IV lab

Data and packages

Before we start, a very important video: https://twitter.com/matt_blackwell/status/1362636561813303297

Acharya, Blackwell and Sen (2016) are interested in the long-run effects of institutions/political events on political attitudes. They argue that the institution of slavery in the American South had an enormous influence on local economies and way of life - an influence that has persisted through history and subsists today.

You can think of their argument as follows: slavery-dependent economy -> backlash against emancipation -> creation of institutions that attempt to protect white domination

It's a story of "institutional path dependence and intergenerational socialization." It's an ambitious argument: part of the present-day variation in white attitudes toward Blacks in the South could be traced back to variation in slavery in the 1800s.

Let's read in the data:

```
library(pacman)
pacman::p_load(ggplot2,
               tidyverse,
               modelsummary,
               devtools,
               viridis,
               AER,
               Formula)

# To make nice plots
# install_github('erocoar/gghalves')
library(gghalves)

# county-level data
wh.counties <- read.csv("data/abs-jop-cces-white-countydata.csv",
                        stringsAsFactors = FALSE)

# vector of Southern state names
st.list <- c("AL", "AR", "GA", "FL", "KY", "LA", "MS", "MO", "NC", "SC", "TN", "TX", "VA","WV")

# Dummy that identifies a county as Southern
wh.counties$abs.sample <- 1 * (wh.counties$state.abb %in% st.list)

# Only South counties
south.counties <- subset(wh.counties, abs.sample == 1)
```

Y_i: attitudes of Southern whites toward blacks today. There are four different measures: Democratic partisanship, racial resentment scale, attitudes toward affirmative action, difference in feeling thermometer scores between whites and blacks.

D_i: Local prevalence of slavery just before emancipation. (proportion of the population that was enslaved in 1860)

Our variables of interest are measured at the county-level.

1. Why not regress Y_i on D_i and be done with it?
2. Why not regress Y_i on D_i and a set of county-level covariates?
3. Why not matching?

The point I'm trying to make is that "out there" (outside the confines of this course), you won't be told which identification strategy to use. A big part of our work is not only to apply some identification strategy, but to think about which one makes sense in the first place.

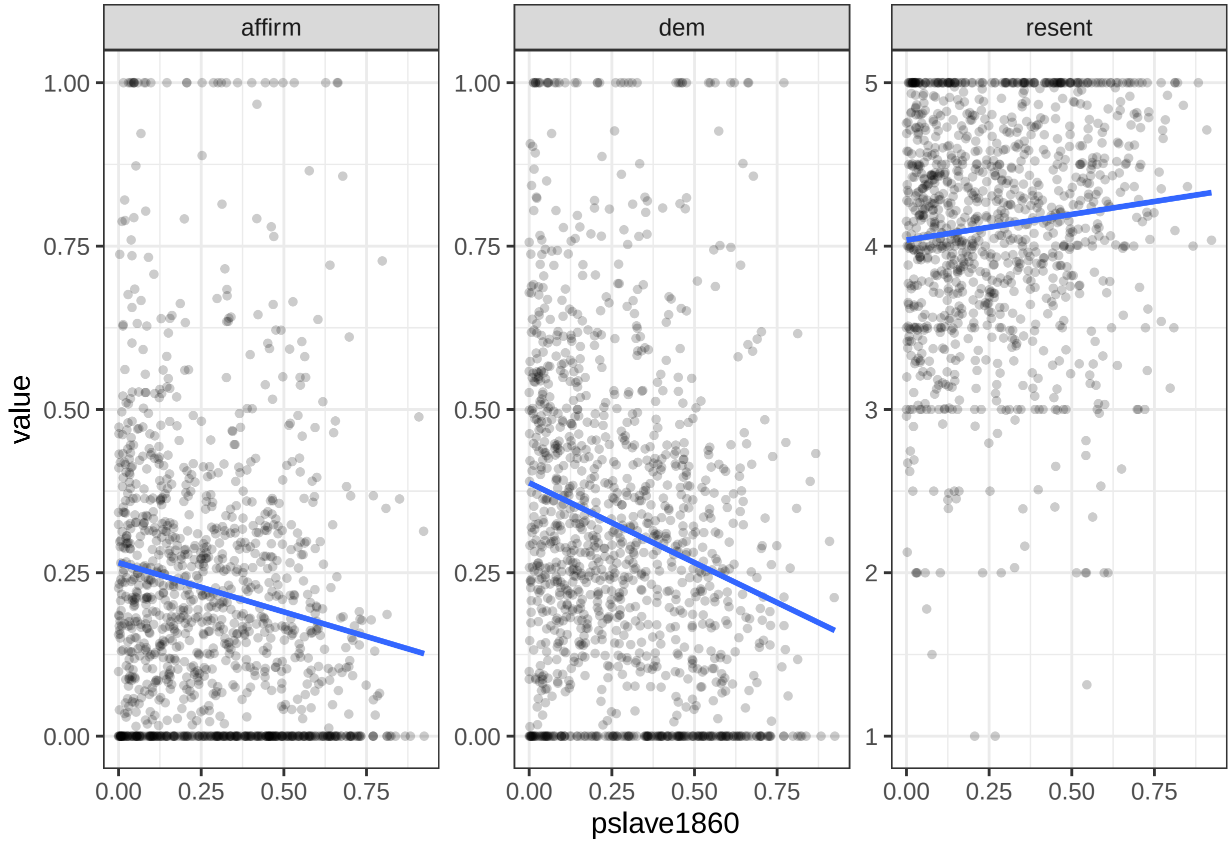
Let's visualize what a naive ATE estimate would look like:

```
# Plotting bivariate relationships
# 3 outcomes: prop. Democrat, support for AA, racial resentment
dplyr::select(
  south.counties, pslave1860, dem, affirm, resent, sample.size
) %>%
  pivot_longer(cols = dem:resent,
               names_to = "outcome") %>%
  ggplot(aes(x = pslave1860, y = value, size = sample.size)) +
  geom_point(alpha = 0.2, size = 1) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw() %+replace%
  theme(legend.position = "none") +
  facet_wrap(~outcome,
             scales = "free_y")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 396 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 396 rows containing missing values (geom_point).
```



```
naive_ate <- lm(affirm ~ pslave1860,
                data = south.counties)
summary(naive_ate)
```

```
##
## Call:
## lm(formula = affirm ~ pslave1860, data = south.counties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26548 -0.16829 -0.04013  0.09335  0.83447
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.26548    0.01033  25.709  < 2e-16 ***
## pslave1860  -0.15057    0.02903  -5.186 2.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2188 on 1240 degrees of freedom
##   (87 observations deleted due to missingness)
## Multiple R-squared:  0.02123,    Adjusted R-squared:  0.02044
## F-statistic:  26.9 on 1 and 1240 DF,  p-value: 2.506e-07
```

Thoughts? What can we tell from this? 1. What's the likely nature of the selection problem? 2. What problem do we solve by instrumenting for D_i (proportion slave)?

The instrument

Acharya, Blackwell and Sen use cotton suitability measured by UNFAO at the county-level.

```
# First stage: regress D on Z
iv.stage1 <- lm(pslave1860 ~ cottonsuit,
   data = south.counties)
summary(iv.stage1)
```

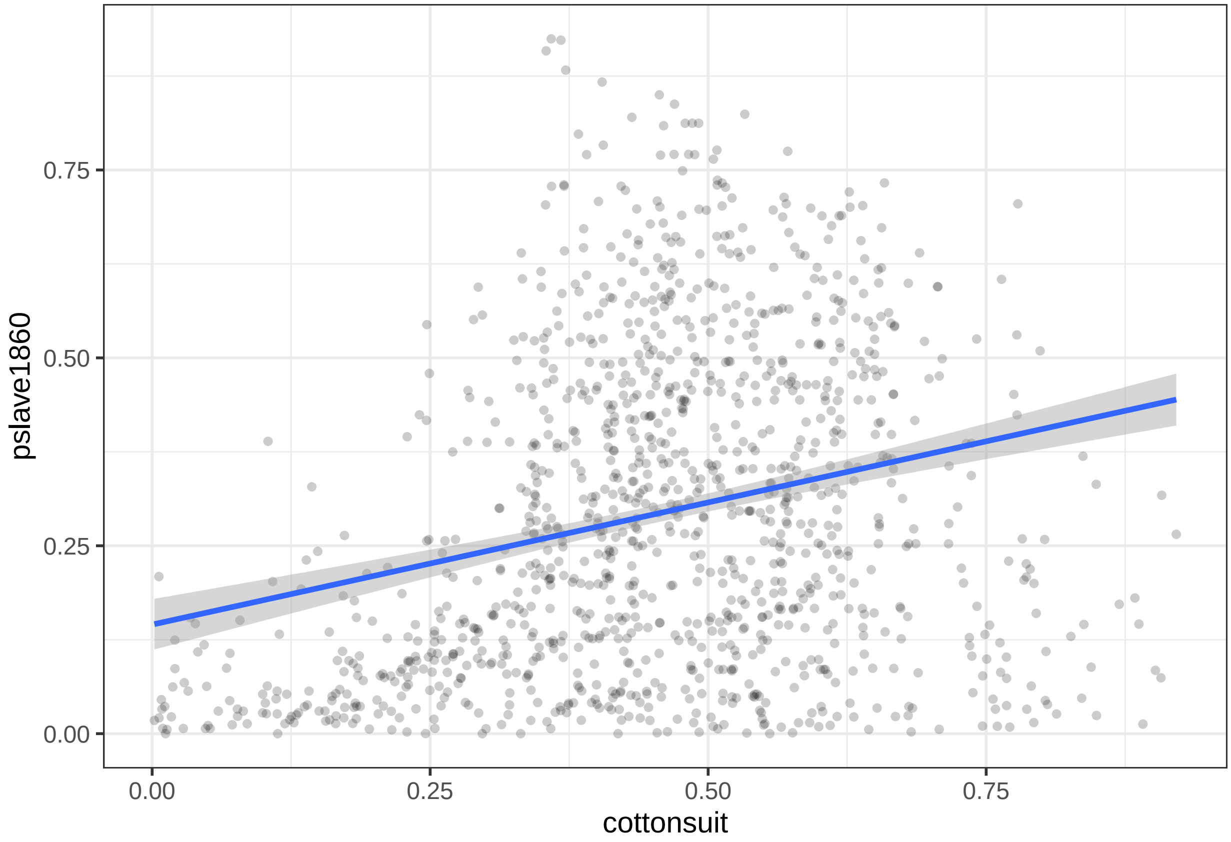
```
##
## Call:
## lm(formula = pslave1860 ~ cottonsuit, data = south.counties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42205 -0.15541 -0.02863  0.14892  0.66271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14508    0.01723   8.420   <2e-16 ***
## cottonsuit   0.32512    0.03556   9.143   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2054 on 1190 degrees of freedom
##   (137 observations deleted due to missingness)
## Multiple R-squared:  0.06564,    Adjusted R-squared:  0.06486
## F-statistic:  83.6 on 1 and 1190 DF,  p-value: < 2.2e-16
```

```
# We can also do this graphically:
rdc_form <- ggplot(south.counties,
       aes(x = cottonsuit, y = pslave1860)) +
  geom_point(alpha = 0.2, size = 1) +
  theme_bw() +
  geom_smooth(method = "lm")
rdc_form
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 137 rows containing non-finite values (stat_smooth).

## Warning: Removed 137 rows containing missing values (geom_point).
```



```r
# Form of the relationship?
rdc_form +
  geom_smooth(col = "red")
```

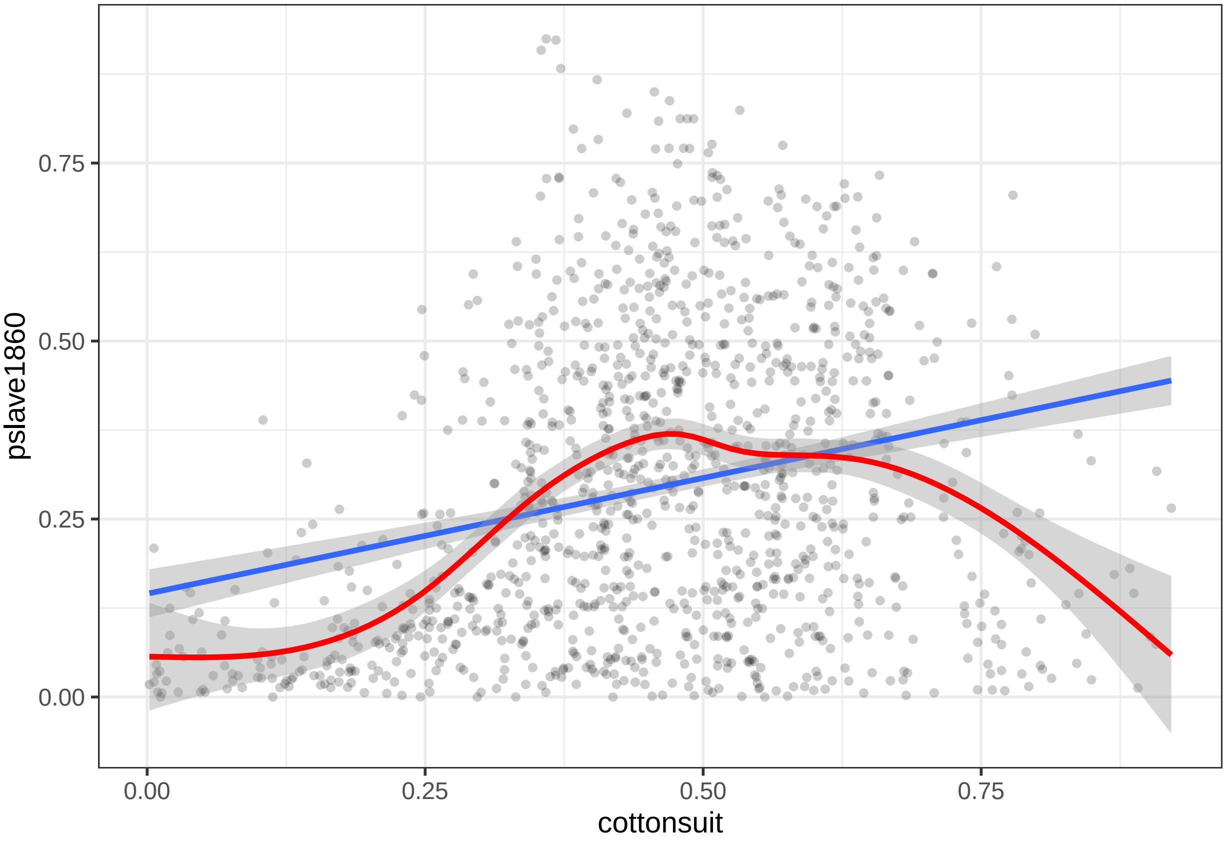
```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 137 rows containing non-finite values (stat_smooth).

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 137 rows containing non-finite values (stat_smooth).

## Warning: Removed 137 rows containing missing values (geom_point).
```



How can we frame this in terms of "encouragement"? Assume a dummy instrument $Z_i \in \{$Low cotton suitability,High cotton s
and a dummmy treatment $D_i \in \{$Weak slavery, Strong slavery$\}$

Who are the "compliers"? Assuming heterogeneous treatment effects, how can we interpret the LATE? What's the "weakness" of the LATE as an estimand?

First stage

Let's start by examining the first stage: the regression of D_i on Z_i.

```
# First stage regression without covariates
iv.stage1 <- lm(pslave1860 ~ cottonsuit,
   data = south.counties)

summary(iv.stage1)
```

```
##
## Call:
## lm(formula = pslave1860 ~ cottonsuit, data = south.counties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42205 -0.15541 -0.02863  0.14892  0.66271
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14508    0.01723   8.420   <2e-16 ***
## cottonsuit   0.32512    0.03556   9.143   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2054 on 1190 degrees of freedom
##   (137 observations deleted due to missingness)
## Multiple R-squared:  0.06564,    Adjusted R-squared:  0.06486
## F-statistic:  83.6 on 1 and 1190 DF,  p-value: < 2.2e-16
```

```
# Producing fitted values for D_i using the first stage regression
south.counties$stage1_fit_nocov <- predict(iv.stage1, newdata = south.counties)

# We can store a formula using the formula() function
# It's then easier to estimate different models with
# the same covariates but different outcomes
stage1.form.covariates <- formula(
  . ~ cottonsuit + log(coarea00) + rugged +
    latitude + I(latitude^2) + longitude +
    I(longitude^2)  + water1860  + state.abb
)

iv.stage1.cov <- lm(
  formula = update(stage1.form.covariates, pslave1860 ~ .),
  data = south.counties,
  weights = sample.size
)

summary(iv.stage1.cov)
```

```
##
## Call:
## lm(formula = update(stage1.form.covariates, pslave1860 ~ .),
##     data = south.counties, weights = sample.size)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0112 -0.2339  0.0278  0.3259  5.1654
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.2436213  1.8853410  -2.781 0.005508 **
## cottonsuit       0.4079082  0.0395272  10.320  < 2e-16 ***
## log(coarea00)    0.0098382  0.0096616   1.018 0.308769
## rugged          -0.0004995  0.0001309  -3.817 0.000143 ***
## latitude         0.4340699  0.0438944   9.889  < 2e-16 ***
## I(latitude^2)   -0.0069634  0.0006817 -10.215  < 2e-16 ***
## longitude        0.0323710  0.0421586   0.768 0.442748
## I(longitude^2)   0.0002039  0.0002422   0.842 0.400039
## water1860        0.0367111  0.0092736   3.959 8.02e-05 ***
## state.abbAR     -0.1500581  0.0338493  -4.433 1.02e-05 ***
## state.abbFL     -0.1495238  0.0301287  -4.963 8.05e-07 ***
## state.abbGA     -0.0256211  0.0251865  -1.017 0.309256
## state.abbKY     0.0814495  0.0358499   2.272 0.023282 *
## state.abbLA     0.0185650  0.0346937   0.535 0.592681
## state.abbMO    -0.0748181  0.0413809  -1.808 0.070875 .
## state.abbMS     0.0824829  0.0336736   2.449 0.014461 *
## state.abbNC     0.0485902  0.0340823   1.426 0.154248
## state.abbSC     0.1173274  0.0318289   3.686 0.000239 ***
## state.abbTN     0.0251802  0.0266218   0.946 0.344433
## state.abbTX    -0.2385709  0.0439660  -5.426 7.08e-08 ***
## state.abbVA     0.2629764  0.0470852   5.585 2.94e-08 ***
## state.abbWV     0.1684690  0.0513956   3.278 0.001079 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.697 on 1098 degrees of freedom
##   (209 observations deleted due to missingness)
## Multiple R-squared:  0.4936, Adjusted R-squared:  0.4839
## F-statistic: 50.95 on 21 and 1098 DF,  p-value: < 2.2e-16
```

```
# Producing fitted values for D_i using the 1st stage with covariates
south.counties$stage1_fit_withcov <- predict(iv.stage1.cov, newdata = south.counties)

# Actually, we only want to use the observations that were not listwise
# deleted in the first stage regression. We can extract those from the lm
# object.
# We also need the Y_i outcomes, which were not part of the first stage model

# I can extract the observations # that were listwise deleted during the
# first stage
iv.stage1.cov$na.action %>% head()
```

```
##  33 167 295 298 300 303
##  33 136 139 142 144 147
```

```
iv.data <- south.counties[-iv.stage1.cov$na.action,] %>%
  mutate(stage1_fit = iv.stage1.cov$fitted.values)
```

Covariates or no?

Should we include covariates in our first stage? Yes, probably. One of our identification assumptions is exogeneity of the instrument - the instrument is as good as random. With observational data, this is a whole lot more likely if we condition on relevant covariates!

Besides, look at the fitted values for D_i that the first stage produces:

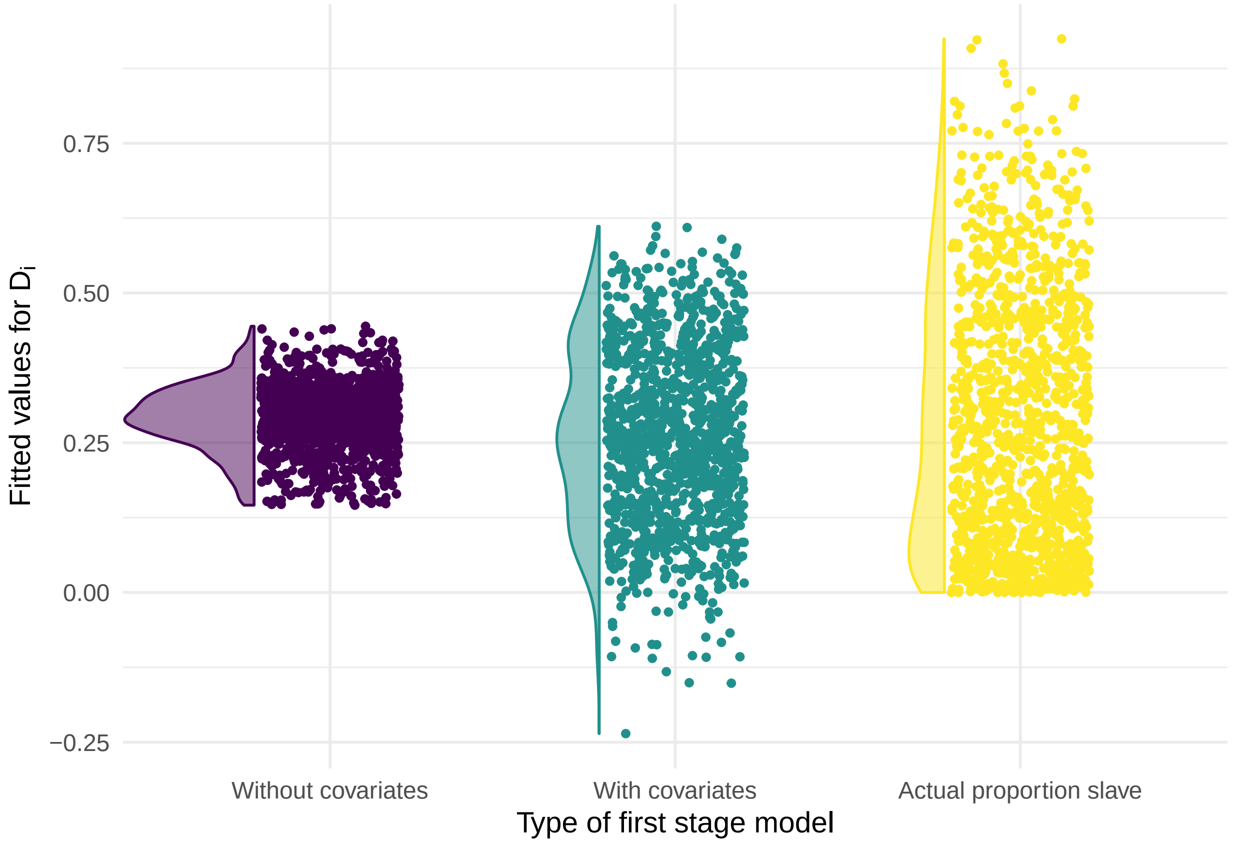
```
fit_differences <- south.counties %>%
  dplyr::select(stage1_fit_nocov, stage1_fit_withcov, pslave1860) %>%
  dplyr::rename(
    `Without covariates` = 1, `With covariates` = 2, `Actual proportion slave` = 3
    ) %>%
  pivot_longer(cols = everything(),
               names_to = "fit_type",
               values_to = "fit") %>%
  mutate(fit_type = factor(fit_type,
                           levels = c("Without covariates", "With covariates", "Actual proportion slave

fit_diff_plot <- ggplot(fit_differences,
       aes(x = fit_type, y = fit, col = fit_type, fill = fit_type)) +
  geom_jitter(size = 1, width = 0.2) +
  geom_half_violin(alpha = 0.5, side = "l", position = position_nudge(x = -.22), trim = T) +
  theme_minimal() %+replace%
  theme(legend.position = "none") +
  # you can use expression() to have math in your ggplot labels
  labs(x = "Type of first stage model",
       y = expression("Fitted values for "*D[i]*"")) +
  scale_color_manual(values = viridis(3)) +
  scale_fill_manual(values = viridis(3))

fit_diff_plot
```

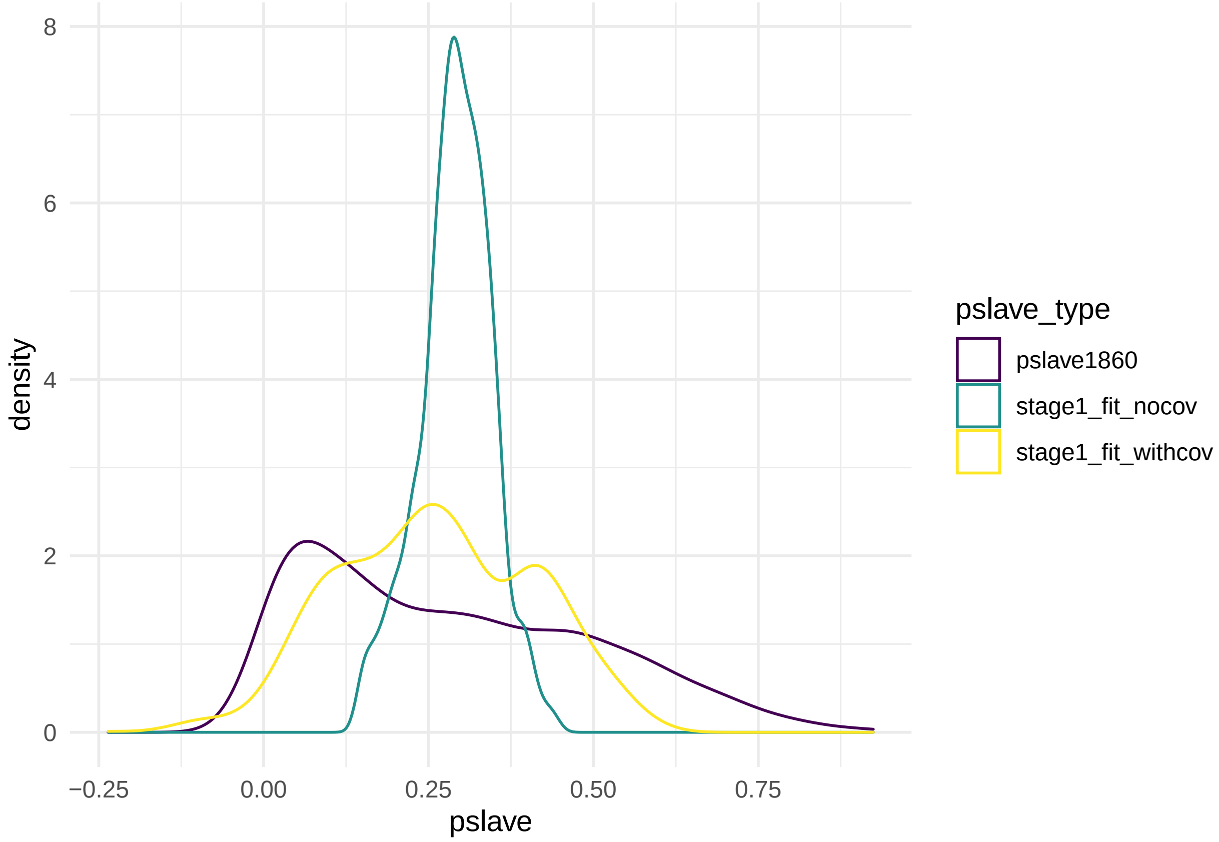
```
## Warning: Removed 304 rows containing non-finite values (stat_half_ydensity).
```

```
## Warning: Removed 304 rows containing missing values (geom_point).
```



```
# Another way of looking at this
south.counties %>%
  dplyr::select(pslave1860, stage1_fit_nocov, stage1_fit_withcov) %>%
  pivot_longer(cols = everything(),
               names_to = "pslave_type",
               values_to = "pslave") %>%
  ggplot(aes(x = pslave, col = pslave_type)) +
  geom_density() +
  scale_color_manual(values = viridis(3)) +
  theme_minimal()
```

```
## Warning: Removed 304 rows containing non-finite values (stat_density).
```



Second stage

```
# Without covariates
# This is somewhat wrong as we only want to use units that were not
# listwise deleted during the first stage
iv.stage2 <- lm(affirm ~ stage1_fit_nocov, data = south.counties)

summary(iv.stage2)
```

```
##
## Call:
```

```
## lm(formula = affirm ~ stage1_fit_nocov, data = south.counties)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24455 -0.20399 -0.04357  0.09923  0.80234
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.27041    0.03435   7.872 7.49e-15 ***
## stage1_fit_nocov -0.17451    0.11593  -1.505    0.132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2264 on 1259 degrees of freedom
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.001797,   Adjusted R-squared:  0.001004
## F-statistic: 2.266 on 1 and 1259 DF,  p-value: 0.1325
```

```r
# Storing my formula
iv.s2.formula <- formula(
  . ~ stage1_fit + log(coarea00) + rugged +
    latitude + I(latitude^2) + longitude +
    I(longitude^2)  + water1860  + state.abb
)

# Running 3 separate second-stage models for the 3 outcomes
iv.s2.affirm <- lm(formula = update(iv.s2.formula, affirm ~ .),
                       data = iv.data,
                       weights = sample.size)
iv.s2.dem <- lm(formula = update(iv.s2.formula, dem ~ .),
                       data = iv.data,
                weights = sample.size)
iv.s2.resent <- lm(formula = update(iv.s2.formula, resent ~ .),
                       data = iv.data,
                   weights = sample.size)
```

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
vars_order <- c("cottonsuit", "stage1_fit")

ch.row <- function(name, yesno) {
    c(name, ifelse(yesno, "$\\checkmark$", ""))
}

stargazer(iv.stage1.cov, iv.s2.affirm, iv.s2.dem, iv.s2.resent,
          keep = c("cottonsuit", "stage1_fit"),
          order = paste0("^", vars_order, "$"),
          header = FALSE,
          omit.stat = c("ll", "rsq", "adj.rsq", "ser", "f"),
          covariate.labels = c("Cotton Suitability", "Prop. Slave, 1860"),
          add.lines = list(c("F-statistic", "80.077* (df = 21; 1,098)", rep("", 4)),
                           ch.row("State Fixed Effects", rep(TRUE,4)),
                           ch.row("Geographic Controls", rep(TRUE,4)), c("Model", rep("2SLS", 4)), c(""
          dep.var.labels = c("Prop Slave", "Affirm. Action", "Prop Democrat", "Racial Resentment"))
```

Table 1:

	Dependent variable:			
	Prop Slave	Affirm. Action		
	(1)	(2)	(3)	(4)
Cotton Suitability	0.408*** (0.040)			
Prop. Slave, 1860		−0.247*** (0.086)	−0.277*** (0.101)	0.784** (0.348)
F-statistic	80.077* (df = 21; 1,098)			
State Fixed Effects	✓	✓	✓	✓
Geographic Controls	✓	✓	✓	✓
Model	2SLS First Stage	2SLS Second Stage	2SLS Second Stage	2SLS Second Stage
Observations	1,120	1,120	1,120	998
Note:				*p<0.1; **p<0.05; ***p<0.01

```
# with ivreg, in a single step
# We first store our formula
base.iv.form <- Formula(. ~ pslave1860 + log(coarea00) + rugged + latitude + I(latitude^2) + longitude

ivreg(formula = update(base.iv.form, affirm ~ .),
      data = south.counties,
      weights = sample.size) %>%
  summary()
```

```
##
## Call:
## ivreg(formula = update(base.iv.form, affirm ~ .), data = south.counties,
##     weights = sample.size)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7927 -0.4523 -0.1453  0.3461  4.2588
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.1992231  1.7355375   1.267  0.20536
## pslave1860     -0.2470599  0.0879711  -2.808  0.00507 **
## log(coarea00)  -0.0225346  0.0088914  -2.534  0.01140 *
## rugged         -0.0002098  0.0001464  -1.433  0.15203
## latitude        0.0154284  0.0464585   0.332  0.73989
## I(latitude^2)  -0.0003648  0.0007229  -0.505  0.61394
## longitude       0.0497215  0.0382835   1.299  0.19430
## I(longitude^2)  0.0002990  0.0002200   1.359  0.17430
## water1860       0.0161035  0.0093681   1.719  0.08590 .
## state.abbAR    -0.0344161  0.0327869  -1.050  0.29409
## state.abbFL    -0.0162649  0.0281239  -0.578  0.56316
## state.abbGA    -0.0090229  0.0227977  -0.396  0.69234
## state.abbKY     0.0717963  0.0317605   2.261  0.02398 *
## state.abbLA    -0.0267724  0.0318369  -0.841  0.40057
## state.abbMO     0.0146291  0.0380986   0.384  0.70107
## state.abbMS    -0.0236406  0.0314432  -0.752  0.45230
## state.abbNC     0.0782381  0.0310351   2.521  0.01184 *
## state.abbSC     0.0853653  0.0316343   2.699  0.00707 **
## state.abbTN     0.0596810  0.0239725   2.490  0.01294 *
## state.abbTX    -0.0713224  0.0420027  -1.698  0.08978 .
## state.abbVA     0.1101115  0.0454180   2.424  0.01549 *
## state.abbWV     0.0566501  0.0442834   1.279  0.20107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6327 on 1098 degrees of freedom
## Multiple R-Squared: 0.05397, Adjusted R-squared: 0.03587
## Wald test: 4.746 on 21 and 1098 DF,  p-value: 1.334e-11
```