R Coding Demonstration Week 12: Uncertainty in Regression (Tidy)

Matthew Blackwell

Gov 51 (Harvard)

Introduction

- Does NIMBYism hurt efforts to expand green energy projects?

    - NIMBY: "not in my backyard"

- Leah Stokes paper on efforts to expand wind power in Ontario, Canada.

    - Liberal Party passed Green Energy Act making wind power easier to build.
    - Did voters where turbines were built punish the Liberal Party?

- For simplicity, focus on a sample of 500 rural precincts.

Data

```
wind <- read.csv("data/stokes_electoral_2015.csv")
```

Name	Description
master_id	Precinct ID number
year	Election year
prop	Binary variable indicating whether there was a proposed turbine in that precinct in that year
perc_lib	Votes cast for Liberal Party divided by the number of voters who cast ballots in precinct

Question 1

First, let's load the data. What years are included? How many precincts are included? How many year-precincts are included?

Answer 1

```r
library(tidyverse)
table(wind$year)
```

```
##
## 2003 2007 2011
##  500  500  500
```

```r
length(unique(wind$master_id))
```

```
## [1] 500
```

```r
nrow(wind)
```
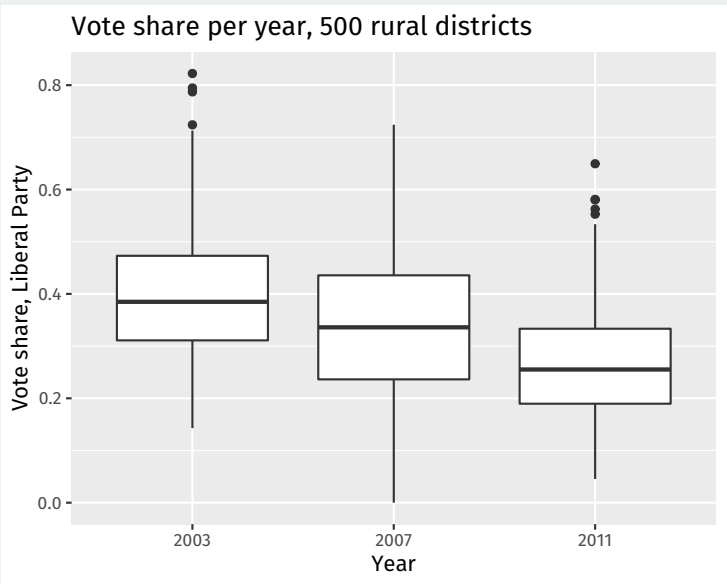
```
## [1] 1500
```

Make a boxplot that shows the distribution of vote share for the Liberal Party in each year. What do you conclude from this plot?

Answer 2

```
ggplot(wind, aes(x = as.factor(year), y = perc_lib)) +
  geom_boxplot() +
  labs(
    x = "Year",
    y = "Vote share, Liberal Party",
    title = "Vote share per year, 500 rural districts"
  )
```

Vote share per year, 500 rural districts

Make a boxplot that shows the distribution of vote share for the Liberal Party in precincts that had a proposed wind turbine and those that did not. What do you conclude from this plot?

Answer 3

```r
wind %>%
  mutate(
    prop_factor = ifelse(prop == 1, "Proposed", "Not Proposed")
  ) %>%
ggplot(aes(x = prop_factor, y = perc_lib)) +
  geom_boxplot() +
  labs(
    x = "Wind Turbine",
    y = "Vote share, Liberal Party",
    title = "Vote share per year, 500 rural districts"
  )
```

Vote share per year, 500 rural districts
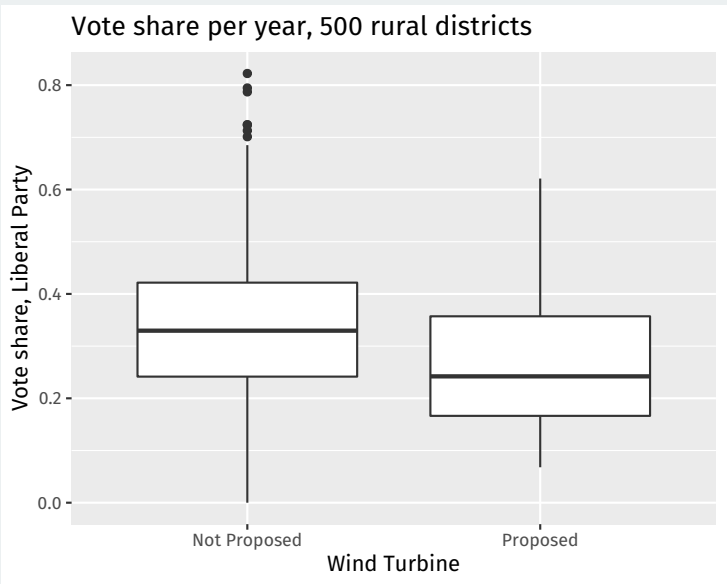
Run a regression of vote share for the Liberal Party on the wind turbine variable. Interpret the coefficient on prop and use summary() to determine if the estimated coefficient is statistically significant at the 0.05 level. What does statistically significant mean in this context?

Answer 4

```
fit1 <- lm(perc_lib ~ prop, data = wind)
summary(fit1)
```

```
##
## Call:
## lm(formula = perc_lib ~ prop, data = wind)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3366 -0.0952 -0.0076  0.0851  0.4858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.33655    0.00332  101.41  < 2e-16 ***
## prop        -0.06818    0.01960   -3.48  0.00052 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.127 on 1498 degrees of freedom
## Multiple R-squared:  0.00801,    Adjusted R-squared:  0.00735
## F-statistic: 12.1 on 1 and 1498 DF,  p-value: 0.000519
```

Add year as a factor to the previous regression and interpret the effect of prop. Does this change the magnitude of the effect?

Answer 5

```
fit2 <- lm(perc_lib ~ prop + factor(year),
           data = wind)
summary(fit2)
```

```
##
## Call:
## lm(formula = perc_lib ~ prop + factor(year), data = wind)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3391 -0.0894 -0.0096  0.0794  0.4225
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.39986    0.00513   77.96   <2e-16 ***
## prop              -0.01653    0.01798   -0.92     0.36
## factor(year)2007  -0.06077    0.00726   -8.36   <2e-16 ***
## factor(year)2011  -0.13359    0.00734  -18.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.115 on 1496 degrees of freedom
## Multiple R-squared:  0.188,  Adjusted R-squared:  0.186
## F-statistic:  115 on 3 and 1496 DF,  p-value: <2e-16
```

Use `modelsummary::modelsummary()` to create a regression table with the two regressions in columns to nicely present the results.

Answer 6

```
mods <- list(
  "No Controls" = fit1,
  "Election Year FEs" = fit2
)
coefs <- c(
  "prop" = "Proposed Turbine",
  "factor(year)2007" = "Election Year 2007",
  "factor(year)2011" = "Election Year 2011",
  "(Intercept)" = "Constant"
)
modelsummary::modelsummary(mods, coef_map = coefs,
                           gof_omit = "AIC|BIC|Log.Lik.|F")
```

Answer 6 (table)

	No Controls	Election Year FEs
Proposed Turbine	-0.068	-0.017
	(0.020)	(0.018)
Election Year 2007		-0.061
		(0.007)
Election Year 2011		-0.134
		(0.007)
Constant	0.337	0.400
	(0.003)	(0.005)
Num.Obs.	1500	1500
R2	0.008	0.188
R2 Adj.	0.007	0.186

Question 7

Run a final model that includes a fixed effect for year and precinct (that is, add a master_id as a factor to the last model). Create a new modelsummary table with all three models. How does the effect of proposed turbines change across the models? How does the model fit change?

Answer 7

```
fit3 <- lm(perc_lib ~ prop + factor(master_id) + factor(year),
           data = wind)
mods <- list("No Controls" = fit1,
             "Election Year" = fit2,
             "Election Year + Precinct" = fit3)

## using the same coef_map as before will omit any of the (many!)
## precinct fixed effects parameters
modelsummary::modelsummary(mods, coef_map = coefs,
                           gof_omit = "AIC|BIC|Log.Lik.|F")
```

Answer 7 (table)

	No Controls	Election Year	Election Year + Precinct
Proposed Turbine	-0.068	-0.017	-0.064
	(0.020)	(0.018)	(0.018)
Election Year 2007		-0.061	-0.060
		(0.007)	(0.005)
Election Year 2011		-0.134	-0.131
		(0.007)	(0.005)
Constant	0.337	0.400	0.365
	(0.003)	(0.005)	(0.046)
Num.Obs.	1500	1500	1500
R2	0.008	0.188	0.741
R2 Adj.	0.007	0.186	0.610