PS3

Haylee Ham

5/11/2017

```
biden_df = read.csv('biden.csv')
biden_df = na.omit(biden_df)
```

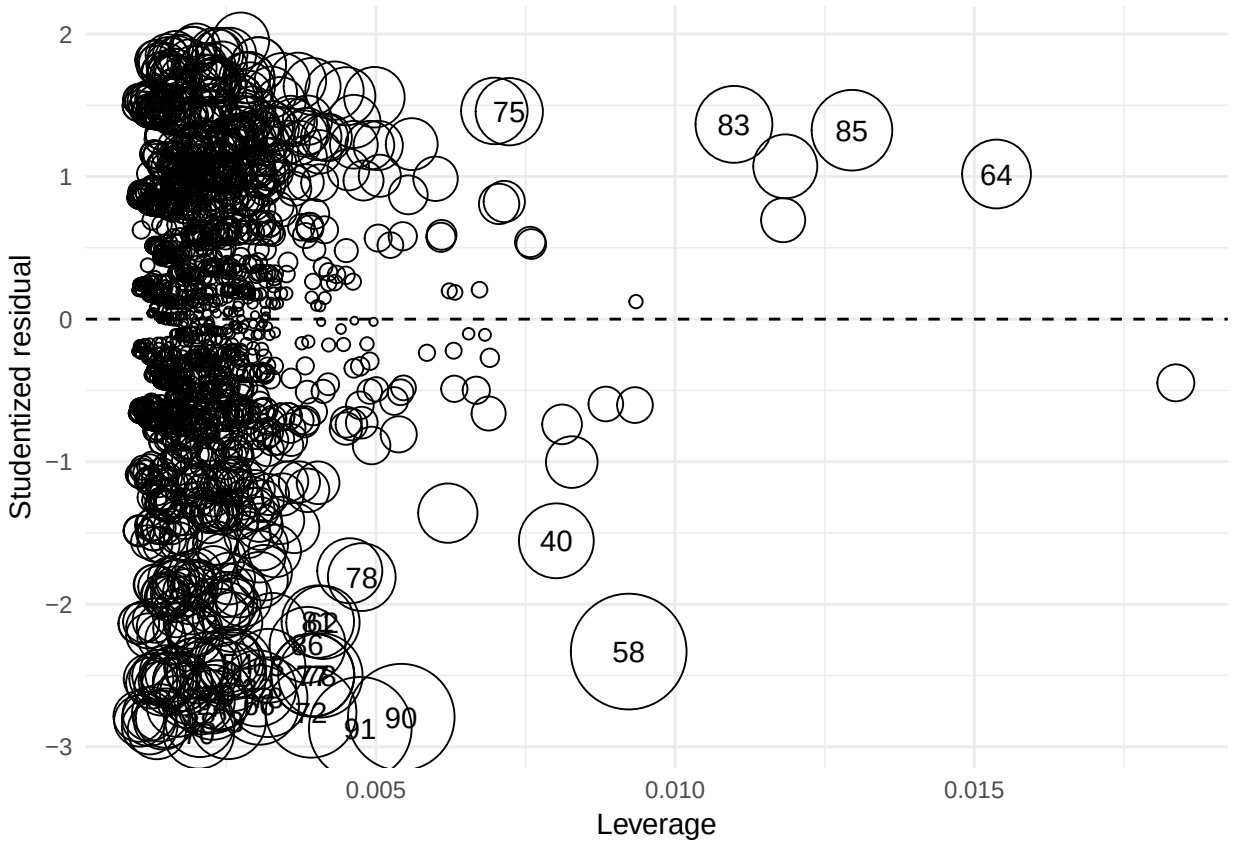
Regression Diagnostics

```
biden_mod <- lm(biden ~ age + female + educ, data = biden_df)
tidy(biden_mod)
```

```
##          term estimate std.error statistic  p.value
## 1 (Intercept)  68.6210    3.5960     19.08 4.34e-74
## 2         age   0.0419    0.0325      1.29 1.98e-01
## 3      female   6.1961    1.0967      5.65 1.86e-08
## 4        educ  -0.8887    0.2247     -3.96 7.94e-05
```

1. From the plot below, we see a mixture of observations that have various levels of leverage and discrepancy. There is a very large grouping of observations that have very low leverage and grouped on the left hand side of the graph. With the observations being labeled by the age of the respondent, one can see that 91, 90, and 58 appear to be noticeably unusual observations. The cluster of observation above the line (83, 85, and 64) do have high amounts of leverage, but with a small Cook's D value they are not very discrepent.

```
# add key statistics
biden_augment <- biden_df %>%
  mutate(hat = hatvalues(biden_mod),
         student = rstudent(biden_mod),
         cooksd = cooks.distance(biden_mod))

# draw bubble plot
ggplot(biden_augment, aes(hat, student)) +
  geom_hline(yintercept = 0, linetype = 2) +
  geom_point(aes(size = cooksd), shape = 1) +
  geom_text(data = biden_augment %>%
              arrange(-cooksd) %>%
              slice(1:30),
            aes(label = age)) +
  scale_size_continuous(range = c(1, 20)) +
  labs(x = "Leverage",
       y = "Studentized residual") +
  theme(legend.position = "none")
```



These tables show observations that are filtered by their respective amount of leverage, discrepency and, finally, influence. When finding observations that have higher than the rule of thumb amount of influence, 90 observations are flagged as having a high amount of observations. Some of these strange observations appear to be females who have a Biden rating of 0. In light of this, I may respecify the model in an attempt to find the missing predictor variable. In this case, party affiliation may prove to be very important.

```
biden_augment %>%
  filter(hat > 2 * mean(hat))
```

```
##    biden female age educ dem rep     hat student   cooksd
## 1     70      0  80   17   0   0 0.00504  0.5686 4.09e-04
## 2     70      1  44    7   1   0 0.00496 -0.0190 4.49e-07
## 3    100      1  64    1   1   0 0.01537  1.0179 4.04e-03
## 4    100      1  76    3   1   0 0.01184  1.0715 3.44e-03
## 5     60      1  84   16   0   0 0.00446 -0.1781 3.55e-05
## 6     60      1  63    4   0   0 0.00933 -0.6029 8.56e-04
## 7     85      0  18    8   1   0 0.00600  0.9846 1.46e-03
## 8     70      0  79    9   1   0 0.00461  0.2626 8.00e-05
## 9     50      1  22    9   0   0 0.00450 -0.7676 6.66e-04
## 10    50      1  23    8   0   0 0.00538 -0.8083 8.83e-04
## 11    85      0  73    3   0   0 0.01180  0.6943 1.44e-03
## 12    60      1  20    9   0   0 0.00474 -0.3313 1.31e-04
## 13    60      1  82    8   0   0 0.00546 -0.4823 3.19e-04
## 14    60      1  43    6   0   0 0.00631 -0.4887 3.79e-04
## 15    20      1  58    4   0   1 0.00922 -2.3319 1.26e-02
## 16    50      0  18    9   0   1 0.00500 -0.4923 3.05e-04
## 17    85      1  71    7   0   1 0.00545  0.5814 4.63e-04
## 18    70      1  73    6   1   0 0.00682 -0.1101 2.08e-05
```

```
## 19   100      0  82    9   1   0 0.00497  1.5565 3.02e-03
## 20    85      1  54    6   1   0 0.00609  0.5739 5.04e-04
## 21    70      0  33    8   0   0 0.00450  0.3075 1.07e-04
## 22    60      0  81    9   1   0 0.00485 -0.1737 3.68e-05
## 23    30      0  40    5   0   0 0.00801 -1.5549 4.88e-03
## 24    85      0  88   11   1   0 0.00479  0.9725 1.14e-03
## 25    50      1  85   17   0   1 0.00530 -0.5743 4.40e-04
## 26     0      1  90   16   0   1 0.00542 -2.7921 1.06e-02
## 27    60      1  91   12   1   0 0.00463 -0.3446 1.38e-04
## 28    40      0  78   17   0   1 0.00476 -0.7261 6.31e-04
## 29    50      1  91   13   0   1 0.00459 -0.7389 6.29e-04
## 30    85      1  59    5   1   0 0.00759  0.5268 5.31e-04
## 31    70      0  86   16   1   0 0.00524  0.5193 3.55e-04
## 32    50      0  68    5   0   0 0.00811 -0.7380 1.11e-03
## 33    15      0  78   17   0   0 0.00476 -1.8094 3.91e-03
## 34    50      0  93   12   0   1 0.00541 -0.5130 3.58e-04
## 35    85      0  88   13   1   0 0.00454  1.0494 1.26e-03
## 36    85      0  79   17   0   0 0.00490  1.2199 1.83e-03
## 37    70      0  51    6   1   0 0.00622  0.1981 6.15e-05
## 38    85      1  51    5   0   1 0.00758  0.5413 5.59e-04
## 39    40      1  46    6   1   0 0.00620 -1.3608 2.89e-03
## 40    40      0  76   17   0   1 0.00450 -0.7224 5.90e-04
## 41    30      1  73    8   1   0 0.00456 -1.7652 3.57e-03
## 42     0      1  91   14   0   1 0.00474 -2.8704 9.76e-03
## 43    60      0  73    7   1   0 0.00585 -0.2363 8.22e-05
## 44    50      0  80    9   0   1 0.00473 -0.6047 4.35e-04
## 45    70      1  82    9   1   0 0.00464 -0.0109 1.39e-07
## 46    85      0  73    6   1   0 0.00706  0.8082 1.16e-03
## 47    50      1  45    7   0   0 0.00493 -0.8865 9.73e-04
## 48   100      1  73    7   1   0 0.00560  1.2276 2.12e-03
## 49    70      1  69    6   0   0 0.00655 -0.1028 1.74e-05
## 50    70      0  89    8   0   1 0.00673  0.2062 7.21e-05
## 51   100      0  72    6   1   0 0.00698  1.4605 3.74e-03
## 52    85      0  85    7   1   0 0.00714  0.8250 1.22e-03
## 53    60      1  80    5   1   0 0.00884 -0.5951 7.90e-04
## 54    60      0  39    0   1   0 0.01837 -0.4468 9.34e-04
## 55   100      0  75    6   1   0 0.00723  1.4552 3.85e-03
## 56    70      0  50    4   1   0 0.00934  0.1232 3.58e-05
## 57    70      0  78    7   1   0 0.00632  0.1877 5.60e-05
## 58    50      1  87    6   1   0 0.00827 -1.0028 2.10e-03
## 59    85      0  80   17   0   0 0.00504  1.2181 1.88e-03
## 60   100      0  83    4   1   0 0.01098  1.3662 5.18e-03
## 61    75      0  78    9   0   1 0.00450  0.4807 2.61e-04
## 62    50      0  79   17   1   0 0.00490 -0.2952 1.07e-04
## 63    85      0  77   17   0   0 0.00463  1.2233 1.74e-03
## 64    60      0  71    6   0   1 0.00690 -0.2713 1.28e-04
## 65    60      0  86    8   1   0 0.00630 -0.2214 7.77e-05
## 66    85      0  91   12   1   0 0.00506  1.0057 1.29e-03
## 67    85      0  80    8   1   0 0.00554  0.8719 1.06e-03
## 68   100      1  91   12   1   0 0.00463  1.3869 2.24e-03
## 69    50      0  90   12   0   1 0.00489 -0.5074 3.17e-04
## 70   100      0  85    3   1   0 0.01295  1.3252 5.76e-03
## 71    50      0  90    8   0   1 0.00688 -0.6620 7.59e-04
## 72   100      0  78    9   1   0 0.00450  1.5634 2.76e-03
```

```
## 73    85      1  87    8   1   0 0.00610  0.5910 5.36e-04
## 74    60      1  91    8   1   0 0.00668 -0.4989 4.18e-04
```

```
biden_augment %>%
  filter(abs(student) > 2)
```

```
##    biden female age educ dem rep     hat student  cooksd
## 1      0      1  70   12   0   1 0.00204   -2.91 0.00429
## 2      0      0  45   12   0   1 0.00142   -2.59 0.00237
## 3      0      0  40   14   0   0 0.00136   -2.50 0.00213
## 4     15      0  62    8   0   1 0.00411   -2.13 0.00466
## 5     15      1  20   13   0   0 0.00260   -2.12 0.00294
## 6      0      1  38   14   1   0 0.00122   -2.77 0.00233
## 7      0      0  34   12   0   0 0.00178   -2.57 0.00293
## 8      0      0  21   13   0   1 0.00259   -2.51 0.00407
## 9     15      1  29   12   0   1 0.00198   -2.18 0.00235
## 10     0      0  36   13   0   1 0.00149   -2.53 0.00239
## 11    15      1  86   12   0   0 0.00386   -2.28 0.00504
## 12    20      1  58    4   0   1 0.00922   -2.33 0.01262
## 13     0      0  56   11   0   0 0.00185   -2.65 0.00323
## 14     0      0  60   16   0   0 0.00236   -2.46 0.00358
## 15     0      1  28   12   1   0 0.00206   -2.83 0.00412
## 16     0      0  41   17   0   1 0.00252   -2.39 0.00360
## 17     0      1  90   16   0   1 0.00542   -2.79 0.01058
## 18     0      0  77   16   0   1 0.00394   -2.50 0.00615
## 19     0      1  51   16   0   1 0.00168   -2.72 0.00309
## 20     0      0  50   17   0   1 0.00257   -2.41 0.00372
## 21    15      1  81   16   0   1 0.00403   -2.12 0.00454
## 22     0      0  53   15   0   1 0.00161   -2.49 0.00249
## 23     8      1  52   12   1   0 0.00120   -2.52 0.00190
## 24     0      1  48   14   0   1 0.00104   -2.79 0.00200
## 25     0      0  64   12   0   1 0.00191   -2.62 0.00329
## 26     0      0  51   16   0   0 0.00198   -2.45 0.00296
## 27     0      1  31   16   0   1 0.00208   -2.68 0.00373
## 28    15      1  39   13   0   0 0.00119   -2.16 0.00138
## 29     0      0  46   13   0   1 0.00125   -2.55 0.00203
## 30    15      1  52   12   0   1 0.00120   -2.22 0.00147
## 31     5      0  51   16   0   1 0.00198   -2.23 0.00246
## 32    15      1  48   14   1   0 0.00104   -2.14 0.00118
## 33    15      1  36   14   0   0 0.00130   -2.11 0.00145
## 34     0      0  58   14   0   1 0.00155   -2.54 0.00249
## 35     0      1  23   12   0   0 0.00253   -2.82 0.00502
## 36     0      1  57   14   0   1 0.00121   -2.80 0.00237
## 37     0      0  70   12   0   1 0.00236   -2.64 0.00411
## 38    15      1  79   15   0   1 0.00327   -2.16 0.00381
## 39     0      0  35   13   0   0 0.00154   -2.53 0.00246
## 40     0      0  50   16   0   1 0.00196   -2.45 0.00293
## 41     0      0  78   16   0   0 0.00407   -2.50 0.00636
## 42     0      0  57   16   0   0 0.00220   -2.46 0.00332
## 43    15      1  42   17   0   0 0.00223   -2.01 0.00226
## 44     0      0  22   15   0   1 0.00260   -2.43 0.00384
## 45     0      0  78   12   0   1 0.00319   -2.65 0.00560
## 46     0      0  72    9   0   0 0.00392   -2.76 0.00745
## 47     0      0  62   14   0   1 0.00176   -2.54 0.00285
## 48    15      1  66   14   0   1 0.00170   -2.17 0.00200
```

```
## 49     0      1  91   14   0   1 0.00474   -2.87 0.00976
## 50    15      1  61   14   0   1 0.00139   -2.16 0.00162
## 51     0      0  50   14   0   0 0.00131   -2.52 0.00207
## 52     0      0  46   15   0   1 0.00150   -2.48 0.00230
## 53     0      0  54   17   0   1 0.00269   -2.41 0.00392
## 54     0      1  44   13   0   1 0.00105   -2.82 0.00209
## 55     0      1  58   12   0   0 0.00134   -2.88 0.00277
## 56     0      0  65   11   0   1 0.00227   -2.67 0.00402
## 57     0      0  63   17   0   0 0.00320   -2.43 0.00474
## 58    15      1  66   16   0   1 0.00241   -2.09 0.00264
## 59     0      1  34   14   0   1 0.00140   -2.76 0.00266
## 60     0      0  77   16   0   1 0.00394   -2.50 0.00615
## 61     0      0  62   14   0   1 0.00176   -2.54 0.00285
## 62    15      1  46   11   1   0 0.00156   -2.25 0.00197
## 63    15      1  48   14   0   1 0.00104   -2.14 0.00118
## 64    15      1  60   12   0   1 0.00141   -2.23 0.00176
## 65     0      0  39   12   0   0 0.00155   -2.58 0.00258
## 66     0      1  66   17   0   0 0.00305   -2.71 0.00558
## 67    15      1  41   14   0   1 0.00112   -2.12 0.00126
## 68    15      1  69   14   0   0 0.00193   -2.17 0.00229
## 69     0      0  32   16   1   0 0.00223   -2.41 0.00324
## 70     0      1  33   13   0   0 0.00148   -2.80 0.00289
## 71     0      1  24   15   0   0 0.00227   -2.71 0.00415
## 72     0      1  45   12   0   1 0.00122   -2.86 0.00248
## 73     0      0  27   14   0   0 0.00202   -2.48 0.00310
## 74     0      0  77   16   0   1 0.00394   -2.50 0.00615
## 75    15      1  57   17   0   0 0.00248   -2.04 0.00257
## 76    15      1  24   16   0   0 0.00260   -2.02 0.00264
## 77    15      1  65   15   0   0 0.00189   -2.13 0.00214
## 78    15      1  50   16   0   0 0.00166   -2.06 0.00177
## 79     0      1  62   14   0   0 0.00144   -2.81 0.00284
## 80     0      0  23   11   0   1 0.00303   -2.59 0.00508
## 81     0      0  70   12   1   0 0.00236   -2.64 0.00411
## 82    15      1  34   16   0   0 0.00192   -2.03 0.00199
```

```
biden_augment %>%
  filter(cooksd > 4 / (nrow(.) - (length(coef(biden_mod)) - 1) - 1))
```

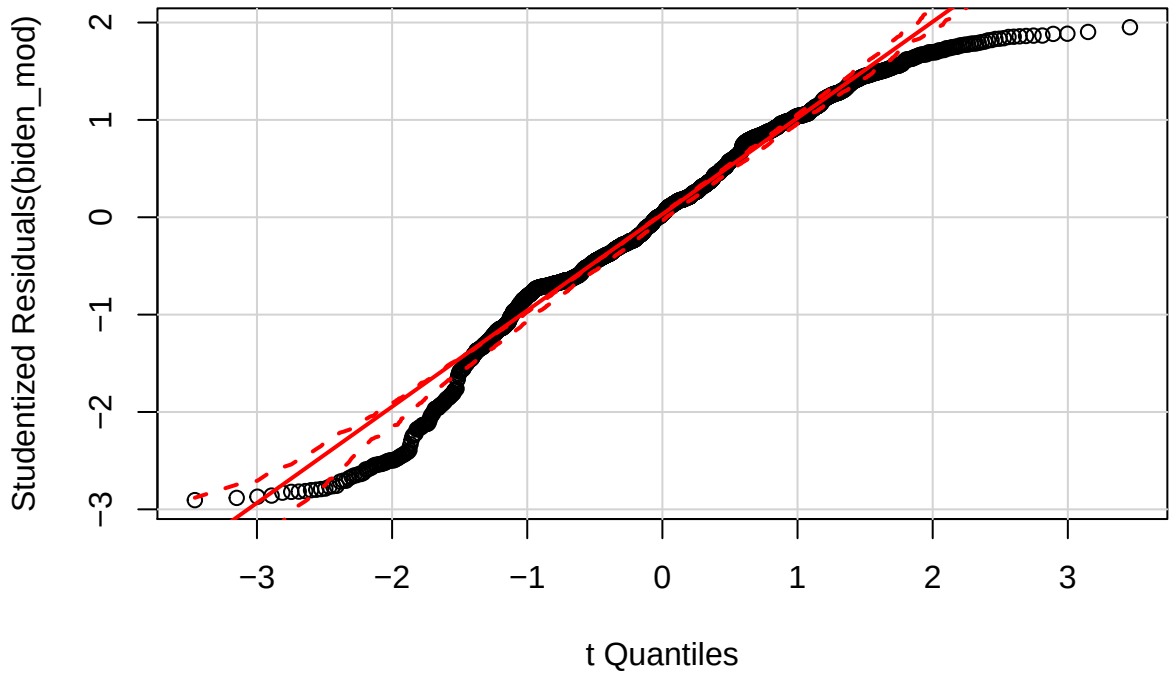
```
##    biden female age educ dem rep     hat student  cooksd
## 1      0      1  70   12   0   1 0.00204   -2.91 0.00429
## 2      0      0  45   12   0   1 0.00142   -2.59 0.00237
## 3     15      0  62    8   0   1 0.00411   -2.13 0.00466
## 4     15      1  20   13   0   0 0.00260   -2.12 0.00294
## 5    100      1  64    1   1   0 0.01537    1.02 0.00404
## 6    100      0  19   12   0   0 0.00304    1.78 0.00242
## 7    100      0  19   12   1   0 0.00304    1.78 0.00242
## 8      0      1  38   14   1   0 0.00122   -2.77 0.00233
## 9    100      1  76    3   1   0 0.01184    1.07 0.00344
## 10     0      0  34   12   0   0 0.00178   -2.57 0.00293
## 11     0      0  21   13   0   1 0.00259   -2.51 0.00407
## 12    15      1  29   12   0   1 0.00198   -2.18 0.00235
## 13     0      0  36   13   0   1 0.00149   -2.53 0.00239
## 14    15      1  86   12   0   0 0.00386   -2.28 0.00504
## 15    20      1  58    4   0   1 0.00922   -2.33 0.01262
## 16     0      0  56   11   0   0 0.00185   -2.65 0.00323
```

```
## 17   100       0  82    9   1    0 0.00497    1.56 0.00302
## 18     0       0  60   16   0    0 0.00236   -2.46 0.00358
## 19    30       0  40    5   0    0 0.00801   -1.55 0.00488
## 20     0       1  28   12   1    0 0.00206   -2.83 0.00412
## 21    15       0  22   12   0    1 0.00271   -1.90 0.00245
## 22     0       0  41   17   0    1 0.00252   -2.39 0.00360
## 23     0       1  90   16   0    1 0.00542   -2.79 0.01058
## 24     0       0  77   16   0    1 0.00394   -2.50 0.00615
## 25     0       1  51   16   0    1 0.00168   -2.72 0.00309
## 26     0       0  50   17   0    1 0.00257   -2.41 0.00372
## 27   100       1  78   17   1    0 0.00431    1.60 0.00278
## 28    15       1  81   16   0    1 0.00403   -2.12 0.00454
## 29     0       0  53   15   0    1 0.00161   -2.49 0.00249
## 30     0       0  64   12   0    1 0.00191   -2.62 0.00329
## 31     0       0  51   16   0    0 0.00198   -2.45 0.00296
## 32     0       1  31   16   0    1 0.00208   -2.68 0.00373
## 33     5       0  51   16   0    1 0.00198   -2.23 0.00246
## 34     0       0  58   14   0    1 0.00155   -2.54 0.00249
## 35    15       0  78   17   0    0 0.00476   -1.81 0.00391
## 36   100       0  82   12   1    0 0.00369    1.67 0.00259
## 37     0       1  23   12   0    0 0.00253   -2.82 0.00502
## 38    15       0  69   16   0    1 0.00306   -1.83 0.00256
## 39     0       1  57   14   0    1 0.00121   -2.80 0.00237
## 40    15       0  75   13   0    1 0.00278   -1.96 0.00266
## 41     0       0  70   12   0    1 0.00236   -2.64 0.00411
## 42    15       1  79   15   0    1 0.00327   -2.16 0.00381
## 43     0       0  35   13   0    0 0.00154   -2.53 0.00246
## 44     0       0  50   16   0    1 0.00196   -2.45 0.00293
## 45    40       1  46    6   1    0 0.00620   -1.36 0.00289
## 46     0       0  78   16   0    0 0.00407   -2.50 0.00636
## 47     0       0  57   16   0    0 0.00220   -2.46 0.00332
## 48    15       1  42   17   0    0 0.00223   -2.01 0.00226
## 49    30       1  73    8   1    0 0.00456   -1.77 0.00357
## 50     0       0  22   15   0    1 0.00260   -2.43 0.00384
## 51     0       0  78   12   0    1 0.00319   -2.65 0.00560
## 52     0       0  72    9   0    0 0.00392   -2.76 0.00745
## 53     0       0  62   14   0    1 0.00176   -2.54 0.00285
## 54     0       1  91   14   0    1 0.00474   -2.87 0.00976
## 55     0       0  46   15   0    1 0.00150   -2.48 0.00230
## 56     0       0  54   17   0    1 0.00269   -2.41 0.00392
## 57   100       0  72    6   1    0 0.00698    1.46 0.00374
## 58     0       1  58   12   0    0 0.00134   -2.88 0.00277
## 59     0       0  65   11   0    1 0.00227   -2.67 0.00402
## 60   100       0  75    6   1    0 0.00723    1.46 0.00385
## 61     0       0  63   17   0    0 0.00320   -2.43 0.00474
## 62    15       1  66   16   0    1 0.00241   -2.09 0.00264
## 63     0       1  34   14   0    1 0.00140   -2.76 0.00266
## 64    15       0  62   17   0    0 0.00313   -1.78 0.00248
## 65    10       0  46   17   0    1 0.00250   -1.97 0.00242
## 66   100       0  33   17   1    0 0.00274    1.95 0.00261
## 67     0       0  77   16   0    1 0.00394   -2.50 0.00615
## 68     0       0  62   14   0    1 0.00176   -2.54 0.00285
## 69    15       0  24   12   0    0 0.00252   -1.90 0.00228
## 70     0       0  39   12   0    0 0.00155   -2.58 0.00258
```

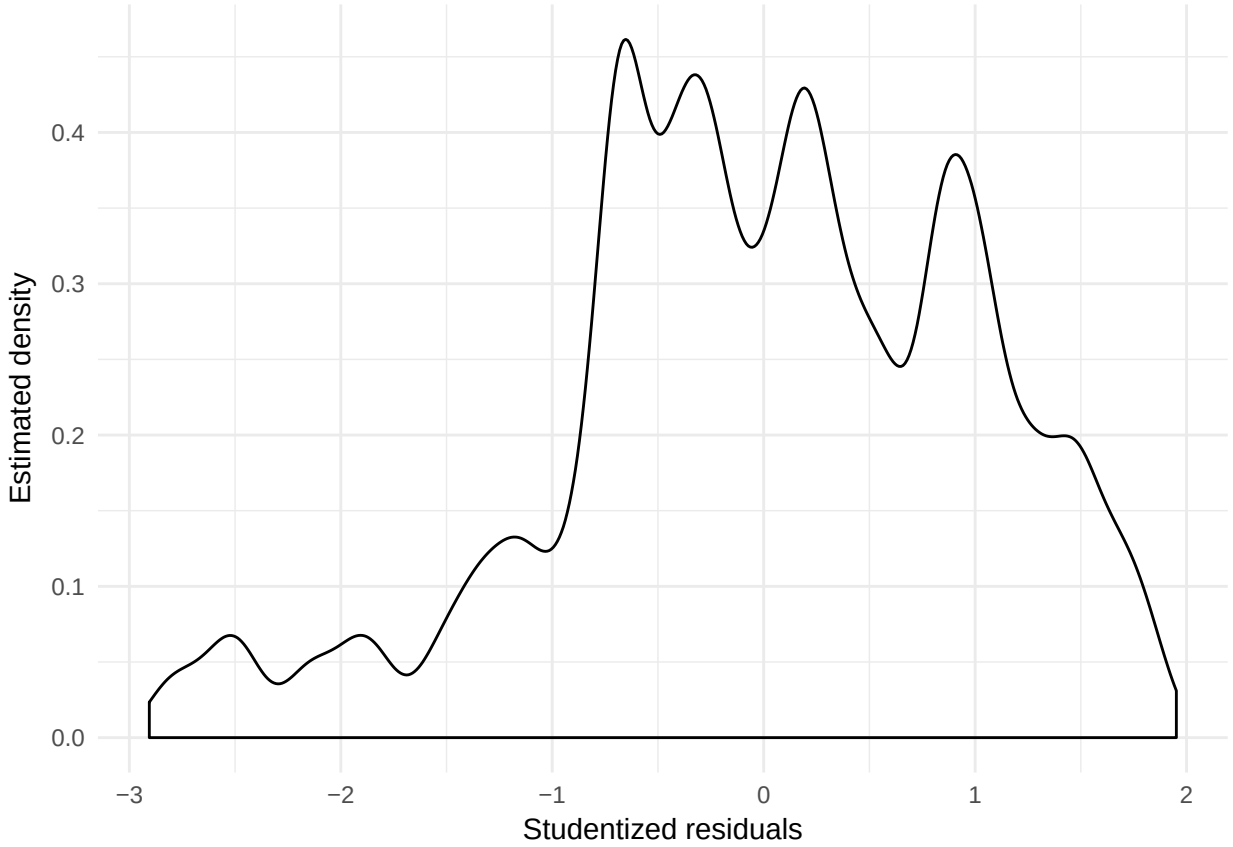
```
## 71     0       1  66   17    0    0 0.00305   -2.71 0.00558
## 72    15       1  69   14    0    0 0.00193   -2.17 0.00229
## 73     0       0  32   16    1    0 0.00223   -2.41 0.00324
## 74   100       0  83    4    1    0 0.01098    1.37 0.00518
## 75    15       0  72   12    0    1 0.00255   -1.99 0.00252
## 76   100       0  82   11    1    0 0.00393    1.63 0.00263
## 77   100       0  80   12    1    0 0.00343    1.67 0.00241
## 78     0       1  33   13    0    0 0.00148   -2.80 0.00289
## 79     0       1  24   15    0    0 0.00227   -2.71 0.00415
## 80     0       1  45   12    0    1 0.00122   -2.86 0.00248
## 81     0       0  27   14    0    0 0.00202   -2.48 0.00310
## 82     0       0  77   16    0    1 0.00394   -2.50 0.00615
## 83    15       1  57   17    0    0 0.00248   -2.04 0.00257
## 84   100       1  91   12    1    0 0.00463    1.39 0.00224
## 85   100       0  85    3    1    0 0.01295    1.33 0.00576
## 86    15       1  24   16    0    0 0.00260   -2.02 0.00264
## 87     0       1  62   14    0    0 0.00144   -2.81 0.00284
## 88   100       0  78    9    1    0 0.00450    1.56 0.00276
## 89     0       0  23   11    0    1 0.00303   -2.59 0.00508
## 90     0       0  70   12    1    0 0.00236   -2.64 0.00411
```

2. In testing for non-normally distributed errors, the quantile-comparison plot clearly shows very non-normal residuals that fall outside of the 95% confidence intervals (the dashed lines). The density also shows very skewed residuals with a long right tail and many modes.

```
car::qqPlot(biden_mod)
```




```
biden_copy <- biden_mod
augment(biden_copy, biden_df) %>%
  mutate(.student = rstudent(biden_mod)) %>%
  ggplot(aes(.student)) +
  geom_density(adjust = .5) +
  labs(x = "Studentized residuals",
       y = "Estimated density")
```

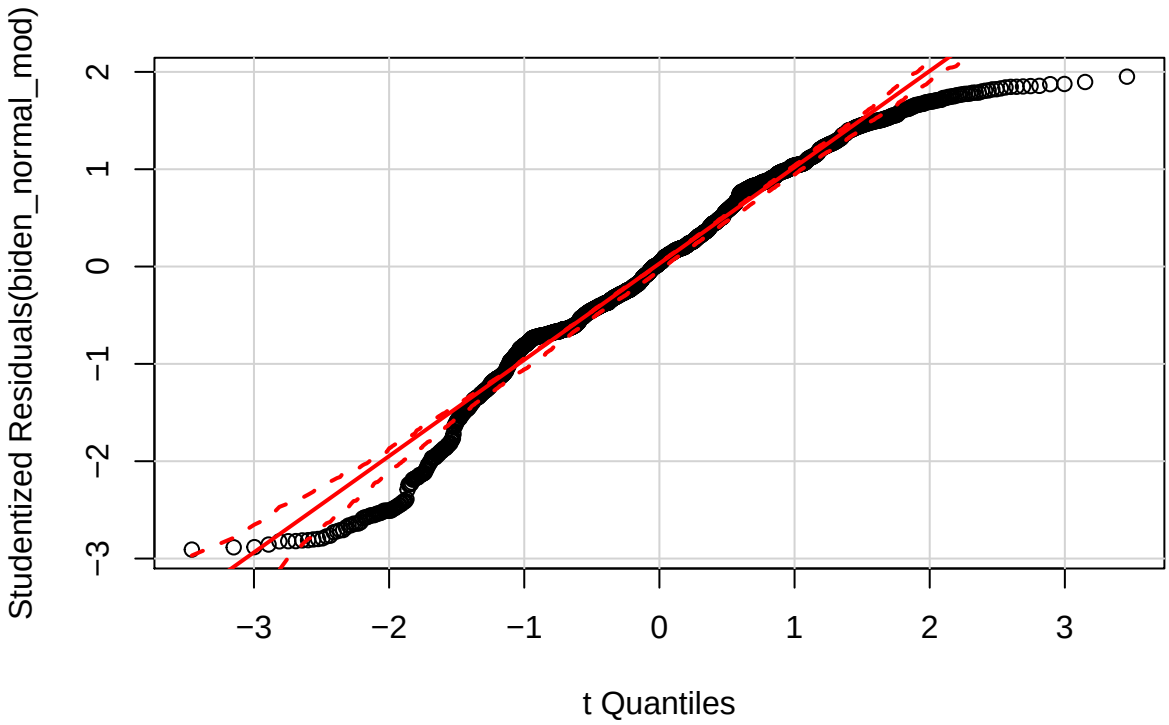


In an attempt to fix this non-normality, power transformations on the response variable are not appropriate, they only exacerbate the problem. Since some of values of the response variable `biden` are 0, a log transformation cannot be done either. Performing a power transformation on the `educ` variable to the power of 2, since the variable is negatively skewed, seems to be an appropriate start to correcting for non-normally distributed errors.

```
biden_normal <- biden_df %>%
  mutate(educ_power = educ^2)

biden_normal_mod <- lm(biden ~ age + female + educ_power, data = biden_normal)
tidy(biden_normal_mod)
```

```
##          term estimate std.error statistic   p.value
## 1 (Intercept)  62.1906   2.44562     25.43 3.67e-122
## 2         age   0.0464   0.03244      1.43  1.53e-01
## 3      female   6.1695   1.09767      5.62  2.20e-08
## 4  educ_power  -0.0306   0.00877     -3.49  4.99e-04
```

```
car::qqPlot(biden_normal_mod)
```



3. In testing for heteroscedasticity, the Breusch-Pagan test rejects the null hypothesis with a p-value of 0.00005, thus indicating that heteroscedasticity is present. Graphically, this non-constant variance can be seen especially on the left side of the plot. Heteroscedasticity could impact our inference by either inflating or deflating our standard errors.

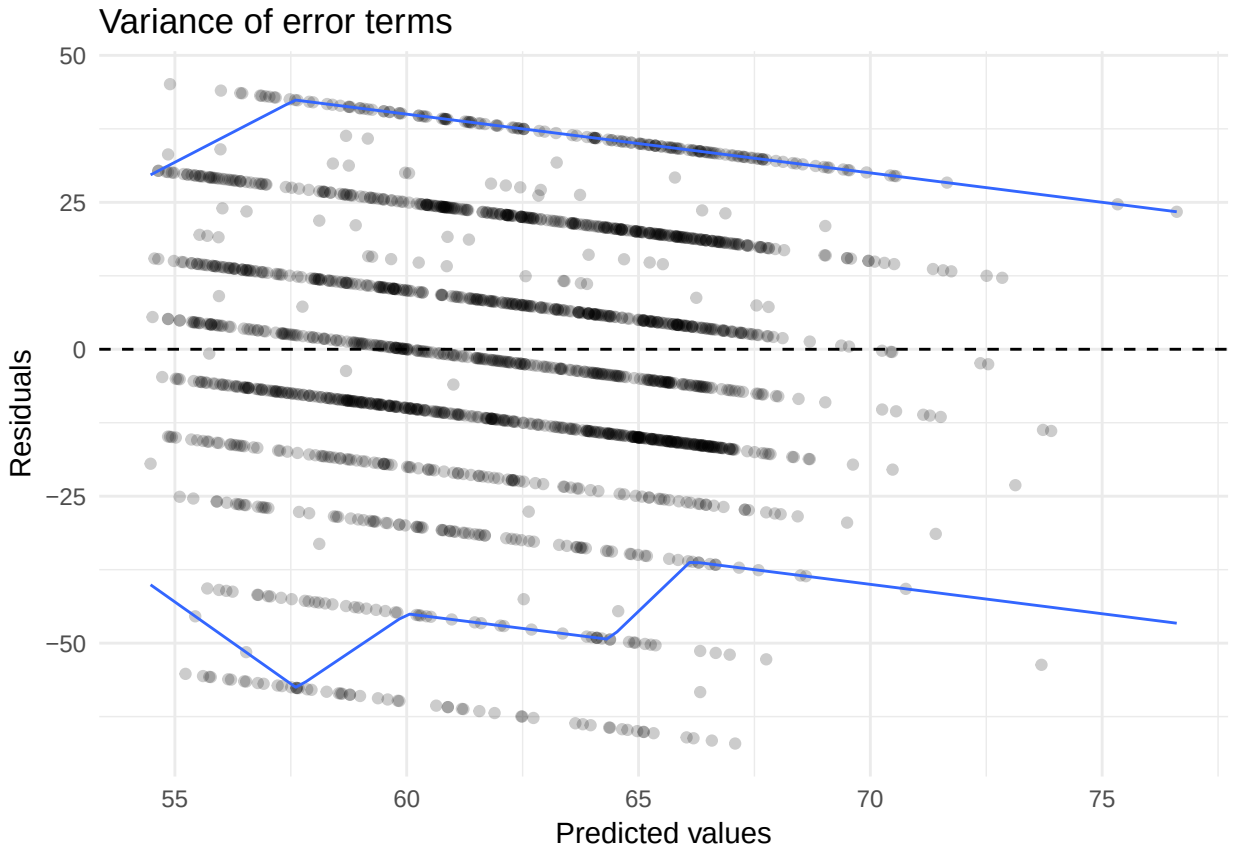
```
biden_df %>%
  add_predictions(biden_copy) %>%
  add_residuals(biden_copy) %>%
  ggplot(aes(pred, resid)) +
  geom_point(alpha = .2) +
  geom_hline(yintercept = 0, linetype = 2) +
  geom_quantile(method = "rqss", lambda = 5, quantiles = c(.05, .95)) +
  labs(title = "Variance of error terms",
       x = "Predicted values",
       y = "Residuals")
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
## Smoothing formula not specified. Using: y ~ qss(x, lambda = 5)
```



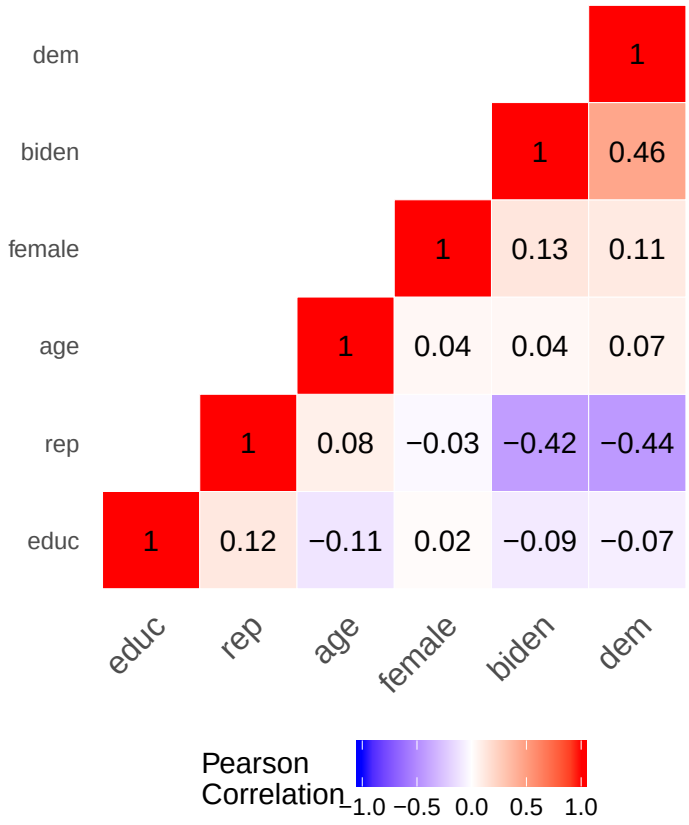

```
bptest(biden_mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  biden_mod
## BP = 20, df = 3, p-value = 5e-05
```

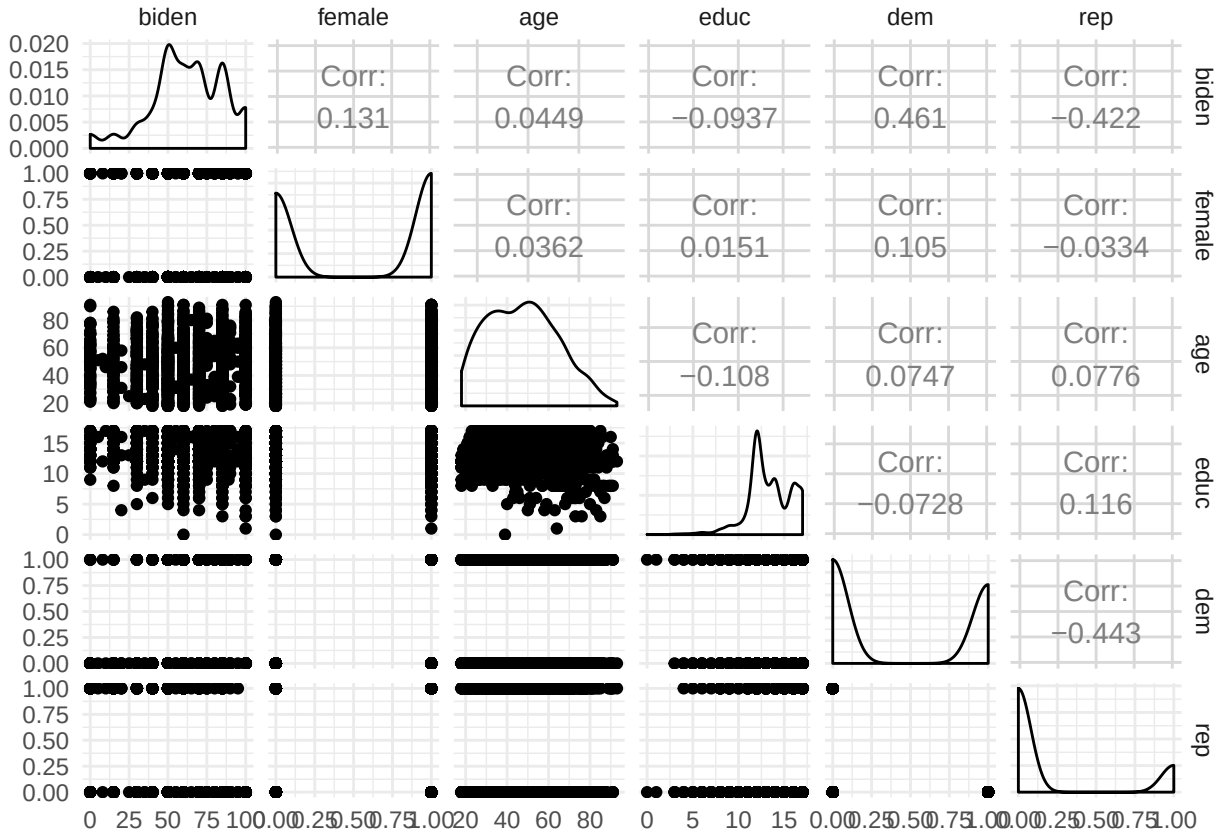
4. The correlation matrices below do not indicate that there are any two variables with extremely high collinearity. Within our model, I have tested calculated variance inflation factors (VIF) for each pair of variables and have found no evidence of multicollinearity (which would be indicated if there were any values over 10). If multicollinearity had been indicated, methods such as adding more data (impractical in this case unless we can collect more data), transforming the covariates, or shrinking the estimated coefficients could be used.

```
cormat_heatmap <- function(data){
  # generate correlation matrix
  cormat <- round(cor(data), 2)

  # melt into a tidy table
  get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }

  upper_tri <- get_upper_tri(cormat)

  # reorder matrix based on coefficient value
```

```
  reorder_cormat <- function(cormat){
    # Use correlation between variables as distance
    dd <- as.dist((1-cormat)/2)
    hc <- hclust(dd)
    cormat <-cormat[hc$order, hc$order]
  }

  cormat <- reorder_cormat(cormat)
  upper_tri <- get_upper_tri(cormat)

  # Melt the correlation matrix
  melted_cormat <- reshape2::melt(upper_tri, na.rm = TRUE)

  # Create a ggheatmap
  ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
    geom_tile(color = "white")+
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
                         name="Pearson\nCorrelation") +
    theme_minimal()+ # minimal theme
    theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                     size = 12, hjust = 1))+
    coord_fixed()

  # add correlation values to graph
  ggheatmap +
    geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
    theme(
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      panel.grid.major = element_blank(),
      panel.border = element_blank(),
      panel.background = element_blank(),
      axis.ticks = element_blank(),
      legend.position = "bottom")
}

cormat_heatmap(select_if(biden_df, is.numeric))
```




```
ggpairs(select_if(biden_df, is.numeric))
```



```
age_female <- lm(biden ~ age + female, data = biden_df)
car::vif(age_female)
```

```
##    age female
##      1      1
```

```
female_educ <- lm(biden ~ educ + female, data = biden_df)
car::vif(female_educ)
```

```
##   educ female
##      1      1
```

```
age_educ <- lm(biden ~ age + educ, data = biden_df)
car::vif(age_educ)
```

```
##  age educ
## 1.01 1.01
```

Interaction Terms

1. The results of this interaction term model, as shown in the plot, show that the marginal effect of age changes as the education level of the respondent changes. The marginal effect of age is postive until the respondents reach an education level of around 13. At that point the effect either becomes negligible or negative.

```
inter_biden <- lm(biden ~ age * educ, data=biden_df)
tidy(inter_biden)
```

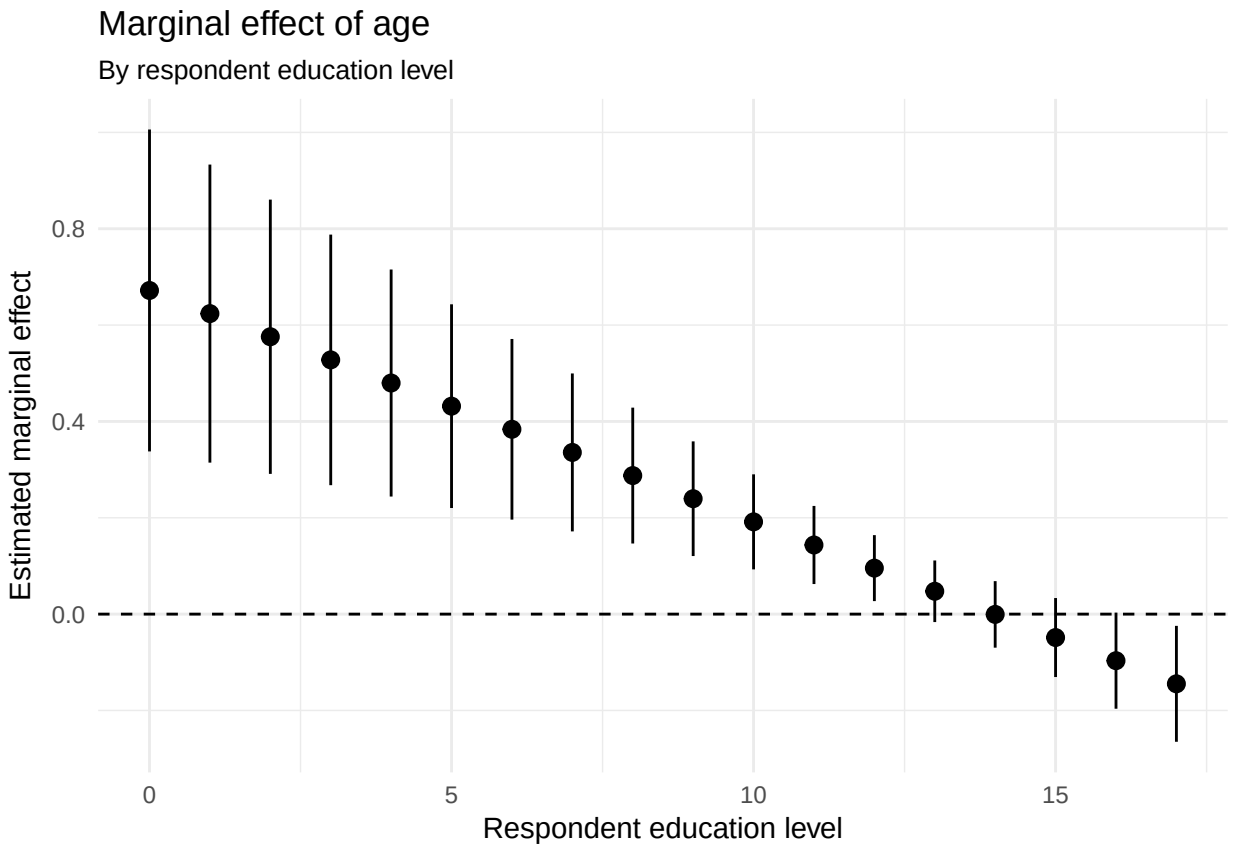
```
##          term estimate std.error statistic  p.value
## 1 (Intercept)   38.374    9.5636      4.01 6.25e-05
## 2         age    0.672    0.1705      3.94 8.43e-05
## 3        educ    1.657    0.7140      2.32 2.04e-02
## 4    age:educ   -0.048    0.0129     -3.72 2.03e-04
```

```
glance(inter_biden)
```

```
##   r.squared adj.r.squared sigma statistic  p.value df logLik   AIC   BIC
## 1    0.0176        0.0159  23.3      10.7 5.37e-07  4  -8249 16509 16536
##   deviance df.residual
## 1   976688        1803
```

```
# function to get point estimates and standard errors
# model - lm object
# mod_var - name of moderating variable in the interaction
instant_effect <- function(model, mod_var){
  # get interaction term name
  int.name <- names(model$coefficients)[[which(str_detect(names(model$coefficients), ":"))]]

  marg_var <- str_split(int.name, ":")[[1]][[which(str_split(int.name, ":")[[1]] != mod_var)]]

  # store coefficients and covariance matrix
  beta.hat <- coef(model)
  cov <- vcov(model)

  # possible set of values for mod_var
  if(class(model)[[1]] == "lm"){
```

```
    z <- seq(min(model$model[[mod_var]]), max(model$model[[mod_var]]))
  } else {
    z <- seq(min(model$data[[mod_var]]), max(model$data[[mod_var]]))
  }

  # calculate instantaneous effect
  dy.dx <- beta.hat[[marg_var]] + beta.hat[[int.name]] * z

  # calculate standard errors for instantaeous effect
  se.dy.dx <- sqrt(cov[marg_var, marg_var] +
                     z^2 * cov[int.name, int.name] +
                     2 * z * cov[marg_var, int.name])

  # combine into data frame
  data_frame(z = z,
             dy.dx = dy.dx,
             se = se.dy.dx)
}

# point range plot
instant_effect(inter_biden, "educ") %>%
  ggplot(aes(z, dy.dx,
             ymin = dy.dx - 1.96 * se,
             ymax = dy.dx + 1.96 * se)) +
  geom_pointrange() +
  geom_hline(yintercept = 0, linetype = 2) +
  labs(title = "Marginal effect of age",
       subtitle = "By respondent education level",
       x = "Respondent education level",
       y = "Estimated marginal effect")
```



The ratio of the point estimate to the standard error yields a t-stat of 10.19, which indicates that the marginal effect of age on biden rating conditional on education is very accurately estimated.

```
coef(inter_biden)[["educ"]] + coef(inter_biden)[["age:educ"]]
```

```
## [1] 1.61
```

```
vcov(inter_biden)
```

```
##             (Intercept)      age     educ  age:educ
## (Intercept)      91.462 -1.54528 -6.72588  0.114416
## age              -1.545  0.02907  0.11415 -0.002159
## educ             -6.726  0.11415  0.50978 -0.008739
## age:educ          0.114 -0.00216 -0.00874  0.000166
```

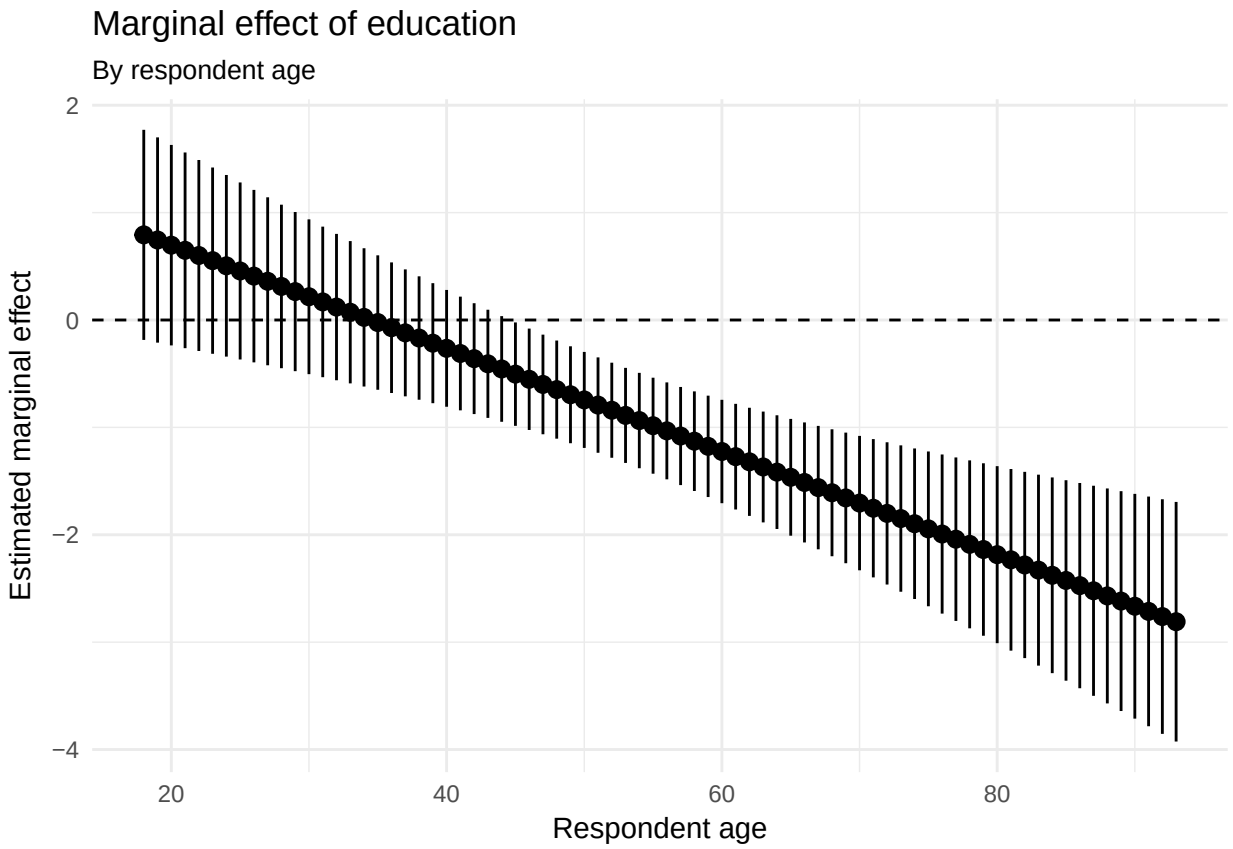
```
sqrt(vcov(inter_biden)["age", "age"] +
       (1)^2 * vcov(inter_biden)["age:educ", "age:educ"] +
       2 * 1 * vcov(inter_biden)["age", "age:educ"])
```

```
## [1] 0.158
```

2. The results of this interaction term model, as shown in the plot, show that the marginal effect of education changes as the age of the respondent changes. The marginal effect of education is never definitively different from zero (due to the confidence intervals passing through 0) until the respondent reaches an age of around 45. At that point, the marginal effect of education is negative and increasingly so as the respondents continue to be older.

```
# point range plot
instant_effect(inter_biden, "age") %>%
  ggplot(aes(z, dy.dx,
```

```
            ymin = dy.dx - 1.96 * se,
            ymax = dy.dx + 1.96 * se)) +
  geom_pointrange() +
  geom_hline(yintercept = 0, linetype = 2) +
  labs(title = "Marginal effect of education",
       subtitle = "By respondent age",
       x = "Respondent age",
       y = "Estimated marginal effect")
```



The ratio of the point estimate to the standard error yields a t-stat of 0.888, which indicates that the marginal effect of education on biden rating conditional on age is not accurately estimated.

```
coef(inter_biden)[["age"]] + coef(inter_biden)[["age:educ"]]
```

```
## [1] 0.624
```

```
vcov(inter_biden)
```

```
##             (Intercept)      age     educ  age:educ
## (Intercept)      91.462 -1.54528 -6.72588  0.114416
## age              -1.545  0.02907  0.11415 -0.002159
## educ             -6.726  0.11415  0.50978 -0.008739
## age:educ          0.114 -0.00216 -0.00874  0.000166
```

```
sqrt(vcov(inter_biden)["educ", "educ"] +
       (1)^2 * vcov(inter_biden)["age:educ", "age:educ"] +
       2 * 1 * vcov(inter_biden)["educ", "age:educ"])
```

```
## [1] 0.702
```

Missing Data

```
df = read.csv('biden.csv')
missing_biden <- lm(biden ~ age + female + educ, data=df)
tidy(missing_biden)
```

```
##          term estimate std.error statistic  p.value
## 1 (Intercept)  67.5579    3.5638     18.96 2.76e-73
## 2         age   0.0432    0.0323      1.34 1.81e-01
## 3      female   6.0221    1.0899      5.53 3.76e-08
## 4        educ  -0.8146    0.2222     -3.67 2.53e-04
```

There is significant missingness in this data.

```
df %>%
  select(biden, age, female, educ) %>%
  summarize_all(funs(sum(is.na(.)))) %>%
  knitr::kable()
```

biden	age	female	educ
460	46	0	11

Observing the plots below, I will transform the left-skewed variables `biden` and `educ` by taking the square and the right-skewed `age` by a log transformation.

```
df_lite <- df %>%
  select(biden, age, female, educ)

GGally::ggpairs(df_lite)
```

```
## Warning: Removed 460 rows containing non-finite values (stat_density).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 493 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 460 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 469 rows containing missing values
```

```
## Warning: Removed 493 rows containing missing values (geom_point).
```

```
## Warning: Removed 46 rows containing non-finite values (stat_density).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 46 rows containing missing values
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 51 rows containing missing values
```

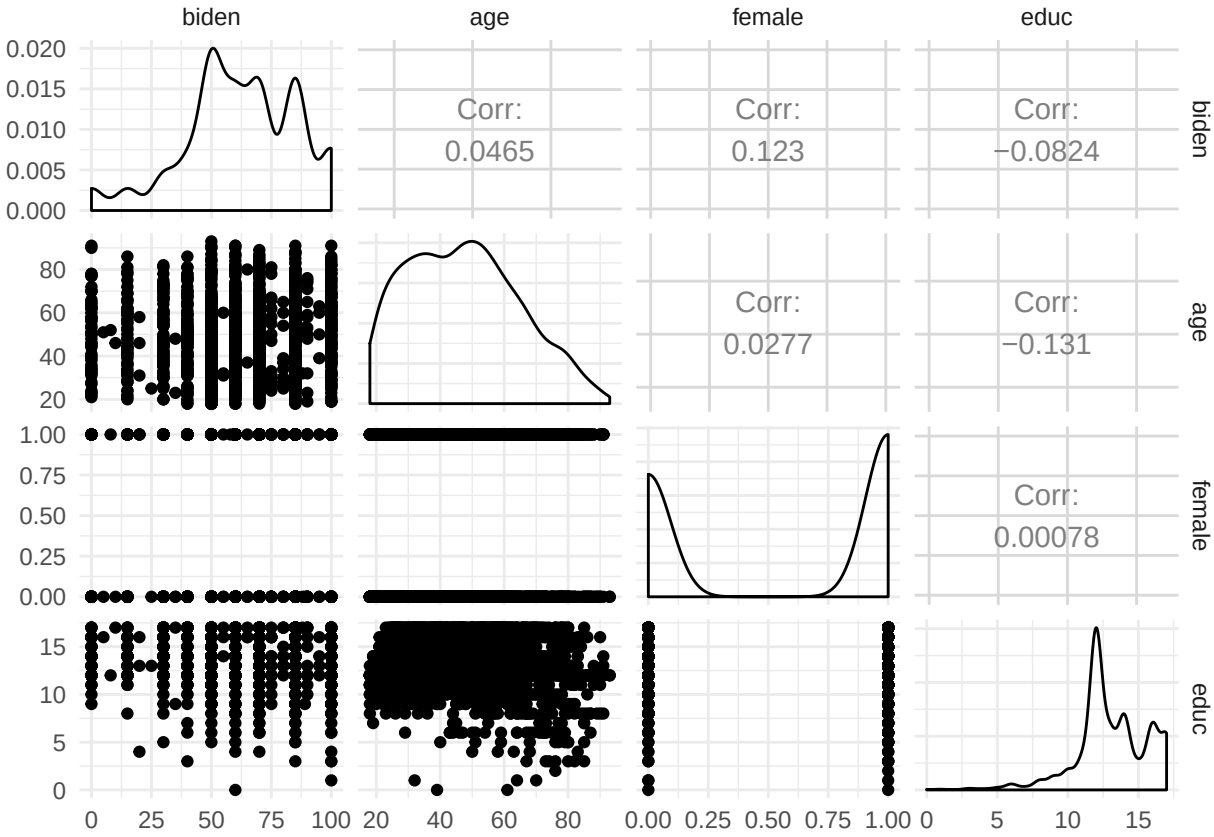
```
## Warning: Removed 460 rows containing missing values (geom_point).
```

```
## Warning: Removed 46 rows containing missing values (geom_point).
```

```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 11 rows containing missing values
```

```
## Warning: Removed 469 rows containing missing values (geom_point).
```

```
## Warning: Removed 51 rows containing missing values (geom_point).
## Warning: Removed 11 rows containing missing values (geom_point).
## Warning: Removed 11 rows containing non-finite values (stat_density).
```




```
df_lite.out <- amelia(df_lite, m = 5,
                      logs = c("age"),
                      sqrt = c("biden", "educ"))
```

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

```r
models_trans_imp <- data_frame(data = df_lite.out$imputations) %>%
  mutate(model = map(data, ~ lm(biden ~ age +
                                  female + educ,
                                data = .x)),
         coef = map(model, tidy)) %>%
  unnest(coef, .id = "id")
models_trans_imp
```

```
## # A tibble: 20 × 6
##       id        term estimate std.error statistic   p.value
##    <chr>       <chr>    <dbl>     <dbl>     <dbl>     <dbl>
## 1   imp1 (Intercept)  70.5235    3.1419    22.446 3.89e-101
## 2   imp1         age   0.0372    0.0290     1.281  2.00e-01
## 3   imp1      female   5.6228    1.0157     5.536  3.44e-08
## 4   imp1        educ  -1.0143    0.1955    -5.188  2.32e-07
## 5   imp2 (Intercept)  66.6878    3.2113    20.767  5.40e-88
## 6   imp2         age   0.0529    0.0298     1.775  7.60e-02
## 7   imp2      female   6.3475    1.0366     6.123  1.07e-09
## 8   imp2        educ  -0.8154    0.1993    -4.091  4.45e-05
## 9   imp3 (Intercept)  64.1689    3.1553    20.337  9.67e-85
## 10  imp3         age   0.0685    0.0292     2.348  1.90e-02
## 11  imp3      female   5.9921    1.0181     5.885  4.55e-09
## 12  imp3        educ  -0.6471    0.1958    -3.305  9.63e-04
## 13  imp4 (Intercept)  69.0697    3.2460    21.278  6.36e-92
## 14  imp4         age   0.0341    0.0300     1.136  2.56e-01
## 15  imp4      female   6.1148    1.0506     5.821  6.68e-09
## 16  imp4        educ  -0.9340    0.2020    -4.623  3.99e-06
## 17  imp5 (Intercept)  66.8567    3.1624    21.141  7.27e-91
## 18  imp5         age   0.0141    0.0291     0.484  6.29e-01
## 19  imp5      female   6.8444    1.0184     6.721  2.26e-11
## 20  imp5        educ  -0.6920    0.1964    -3.524  4.33e-04
```

```r
# compare results
mi.meld.plus <- function(df_tidy){
  # transform data into appropriate matrix shape
  coef.out <- df_tidy %>%
    select(id:estimate) %>%
    spread(term, estimate) %>%
    select(-id)

  se.out <- df_tidy %>%
    select(id, term, std.error) %>%
    spread(term, std.error) %>%
    select(-id)

  combined.results <- mi.meld(q = coef.out, se = se.out)

  data_frame(term = colnames(combined.results$q.mi),
             estimate.mi = combined.results$q.mi[1, ],
             std.error.mi = combined.results$se.mi[1, ])
}

# compare results
tidy(missing_biden) %>%
```

```
  left_join(mi.meld.plus(models_trans_imp)) %>%
  select(-statistic, -p.value)
```

```
## Joining, by = "term"
```

```
##          term estimate std.error estimate.mi std.error.mi
## 1 (Intercept)  67.5579    3.5638     67.4613        4.155
## 2         age   0.0432    0.0323      0.0413        0.037
## 3      female   6.0221    1.0899      6.1843        1.141
## 4        educ  -0.8146    0.2222     -0.8206        0.261
```

From the estimates above, we notice differences from our original model to the new model made with multipled imputed data. The largest differences can be seen with the `female` and `educ` variables. In particular, the effect of the `female` variable is dampened from 6.02 to 5.64 once the data has been imputed. The coefficient on `educ` has become slightly more negative from -0.814 to -0.841. The newly imputed data also results in large standard errors for all three variables.