PS3

Xinzhu Sun

5/13/2017

```
library(tidyverse)
library(modelr)
library(broom)
library(dplyr)
library(ggplot2)
library(readr)
library(forcats)
library(pROC)
library(lmtest)
library(GGally)
library(stringr)
library(car)
library(titanic)
library(haven)
library(plotly)
library(coefplot)
library(rcfss)
library(RColorBrewer)
library(MVN)
library(Amelia)
library(purrr)
options(digits = 3)
options(na.action = na.warn)
set.seed(1234)
theme_set(theme_minimal())
```

Regression diagostics

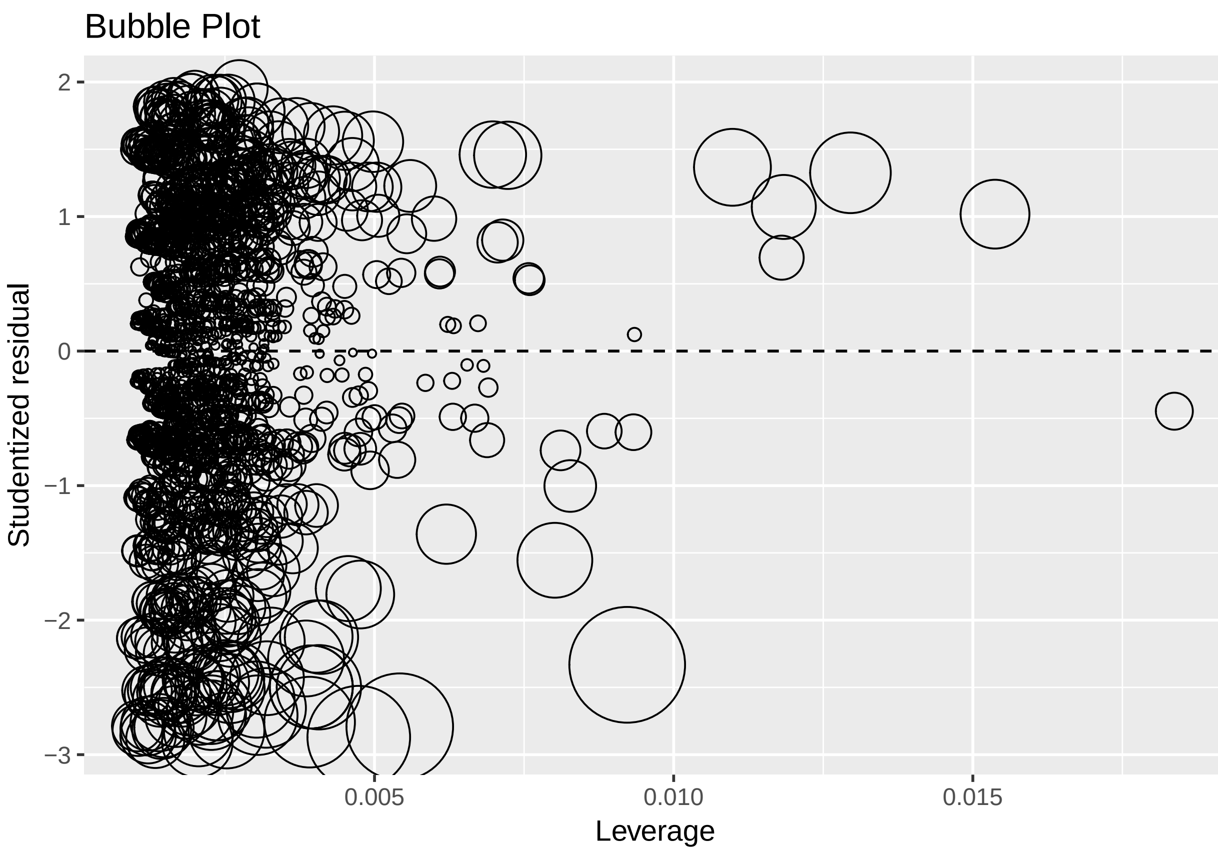
```
biden_dat <- read_csv("biden.csv") %>%
  na.omit()
```

```
## Parsed with column specification:
## cols(
##   biden = col_integer(),
##   female = col_integer(),
##   age = col_integer(),
##   educ = col_integer(),
##   dem = col_integer(),
##   rep = col_integer()
## )
```

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

1. Test the model to identify any unusual and/or influential observations. Identify how you would treat these observations moving forward with this research.

```r
# add key statistics
biden_augment <- biden_dat %>%
  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) +
  scale_size_continuous(range = c(1, 20)) +
  labs(title = "Bubble Plot",
       x = "Leverage",
       y = "Studentized residual") +
  theme(legend.position = "none")
```




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

```
## # A tibble: 74 × 9
##    biden female   age  educ   dem   rep         hat     student
##    <int>  <int> <int> <int> <int> <int>       <dbl>       <dbl>
## 1     70      0    80    17     0     0 0.005036344  0.56855300
## 2     70      1    44     7     1     0 0.004958796 -0.01898772
```

```
## 3     100      1    64     1     1     0 0.015371125  1.01786735
## 4     100      1    76     3     1     0 0.011840543  1.07145720
## 5      60      1    84    16     0     0 0.004456532 -0.17805204
## 6      60      1    63     4     0     0 0.009328015 -0.60292076
## 7      85      0    18     8     1     0 0.005995042  0.98460489
## 8      70      0    79     9     1     0 0.004614309  0.26258680
## 9      50      1    22     9     0     0 0.004496749 -0.76762784
## 10     50      1    23     8     0     0 0.005378701 -0.80826794
## # ... with 64 more rows, and 1 more variables: cooksd <dbl>
```

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

```
## # A tibble: 82 × 9
##    biden female   age  educ   dem   rep         hat   student      cooksd
##    <int>  <int> <int> <int> <int> <int>       <dbl>     <dbl>       <dbl>
## 1      0      1    70    12     0     1 0.002038099 -2.905524 0.004292512
## 2      0      0    45    12     0     1 0.001415510 -2.589766 0.002369281
## 3      0      0    40    14     0     0 0.001359874 -2.503407 0.002127285
## 4     15      0    62     8     0     1 0.004112951 -2.126852 0.004661332
## 5     15      1    20    13     0     0 0.002600197 -2.124956 0.002937181
## 6      0      1    38    14     1     0 0.001217550 -2.768766 0.002327692
## 7      0      0    34    12     0     0 0.001777029 -2.570223 0.002930895
## 8      0      0    21    13     0     1 0.002587140 -2.508993 0.004070149
## 9     15      1    29    12     0     1 0.001979288 -2.179191 0.002349621
## 10     0      0    36    13     0     1 0.001489411 -2.534889 0.002388994
## # ... with 72 more rows
```

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

```
## # A tibble: 90 × 9
##    biden female   age  educ   dem   rep         hat   student      cooksd
##    <int>  <int> <int> <int> <int> <int>       <dbl>     <dbl>       <dbl>
## 1      0      1    70    12     0     1 0.002038099 -2.905524 0.004292512
## 2      0      0    45    12     0     1 0.001415510 -2.589766 0.002369281
## 3     15      0    62     8     0     1 0.004112951 -2.126852 0.004661332
## 4     15      1    20    13     0     0 0.002600197 -2.124956 0.002937181
## 5    100      1    64     1     1     0 0.015371125  1.017867 0.004043401
## 6    100      0    19    12     0     0 0.003037348  1.784812 0.002423340
## 7    100      0    19    12     1     0 0.003037348  1.784812 0.002423340
## 8      0      1    38    14     1     0 0.001217550 -2.768766 0.002327692
## 9    100      1    76     3     1     0 0.011840543  1.071457 0.003438734
## 10     0      0    34    12     0     0 0.001777029 -2.570223 0.002930895
## # ... with 80 more rows
```

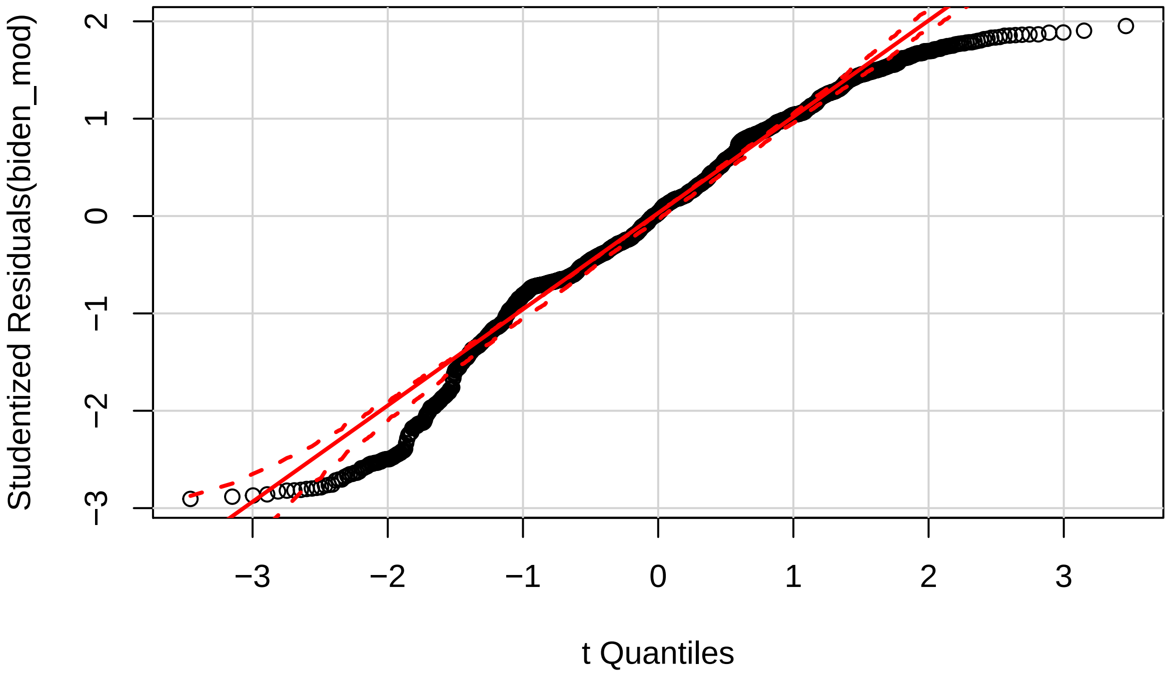
The bubble plot shows that there are observation has high leverage and low discrepancy, observation has high leverage and high discrepancy, and observations have low leverage but very high discrepancy.That is, there are unusual and influential observations in the data.

If this is because the data is just wrong (miscoded, mismeasured, misentered, etc.), then either fix the error, impute a plausible value for the observation, or drop the observations.

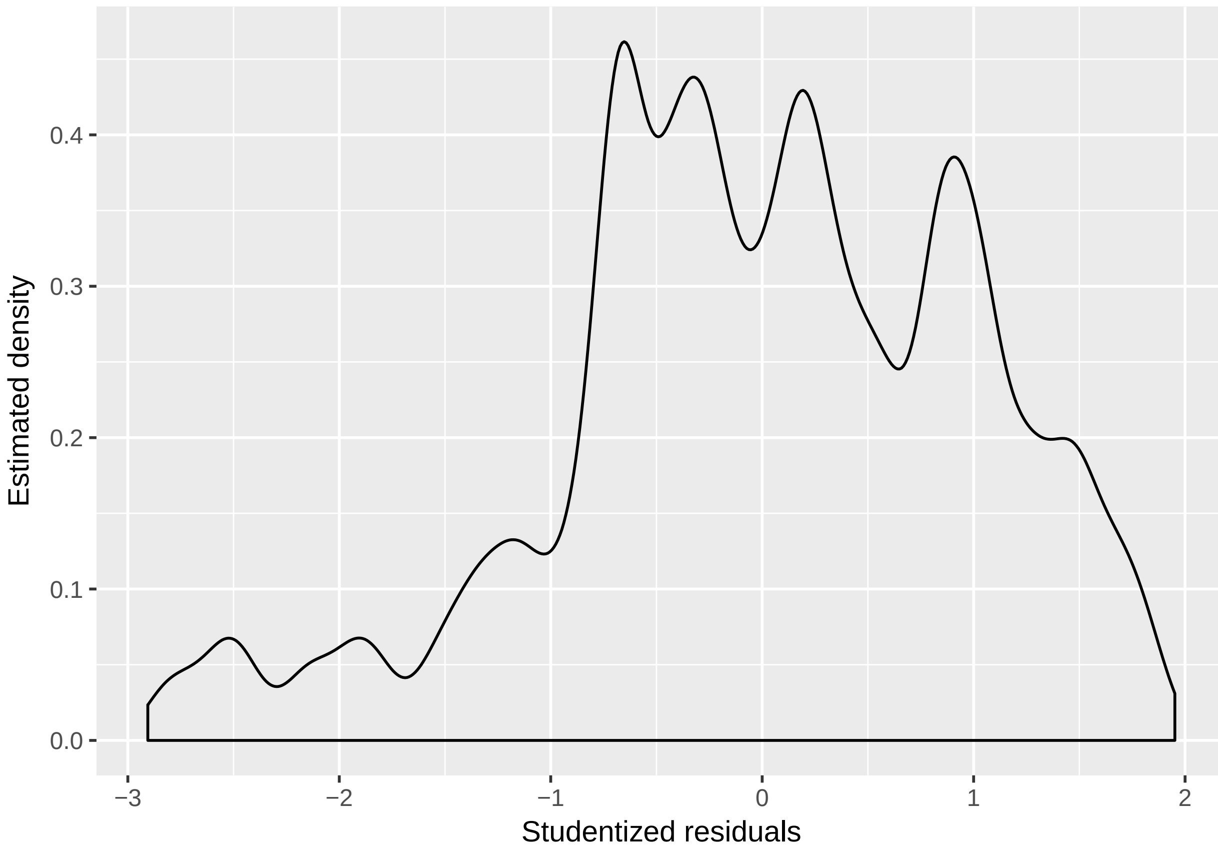
If this is because the data for a particular observation is just strange, then I'll first identify whether it is because something unusual/weird/singular happened to that data point. If the answer is yes and that "something" is important to the theory being tested, then I'd respecify the model. If the answer is no, then I'd drop the offending observation from the analysis. If the data are strange for no apparent reason, then I'd drop the observation and do robustness check.

2. Test for non-normally distributed errors.

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




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

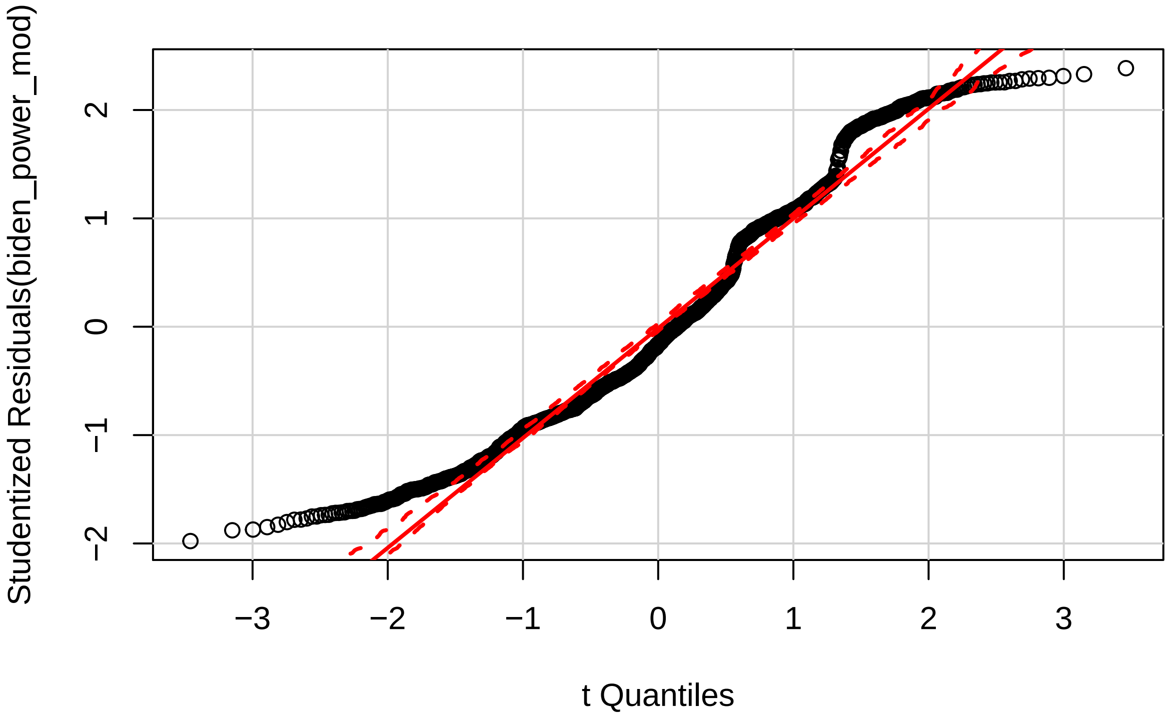


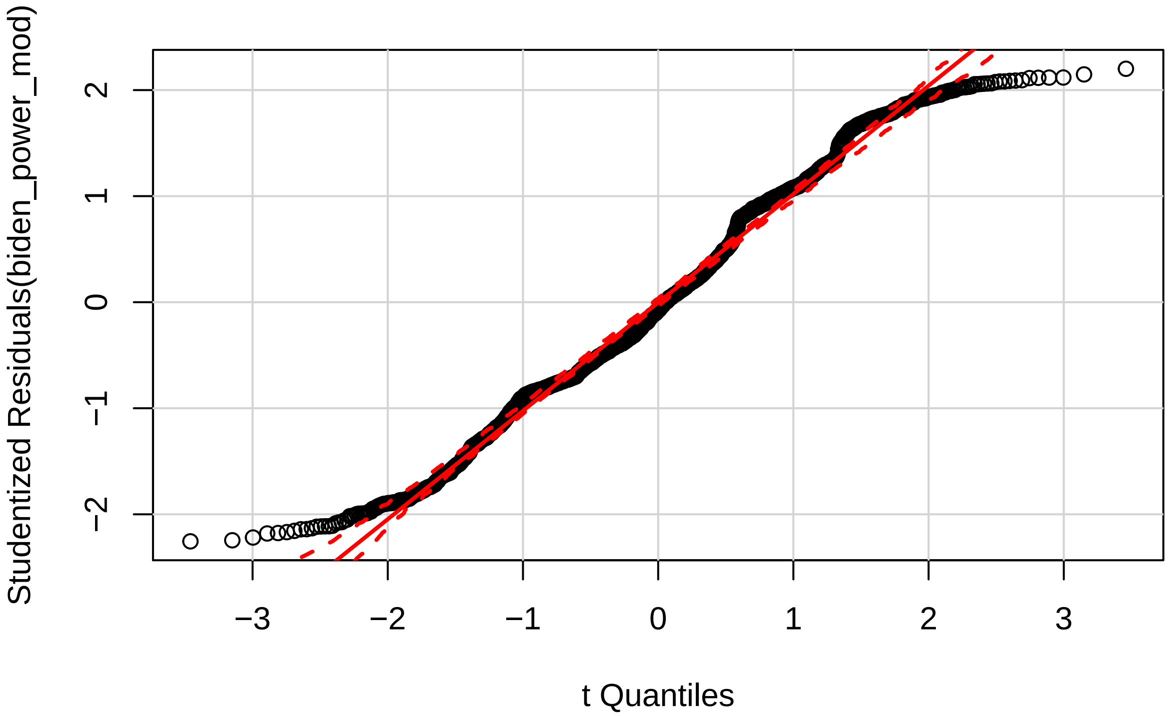
The quantile-comparison plot shows that the assumption of normally ditribution has been violated. From the density plot of the studentized residuals, we can also see that the residuals are skewed.

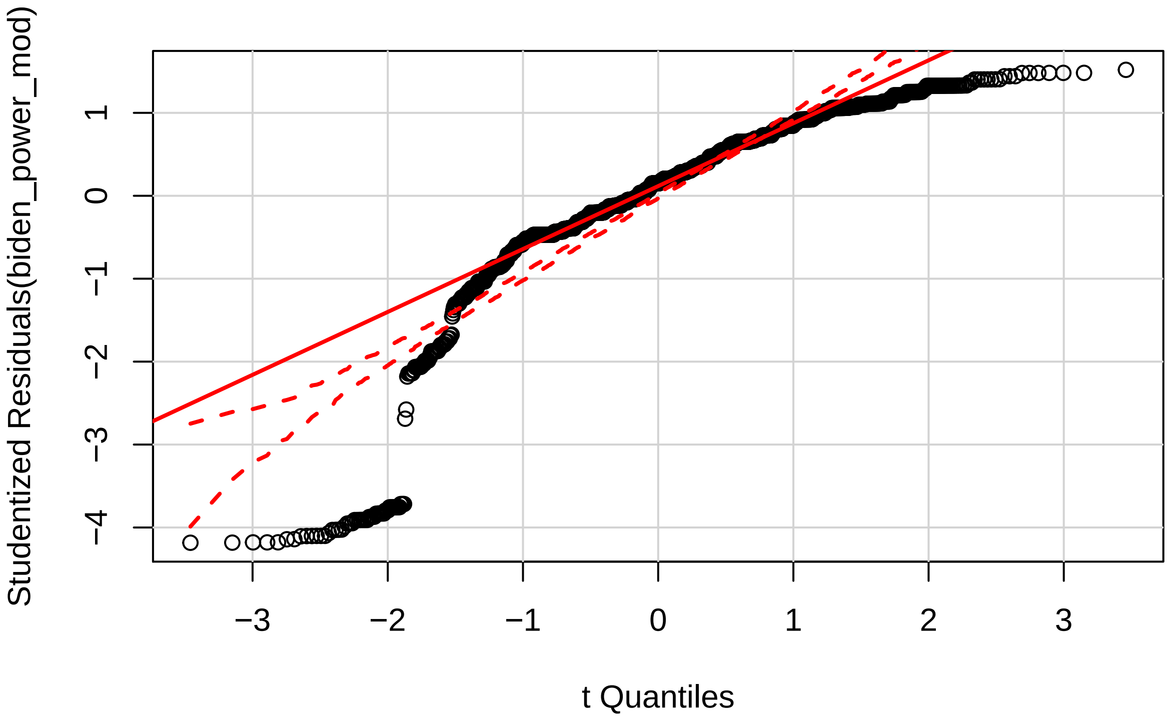
Power and log transformations are typically used to correct this problem. Here, trial and error reveals that by power transforming the biden variable, the distribution of the residuals becomes much more symmetric:

```
biden_fix <- function(power){
  if (power < 0){
  temp_biden <- biden_dat %>%
    mutate(biden_power = - 1 / (biden ^ power))

  biden_power_mod <- temp_biden %>%
    lm(biden_power ~ age + female + educ, data = .)
  } else {
    temp_biden <- biden_dat %>%
      mutate(biden_power = (biden ^ power))

    biden_power_mod <- temp_biden %>%
      lm(biden_power ~ age + female + educ, data = .)
  }

  car::qqPlot(biden_power_mod)
}

powers <- c(2, 1.5, 0.5)

for (power in powers){
  biden_fix(power)
```

```
}
```


Studentized Residuals(biden_power_mod)
t Quantiles
−3
−2
−1
0
1
2
3
−2
−1
0
1
2

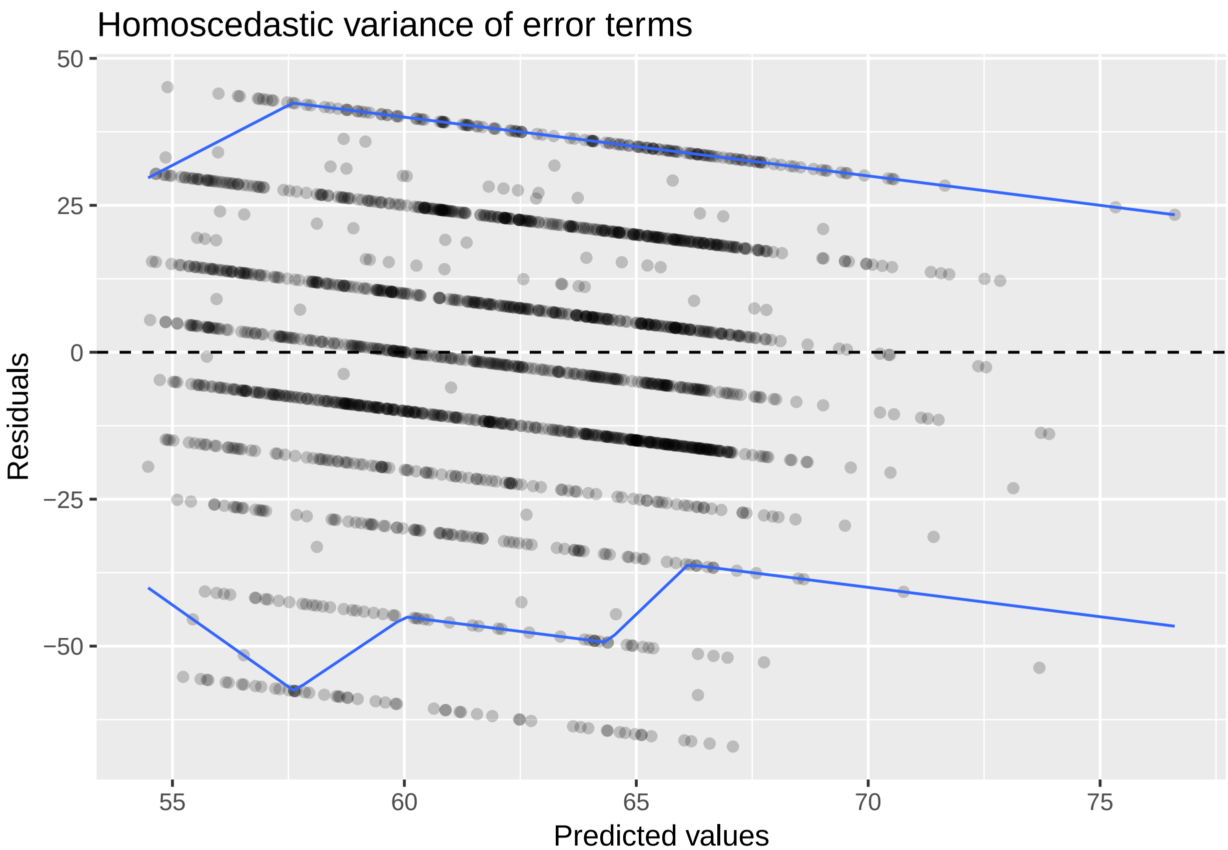

Studentized Residuals(biden_power_mod)
t Quantiles
−3
−2
−1
0
1
2
3
−2
−1
0
1
2



3. Test for heteroscedasticity in the model.

```
biden_dat %>%
  add_predictions(biden_mod) %>%
  add_residuals(biden_mod) %>%
  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 = "Homoscedastic variance of error terms",
       x = "Predicted values",
       y = "Residuals")
```

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




```
bptest(biden_mod)
```

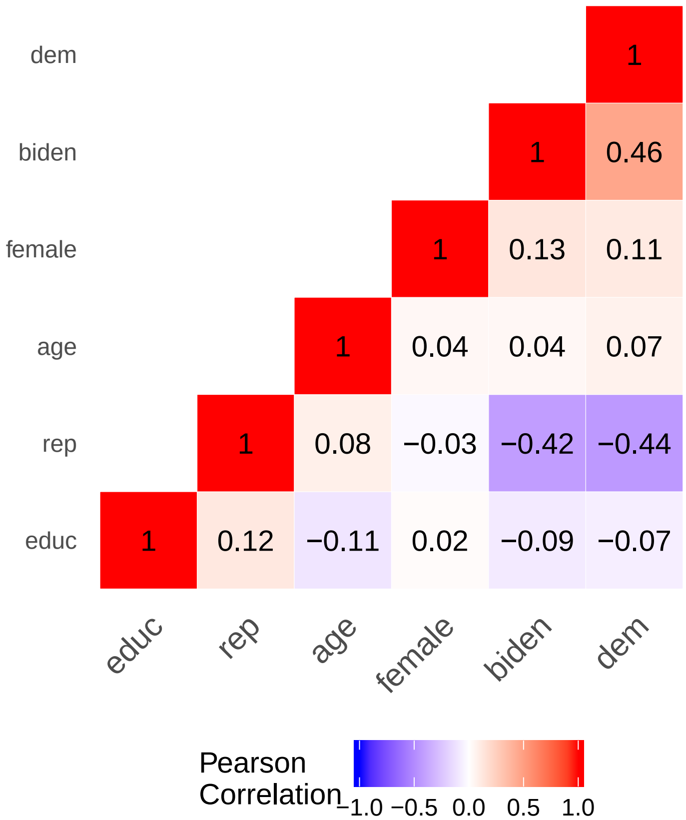
```
##
##  studentized Breusch-Pagan test
##
## data:  biden_mod
## BP = 22.559, df = 3, p-value = 4.989e-05
```

From the residual plot and Breusch-Pagan test (P-value is very low), we can learn that there is heteroskedasticity present in the errors. If left unaccounted for, this could distort the estimates for the standard error for each coefficient either up or down.

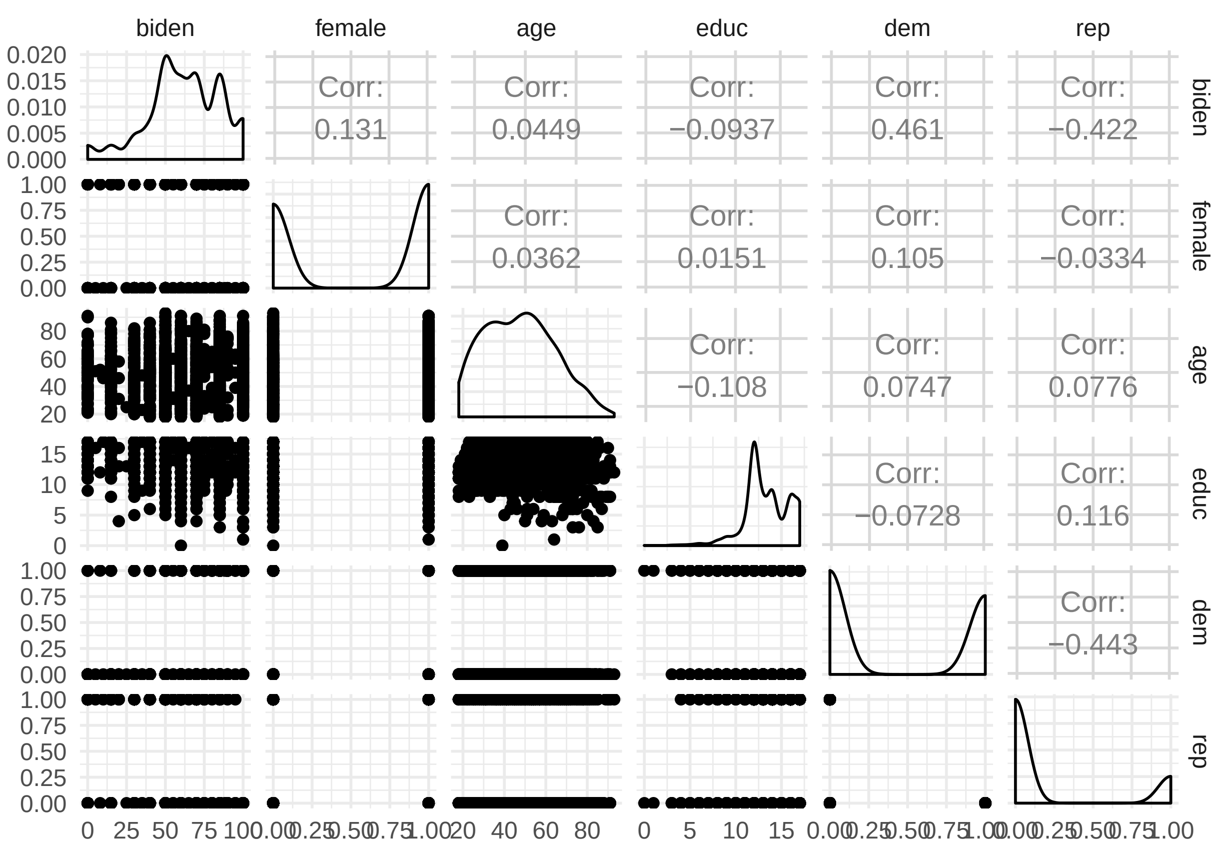
4. Test for multicollinearity.

```
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)
```

```r
  # 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_dat, is.numeric))
```




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



```r
vif(biden_mod)
```

```
##    age female   educ
##   1.01   1.00   1.01
```

Thus, there is no multicollinearity exists in the model.

Interaction terms

```r
biden_dat <- read_csv("biden.csv") %>%
   na.omit()
```

```
## Parsed with column specification:
## cols(
##   biden = col_integer(),
##   female = col_integer(),
##   age = col_integer(),
##   educ = col_integer(),
##   dem = col_integer(),
##   rep = col_integer()
## )
```

```r
biden_mod_2 <- lm(biden ~ age * educ, data = biden_dat)
```

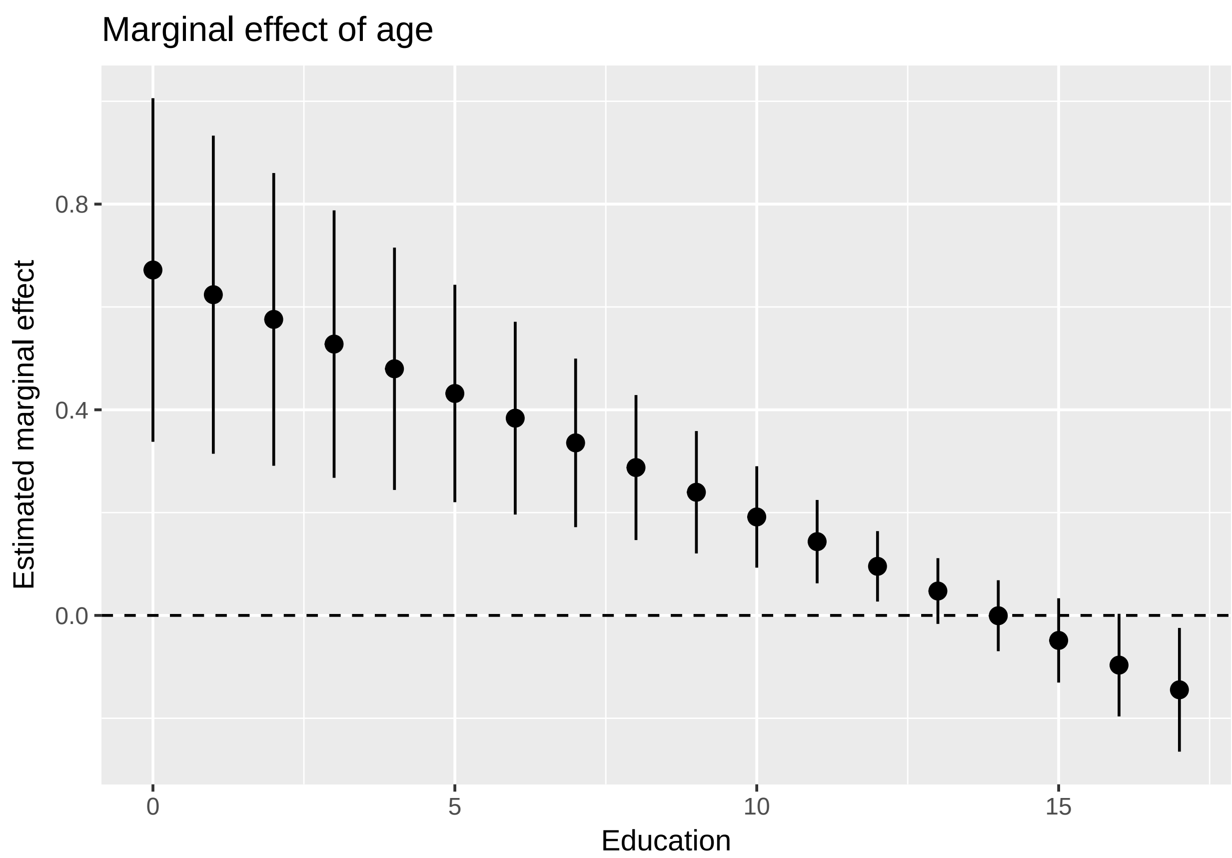
1. Evaluate the marginal effect of age on Joe Biden thermometer rating, conditional on education.

```r
coef(biden_mod_2)[["educ"]] + coef(biden_mod_2)[["age:educ"]]
```

```
## [1] 1.609391
```

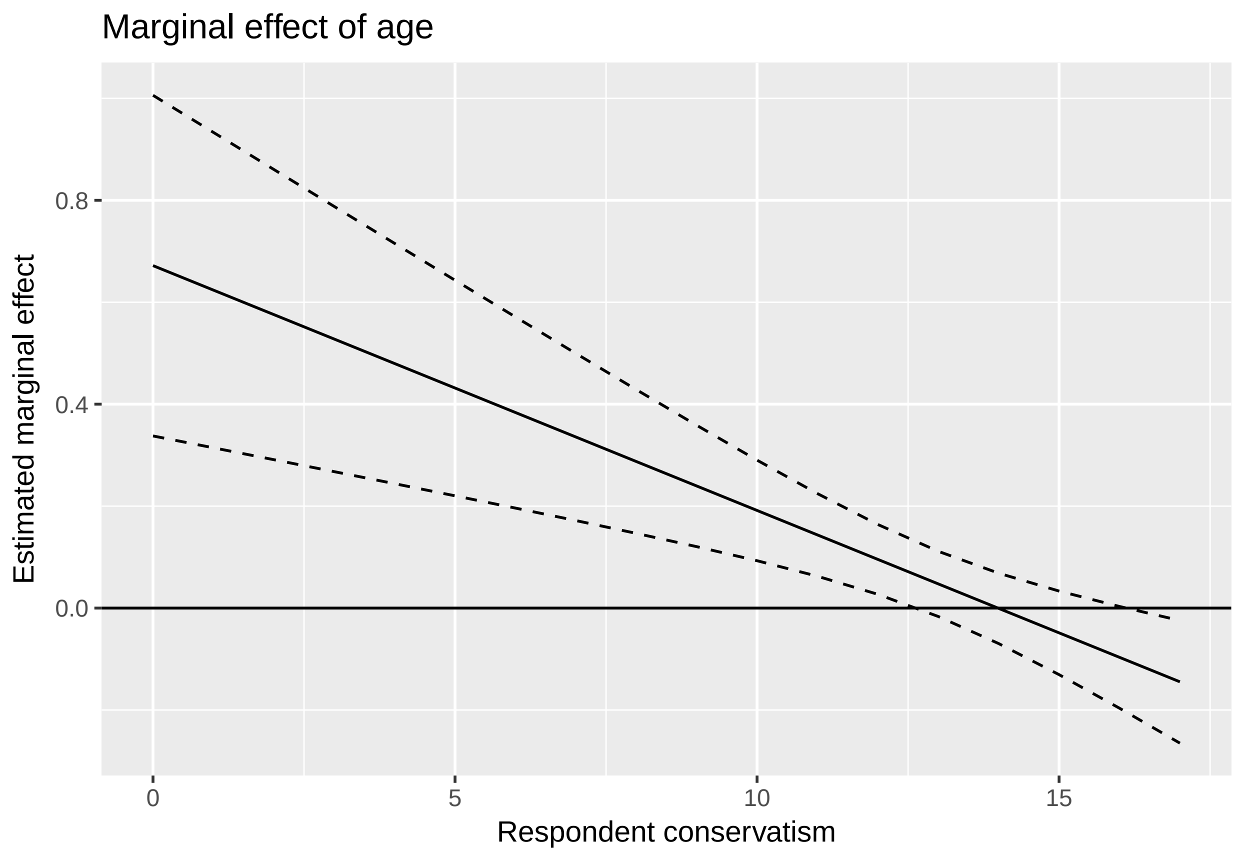
```r
# function to get point estimates and standard errors
# model - lm object
# mod_var - name of moderating variable in the interaction
instant_effect <- function(model, mod_var){
  # get interaction term name
  int.name <- names(model$coefficients)[[which(str_detect(names(model$coefficients), ":"))]]

  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
```

```r
  # 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)
}

instant_effect(biden_mod_2, "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",
       x = "Education",
       y = "Estimated marginal effect")
```




```r
# line plot
instant_effect(biden_mod_2, "educ") %>%
  ggplot(aes(z, dy.dx)) +
  geom_line() +
  geom_line(aes(y = dy.dx - 1.96 * se), linetype = 2) +
  geom_line(aes(y = dy.dx + 1.96 * se), linetype = 2) +
```

```r
  geom_hline(yintercept = 0) +
  labs(title = "Marginal effect of age",
       x = "Respondent conservatism",
       y = "Estimated marginal effect")
```



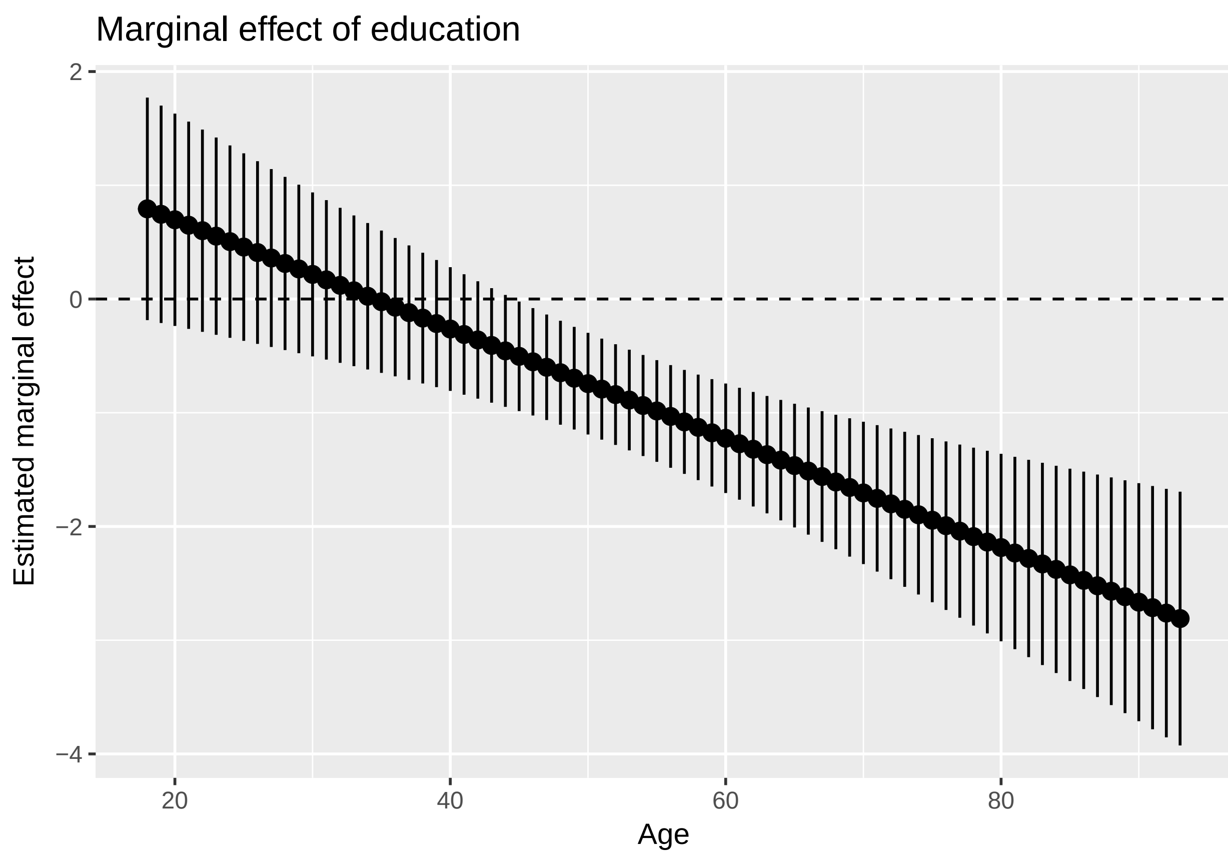

```r
linearHypothesis(biden_mod_2, "age + age:educ")
```

```
## Linear hypothesis test
##
## Hypothesis:
## age  + age:educ = 0
##
## Model 1: restricted model
## Model 2: biden ~ 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
```

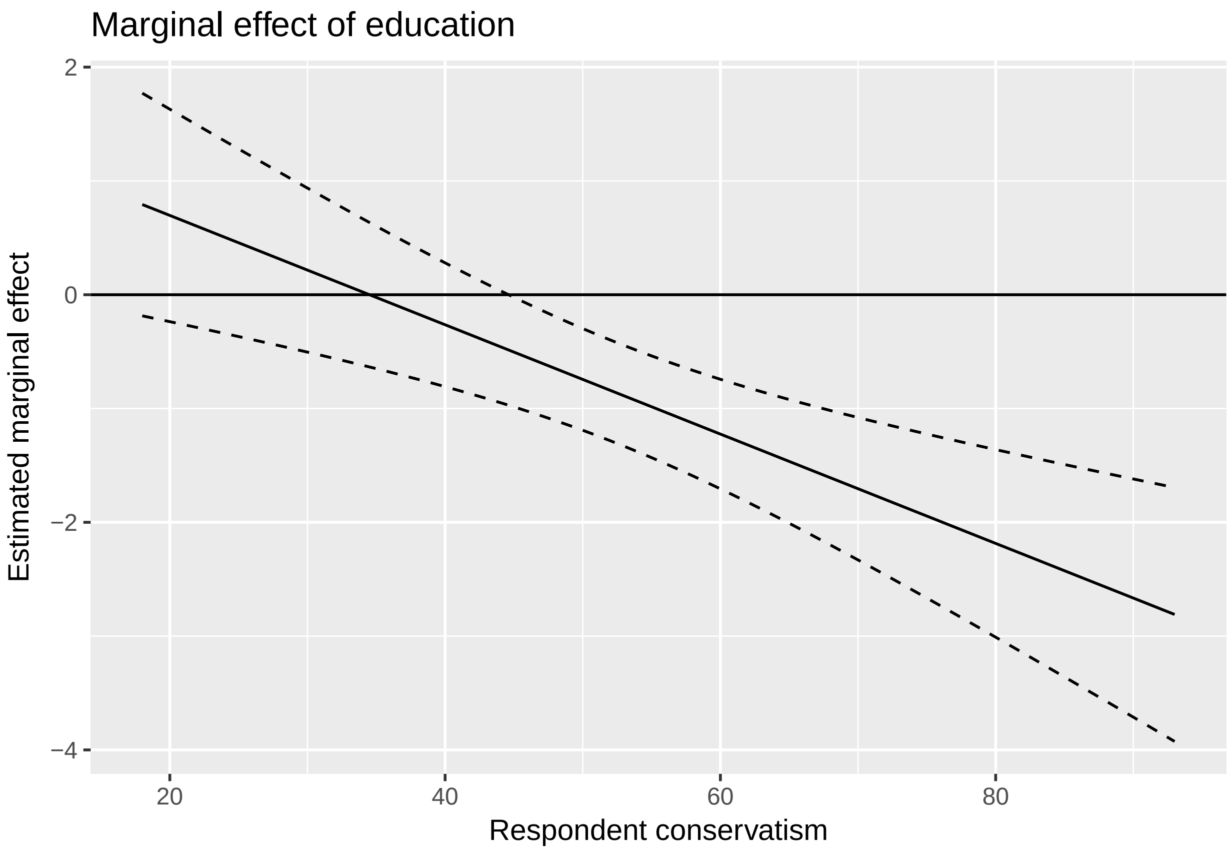
The p-value of the marginal effect of age is significant. The magnitude and direction are shown in the plots.As education level of respondent increase, the marginal effect of age decreases from 0.7 to almost -0.1. The 95% confidence interval is shown in the graph.

2. Evaluate the marginal effect of education on Joe Biden thermometer rating, conditional on age.

```
instant_effect(biden_mod_2, "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",
       x = "Age",
       y = "Estimated marginal effect")
```




```
# line plot
instant_effect(biden_mod_2, "age") %>%
  ggplot(aes(z, dy.dx)) +
  geom_line() +
  geom_line(aes(y = dy.dx - 1.96 * se), linetype = 2) +
  geom_line(aes(y = dy.dx + 1.96 * se), linetype = 2) +
  geom_hline(yintercept = 0) +
  labs(title = "Marginal effect of education",
       x = "Respondent conservatism",
       y = "Estimated marginal effect")
```




```r
linearHypothesis(biden_mod_2, "educ + age:educ")
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ  + age:educ = 0
##
## Model 1: restricted model
## Model 2: biden ~ 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 p-value of the marginal effect of eduction is significant. The magnitude and direction are shown in the plots.As age of respondent increase, the marginal effect of age decreases from 0.8 to almost -2.8. The 95% confidence interval is shown in the graph.

Missing data

Note: female is a binary variable.

```r
biden_raw <- read_csv("biden.csv")
```

```
## Parsed with column specification:
## cols(
```

```
##   biden = col_integer(),
##   female = col_integer(),
##   age = col_integer(),
##   educ = col_integer(),
##   dem = col_integer(),
##   rep = col_integer()
## )
```

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

biden	age	female	educ
460	46	0	11

```r
hzTest(biden_dat %>%
         select(biden, age, educ), cov = TRUE, qqplot = FALSE)
```

```
##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : biden_dat %>% select(biden, age, educ)
##
##   HZ      : 7.859611
##   p-value : 0
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------
```

```r
uniNorm(biden_dat %>%
          select(biden, age, educ), type = "SW", desc = FALSE)
```

```
## $`Descriptive Statistics`
## NULL
##
## $`Shapiro-Wilk's Normality Test`
##    Variable Statistic   p-value Normality
## 1   biden      0.9475         0    NO
## 2    age       0.9795         0    NO
## 3   educ       0.9180         0    NO
```

Henze-Zirkler's Multivariate Normality Test and Shapiro-Wilk's Normality Test both tell us that the biden data are not multivariate normal. To fix this problem, I'll try power transformation to coerce all of predictors to be MVN distributed.

```r
# square
biden_dat <- biden_dat %>%
  mutate(sq_biden = biden^2,
         sq_educ = educ^2,
         sq_age = age^2)

hzTest(biden_dat %>%
         select(sq_biden, sq_educ, sq_age))
```

```
##   Henze-Zirkler's Multivariate Normality Test
```

```
## ---------------------------------------------
##   data : biden_dat %>% select(sq_biden, sq_educ, sq_age)
##
##   HZ      : 16.71075
##   p-value : 0
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------
```

```
uniNorm(biden_dat %>%
          select(sq_biden, sq_educ, sq_age), type = "SW", desc = FALSE)
```

```
## $`Descriptive Statistics`
## NULL
##
## $`Shapiro-Wilk's Normality Test`
##    Variable Statistic   p-value Normality
## 1 sq_biden     0.9302         0    NO
## 2  sq_educ     0.9296         0    NO
## 3  sq_age      0.9270         0    NO
```

```
# 1.5 power
biden_dat <- biden_dat %>%
  mutate(power_biden = biden^1.5,
         power_educ = educ^1.5,
         power_age = age^1.5)

hzTest(biden_dat %>%
         select(power_biden, power_educ, power_age))
```

```
##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : biden_dat %>% select(power_biden, power_educ, power_age)
##
##   HZ      : 10.58166
##   p-value : 0
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------
```

```
uniNorm(biden_dat %>%
          select(power_biden, power_educ, power_age), type = "SW", desc = FALSE)
```

```
## $`Descriptive Statistics`
## NULL
##
## $`Shapiro-Wilk's Normality Test`
##      Variable Statistic   p-value Normality
## 1 power_biden    0.9542         0    NO
## 2 power_educ     0.9314         0    NO
## 3  power_age     0.9594         0    NO
```

```
# square root
biden_dat <- biden_dat %>%
  mutate(sqrt_biden = sqrt(biden),
         sqrt_educ = sqrt(educ),
         sqrt_age = sqrt(age))
```

```r
hzTest(biden_dat %>%
         select(sqrt_biden, sqrt_educ, sqrt_age))
```

```
##   Henze-Zirkler's Multivariate Normality Test
## ---------------------------------------------
##   data : biden_dat %>% select(sqrt_biden, sqrt_educ, sqrt_age)
##
##   HZ      : 12.85594
##   p-value : 0
##
##   Result  : Data are not multivariate normal.
## ---------------------------------------------
```

```r
uniNorm(biden_dat %>%
          select(sqrt_biden, sqrt_educ, sqrt_age), type = "SW", desc = FALSE)
```

```
## $`Descriptive Statistics`
## NULL
##
## $`Shapiro-Wilk's Normality Test`
##     Variable Statistic   p-value Normality
## 1 sqrt_biden    0.8146         0    NO
## 2 sqrt_educ     0.8639         0    NO
## 3  sqrt_age     0.9841         0    NO
```

Although after all the power transformation I tried, the data is still not multivariate distributed. But from the results and plots int the first section part 2, power 1.5 provides the best adjustment.

```r
biden_transform = biden_raw %>%
   mutate(power_biden = biden^1.5,
         power_educ = educ^1.5,
         power_age = age^1.5)

biden.out <- amelia(as.data.frame(biden_transform), m = 5)
```

```
## -- Imputation 1 --
##
##   1  2  3  4  5  6  7
##
## -- Imputation 2 --
##
##   1  2  3  4  5  6  7  8  9
##
## -- Imputation 3 --
##
##   1  2  3  4  5
##
## -- Imputation 4 --
##
##   1  2  3  4  5  6  7  8  9
##
## -- Imputation 5 --
##
##   1  2  3  4  5  6  7  8
```

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

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

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

##          term    estimate  std.error estimate.mi std.error.mi
## 1 (Intercept) 68.62101396 3.59600465 68.03239896   3.52806602
## 2         age  0.04187919 0.03248579  0.05507125   0.03047478
## 3      female  6.19606946 1.09669702  5.47057898   1.03216379
## 4        educ -0.88871263 0.22469183 -0.87057296   0.21193019
```

In conclusion, it could be shown from the table of comparison above, conducting imputation after putting a power transformation on the educ variable for the sake of the normality assumption and then comparing the result with the non-imputed model, it could be seen that except female's coefficient and standard error remains almost identical, the rest of the coefficients reduced and the standard error of those coefficients are also reduced.