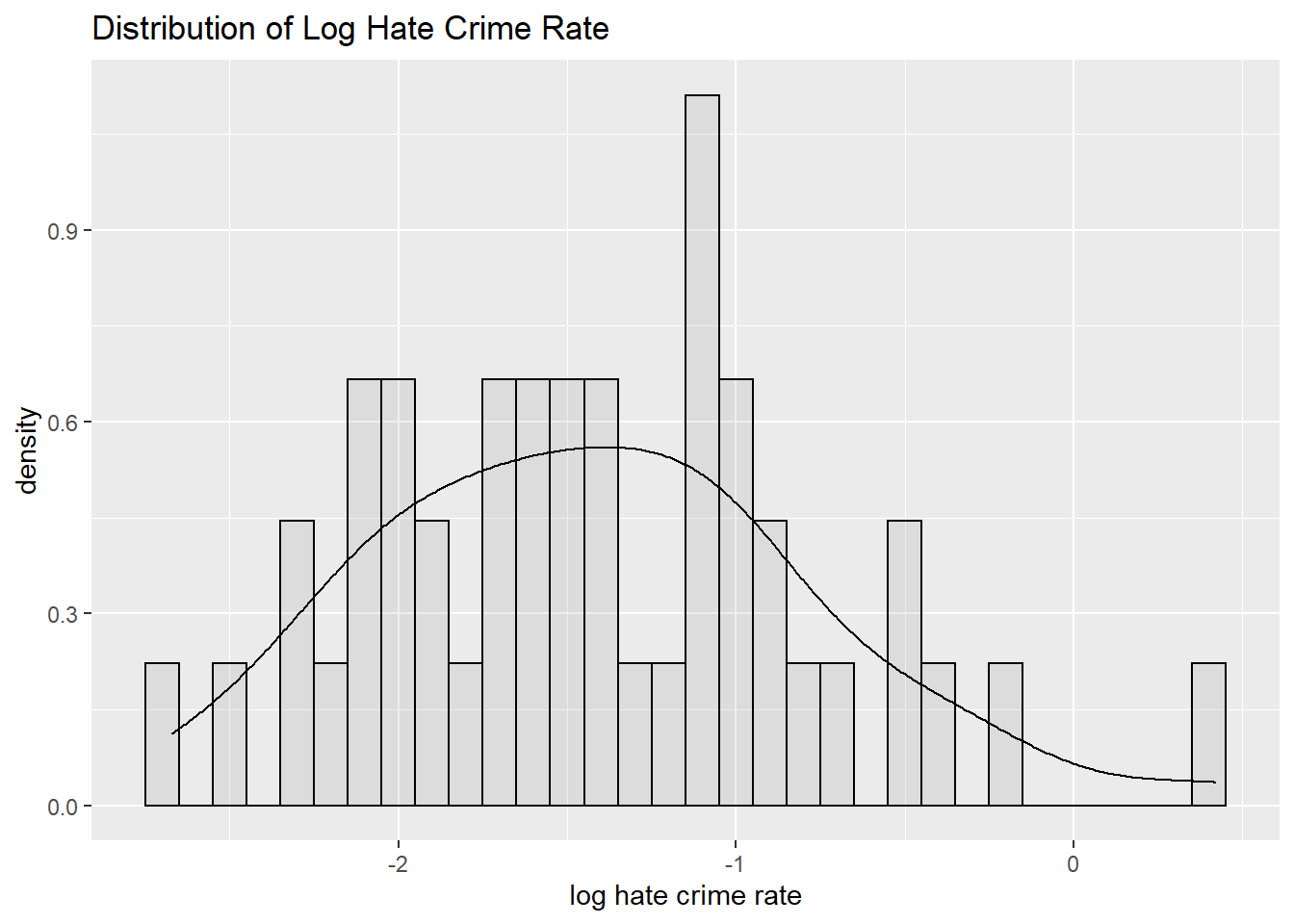
Appendix



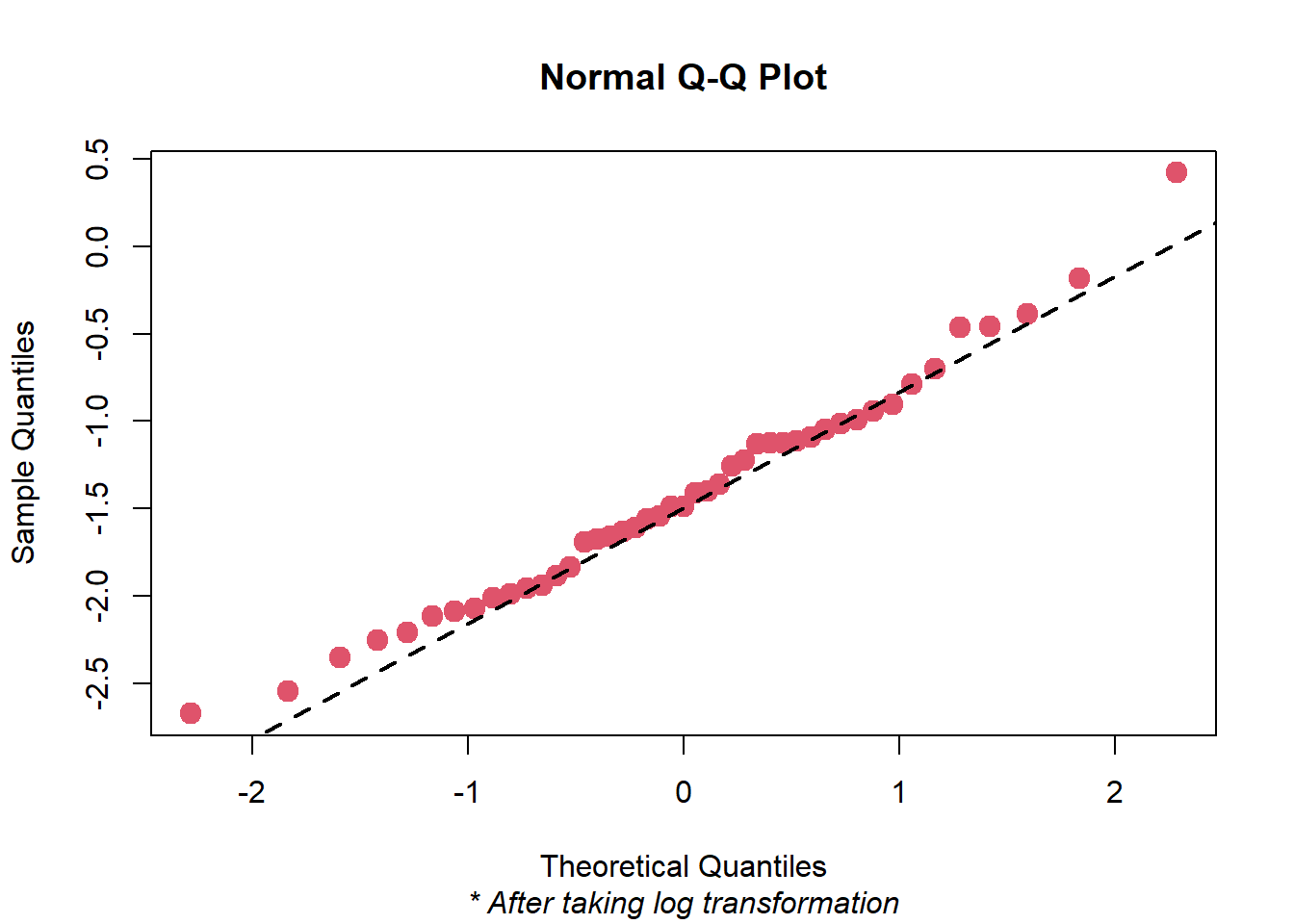
**Fig. 1**



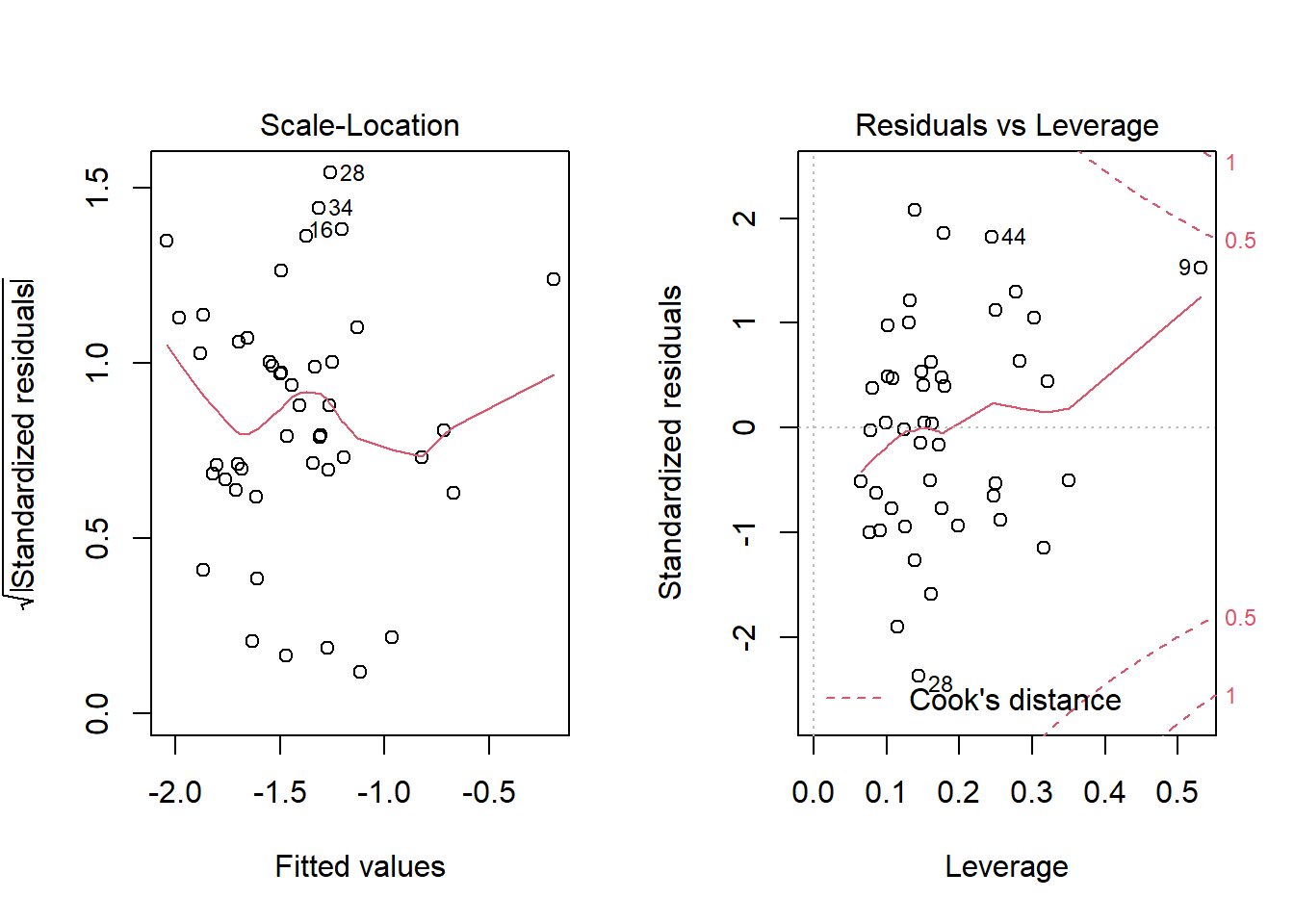
**Fig. 2**



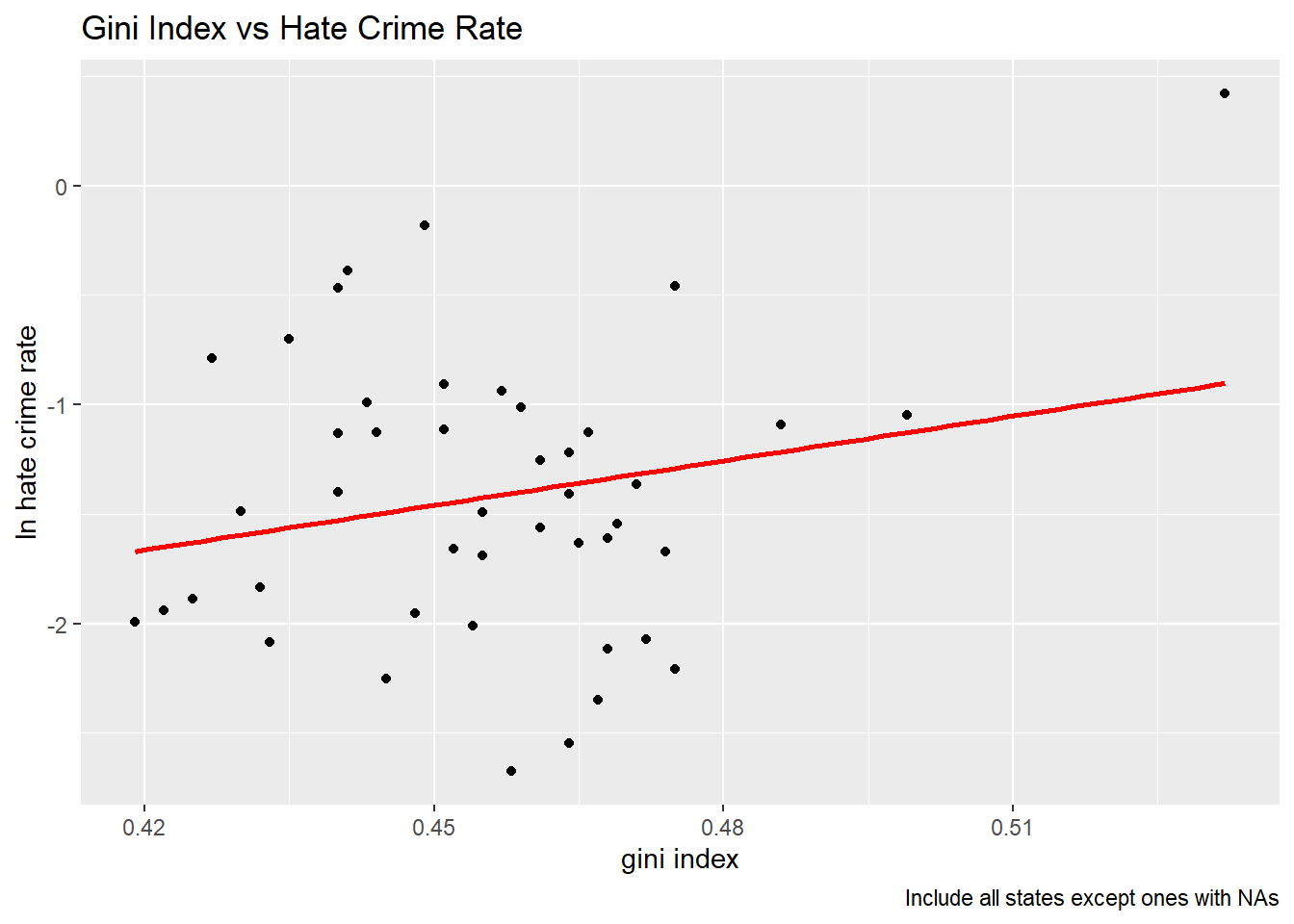
**Fig. 3**



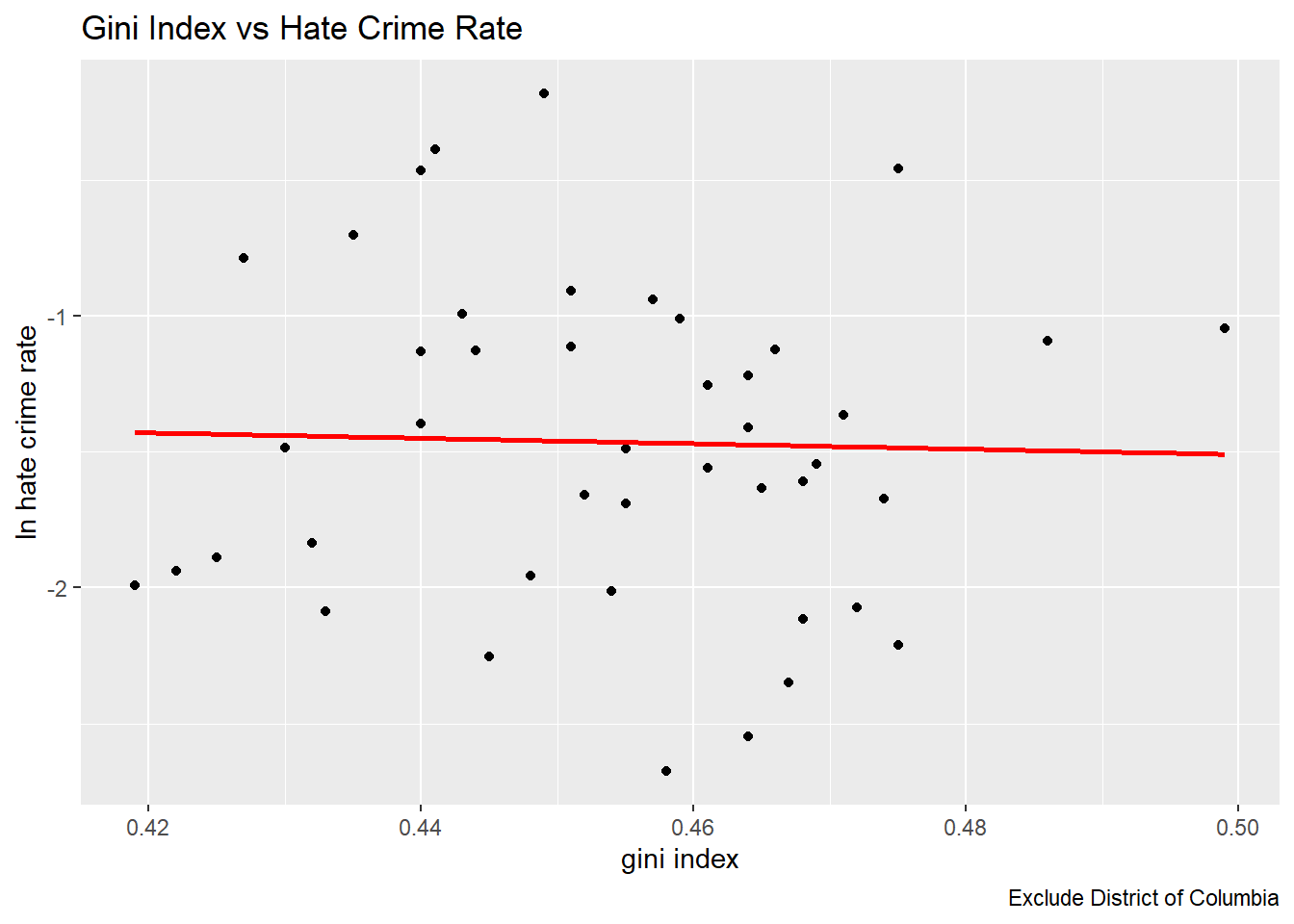
**Fig. 4**



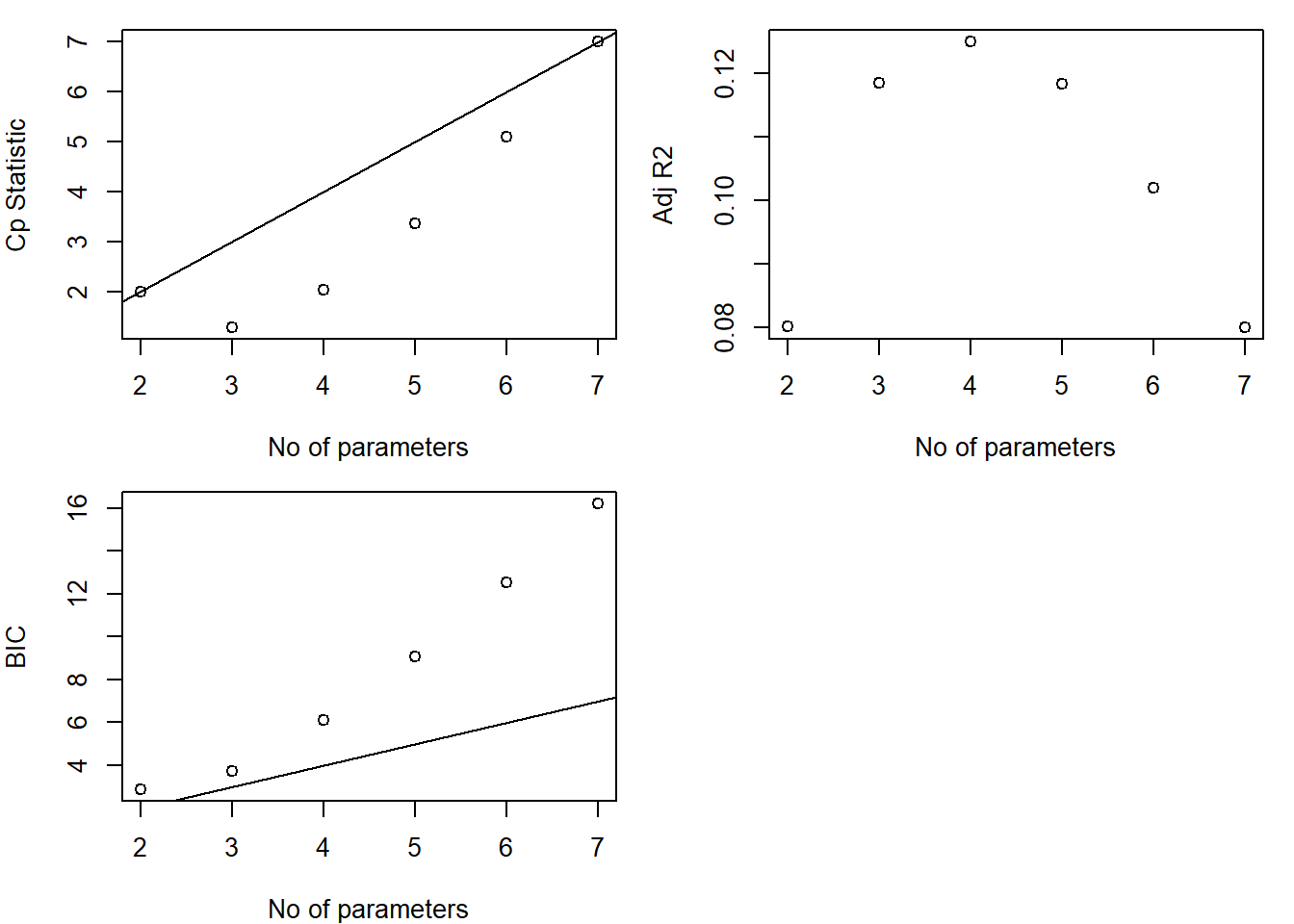
**Fig. 5**



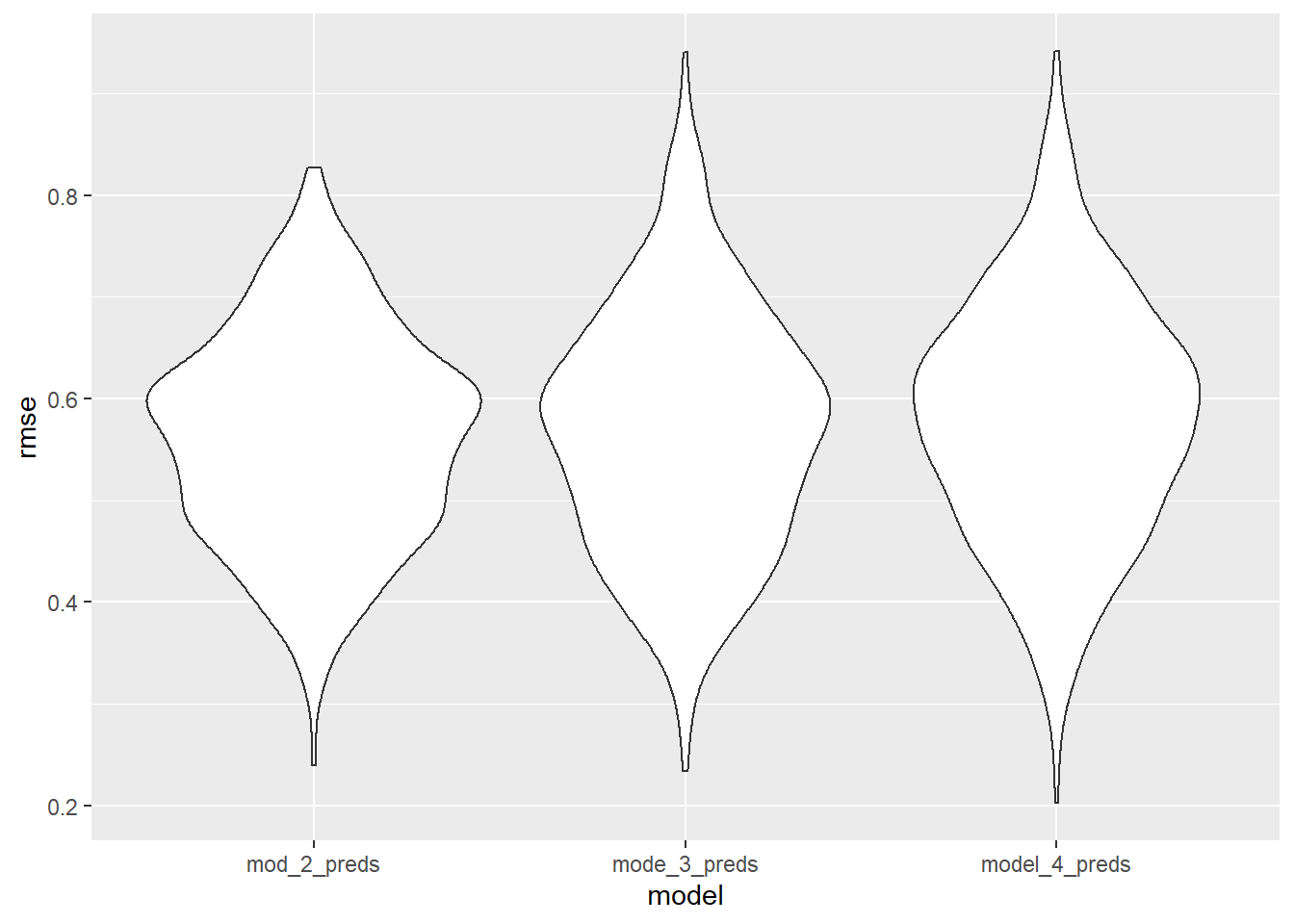
**Fig. 6**



**Fig. 7**

****

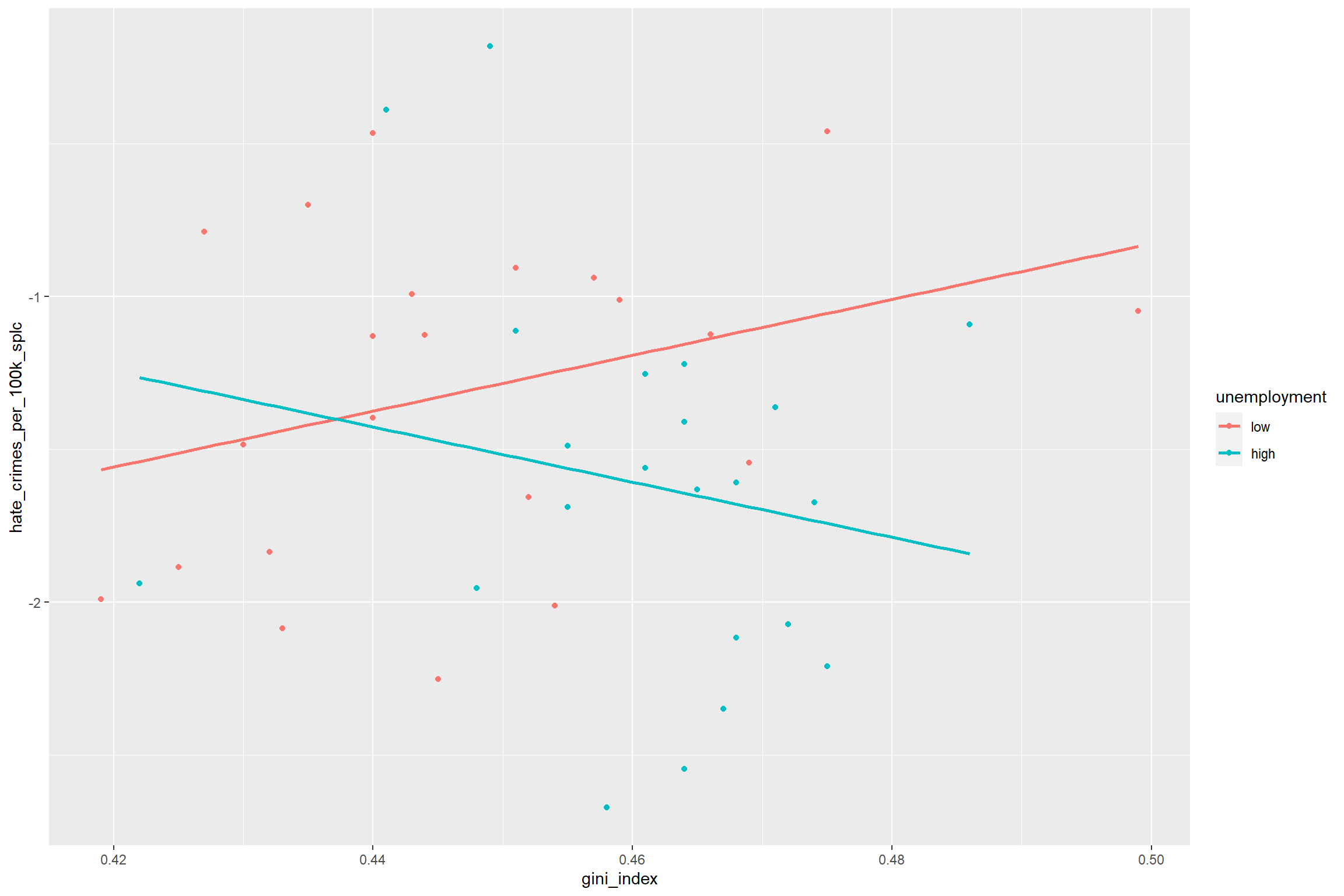
**Fig. 8**



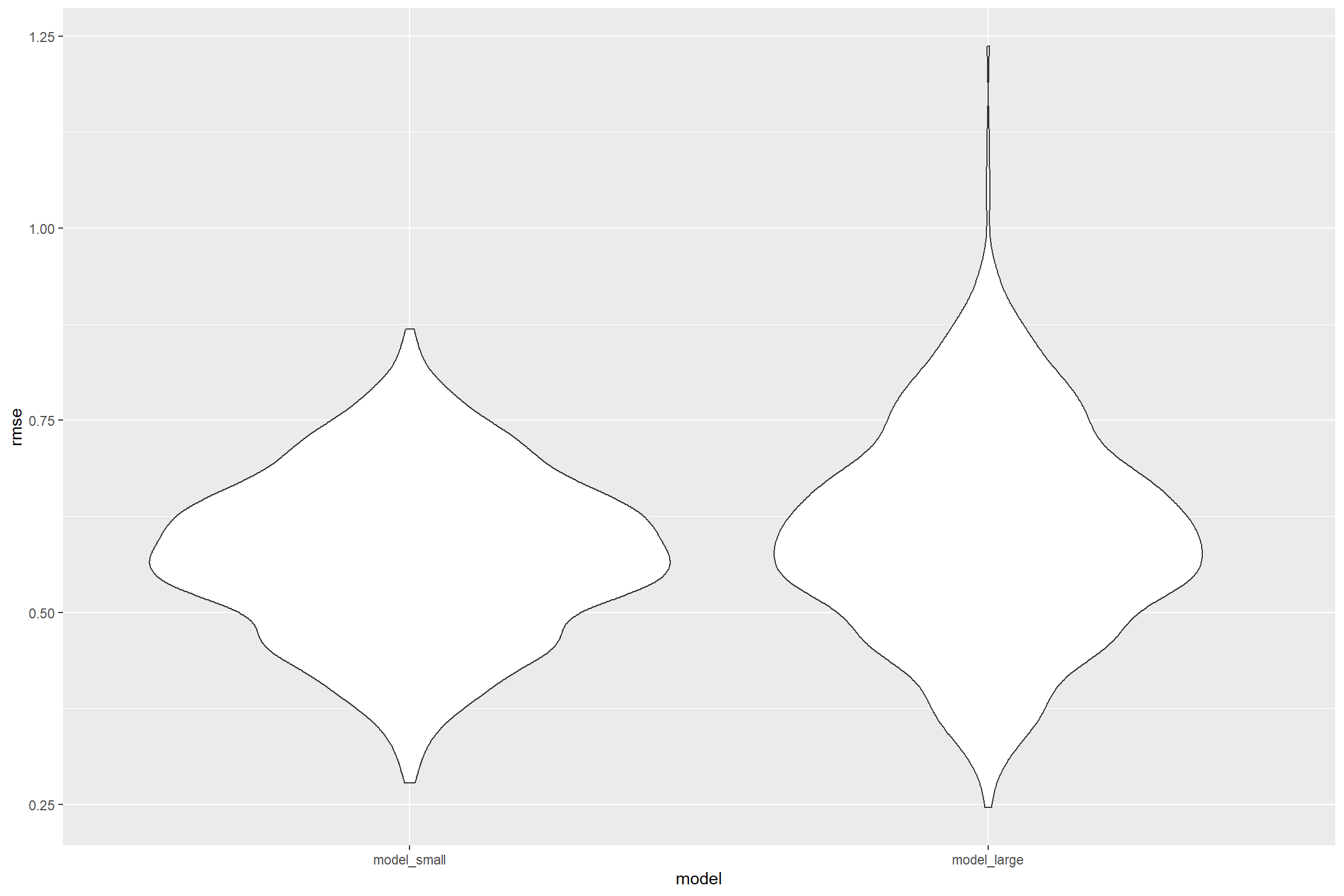
**Fig. 9**

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

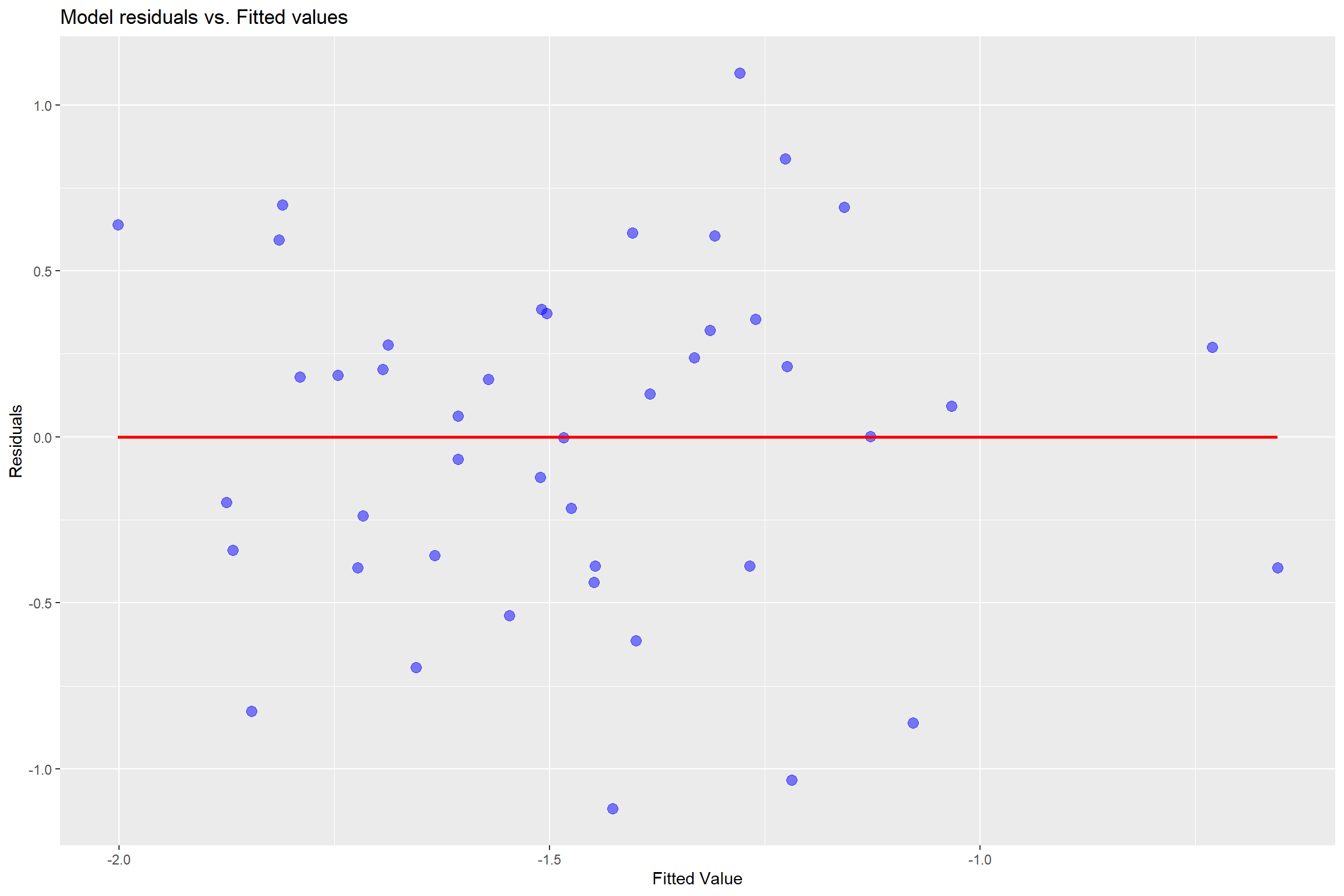
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Estimate** | **Std.error** | **Statistic** | **P.value** |
| (Intercept) | -14.611 | 5.359 | -2.726 | 0.009 |
| perc\_population\_with\_high\_school\_degree | 9.509 | 3.419 | 2.781 | 0.008 |
| gini\_index | 10.811 | 6.429 | 1.682 | 0.100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

****

**Fig. 10**

****

**Fig. 11**

****

**Fig. 12**

**Table 2: Regression Results for the Modified Model**

| **term** |  | **estimate** | **std.error** | **statistic** | **p.value** |
| --- | --- | --- | --- | --- | --- |
| (Intercept) |  | -16.897 | 5.517 | -3.063 | 0.004 |
| perc\_population\_with\_high\_school\_degree |  | 8.427 | 3.474 | 2.426 | 0.020 |
| gini\_index |  | 18.247 | 7.338 | 2.487 | 0.017 |
| unemploymenthigh |  | 8.235 | 4.933 | 1.669 | 0.103 |
| gini\_index:unemploymenthigh |  | -18.526 | 10.804 | -1.715 | 0.094 |