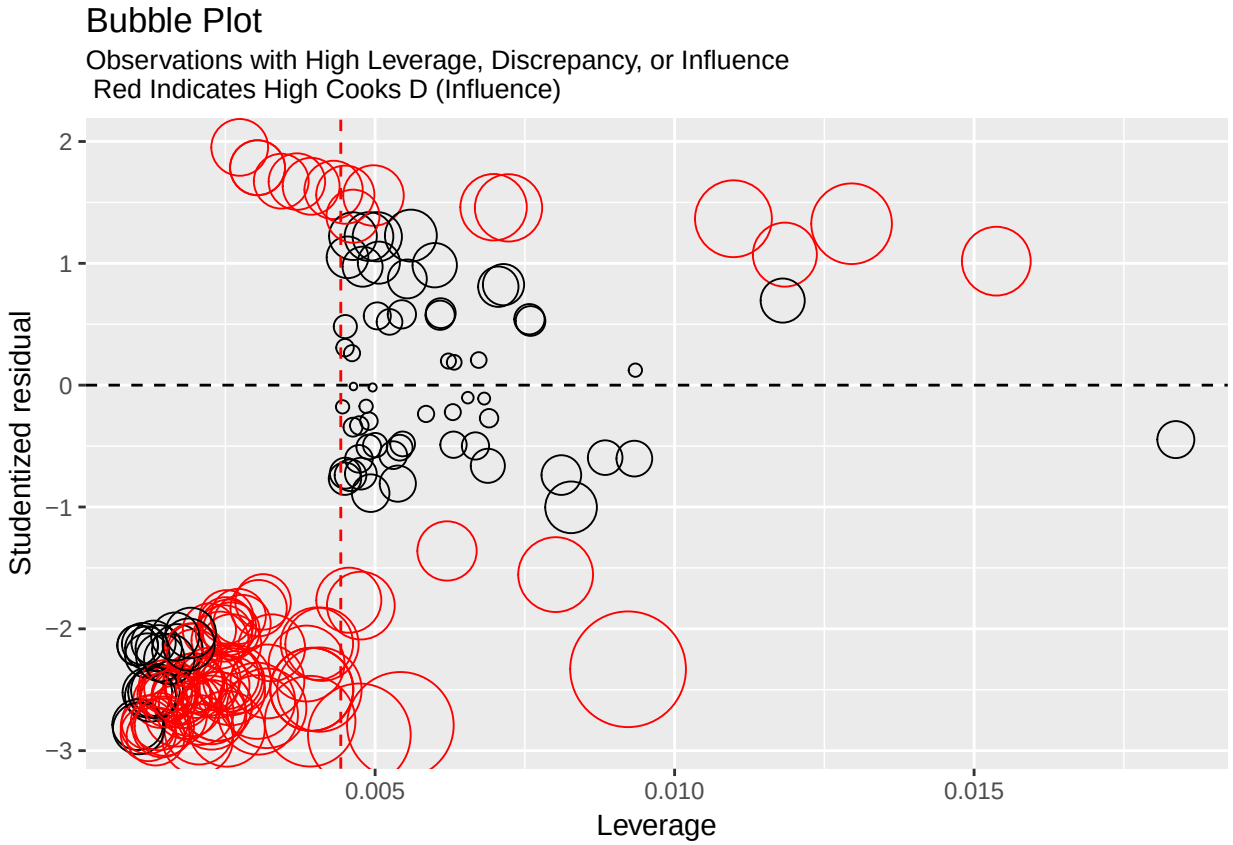
R Notebook

```
knitr::opts_chunk$set(message = FALSE,
                      warning = FALSE,
                      echo = FALSE)
```

```
##
## Call:
## lm(formula = biden ~ age + female + educ, data = .)
##
## Coefficients:
## (Intercept)          age       female         educ
##    68.62101      0.04188      6.19607     -0.88871
```

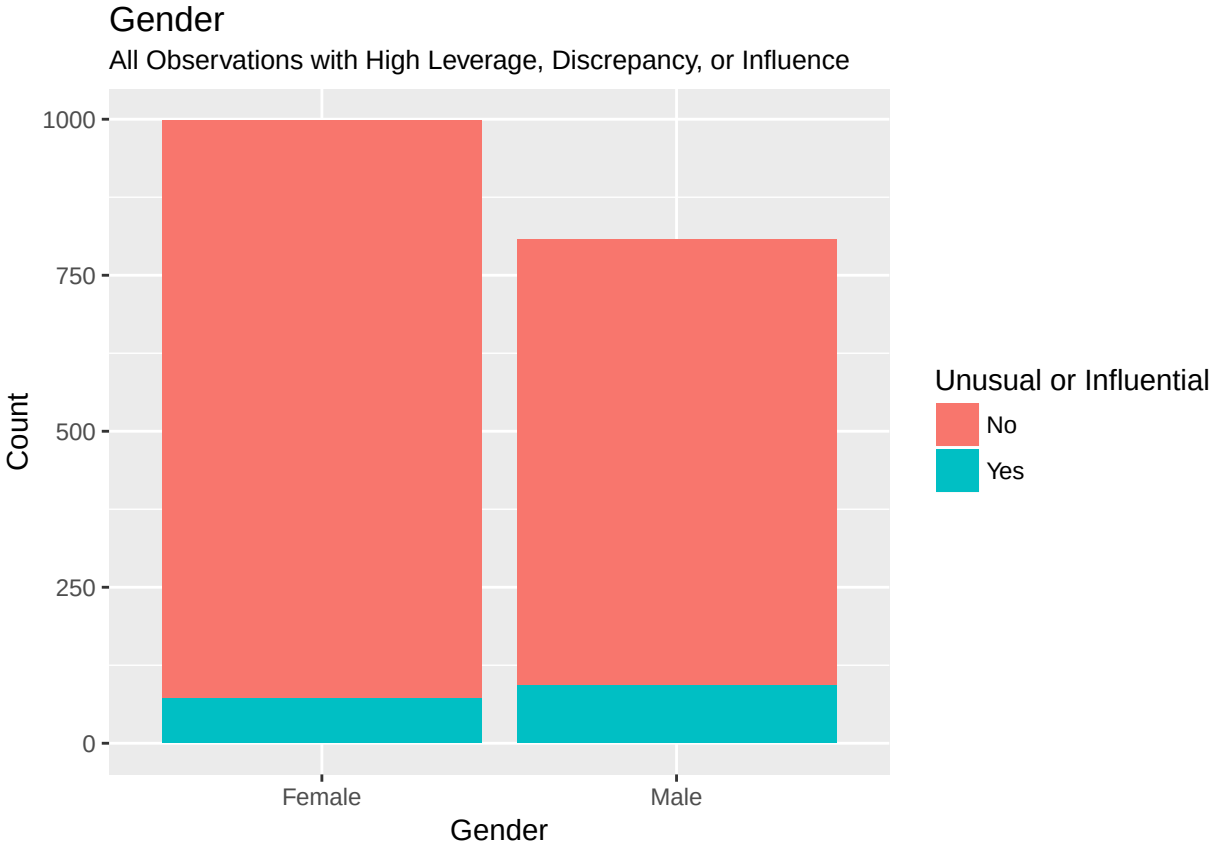
1.

One thing that can be done to find higly influential observations is to look at the effect each observation has on the coefficients by a few different measures: leverage, discrepency, and Cook's D (leverage x discrepency). Below I calculate these values for each observation and plot the leverage and residual of each observation. Points in red have the high Cook's D values.

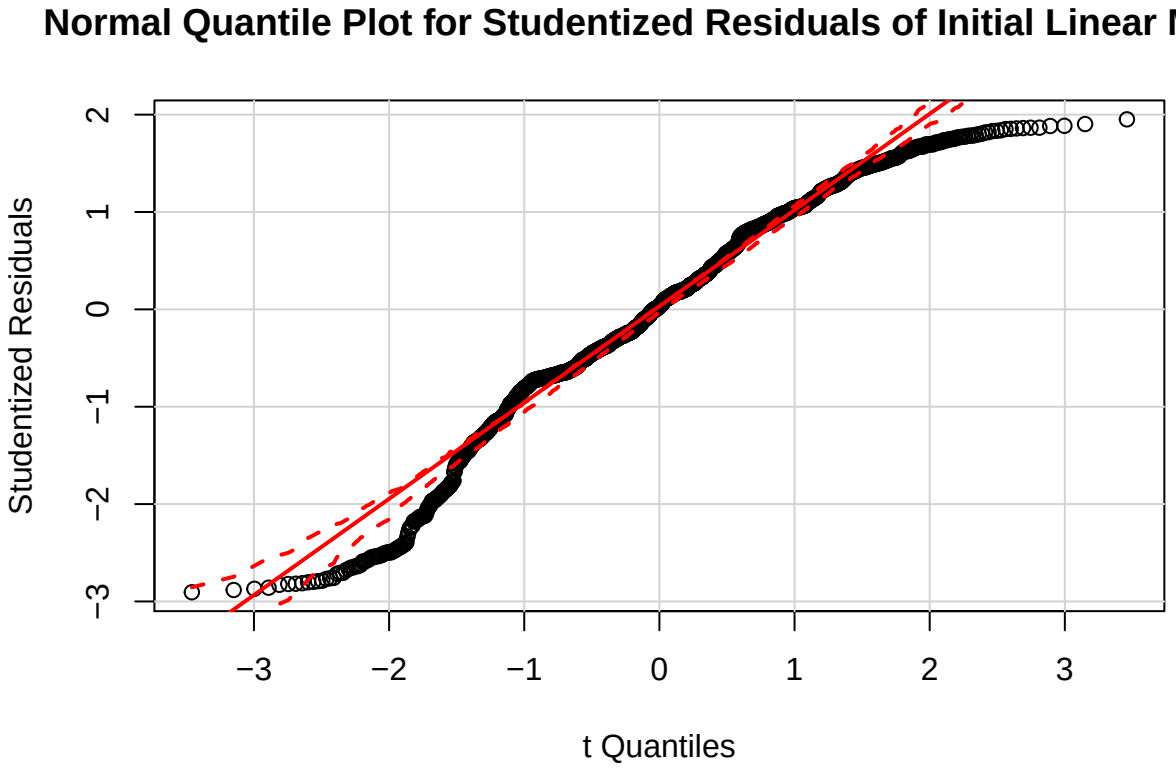


From these results I would suggest to look at observations with a high Cook's D value to see if there is an obvious reason why. One issue with our dataset is that there are not a lot of observations (1,807 after omitting variables). I would want to make sure the makeup of our dataset is also representative of the whole population. For example if we find that a certain subset of observaions with a high Cook's D value are not

common in the entire population it might make sense to drop those observations. Below I plot a histogram of male's and females observations with the number of inflential observations.



For example, in the histogram above you can see that 11.6% of males have a high discrepency, influence, or leverage whereas only 7.3% of females do. This is concerning to me because if I want to extrapolate my results to a population with a 50/50 male/female ratio the results of my regression might now because male observations are having a higher influence on the results. To correct for this I might omit some of my male observations.



From the plot of the error above, it does not look normally distributed (especially on the outer quantiles). In order to fix this I would attempt to add or combine or transform some of the independent variables to see if the error term is closer to a normal distribution.

3. To test for heteroscedasticity in the model I will perform the Bresuch-Pagan test.

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_init_biden
## BP = 22.559, df = 3, p-value = 4.989e-05
```

My value is less than 0.05 which means that there is heterscedasticity in the model or that the errors of the coeficcient estimates are not of constant variance. This could influence the estimates for the coefiecients and standard errors in our model.

4. To test for multicollinearity I will look at the VIF for each coefficient.

```
##      age   female     educ
## 1.013369 1.001676 1.012275
```

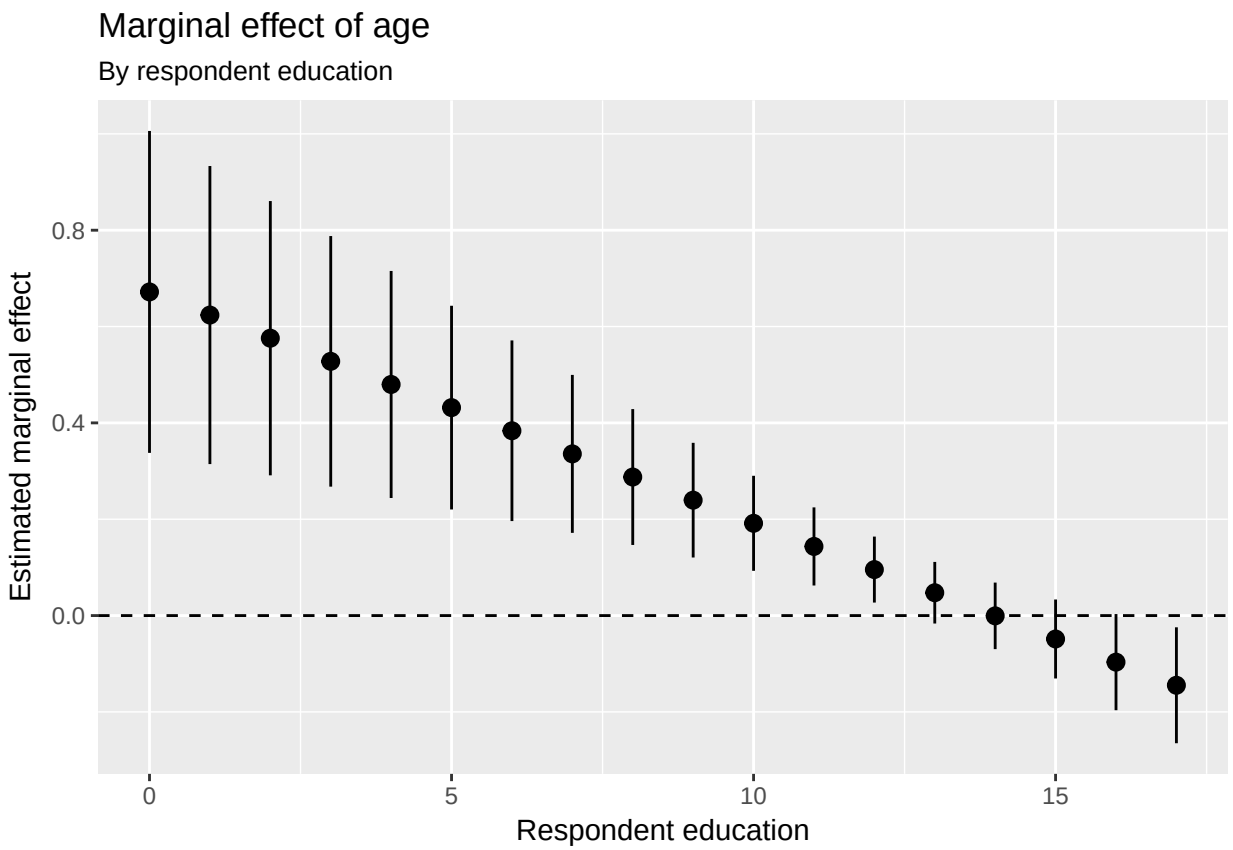
Since the values for each variable are less than 10 (which is a good rule of thumb) I do not believe there is any colinearity in this simple model.

2. Interaction Terms

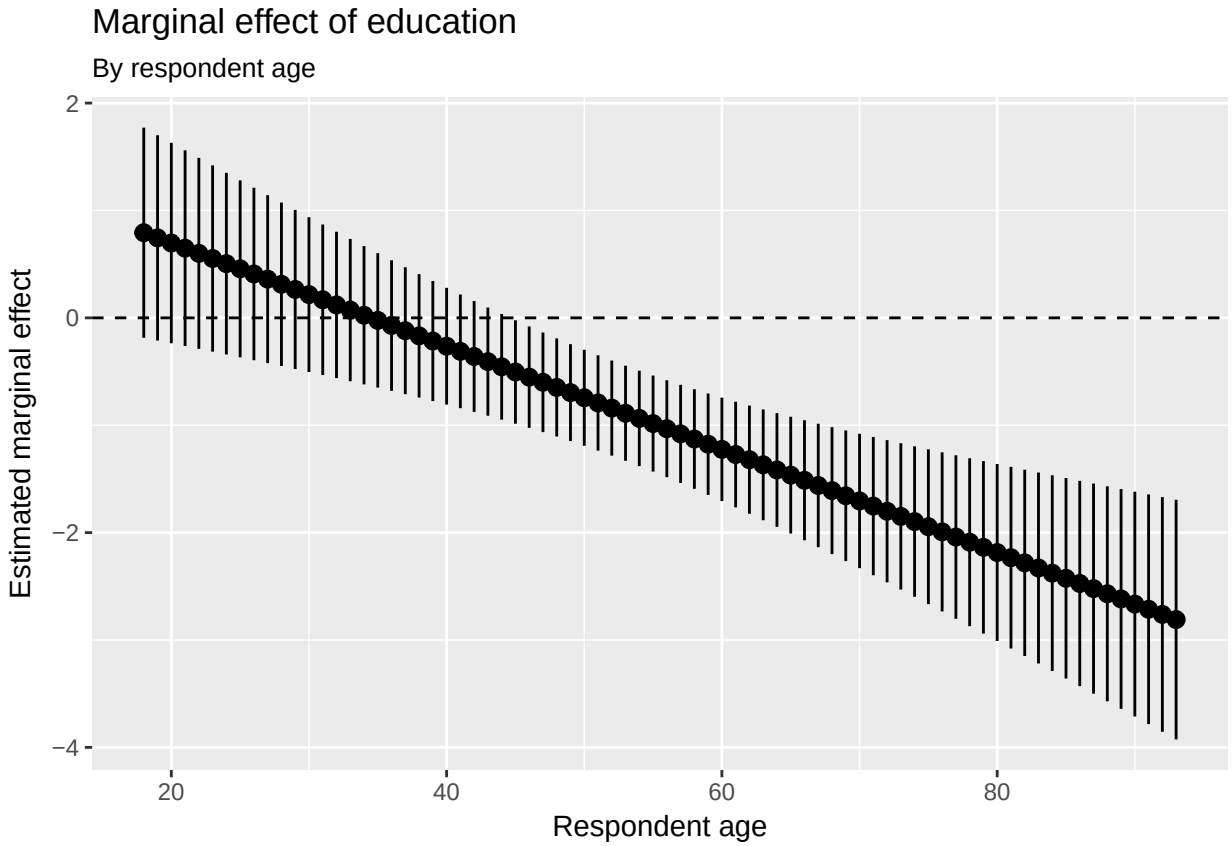
```
##
## Call:
## lm(formula = biden ~ age + educ + age * educ, data = .)
##
## Coefficients:
## (Intercept)          age         educ     age:educ
```

```
##    38.37351      0.67187      1.65743     -0.04803
```

```
## Linear hypothesis test
##
## Hypothesis:
## age  + age:educ = 0
##
## Model 1: restricted model
## Model 2: biden ~ age + educ + age * educ
##
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1   1804 985149
## 2   1803 976688  1    8461.2 15.62 8.043e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



To evaluate the marginal effect of age on Joe Biden thermometer, conditional on education I plotted the estimated marginal effect of age by education level. It can be seen here that the marginal effect decreases as education increases. I also ran the Wald Test to find the pvalue to find that the the marginal effect of education on age's impact is significant




```
## Linear hypothesis test
##
## Hypothesis:
## age  + age:educ = 0
##
## Model 1: restricted model
## Model 2: biden ~ age + educ + age * educ
##
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1   1804 985149
## 2   1803 976688  1    8461.2 15.62 8.043e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similarlly, I plotted the marginal effect of education on biden feeling conditional on respondant age. It also decreases as age increases. The significance is also confirmed via the Wald Test because the p-value is less than 0.05.

3. Missing Data

```
## # A tibble: 1,807 × 7
##    biden female   age  educ   dem   rep    ID
##    <int>  <int> <int> <int> <int> <int> <int>
## 1     90      0    19    12     1     0     1
## 2     70      1    51    14     1     0     2
## 3     60      0    27    14     0     0     3
## 4     50      1    43    14     1     0     4
```

```
## 5     60      1    38    14     0     1     5
## 6     85      1    27    16     1     0     6
## 7     60      1    28    12     0     0     7
## 8     50      0    31    15     1     0     8
## 9     50      1    32    13     0     0     9
## 10    70      0    51    14     1     0    10
## # ... with 1,797 more rows
```

```
##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : biden_data %>% dplyr::select(-c(biden, female, dem, rep))
##
##   HZ      : 11.96555
##   p-value : 0
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------
```

```
## $`Descriptive Statistics`
## NULL
##
## $`Shapiro-Wilk's Normality Test`
##    Variable Statistic   p-value Normality
## 1    age       0.9795         0    NO
## 2   educ       0.9180         0    NO
## 3    ID        0.9549         0    NO
```

I first test for multivariate normality using the Henze-Zrkler Test to see if the variables are distributed as a multivariate normal distribution. The result of the test states that the data are not multivariate normal. In order to fix this I will try to transform the variables and retest. For example, below I take the sqrt of age and see that my HZ value improves, though it still don't pass the Shapiro-Wilk test.

```
## [1] "Sqrt age and educ"
```

```
##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : biden_data %>% dplyr::select(sqrt_educ, sqrt_age)
##
##   HZ      : 15.33627
##   p-value : 0
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------
```

```
## $`Descriptive Statistics`
## NULL
##
## $`Shapiro-Wilk's Normality Test`
##    Variable Statistic   p-value Normality
## 1 sqrt_educ    0.8639         0    NO
## 2 sqrt_age     0.9841         0    NO
```

I will now compre this model with the previous one.

```
## -- Imputation 1 --
##
##   1  2  3  4  5  6
##
```

```
## -- Imputation 2 --
##
##   1  2  3  4  5  6
##
## -- Imputation 3 --
##
##   1  2  3  4  5  6
##
## -- Imputation 4 --
##
##   1  2  3  4  5
##
## -- Imputation 5 --
##
##   1  2  3  4  5  6
```

```
##          term    estimate  std.error estimate.mi std.error.mi
## 1 (Intercept) 68.62101396 3.59600465 64.54557579   3.27172488
## 2         age  0.04187919 0.03248579  0.05604686   0.02892925
## 3      female  6.19606946 1.09669702  5.79697871   1.03588878
## 4        educ -0.88871263 0.22469183 -0.62938405   0.20569505
```

Using the imputed model does not give us very different results than the original model.