Multivariate regression, part b

(Multi)collinearity

- Collinearity is when two or more independent variables are highly correlated
  - More precisely: when one independent variable is a linear combination of the other independent variables
- Effects:
  - Coefficients of the affected variables may will be unstable
  - Standard errors (for after spring break) will be inflated
  - BUT, the overall predictive power of the regression will not be affected

Simplest example: Sex

- Say I have a variable named **female**, coded 1 if the respondent is female, 0 if the respondent is male
- A second variable named male coded 1 if the respondent is male, 0 if the respondent is female is ***collinear*** with **female** because

  male = 1-female

What the data matrix looks like

constant	obama_approve	Female	male
1	5	1	0
1	4	1	0
1	1	0	1
1	3	0	1
1	4	1	0
…	…	…	…
1	5	1	1

What the estimation matrix looks like

$\boldsymbol{B} = (X'X)^{-1}X'y$

What the estimation matrix looks like

$$\boldsymbol{B} = (X'X)^{-1}X'y$$

1	1	0
1	1	0
1	0	1
1	0	1
1	1	0
…	…	…
1	1	0

What the estimation matrix looks like

$$\boldsymbol{B} = (X'X)^{-1}X'y$$

1	1	1	1	1	…	1
1	1	0	0	1	…	1
0	0	1	1	0	…	0

1	1	0
1	1	0
1	0	1
1	0	1
1	1	0
…	…	…
1	1	0

What the estimation matrix looks like

$$\boldsymbol{B} = (X'X)^{-1}X'y$$

$$\left(\begin{matrix} 1 & 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & 0 & 1 & \dots & 1 \\ 0 & 0 & 1 & 1 & 0 & \dots & 0 \end{matrix} \quad \begin{matrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ \dots & \dots & \dots \\ 1 & 1 & 0 \end{matrix}\right)^{-1}$$

What the estimation matrix looks like

$$\boldsymbol{B} = (X'X)^{-1}X'y$$

$$\left(\begin{matrix} 1 & 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & 0 & 1 & \dots & 1 \\ 0 & 0 & 1 & 1 & 0 & \dots & 0 \end{matrix} \quad \begin{matrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ \dots & \dots & \dots \\ 1 & 1 & 1 \end{matrix} \right)^{-1} = \begin{pmatrix} N & \sum F & N - \sum F \\ \sum F & \sum F & 0 \\ N - \sum F & 0 & N - \sum F \end{pmatrix}^{-1}$$

Without the redundant variable

```
. reg approve female  [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  1,   975) =    5.70
       Model |  8.42953813     1  8.42953813           Prob > F      =  0.0172
    Residual |  1442.49523   975  1.47948229           R-squared     =  0.0058
-------------+------------------------------           Adj R-squared =  0.0048
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.2163

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .1858039   .0778409     2.39   0.017     .0330489    .3385589
       _cons |   2.292257   .0555348    41.28   0.000     2.183275    2.401238
------------------------------------------------------------------------------
```

Without the redundant variable

```
. reg approve female  [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  1,   975) =    5.70
       Model |  8.42953813     1  8.42953813           Prob > F      =  0.0172
    Residual |  1442.49523   975  1.47948229           R-squared     =  0.0058
-------------+------------------------------           Adj R-squared =  0.0048
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.2163

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .1858039   .0778409     2.39   0.017     .0330489    .3385589
       _cons |   2.292257   .0555348    41.28   0.000     2.183275    2.401238
------------------------------------------------------------------------------
```

When the respondent is male, *female = 0,* therefore
E(approve_obama) = 2.29 + 0*.19 = 2.29

When the respondent is female, female = 1therefore
E(*approve_obama)* = 2.29 + 1*.19 = 2.48

With the redundant variable

```
. reg approve female  male [aw=v102]
(sum of wgt is   9.5361e+02)
note: male omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  1,   975) =    5.70
       Model |  8.42953813     1  8.42953813           Prob > F      =  0.0172
    Residual |  1442.49523   975  1.47948229           R-squared     =  0.0058
-------------+------------------------------           Adj R-squared =  0.0048
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.2163

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .1858039   .0778409     2.39   0.017     .0330489    .3385589
        male |          0  (omitted)
       _cons |   2.292257   .0555348    41.28   0.000     2.183275    2.401238
------------------------------------------------------------------------------
```

With a fake collinear version of female

```
. gen female2=female

. replace female2=0 if female==1&_n<=50
(15 real changes made)


. tab female female2

           |        female2
    female |         0          1 |     Total
-----------+----------------------+----------
         0 |       461          0 |       461
         1 |        15        524 |       539
-----------+----------------------+----------
     Total |       476        524 |     1,000


. corr female female2
(obs=1000)

             |   female  female2
-------------+------------------
      female |   1.0000
     female2 |   0.9703   1.0000
```

With a fake collinear version of female

```
. reg approve_o female2 [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  1,   975) =    2.86
       Model |  4.23969796     1  4.23969796           Prob > F      =  0.0913
    Residual |  1446.68507   975  1.48377956           R-squared     =  0.0029
-------------+------------------------------           Adj R-squared =  0.0019
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.2181

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     female2 |   .1317574   .0779457     1.69   0.091    -.0212033     .284718
       _cons |   2.321656   .0548207    42.35   0.000     2.214075    2.429236
------------------------------------------------------------------------------
```

.1858039 originally

With a fake collinear version of female

```
. reg approve_o female female2 [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  2,   974) =    6.41
       Model |  18.8520935     2  9.42604675           Prob > F      =  0.0017
    Residual |  1432.07268   974  1.47030049           R-squared     =  0.0130
-------------+------------------------------           Adj R-squared =  0.0110
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.2126

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    1.03613   .3286674     3.15   0.002     .3911523    1.681108
     female2 |   -.874974   .3286329    -2.66   0.008    -1.519884   -.2300639
       _cons |   2.292257   .0553622    41.40   0.000     2.183614      2.4009
------------------------------------------------------------------------------
```

.1858039 originally

With a fake collinear version of female

```
. reg approve_o female female2 [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  2,   974) =    6.41
       Model |  18.8520935     2  9.42604675           Prob > F      =  0.0017
    Residual |  1432.07268   974  1.47030049           R-squared     =  0.0130
-------------+------------------------------           Adj R-squared =  0.0110
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.2126

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    1.03613   .3286674     3.15   0.002     .3911523    1.681108
     female2 |   -.874974   .3286329    -2.66   0.008    -1.519884   -.2300639
       _cons |   2.292257   .0553622    41.40   0.000     2.183614      2.4009
------------------------------------------------------------------------------
```

.1858039 originally

Recall that every coded as follows is a “real” female:
female ==1 & female2==1

With a fake collinear version of female

```
. reg approve_o female female2 [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  2,   974) =    6.41
       Model |  18.8520935     2  9.42604675           Prob > F      =  0.0017
    Residual |  1432.07268   974  1.47030049           R-squared     =  0.0130
-------------+------------------------------           Adj R-squared =  0.0110
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.2126

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    1.03613   .3286674     3.15   0.002     .3911523    1.681108
     female2 |   -.874974   .3286329    -2.66   0.008    -1.519884   -.2300639
       _cons |   2.292257   .0553622    41.40   0.000     2.183614      2.4009
------------------------------------------------------------------------------
```

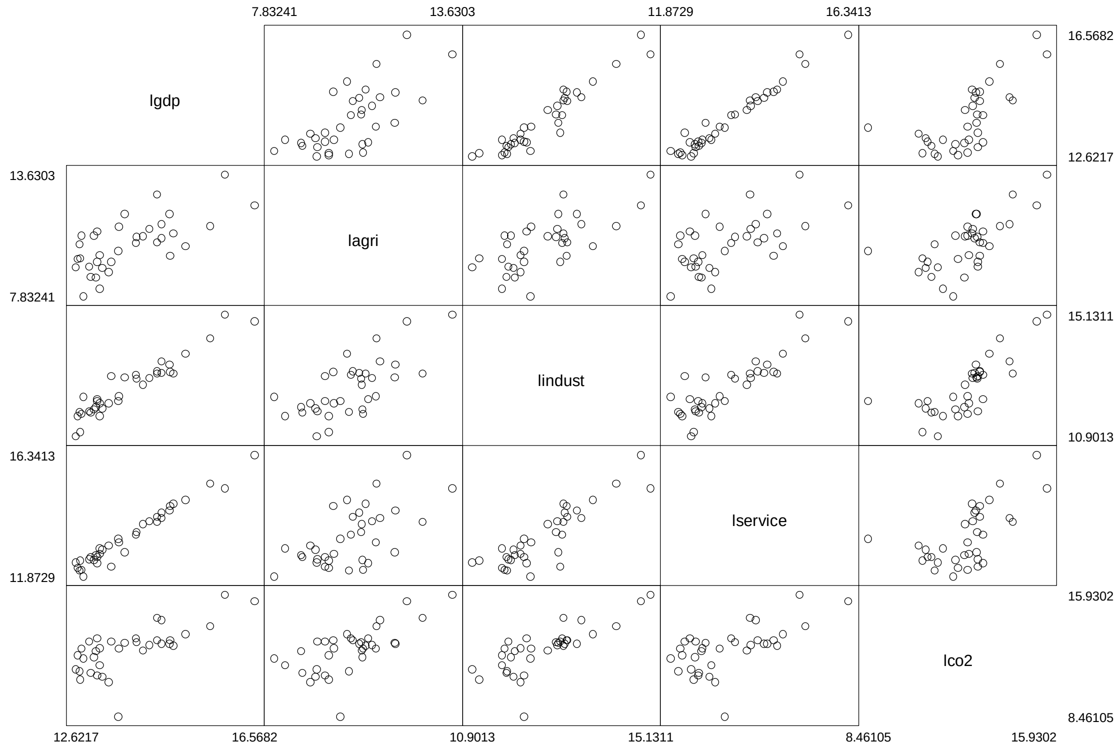
female + female2 = .161156

.1858039 originally

Recall that every coded as follows is a “real” female:
female ==1 & female2==1

The real-life problem is more like this

CO2 emissions = f(Population, GDP, Industrial Production, Agricultural Production, Number of livestock)



```
. reg lco2 lgdp

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  1,    33) =   32.89
       Model |  34.5560587     1  34.5560587           Prob > F      =  0.0000
    Residual |  34.6714992    33  1.05065149           R-squared     =  0.4992
-------------+------------------------------           Adj R-squared =  0.4840
       Total |   69.227558    34  2.03610465           Root MSE      =   1.025

------------------------------------------------------------------------------
        lco2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |   1.004697   .1751872     5.73   0.000     .6482759    1.361118
       _cons |   -1.28411   2.423172    -0.53   0.600    -6.214091    3.645871
------------------------------------------------------------------------------
```

```
. reg lco2 lgdp lindust lagri

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  3,    31) =   18.97
       Model |  44.8134696     3  14.9378232           Prob > F      =  0.0000
    Residual |  24.4140884    31  .787551237           R-squared     =  0.6473
-------------+------------------------------           Adj R-squared =  0.6132
       Total |   69.227558    34  2.03610465           Root MSE      =  .88744

------------------------------------------------------------------------------
        lco2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |  -.5408816    .496644    -1.09   0.285    -1.553794    .4720305
     lindust |   1.365427   .5153901     2.65   0.013     .3142816    2.416572
       lagri |    .323599   .1718835     1.88   0.069    -.0269597    .6741578
       _cons |   -.496777   2.113246    -0.24   0.816    -4.806771    3.813217
------------------------------------------------------------------------------
```

Auxiliary regression of lgdp on lindust and lagri

```
. reg lgdp lindust lagri

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  2,    32) =  155.55
       Model |  31.0407956     2  15.5203978           Prob > F      =  0.0000
    Residual |  3.19292307    32  .099778846           R-squared     =  0.9067
-------------+------------------------------           Adj R-squared =  0.9009
       Total |  34.2337187    34  1.00687408           Root MSE      =  .31588

------------------------------------------------------------------------------
        lgdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lindust |   .9415421   .0771392    12.21   0.000     .7844147    1.098669
       lagri |   .0364249   .0608409     0.60   0.554    -.0875038    .1603537
       _cons |   1.560045   .6998153     2.23   0.033     .1345674    2.985522
------------------------------------------------------------------------------
```

Return to focus on dummy variables

- A dummy variable in general is an *indicator variable* equal to 1 if a condition is true, 0 otherwise.
- The dummy variable coefficient measures the *difference* in the mean effects between some *omitted category* and the *indicator category* (controlling for everything else in the regression)

Without the redundant variable

```
. reg approve female  [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  1,   975) =    5.70
       Model |  8.42953813     1  8.42953813           Prob > F      =  0.0172
    Residual |  1442.49523   975  1.47948229           R-squared     =  0.0058
-------------+------------------------------           Adj R-squared =  0.0048
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.2163

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .1858039   .0778409     2.39   0.017     .0330489    .3385589
       _cons |   2.292257   .0555348    41.28   0.000     2.183275    2.401238
------------------------------------------------------------------------------
```

A better specified model

```
. reg obama_approve dem01 liberal01 female [aw=v102]
(sum of wgt is   8.5597e+02)

      Source |       SS       df       MS              Number of obs =     909
-------------+------------------------------           F(  3,   905) =  333.91
       Model |  698.216712     3  232.738904           Prob > F      =  0.0000
    Residual |  630.800959   905  .697017635           R-squared     =  0.5254
-------------+------------------------------           Adj R-squared =  0.5238
       Total |  1329.01767   908  1.46367585           Root MSE      =  .83488

------------------------------------------------------------------------------
obama_appr~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dem01 |   1.687039   .0790294    21.35   0.000     1.531937    1.842142
   liberal01 |   1.195135   .1240645     9.63   0.000     .9516476    1.438623
      female |   .0489068   .0557469     0.88   0.381    -.0605015    .1583151
       _cons |   .8833574   .0631853    13.98   0.000     .7593507    1.007364
------------------------------------------------------------------------------
```

Another model, to get party ID better

Rather than code party ID as a linear variable (1 = Dem, 0 = Ind, -1 = Rep),

code 3 dummy variables:

dem = 1 if party ID == Democrat, 0 otherwise
ind = 1 if party ID = Independent, 0 otherwise
Rep = 1 if party ID = Republican, 0 otherwise

Note that dem + ind + rep = 1, so we have to exclude a category to put these into the regression (let’s choose to exclude ind)

```
. reg obama_approve  dem rep liberal01 female [aw=v102]
(sum of wgt is   8.5597e+02)

      Source |       SS       df       MS              Number of obs =     909
-------------+------------------------------           F(  4,   904) =  250.63
       Model |  698.845875     4  174.711469           Prob > F      =  0.0000
    Residual |  630.171797   904  .697092695           R-squared     =  0.5258
-------------+------------------------------           Adj R-squared =  0.5237
       Total |  1329.01767   908  1.46367585           Root MSE      =  .83492

------------------------------------------------------------------------------
obama_appr~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dem |   .8952153   .0672499    13.31   0.000     .7632311    1.027199
         rep |  -.7831749    .074808   -10.47   0.000    -.9299925   -.6363573
   liberal01 |   1.203246   .1243646     9.68   0.000     .9591695    1.447323
      female |   .0463874    .055813     0.83   0.406    -.0631507    .1559255
       _cons |    1.68793   .0773509    21.82   0.000     1.536121    1.839738
------------------------------------------------------------------------------
```

Multiple dummy variables

- Dummy variables are how we deal with categorical independent variables
  - Race (white, black, Hispanic, Asian-Amer., other)
  - Alliance (NATO, Warsaw Pact, other)
  - Religion (Christian, Jewish, Muslim, Hindu, other)
- To code dummy variables, create a new variable for each category
  - In the regression, exclude one of the variables (usually the most numerous) --- **remember the lecture on collinearity**

Example: predicting Obama approval by race (white is omitted category)

```
. gen white=race==1
. gen black=race==2
. gen hisp=race==3
. gen other_race=1-white-black-hisp
. reg approve_o black hisp other_race [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  3,   973) =   33.34
       Model |  135.231871     3  45.0772903           Prob > F      =  0.0000
    Residual |   1315.6929   973  1.35220236           R-squared     =  0.0932
-------------+------------------------------           Adj R-squared =  0.0904
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.1628

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   1.097946   .1125408     9.76   0.000     .8770956    1.318797
        hisp |   .4846481   .1577768     3.07   0.002     .1750261      .79427
  other_race |   .0429154   .2131171     0.20   0.840    -.3753068    .4611375
       _cons |   2.216973   .0420862    52.68   0.000     2.134383    2.299563
------------------------------------------------------------------------------

.
```

Example: predicting Obama approval by race (forgetting omitted category)

```
. reg approve_o white black hisp other_race [aw=v102]
(sum of wgt is   9.5361e+02)
note: other_race omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  3,   973) =   33.34
       Model |  135.231871     3  45.0772903           Prob > F      =  0.0000
    Residual |   1315.6929   973  1.35220236           R-squared     =  0.0932
-------------+------------------------------           Adj R-squared =  0.0904
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.1628

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |  -.0429154   .2131171    -0.20   0.840    -.4611375    .3753068
       black |   1.055031    .233542     4.52   0.000     .5967269    1.513335
        hisp |   .4417327   .2583988     1.71   0.088    -.0653504    .9488157
  other_race |          0  (omitted)
       _cons |   2.259888   .2089202    10.82   0.000     1.849902    2.669875
------------------------------------------------------------------------------
```

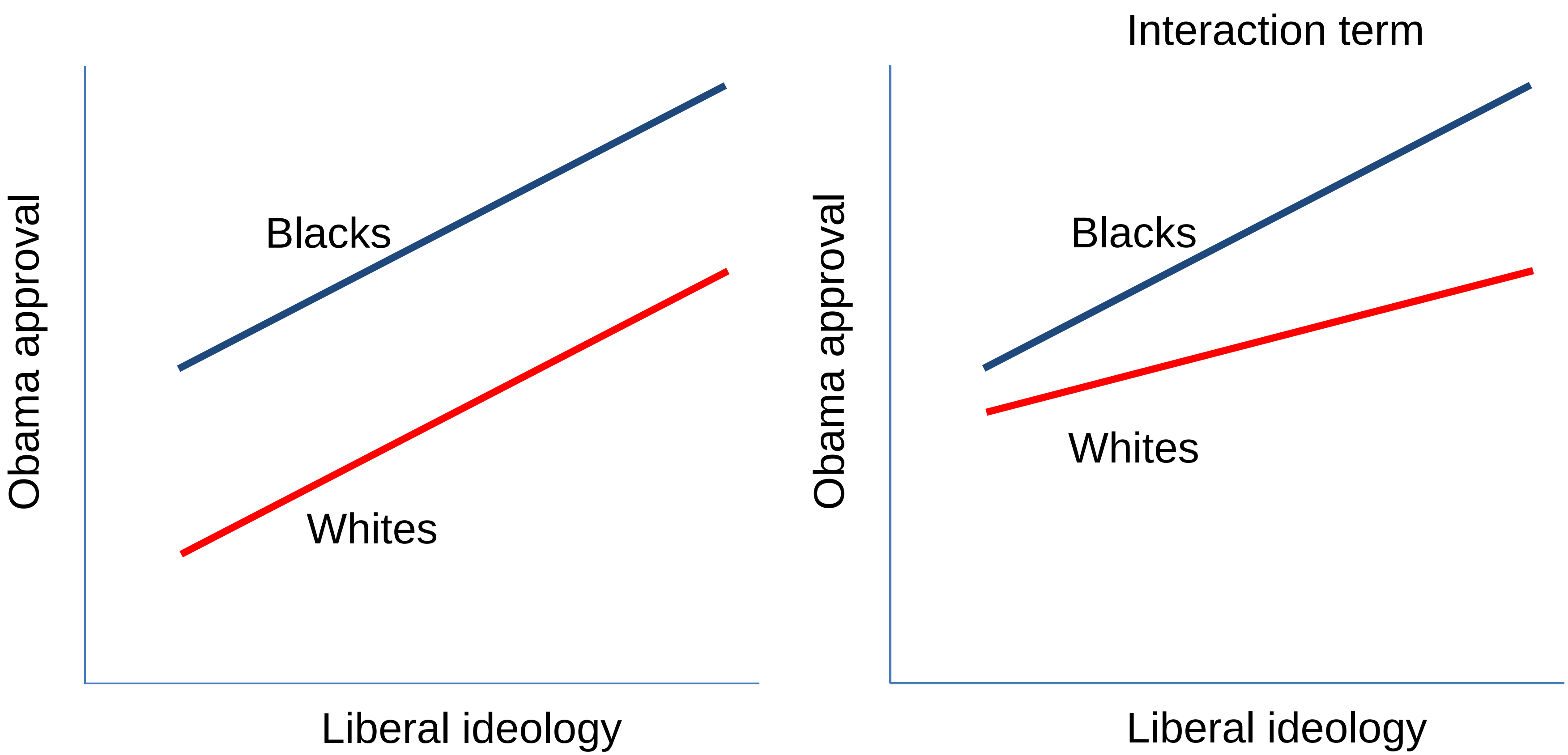
The lazy person’s way to enter in a bunch of categorical variables

```
. reg approve_o i.race [aw=v102]
(sum of wgt is   9.5361e+02)

      Source |       SS       df       MS              Number of obs =     977
-------------+------------------------------           F(  6,   970) =   17.33
       Model |  140.469008     6  23.4115013           Prob > F      =  0.0000
    Residual |  1310.45576   970  1.35098532           R-squared     =  0.0968
-------------+------------------------------           Adj R-squared =  0.0912
       Total |  1450.92477   976  1.48660325           Root MSE      =  1.1623

------------------------------------------------------------------------------------
   approve_obama |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------------
            race |
           black |   1.097946   .1124902     9.76   0.000     .8771941    1.318698
        hispanic |   .4846481   .1577057     3.07   0.002     .1751643    .7941318
           asian |   .4934667    .669953     0.74   0.462    -.8212575    1.808191
 native american |   .1003945   1.144268     0.09   0.930    -2.145132    2.345921
           mixed |   .5948015   .4007686     1.48   0.138    -.1916718    1.381275
           other |   -.289047   .2740686    -1.05   0.292    -.8268827    .2487888
                 |
           _cons |   2.216973   .0420672    52.70   0.000      2.13442    2.299526
------------------------------------------------------------------------------------
```

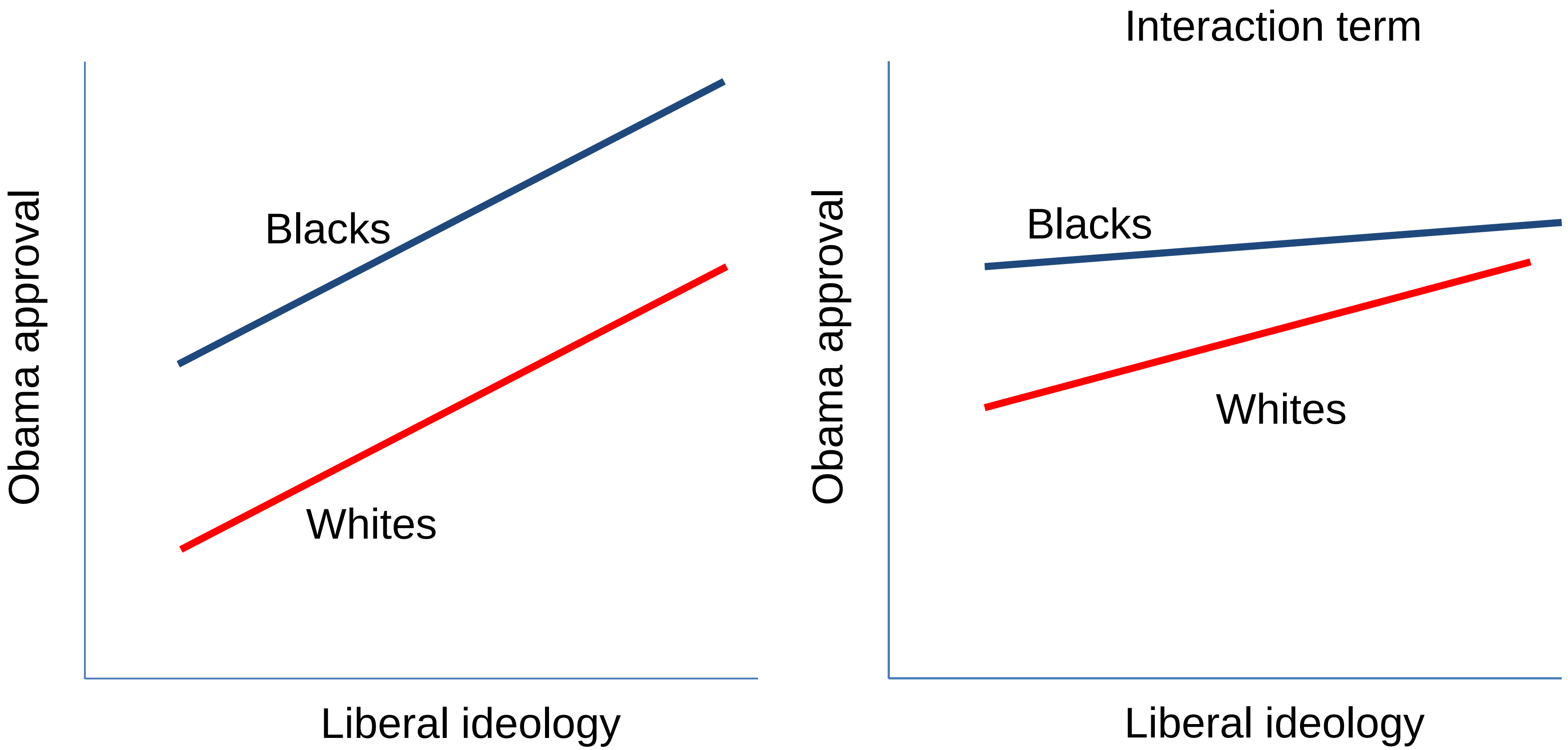
Interaction Terms

- Sometimes we think that one set of regression coefficients apply to one population and another set to another population
- Example: evaluations of Barak Obama
  - Assume ideology and party ID is the baseline factor leading to approval of Obama
  - We know that African-Americans approve of Obama more than whites, but is it:
    - African-Americans simply like Obama better, controlling for Ideology and party ID **or**
    - African-Americans weigh ideology and party ID differently?

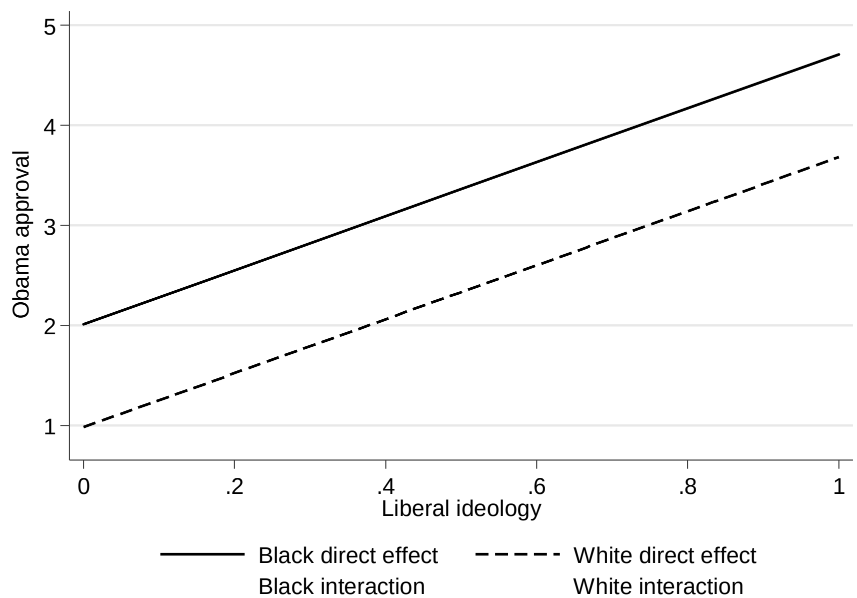
What those explanations look like



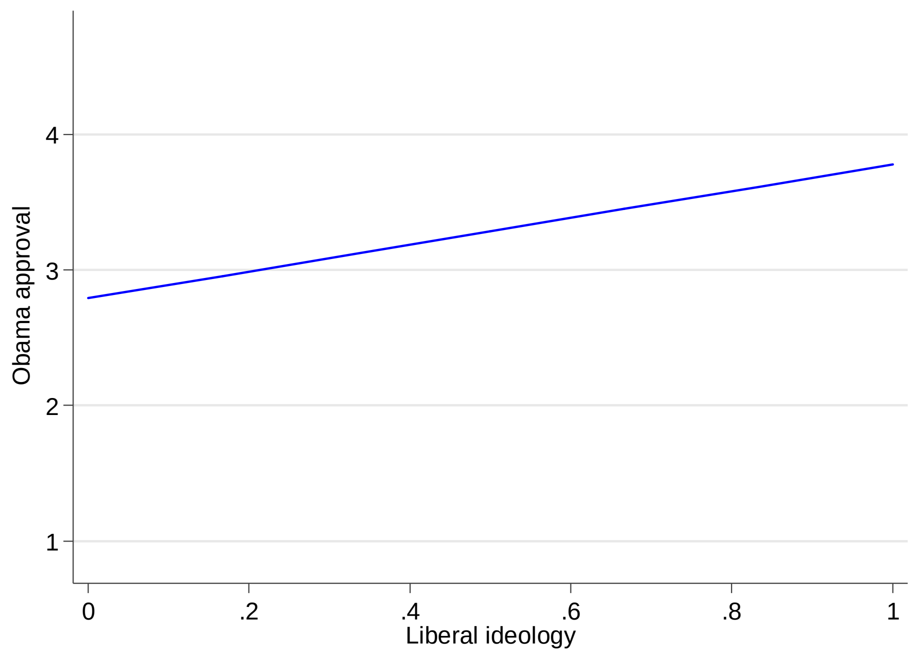
What those explanations look like



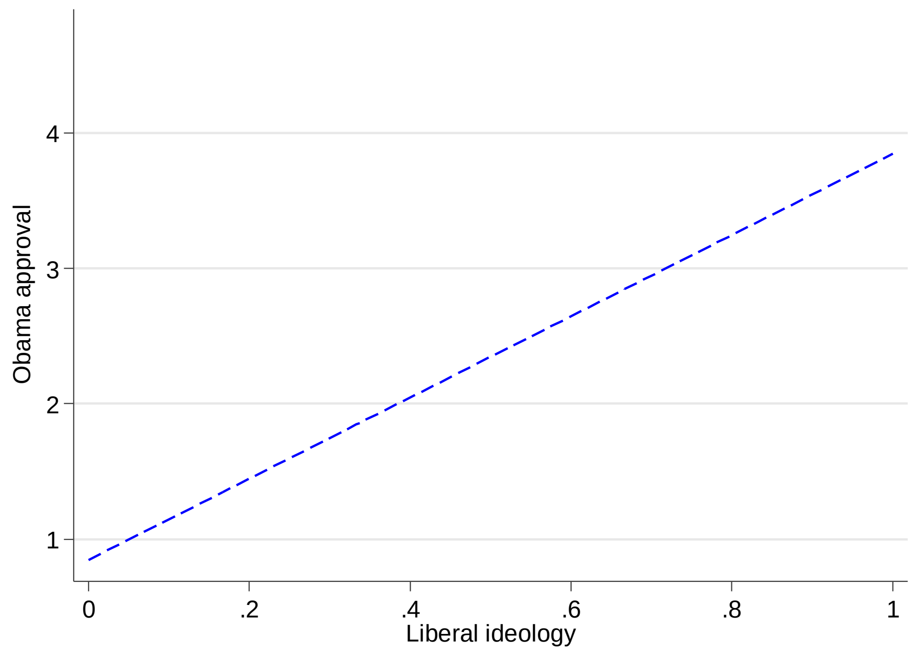
```
 reg approve_obama liberal01 african_am [aw=v102]
(sum of wgt is   8.1363e+02)

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   liberal01 |   2.698301   .1121764    24.05   0.000     2.478114    2.918488
  african_am |   1.029124   .0920621    11.18   0.000     .8484191    1.209829
       _cons |   .9838004    .060948    16.14   0.000      .864168    1.103433
------------------------------------------------------------------------------
```



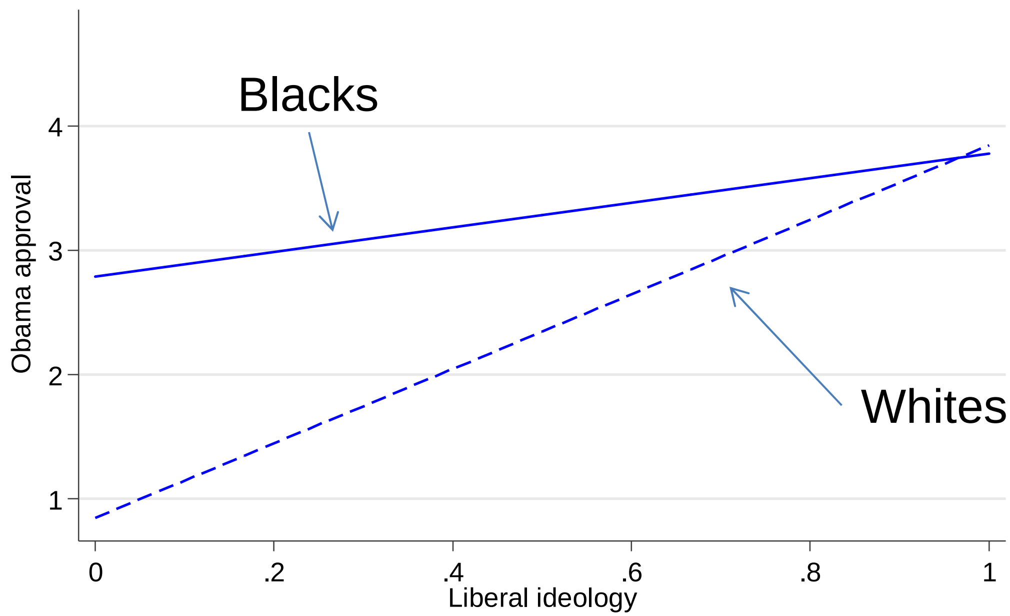
```
. reg approve_obama liberal01 african_am [aw=v102] if african_am==1
(sum of wgt is   1.0599e+02)
note: african_am omitted because of collinearity

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   liberal01 |   .9865412   .3171368     3.11   0.002     .3580505    1.615032
  african_am |          0  (omitted)
       _cons |   2.791209   .1717943    16.25   0.000     2.450753    3.131665
```

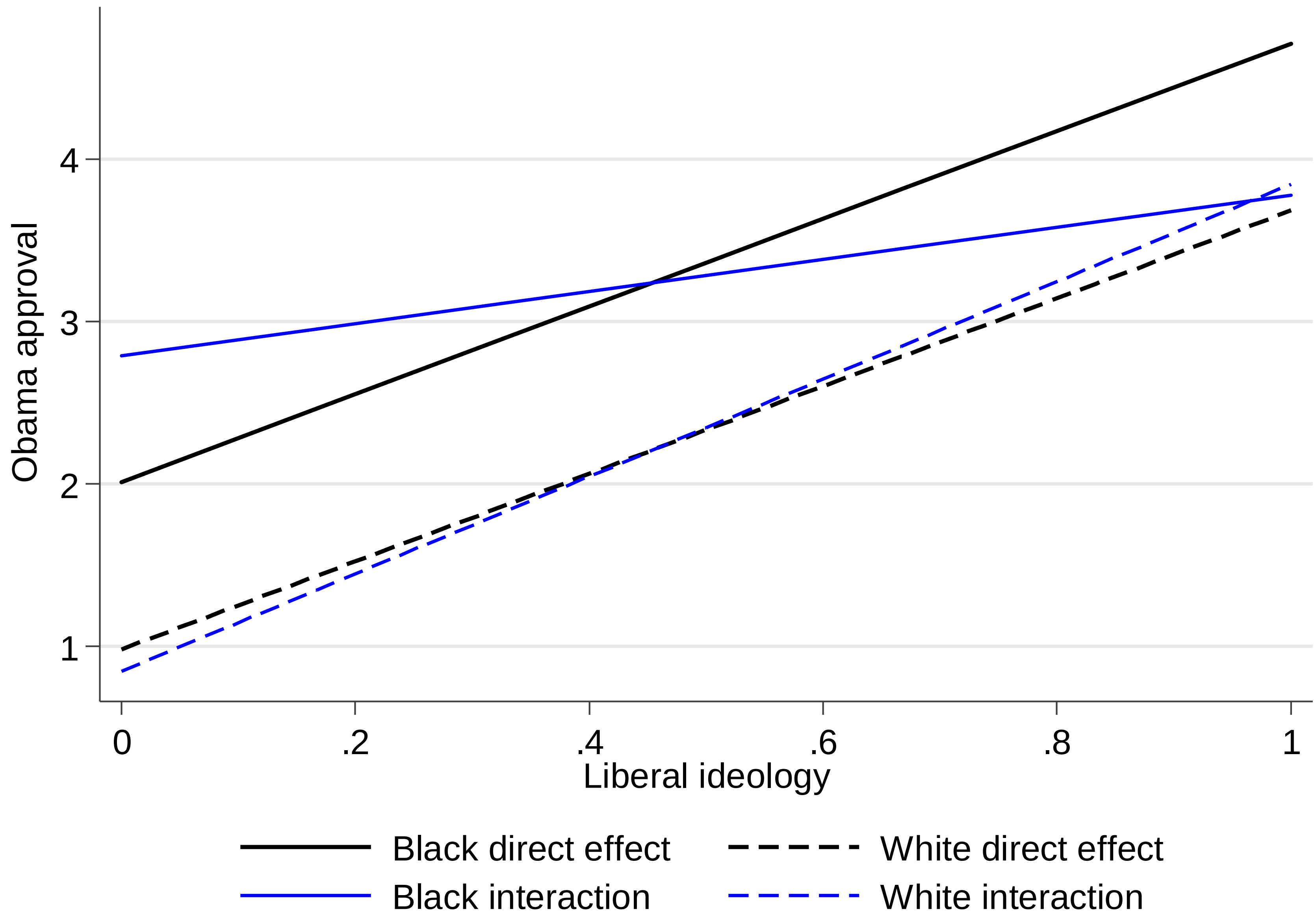


```
. reg approve_obama liberal01 african_am [aw=v102] if african_am==0
(sum of wgt is   7.0764e+02)
note: african_am omitted because of collinearity


------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   liberal01 |   2.995738   .1163107    25.76   0.000     2.767384    3.224093
  african_am |          0  (omitted)
       _cons |   .8483249   .0617883    13.73   0.000      .727015    .9696349
------------------------------------------------------------------------------
```



```
. reg approve_obama liberal01 african_am [aw=v102] if african_am==1
(sum of wgt is   1.0599e+02)
note: african_am omitted because of collinearity

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   liberal01 |   .9865412   .3171368     3.11   0.002     .3580505    1.615032
  african_am |          0  (omitted)
       _cons |   2.791209   .1717943    16.25   0.000     2.450753    3.131665


. reg approve_obama liberal01 african_am [aw=v102] if african_am==0
(sum of wgt is   7.0764e+02)
note: african_am omitted because of collinearity


------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   liberal01 |   2.995738   .1163107    25.76   0.000     2.767384    3.224093
  african_am |          0  (omitted)
       _cons |   .8483249   .0617883    13.73   0.000      .727015    .9696349
------------------------------------------------------------------------------
```


Blacks
Whites
Obama approval
1
2
3
4
0
.2
.4
.6
.8
1
Liberal ideology



Obama approval
4
3
2
1
0
.2
.4
.6
.8
1
Liberal ideology
Black direct effect
White direct effect
Black interaction
White interaction

- We can do this comparison all in one regression
- obama_approve=b_0 +
- b_1*african_am +
- b_2*liberal01 +
- b_3*african_am×liberal01

- We can do this comparison all in one regression
- obama_approve=b_0 +
- b_1*african_am +
- b_2*liberal01 +
- **b3*african_am×liberal01**

Interaction term

- Note, if the R is white, african_am = 0, therefore:

$$obama\_approve = b_0 +$$
$$b_1 * 0 +$$
$$b_2 * liberal01 +$$
$$b_3 * 0 \times liberal01$$

- Note, if the R is white, african_am = 0, therefore:

obama_approve=b_0 +

b_1*0+

b_2*liberal01 +

b_3*0×liberal01

- BECOMES
- obama_approve=b_0 + b_2*liberal01

- Note, if the R is black, african_am = 1, therefore:

obama_approve=b_0 +

b_1*1 +

b_2*liberal01 +

b_3*1×liberal01

- Note, if the R is black, african_am = 1, therefore:

obama_approve=b_0 +

b_1*1 +

b_2*liberal01 +

b_3*1×liberal01

Group terms

- BECOMES
- obama_approve=(b_0 + b_1)+ (b_2+b_3)liberal01

```
. gen african_amXliberal01=african_am*liberal01
(166 missing values generated)

. reg approve_obama african_am liberal01 african_amXliberal01 [aw=v102]

------------------------------------------------------------------------------------
       approve_obama |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------+--------------------------------------------------------------
          african_am |   1.942884   .1664602    11.67   0.000     1.616146    2.269623
           liberal01 |   2.995738   .1185683    25.27   0.000     2.763005    3.228472
african_amXliberal01 |  -2.009197   .3081642    -6.52   0.000    -2.614082   -1.404312
               _cons |   .8483249   .0629876    13.47   0.000     .7246888    .9719611
------------------------------------------------------------------------------------
```

approve_obama	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
african_am	1.942884	.1664602	11.67	0.000	1.616146	2.269623
liberal01	2.995738	.1185683	25.27	0.000	2.763005	3.228472
african_amXliberal01	-2.009197	.3081642	-6.52	0.000	-2.614082	-1.404312
_cons	.8483249	.0629876	13.47	0.000	.7246888	.9719611

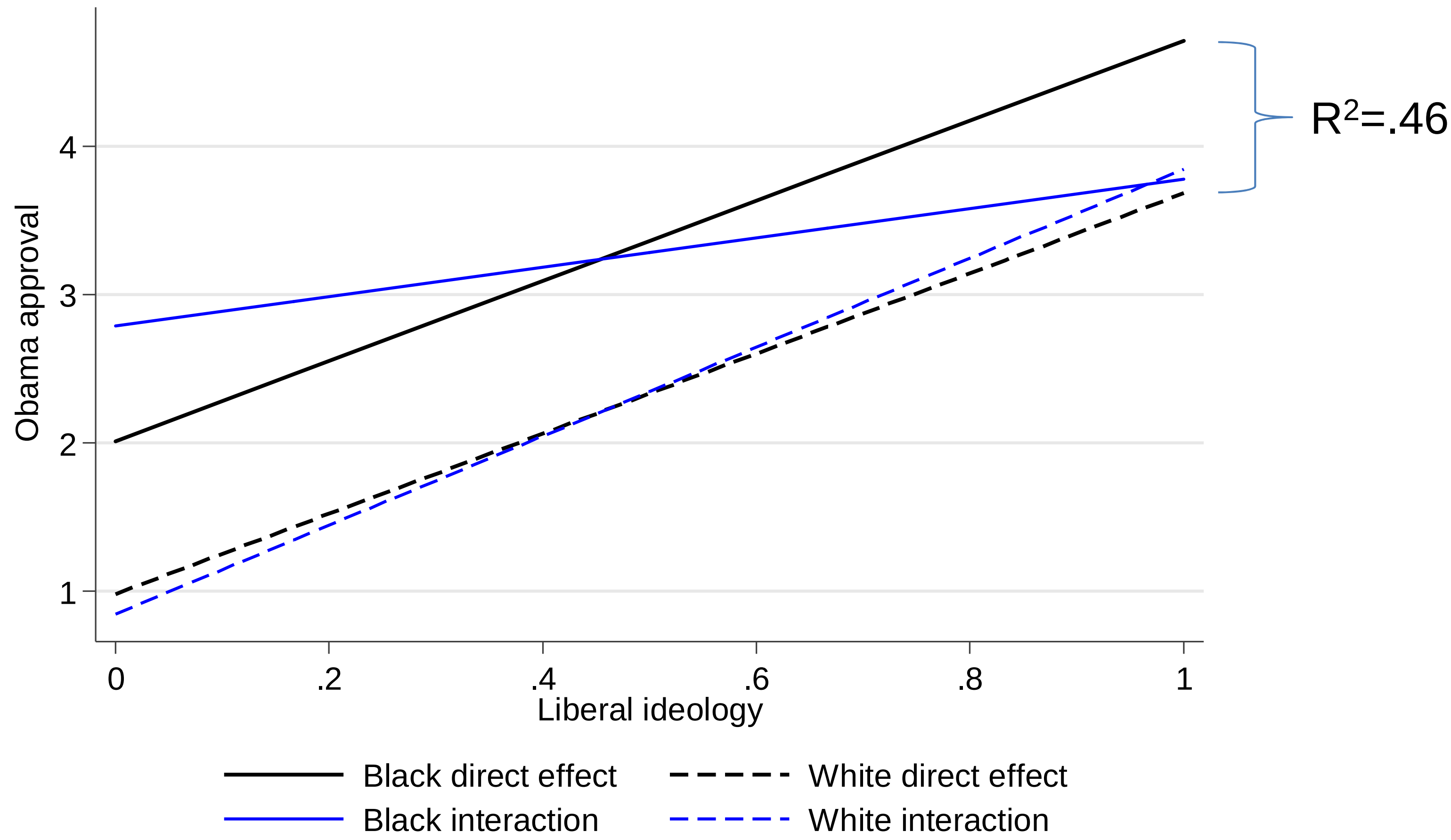
african_am==0

approve_ob~a	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
liberal01	2.995738	.1163107	25.76	0.000	2.767384	3.224093
african_am	0	(omitted)				
_cons	.8483249	.0617883	13.73	0.000	.727015	.9696349

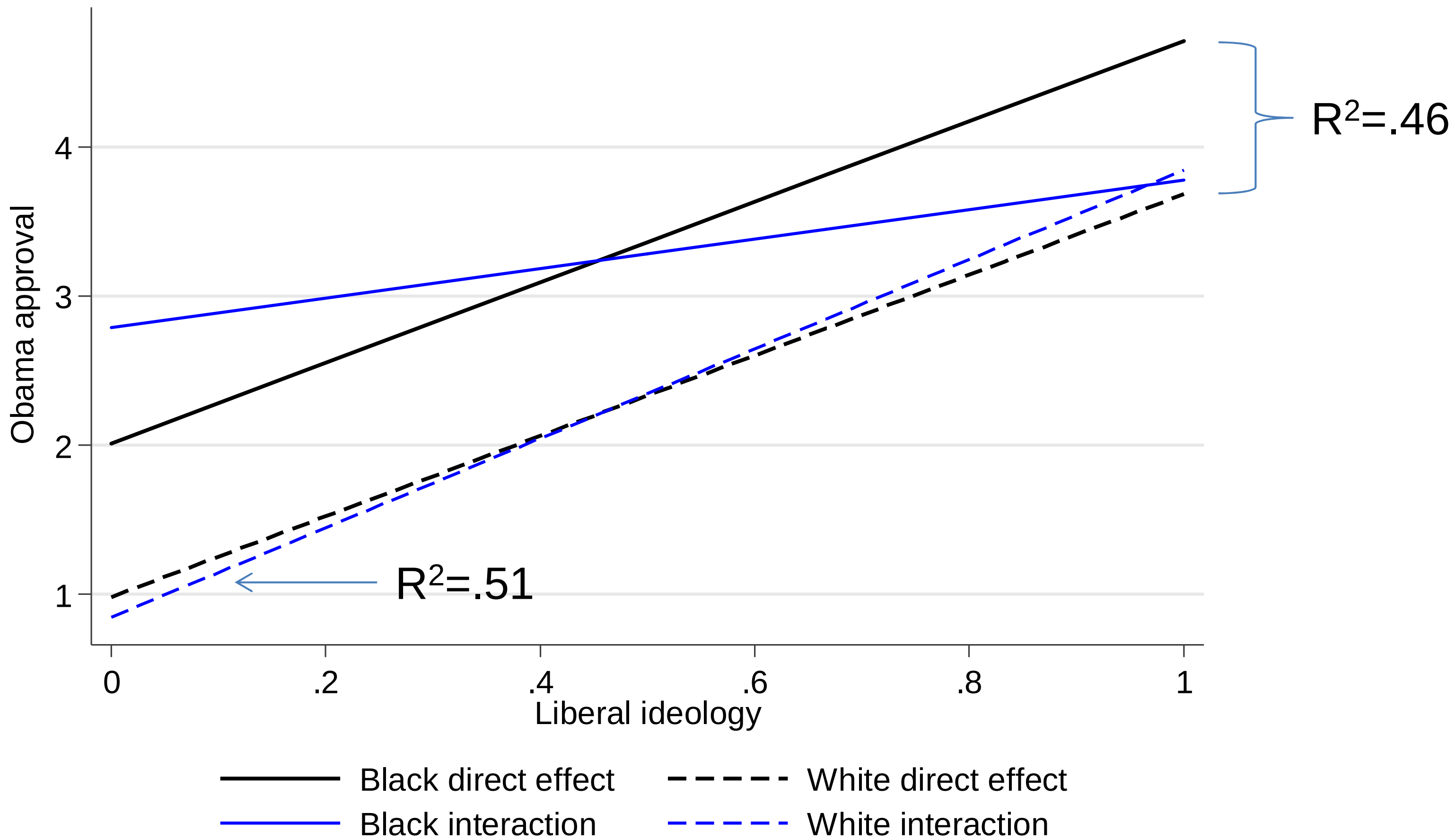
african_am==1

approve_ob~a	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
liberal01	.9865412	.3171368	3.11	0.002	.3580505	1.615032
african_am	0	(omitted)				
_cons	2.791209	.1717943	16.25	0.000	2.450753	3.131665

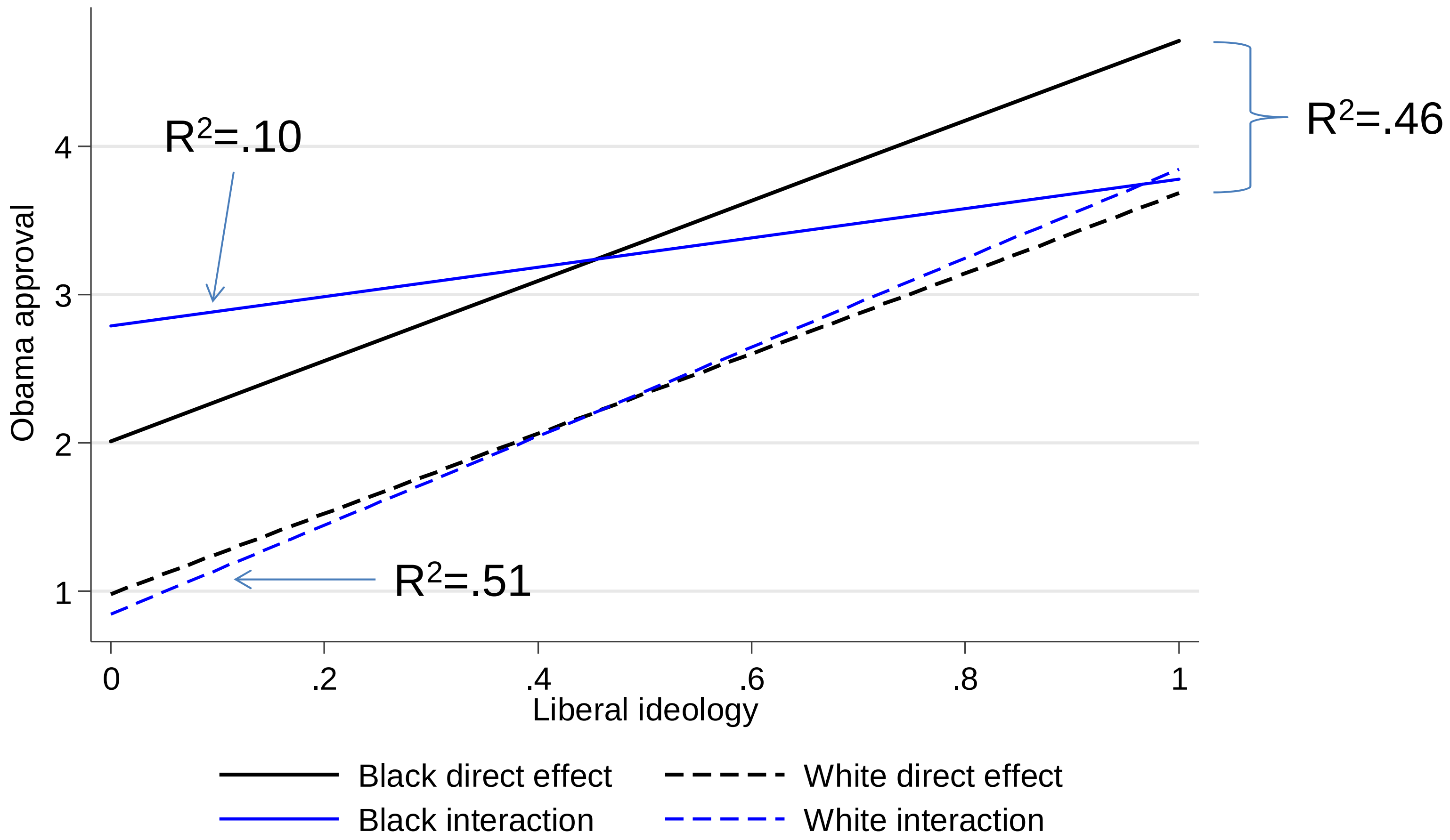
Aside about R^2



Aside about R^2

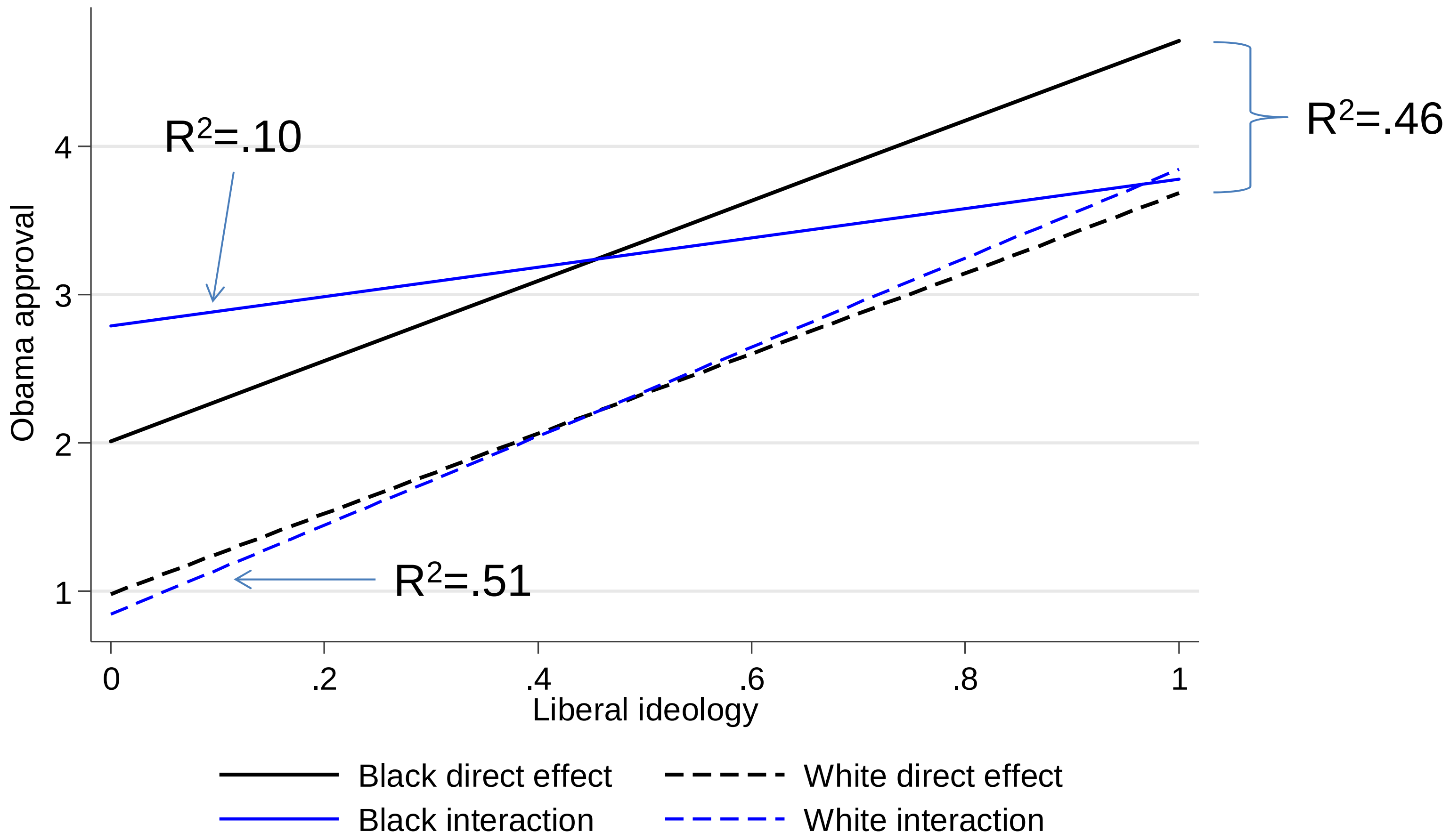


Aside about R^2



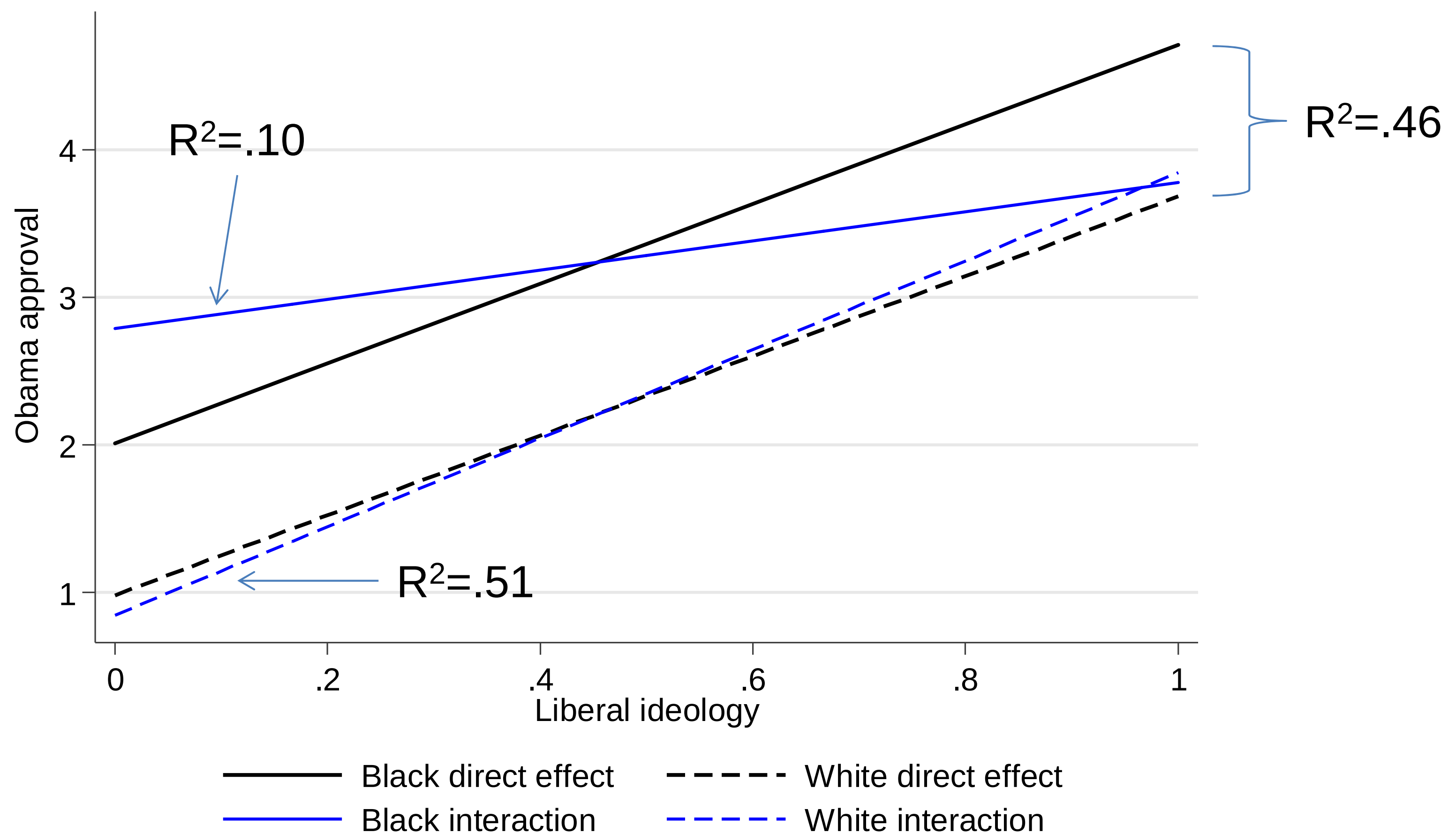
Aside about R^2

R^2 of interaction term regression =.49



Aside about R^2

R^2 of interaction term regression = .49



Fixed effects regression: Another application of dummy variables

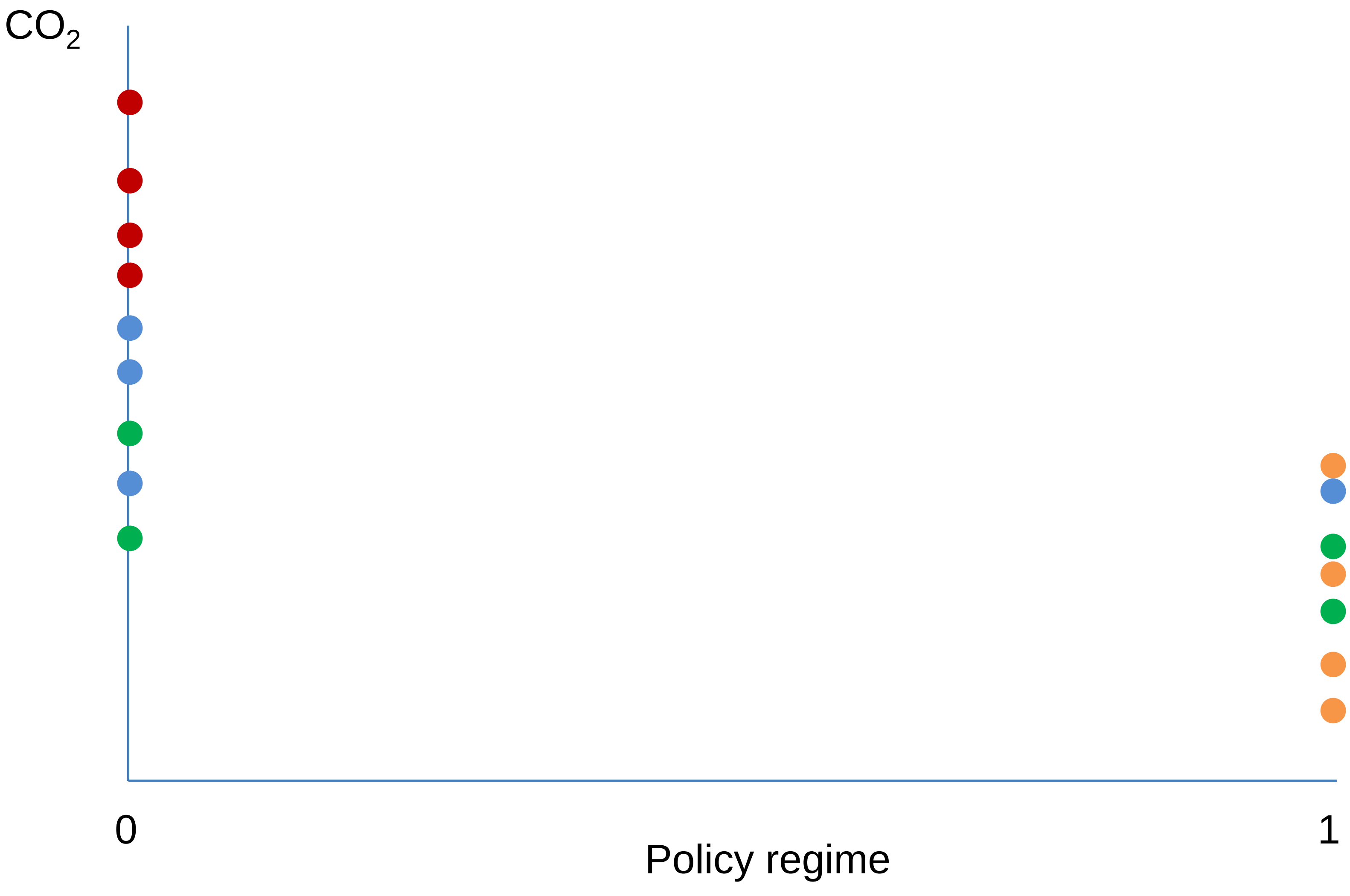
- Suppose we want to know if certain types of national environmental regulations have been effective in reducing CO_2 emissions
- Kitchen sink approach:
  - CO_2 emissions$_t$ = f(policy dummy$_t$, $\sum kitchen\ sink_t$)
- Problems with this approach
  - Specification of the kitchen sink
  - Measurement of the kitchen sink
  - Cross-sectional analysis bad way to interrogate causality

Distracting from the interesting question

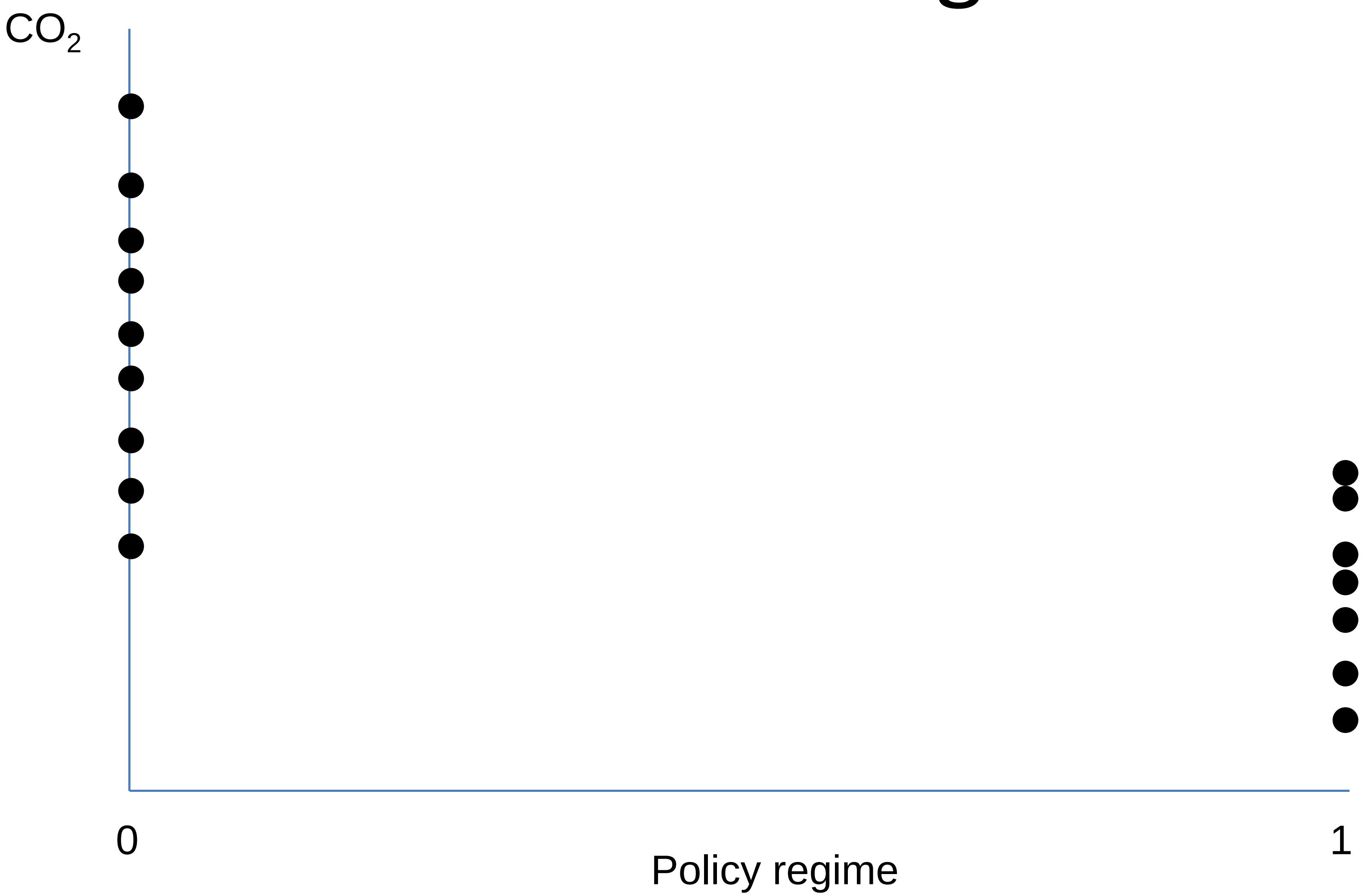
Solution

- Collect panel data
- Conduct a fixed effects regression, where the "country effects" are accounted for by dummy variables

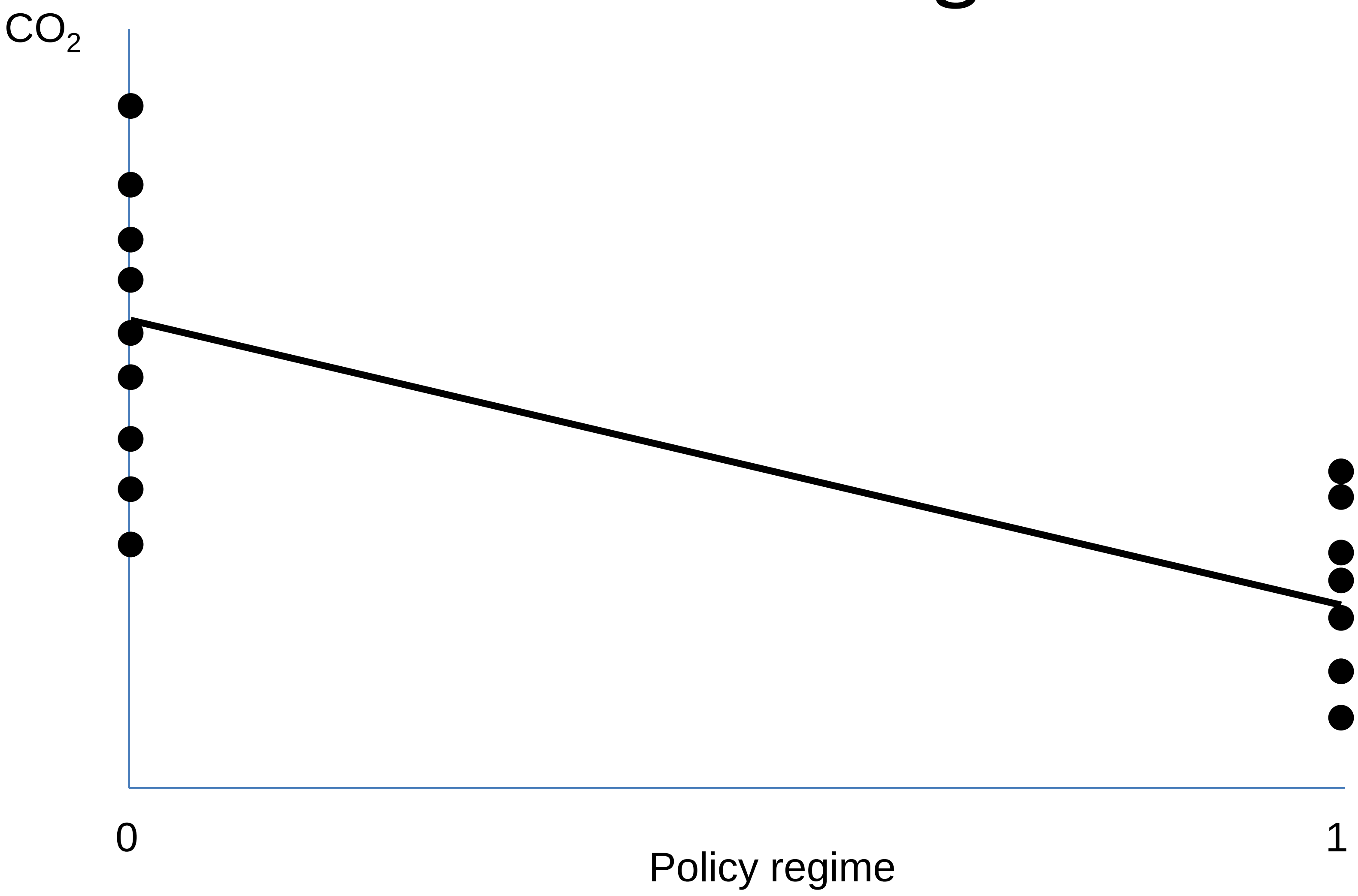
The picture



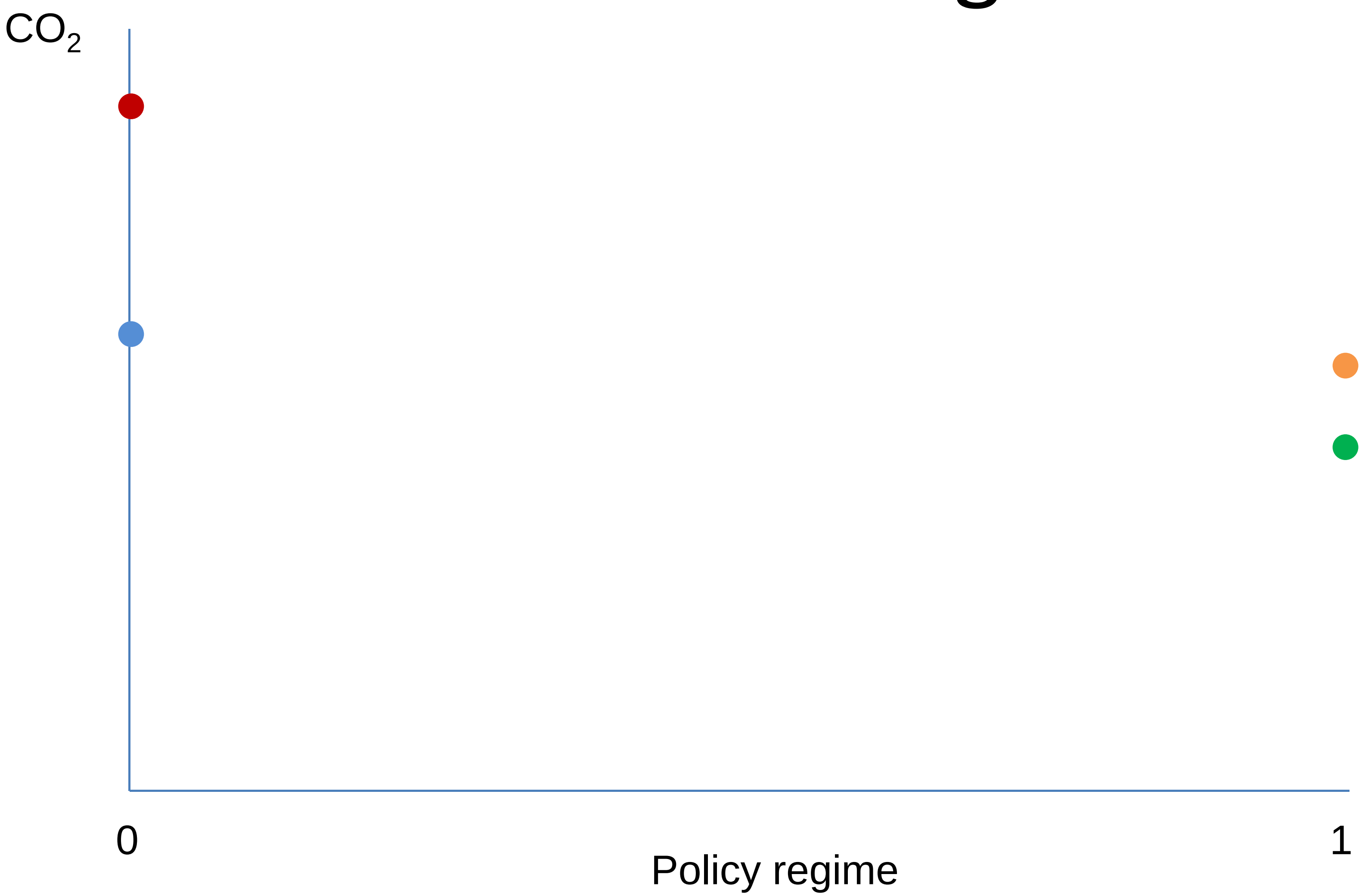
The picture:
Indiscriminate regression



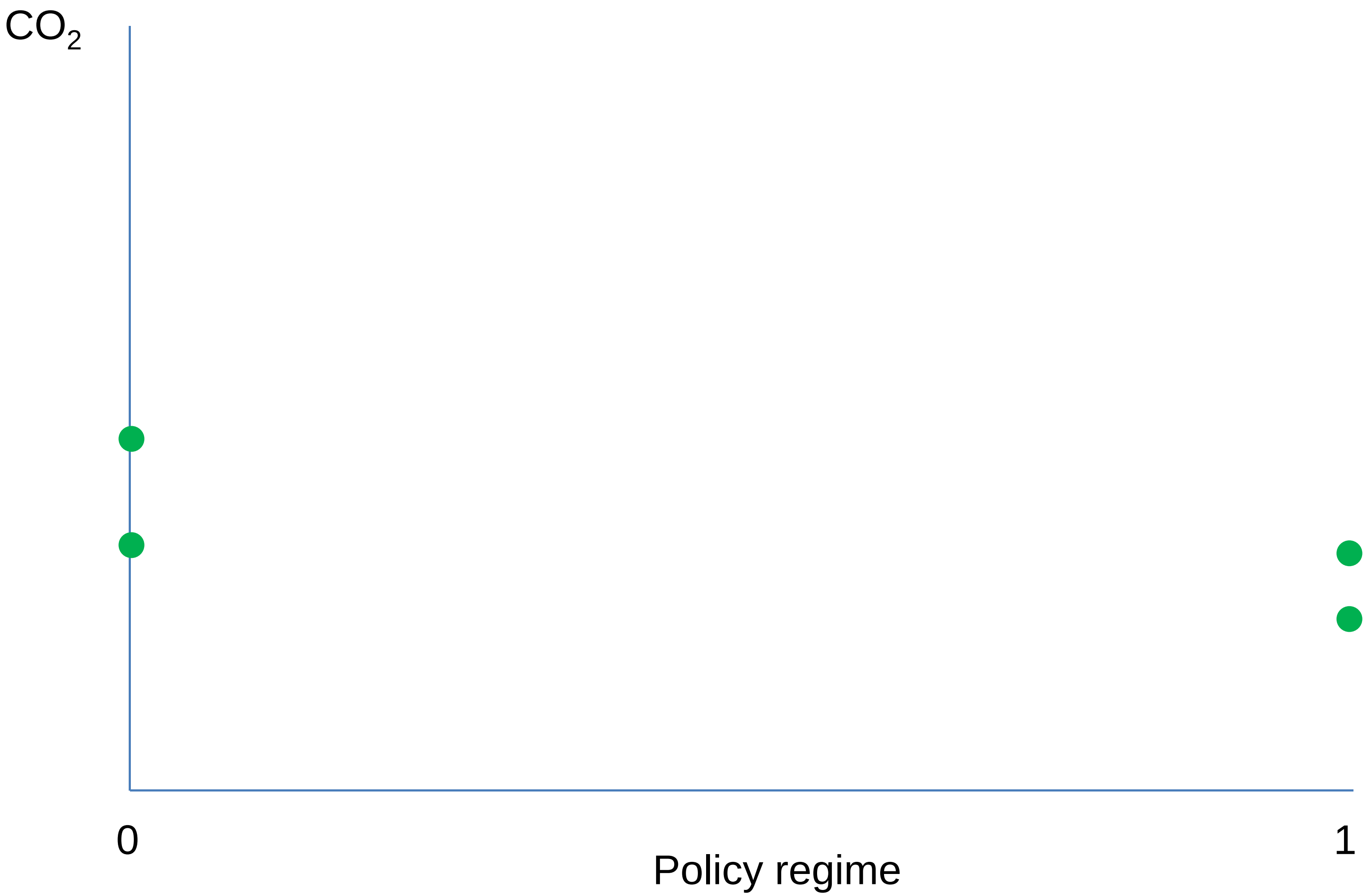
The picture: Indiscriminate regression



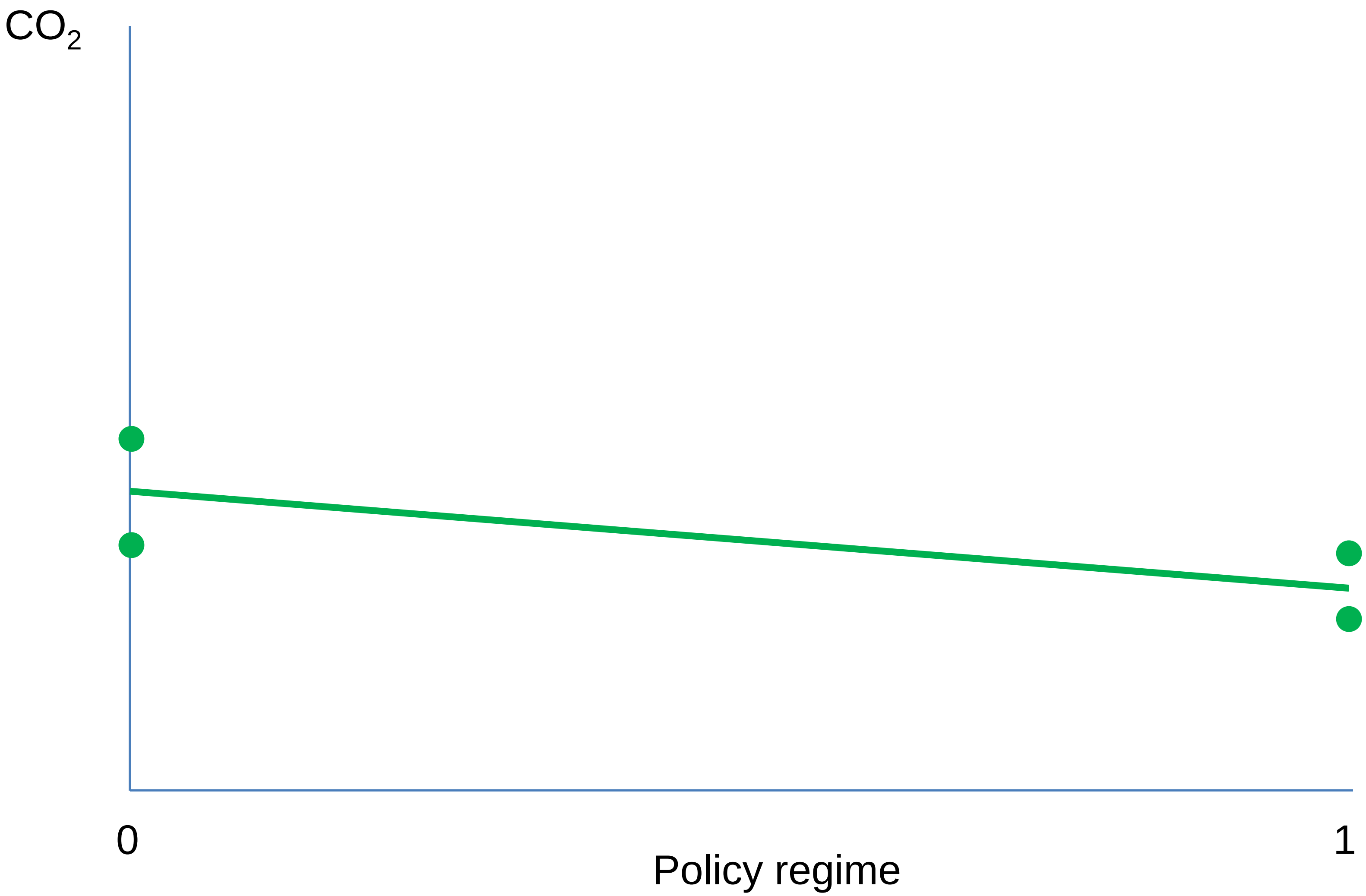
The picture: Cross-sectional regression



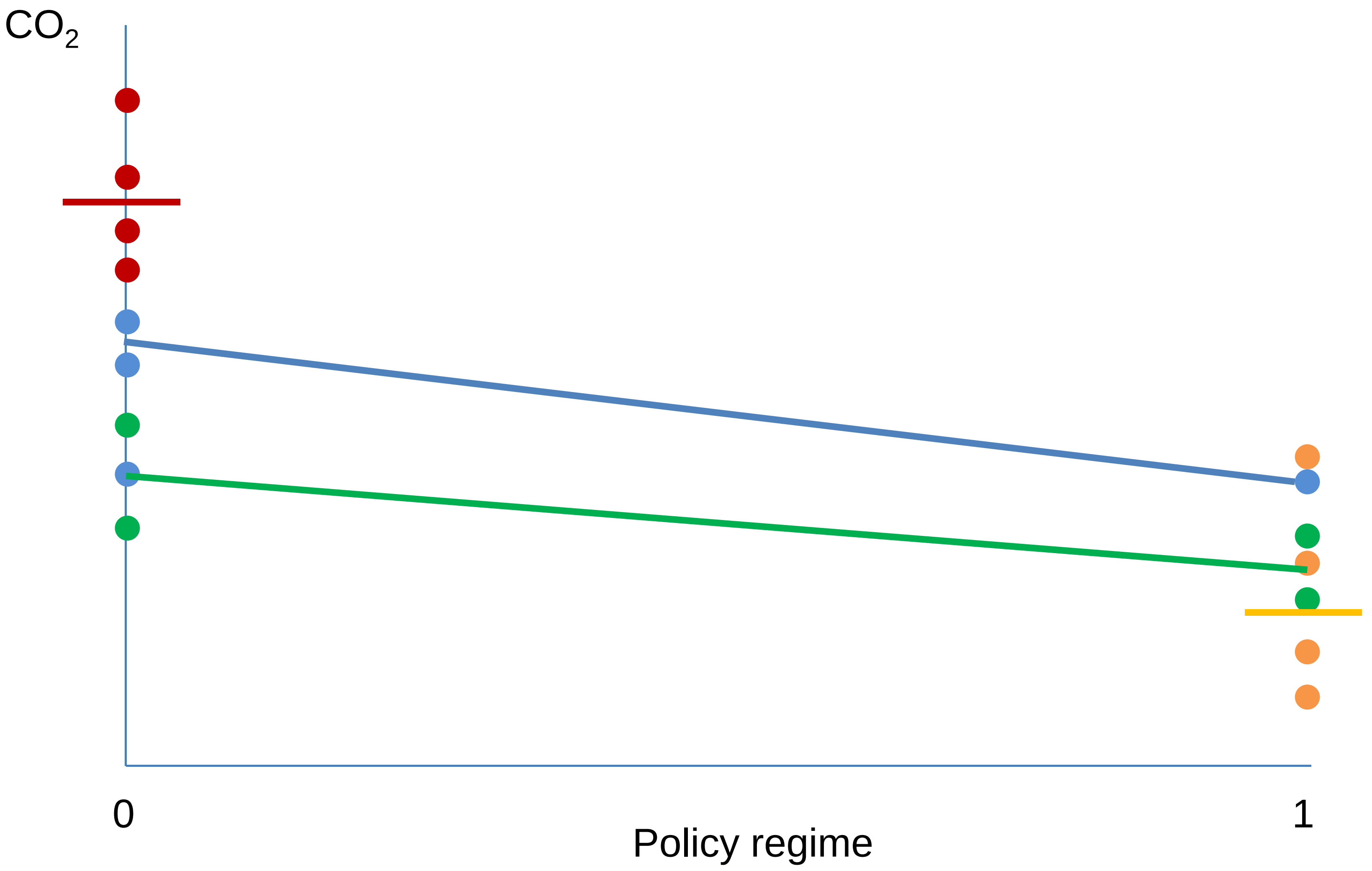
The picture
Longitudinal regression, 1 country



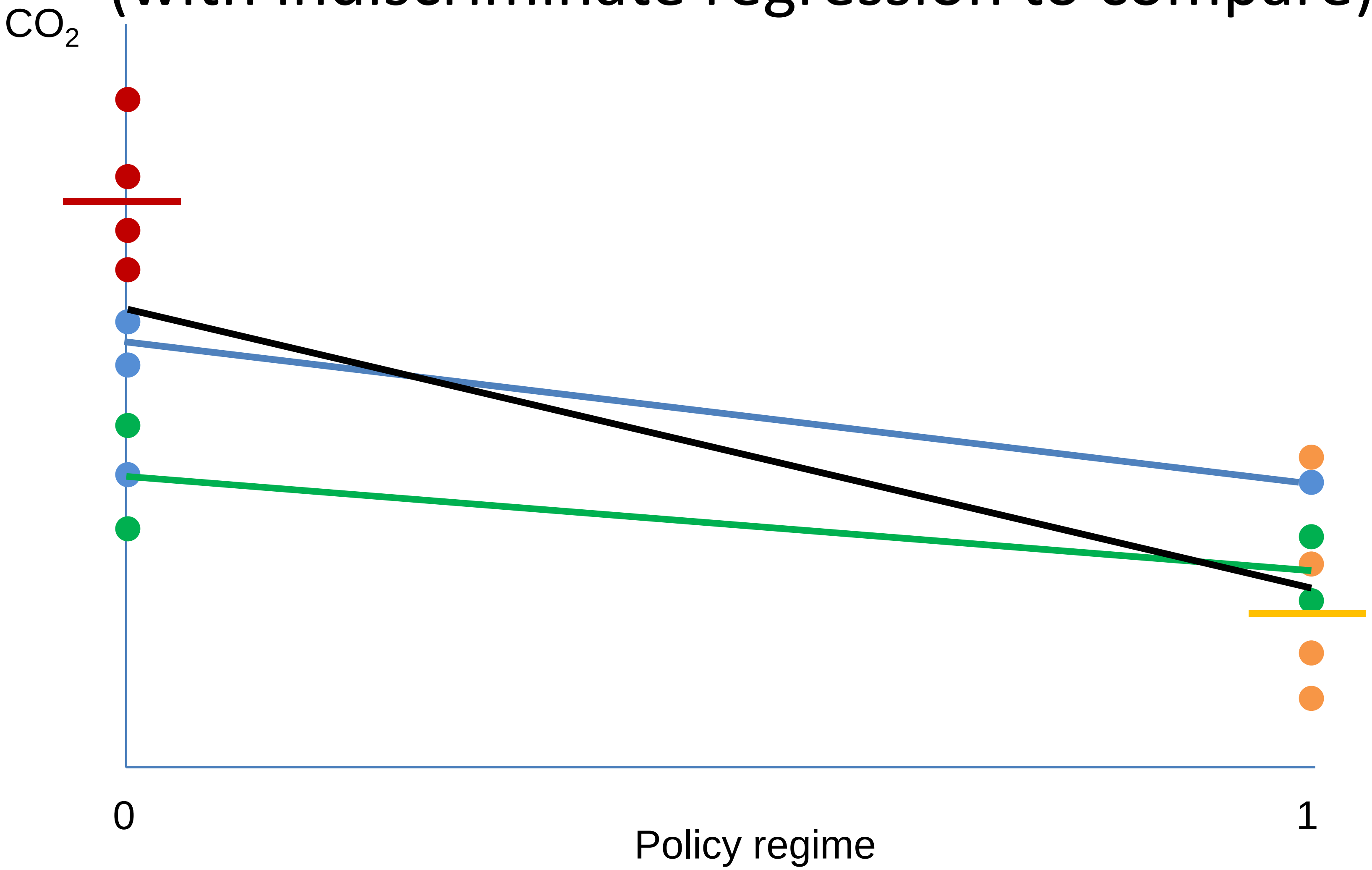
The picture
Longitudinal regression, 1 country



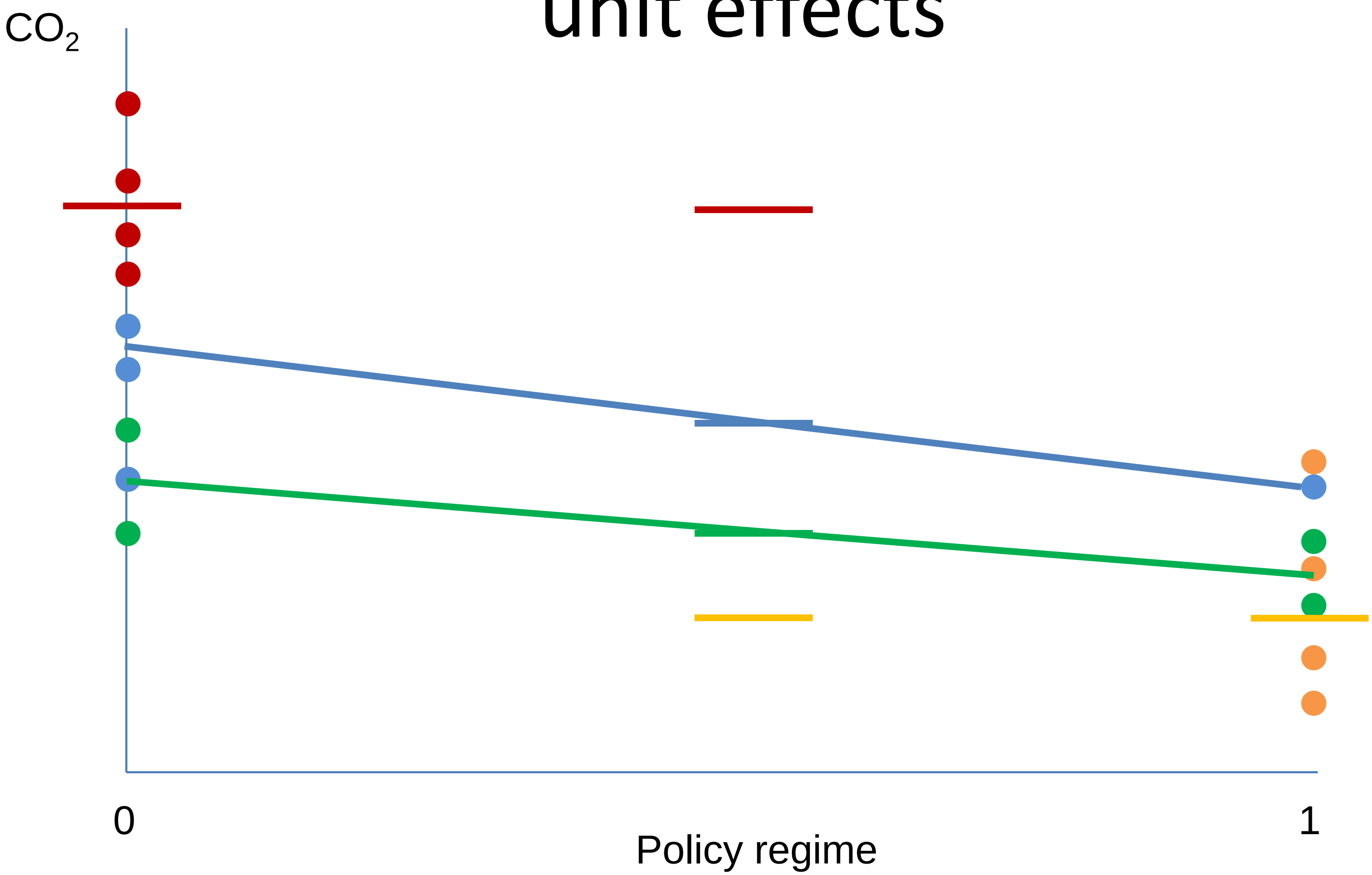
The picture:
Longitudinal regression, 4 countries



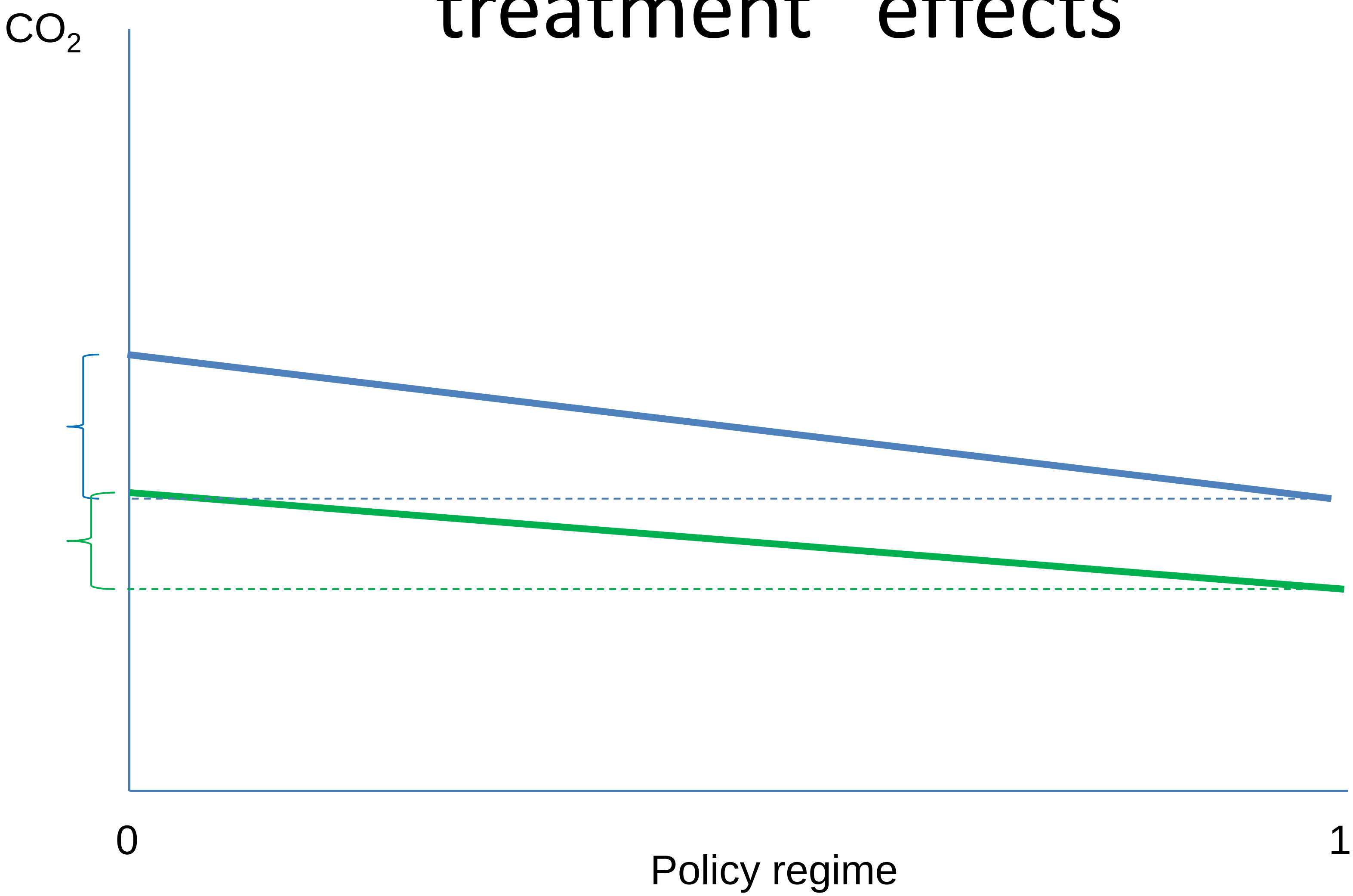
The picture:
Longitudinal regression, 4 countries
(with indiscriminate regression to compare)



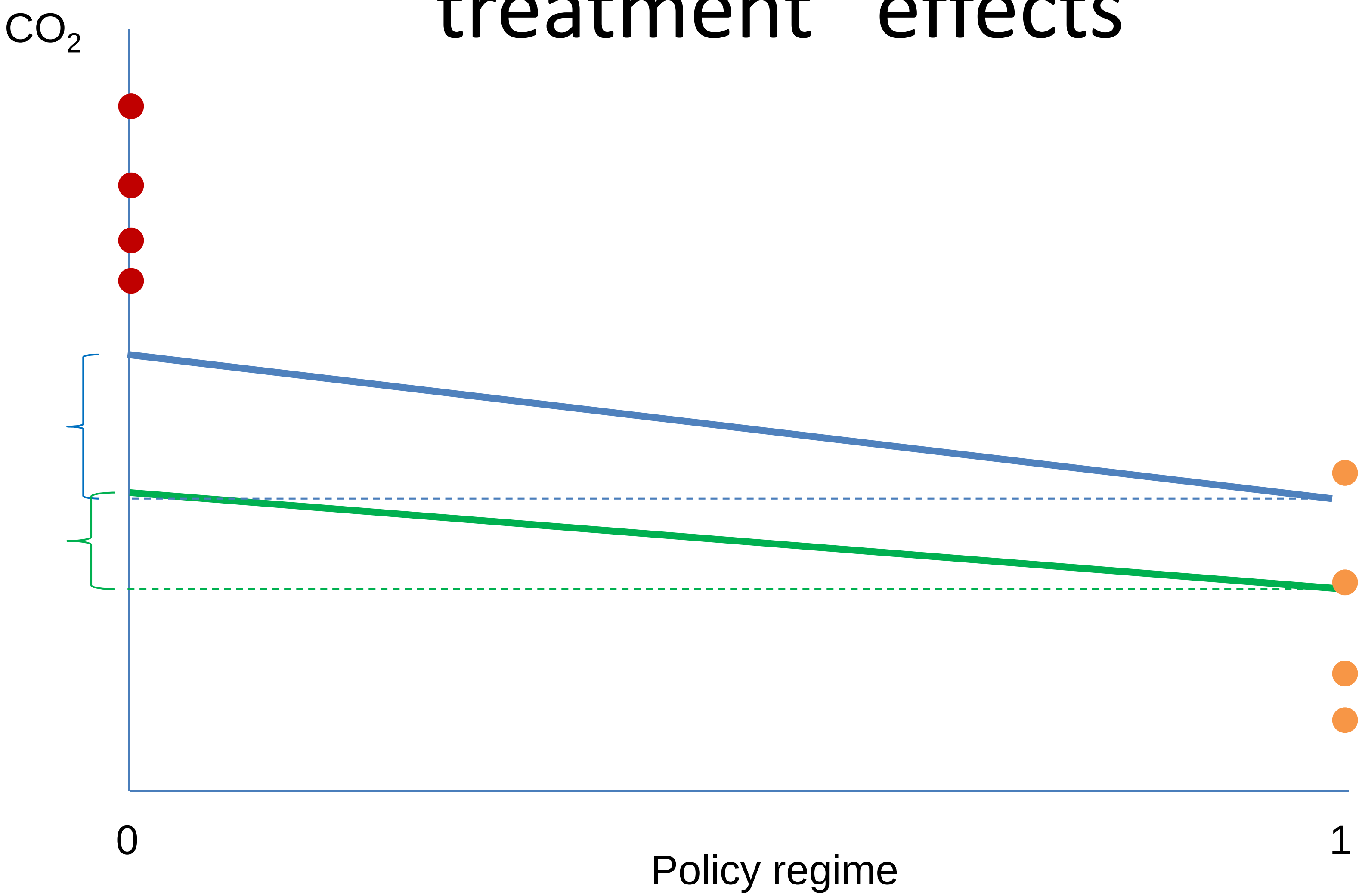
The picture:
Longitudinal regression, 4 countries, unit effects



The picture: Cross-sectional regression, 4 countries, "treatment" effects



The picture: Cross-sectional regression, 4 countries, "treatment" effects



Regression framework for previous example

$CO2_{c,t} = B_0 + B_1 policy_{c,t} + B_2 red_c + B_3 green_c + B_4 blue_c + e_{c,t}$

(orange is the omitted category)

country	year	co2	policy	red	blue	green	orange
blue	1	12	0	0	1	0	0
blue	2	11	0	0	1	0	0
blue	3	7	0	0	1	0	0
blue	4	9	1	0	1	0	0
green	1	8	0	0	0	1	0
green	2	4	0	0	0	1	0
green	3	6	1	0	0	1	0
green	4	3	1	0	0	1	0
orange	1	10	1	0	0	0	1
orange	2	5	1	0	0	0	1
orange	3	2	1	0	0	0	1
orange	4	1	1	0	0	0	1
red	1	16	0	1	0	0	0
red	2	15	0	1	0	0	0
red	3	14	0	1	0	0	0
red	4	13	0	1	0	0	0

Regression framework for previous example

$$CO2_{c,t} = B_0 + B_1 policy_{c,t} + B_2 red_c + B_3 green_c + B_4 blue_c + B_5 year2_t + B_6 year3_t + B_7 year4_t + e_{c,t}$$

(orange is the omitted category)
(year1 = omitted year)

country	year	co2	policy	red	blue	green	orange	year1	year2	year3	year4
blue	1	12	0	0	1	0	0	1	0	0	0
blue	2	11	0	0	1	0	0	0	1	0	0
blue	3	7	0	0	1	0	0	0	0	1	0
blue	4	9	1	0	1	0	0	0	0	0	1
green	1	8	0	0	0	1	0	1	0	0	0
green	2	4	0	0	0	1	0	0	1	0	0
green	3	6	1	0	0	1	0	0	0	1	0
green	4	3	1	0	0	1	0	0	0	0	1
orange	1	10	1	0	0	0	1	1	0	0	0
orange	2	5	1	0	0	0	1	0	1	0	0
orange	3	2	1	0	0	0	1	0	0	1	0
orange	4	1	1	0	0	0	1	0	0	0	1
red	1	16	0	1	0	0	0	1	0	0	0
red	2	15	0	1	0	0	0	0	1	0	0
red	3	14	0	1	0	0	0	0	0	1	0
red	4	13	0	1	0	0	0	0	0	0	1

Regression framework for previous example

$CO2_{c,t} = B_0 + B_1 policy_{c,t} + B_2 red_c + B_3 green_c + B_4 blue_c +$
$B_5 year2_t + B_6 year3_t + B_7 year4_t + e_{c,t}$

(orange is the omitted category)
(year1 = omitted year)

```
areg co2 policy,a(country)

tabulate year,gen(y_)
areg co2 poicy,a(country)
areg co2 policy y_,a(country)
```

```
encode country,gen(cntry)
xtset cntry year

xtreg co2 policy
```

```
. encode country,gen(cntry)
. reg co2 policy i.cntry

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  4,    11) =    8.85
       Model |  259.392857     4  64.8482143           Prob > F      =  0.0019
    Residual |  80.6071429    11  7.32792208           R-squared     =  0.7629
-------------+------------------------------           Adj R-squared =  0.6767
       Total |         340    15  22.6666667           Root MSE      =   2.707

------------------------------------------------------------------------------
         co2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      policy |  -1.285714    2.04631    -0.63   0.543    -5.789612    3.218183
             |
       cntry |
      green  |  -4.178571   1.981331    -2.11   0.059    -8.539452    .1823087
     orange  |  -4.285714   2.453439    -1.75   0.108    -9.685698    1.114269
        red  |   4.428571   1.981331     2.24   0.047     .0676913    8.789452
             |
       _cons |   10.07143    1.44696     6.96   0.000     6.886692    13.25617
------------------------------------------------------------------------------
```

```
. areg co2 policy,a(country)

Linear regression, absorbing indicators          Number of obs   =         16
                                                 F(   1,     11) =       0.39
                                                 Prob > F        =     0.5426
                                                 R-squared       =     0.7629
                                                 Adj R-squared   =     0.6767
                                                 Root MSE        =     2.7070

------------------------------------------------------------------------------
         co2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      policy |  -1.285714    2.04631    -0.63   0.543    -5.789612    3.218183
       _cons |     9.0625   1.122269     8.08   0.000     6.592404     11.5326
-------------+----------------------------------------------------------------
     country |          F(3, 11) =      5.419   0.016          (4 categories)
```

```
. areg co2 policy i.year,a(country)

Linear regression, absorbing indicators         Number of obs   =         16
                                                F(   4,      8) =       8.10
                                                Prob > F        =     0.0065
                                                R-squared       =     0.9514
                                                Adj R-squared   =     0.9088
                                                Root MSE        =     1.4374

------------------------------------------------------------------------------
         co2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      policy |   2.823529   1.394502     2.02   0.077    -.3921987    6.039257
             |
        year |
          2  |      -2.75   1.016409    -2.71   0.027    -5.093844   -.4061555
          3  |  -4.955882   1.074536    -4.61   0.002    -7.433767   -2.477997
          4  |  -6.411765   1.232578    -5.20   0.001    -9.254094   -3.569436
             |
       _cons |   10.79412   .7988015    13.51   0.000     8.952078    12.63616
-------------+----------------------------------------------------------------
     country |          F(3, 8) =     26.214   0.000           (4 categories)
```

```
. . xtset cntry year
       panel variable:  cntry (strongly balanced)
        time variable:  year, 1 to 4
                delta:  1 unit

xtreg co2 policy,fe

Fixed-effects (within) regression               Number of obs      =        16
Group variable: cntry                           Number of groups   =         4

R-sq:  within  = 0.0346                         Obs per group: min =         4
       between = 0.8048                                        avg =       4.0
       overall = 0.4125                                        max =         4

                                                F(1,11)            =      0.39
corr(u_i, Xb)  = 0.6482                         Prob > F           =    0.5426

------------------------------------------------------------------------------
         co2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      policy |  -1.285714    2.04631    -0.63   0.543    -5.789612    3.218183
       _cons |     9.0625   1.122269     8.08   0.000     6.592404     11.5326
-------------+----------------------------------------------------------------
     sigma_u |  4.1379666
     sigma_e |  2.7070135
         rho |  .70029794   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(3, 11) =     5.42               Prob > F = 0.0156
```

Residual Votes Attributable to Technology

Stephen Ansolabehere
Massachusetts Institute of Technology
Charles Stewart III
Massachusetts Institute of Technology

$$\text{residual vote pct.}_{c,t} = (\text{turnout}_{c,t} - \text{pvotes}_{c,t})/\text{turnout}_{c,t}$$

TABLE 2

Residual Vote in Presidential Elections, by Machine Type, U.S Cou
1988–2000. (Numbers in parentheses are standard deviations

	Counties			Voters	
Machine Type	Pres.	Gov.	Sen.	Pres.	Gov.
Paper ballot	1.8%	3.3%	3.6%	1.9%	3.2%
	(2.1%)	(2.0%)	(3.2%)	(2.0%)	(2.2%)
Lever machine	1.9%	5.1%	9.5%	1.8%	4.2%
	(1.8%)	(3.1%)	(5.3%)	(1.8%)	(2.6%)
Punch card	2.9%	3.3%	4.7%	2.5%	3.3%
	(1.1%)	(2.1%)	(3.1%)	(1.5%)	(1.6%)
Optically scanned	2.1%	3.1%	3.4%	1.6%	2.1%
	(2.7%)	(1.9%)	(3.8)%	(2.4%)	(1.7%)
Electronic (DRE)	3.0%	4.3%	8.2%	2.5%	3.7%
	(3.0%)	(1.2%)	(4.0%)	(3.6%)	(1.9%)
Mixed	2.0%	5.0%	6.1%	1.5%	3.0%
	(1.7%)	(2.8%)	(3.9%)	(1.3%)	(1.5%)
Overall	2.2%	3.6%	5.5%	2.1%	3.2%
	(2.3%)	(2.3%)	(4.5%)	(1.9%)	(1.9%)

Residual Vote Multivariate Analysis, Presidential, Gubernatorial, and Senatorial Elections, 1988–2000

	President	President
Equipment effects:		
Punch card	.0082	.0077
	(.0015)	(.0005)
Lever machine	Excluded equip. category	Excluded equip. category
Paper	−.014	−.0012
	(.002)	(.0014)
Optical scan	−.0045	.00071
	(.0014)	(.00070)
Electronic (DRE)	.0022	.0080
	(.0015)	(.0010)
Shift in tech.	.0010	.00005
	(.0007)	(.00067)
Log (turnout)	.0095	−.0004
	(.0026)	(.0001)
Gov. or Sen. on ballot	−.0011	−.003
	(.0007)	(.001)
Senator	—	—
Percent Over 65	—	.047
		(.008)
Percent 18–24	—	−.012
		(.009)
Percent White	—	−.030
		(.002)
Percent Hispanic	—	.011
		(.004)
Median Income (10,000s)	—	−.002
		(.001)
Constant	−.11	.025
	(.03)	(.002)
N	8,982	8,982
R^2	.79	.14
Fixed effect: (not shown)	Year × State County	Year × State
Number of categories	3,346	—