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-2")`

The Instrumental Variable Rises

API 203: TF Session 2

R

Nolan M. Kavanagh
March 8, 2024

Goals for today

1. **Review the principles of instrumental variables (IV).**
2. **Learn how to run IV regressions in R.**
3. **Practice interpreting IV regressions.**

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

Overview of our sample data

Dataset of U.S. county characteristics in 2016

`county_fips`	**County FIPS identifier**	*Administrative*
`state`	**State**	*Administrative*
`pop_over_18`	**County voting age population (i.e. over 18 years)**	*American Community Survey (5-year estimates)*
`med_inc_000s`	**County median income (in $1,000s)**	*American Community Survey (5-year estimates)*
`unemploy_rate`	**County unemployment rate (0–100)**	*American Community Survey (5-year estimates)*
`pc_uninsured`	**Percent of county without health insurance (0–100)**	*American Community Survey (5-year estimates)*
`all_votes`	**Total number of votes cast in election**	*MIT Election Lab*
`rep_votes`	**Total number of votes for Trump in 2016**	*MIT Election Lab*
`turnout`	**Percent of county VAP that voted in 2016 (0–100)**	*MIT Election Lab/Constructed*
`pc_rep`	**Percent of county votes for Trump (0–100)**	*MIT Election Lab/Constructed*
`rain_election`	**Simulated rainfall (1) or not (0) on election day**	*Simulated data*
`rain_historical`	**Average county rainfall in October 1901–2000 (in.)**	*National Oceanic and Atmospheric Administration*

Brief disclosure

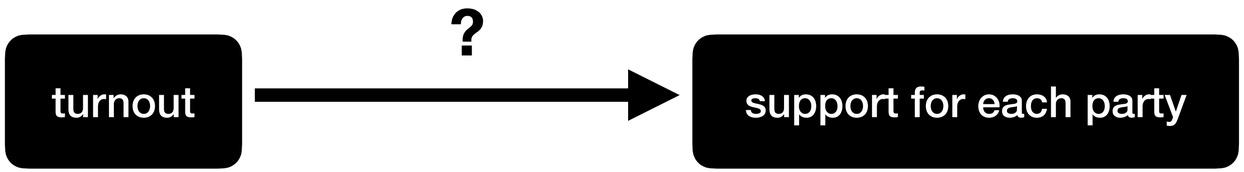
Unlike previous sessions, we will be using partially simulated data. I couldn’t get great rainfall data on election day, so I simulated it.

The concepts all work the same, but don’t take the conclusions of the models to heart!

It’s election season!

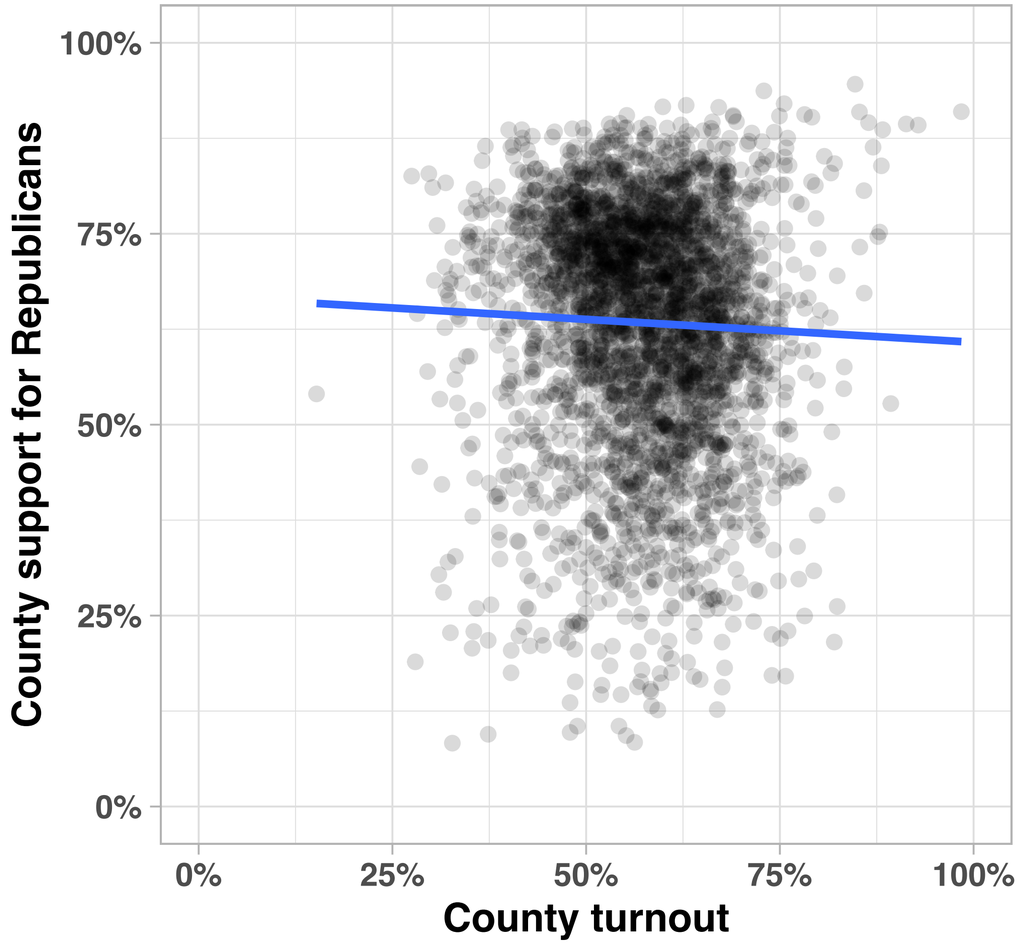
Some elections have higher turnout than others.

When more people turn out to vote, which party benefits: Democrats or Republicans?



We could start with a naive regression.

```
# Graph turnout and percent Republican
plot_1 <- ggplot() +

  # Add scatterplot points
  geom_point(data=df, aes(x=turnout, y=pc_rep), alpha=0.15) +

  # Labels of axes
  xlab("County turnout") +
  ylab("County support for Republicans") +

  # Cosmetic changes
  theme_light() +
  theme(text = element_text(face="bold")) +
  scale_y_continuous(limits=c(0,100),
                     labels = function(x) paste0(x,"%")) +
  scale_x_continuous(limits=c(0,100),
                     labels = function(x) paste0(x,"%")) +

  # Add line of best fit
  geom_smooth(data=df, aes(x=turnout, y=pc_rep),
              method="lm", se=F, formula = y~x)

# Print graph
print(plot_1)
```



We could start with a naive regression.

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

# Estimate between-county regression
model_1 <- feols(pc_rep ~ turnout | 0, df)
summary(model_1)
```

```
OLS estimation, Dep. Var.: pc_rep
Observations: 3,106
Standard-errors: IID
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 66.96115   1.709410 39.17208 < 2.2e-16 ***
turnout     -0.06312   0.029339 -2.15137  0.031524 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.6   Adj. R2: 0.001167
```

Across all counties in 2016, a 1 pp increase in turnout was significantly associated with 0.06 pp less support for Trump.

Note: Both variables are coded 0–100.

We could try a few other regressions.

	Model 1	Model 2	Model 3
County turnout (0–100)	**–0.06***	**–0.08****	**0.22*****
	(0.03)	(0.03)	(0.03)
County median income ($1,000s)		**–0.49*****	**–0.35*****
		(0.02)	(0.02)
County unemployment rate (0–100)		**–2.21*****	**–1.66*****
		(0.09)	(0.09)
Num.Obs.	3,106	3,106	3,106
State fixed effects	No	No	Yes
Standard errors	Robust	Robust	Robust
R2 Adj.	0.001	0.198	0.399

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

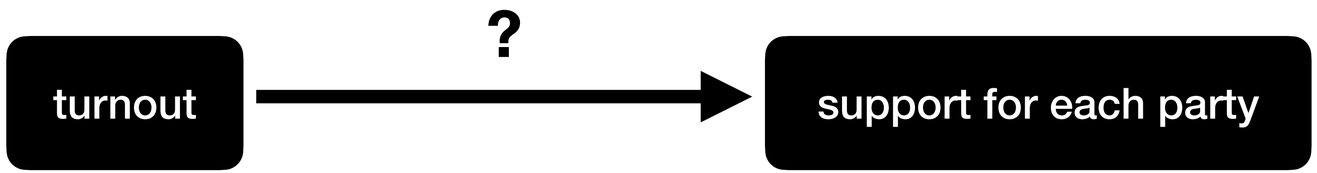
<u>Within</u> each state, a 1 pp increase in county turnout was significantly associated with a 0.22 pp increase in support for Donald Trump.

But does this deal with all omitted variables we're worried about?


omitted variables
me and my silly little controls

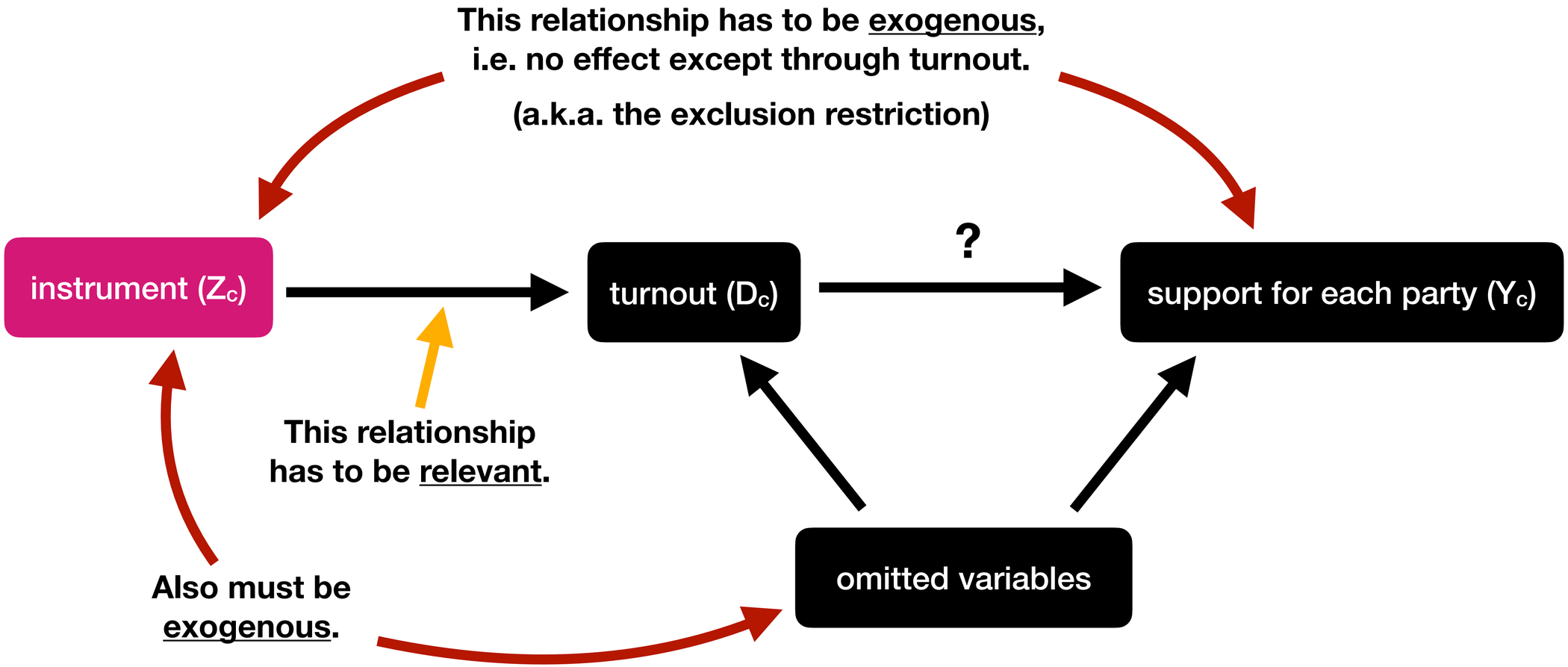
Enter: Instrumental variables

Some of this relationship is causal; some is due to omitted variables.



IVs basically split the relationship into causal and non-causal parts by taking advantage of random variation in a third variable.

What's up with instrumental variables?



Let’s say our instrument is rain.

The idea is that random variation in rain = random variation in turnout.

<u>Stage 1:</u> Regress D on Z and take the predicted values.
This captures the part of our predictor that is “random.”

$$(turnout)_c = \alpha_0 + \alpha_{FS}(rain)_c + \alpha_2(average\_rain)_c + v_c$$

I’ve included a control for the average rainy-ness of each county.

Let's say our instrument is rain.

The idea is that random variation in rain = random variation in turnout.

<u>Stage 1:</u> Regress D on Z and take the predicted values.
This captures the part of our predictor that is "random."

I've included a control for the average rainy-ness of each county.

$$(turnout)_c = \alpha_0 + \alpha_{FS}(rain)_c + \alpha_2(average\_rain)_c + v_c$$

<u>Stage 2:</u> Then, regress Y on the <u>predicted</u> values of D.
This estimates the change in our outcome due to random variation in our predictor.

$$(pc\_Republican)_c = \beta_0 + \beta_{IV}(\widehat{turnout})_c + \beta_2(average\_rain)_c + u_c$$

This process is called "two-stage least squares."

Let's estimate the first stage.

<u>Stage 1:</u> Regress D on Z and take the predicted values.

$$(turnout)_c = \alpha_0 + \alpha_{FS}(rain)_c + \alpha_2(average\_rain)_c + v_c$$

```
# Estimate first stage
model_fs <- feols(turnout ~ rain_election + rain_historical | 0, data=df)
summary(model_fs)
```

```
OLS estimation, Dep. Var.: turnout
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error   t value  Pr(>|t|)
(Intercept)     61.109689   0.476529 128.23929 < 2.2e-16 ***
rain_election   -8.058613   0.355154 -22.69047 < 2.2e-16 ***
rain_historical -0.540862   0.169689  -3.18737   0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 8.80973   Adj. R2: 0.147043
```

Let's estimate the first stage.

<u>Stage 1:</u> Regress D on Z and take the predicted values.

$$(turnout)_c = \alpha_0 + \alpha_{FS}(rain)_c + \alpha_2(average\_rain)_c + v_c$$

```
# Estimate first stage
model_fs <- feols(turnout ~ rain_election + rain_historical | 0, data=df)
summary(model_fs)
```

```
OLS estimation, Dep. Var.: turnout
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error   t value  Pr(>|t|)
(Intercept)     61.109689   0.476529 128.23929 < 2.2e-16 ***
rain_election   -8.058613   0.355154 -22.69047 < 2.2e-16 ***
rain_historical -0.540862   0.169689  -3.18737   0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 8.80973   Adj. R2: 0.147043
```

We can interpret this causally since rain is exogenous, after controlling for average rain.

Controlling for historical average rain, rain on election day caused a decrease in turnout of 8 pp.

It's significant at the 5% level.

Let's estimate the first stage.

<u>Stage 1:</u> Regress D on Z and take the predicted values.

$$(turnout)_c = \alpha_0 + \alpha_{FS}(rain)_c + \alpha_2(average\_rain)_c + v_c$$

```
# Estimate first stage
model_fs <- feols(turnout ~ rain_election + rain_historical | 0, data=df)
summary(model_fs)
```

```
OLS estimation, Dep. Var.: turnout
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error   t value  Pr(>|t|)
(Intercept)     61.109689   0.476529 128.23929 < 2.2e-16 ***
rain_election   -8.058613   0.355154 -22.69047 < 2.2e-16 ***
rain_historical -0.540862   0.169689  -3.18737   0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 8.80973   Adj. R2: 0.147043
```

This is a large t-stat. Looks relevant!

A common rule of thumb is that the F-statistic = (t-statistic)^2 for our instrument should be ≥10.

(-22.7)^2 = 514. We're good!

Let's estimate the first stage.

<u>Stage 1:</u> Regress D on Z and take the predicted values.

$$(turnout)_c = \alpha_0 + \alpha_{FS}(rain)_c + \alpha_2(average\_rain)_c + v_c$$

```
# Estimate first stage
model_fs <- feols(turnout ~ rain_election + rain_historical | 0, data=df)
summary(model_fs)
```

```
OLS estimation, Dep. Var.: turnout
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error   t value  Pr(>|t|)
(Intercept)     61.109689   0.476529 128.23929 < 2.2e-16 ***
rain_election   -8.058613   0.355154 -22.69047 < 2.2e-16 ***
rain_historical -0.540862   0.169689  -3.18737   0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 8.80973   Adj. R2: 0.147043
```

Can we interpret historical rain as causal?

No! Its coefficient reflects the fact that places with more historical rain are less supportive of Trump.

But we haven't specified its causal model. We only have a causal model for rain on election day.

Now, let's estimate the second stage.

Stage 2: Regress Y on the predicted values of D.

$$(pc\_Republican)_c = \beta_0 + \beta_{IV}(\widehat{turnout})_c + \beta_2(average\_rain)_c + u_c$$

```
# Save predicted turnout
df$turnout_pred <- predict(reg_fs)

# Estimate second stage
model_ss <- feols(pc_rep ~ turnout_pred + rain_historical | 0, data=df)
summary(model_ss)

OLS estimation, Dep. Var.: pc_rep
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)     78.496704   4.701543 16.69595  < 2.2e-16 ***
turnout_pred    -0.185110   0.077666 -2.38340 1.7213e-02 *
rain_historical -1.723287   0.304782 -5.65416 1.7082e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.5   Adj. R2: 0.010057
```

Now, let's estimate the second stage.

Stage 2: Regress Y on the predicted values of D.

$$(pc\_Republican)_c = \beta_0 + \beta_{IV}(\widehat{turnout})_c + \beta_2(average\_rain)_c + u_c$$

```
# Save predicted turnout
df$turnout_pred <- predict(reg_fs)

# Estimate second stage
model_ss <- feols(pc_rep ~ turnout_pred + rain_historical | 0, data=df)
summary(model_ss)

OLS estimation, Dep. Var.: pc_rep
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)     78.496704   4.701543 16.69595  < 2.2e-16 ***
turnout_pred    -0.185110   0.077666 -2.38340 1.7213e-02 *
rain_historical -1.723287   0.304782 -5.65416 1.7082e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.5   Adj. R2: 0.010057
```

We can interpret this causally since we're using the "portion" of turnout due to the random presence of rain on election day.

Using rain on election day as an instrument, a 1 pp increase in turnout led to a 0.2 pp decline in support for Trump in 2016, controlling for average county rain.

It's significant at the 5% level.

Now, let's estimate the second stage.

Stage 2: Regress Y on the predicted values of D.

$$(pc\_Republican)_c = \beta_0 + \beta_{IV}(\widehat{turnout})_c + \beta_2(average\_rain)_c + u_c$$

```
# Save predicted turnout
df$turnout_pred <- predict(reg_fs)

# Estimate second stage
model_ss <- feols(pc_rep ~ turnout_pred + rain_historical | 0, data=df)
summary(model_ss)
```

```
OLS estimation, Dep. Var.: pc_rep
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)     78.496704   4.701543 16.69595  < 2.2e-16 ***
turnout_pred    -0.185110   0.077666 -2.38340 1.7213e-02 *
rain_historical -1.723287   0.304782 -5.65416 1.7082e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.5   Adj. R2: 0.010057
```

As before, we don't want to interpret historical rain causally.

Most researchers will not verbally interpret it in their papers.

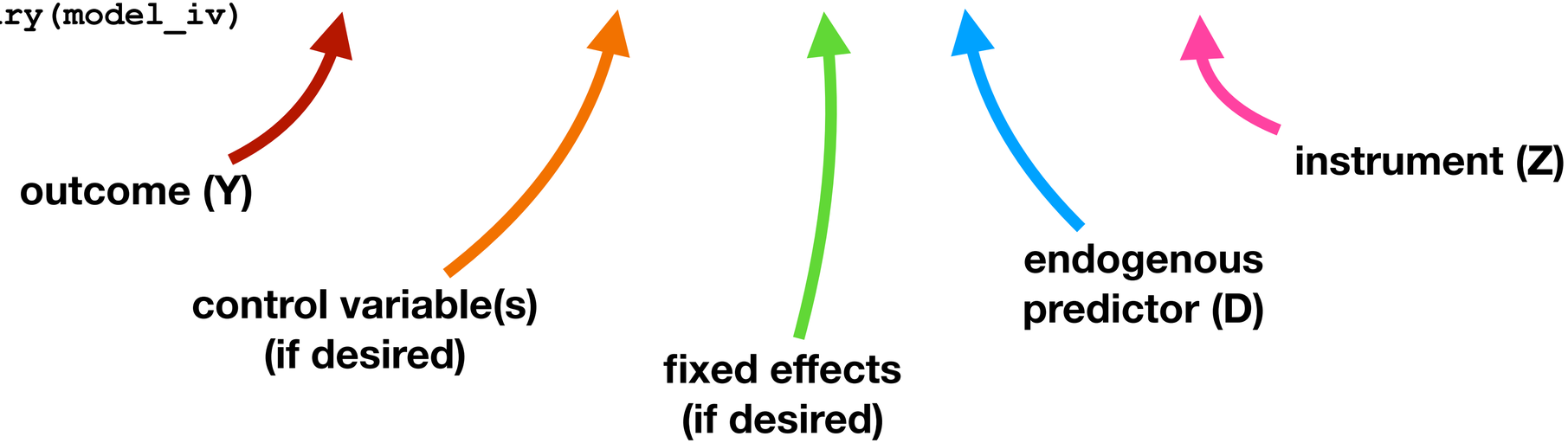
Some (like Nolan) won't even report the coefficient at all!

If tomorrow, I do a low-effort regression with lots of omitted variables, nobody panics.

But when I say “causal two-stage least squares regression,” everyone loses their minds!

We can also have R do it all at once!

```
# Estimate all-in-one IV model
# This includes a degree of freedom correction
model_iv <- feols(pc_rep ~ rain_historical | 0 | turnout ~ rain_election, data=df)
summary(model_iv)
```



`feols()` can do it all for us!

We can also have R do it all at once!

```
# Estimate all-in-one IV model
# This includes a degree of freedom correction
model_iv <- feols(pc_rep ~ rain_historical | 0 | turnout ~ rain_election, data=df)
summary(model_iv)

TSLS estimation, Dep. Var.: pc_rep, Endo.: turnout, Instr.: rain_election
Second stage: Dep. Var.: pc_rep
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)     78.496704   4.711348 16.66120  < 2.2e-16 ***
fit_turnout     -0.185110   0.077828 -2.37844 1.7446e-02 *
rain_historical -1.723287   0.305418 -5.64239 1.8279e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.6   Adj. R2: 0.005924
F-test (1st stage), turnout: stat = 514.9    , p < 2.2e-16 , on 1 and 3,103 DoF.
                Wu-Hausman: stat =   2.31607, p = 0.128145, on 1 and 3,102 DoF.
```

We'll notice that this is the same coefficient but slightly different SE and P-value since `feols()` includes a degree of freedom correction that the two-stage version didn't.

Whom does it apply to?

The “local average treatment effect” (LATE) is only for compliers.

We must ask: Is this group generalizable?

In our case, the IV estimate applies to counties that decrease turnout in response to rain.

Are some counties’ voters “hardier” than others?

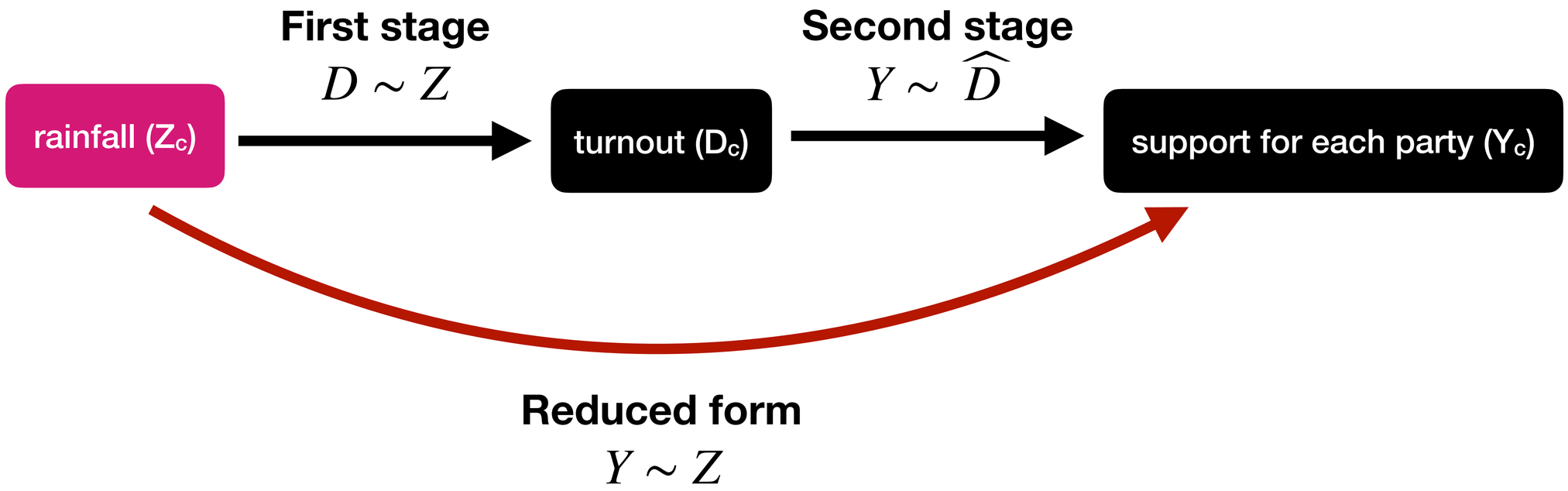
		Offered (Z = 1)	
		Don’t get treatment (D = 0)	**Get treatment (D = 1)**
Not offered (Z = 0)	**Don’t get treatment (D = 0)**	Never takers	Compliers
	Get treatment (D = 1)	Defiers	Always takers

We assume this group doesn’t exist. Is that a good assumption?

What about this “reduced form”?

The reduced form is our intention-to-treat estimate.

It gives us the direct effect of the instrument (Z) on the outcome (Y).



What about this “reduced form”?

```
# Estimate reduced form
reg_rf <- feols(pc_rep ~ rain_election + rain_historical | 0, data=df)
summary(reg_rf)

OLS estimation, Dep. Var.: pc_rep
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)      67.18468   0.839781 80.00266  < 2.2e-16 ***
rain_election     1.49173   0.625884  2.38340 1.7213e-02 *
rain_historical -1.62317   0.299042 -5.42790 6.1411e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.5   Adj. R2: 0.010057
```

What about this “reduced form”?

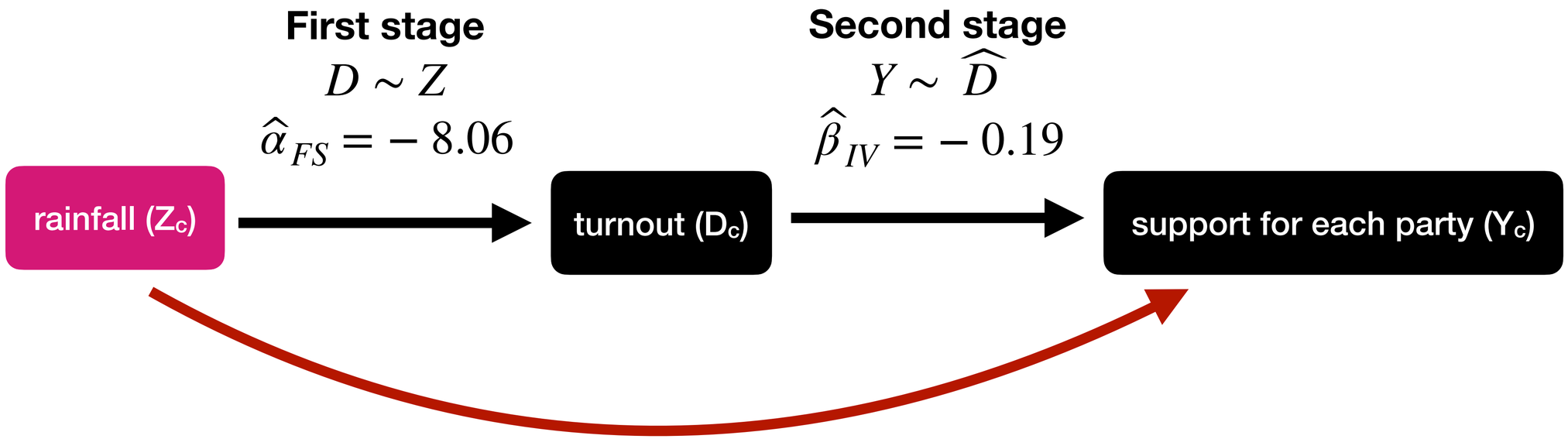
```
# Estimate reduced form
reg_rf <- feols(pc_rep ~ rain_election + rain_historical | 0, data=df)
summary(reg_rf)

OLS estimation, Dep. Var.: pc_rep
Observations: 3,106
Standard-errors: IID
                 Estimate Std. Error  t value   Pr(>|t|)
(Intercept)      67.18468   0.839781 80.00266  < 2.2e-16 ***
rain_election     1.49173   0.625884  2.38340 1.7213e-02 *
rain_historical -1.62317   0.299042 -5.42790 6.1411e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 15.5   Adj. R2: 0.010057
```

Rain on election day caused a significant 1.5 pp increase in support for Trump in 2016, controlling for average county rain.

It’s significant at the 5% level.


“Do you want to see a magic trick?”

Do you want to see a magic trick?



It works either way!

The IV estimate is just the reduced form scaled by the first stage.

Reduced form

$$Y \sim Z$$
$$\hat{\gamma}_{RF} = 1.49$$

Wald estimate

$$\frac{\hat{\gamma}_{RF}}{\hat{\alpha}_{FS}} = \frac{1.49}{-8.06} = -0.19$$

What can go wrong?

As it turns out, potentially a lot:

1. **Weak IV: The first stage just isn't very strong ("irrelevant").**
2. **Violations of exclusion restriction: Does rain *only* affect turnout?**
3. **Not "independent": The instrument might not actually be random.**
4. **Compliers? This group might not be our policy's target.**

We can only prove 1 and maybe 3. And 4 is a matter of judgment.

IVs can be powerful…

…but we have to nail the assumptions.

If we don’t, IVs are not especially informative.

Be especially skeptical of the exclusion restriction.

There have been hundreds of IV rainfall papers.
Can they all pass the exclusion restriction?