hw7

```r
library(here)
library(tidyverse)
library(sandwich)
data <- read.table(here("data", "rdcongress.txt"))
# response = dwnom
# forcing variable demvoteshare
```

a

```r
data_clean <- data %>%
  select(dwnom, demvoteshare) %>%
  drop_na() %>%
  mutate(demvoteshare_t = demvoteshare - 0.5)
```

b

First fit the models and compute robust variance estiamtes for their parameters

```r
lm_lower <- lm(
  dwnom ~ demvoteshare_t,
  data = filter(data_clean, between(demvoteshare_t, -0.02, 0))
  )
lower_vcov <- vcovHC(lm_lower, type="HC")

lm_upper <- lm(
  dwnom ~ demvoteshare_t,
  data = filter(data_clean, between(demvoteshare_t, 0, 0.02))
)
upper_vcov <- vcovHC(lm_upper, type="HC")
```

Using the provided formula, compute 95% confidence intervals using robust standard errors for the two regression lines

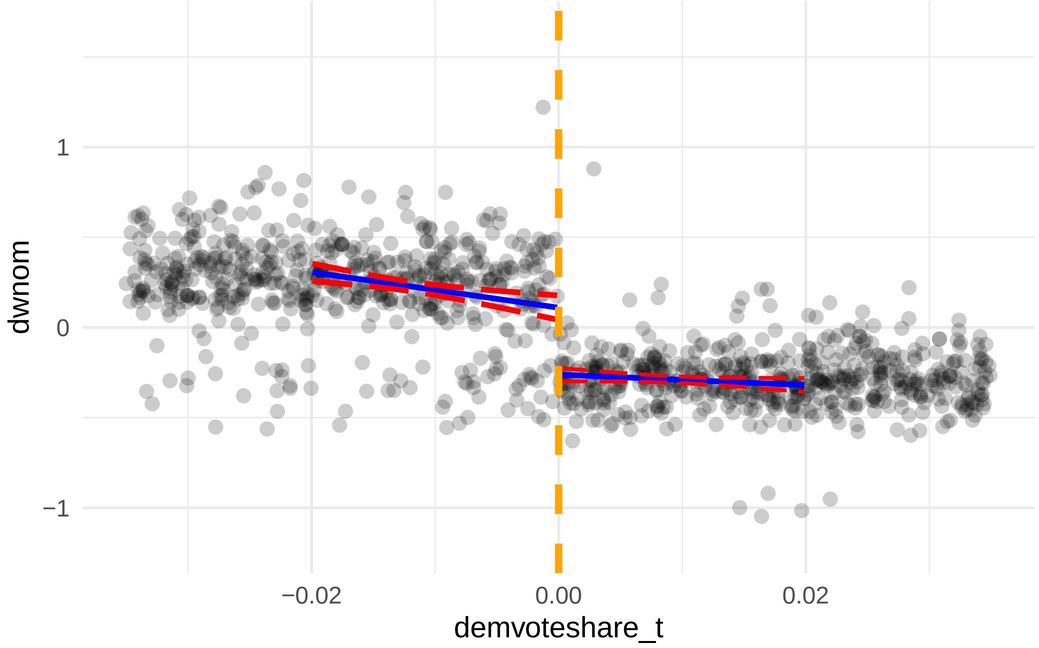
```
# var(intercept) + x2 var(slope) + 2*x*cov(slope, intercept)

se_lower <- fortify(lm_lower) %>%
  mutate(
    var_l = lower_vcov[1, 1] + demvoteshare_t^2*lower_vcov[2, 2] +
      2*demvoteshare_t*lower_vcov[1, 2],
    se_l = sqrt(var_l)
    ) %>%
  mutate(lower = .fitted - 1.96*se_l,
         upper = .fitted + 1.96*se_l) %>%
  select(demvoteshare_t, upper, lower)

se_upper <- fortify(lm_upper) %>%
  mutate(
    var_u = upper_vcov[1, 1] + demvoteshare_t^2*upper_vcov[2, 2] +
      2*demvoteshare_t*upper_vcov[1, 2],
    se_u = sqrt(var_u)
    ) %>%
  mutate(lower = .fitted - 1.96*se_u,
         upper = .fitted + 1.96*se_u) %>%
  select(demvoteshare_t, upper, lower)
```

Plot the reggression lines, the confidence intervals and the cutoff all on one plot. We'll look at a restricted range of the forcing variable so we can see these things more easily. My substantive interpretation is that we do see a discontinuity in our forcing variable and that lower proportions of democratic vote share are associated with higher values of dwnom while higher proportions of democratic vote share are associated with lower values of dwnom.

```
data_clean %>%
  ggplot(aes(x = demvoteshare_t, y = dwnom)) +
  geom_point(size = 2, alpha = 0.2) +
  geom_line(data = fortify(lm_lower),
            aes(x = demvoteshare_t, y = .fitted),
            size = 1, color = "blue") +
  geom_line(data = se_lower,
            aes(x = demvoteshare_t, y = lower),
            size = 1, color = "red", linetype = "longdash") +
  geom_line(data = se_lower,
            aes(x = demvoteshare_t, y = upper),
            size = 1, color = "red", linetype = "longdash") +
  geom_line(data = fortify(lm_upper),
            aes(x = demvoteshare_t, y = .fitted),
            size = 1, color = "blue") +
  geom_line(data = se_upper,
            aes(x = demvoteshare_t, y = lower),
            size = 0.7, color = "red", linetype = "longdash") +
  geom_line(data = se_upper,
            aes(x = demvoteshare_t, y = upper),
            size = 0.7, color = "red", linetype = "longdash") +
  geom_vline(xintercept = 0, color = "orange", size = 1.3, linetype = "dashed") +
  theme_minimal() +
  xlim(c(-0.035, 0.035))
```



c

```
model_data <- data_clean %>%
  mutate(D = ifelse(demvoteshare > 0.5, 1, 0)) %>%
  filter(between(demvoteshare, 0.48, 0.52))

mod <- lm(dwnom ~ demvoteshare_t + D + demvoteshare_t*D, data = model_data)
```

```
se <- sqrt(vcovHC(mod, type="HC")["D", "D"])
ci <- mod$coefficients["D"] + 1.96*c(-1,1)*se

data.frame(
  est = mod$coefficients["D"],
  lower = ci[1],
  upper = ci[2]
)
```

```
         est      lower      upper
D -0.3710298 -0.4481374 -0.2939222
```

Significant at the $\alpha = 0.05$ level

```
z <- (mod$coefficients["D"])/se
p_val <- 2*pnorm(z, mean = 0, sd = 1)
p_val
```

```
           D
4.053552e-21
```

d

Create this "lagged" DW-NOMINATE variable that within each district (district) and congress number (cngrsnum) takes the value of dwnom from the previous congress number. Note that we need to delete the observations where there was a redistricting prior to the election. The year variable of these observations take the value of 1932, 1942, 1952, 1962, 1972, 1982, 1992, and 2002.

```
data_lagged <- data %>%
  select(dwnom, demvoteshare, district, cngrsnum, year, state) %>%
  mutate(demvoteshare_t = demvoteshare - 0.5) %>%
  filter(!(year %in% c(1932, 1942, 1952, 1962, 1972, 1982, 1992, 2002))) %>%
  drop_na() %>%
  arrange(state, district, cngrsnum) %>%
  group_by(state, district) %>%
  mutate(dwnom_lag = lag(dwnom)) %>%
  ungroup()
```

e

Again fit the separate models

```
lm_lower_lag <- lm(
  dwnom_lag ~ demvoteshare_t,
  data = filter(data_lagged, between(demvoteshare_t, -0.02, 0))
  )

lower_vcov_lag <- vcovHC(lm_lower_lag, type="HC")

lm_upper_lag <- lm(
  dwnom_lag ~ demvoteshare_t,
  data = filter(data_lagged, between(demvoteshare_t, 0, 0.02))
)
upper_vcov_lag <- vcovHC(lm_upper_lag, type="HC")
```

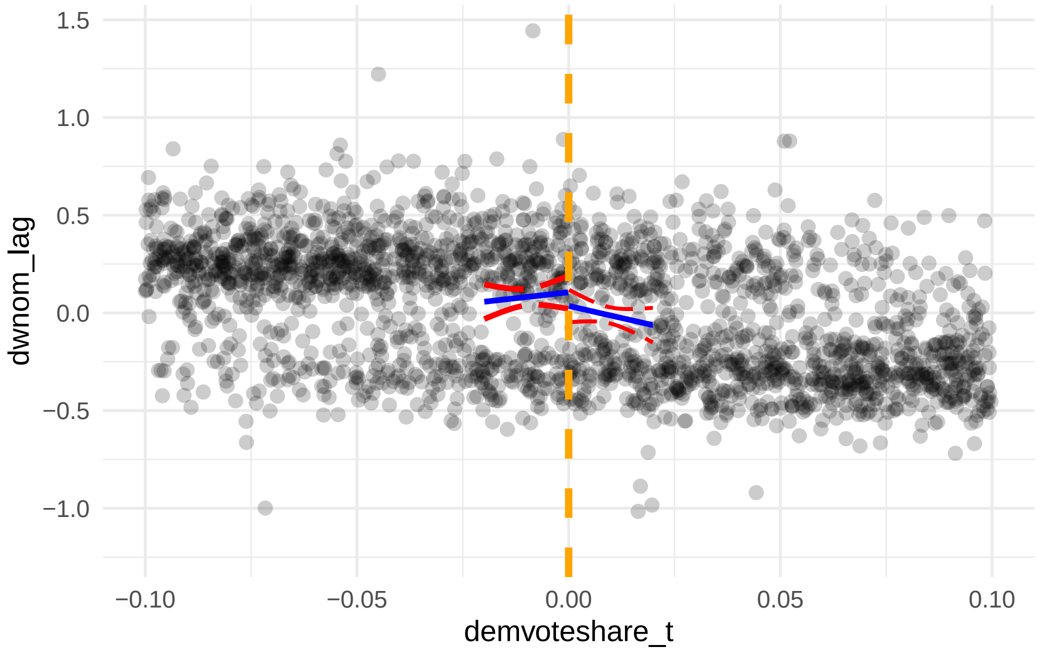
compute confidence intervals

```
se_lower_lag <- fortify(lm_lower_lag) %>%
  mutate(
    var_l = lower_vcov_lag[1, 1] + demvoteshare_t^2*lower_vcov_lag[2, 2] +
      2*demvoteshare_t*lower_vcov_lag[1, 2],
    se_l = sqrt(var_l)
    ) %>%
  mutate(lower = .fitted - 1.96*se_l,
         upper = .fitted + 1.96*se_l) %>%
  select(demvoteshare_t, upper, lower)

se_upper_lag <- fortify(lm_upper_lag) %>%
  mutate(
    var_u = upper_vcov_lag[1, 1] + demvoteshare_t^2*upper_vcov_lag[2, 2] +
      2*demvoteshare_t*upper_vcov_lag[1, 2],
    se_u = sqrt(var_u)
    ) %>%
  mutate(lower = .fitted - 1.96*se_u,
         upper = .fitted + 1.96*se_u) %>%
  select(demvoteshare_t, upper, lower)
```

and plot everything

```r
data_lagged %>%
  ggplot(aes(x = demvoteshare_t, y = dwnom_lag)) +
  geom_point(size = 2, alpha = 0.2) +
  geom_line(data = fortify(lm_lower_lag),
            aes(x = demvoteshare_t, y = .fitted),
            size = 1, color = "blue") +
  geom_line(data = se_lower_lag,
            aes(x = demvoteshare_t, y = lower),
            size = 1, color = "red", linetype = "longdash") +
  geom_line(data = se_lower_lag,
            aes(x = demvoteshare_t, y = upper),
            size = 1, color = "red", linetype = "longdash") +
  geom_line(data = fortify(lm_upper_lag),
            aes(x = demvoteshare_t, y = .fitted),
            size = 1, color = "blue") +
  geom_line(data = se_upper_lag,
            aes(x = demvoteshare_t, y = lower),
            size = 0.7, color = "red", linetype = "longdash") +
  geom_line(data = se_upper_lag,
            aes(x = demvoteshare_t, y = upper),
            size = 0.7, color = "red", linetype = "longdash") +
  geom_vline(xintercept = 0, color = "orange", size = 1.3, linetype = "dashed") +
  theme_minimal() +
  xlim(c(-0.1, 0.1))
```



compute a confidence interval for the SRDD estimate

```
model_data_lag <- data_lagged %>%
  mutate(D = ifelse(demvoteshare > 0.5, 1, 0)) %>%
  filter(between(demvoteshare, 0.48, 0.52))

mod_lag <- lm(dwnom_lag ~ demvoteshare_t + D + demvoteshare_t*D, data = model_data_lag)
se <- sqrt(vcovHC(mod_lag, type="HC")["D", "D"])
ci <- mod_lag$coefficients["D"] + 1.96*c(-1,1)*se

data.frame(
  est = mod_lag$coefficients["D"],
  lower = ci[1],
  upper = ci[2]
)
```

```
        est      lower      upper
D -0.069983 -0.1896439 0.04967791
```

compute the p-value (not significant at the $\alpha = 0.05$ level)

```
z <- (mod_lag$coefficients["D"])/se
p_val <- 2*pnorm(z, mean = 0, sd = 1)

p_val
```

```
        D 
0.2516732 
```

We don't really see a clear discontinuity in the scatter plot and indeed this is backed by the fact that our confidence interval for the SRDD contains zero and the fact that we failed to reject the null hypothesis that the local average treatment effect is zero. This result adds credibility to our findings in parts (b) and (c) since we see no clear discontinuity in the lagged dwnom scores (which is what we'd expect in a truly randomized experiment).

f

```
l <- seq(44.75, 49.75, by = 0.5)/100
u <- rev(seq(50.25, 55.25, by = 0.5)/100)

results <- data.frame()

for(i in 1:length(l)) {

  size <- u[i] - l[i]
  model_data <- data_lagged %>%
    drop_na() %>%
    mutate(D = ifelse(demvoteshare > 0.5, 1, 0)) %>%
    filter(between(demvoteshare, l[i], u[i]))

  mod_lag <- lm(dwnom_lag ~ demvoteshare_t + D + demvoteshare_t*D, data = model_data)
  se_lag <- sqrt(vcovHC(mod_lag, type="HC")["D", "D"])
  ci_lag <- mod_lag$coefficients["D"] + 1.96*c(-1,1)*se_lag

  mod <- lm(dwnom ~ demvoteshare_t + D + demvoteshare_t*D, data = model_data)
  se <- sqrt(vcovHC(mod, type="HC")["D", "D"])
  ci <- mod$coefficients["D"] + 1.96*c(-1,1)*se

  ret <- data.frame(
    est = c(mod_lag$coefficients["D"], mod$coefficients["D"]),
    lower = c(ci_lag[1], ci[1]),
```

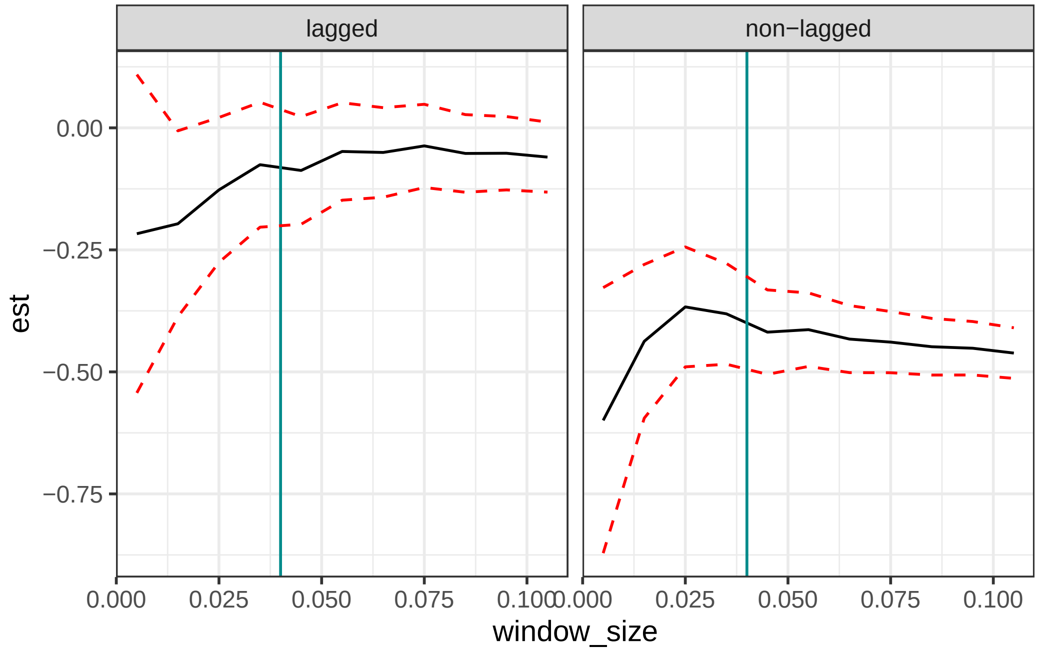
```r
    upper = c(ci_lag[2], ci[2]),
    type = c("lagged", "non-lagged"),
    window_size = c(size, size)
  )

  results <- rbind(results, ret)

}
```

After examining the sensitivity in this way I still feel pretty good about both conclusions that we reached. In the non-lagged case we do see an increase in the magnitude of the estimate as we decrease the window size but not a terribly drastic one, and it seems like for any of the window sizes that we checked we would have come to the same conclusion. Similarly we wee the same thing in the lagged version except we see that our confidence interval actually doesn't include zero in one of the smaller window sizes which is a little bit concerning. That being said, we generally would reach the same conclusion across the various window sizes.

```r
results %>%
  ggplot(aes(x = window_size, y = est)) +
  geom_line() +
  geom_vline(xintercept = 0.04, color = "cyan4") +
  geom_line(aes(x = window_size, y = upper), color = "red", linetype = "dashed") +
  geom_line(aes(x = window_size, y = lower), color = "red", linetype = "dashed") +
  facet_wrap(~type) +
  theme_bw()
```


lagged
non−lagged
est
0.00
−0.25
−0.50
−0.75
0.000
0.025
0.050
0.075
0.100
window_size