Hodge Podge Problem Set

Julian McClellan

Due 5/15/17

Regression Diagnostics

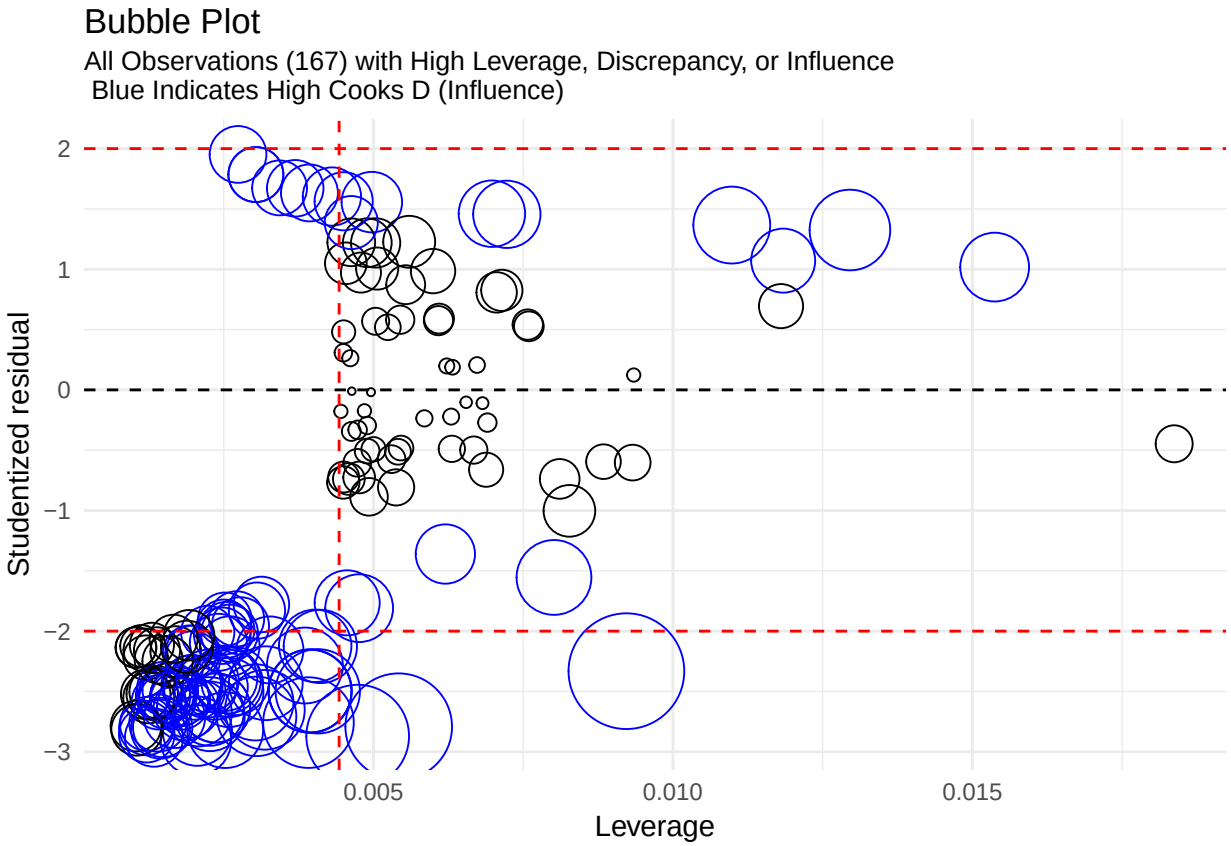
```
##
## Call:
## lm(formula = biden ~ age + female + educ, data = .)
##
## Coefficients:
## (Intercept)          age       female         educ
##    68.62101      0.04188      6.19607     -0.88871
```

1. Test the model to identify any unusual and/or influential observations. Identify how you would treat these observations moving forward with this research. Note you do not actually have to estimate a new model, just explain what you would do. This could include things like dropping observations, respecifying the model, or collecting additional variables to control for this influential effect.

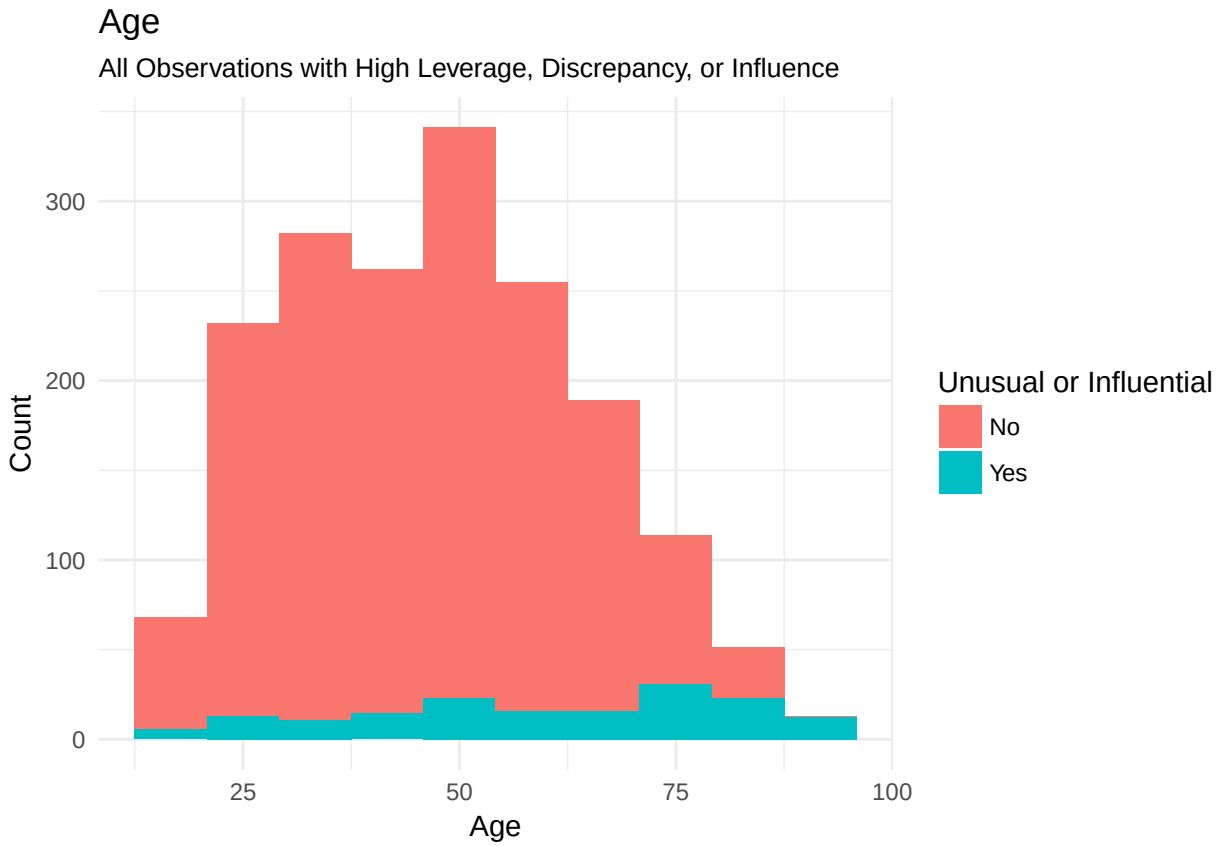
I search the data for observations with high leverage, discrepancy, or influence, displaying them in the bubble plot below. Additionally, since

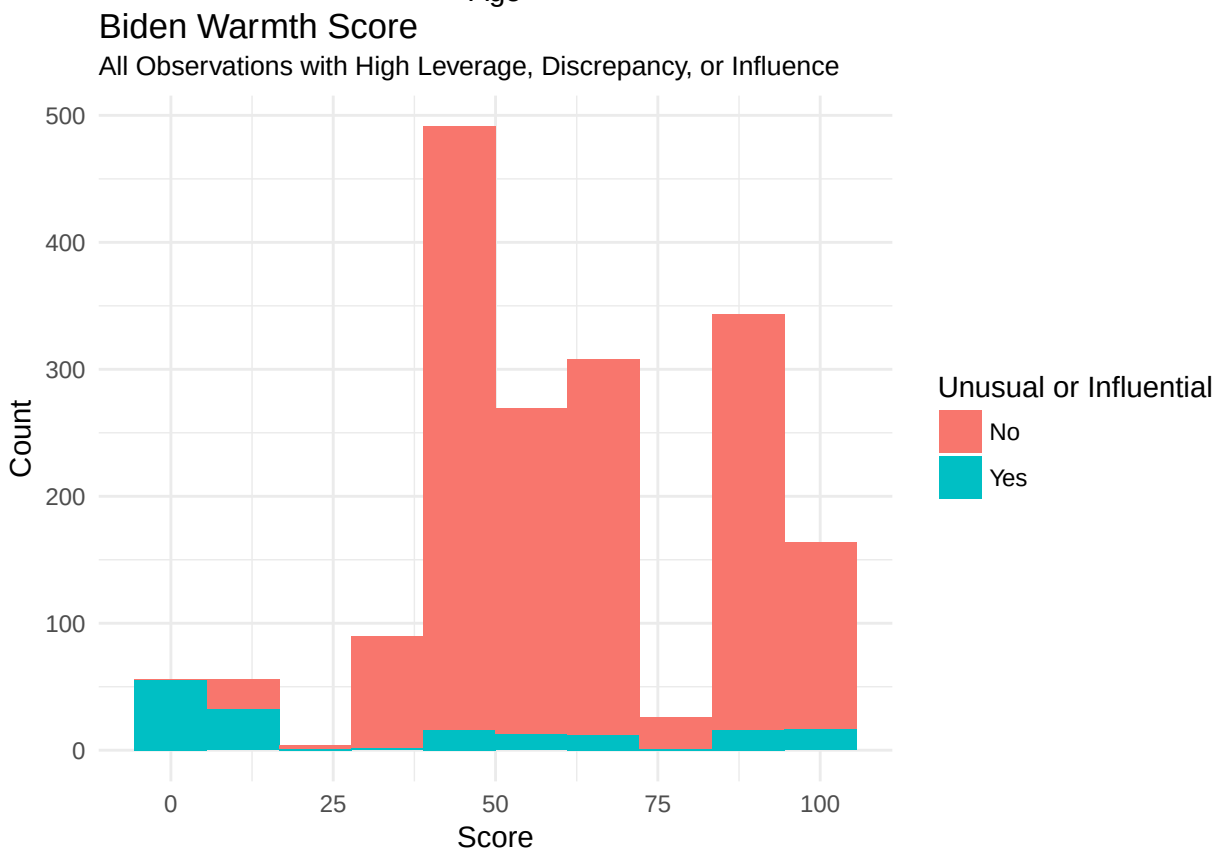
$$Influence = Leverage\, x\, Discrepancy$$

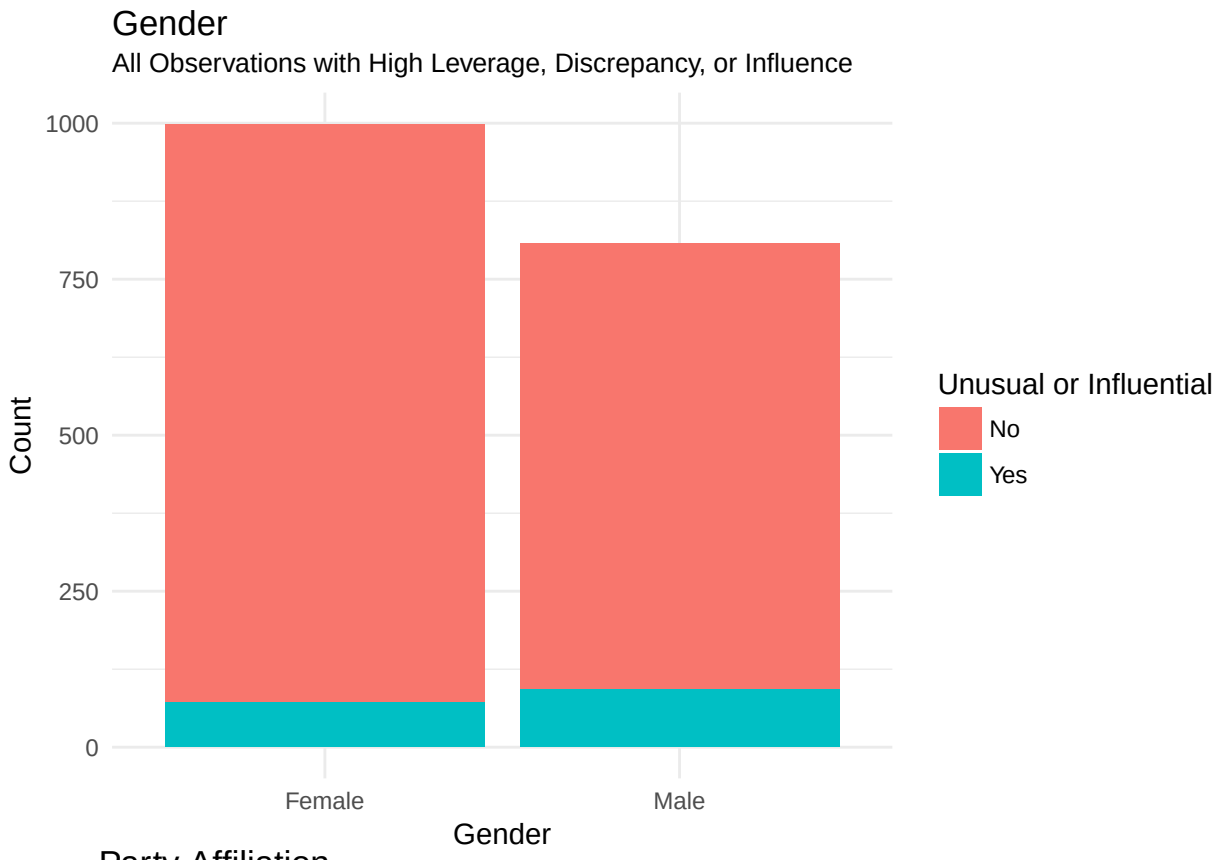
we indicate high influence (Cooks's D) values in red.

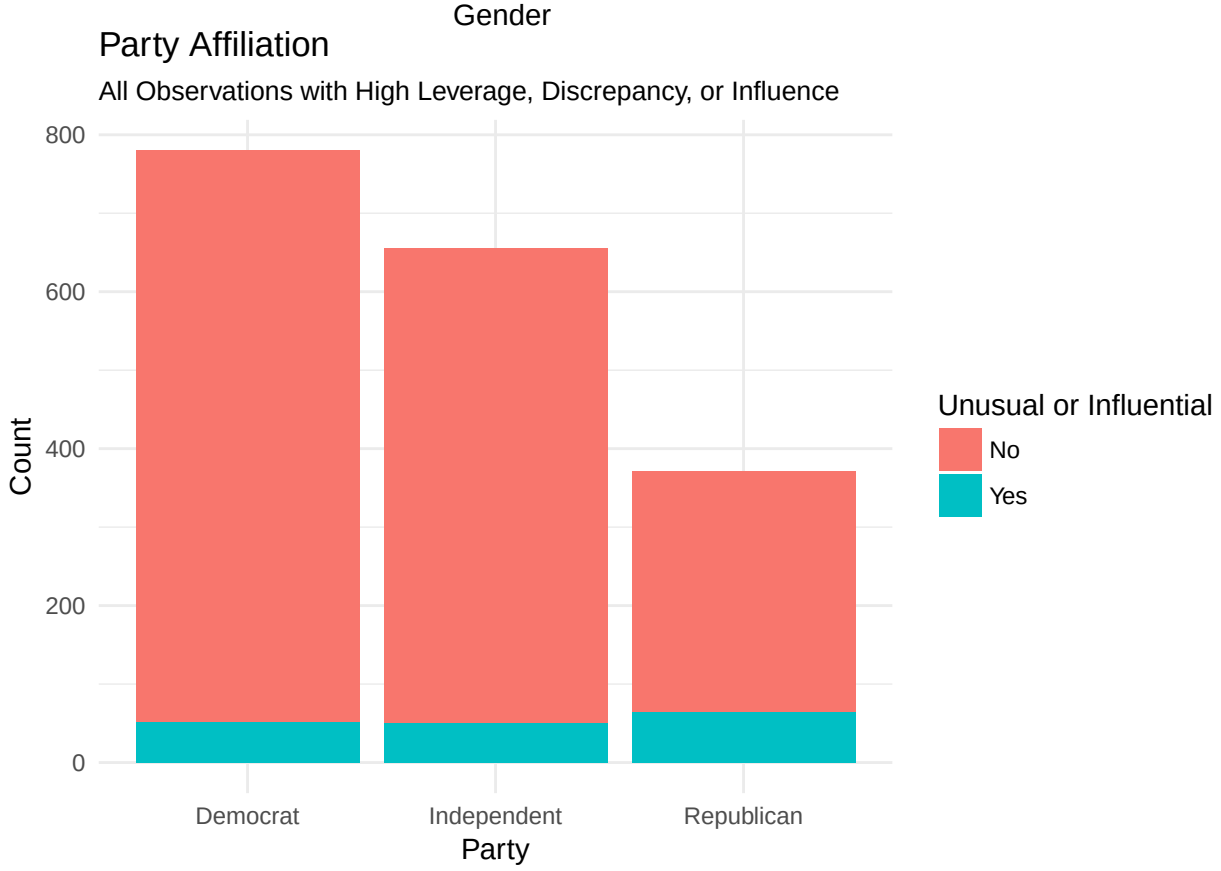


As we can see from the Bubble plot, there are 167 values with high values of leverage, discrepancy, or influence. Now we want to investigate whether these points are strange because unusual is happening to these data points. Let's look at histograms for their values of biden score, age, education level, party affiliation, and gender.


Age
All Observations with High Leverage, Discrepancy, or Influence
Count
300
200
100
0
25
50
75
100
Age
Unusual or Influential
No
Yes


Biden Warmth Score
All Observations with High Leverage, Discrepancy, or Influence
Count
500
400
300
200
100
0
0
25
50
75
100
Score
Unusual or Influential
No
Yes

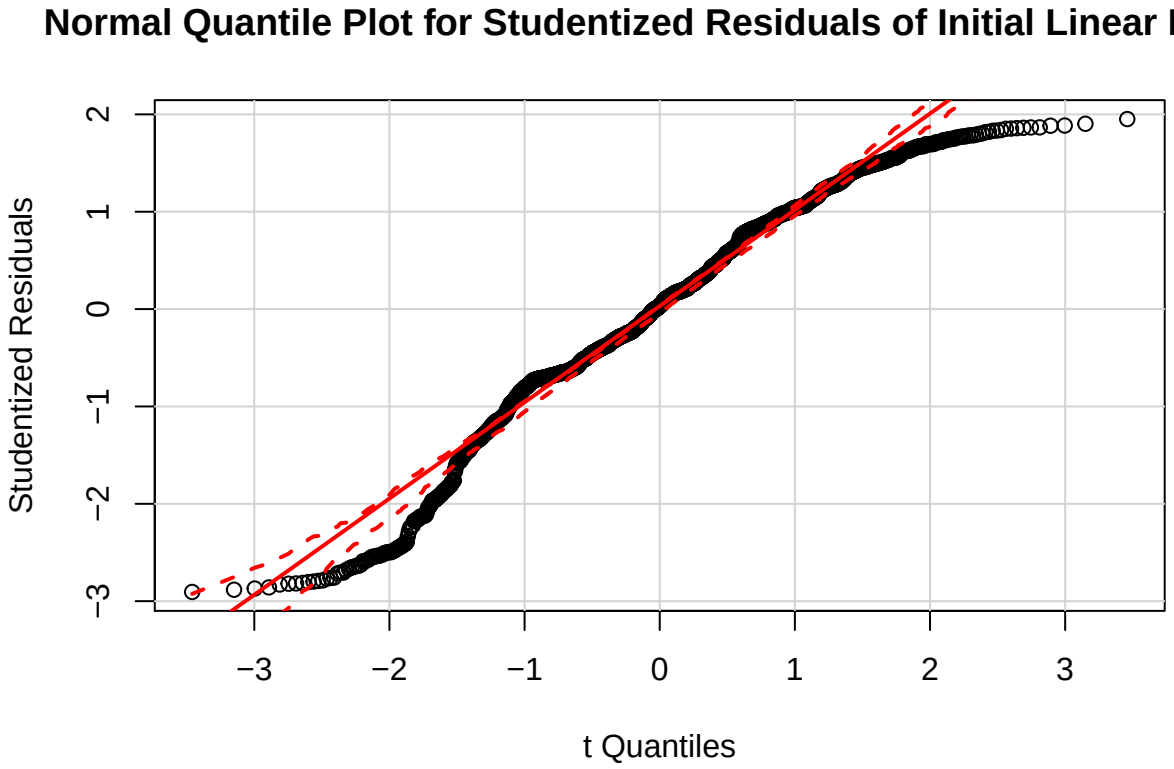




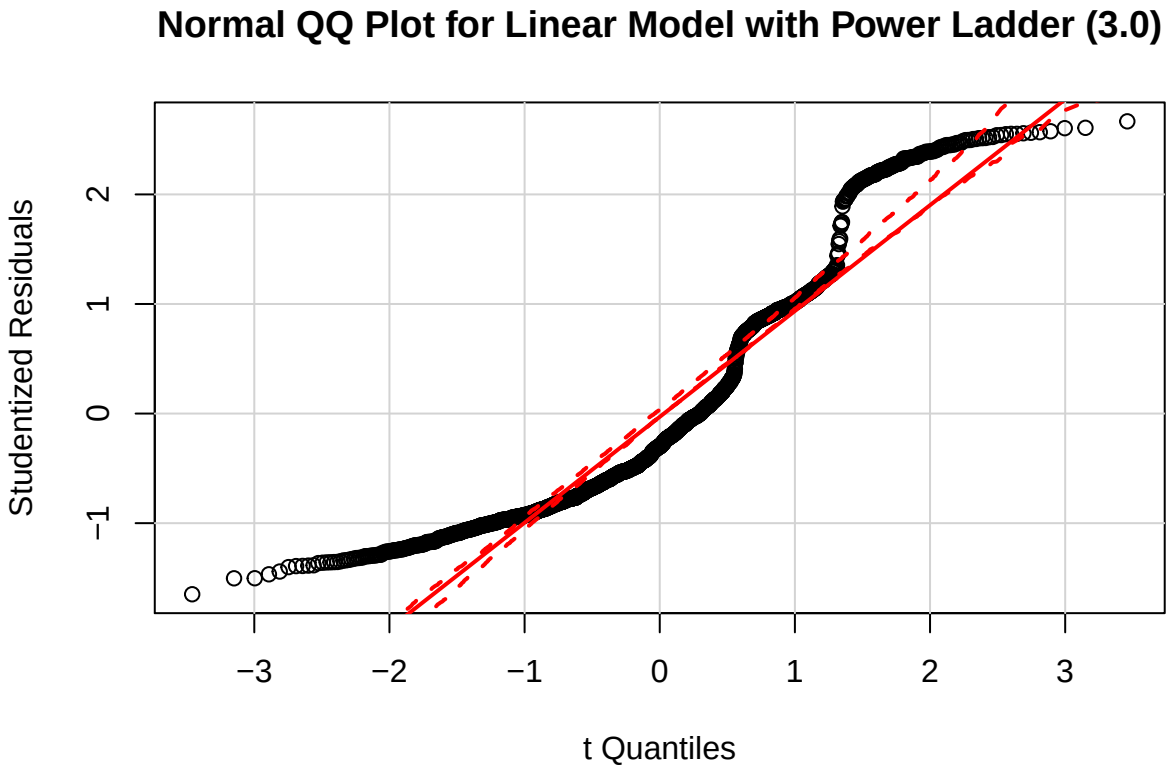
Looking at the above histograms, lower scores, older males, and Republicans seem to be more represented in

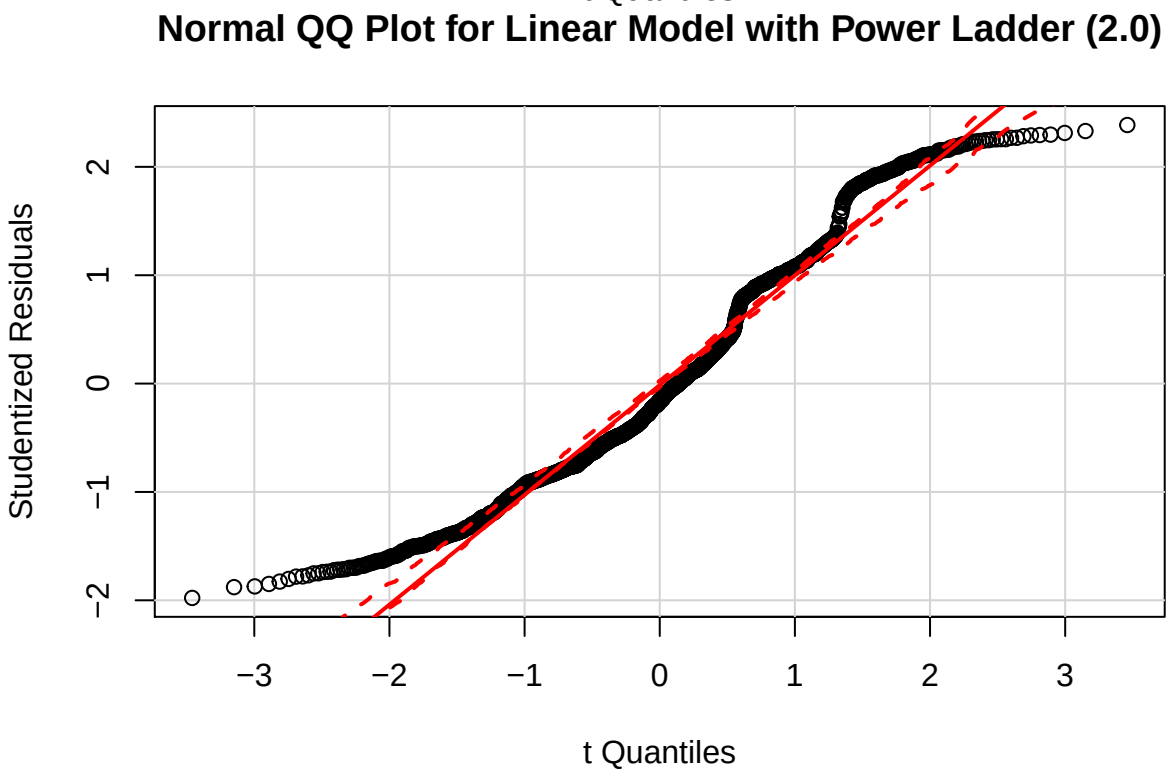
the the unusual or influential observations. To account for this, I might want to respecify my model to include for interaction terms between Republicans party affiliation and age. Note, that the given model doesn't have Republican party affiliation in the first place, so a first step could be to add it as a predictor, carry out the same process to look for unusual or influential observations, and then see try the aforementioned interaction term.

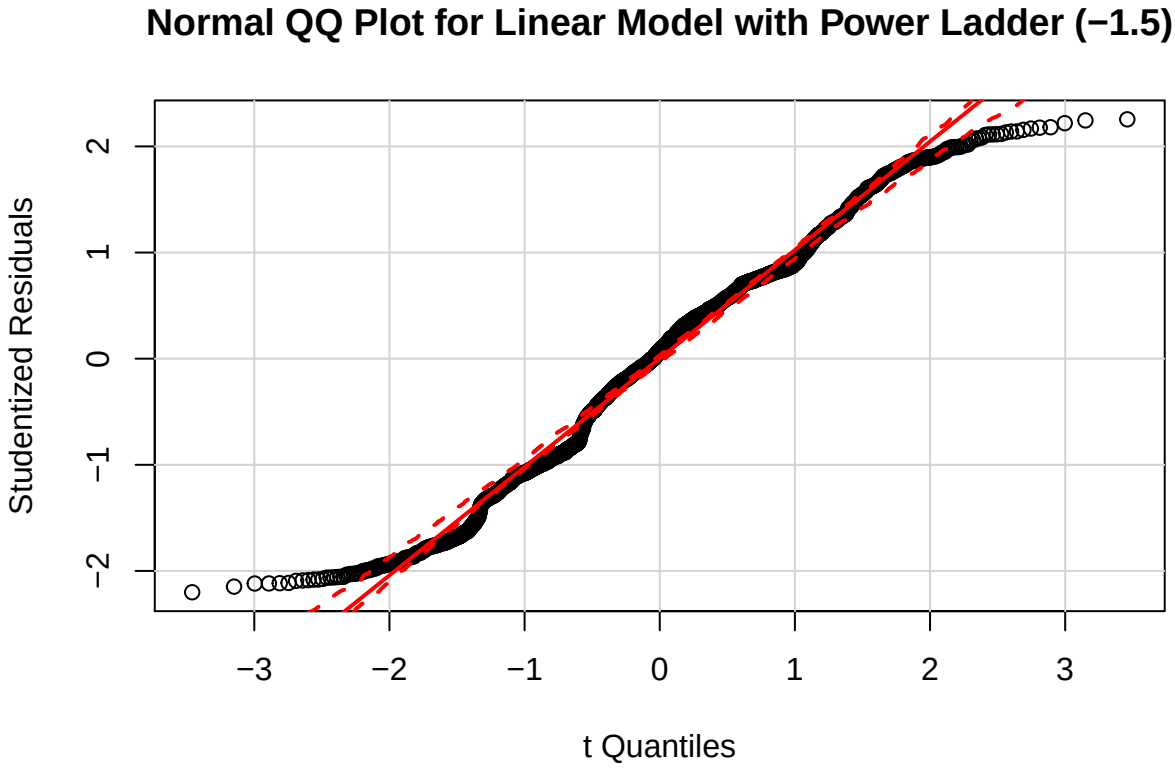
2. Test for non-normally distributed errors. If they are not normally distributed, propose how to correct for them.

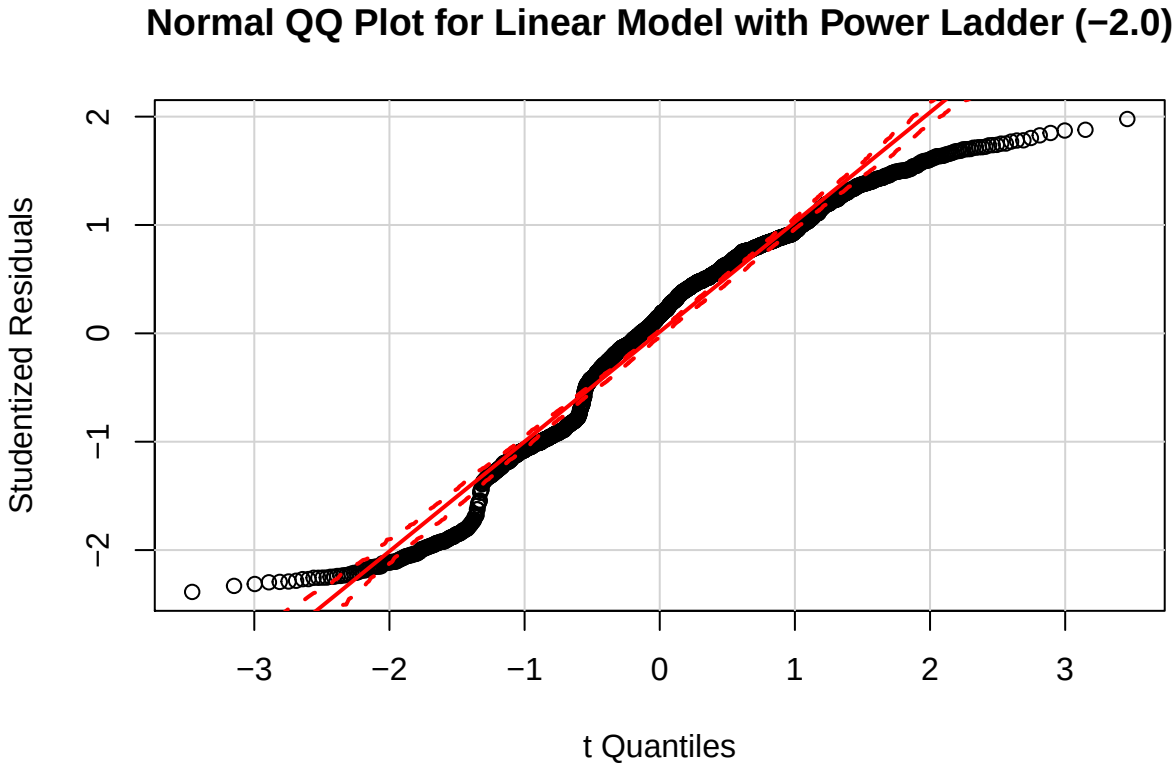


Looking at the normal Quintilian plot above, se see clear deviation from the normal distribution. In order to correct for this, the response variable, `biden`, can be transformed using Tukey's Ladder of Powers transformations. Experimenting with these transformations, I can make the errors of the linear model more normally distributed. Examples, of several transformations and their normal QQ plots are given below:


Normal QQ Plot for Linear Model with Power Ladder (3.0)
Studentized Residuals
t Quantiles
−3
−2
−1
0
1
2
3
−1
0
1
2


Normal QQ Plot for Linear Model with Power Ladder (2.0)
Studentized Residuals
t Quantiles
−3
−2
−1
0
1
2
3
−2
−1
0
1
2





3. Test for heteroscedasticity in the model. If present, explain what impact this could have on inference.

For this, we will conduct a Breusch-Pagan test.

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

With a p-value below `.05`, this suggest that there is heteroskedasticity present in the errors for our model. If left unaccounted for, this could distort the estimates for the standard error for each coefficient either up or down.

4. Test for multicollinearity. If present, propose if/how to solve the problem.

For this, let's simply take a look at the variance inflation factors for our three coefficients in our model.

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

Since none of the variance inflation factors are above `10`, this suggests that we do not have to take steps to account for multicollinearity in the model.

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
```

1. Evaluate the marginal effect of age on Joe Biden thermometer rating, conditional on education. Consider the magnitude and direction of the marginal effect, as well as its statistical significance.

Firstly, we note that the model estimated above has the following form:

$$E(biden) = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 age * educ$$

In order to evaluate the marginal effect of `age` on `biden` conditional on `educ` we take:

$$\frac{\delta E(biden)}{\delta age} = \beta_1 + \beta_3 educ$$

From the model summary, we have the values for β_1 and β_3, inserting them into the equation, we have:

$$\frac{\delta E(biden)}{\delta age} = 0.67187 + -0.04803 educ$$

We see that the marginal effect of `age` on `biden` conditional on `educ` has variable magnitude. For values of $educ < 14$ the effect on `biden` is positive, but for $educ \geq 14$ the effect is negative.

Now, is this marginal effect significant? To find out we conduct a hypothesis test.

```
## 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
```

With a p-value way below .05 we can conclude that the marginal effect of is indeed statistically significant.

2. Evaluate the marginal effect of education on Joe Biden thermometer rating, conditional on age. Consider the magnitude and direction of the marginal effect, as well as its statistical significance.

Firstly, we note again that the model estimated above has the following form:

$$E(biden) = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 age * educ$$

In order to evaluate the marginal effect of `educ` on `biden` conditional on `age` we take:

$$\frac{\delta E(biden)}{\delta educ} = \beta_2 + \beta_2 age$$

From the model summary, we have the values for β_2 and β_3, inserting them into the equation, we have:

$$\frac{\delta E(biden)}{\delta educ} = 1.65743 + -0.04803 educ$$

We see that the marginal effect of `educ` on `biden` conditional on `age` has variable magnitude. For values of $age < 35$ the effect on `biden` is positive, but for $age \geq 35$ the effect is negative.

Now, is this marginal effect significant? To find out we conduct a hypothesis test.

```
## Linear hypothesis test
##
## Hypothesis:
## educ  + 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 979537
## 2   1803 976688  1    2849.1 5.2595 0.02194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Hypothesis test tells us that the marginal effect is indeed statistically significant, at least the the .05 level.

Missing Data

Before conducting a multiple imputation process we will conduct a Henze-Zirkler' Multivariate Normality Test to see if our predictors are distributed as a multivariate normal distribution. Since `female` is a binary variable, we will see if `age` and `educ` are together distributed multivariate normally and whether they are individually distributed normally.

```
preds <- biden_dat %>%
  select(biden, age, educ, female, dem, rep)

hzTest(preds %>%
         select(-c(biden, female, dem, rep)))
```

```
##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : preds %>% select(-c(biden, female, dem, rep))
##
##   HZ      : 22.08529
##   p-value : 0
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------
```

```
uniNorm(preds %>%
          na.omit() %>%
          select(-c(biden, female, dem, rep)), type = "SW", desc = FALSE)
```

```
## $`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
```

We see that our predictors are indeed not distributed multivariate normally and that they are not distributed normally on their own either under the Shapiro-Wilk test. Let's try and use either a square root, log, or $-\frac{1}{2}$ power transformation to coerce some of the variables to be more normal individually and see if this can coerce all of our predictors to be MVN distributed.

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

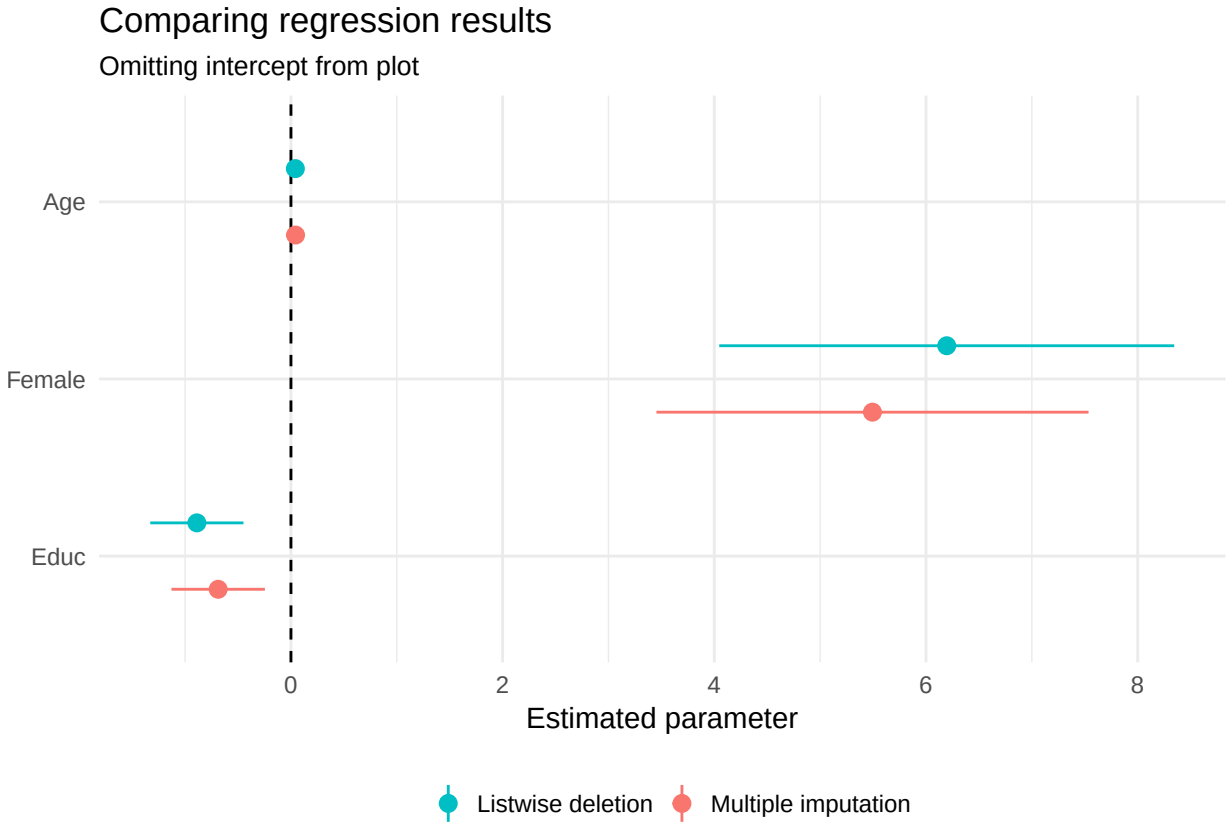
```
##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : biden_omit %>% select(sqrt_educ, sqrt_age)
##
##   HZ      : 15.33627
##   p-value : 0
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------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

```
## $`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
```

After experimenting a bit, square-root transforming provides the most improved, though not ideal results. It reduces the HZ statistics for MVN testing from `~22` to `~15`, although individually, `sqrt_age` and `sqrt_educ` aren't normal according to the Shapiro-Wilk test.

Now let's compute the new linear model with the imputed values and compare its coefficient estimates and standard errors against the original model with the missing values removed.



As we can see, there does not appear to be any statistically significant difference between the coefficients of the linear model where the multiple imputation procedure was conducted and the original model where rows with `NA` values were simply removed.