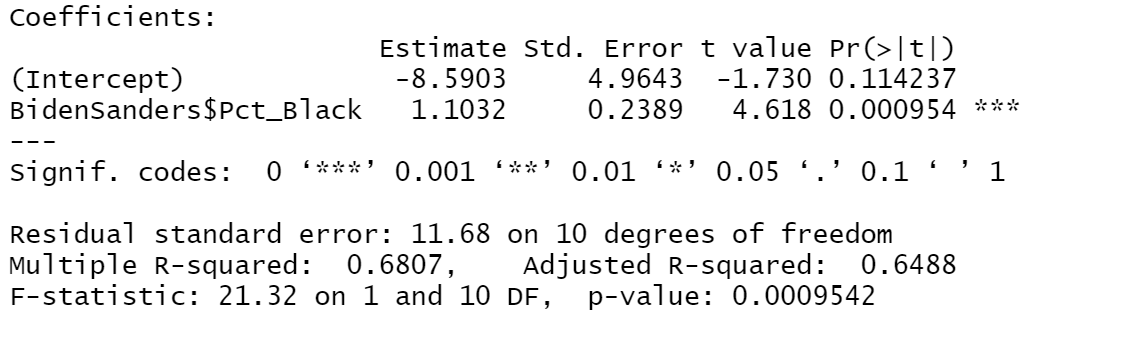
**Problem Set 5**

This problem set covers material from the two classes after the exam, including material from Chapter 7. It makes use similar data to that we discussed in class. Partial credit may be given for answers that are correct in part, but not in full.

**Part I: Super Tuesday 2020 (48 pts.)**

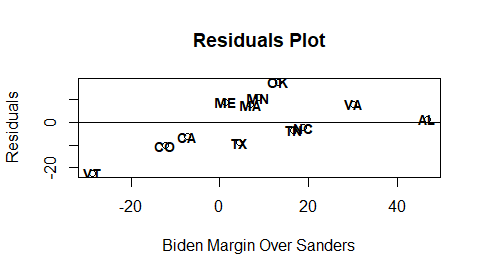
1. In class, we looked at a model of how the percentage of a state’s population that was Black related to Hillary Clinton’s margin over Bernie Sanders in 2016. Below is a model where the dependent variable (response) is Joe Biden’s margin over Bernie Sanders in 2020 (negative values of this variable would correspond to states where Sanders received a higher percentage of the vote than Biden) and the independent (explanatory) variable is the percentage of the electorate that was Black, according to exit polls. Several states (and American Samoa) did not have exit polls, so they are not included here.



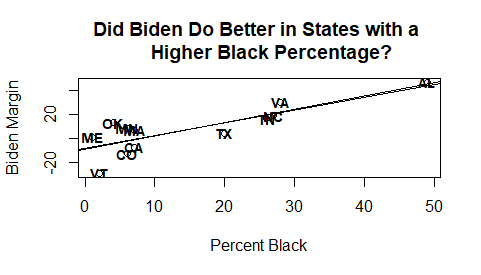
1. Please interpret the coefficient and p-value for the Pct\_Black variable and the R2 value for the model. What does this R2 value suggest about how well Pct\_Black explains the variation in how well Joe Biden performed compared to Bernie Sanders on Super Tuesday? (12 pts.)
2. Below is a scatterplot of the data and a residuals plot. Please assess each of the four requirements for the least squares line. (12 pts.)

|  |  |
| --- | --- |
|  |  |

1. Here is a scatterplot that includes state labels. Do there appear to be any outliers here? Why might one datapoint here not fit the least squares particularly well? (Hint: This requires a small amount of American politics knowledge; if you aren’t sure here, search the state’s name or abbreviation + Bernie Sanders on Google) Should we just throw out this observation? (6 pts.)



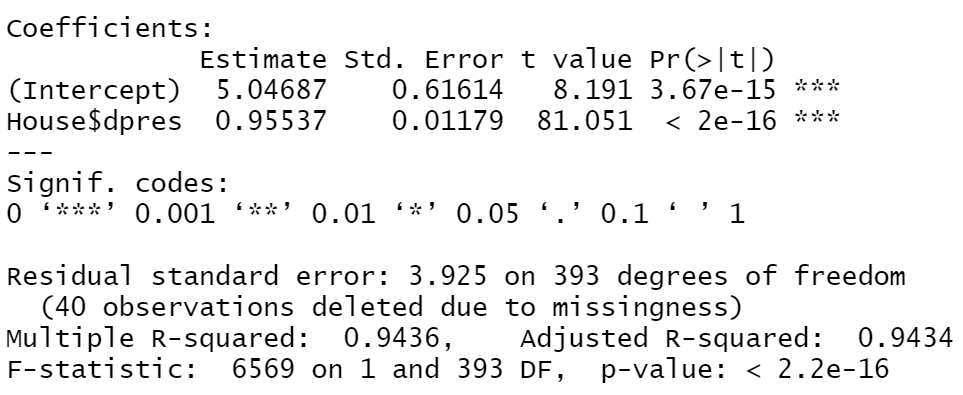
1. One state (Alabama) had a much larger Black percentage than any other state. What kind of point could this potentially be based upon its location compared to other points? Below, I’ve included a scatterplot with the least squares line with all states included and the least squares line for a model that omitted the Alabama observation. Based upon this, should we be especially concerned about this point? (4 pts.)



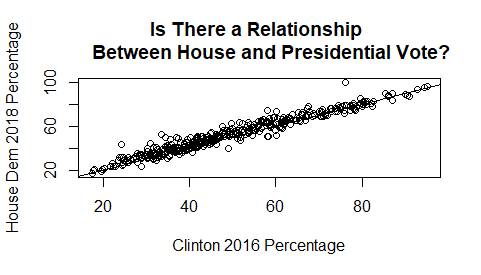
1. The correlation (Pearson’s r) between the percentage of a state’s electorate that is Black and Joe Biden’s margin over Bernie Sanders is 0.825. The standard deviation of a state’s Percent Black is 19.71. What is the standard deviation of Joe Biden’s margin over Bernie Sanders? How do you know? (6 pts.)
2. Sometimes, rather than using the “one-unit increase” language, a researcher will look at what a one standard deviation increase in an independent variable translate to in terms of predicted change in the dependent variable. Using your value from part d., carry out and interpret that analysis here. (4 pts.)
3. Let’s say that instead of Percent Black, we had used a dummy variable for whether the state was located in the South as our predictor. Why would we not want to look at one-standard deviation increases in this variable in our interpretation? (4 pts.)

**Part II: The 2018 House Elections (52 pts.)**

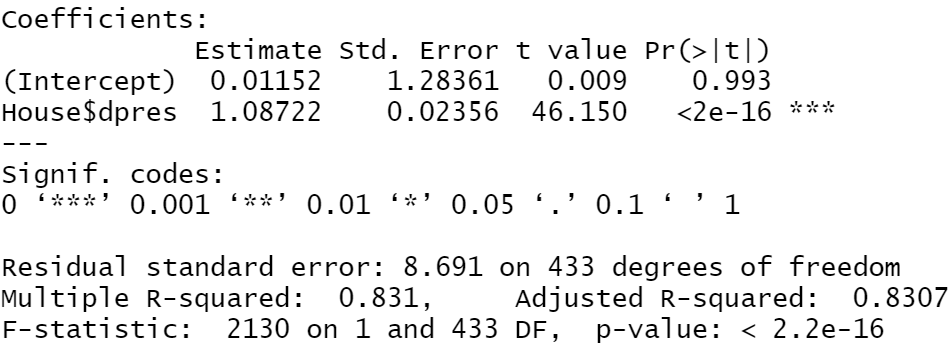
1. In this part of the problem set, you will be working with a regression I ran on the performance of Democratic candidates in 2018 House Elections. Here, the dependent variable is the percentage of the two-party vote received by the Democratic House candidate and the independent variable is the percentage of the two-party vote received by Hillary Clinton in 2016. (22 pts.)

****

* 1. Please interpret the coefficient and p-value for the dpres variable and the R2 value for the model. What does this R2 value suggest about how well presidential vote explains performance in House elections? (12 pts.)
  2. What is the correlation (Pearson’s r) between Democratic two-party presidential vote and the Democratic two-party House vote? How do you know? (8 pts.)
  3. Below is a scatterplot of these two variables. How well does the model appear to fit the data? Do there appear to be any outliers? (4 pts.)



1. The previous model did not include unopposed races (i.e., races that either did not have a Democrat or a Republican. Now, I ran a model with all 435 districts, assigning districts with no Democrat a 0 for this variable and districts with no Republican a 100. (28 pts.)



* 1. Please interpret the coefficient and p-value for the dpres variable and the R2 value for the model. What does this R2 value suggest about how well presidential vote explains performance in House elections? Why is it lower than in the previous model? (Hint: Think about what percentage of the vote a presidential candidate would usually receive, even in their worst districts.) (14 pts.)
  2. Washington, D.C. does not currently have a full voting member in the House of Representatives, but there is a move to grant [DC statehood.](https://www.nytimes.com/2020/06/26/us/politics/dc-statehood-house-vote.html) In 2016, Hillary Clinton received 95.7% of the two-party vote in Washington, D.C. Using this model, predict what percent of the vote a Democratic candidate for Congress would receive in Washington, D.C. if it were a state. What does this suggest about extrapolation? Please show your work. (8 pts.)
  3. What is the 95% confidence interval around the coefficient for the Democratic presidential vote percentage for this variable? What does this suggest about the statistical significance of this predictor. Please show your work. (6 pts.)