Welcome!

Nameplates please. And technology encouraged today!

All TF materials are available at github.com/nolankav/api-203.

If you want to follow along, download the dataset here:

In R: `df <- read.csv("http://tinyurl.com/api-203-tf-3")`

Regression discontinuity

API 203: TF Session 3

R

Nolan M. Kavanagh
March 22, 2024

Goals for today

1. **Review the principles of regression discontinuity (RD).**
2. **Learn how to run basic RD models in R.**
3. **Practice interpreting basic RD models.**
4. **Preview advanced methods in RD (optional!).**

We'll treat this session like a workshop with an interactive example.

Overview of our sample data

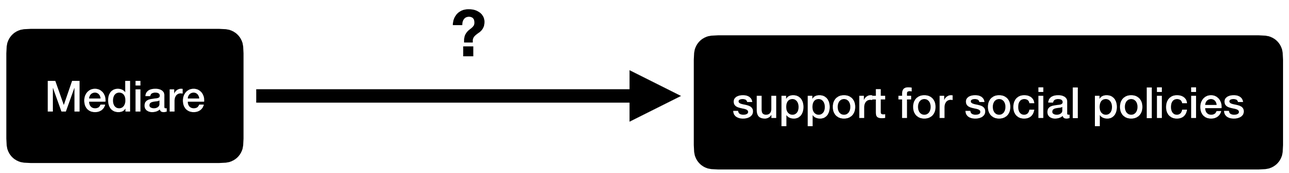
Dataset of 17,500 U.S. adults surveyed in 2019

`casein`	**Respondent identifier**	*Cooperative Election Study*
`commonweight`	**Post-stratification survey weight**	*Cooperative Election Study*
`age`	**Respondent age (in years)**	*Cooperative Election Study*
`age_center`	**Respondent age, recentered at 65**	*Cooperative Election Study*
`age_over65`	**Indicator for being 65 years or older (1) or not (0)**	*Cooperative Election Study*
`gender`	**Respondent gender**	*Cooperative Election Study*
`race_eth`	**Respondent race/ethnicity**	*Cooperative Election Study*
`education`	**Respondent educational attainment**	*Cooperative Election Study*
`marital`	**Respondent marital status**	*Cooperative Election Study*
`public_ins`	**Respondent has public health insurance (1) or not (0)**	
`public_option`	**Supports a public option in Medicare (1) or not (0)**	*Cooperative Election Study*

Finally, Nolan's time to shine.

What are the political effects of social policies?

For example, when people get Medicare, the health insurance program for people over 65, how does it change their political preferences?

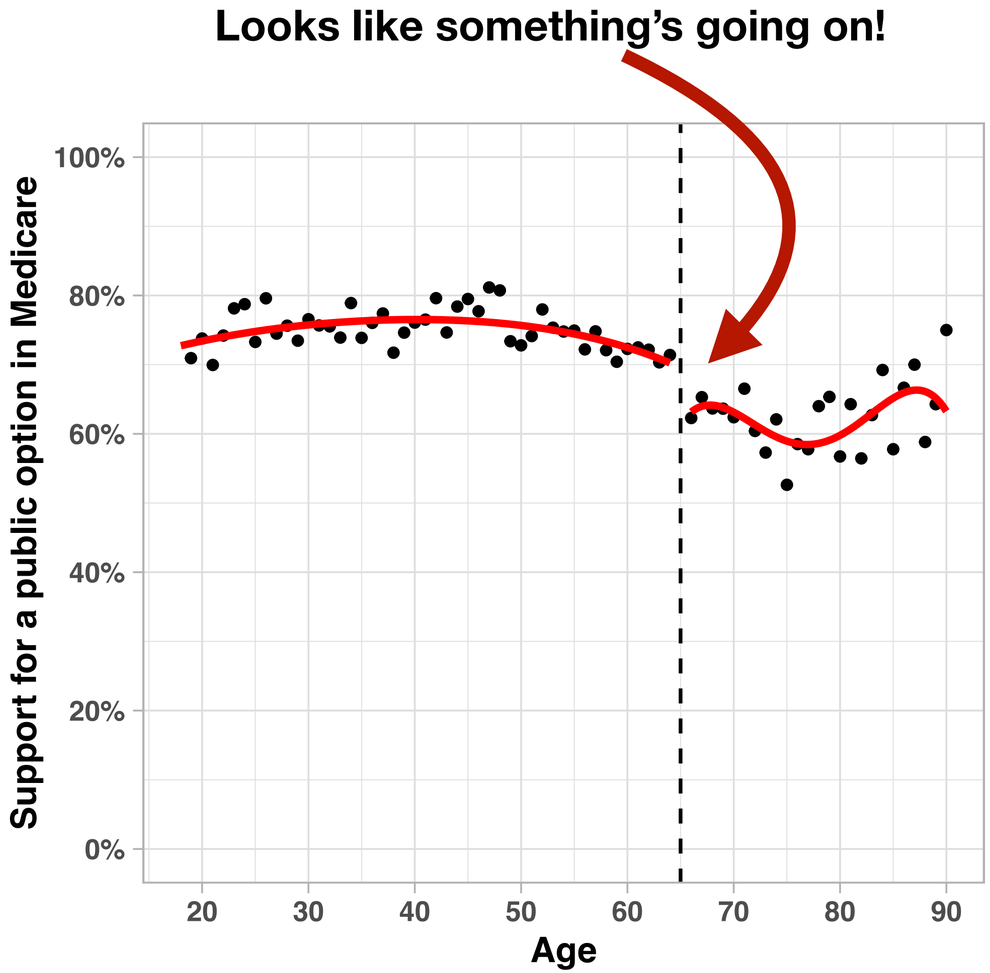


We'll look at support for expanding the Medicare program to more people.

Why yes, this was an API 202 exam question!

Let's start with the graph.

```
# Generate RDD plot
plot_rd_1 <- ggplot() +

  # Threshold line
  geom_vline(xintercept=THRESHOLD, linetype="dashed") +

  # Binned scatterplots for each side of 65
  stat_binmean(n=1000, data=subset(df, age < 65),
               aes(x=age, y=public_option), size=1) +
  stat_binmean(n=1000, data=subset(df, age > 65 & age <= 90),
               aes(x=age, y=public_option), size=1) +

  # Global polynomial fit lines
  # Note: Limit to ≤90 years given sparse data over 90
  geom_smooth(data=subset(df, age < 65),
              aes(x=age, y=public_option), formula=y ~ poly(x, 4, raw=TRUE),
              method = "lm", se = F, color="red") +
  geom_smooth(data=subset(df, age > 65 & age <= 90),
              aes(x=age, y=public_option), formula=y ~ poly(x, 4, raw=TRUE),
              method = "lm", se = F, color="red") +

  # Cosmetic modifications
  xlab("Age") + ylab("Support for a public option in Medicare") +
  coord_cartesian(xlim=c(18,90), ylim=c(0,1)) +
  theme_light() +
  theme(text = element_text(size = 10, face = "bold")) +
  scale_x_continuous(breaks = seq(0, 100, 10)) +
  scale_y_continuous(breaks = seq(0, 1, 0.2),
                     labels = function(x) paste0(x*100,"%"))
```



Let's run the naive regression.

```
# Load package
library(fixest) # Modeling tools

# OLS regression without controls
model_1 <- feols(public_option ~ age_over65 | 0, data=df, vcov="HC1")
summary(model_1)
```

```
OLS estimation, Dep. Var.: public_option
Observations: 17,540
Standard-errors: Heteroskedasticity-robust
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)  0.746784   0.003646 204.8109 < 2.2e-16 ***
age_over65  -0.127176   0.009187 -13.8429 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.444864   Adj. R2: 0.012315
```

U.S. adults over 65 years are 12.7 pp less supportive of a public option in Medicare than adults under 65.

Note: Both variables are coded 0–1.

Let’s add some controls.

	Model 1	Model 2
Being over age 65	**–0.127*** **	**–0.107*** **
	(0.009)	(0.010)
Age (centered)		
Being over 65 * age (centered)		
Num.Obs.	17,540	17,540
Demographic controls	No	Yes
Standard errors	Robust	Robust
R2 Adj.	0.012	0.198

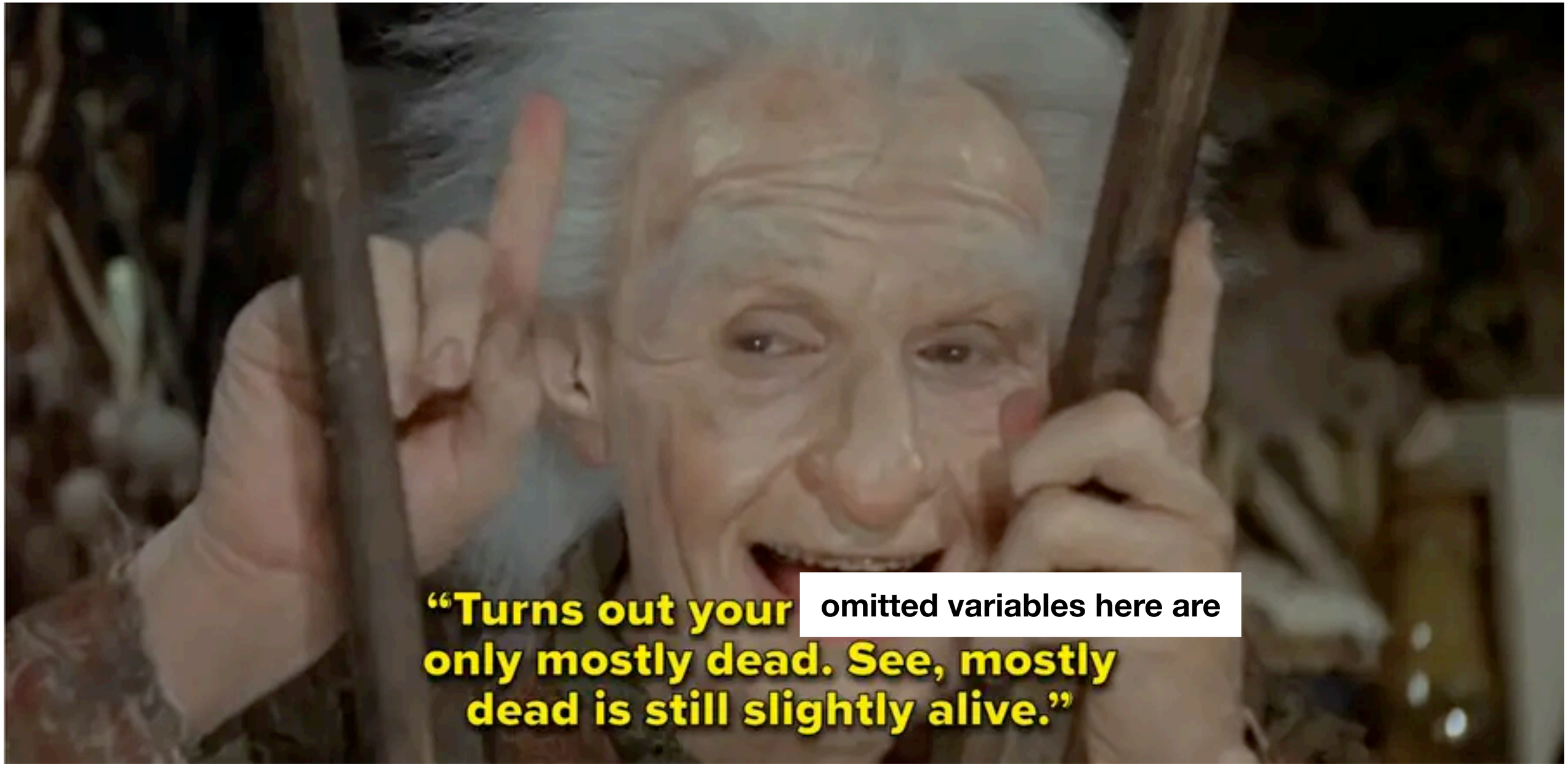
*p<0.05, **p<0.01, ***p<0.001

After demographic controls, being over age 65 is associated with 10.7 pp less support for a public option in Medicare.

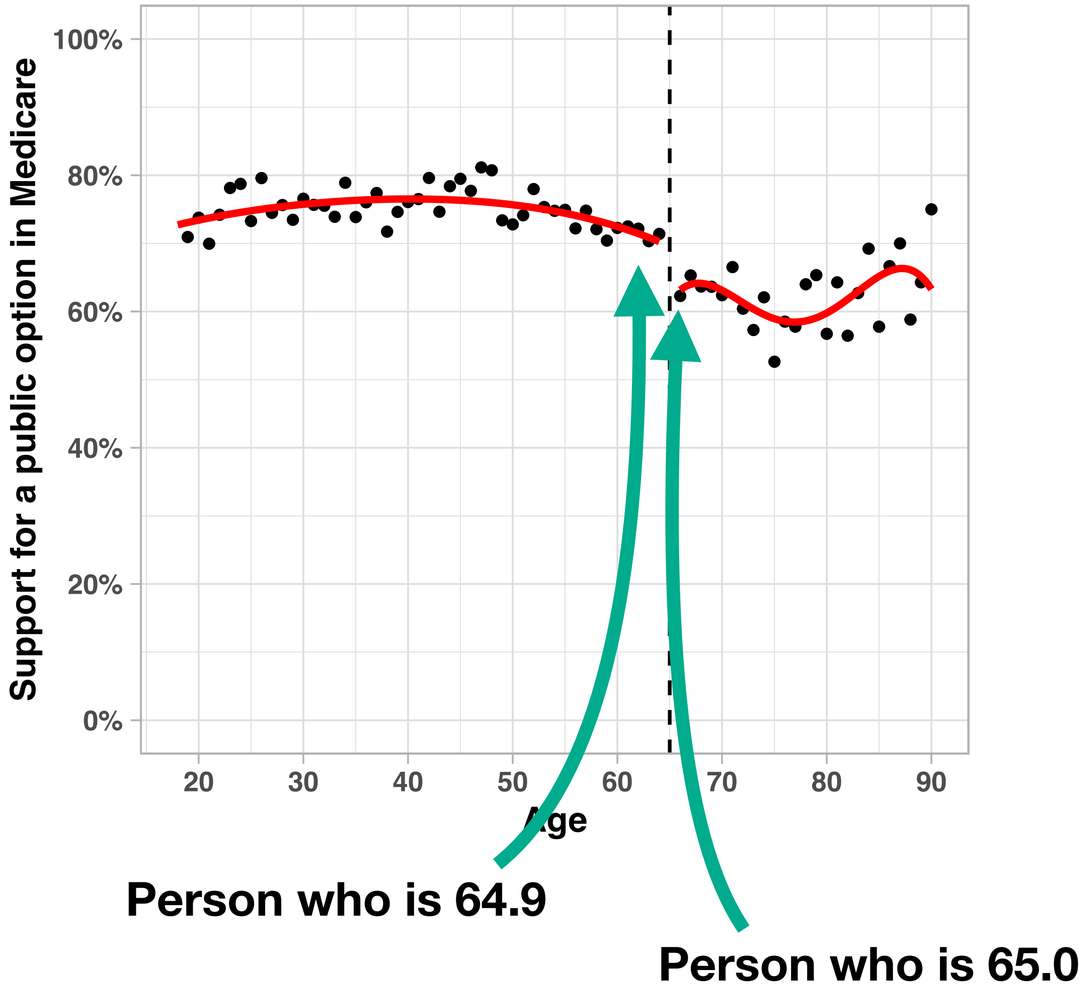
It’s significant at the 5% level.

Here, demographic controls are gender, race/ethnicity, educational attainment, and marital status.

But are there more omitted variables?


omitted variables here are
“Turns out your
only mostly dead. See, mostly
dead is still slightly alive.”

Enter: Regression discontinuity



The fundamental assumption of RD designs is that the people who are 64.9 years old are the same as the people who are 65.0 years old, except in one key way: Medicare eligibility.

As such, the threshold is as-if random, and all omitted variables are no more!


Hello. My name is
regression discontinuity
omitted variables
You killed my
basic regression
Prepare to die.

How do we specify an RD model?

$$(support)_i = \beta_0 + \beta_1(age\_over65)_i + \beta_2(age\_center)_i + \beta_3(age\_over65 \times age\_center)_i + u_i$$

outcome (Y)

indicator for threshold
(i.e. 1 if over, 0 if under)

running variable
(must be continuous; can recenter at the threshold for ease of interpretation)

interaction of indicator and running variable
(allows slope to vary at threshold)

The estimate we're most interested in is usually given by the indicator. Here, that's β_1.

Let's do it!

```
# Regression discontinuity: Manual, linear
model_3 <- feols(public_option ~ age_over65 + age_center +
                   age_over65*age_center | 0, data=df, vcov="HC1")
summary(model_3)

OLS estimation, Dep. Var.: public_option
Observations: 17,540
Standard-errors: Heteroskedasticity-robust
                       Estimate Std. Error    t value   Pr(>|t|)
(Intercept)            0.731069   0.006978 104.770065  < 2.2e-16 ***
age_over65            -0.106933   0.016128  -6.630252 3.4489e-11 ***
age_center            -0.000726   0.000271  -2.680806 7.3514e-03 **
age_over65:age_center  0.000165   0.001495   0.110581 9.1195e-01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.444773   Adj. R2: 0.01261
```

Let's do it!

```
# Regression discontinuity: Manual, linear
model_3 <- feols(public_option ~ age_over65 + age_center +
                   age_over65*age_center | 0, data=df, vcov="HC1")
summary(model_3)

OLS estimation, Dep. Var.: public_option
Observations: 17,540
Standard-errors: Heteroskedasticity-robust
                        Estimate Std. Error    t value   Pr(>|t|)
(Intercept)             0.731069   0.006978 104.770065  < 2.2e-16 ***
age_over65             -0.106933   0.016128  -6.630252 3.4489e-11 ***
age_center             -0.000726   0.000271  -2.680806 7.3514e-03 **
age_over65:age_center   0.000165   0.001495   0.110581 9.1195e-01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.444773   Adj. R2: 0.01261
```

It appears that becoming eligible for Medicare leads to a 10.7 pp decline in support for a public option in the program!

Looks pretty similar!

	Model 1	Model 2	Model 3
Being over age 65	**–0.127*** **	**–0.107*** **	**–0.107*** **
	(0.009)	(0.010)	(0.016)
Age (centered)			**–0.001** **
			(0.001)
Being over 65 * age (centered)			**0.000**
			(0.001)
Num.Obs.	17,540	17,540	17,540
Demographic controls	No	Yes	No
Standard errors	Robust	Robust	Robust
R2 Adj.	0.012	0.198	0.013

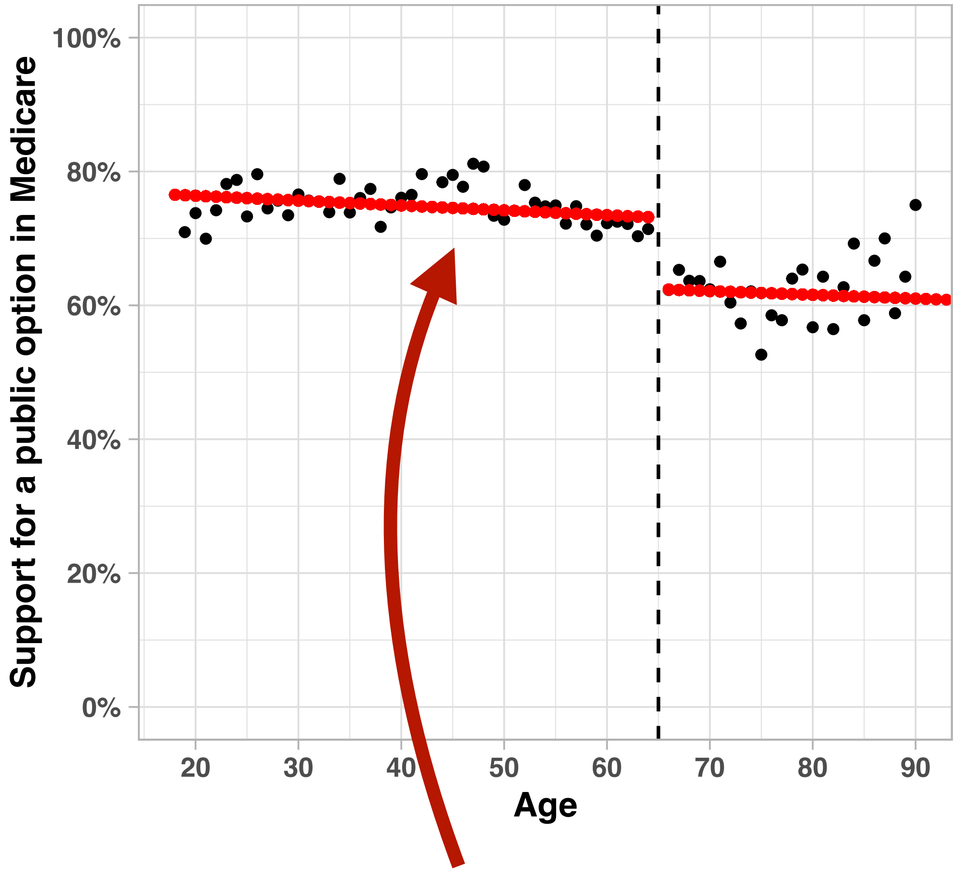
*p<0.05, **p<0.01, ***p<0.001

But wait. Are we done?

What might we be missing?

Hmm, show me the predicted values.

```
# Save predictions
df <- mutate(df, predictions = model_3$fitted.values)

# Plot predictions against true values
plot_rd_2 <- ggplot() +

  # Threshold line
  geom_vline(xintercept=THRESHOLD, linetype="dashed") +

  # Binned scatterplots for each side of 65
  stat_binmean(n=1000, data=subset(df, age < 65),
               aes(x=age, y=public_option), size=1) +
  stat_binmean(n=1000, data=subset(df, age > 65 & age <= 90),
               aes(x=age, y=public_option), size=1) +

  # Predicted support
  geom_point(data=df, aes(x=age, y=predictions), color="red", size=1)

  # Cosmetic modifications
  xlab("Age") + ylab("Support for a public option in Medicare") +
  coord_cartesian(xlim=c(18,90), ylim=c(0,1)) +
  theme_light() +
  theme(text = element_text(size = 10, face = "bold")) +
  scale_x_continuous(breaks = seq(0, 100, 10)) +
  scale_y_continuous(breaks = seq(0, 1, 0.2),
                     labels = function(x) paste0(x*100,"%"))
```



It doesn't look like we've quite gotten the functional form (shape) right.

Let's add quadratic terms.

outcome (Y)

indicator for threshold
(i.e. 1 if over, 0 if under)

running variable

running variable squared

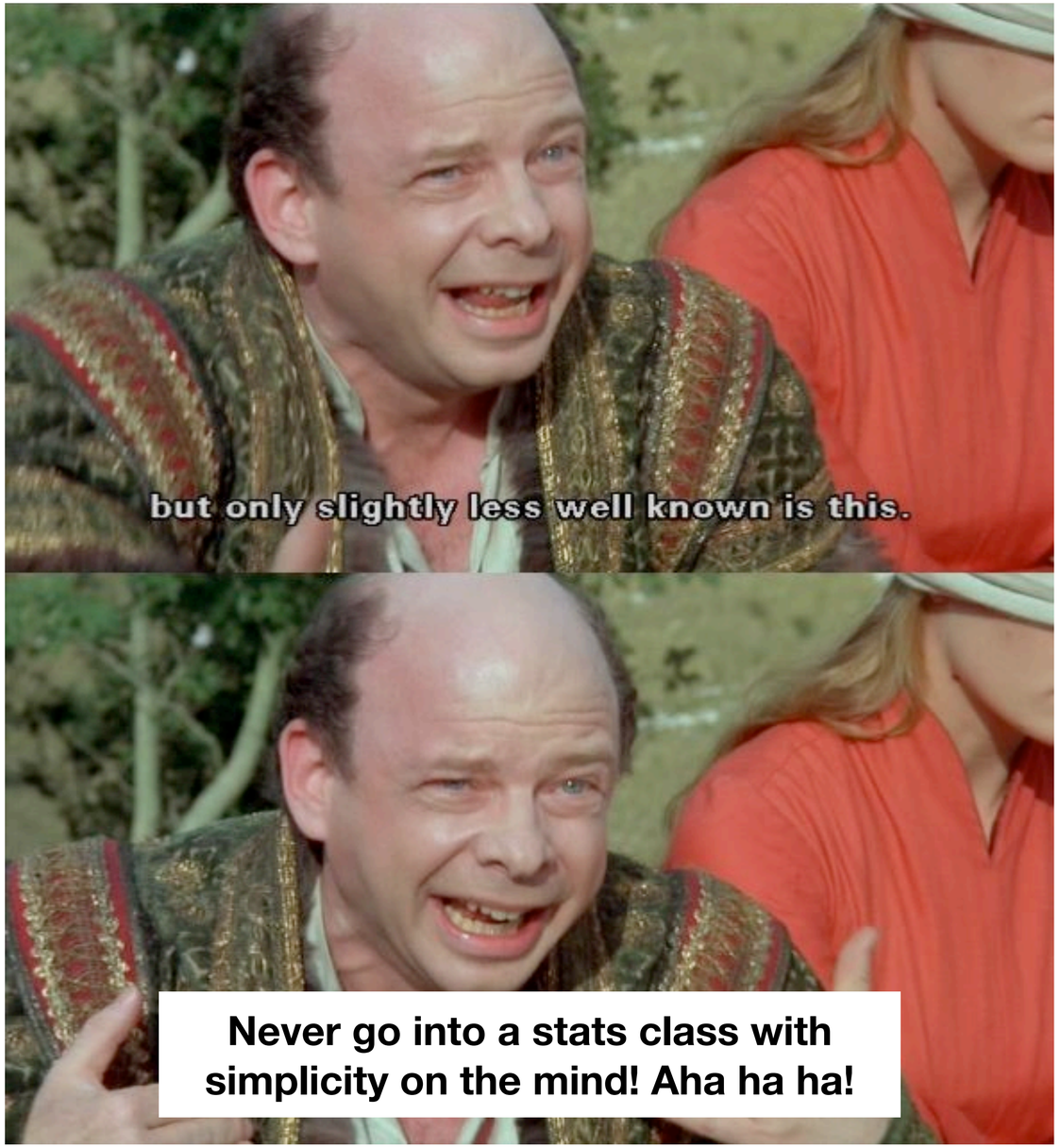
$$(support)_i = \beta_0 + \beta_1(age\_over65)_i + \beta_2(age\_center)_i + \beta_3(age\_center^2)_i$$
$$+\beta_4(age\_over65 \times age\_center)_i + \beta_5(age\_over65 \times age\_center^2)_i + u_i$$

interaction of indicator and running variable
(allows slope to vary at threshold)

interaction of indicator and running variable squared
(allows slope to vary at threshold)

We're still interested in the indicator's coefficient, which is β_1.


You fool, you fell victim
to one of the classic blunders.
The most famous is
never get involved in a land war in Asia,


but only slightly less well known is this.
Never go into a stats class with
simplicity on the mind! Aha ha ha!

Let's try it again.

```
# Square centered age variable
df$age_center_sq <- (df$age_center)^2

# Regression discontinuity: Manual, quadratic
model_4 <- feols(public_option ~ age_over65 + age_center + age_center_sq +
                   age_over65*age_center + age_over65*age_center_sq |
                   0, data=df, vcov="HC1")
summary(model_4)

OLS estimation, Dep. Var.: public_option
Observations: 17,540
Standard-errors: Heteroskedasticity-robust
                          Estimate Std. Error  t value   Pr(>|t|)
(Intercept)               0.701826   0.010256 68.43136  < 2.2e-16 ***
age_over65               -0.034951   0.023181 -1.50772 1.3164e-01
age_center               -0.004977   0.001099 -4.53069 5.9179e-06 ***
age_center_sq            -0.000096   0.000024 -4.00920 6.1176e-05 ***
age_over65:age_center    -0.007925   0.004607 -1.72026 8.5403e-02 .
age_over65:age_center_sq  0.000681   0.000196  3.46848 5.2467e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.44446   Adj. R2: 0.013883
```

Let's try it again.

```
# Square centered age variable
df$age_center_sq <- (df$age_center)^2

# Regression discontinuity: Manual, quadratic
model_4 <- feols(public_option ~ age_over65 + age_center + age_center_sq +
                   age_over65*age_center + age_over65*age_center_sq |
                   0, data=df, vcov="HC1")
summary(model_4)
```

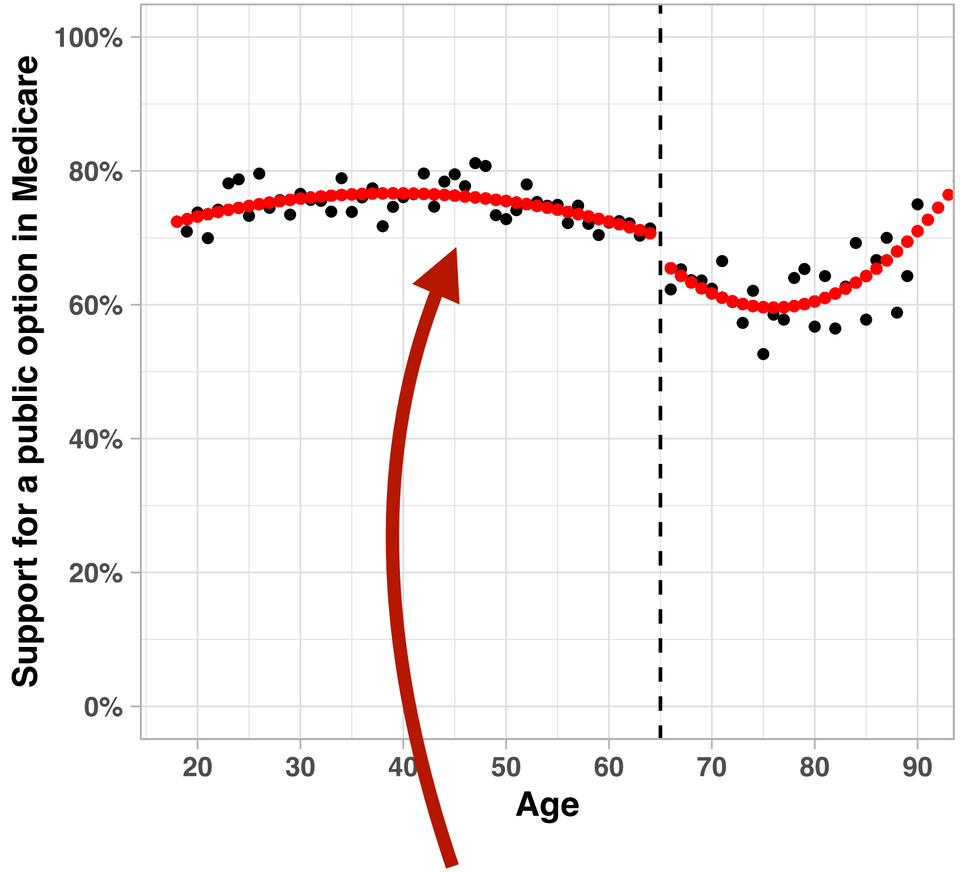
```
OLS estimation, Dep. Var.: public_option
Observations: 17,540
Standard-errors: Heteroskedasticity-robust
                          Estimate Std. Error  t value   Pr(>|t|)
(Intercept)               0.701826   0.010256 68.43136  < 2.2e-16 ***
age_over65               -0.034951   0.023181 -1.50772 1.3164e-01
age_center               -0.004977   0.001099 -4.53069 5.9179e-06 ***
age_center_sq            -0.000096   0.000024 -4.00920 6.1176e-05 ***
age_over65:age_center    -0.007925   0.004607 -1.72026 8.5403e-02 .
age_over65:age_center_sq  0.000681   0.000196  3.46848 5.2467e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.44446   Adj. R2: 0.013883
```

Now, it looks like the estimate is much softer: closer to –3.4 pp.

Using a quadratic RD model, becoming eligible for Medicare did not cause a significant decrease in support for a public option, at least at the 5% level.

Show me the predicted values.

```
# Save predictions
df <- mutate(df, predictions_sq = model_4$fitted.values)

# Plot predictions against true values
plot_rd_3 <- ggplot() +

  # Threshold line
  geom_vline(xintercept=THRESHOLD, linetype="dashed") +

  # Binned scatterplots for each side of 65
  stat_binmean(n=1000, data=subset(df, age < 65),
               aes(x=age, y=public_option), size=1) +
  stat_binmean(n=1000, data=subset(df, age > 65 & age <= 90),
               aes(x=age, y=public_option), size=1) +

  # Predicted support
  geom_point(data=df, aes(x=age, y=predictions_sq),
             color="red", size=1) +

  # Cosmetic modifications
  xlab("Age") + ylab("Support for a public option in Medicare") +
  coord_cartesian(xlim=c(18,90), ylim=c(0,1)) +
  theme_light() +
  theme(text = element_text(size = 10, face = "bold")) +
  scale_x_continuous(breaks = seq(0, 100, 10)) +
  scale_y_continuous(breaks = seq(0, 1, 0.2),
                     labels = function(x) paste0(x*100,"%"))
```

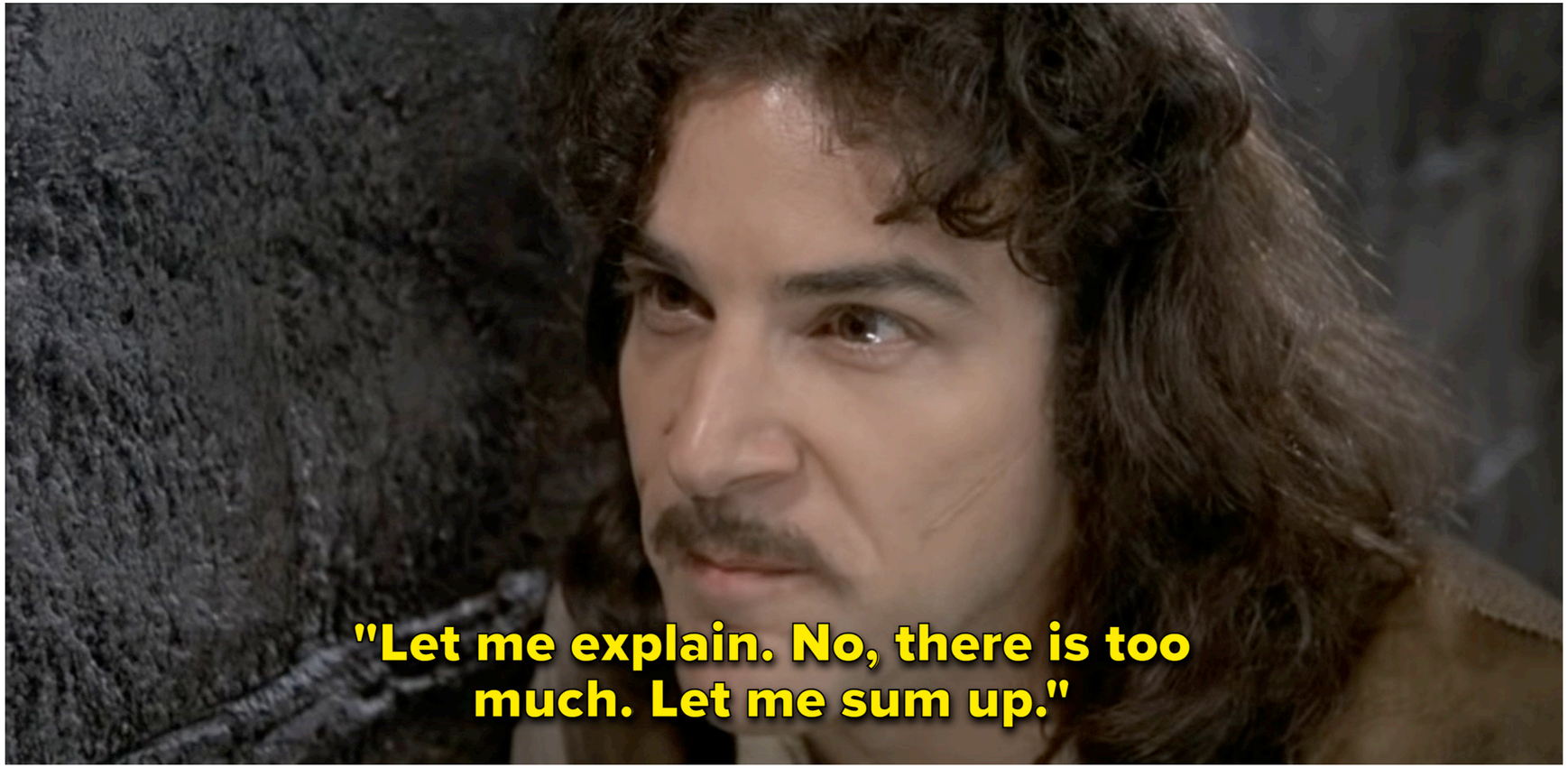


This is a bit better but not perfect. Why?

What might we be missing?

Optional: Advanced RD.

RD methods have exploded in complexity in recent years.



Optional: Advanced RD.

The latest and greatest in RD is to use local linear regression, which fits a smoother line that overweights values nearest to the threshold.

The lingo is a “local linear regression with data-driven, mean squared error-optimal bandwidths.”

This is the “best” approach… at the moment.

OK, one last time. I promise.

```
# Load library
library(rdrobust)

# Regression discontinuity: Data-driven bandwidths
model_5 <- rdrobust(df$public_option, df$age, c=65,
                    p=1, kernel="triangular")
summary(model_5)

Sharp RD estimates using local polynomial regression.

Number of Obs.                17540
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                14225         3315
Eff. Number of Obs.            3075         1803
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   7.037        7.037
BW bias (b)                  11.718       11.718
rho (h/b)                     0.601        0.601
Unique Obs.                      47           31

=============================================================================
        Method     Coef. Std. Err.          z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional    -0.080      0.035    -2.248     0.025     [-0.149 , -0.010]
        Robust          -          -    -2.132     0.033     [-0.183 , -0.008]
=============================================================================
```

OK, one last time. I promise.

```
# Load library
library(rdrobust)

# Regression discontinuity: Data-driven bandwidths
model_5 <- rdrobust(df$public_option, df$age, c=65,
                    p=1, kernel="triangular")
summary(model_5)

Sharp RD estimates using local polynomial regression.

Number of Obs.                17540
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                14225         3315
Eff. Number of Obs.            3075         1803
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   7.037        7.037
BW bias (b)                  11.718       11.718
rho (h/b)                     0.601        0.601
Unique Obs.                      47           31

=============================================================================
        Method     Coef. Std. Err.          z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional    -0.080      0.035    -2.248     0.025     [-0.149 , -0.010]
        Robust         -          -    -2.132     0.033     [-0.183 , -0.008]
=============================================================================
```

OK, so in the “best” model we have, becoming eligible for Medicare causes an 8.0 pp decline in support for a public option (conventional 95% CI, –14.9 to –1.0 pp).

OK, one last time. I promise.

```
# Load library
library(rdrobust)

# Regression discontinuity: Data-driven bandwidths
model_5 <- rdrobust(df$public_option, df$age, c=65,
                    p=1, kernel="triangular")
summary(model_5)

Sharp RD estimates using local polynomial regression.

Number of Obs.                17540
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs                 14225         3315
Eff. Number of Obs.            3075         1803
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   7.037        7.037
BW bias (b)                  11.718       11.718
rho (h/b)                     0.601        0.601
Unique Obs.                      47           31

=============================================================================
        Method     Coef. Std. Err.          z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional    -0.080      0.035    -2.248     0.025    [-0.149 , -0.010]
        Robust          -          -    -2.132     0.033    [-0.183 , -0.008]
=============================================================================
```

This model selects a data-driven bandwidth that balances bias (due to functional form) and variance (or precision) in the estimate.

Only about 3,075 + 1,803 = 4,878 respondents near the threshold contributed to the RD estimate.

Specifically, folks 65 ± 7 years were included.

Research Letter | Health and the 2024 US Election

February 12, 2024

Medicare Eligibility and Reported Support for Proposals to Expand Medicare

Nolan M. Kavanagh, MPH[1]; Andrea L. Campbell, PhD[2]; Adrianna McIntyre, PhD, MPP, MPH[3]

Author Affiliations | Article Information

JAMA. 2024;331(10):882-884. doi:10.1001/jama.2024.0379

Every year, millions of US adults turn 65 years of age and initiate Medicare enrollment. The program's broad popularity makes it—and its beneficiaries—electorally consequential.[1]

Direct experiences with programs may influence beneficiaries' opinions.[2,3] Positive experiences might motivate beneficiaries to support a program's expansion—or to shield it from competing groups.[2] Medicare eligibility has been associated with increased opposition to Medicare cuts,[4] but, to our knowledge, little work has been done to examine how eligibility shapes support for program expansion. We estimated the association between Medicare eligibility and support for recent proposals to expand program participation and benefits.

Methods

We pooled the 2018-2022 waves of the annual Cooperative Election Study (CES), a large national survey of political attitudes.[5] The CES uses a matched random sample design. It simulates a "true" random sample of US adults using the US Census and other authoritative sources. Then, it matches the simulated sample on demographics to panelists of the polling firm YouGov. The response rate to YouGov's panel recruitment was 44% to 51% in the study

Table. Association Between Medicare Eligibility and Public Insurance Coverage or Support for Proposals to Expand Medicare[a]

Variable	Reported outcome (ie, coverage or support)				
	Public health insurance coverage	Expand Medicare to "all Americans"	Allow a public option in Medicare	Lower Medicare eligibility age to 50 y	Allow government to negotiate drug prices
Rate for all respondents, No./total No. (%)	87 988/220 728 (39.9)	146 068/220 388 (66.3)	12 677/17 540 (72.3)	46 758/77 377 (60.4)	105 059/118 589 (88.6)
Rate for ages 18-64 y, No./total No. (%)	48 139/174 931 (27.5)	124 066/174 689 (71.0)	10 623/14 225 (74.7)	41 970/62 244 (67.4)	82 406/93 317 (88.3)
Rate for ages ≥66 y, No./total No. (%)	39 849/45 797 (87.0)	22 002/45 699 (48.1)	2054/3315 (62.0)	4788/15 133 (31.6)	22 653/25 272 (89.6)
Unadjusted difference (95% CI), percentage points[b]	59.5 (59.1 to 59.9)	−22.9 (−23.4 to −22.4)	−12.7 (−14.5 to −10.9)	−35.8 (−36.6 to −35.0)	1.3 (0.9 to 1.8)
P value	<.001	<.001	<.001	<.001	<.001
Adjusted difference at 65 y (95% CI), percentage points[c]	41.3 (38.3 to 44.3)	−0.9 (−2.8 to 1.1)	−8.0 (−14.9 to −1.0)	1.2 (−3.2 to 5.6)	1.2 (−0.3 to 2.7)
P value	<.001	.37	.02	.60	.11
Included ages, y[d]	62 to 68	57 to 73	58 to 72	60 to 70	56 to 74
Effective sample size[e]	25 305	64 616	4878	14 274	39 565
Years when polled	2018 to 2022	2018 to 2022	2019	2019 to 2020	2020 and 2022

[a] Adults aged 66 years or older were less supportive of most proposals to expand Medicare than adults aged 18 to 64 years. Medicare eligibility at exactly 65 years of age was associated with decreased support for one of the proposals. All estimates are based on unweighted, pooled responses to the 2018-2022 waves of the Cooperative Election Study. Respondents who turned 65 years during the survey year were excluded.

[b] Estimates provide the simple unadjusted difference in each outcome for adults aged 66 years or older compared with those aged 18 to 64 years.

[c] Estimates provide the percentage point difference in each outcome at the Medicare eligibility cutoff (ie, exactly 65 years) using regression discontinuity (RD) models. RD models help control for observed and unobserved confounders that can bias simple differences. We fit local linear regressions on either side of 65 years of age with survey wave fixed effects; data-driven, mean squared error (MSE)–optimal bandwidths; and triangular kernels. Our conclusions were not affected by the use of robust bias-corrected 95% CIs with a polynomial order of 2, so we present conventional 95% CIs. Full details are provided in Supplement 1.

[d] Included ages refer to the MSE-optimal bandwidths of the RD models. The model selected the range of ages that best balanced bias and variance in the estimate for each outcome. These bandwidths were ±3.5, ±8.0, ±7.0, ±5.2, and ±9.1 years, respectively.

[e] The total numbers of US adults (ie, full sample minus those aged 65 years) with valid data for each outcome are 220 728, 220 388, 17 540, 77 377, and 118 589, respectively. Effective sample sizes refer to the number of respondents included within the MSE-optimal bandwidths of the RD models.

These are the published estimates, based on the data-driven approach.

What's the catch?

RDs only estimate a local average treatment effect (or LATE) near the threshold.

Is this group of people meaningful?

It's harder to make inferences about a broader population. What about folks who have been on Medicare for years?

What's the catch?

Also, we have to be very sure that no other treatment happens at the threshold.

And that there's not manipulation of the running variable. (Here's that's unlikely since people can't change their ages.)

Lastly, modern RDs need very large samples near the threshold to be well-powered!