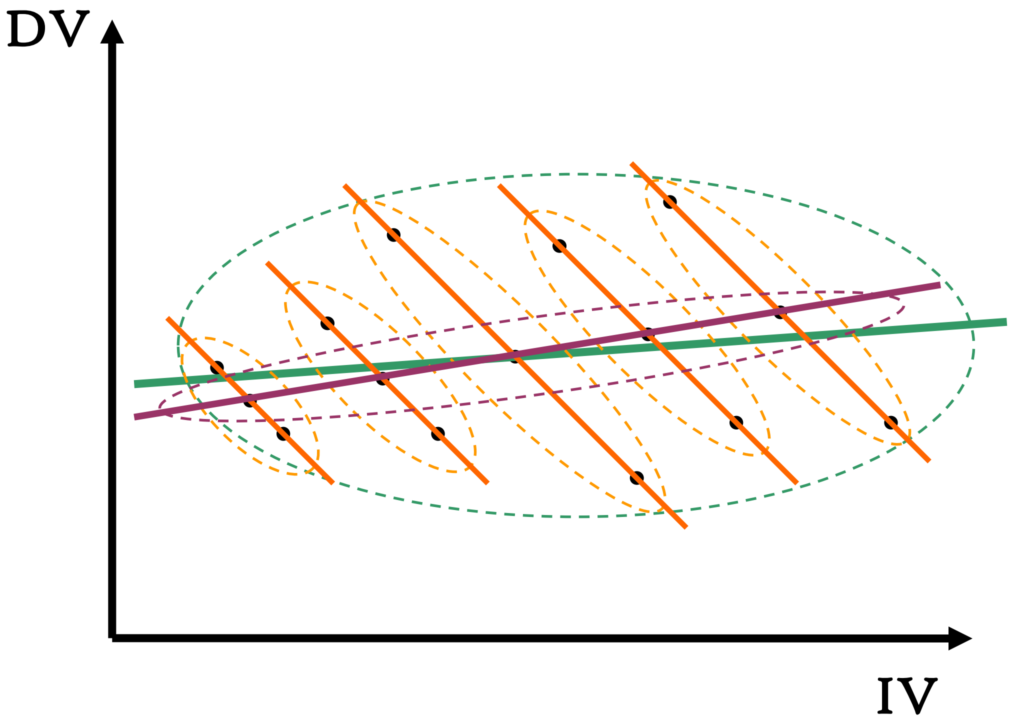
# Special topics

# Fixed effects

- When trying to compare apples with apples, we worry about the numerous potential differences on confounding variables
- If differences on confounding variables are stable over time, we can eliminate bias from them by only analyzing variation within the same unit over time
  - E.g., breast-feeding study (unit is woman)
  - E.g., currency unions and euro (unit is country)
- To only analyze variation within the same unit over time, we use fixed effects
  - Stata commands `areg` and `xtreg`
  - Equivalent to adding indicator (or dummy variables) variables for units
  - Equivalent to between subjects design (as opposed to within subjects)

# Intra- or Inter-country Variation?

(animated slide, see summary on next slide)




Inter-country Variation

xtreg [DV] [IV], be

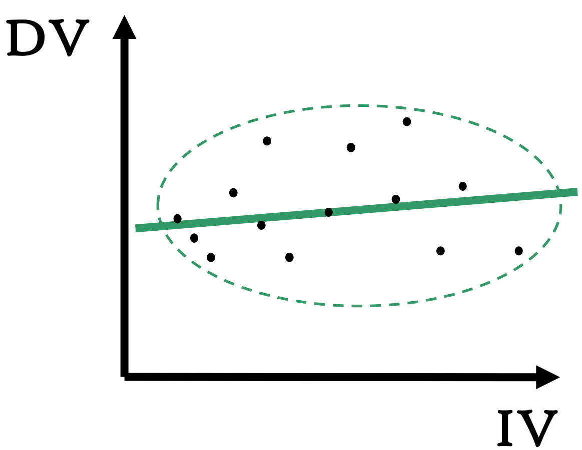
(or)

collapse [DV] [IV], by(country)
reg [DV] [IV]

# Intra- or Inter-country Variation?

Aggregate Panel Variation

*reg [DV] [IV]*



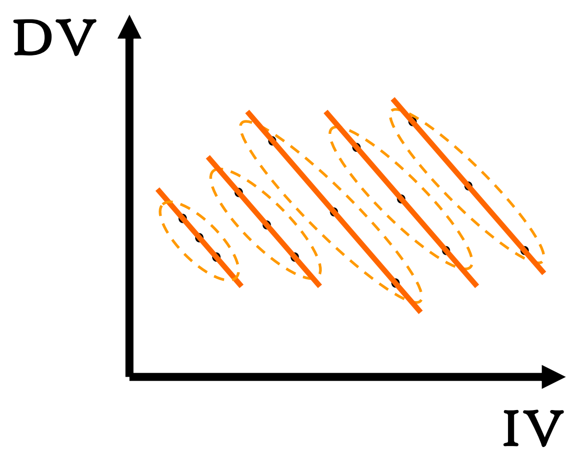

**Fixed effects** (fe)

Intra-Country Variation

*xtreg [DV] [IV], fe*

*(or)*

*areg [DV] [IV], a(country)*




**Between effects** (be)

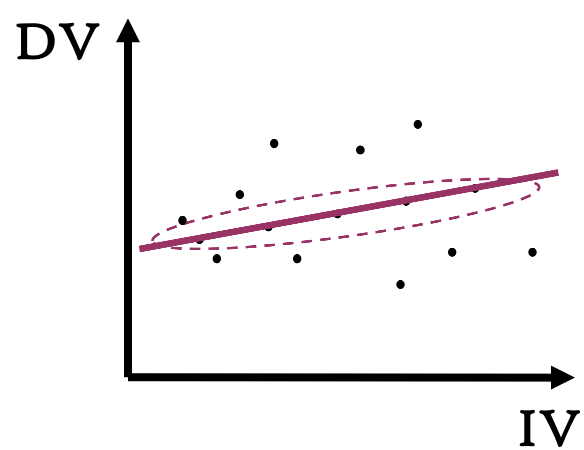
Inter-Country Variation

*xtreg [DV] [IV], be*

*(or)*

*collapse [DV] [IV], by(country)*

*reg [DV] [IV]*



# Fixed effects

- Problems
  - Throws away potentially relevant variation (alternative: random effects)
  - Variation over time may be primarily from random measurement error (e.g., unions and wages)
  - Unusual factors may drive changes in explanatory variables over time and also influence the dependent variable (e.g., currency unions)

# Interactions

# Interactions

- Interactions test whether the combination of variables affects the outcome differently than the sum of the main (or individual) effects.
- For example, how would we test whether defendants are sentenced to death more frequently for killing white strangers than you would expect from the coefficients on white victim and on victim stranger?

```
. tab wv vs

           |          vs
        wv |         0          1 |     Total
-----------+----------------------+----------
         0 |        12         14 |        26
         1 |        37         37 |        74
-----------+----------------------+----------
     Total |        49         51 |       100
```

# Interactions

```
. g wvXvs = wv* vs
. reg death bd yv ac fv v2 ms wv vs wvXvs
----------------------------------------------
 death |       Coef.   Std. Err.      t   P>|t|
-------+--------------------------------------
 (omitted)
    wv |   .0985493    .1873771    0.53   0.600
    vs |   .1076086    .2004193    0.54   0.593
 wvXvs |   .3303334    .2299526    1.44   0.154
 _cons |   .0558568    .2150039    0.26   0.796
----------------------------------------------
```

- To interpret interactions, substitute the appropriate values for each variable
- E.g., what's the effect for
  - .099 wv+.108 vs+.330 wvXvs
  - White, non-stranger: .099(1)+.108(0)+.330(1)*(0) = .099
  - White, stranger: .099(1)+.108(1)+.330(1)*(1) = .537
  - Black, non-stranger: .099(0)+.108(0)+.330(0)*(0) = comparison
  - Black, stranger: .099(0)+.108(1)+.330(1)*(0) = .108

# Interactions

```
. tab wv vs, sum(death)

Means, Standard Deviations and Frequencies of death

           |          vs
        wv |         0          1 |     Total
-----------+----------------------+----------
         0 | .16666667  .28571429 | .23076923
           | .38924947  .46880723 | .42966892
           |        12         14 |        26
-----------+----------------------+----------
         1 | .40540541  .75675676 | .58108108
           | .49774265  .43495884 |  .4967499
           |        37         37 |        74
-----------+----------------------+----------
     Total | .34693878  .62745098 |       .49
           | .48092881  .48829435 | .50241839
           |        49         51 |       100
```

# Importance of a variable

# Death penalty example

```
. sum death bd- yv

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |       100         .49    .5024184          0          1
          bd |       100         .53    .5016136          0          1
          wv |       100         .74     .440844          0          1
          ac |       100    .4366667     .225705          0          1
          fv |       100         .31    .4648232          0          1
-------------+--------------------------------------------------------
          vs |       100         .51    .5024184          0          1
          v2 |       100         .14    .3487351          0          1
          ms |       100         .12    .3265986          0          1
          yv |       100         .08    .2726599          0          1
```

# Death penalty example

```
. reg death bd-yv , beta
```

| death | Coef. | Std. Err. | P>\|t\| | Beta |
|---|---|---|---|---|
| bd | -.0869168 | .1102374 | 0.432 | -.0867775 |
| wv | .3052246 | .1207463 | 0.013 | .2678175 |
| ac | .4071931 | .2228501 | 0.071 | .1829263 |
| fv | .0790273 | .1061283 | 0.458 | .0731138 |
| vs | .3563889 | .101464 | 0.001 | .3563889 |
| v2 | .0499414 | .1394044 | 0.721 | .0346649 |
| ms | .2836468 | .1517671 | 0.065 | .1843855 |
| yv | .050356 | .1773002 | 0.777 | .027328 |
| _cons | -.1189227 | .1782999 | 0.506 | . |

# Importance of a variable

- Three potential answers
  - Theoretical importance
  - Level importance
  - Dispersion importance

# Importance of a variable

- Theoretical importance
  - Theoretical importance = Regression coefficient (b)
  - To compare explanatory variables, put them on the same scale
    - E.g., vary between 0 and 1

# Importance of a variable

- Level importance: most important in particular times and places
  - E.g., did the economy or presidential popularity matter more in congressional races in 2006?
  - Level importance= $b_j * x_j$

# Importance of a variable

- Dispersion importance: what explains the variance on the dependent variable
  - E.g., given that the GOP won in this particular election, why did some people vote for them and others against?
  - Dispersion importance =
    - Standardized coefficients, or alternatively
    - Regression coefficient times standard deviation of explanatory variable
    - In bivariate case, correlation

# Which to use?

- Depends on the research question
  - Usually theoretical importance
  - Sometimes level importance
  - Dispersion importance not usually relevant

# Partial residual scatter plots

# Partial residual scatter plots

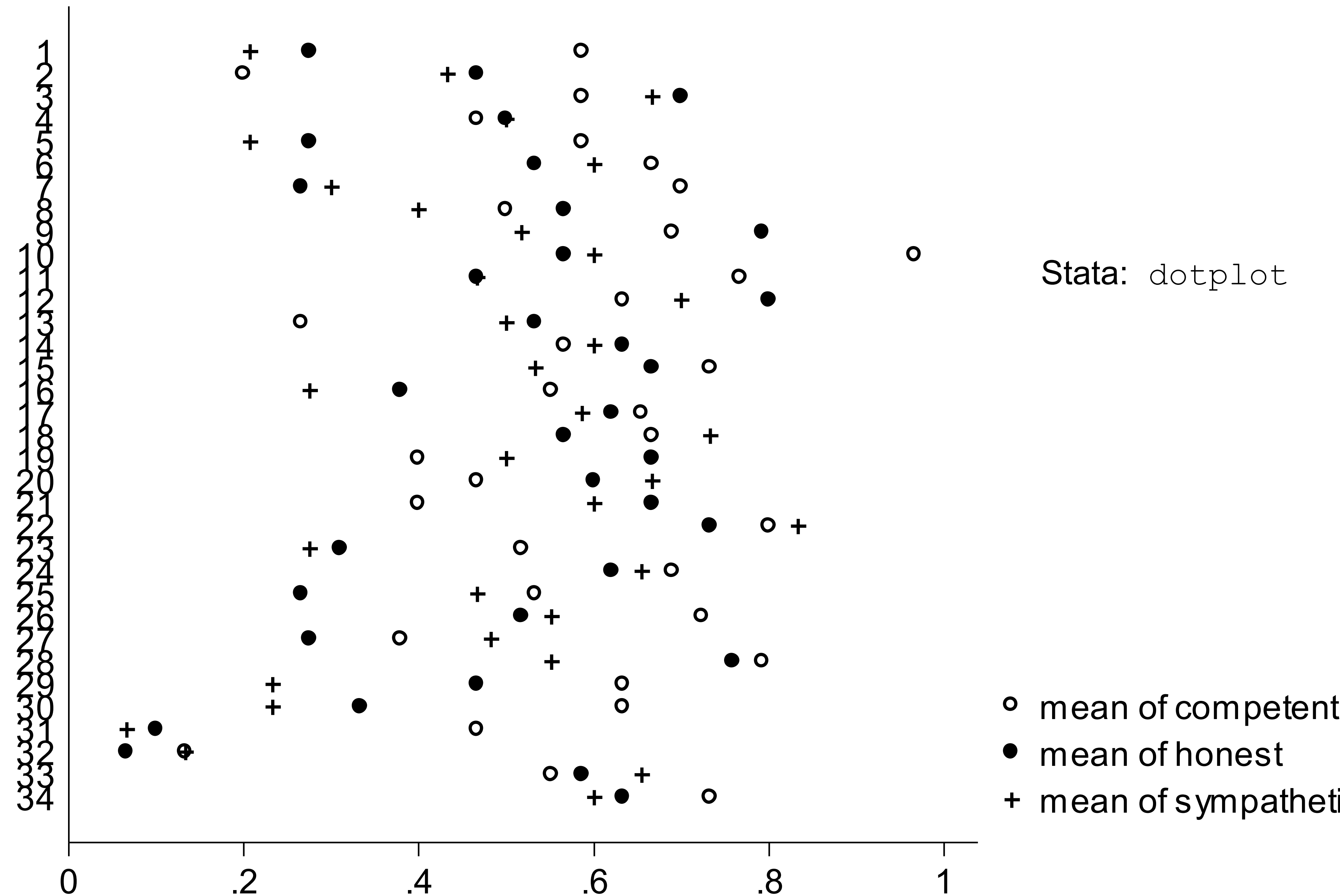
- Importance of plotting your data
- Importance of controls
- How do you plot your data after you've adjusted it for control variables?
- Example: inferences about candidates in Mexico from faces

# Greatest competence disparity: pairing 10

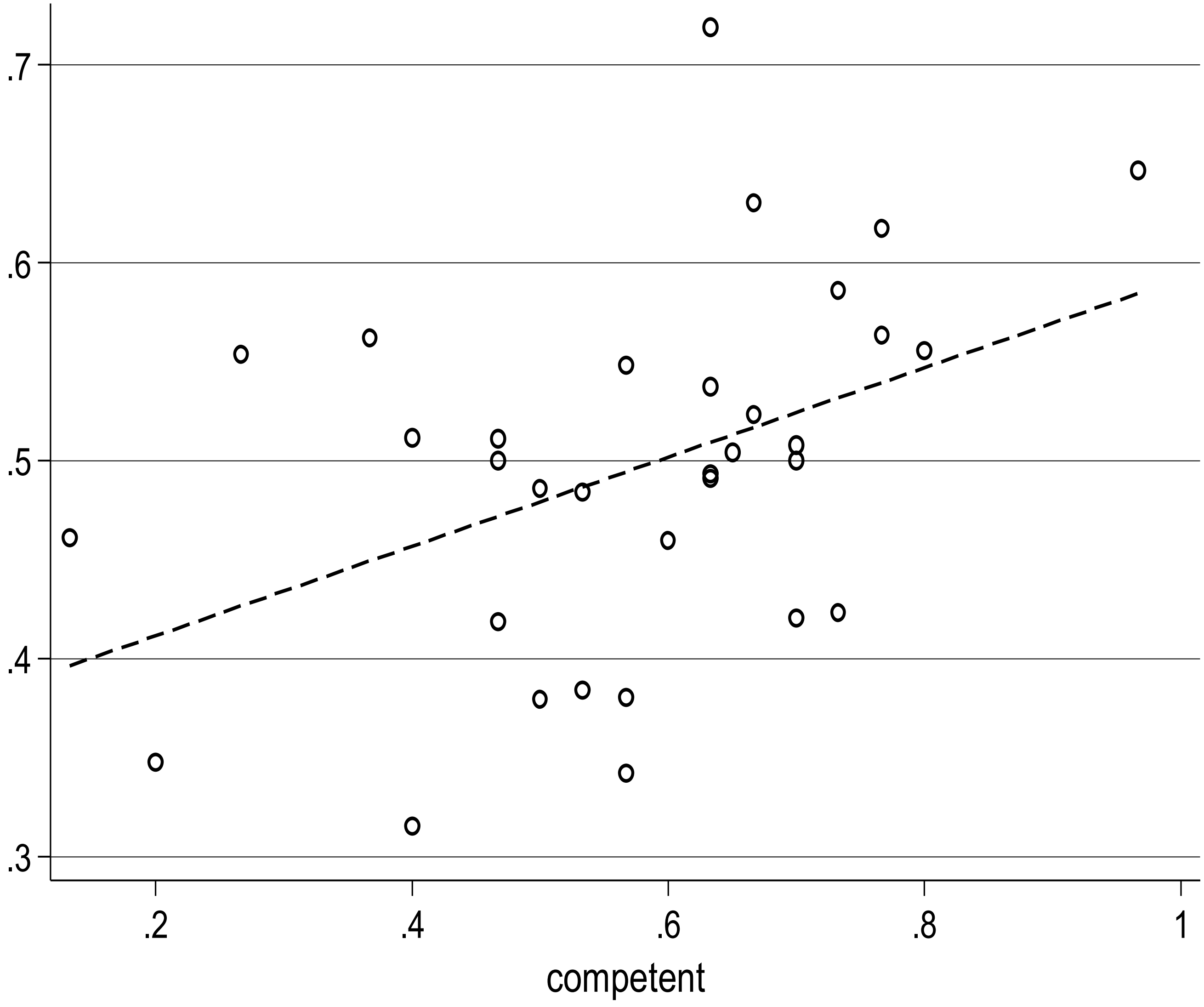
A

B

- Gubernatorial race
- A more competent
- Who won?
  - A by 65%



Stata: dotplot
mean of competent
mean of honest
mean of sympatheti
0
.2
.4
.6
.8
1



.7
.6
.5
.4
.3
.2
.4
.6
.8
1
competent

# Regression

```
      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  3,    29) =    4.16
       Model |  .082003892     3  .027334631           Prob > F      =  0.0144
    Residual |  .190333473    29  .006563223           R-squared     =  0.3011
-------------+------------------------------           Adj R-squared =  0.2288
       Total |  .272337365    32  .008510543           Root MSE      =  .08101

------------------------------------------------------------------------------
      vote_a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   competent |   .1669117   .0863812     1.93   0.063    -.0097577     .343581
   incumbent |   .0110896   .0310549     0.36   0.724    -.0524248     .074604
     party_a |   .2116774   .1098925     1.93   0.064     -.013078    .4364327
       _cons |   .2859541   .0635944     4.50   0.000     .1558889    .4160194
------------------------------------------------------------------------------
```

- vote_a is vote share for Candidate A
- incumbent is a dummy variable for whether the party currently holds the office
- party_a is the vote share for the party of Candidate A in the previous election

- We want to create a scatter plot of vote_a by competent controlling for incumbent and party_a

# Calculating partial residuals

First run your regression with all the relevant variables

```
. reg vote_a competent incumbent party_a
```

To calculate the residual for the full model, use

```
. predict e, res
```

(This creates a new variable “e”, which equals to the residual.)

```
Source |
------------+-------
```

Here, however, we want to generate the residual controlling <u>only</u> for some variables. To do this, we could manually predict vote_a based only on incumbent and party_a:

```
. g y_hat = (0)*.167+ incumbent*.011 + party_a*.212
```

We can then generate the partial residual

```
. g partial_e = vote_a - y_hat
```

Instead, can use the Stata adjust

```
. adjust competent = 0, by(incumbent party_a) gen(y_hat)
. g partial_e = vote_a - y_hat
```

# Calculating partial residuals

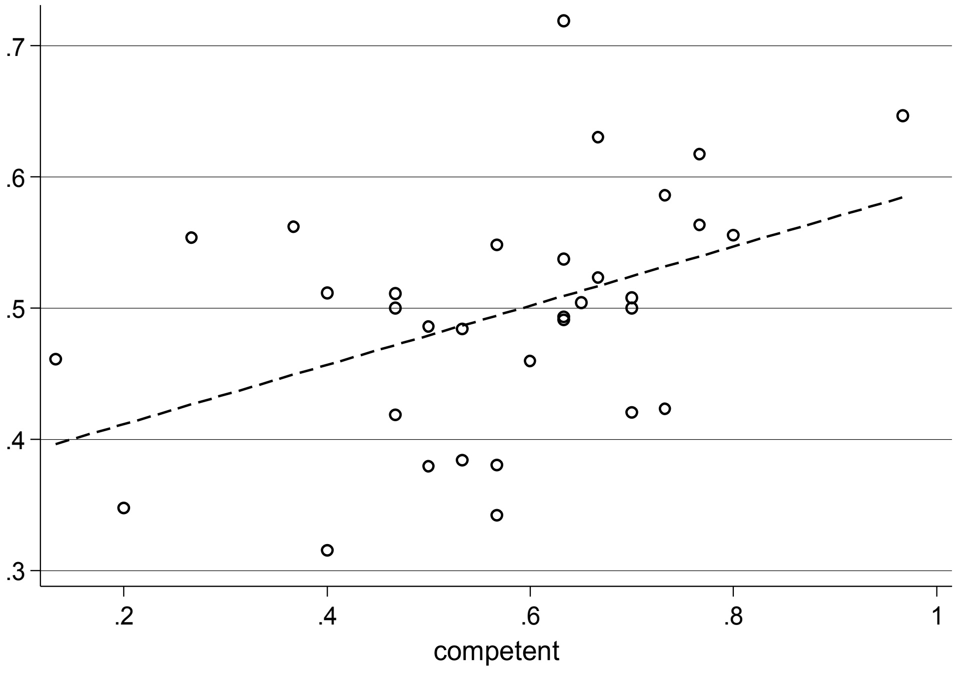
- Regression of the partial residual on competent. (Should not necessarily give you the same coefficient estimate because competent is not residualized.)

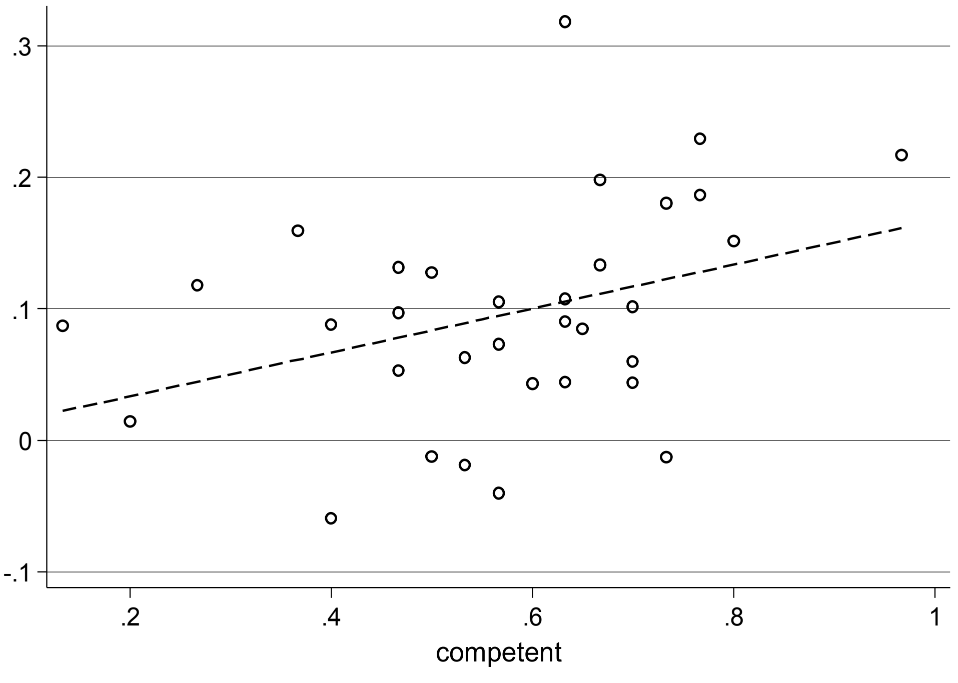
```
. reg partial_e competent
      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  1,    31) =    4.52
       Model |  .027767147     1  .027767147           Prob > F      =  0.0415
    Residual |  .190333468    31  .006139789           R-squared     =  0.1273
-------------+------------------------------           Adj R-squared =  0.0992
       Total |  .218100616    32  .006815644           Root MSE      =  .07836

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   competent |   .1669117    .078487     2.13   0.042     .0068364     .326987
       _cons |  -7.25e-09   .0470166    -0.00   1.000    -.0958909    .0958909
------------------------------------------------------------------------------
```

- Compare scatter plot (top) with residual scatter plot (bottom)

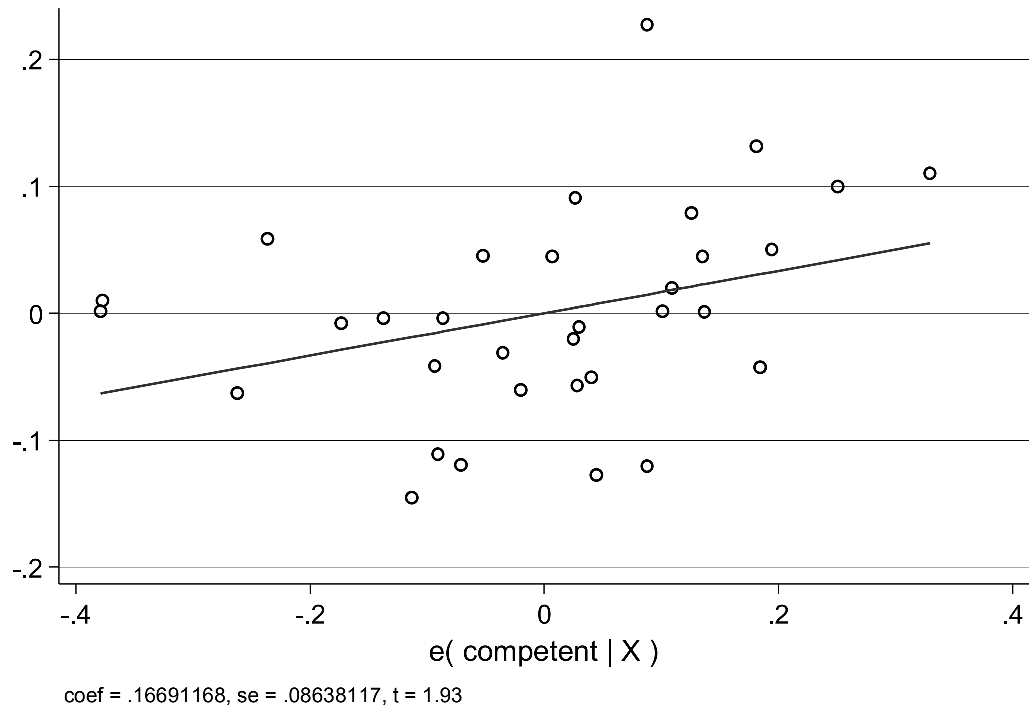
- Residual plots especially important if results change when adding controls



# Avplot & cprplot

- You can also use `avplot` to generate residual scatter plots
- After you run your regression use the following command
  - `avplot competent`
- Unlike the method above, `avplot` also conditions (residualizes) your explanatory variable
- Good for detecting outliers
- Bad for detecting functional form
- For functional form, use `cprplot`, which does not residualize the explanatory variable



# Imputing missing data (on controls

# Imputing missing data

- Variables often have missing data
- Sources of missing data
- Missing data
  - May reduce estimate precision (wider confidence intervals b/c smaller sample)
  - May bias estimates if data is not missing a random
- To rescue data with missing cases on control variables: impute using other variables
- Imputing data can
  - Increase sample size and so increase precision of estimates
  - Reduce bias if data is not missing at random

# Imputation example

- Car ownership in 1948
- Say that some percentage of sample forgot to answer a question about whether they own a car
- The data set contains variables that predict car ownership: family_income, family_size, rural, urban, employed

# Stata imputation command

- `impute depvar varlist [weight] [if exp] [in range], generate(newvar1)`
  - depvar is the variable whose missing values are to be imputed.
  - varlist is the list of variables on which the imputations are to be based
  - newvar1 is the new variable to contain the imputations

- Example
  - `impute own_car family_income family_size rural suburban employed, g(i_own_car)`

# Rules about imputing

- Before you estimate a regression model, use the summary command to check for missing data
- Before you impute, check that relevant variables actually predict the variable with missing values (use regression or other estimator)
- <u>Don’t use your studies’ dependent variable or key explanatory variable to make the imputation’s (exceptions)</u>
  - Use demographic variables
  - Use variables exogenous to the dependent variable or key explanatory variables
- <u>Don’t impute missing values on your studies’ dependent variable or key explanatory variable (exceptions)</u>
- Always note whether imputation changed results
- If too much data is missing, imputation won’t help