Homework Two

Yael Beshaw

2025-01-03

Using the data set you created in HW 1: create a regression model to explain/predict the Democratic vote share in 2020.

```
library(readr)
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.0.1
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidycensus)
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.3.2
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(huxtable)
```

```
## Warning: package 'huxtable' was built under R version 4.3.3
```

```
##
## Attaching package: 'huxtable'
##
## The following object is masked from 'package:dplyr':
```

```
##
##     add_rownames
##
## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

```r
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(ggdist)
```

```
## Warning: package 'ggdist' was built under R version 4.3.2
```

```r
#democratic vote share == Biden's Share from HW#1
HW1_data <- read_csv("~/Downloads/HW1_data.csv")
```

```
## Rows: 51 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): state
## dbl (4): Donald Trump, Joe Biden, Biden's Share, Median Household Income
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Adding predictors to dataset from HW#1

```r
v20 <- load_variables(2020, "acs5", cache = TRUE)

HW2_data <- get_acs(geography = "state",
                    variables = c(median_income ="B19013_001", #mi,
                                  race = "B02001_001", #race
                                  white_race = "B02001_002",
                                  black_race = "B02001_003",
                                  asian_race = "B02001_005",
                                  education ="B29002_001", #educ
                                  no_diploma = "B29002_002",
                                  high_school = "B29002_004",
                                  bachelors = "B29002_007",
                                  grad = "B29002_008",
                                  age = "B29001_001", #age
                                  eighteen = "B29001_002",
                                  thirty = "B29001_003",
                                  fourty_five = "B29001_004",
                                  sixtyfive_plus="B29001_005",
                                  people = "B01001_001", # sex
                                  male = "B01001_002"),
                    year = 2020)
```

```
## Getting data from the 2016-2020 5-year ACS
```

```r
HW2_data  <- HW2_data|>
  select(NAME,variable,estimate)|>
  pivot_wider(names_from = variable,
              values_from = estimate)|>
  arrange(NAME)
```

```r
HW2_data <-HW1_data|>
  left_join(HW2_data, by=join_by(state == NAME))
```

```r
HW2_data.adj <- HW2_data |>
  mutate (`Prop_Male` = male/people) |>
  mutate (`Prop_White` = white_race/race) |>
  mutate(`Prop_NonWhite`= (black_race + asian_race)/race) |>
  mutate (`Prop_Grad` = (bachelors + grad)/education)|>
  mutate (`Prop_Highschool_Lower`= (no_diploma + high_school)/education)|>
  mutate (`GenZ_Mill` = (eighteen + thirty)/age) |>
  mutate (`GenX` = fourty_five/age) |>
  mutate (`Boomers_Silent` = sixtyfive_plus/age)|>
  mutate (`Median Household Income`= `Median Household Income`/10000)|>
  select (`Biden's Share`, `Median Household Income`, `Prop_Male`,
          `Prop_White`, `Prop_NonWhite`, `Prop_Grad`,
          `Prop_Highschool_Lower`, `GenZ_Mill`, `GenX`)
```

Provide descriptive statistics, regression results and diagnostics, and make a case for why your model is a good one.

Descriptive Statistics

```r
#this dataset contains variables that are listed as proportions of the overall population for interpret
summary(HW2_data.adj)
```

```
##  Biden's Share    Median Household Income   Prop_Male        Prop_White
##  Min.   :0.2752   Min.   :4.651           Min.   :0.4746   Min.   :0.2415
##  1st Qu.:0.4107   1st Qu.:5.750           1st Qu.:0.4880   1st Qu.:0.6697
##  Median :0.5012   Median :6.301           Median :0.4928   Median :0.7656
##  Mean   :0.4971   Mean   :6.505           Mean   :0.4940   Mean   :0.7429
##  3rd Qu.:0.5819   3rd Qu.:7.379           3rd Qu.:0.4983   3rd Qu.:0.8395
##  Max.   :0.9447   Max.   :9.084           Max.   :0.5219   Max.   :0.9368
##  Prop_NonWhite       Prop_Grad      Prop_Highschool_Lower   GenZ_Mill
##  Min.   :0.01374   Min.   :0.1976   Min.   :0.2029        Min.   :0.3841
##  1st Qu.:0.06830   1st Qu.:0.2649   1st Qu.:0.2875        1st Qu.:0.4334
##  Median :0.13420   Median :0.2953   Median :0.3087        Median :0.4456
##  Mean   :0.15544   Mean   :0.3023   Mean   :0.3111        Mean   :0.4497
##  3rd Qu.:0.21797   3rd Qu.:0.3353   3rd Qu.:0.3358        3rd Qu.:0.4583
##  Max.   :0.49488   Max.   :0.5565   Max.   :0.4359        Max.   :0.5884
##       GenX
##  Min.   :0.2544
##  1st Qu.:0.3244
##  Median :0.3341
##  Mean   :0.3318
##  3rd Qu.:0.3419
##  Max.   :0.3685
```

Regression Results

```r
model<-lm(`Biden's Share` ~ ., data= HW2_data.adj)
summary(model)
```

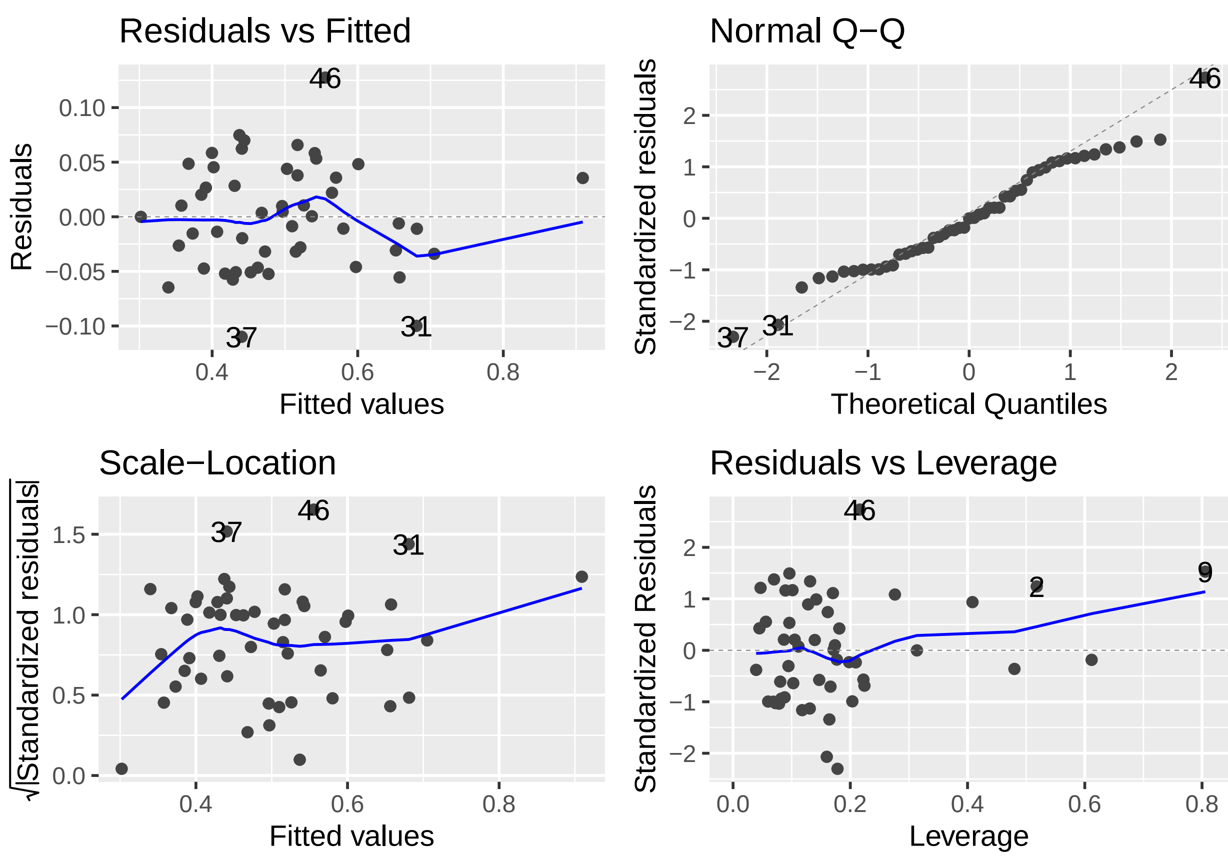
```
##
## Call:
## lm(formula = `Biden's Share` ~ ., data = HW2_data.adj)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.109942 -0.032933 -0.000076  0.036858  0.127501
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.651927   1.086707   2.440 0.018974 *
## `Median Household Income` -0.002225   0.019007  -0.117 0.907369
## Prop_Male                -4.006439   1.972888  -2.031 0.048643 *
## Prop_White               -0.596342   0.160136  -3.724 0.000578 ***
## Prop_NonWhite            -0.514700   0.222792  -2.310 0.025859 *
## Prop_Grad                 1.535779   0.389870   3.939 0.000303 ***
## Prop_Highschool_Lower    -0.042019   0.361187  -0.116 0.907940
## GenZ_Mill                -0.406673   0.487347  -0.834 0.408741
## GenX                      0.281265   0.903012   0.311 0.756981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05269 on 42 degrees of freedom
## Multiple R-squared:  0.8464, Adjusted R-squared:  0.8172
## F-statistic: 28.93 on 8 and 42 DF,  p-value: 1.03e-14
```

Based on the regression results, for every one unit increase in the median household income of a state (including D.C.) in the United States. The expected mean for Biden's Share of the vote decreases by 0.002 units. This is not statistically significant as we observe the p-value to less than 0.05 and we therefore fail to reject the null hypothesis that there is is no effect of median household income. Additionally,for every one unit increase in the proportion of men within a state (including D.C.) in the United States, the expected mean for Biden's Share of the vote decreases by 4.00 units and is statistically significant. Indicating that men are not as likely to contribute to the vote share of the Democratic party. We see a similar pattern when assessing the proportion of White voting age citizens compared, in which for every one unit increase in the proportion of White voters within a state (including D.C.) in the United States, the expected mean for Biden's Share of the vote decreases by 0.596 units and is statistically significant. Interestingly enough, we also see that for Non-White voters (Black and Asian in this analysis), there is a decrease of the expected mean for Biden's Share of the vote by 0.514 units, which is statistically significant. This means that although White voters are less likely to contribute to the vote share of the Democratic party than other races, Non-white voters do not contribute as much as expected by literature and theory. Lastly, for every one unit increase in the proportion of those who graduated undergrad and graduate school, the expected mean for Biden's Share of the vote increases by 1.535 units and is statistically significant. This is expected according to literature and theory regarding voting patterns of increasingly educated voters. We see that voters of lower educational attainment do not have a statisitically significant effect on the vote share of the Democratic party. We also see that no proportion of voting-eligible age group has a statistically significant effect either.

Additionally, the residual standard error of 0.05269 both the Adjusted R^2 and Multiple R^2 explain about 80% of the variance The F-statistic is 28.93 and is statistically significant at 3.482e-09

Diagnostics, Test the Assumptions - Linearity, Homoscedastisticity, Normality of Residuals, Influential Data Points

```r
autoplot(model)
```



Based on these results, we can see that there is a slight curve in the Residuals vs Fitted and Scale-Location plots which may be cause for concern regarding violation of homoscedasticity assumption. In order to take a closer look at this utilizing the Breusch–Pagan test. Additionally, we see that the Q-Q plot flows a straight line between -1.8 to 1.8, indicating that we may need to assess the data points that could have an influential impact on the model. We see that in the Residuals vs Leverage plot that there are several points that could potentially have a great influence.

```r
#check homoscedasticity
ncvTest(model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.09186774, Df = 1, p = 0.76182
```

The results support that there is no evidence of heteroscedasticity as the p-value is greater than 0.05. Therefore the assumption of homoscedasticity is not violated.

```r
#check mulitcolinearity
vif(model)
```

```
## `Median Household Income`                 Prop_Male                Prop_White
##                  7.947900                  4.878106                  8.624682
##             Prop_NonWhite                 Prop_Grad     Prop_Highschool_Lower
##                 11.754775                 10.449493                  3.944462
##                 GenZ_Mill                      GenX
##                  5.286982                  5.291380
```

The results indicate that there are four predictors; Median Household Income, Prop_White, Prop_NonWhite, and Prop_Grad that may indicate strong mulitcolinearity. In order to assess this, we can take a look at these specific predictors to address the issue.

This model can be utilized to make predictions about the vote share of the Democratic party as we have evidence that there is no evidence of heteroscedasticity due to the non-significant test from the Breusch–Pagan test. Additionally, we see that in the Q-Q plot, there is not a significant deviation away from the straight line which supports that the normality assumption is not violated. We see that there is a chance for influential points that may be impacting our estimates. However, we are able to conduct further tests and adjust by removing these points, if necessary. Our main concern here would be multicolinearity, as evidenced by the VIF model. As a result, we may consider transformations and adding/removing predictors that ultimately may not help us make accurate predictions.