Indicators and Interactions

(Categorical Variables in Regression)

Understanding Political Numbers

March 25, 2019

While you were gone...


MIGHTY
CUTE
ROAR!


GREENBUSH BAKERY
Donuts
OPEN
NOW HIRING
GREENBU
BAKER

Research Questions

Only *one* y variable, multiple x variables

Selecting on the Dependent Variable

Selecting cases that meet some criteria, and using *only* those cases as evidence for the criteria

Instead, we need variation in both x and y (or we can't study the relationship)

Selecting on the Dependent Variable

Selecting cases that meet some criteria, and using *only* those cases as evidence for the criteria

Instead, we need variation in both x and y (or we can't study the relationship)

The New York Times

I've Interviewed 300 High Achievers About Their Morning Routines. Here's What I've Learned.

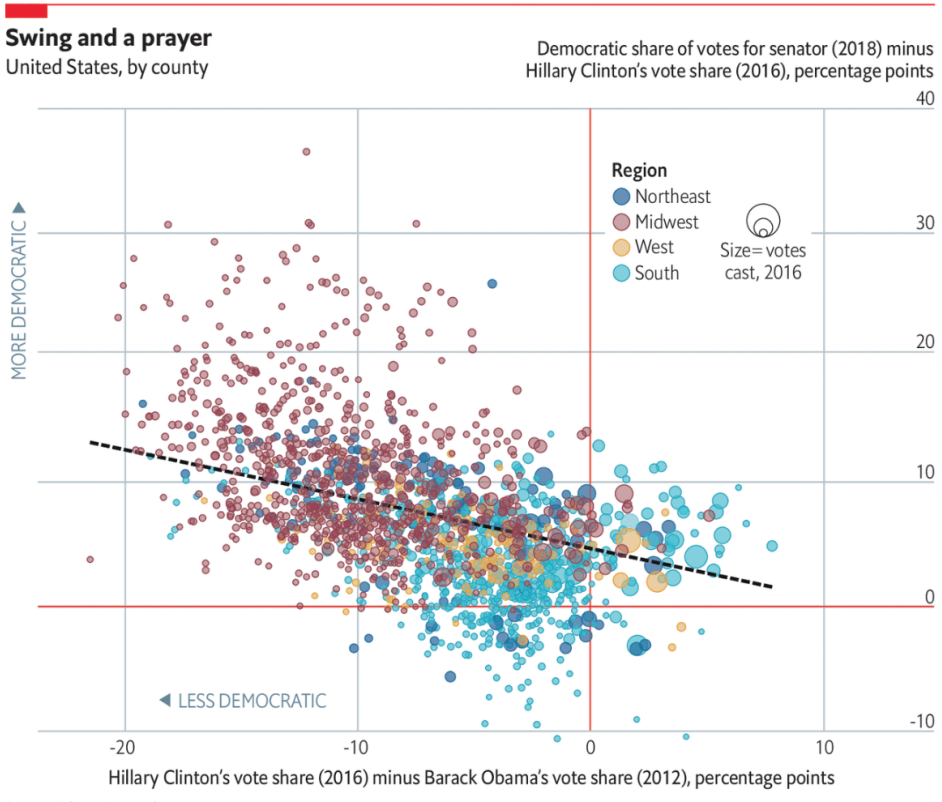
Your morning routine should suit your needs, but there are some habits everyone should try.

Ecological Inference Fallacy

Assuming that group-level patterns apply to individuals within the group

Obama-Trump voters turn back to Democrats

Senate Democrats did especially well where Donald Trump had gained the most ground



Confounders

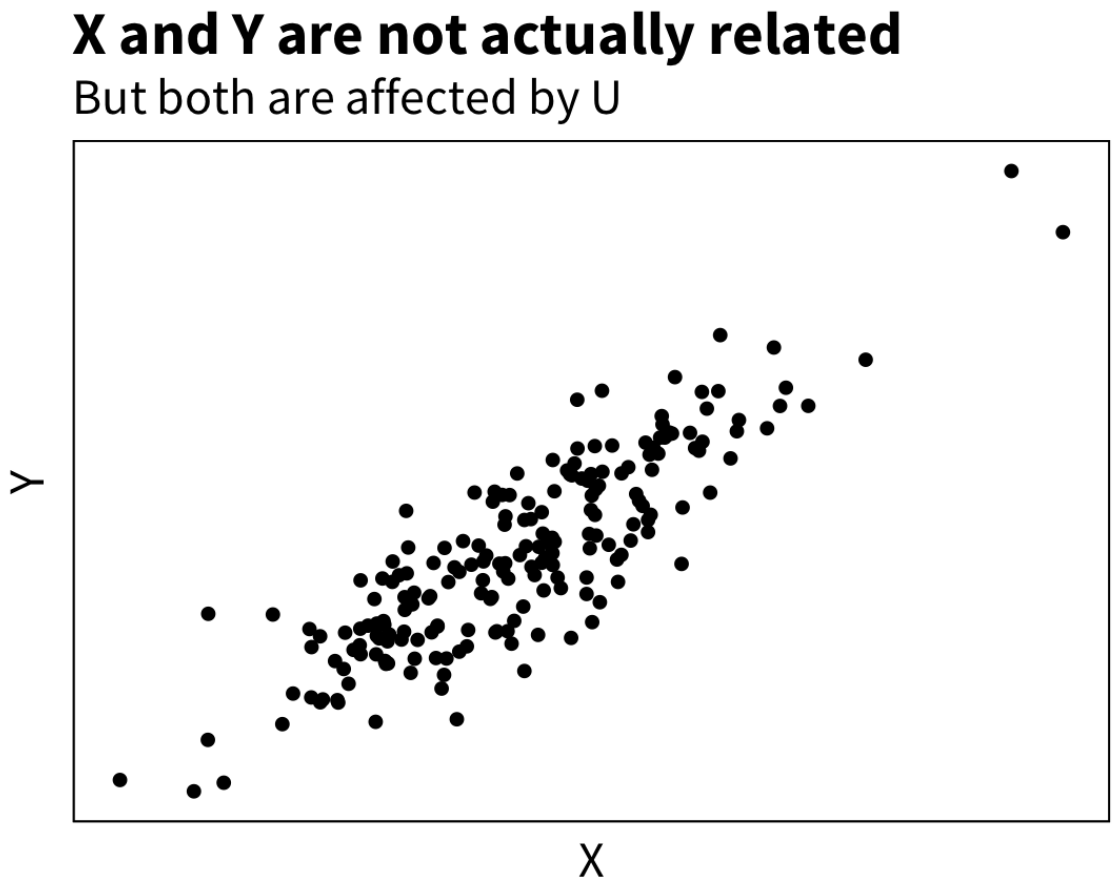
```
# X = f(U) and Y = f(U)
# but Y ≠ f(X)
confound_data <- confound_data %>%
  mutate(
    X = 2 + (3*U) + error_x,
    Y = 1 + (2*U) + error_y
  ) %>%
  select(-starts_with("error")) %>%
  print()
```

```
## # A tibble: 200 x 3
##          U       X       Y
##      <dbl>   <dbl>   <dbl>
##  1  0.624   5.13    3.51
##  2  0.998   3.50    3.24
##  3 -0.910  -0.391  -1.68
##  4  0.0374  1.74   -1.04
##  5  0.433   3.99    0.296
##  6 -1.32   -1.45   -1.27
##  7 -0.811   0.583  -2.68
##  8 -0.492   1.40   -0.802
##  9 -0.353  -0.0152 -0.0595
```

Confounders

```
# X = f(U) and Y = f(U)
# but Y ≠ f(X)
confound_data <- confound_data %>%
  mutate(
    X = 2 + (3*U) + error_x,
    Y = 1 + (2*U) + error_y
  ) %>%
  select(-starts_with("error")) %>%
  print()
```

```
## # A tibble: 200 x 3
##          U       X       Y
##      <dbl>   <dbl>   <dbl>
##  1  0.624   5.13    3.51
##  2  0.998   3.50    3.24
##  3 -0.910  -0.391  -1.68
##  4  0.0374  1.74   -1.04
##  5  0.433   3.99    0.296
##  6 -1.32   -1.45   -1.27
##  7 -0.811   0.583  -2.68
##  8 -0.492   1.40   -0.802
##  9 -0.353  -0.0152 -0.0595
```



Proof-read better

Indicators

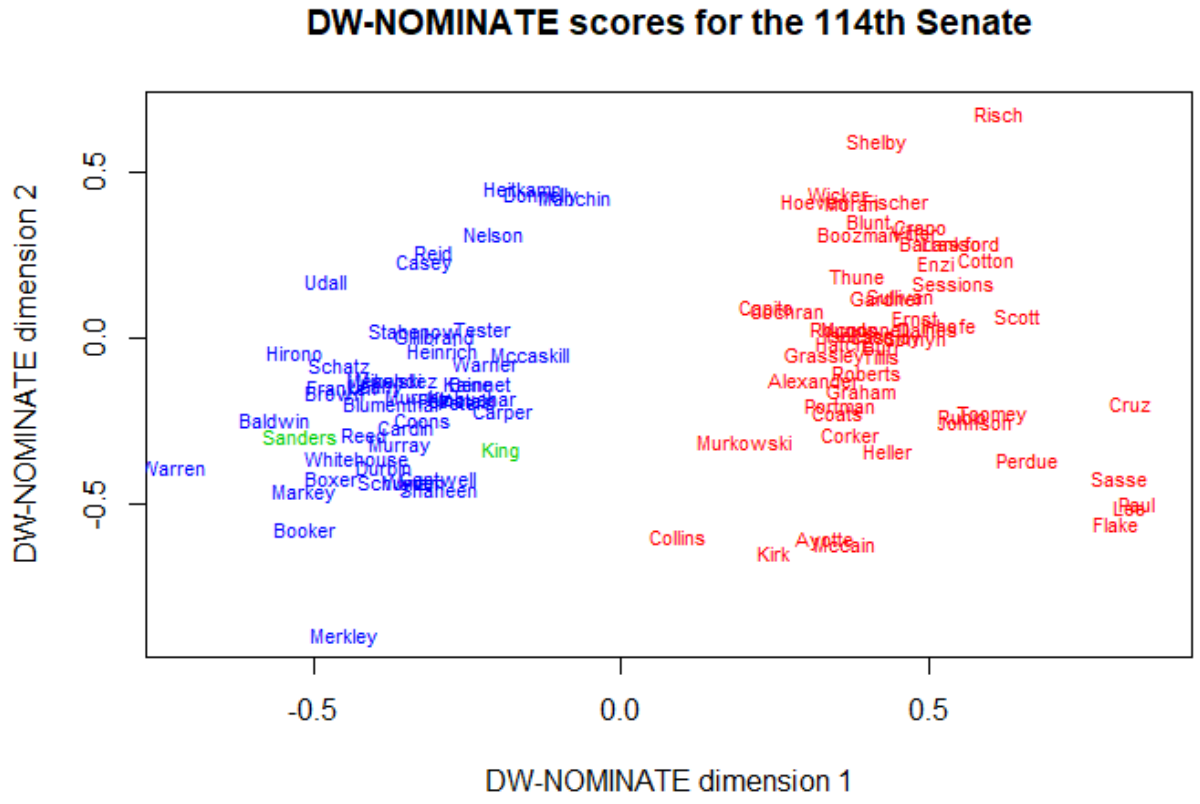
Multiple regression

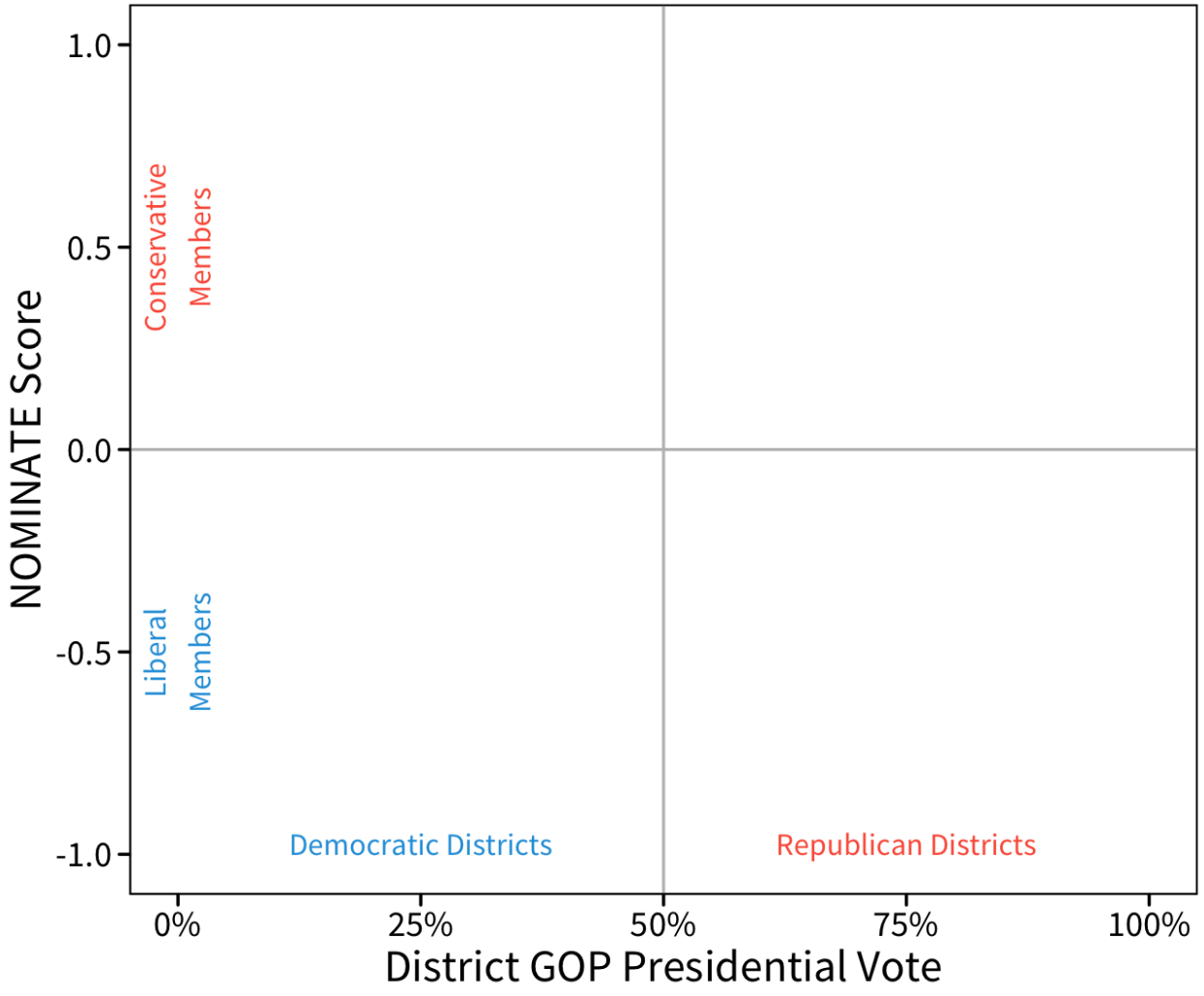
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \epsilon$$

DW-Nominate

Ideological scaling method from *roll call votes*

-1 (left/"liberal") to +1 (right/conservative)




1.0
0.5
0.0
-0.5
-1.0
NOMINATE Score
Conservative
Members
Liberal
Members
Democratic Districts
Republican Districts
0%
25%
50%
75%
100%
District GOP Presidential Vote

Elections data

```
library("tidyverse")
library("here")
library("broom")

# read House elections/ideology data
house <- read_csv(here("data", "house-ideology.csv")) %>%
  print()
```

```
## # A tibble: 450 x 5
##    state_abbrev district_code nominate_dim1 rep_pvote party
##    <chr>                <dbl>         <dbl>     <dbl> <chr>
##  1 AL                       3         0.338     0.669 Republican
##  2 AL                       7        -0.39      0.291 Democrat
##  3 AL                       2         0.367     0.663 Republican
##  4 AL                       5         0.607     0.674 Republican
##  5 AL                       1         0.543     0.651 Republican
##  6 AL                       6         0.773     0.731 Republican
##  7 AL                       4         0.362     0.822 Republican
##  8 AK                       1         0.28      0.584 Republican
##  9 AZ                       8         0.749     0.611 Republican
## 10 AZ                       3        -0.599     0.342 Democrat
## # … with 440 more rows
```

Nominate as f(presidential vote)

```
# estimate the linear model
# lm(y ~ x, data = dataset)
house_reg <- lm(nominate_dim1 ~ rep_pvote,
                data = house)

# intercept and slope estimates
tidy(house_reg)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -0.972    0.0338     -28.8 6.15e-104
## 2 rep_pvote      2.20     0.0653      33.7 5.49e-125
```

$$\hat{\text{Nom}} = -0.97 + 2.2\,(\text{Pres. Vote})$$

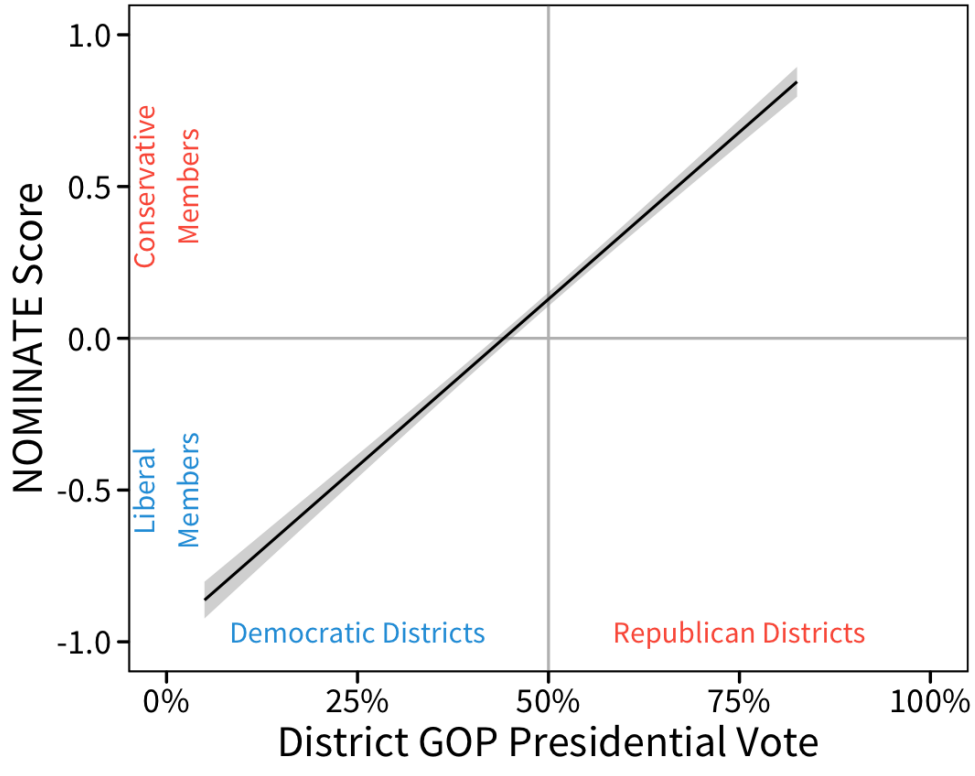
Nominate as f(presidential vote)

```
# estimate the linear model
# lm(y ~ x, data = dataset)
house_reg <- lm(nominate_dim1 ~ rep_pvote,
                data = house)

# intercept and slope estimates
tidy(house_reg)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -0.972    0.0338     -28.8 6.15e-104
## 2 rep_pvote      2.20     0.0653      33.7 5.49e-125
```

$$\hat{\text{Nom}} = -0.97 + 2.2\,(\text{Pres. Vote})$$



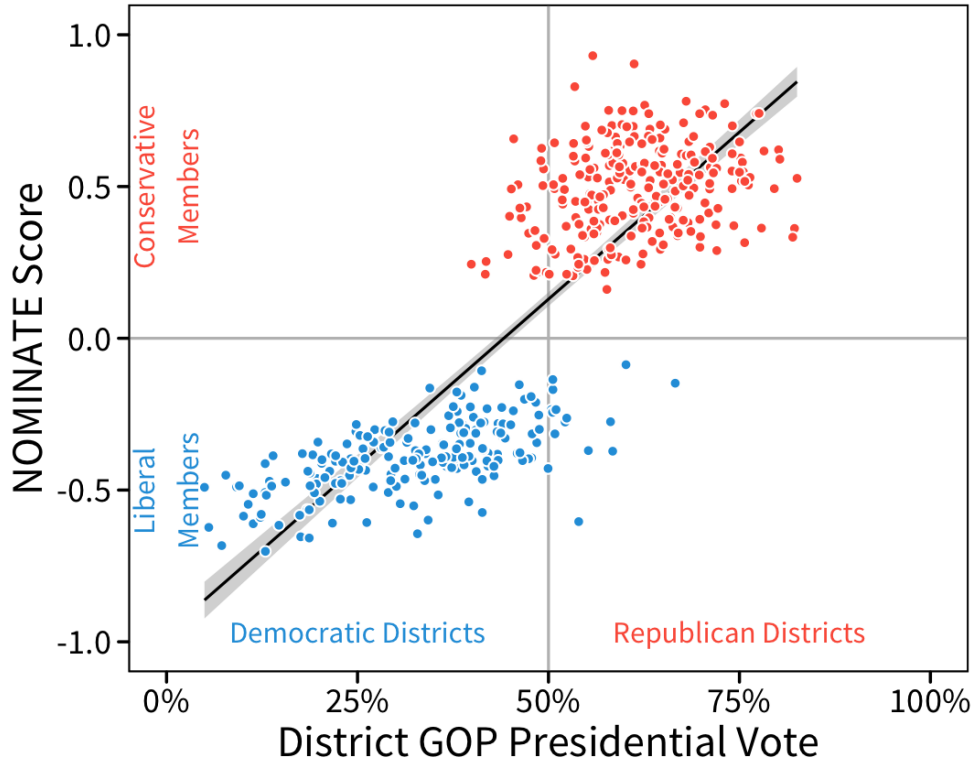
Nominate as f(presidential vote)

```
# estimate the linear model
# lm(y ~ x, data = dataset)
house_reg <- lm(nominate_dim1 ~ rep_pvote,
                data = house)

# intercept and slope estimates
tidy(house_reg)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -0.972    0.0338     -28.8 6.15e-104
## 2 rep_pvote      2.20     0.0653      33.7 5.49e-125
```

$$\hat{\text{Nom}} = -0.97 + 2.2\,(\text{Pres. Vote})$$



Control for party (using indicator/dummy variable)

Control for party (using indicator/dummy variable)

```
# new_variable = case_when(if condition ~ result if TRUE,
#                          if condition2 ~ result if TRUE)
house <- house %>%
  mutate(republican = case_when(party == "Republican" ~ 1,
                                party == "Democrat" ~ 0)) %>%
  print()
```

Control for party (using indicator/dummy variable)

```
# new_variable = case_when(if condition ~ result if TRUE,
#                          if condition2 ~ result if TRUE)
house <- house %>%
  mutate(republican = case_when(party == "Republican" ~ 1,
                                party == "Democrat" ~ 0)) %>%
  print()
```

```
## # A tibble: 450 x 6
##    state_abbrev district_code nominate_dim1 rep_pvote party      republican
##    <chr>                <dbl>         <dbl>     <dbl> <chr>           <dbl>
##  1 AL                       3         0.338     0.669 Republican          1
##  2 AL                       7        -0.39      0.291 Democrat            0
##  3 AL                       2         0.367     0.663 Republican          1
##  4 AL                       5         0.607     0.674 Republican          1
##  5 AL                       1         0.543     0.651 Republican          1
##  6 AL                       6         0.773     0.731 Republican          1
##  7 AL                       4         0.362     0.822 Republican          1
##  8 AK                       1         0.28      0.584 Republican          1
##  9 AZ                       8         0.749     0.611 Republican          1
## 10 AZ                       3        -0.599     0.342 Democrat            0
## # … with 440 more rows
```

Estimate new regression

```
# use `+` to add additional predictors
reg_party <- lm(nominate_dim1 ~ rep_pvote + republican,
                data = house)

tidy(reg_party)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -0.575    0.0204    -28.2  2.97e-101
## 2 rep_pvote      0.559    0.0564      9.92 4.45e- 21
## 3 republican     0.721    0.0200     36.0  4.38e-134
```

Estimate new regression

```
# use `+` to add additional predictors
reg_party <- lm(nominate_dim1 ~ rep_pvote + republican,
                data = house)

tidy(reg_party)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -0.575    0.0204    -28.2  2.97e-101
## 2 rep_pvote      0.559    0.0564      9.92 4.45e- 21
## 3 republican     0.721    0.0200     36.0  4.38e-134
```

Remember... coefficients are "rise over run"

What is the effect of "Republican"?

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

If the member is a Democrat?

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

If the member is a Democrat?

$$\hat{y} = -0.575 + 0.559v + 0.721\,(\mathbf{R = 0})$$

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

If the member is a Democrat?

$$\hat{y} = -0.575 + 0.559v + 0.721\,(\mathbf{R = 0})$$

$$\hat{y} = -0.575 + 0.559v$$

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

If the member is a **Republican**?

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

If the member is a Republican?

$$\hat{y} = -0.575 + 0.559v + 0.721\,(\mathbf{R = 1})$$

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

If the member is a Republican?

$$\hat{y} = -0.575 + 0.559v + 0.721\,(\mathbf{R = 1})$$

$$\hat{y} = -0.575 + 0.559v + 0.721$$

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

If the member is a Republican?

$$\hat{y} = -0.575 + 0.559v + 0.721\,(\mathbf{R = 1})$$

$$\hat{y} = -0.575 + 0.559v + 0.721$$

$$\hat{y} = (-0.575 + 0.721) + 0.559v$$

What is the effect of "Republican"?

$$\hat{y} = a + b_1\,(\text{Pres. Vote}) + b_2\,(\text{Republican})$$

$$\hat{y} = -0.575 + 0.559v + 0.721R$$

If the member is a Republican?

$$\hat{y} = -0.575 + 0.559v + 0.721\,(\mathbf{R = 1})$$

$$\hat{y} = -0.575 + 0.559v + 0.721$$

$$\hat{y} = (-0.575 + 0.721) + 0.559v$$

$$\hat{y} = 0.146 + 0.559v$$

Dummy variables *shift the intercept*

Dummy variables *shift the intercept*

Democratic line:

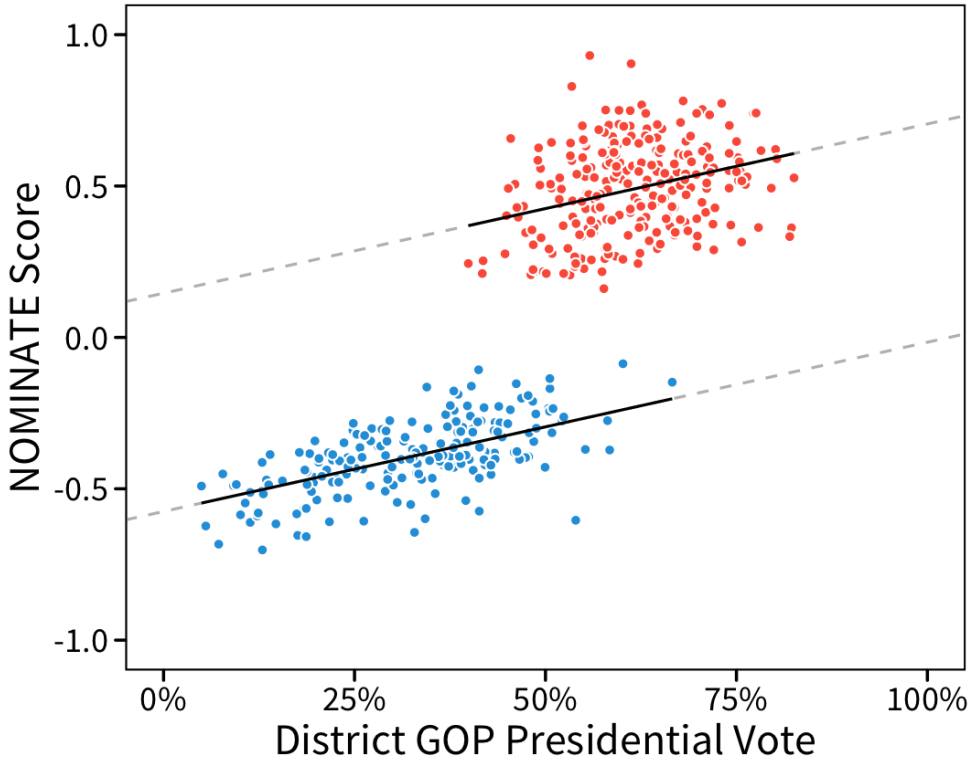
$$\hat{y} = -0.575 + 0.559v$$

Republican line:

$$\hat{y} = (-0.575 + 0.721) + 0.559v$$

Controlling for the voting behavior in the district, Republicans are 0.721 points more conservative than Democrats (on average)

Always leave one category "omitted"



The "omitted category"

Dummy variables are intercept shifts *relative to baseline*, where baseline = omitted category

The "omitted category"

Dummy variables are intercept shifts *relative to baseline*, where baseline = omitted category

```
# car data contain 4, 6, and 8 cylinder engines
mtcars <- mtcars %>%
  mutate(four_cyl = case_when(cyl == 4 ~ 1,
                              cyl != 4 ~ 0),
         six_cyl = case_when(cyl == 6 ~ 1,
                             cyl != 6 ~ 0),
         eight_cyl = case_when(cyl == 8 ~ 1,
                               cyl != 8 ~ 0))

lm(mpg ~ six_cyl + eight_cyl, data = mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ six_cyl + eight_cyl, data = mtcars)
##
## Coefficients:
## (Intercept)      six_cyl    eight_cyl
##      26.664       -6.921      -11.564
```

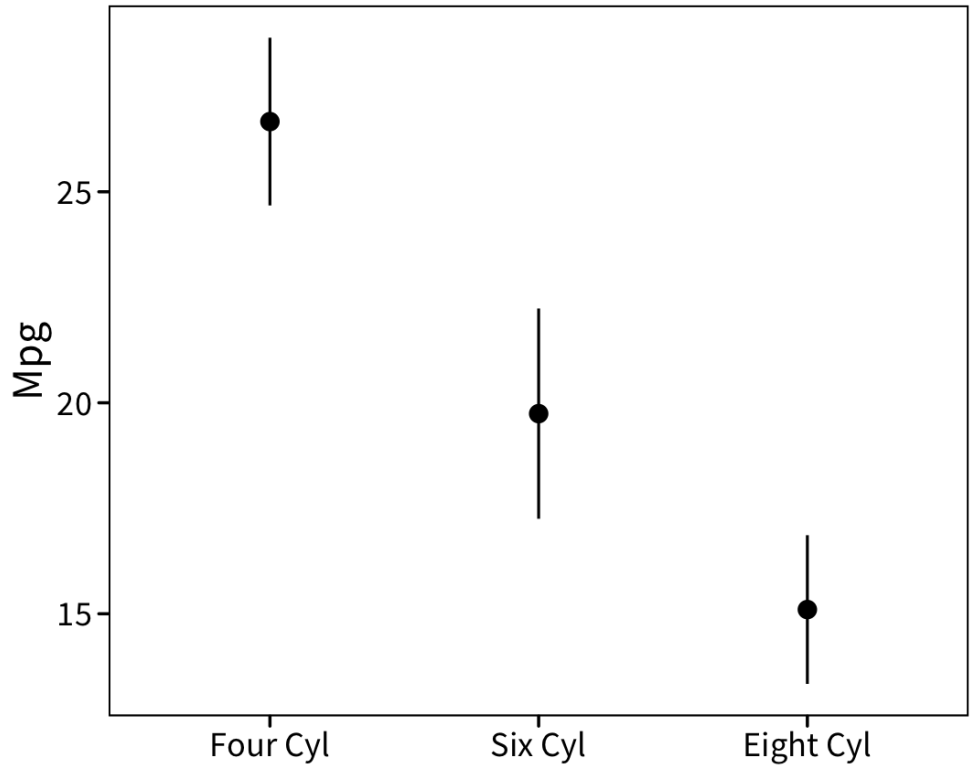
The "omitted category"

Dummy variables are intercept shifts *relative to baseline*, where baseline = omitted category

```
# car data contain 4, 6, and 8 cylinder engines
mtcars <- mtcars %>%
  mutate(four_cyl = case_when(cyl == 4 ~ 1,
                              cyl != 4 ~ 0),
         six_cyl = case_when(cyl == 6 ~ 1,
                             cyl != 6 ~ 0),
         eight_cyl = case_when(cyl == 8 ~ 1,
                               cyl != 8 ~ 0))

lm(mpg ~ six_cyl + eight_cyl, data = mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ six_cyl + eight_cyl, data = mtcars)
##
## Coefficients:
## (Intercept)      six_cyl    eight_cyl
##      26.664       -6.921      -11.564
```



When to use indicators?

Categorical data: What's the average effect of *group membership*? (regime type, gender, race/ethnicity, UNSC member)

Ordinal data: not sure if linear (education/income brackets)

Policy change: is policy enacted (`strict_photo_id == 1` or `strict_photo_id == 0`)

Small number of time periods: before and after (`year_2016 == 1` or `year_2016 == 0`)

Dummies for *varying slopes*

Maybe one party is *more responsive* to district voting?

How? Use an Interaction term

Dummies for *varying slopes*

Maybe one party is *more responsive* to district voting?

How? Use an Interaction term

$$\text{Nom} = \alpha + \beta_1(\text{PresVote}) + \beta_2(\text{Republican}) + \beta_3(\text{PresVote} \times \text{Republican}) + \varepsilon$$

Dummies for *varying slopes*

Maybe one party is *more responsive* to district voting?

How? Use an Interaction term

$$\text{Nom} = \alpha + \beta_1(\text{PresVote}) + \beta_2(\text{Republican}) + \beta_3(\text{PresVote} \times \text{Republican}) + \varepsilon$$

```
house <- house %>%
  mutate(interact_vote_rep = rep_pvote * republican)

interact_reg <- lm(nominate_dim1 ~ rep_pvote + republican + interact_vote_rep,
                   data = house)
```

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

```
tidy(interact_reg)
```

```
## # A tibble: 4 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         -0.589    0.0249    -23.7  6.08e-81
## 2 rep_pvote            0.604    0.0713      8.47 3.72e-16
## 3 republican           0.782    0.0622     12.6  3.46e-31
## 4 interact_vote_rep   -0.119    0.116      -1.02 3.06e- 1
```

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Democrat**?

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Democrat**?

$$\hat{y} = -0.59 + 0.6v + 0.78(0) + -0.12(v \times 0)$$

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Democrat**?

$$\hat{y} = -0.59 + 0.6v + 0.78(0) + -0.12(v \times 0)$$

$$\hat{y} = -0.59 + 0.6v + 0.78(0) + -0.12(0)$$

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Democrat**?

$$\hat{y} = -0.59 + 0.6v + 0.78(0) + -0.12(v \times 0)$$

$$\hat{y} = -0.59 + 0.6v + 0.78(0) + -0.12(0)$$

$$\hat{y} = -0.59 + 0.6v$$

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Republican**?

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Republican**?

$$\hat{y} = -0.59 + 0.6v + 0.78(1) + -0.12(v \times 1)$$

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Republican**?

$$\hat{y} = -0.59 + 0.6v + 0.78(1) + -0.12(v \times 1)$$

$$\hat{y} = -0.59 + 0.6v + 0.78 + -0.12v$$

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Republican**?

$$\hat{y} = -0.59 + 0.6v + 0.78(1) + -0.12(v \times 1)$$

$$\hat{y} = -0.59 + 0.6v + 0.78 + -0.12v$$

Simplify again!

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Republican**?

$$\hat{y} = -0.59 + 0.6v + 0.78(1) + -0.12(v \times 1)$$

$$\hat{y} = -0.59 + 0.6v + 0.78 + -0.12v$$

Simplify again!

$$\hat{y} = (-0.59 + 0.78) + (0.6 + -0.12)v$$

What is the effect of "Republican"?

$$\hat{\text{Nom}} = a + b_1(\text{PresVote}) + b_2(\text{Republican}) + b_3(\text{PresVote} \times \text{Republican})$$

$$\hat{y} = -0.59 + 0.6v + 0.78R + -0.12(v \times R)$$

If the member is a **Republican**?

$$\hat{y} = -0.59 + 0.6v + 0.78(1) + -0.12(v \times 1)$$

$$\hat{y} = -0.59 + 0.6v + 0.78 + -0.12v$$

Simplify again!

$$\hat{y} = (-0.59 + 0.78) + (0.6 + -0.12)v$$

$$\hat{y} = 0.19 + 0.48v$$

Dummy variables *shift the intercept*, and interactions *shift the slope*

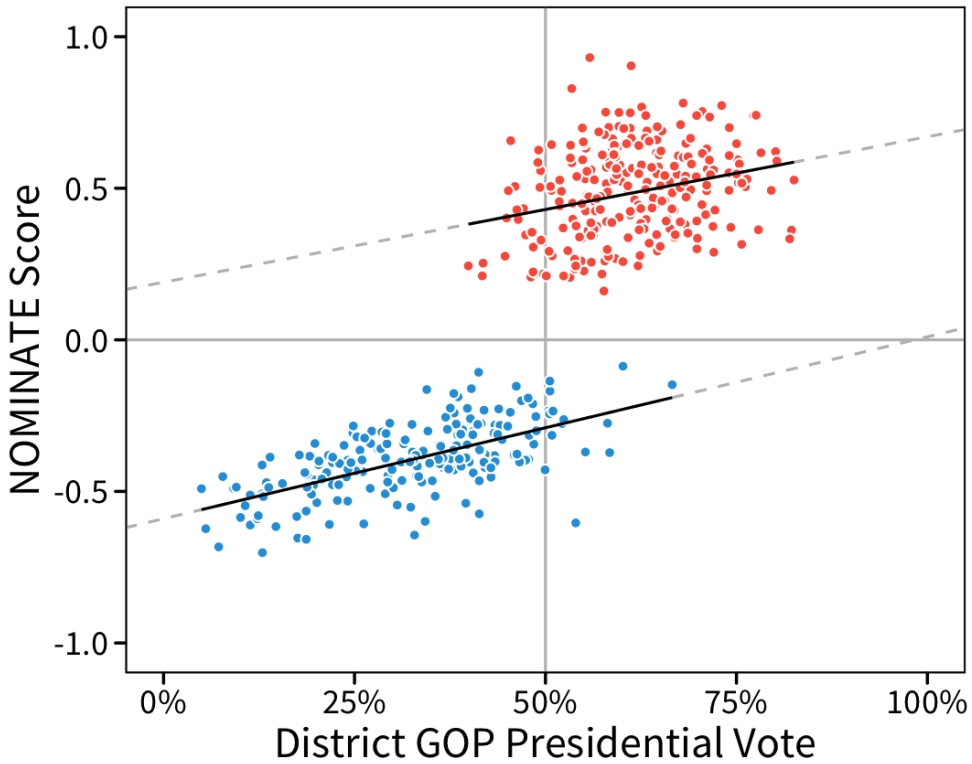
Democratic line:

$$\hat{y} = -0.59 + 0.6v$$

Republican line:

$$\hat{y} = 0.19 + 0.48v$$

Republicans are *slightly* more responsive, but not statistically significant $p = 0.31$



Looking ahead

For today:

- Assigned reading (GOTV)
- Download lecture, code, and data (later today)

For Wednesday:

- Assigned videos on LOGARITHMS
- Essay 2 is incoming

Next week:

- Data due Monday
- Ask us for help