Discussion 10

Week of 04.06

UVA Batten

Agenda

- Quiz Review
- Regression Discontinuity in STATA cont.
- FE Example (police and crime)

Announcements

- Quiz 8 due Tuesday, April 7 @ 9PM
- No class on Wednesday
- Homework 6 due April 14
- Homework 7 due April 18 (no late penalty until April 22)

Quiz Review

Question 2.6

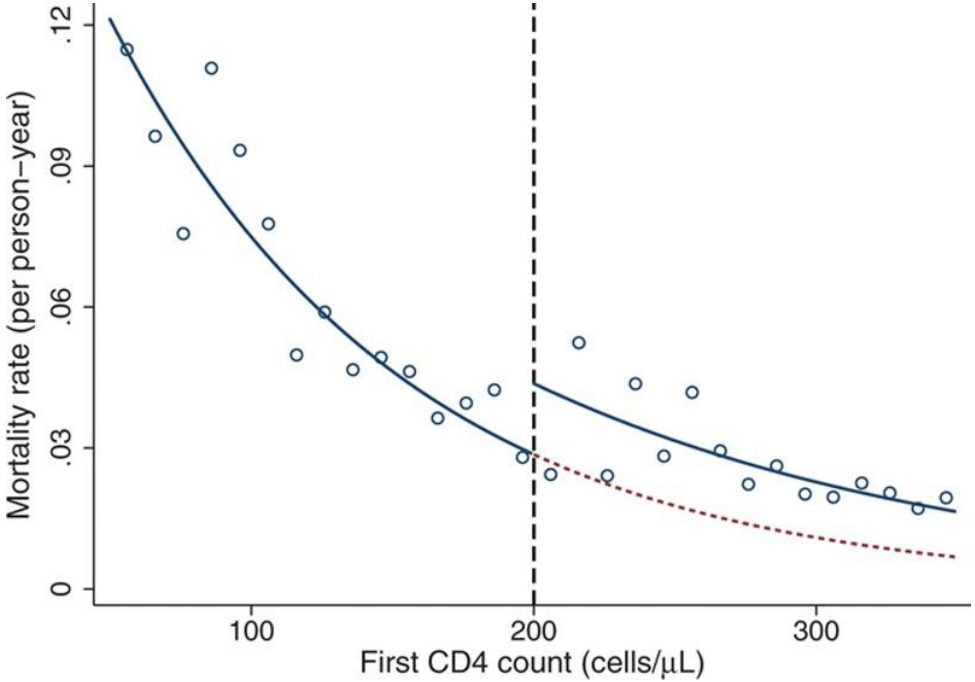
Based on the graph, which regression should we start with?

( ) $Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \varepsilon_i$

( ) $Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \boldsymbol{\beta_3 D(CD4_i-200) + \varepsilon_i}$

( ) $Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \boldsymbol{\beta_3(CD4_i-200)^2 + \varepsilon_i}$

( ) $Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \boldsymbol{\beta_3 I(CD4_i-200) + \beta_4 I(CD4_i-200)^2 + \beta_5(I(CD4_i<200))(CD4_i-200)^2 + \varepsilon_i}$



Question 2.6

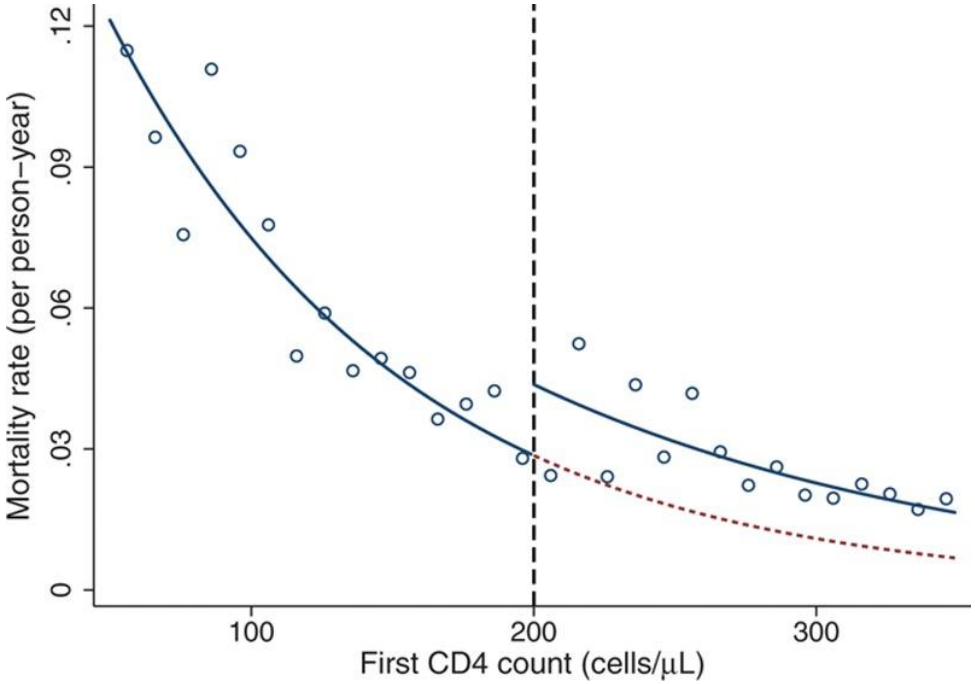
Based on the graph, which regression should we start with?

( ) $Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \varepsilon_i$

( ) $Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \boldsymbol{\beta_3 D(CD4_i-200)} + \boldsymbol{\varepsilon_i}$

(✅) $Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \boldsymbol{\beta_3(CD4_i-200)^2} + \boldsymbol{\varepsilon_i}$

( ) $Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \boldsymbol{\beta_3 I(CD4_i-200)} + \boldsymbol{\beta_4 I(CD4_i-200)^2} + \boldsymbol{\beta_5(I(CD4_i<200))(CD4_i-200)^2} + \boldsymbol{\varepsilon_i}$

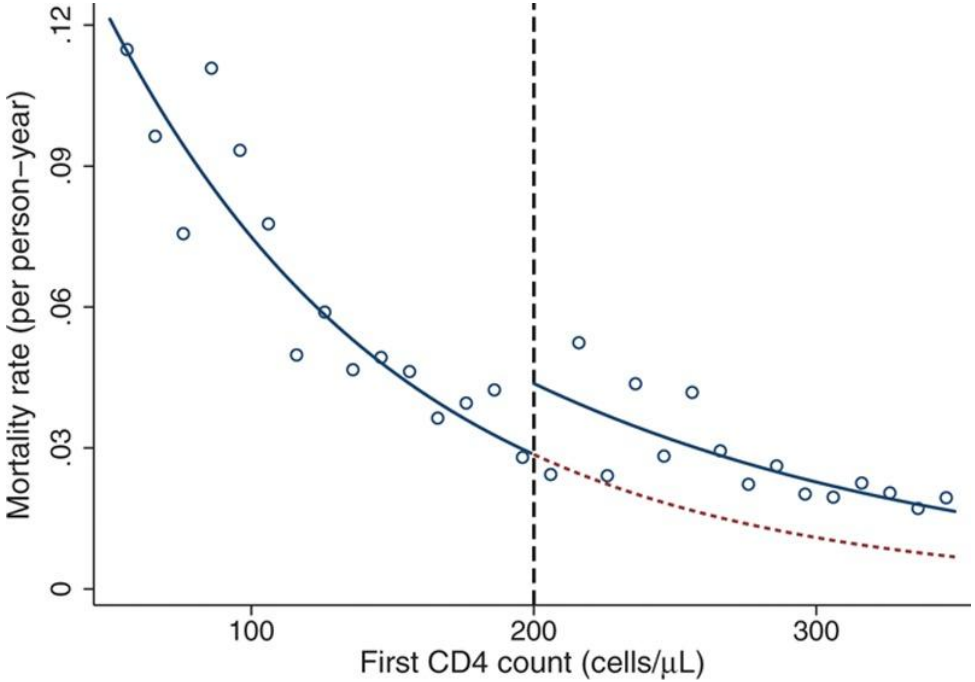


Question 2.6

Using the regression below, what should be expect the value of β_1 to be?

$Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \beta_3(CD4_i-200)^2 + \varepsilon_i$

( ) Positive
( ) Negative

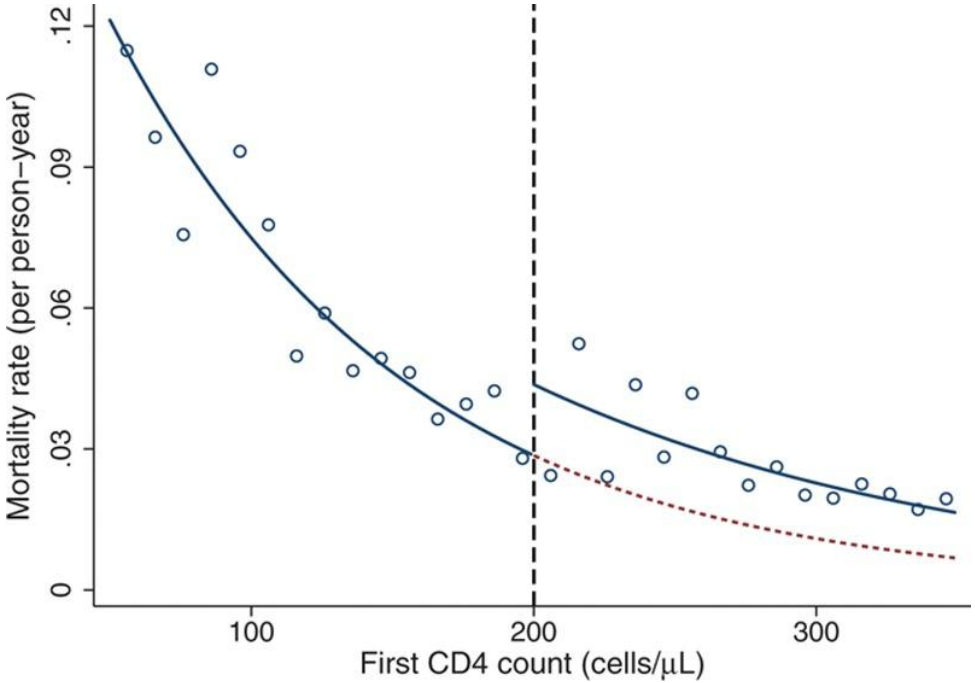


Question 2.6

Using the regression below, what should be expect the value of β_1 to be?

$Mortality = \beta_0 + \beta_1 I(CD4_i<200) + \beta_2(CD4_i-200) + \beta_3(CD4_i-200)^2 + \varepsilon_i$

( ) Positive
(✅) Negative



RD Practice in STATA (Revisit)

RD Practice

Congressional elections are decided by a clear rule: whoever gets the most votes in November wins. Because virtually every congressional race in the United States is between two parties, whoever gets more than 50 percent of the vote wins.

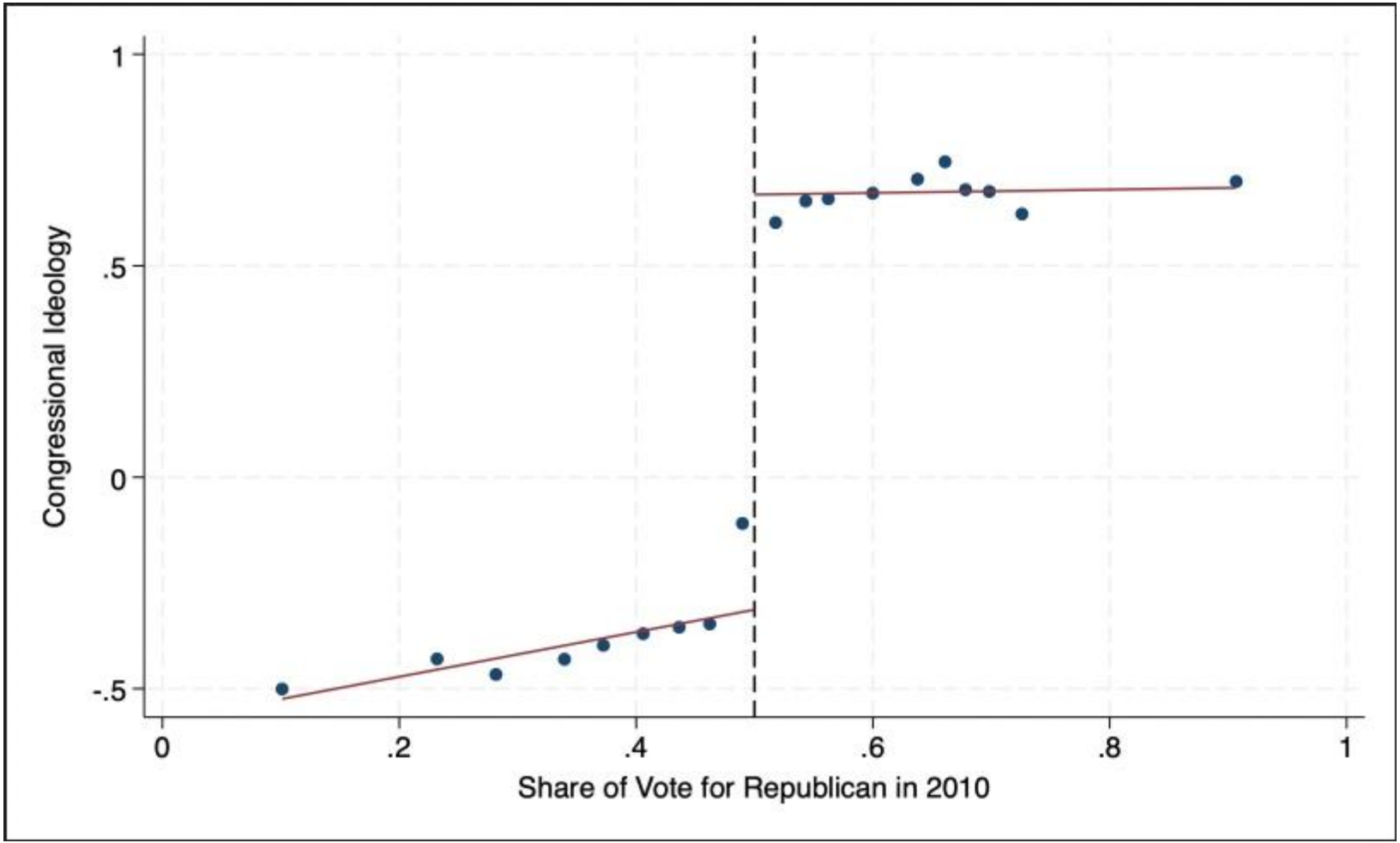
We can use this fact to estimate the effect of political party on ideology. Some argue that Republicans and Democrats are very distinctive; others argue that members of Congress have strong incentives to respond to the median voter in their districts, regardless of party.

We can assess how much party matters by looking at the ideology of members of Congress in the 112th Congress (which covered the years 2011 and 2012).

RD Practice

1. What's the outcome, RV, and cutoff? What would an RD model for this look like?

2. Using the data (found on the class website), use *binscatter* to create an RD plot of vote share and congressional ideology.


Congressional Ideology
1
.5
0
-.5
0
.2
.4
.6
.8
1
Share of Vote for Republican in 2010

RD Practice

Run the RD in Stata. What's the coefficient on our variable of interest? What does it mean?

```
. reg Ideology GOPwin2010 x_c

      Source |       SS           df       MS      Number of obs   =       432
-------------+----------------------------------   F(2, 429)       =   2685.95
       Model |  122.719927         2  61.3599634   Prob > F        =    0.0000
    Residual |  9.80042133       429  .022844805   R-squared       =    0.9260
-------------+----------------------------------   Adj R-squared   =    0.9257
       Total |  132.520348       431  .307471805   Root MSE        =    .15114

------------------------------------------------------------------------------
    Ideology | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  GOPwin2010 |   .9951733    .023815    41.79   0.000     .9483646    1.041982
         x_c |    .230444   .0571083     4.04   0.000     .1181972    .3426908
       _cons |  -.3598854   .0140958   -25.53   0.000    -.3875908     -.33218
------------------------------------------------------------------------------
```

RD Practice

In our original graph, it looks like the relationship between vote share and ideology is different depending on whether the GOP or Democrats win. Estimate that difference in Stata (hint: you'll need to create a new variable!)

```
. reg Ideology GOPwin2010 x_c rv_share

      Source |       SS           df       MS      Number of obs   =       432
-------------+----------------------------------   F(3, 428)       =   1868.39
       Model |  123.119182         3  41.0397274   Prob > F        =    0.0000
    Residual |  9.40116584       428  .021965341   R-squared       =    0.9291
-------------+----------------------------------   Adj R-squared   =    0.9286
       Total |  132.520348       431  .307471805   Root MSE        =    .14821

------------------------------------------------------------------------------
    Ideology | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  GOPwin2010 |   .9815657   .0235692    41.65   0.000     .9352399    1.027892
         x_c |   .5286074   .0895923     5.90   0.000     .3525118     .704703
    rv_share |  -.4893277    .114774    -4.26   0.000    -.7149184    -.263737
       _cons |  -.3132725   .0176232   -17.78   0.000    -.3479113   -.2786337
------------------------------------------------------------------------------
```

RD Practice

Now run the same regression for child poverty, median income, percent white, and 2008 Obama vote share. What do these results tell us?

RD Practice

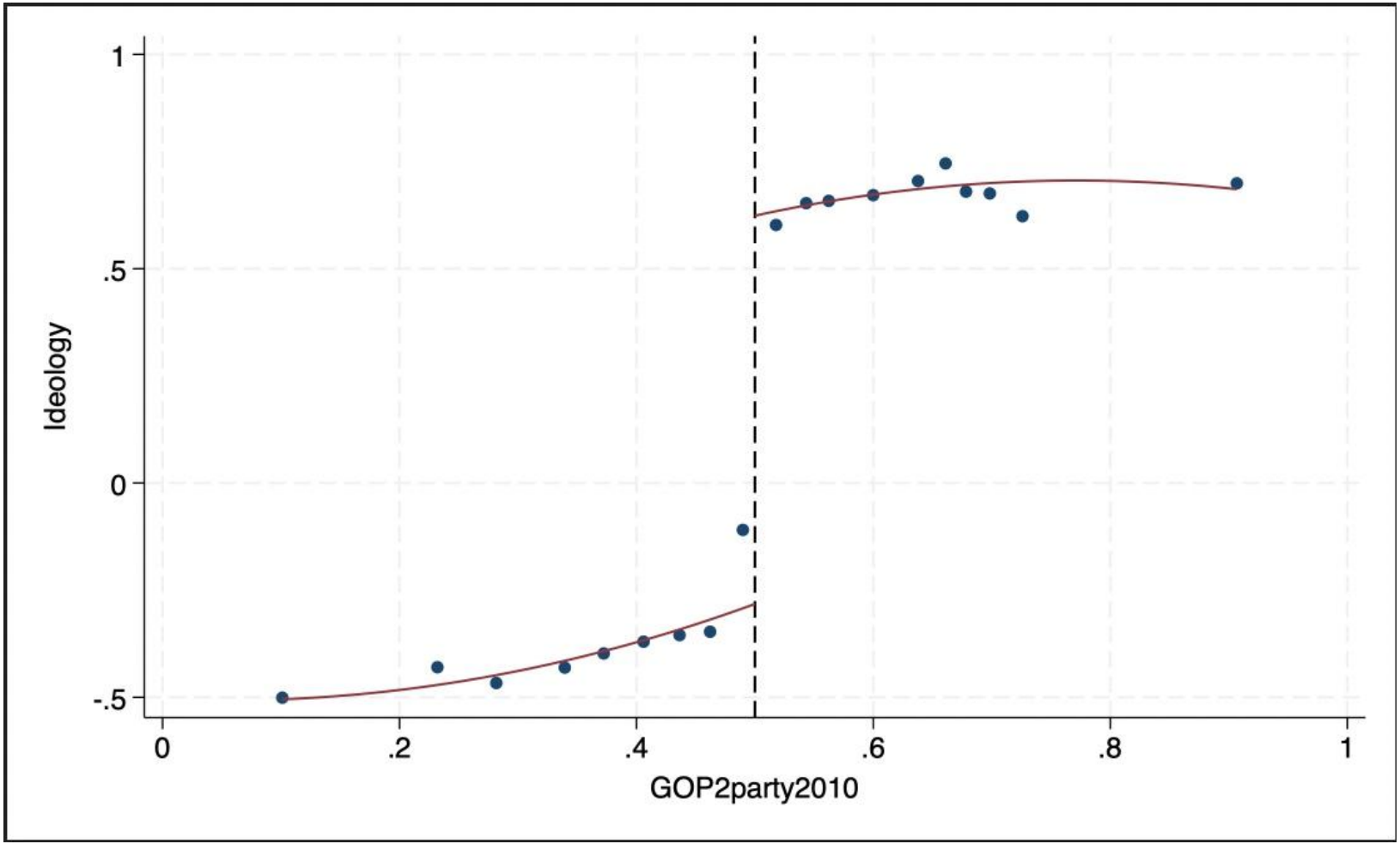
Run the original regression, but include the variables we tested previously as controls. Interpret the results.

```
. reg Ideology GOPwin2010 x_c rv_share ChildPoverty MedianIncome WhitePct Obama2008

      Source |       SS           df       MS      Number of obs   =       432
-------------+----------------------------------   F(7, 424)       =    879.56
       Model |  123.982228         7  17.7117469   Prob > F        =    0.0000
    Residual |  8.53811988       424  .020137075   R-squared       =    0.9356
-------------+----------------------------------   Adj R-squared   =    0.9345
       Total |  132.520348       431  .307471805   Root MSE        =    .14191

------------------------------------------------------------------------------
    Ideology | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  GOPwin2010 |   .9627815   .0228694    42.10   0.000     .9178299    1.007733
         x_c |  -.0476237   .1416603    -0.34   0.737    -.3260676    .2308202
    rv_share |  -.1770715   .1365759    -1.30   0.196    -.4455216    .0913786
ChildPoverty |   .2651014   .1885303     1.41   0.160     -.105469    .6356719
MedianIncome |   2.12e-06   1.01e-06     2.10   0.036     1.38e-07    4.10e-06
   WhitePct  |  -.0608577   .0497801    -1.22   0.222    -.1587041    .0369887
   Obama2008 |  -.7905298   .1242559    -6.36   0.000    -1.034764   -.5462955
       _cons |  -.0134057    .127232    -0.11   0.916    -.2634897    .2366784
------------------------------------------------------------------------------
```

RD Practice

What would our model look like if we used a quadratic fit?

Write the regression down, then try it out in Stata. Start by using binscatter to see how it looks, then run the regression.


1
.5
0
-.5
Ideology
0
.2
.4
.6
.8
1
GOP2party2010

```
. reg Ideology GOPwin2010 x_c x_csq rv_share rv_share_sq
```

Source	SS	df	MS		
Model	123.336513	5	24.6673026	Number of obs	= 432
Residual	9.18383526	426	.021558299	F(5, 426)	= 1144.21
				Prob > F	= 0.0000
Total	132.520348	431	.307471805	R-squared	= 0.9307
				Adj R-squared	= 0.9299
				Root MSE	= .14683

Ideology	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
GOPwin2010	.9058469	.033603	26.96	0.000	.8397986	.9718953
x_c	1.006793	.2751797	3.66	0.000	.465914	1.547672
x_csq	1.126757	.6137553	1.84	0.067	-.0796089	2.333123
rv_share	-.4029129	.3581573	-1.12	0.261	-1.106888	.3010626
rv_share_sq	-2.238689	.7489567	-2.99	0.003	-3.710799	-.7665781
_cons	-.2821135	.0243494	-11.59	0.000	-.3299734	-.2342536

Fixed Effects

Panel data follows units of observation across many time periods.

- Examples: monthly employment rates by state, yearly test scores by school

Dimensions:

i: our panel variable, or unit of observation – e.g. states, households, classrooms

j: our time variable, e.g. month, year

Fixed effects help account for observable and unobservable characteristics in our units.

- They allow us to look within units – e.g. what is the relationship between X and Y for every state for every year
- Two-way fixed effects account for differences in both our panel and time variables.

In order for there to be an effect, there should still be variation within units – e.g. within a year or a state.

Example: Police and Crime

Say we have data on police officers per capita and burglaries per capita across 59 major cities in the US.

Why might we not want to use the following regression?

$$Crime_{it} = \beta_0 + \beta_1 Police_{i,t-1}$$

Fixed effects regressions account for all of the time or panel-level variation that exists.

A fixed effects regression using city-year panel data might look like:

$$Crime_{it} = \beta_0 + \beta_1 Police_{i,t-1} + a_i + \varepsilon_i$$

You can think of a_i as including a dummy control variable for every city.

$$Crime_{it} = \beta_0 + \beta_1 Police_{i,t-1} + Charlottesville_i + Richmond_i + DC_i + \varepsilon_i$$

Stata has three ways to do fixed effects:

1) **Using the i.panel approach**

```
reg y x i.panel i.time
```

* if your data is at the individual level, you may need to collapse it

**use i.varname for categorial variables and c.varname for continuous variables

2) **Using xtreg**

```
xtset panel time
xtreg y x, fe
```

3) **Using areg**

```
areg y x, a(panel)
```

Let’s Try it Out

In Stata, run `webuse nlswork` **to open the dataset.**

Now let’s collapse the variables by year:

```
collapse wks_work hours tenure ttl_exp, by(year)
```

This gives us annual averages of each variable. Now, if we want to see how hours worked changes year-to-year, we can run:

```
reg hours i.year, base
```

Let’s Try it Out

What if we want to account for individual time-invariant characteristics?

```
xtset id year

xtreg hours i.year, fe
```