Bivariate Relationships

17.871

2012

Testing associations (not causation!)

- Continuous data
  - □ Scatter plot (always use first!)
  - □ (Pearson) correlation coefficient (rare, should be rarer!)
  - □ (Spearman) rank-order correlation coefficient (rare)
  - □ Regression coefficient (common)
- Discrete data
  - □ Cross tabulations
  - □ χ^2
  - □ Gamma, Beta, etc.

Continuous DV, continuous EV

- Dependent Variable: DV
- Explanatory (or independent) Variable: EV

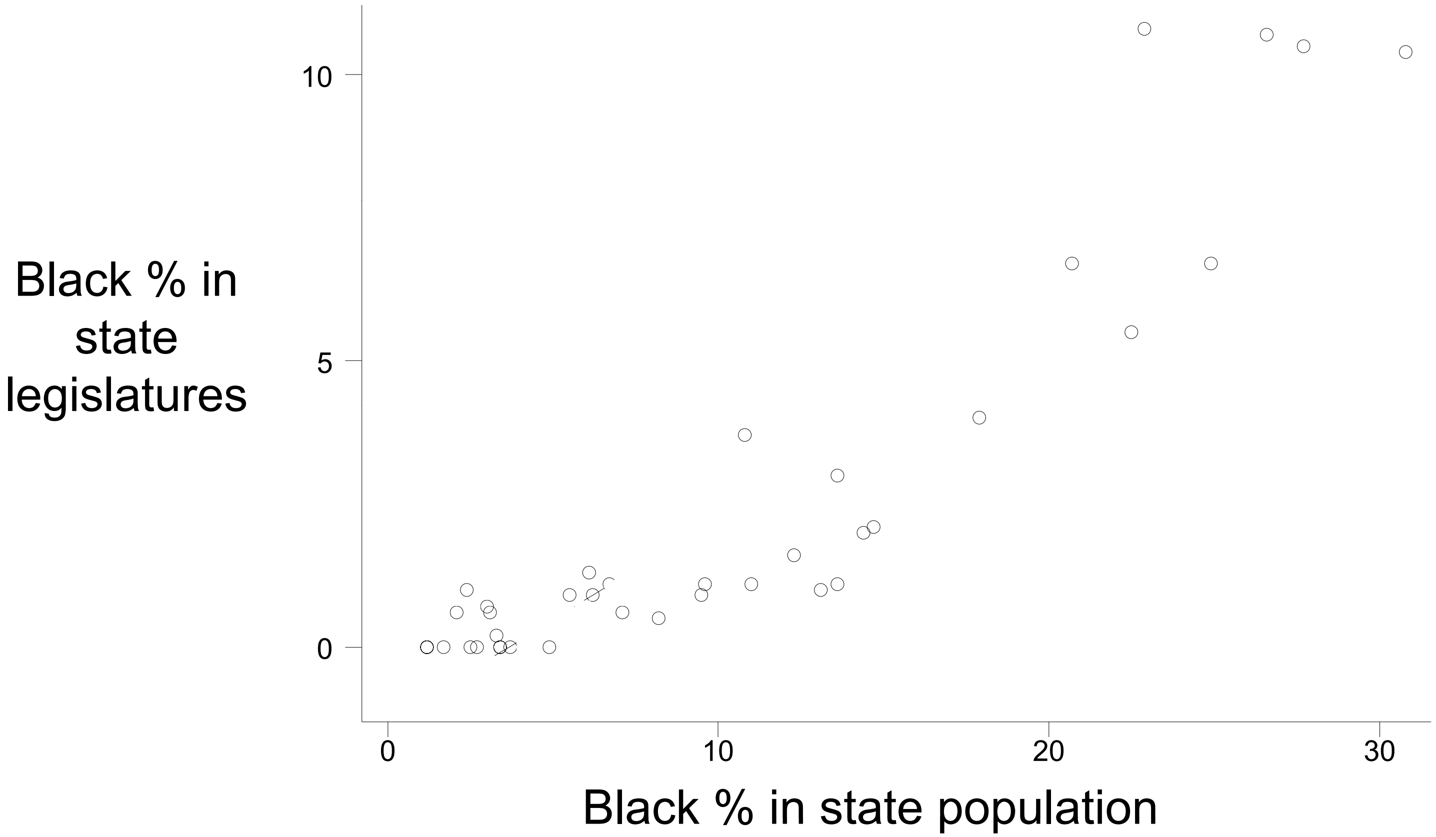
- Example: What is the relationship between Black percent in state legislatures and black percent in state populations

Regression interpretation

Three key things to learn (today)

1. Where does regression come from
2. To interpret the regression coefficient
3. To interpret the confidence interval
    - We will learn how to calculate confidence intervals in a couple of weeks

Linear Relationship between African American Population & Black Legislators

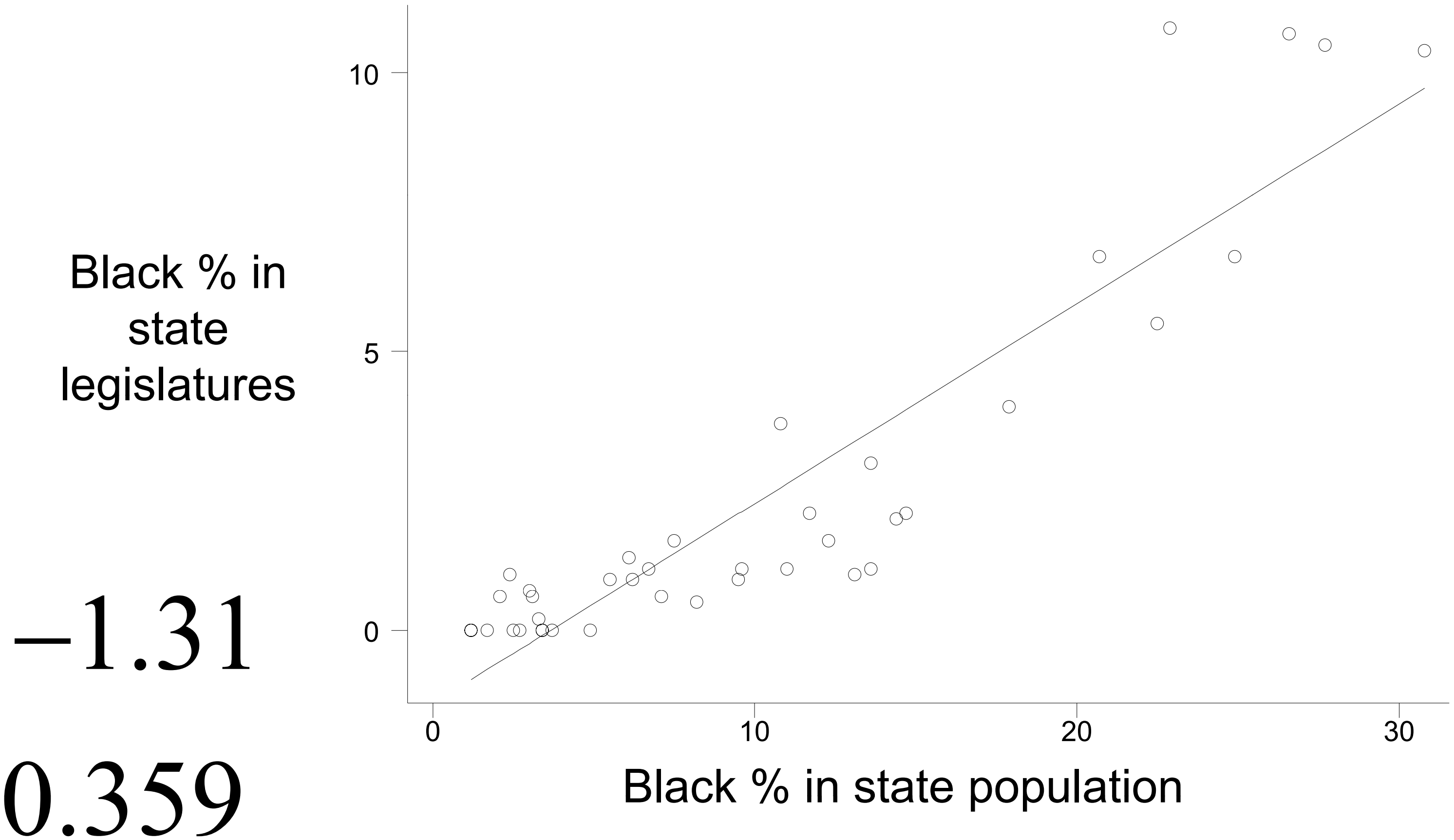


The linear relationship between two variables

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Regression quantifies how one variable can be described in terms of another

Linear Relationship between African American Population & Black Legislators

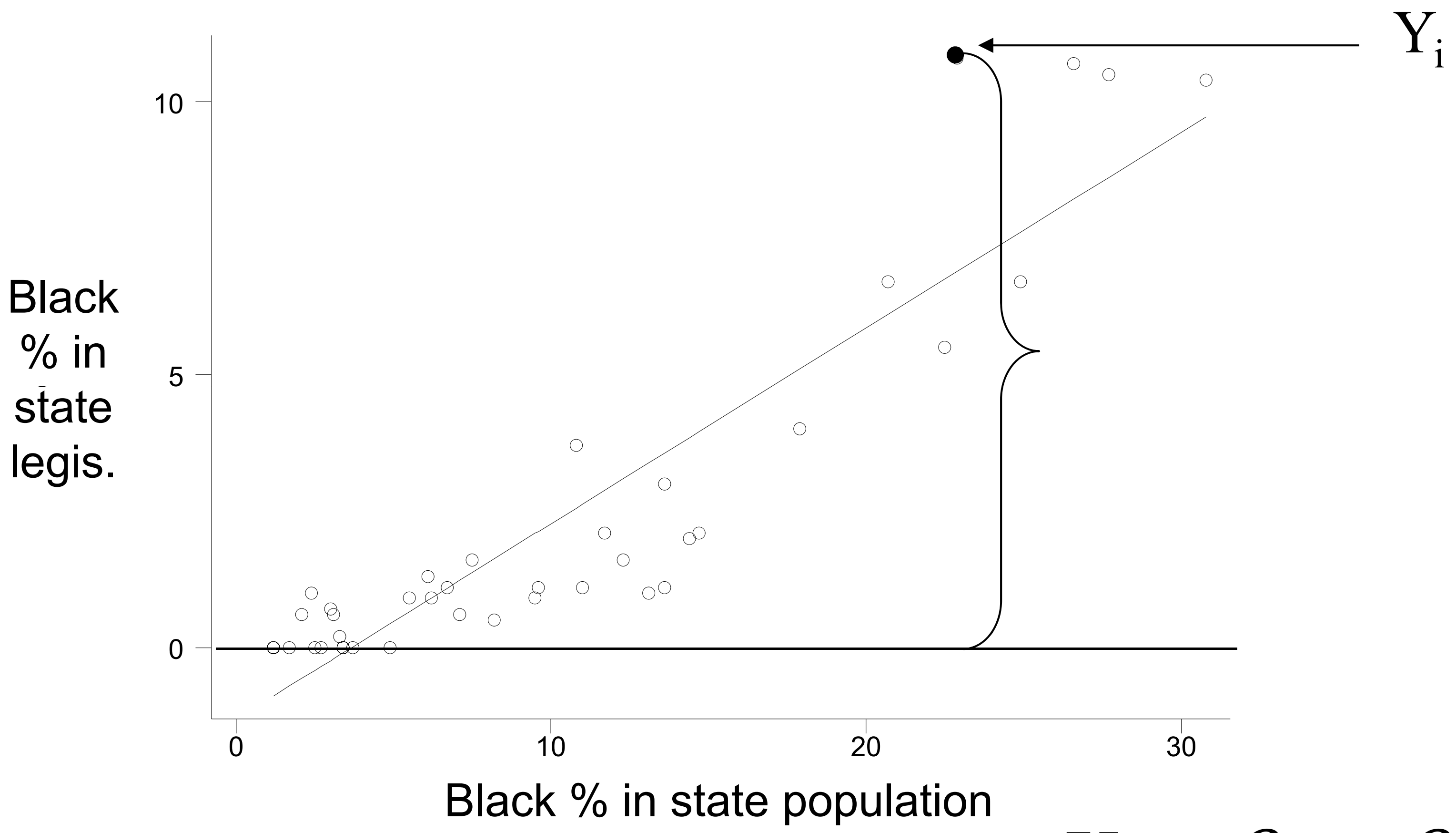


$$\hat{\beta}_0 = -1.31$$

$$\hat{\beta}_1 = 0.359$$

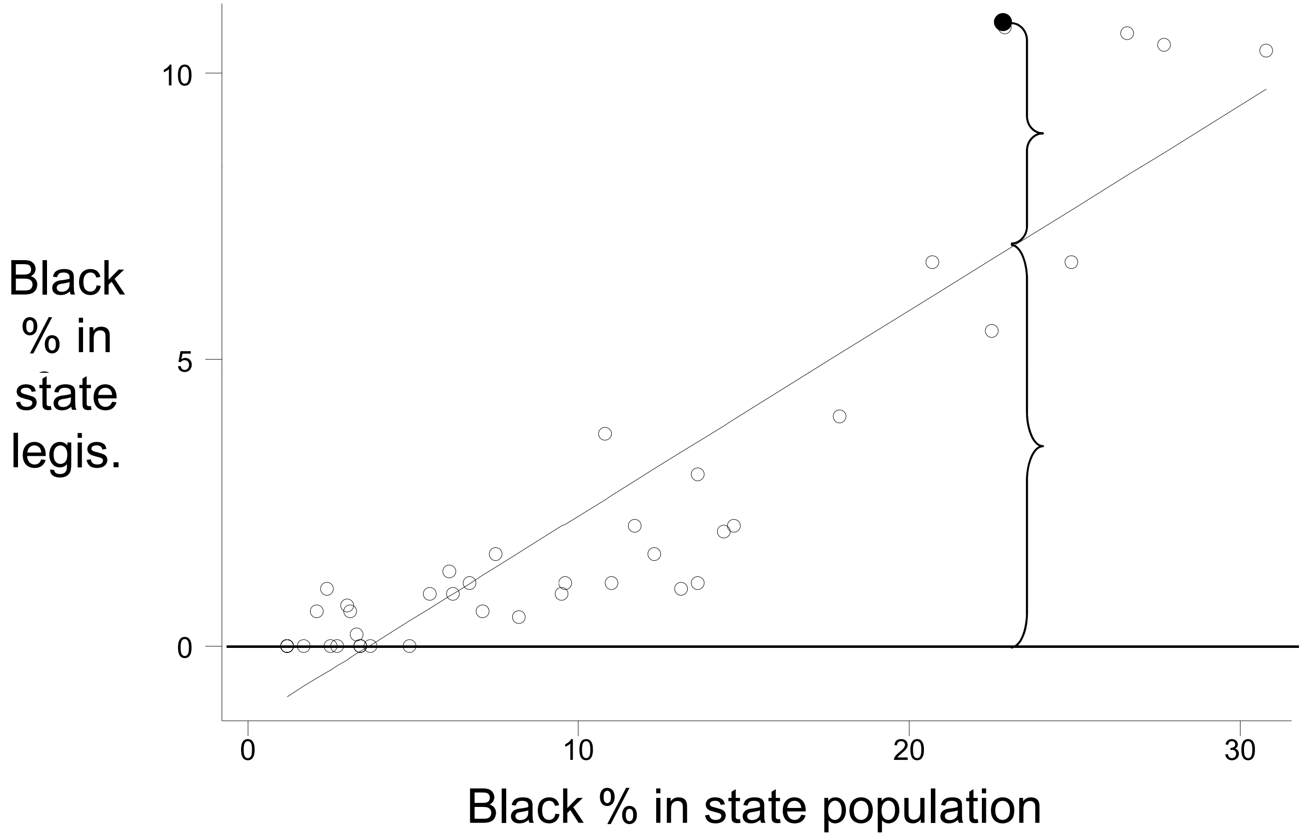
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

How did we get that line?
1. Pick a value of Y_i



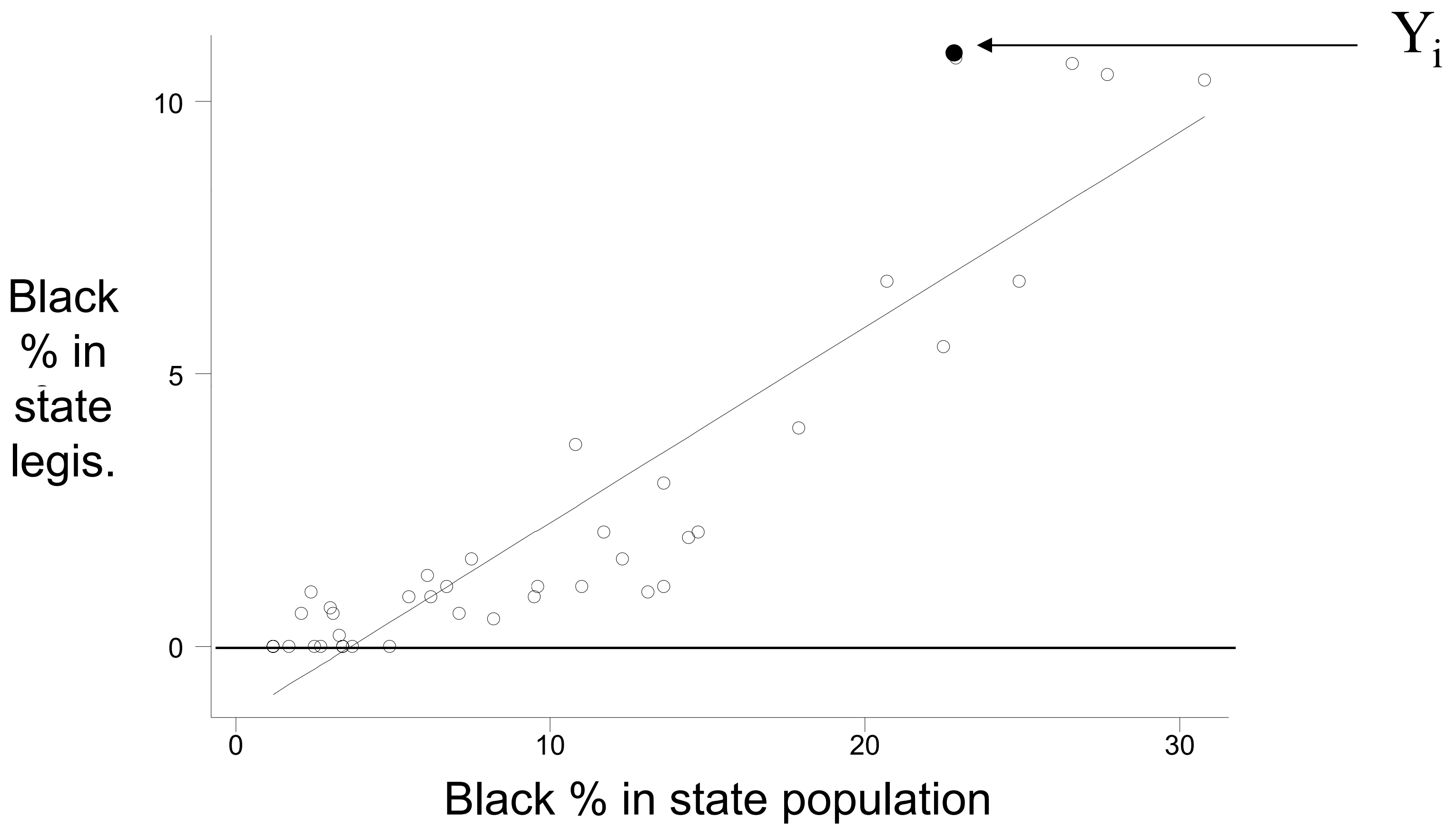
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

How did we get that line?
2. Decompose Y_i into two parts



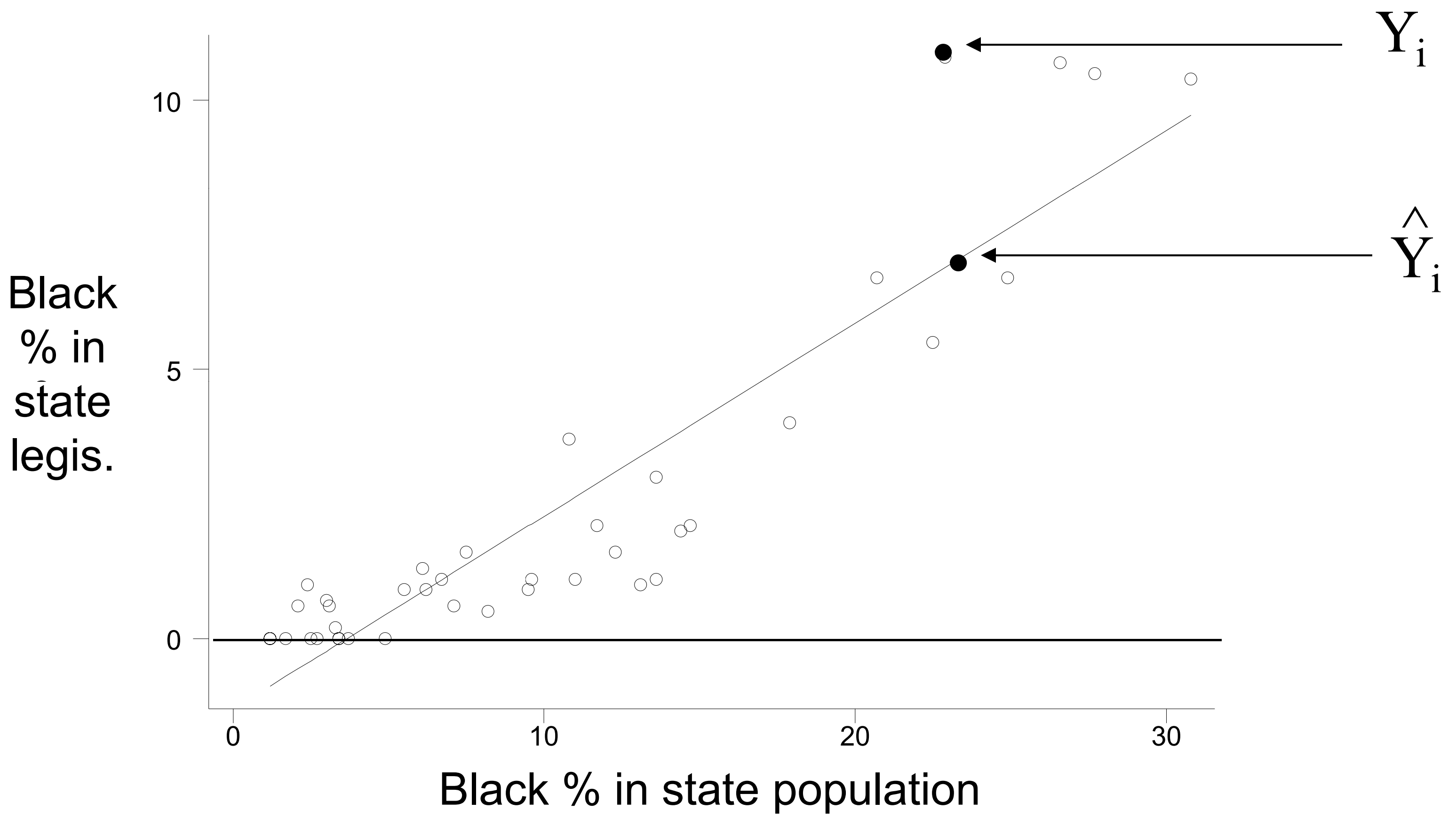
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

How did we get that line?
3. Label the points



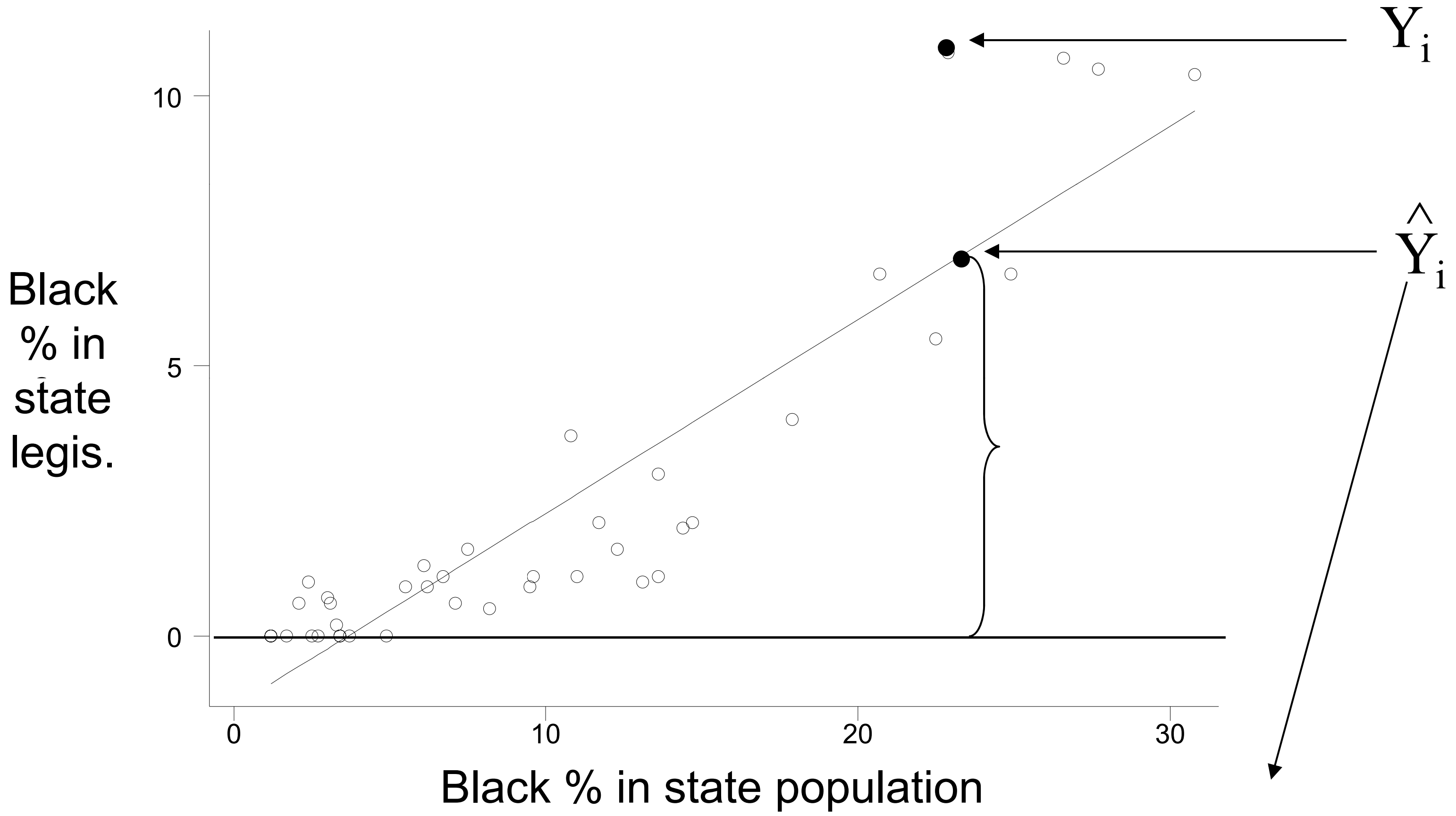
$$Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$$

How did we get that line?
3. Label the points



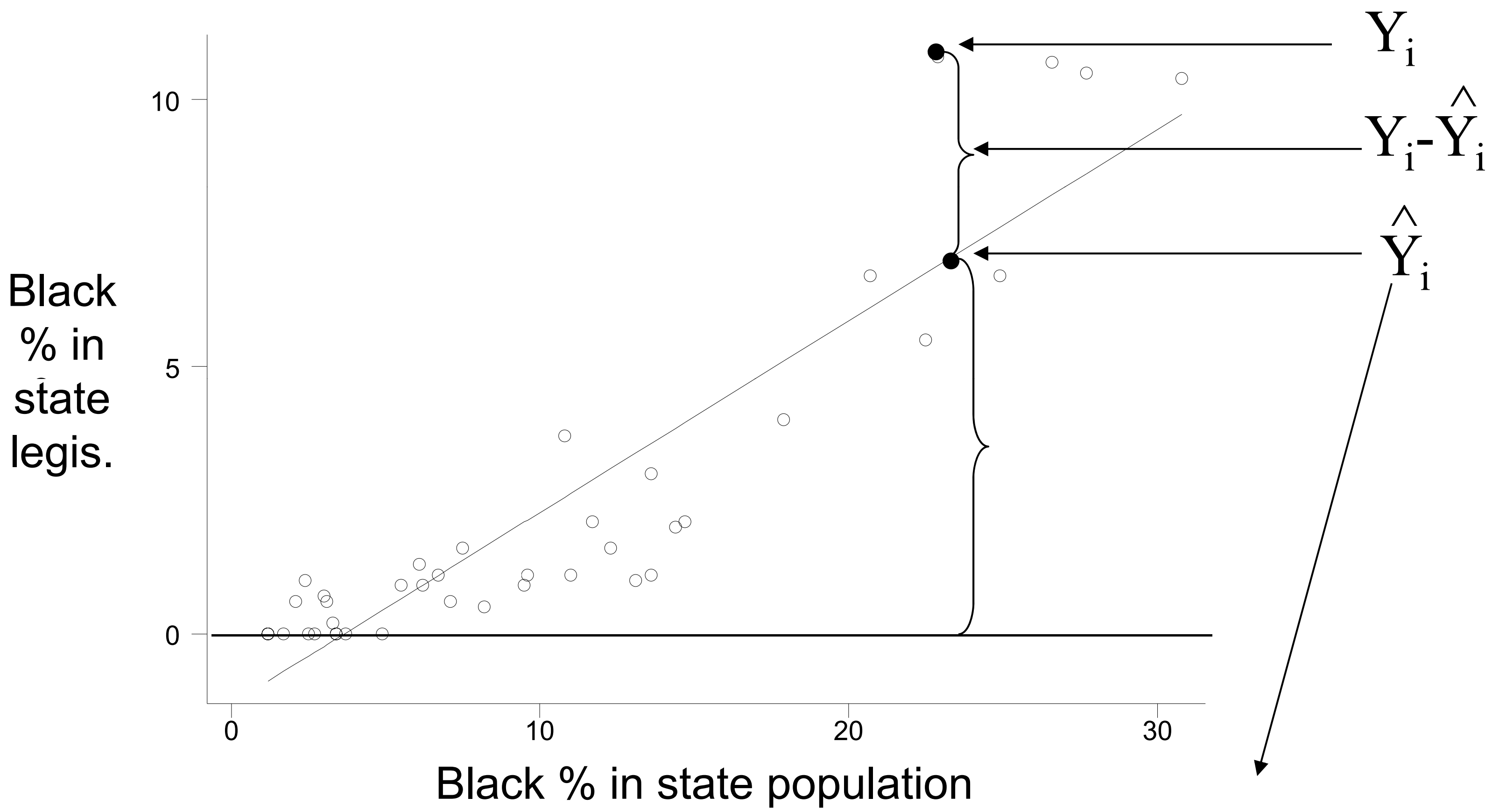
$$Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$$

How did we get that line?
3. Label the points



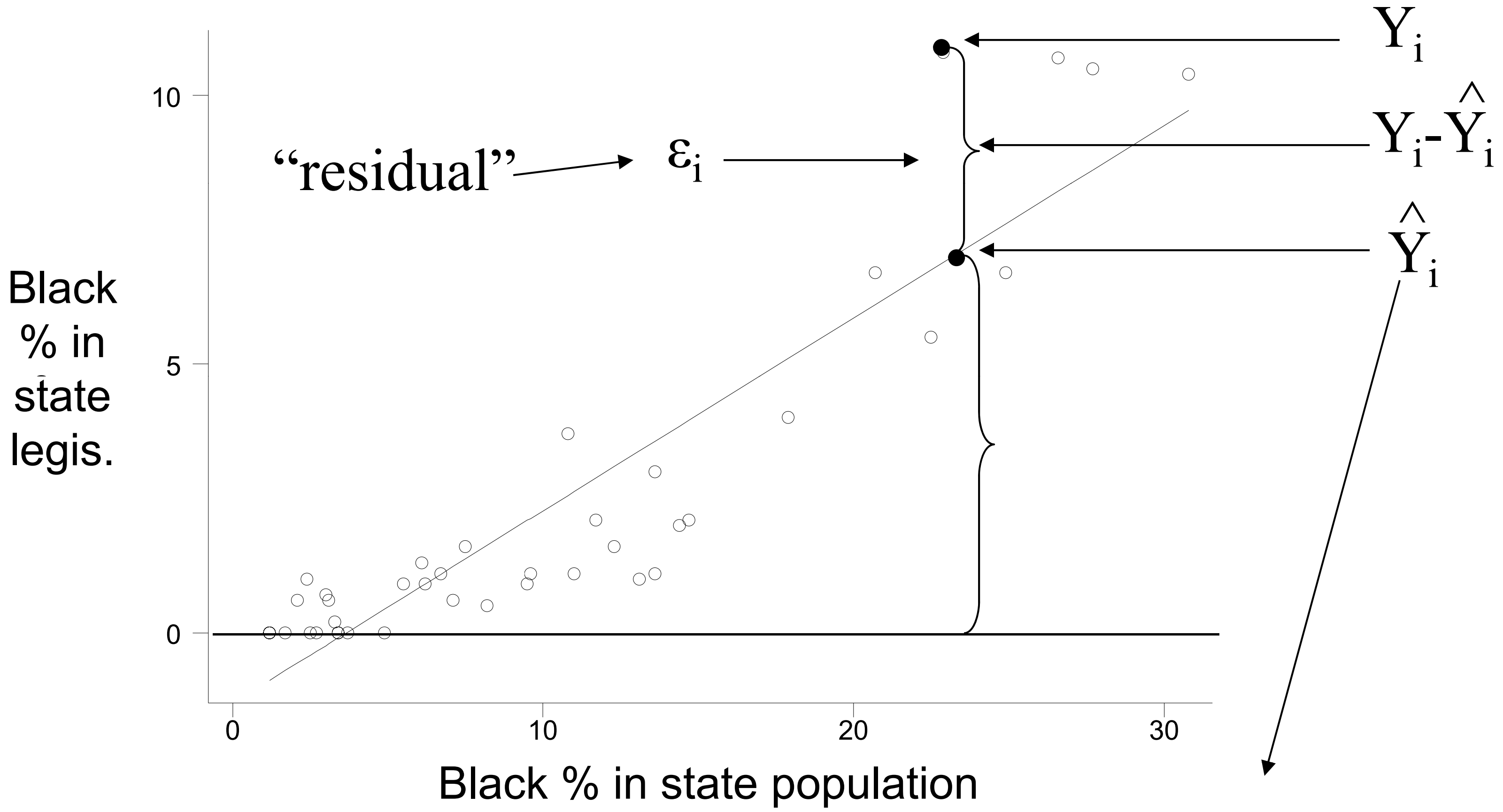
$$Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$$

How did we get that line?
3. Label the points



$$Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$$

How did we get that line?
3. Label the points



$$Y_i = (\beta_0 + \beta_1 X_i) + \varepsilon_i$$

What is ε_i? (sometimes u_i)

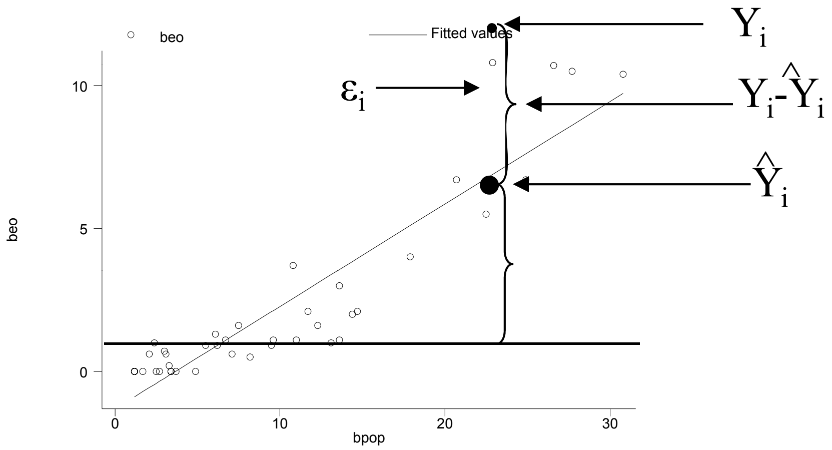
- Wrong functional form
- Measurement error
- Stochastic component in Y
- Unmeasured influences on Y

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The Method of Least Squares

Pick β_0 and β_1 to minimize $\sum_{i=1}^{n} \varepsilon_i^2$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \text{ or}$$



$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Solve for $\frac{\partial \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}{\partial \beta_1} = 0$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (\bar{Y} - Y_i)(\bar{X} - X_i)}{\sum_{i=1}^{n} (\bar{X} - X_i)^2} \text{ or}$$

$$\frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}$$

Remember this for the problem set!

Regression commands in STATA

- `reg` *`depvar expvars`*
  - E.g., `reg y x`
  - E.g., `reg beo bpop`
- Making predictions from regression lines
  - `predict` *`newvar`*
  - `predict` *`newvar,`* `resid`
    - *newvar* will now equal ε_i

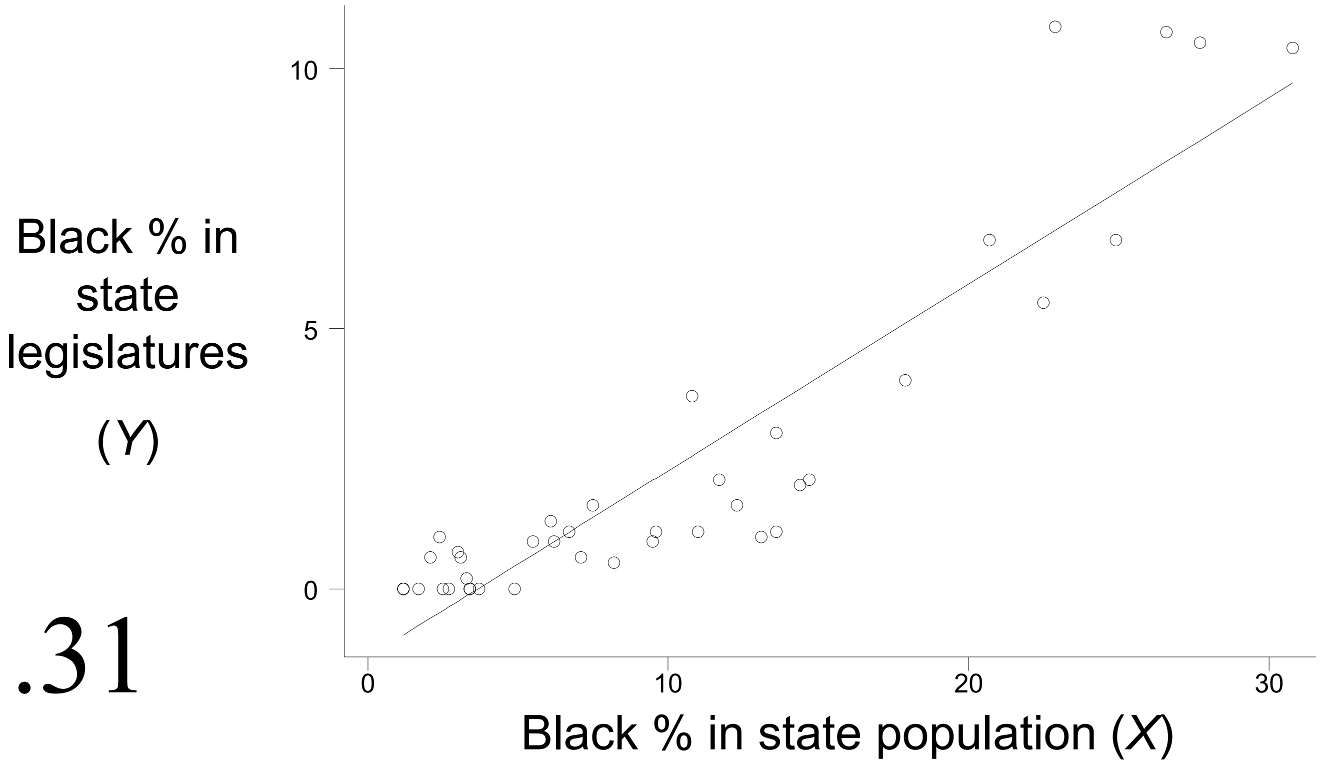
Black elected officials example

```
. reg beo bpop

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =  202.56
       Model |   351.26542     1   351.26542           Prob > F      =  0.0000
    Residual |  67.6326195    39  1.73416973           R-squared     =  0.8385
-------------+------------------------------           Adj R-squared =  0.8344
       Total |   18.898039    40   10.472451           Root MSE      =  1.3169

------------------------------------------------------------------------------
         beo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bpop |   .3586751   .0251876    14.23   0.000     .3075284    .4094219
       _cons |  -1.314892   .3277508    -4.01   0.000    -1.977831   -.6519535
------------------------------------------------------------------------------
```

Always include interpretation in your presentations and papers

Interpretation: a one percentage point increase in black population leads to a .36 percentage point increase in black composition in the legislature

The Linear Relationship between African American Population & Black Legislators



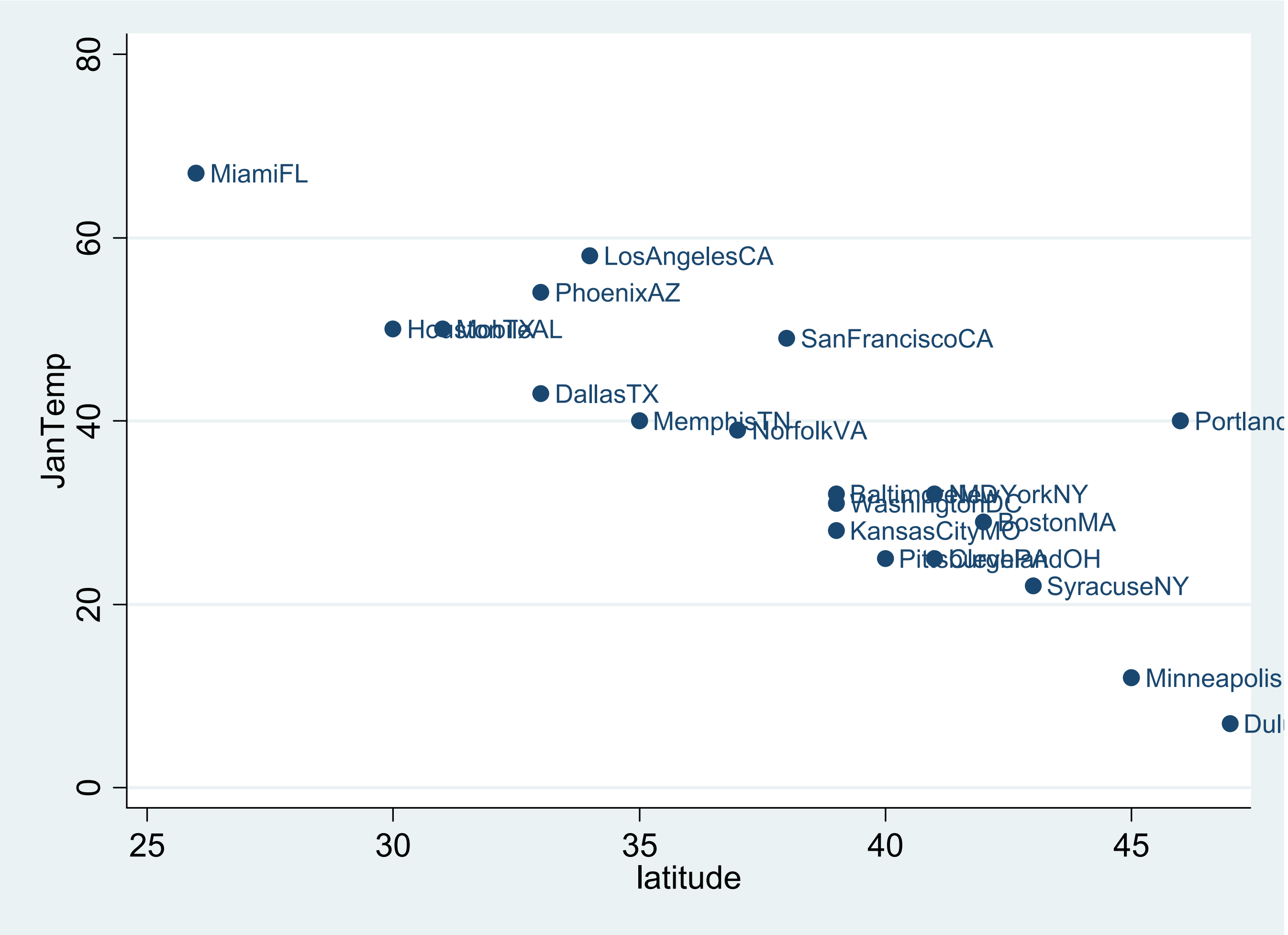
$$\beta_0 = -1.31$$

$$\beta_1 = 0.359$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

More regression examples

Temperature and Latitude



```
scatter JanTemp latitude, mlabel(city)
```

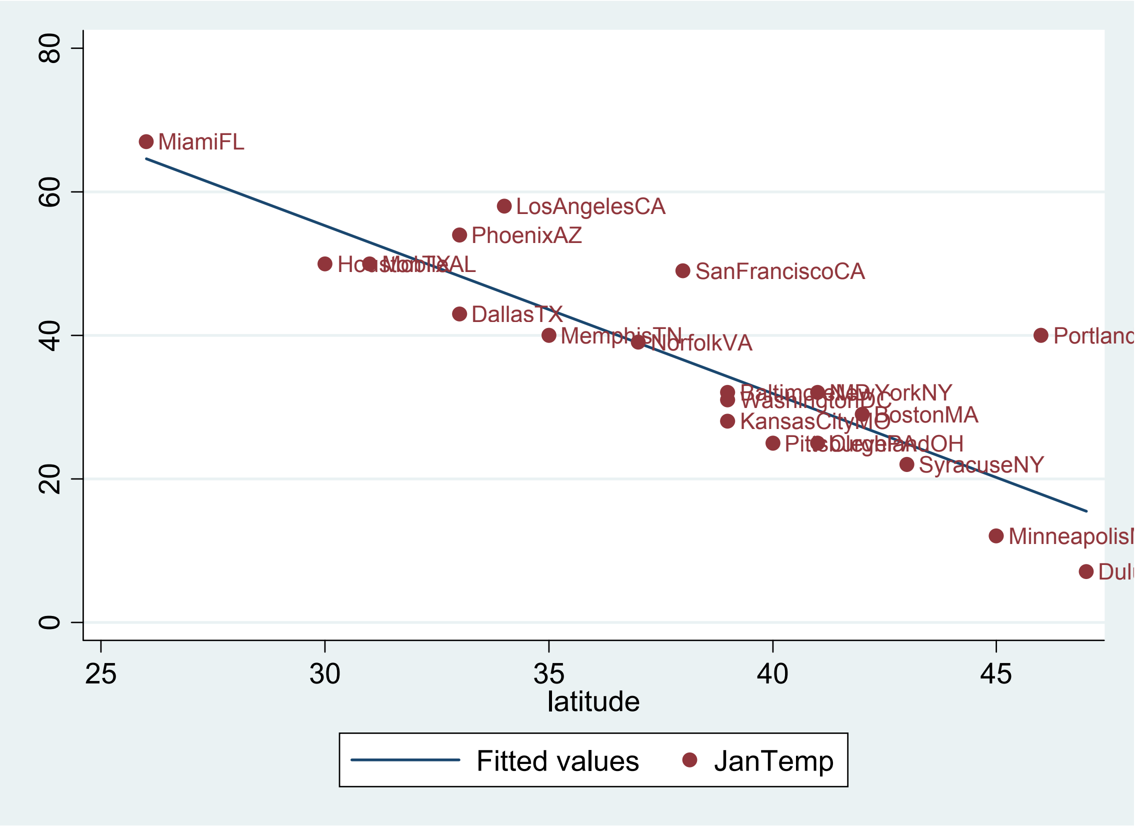
```
. reg jantemp latitude

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =   49.34
       Model |  3250.72219     1  3250.72219           Prob > F      =  0.0000
    Residual |  1185.82781    18  65.8793228           R-squared     =  0.7327
-------------+------------------------------           Adj R-squared =  0.7179
       Total |     4436.55    19  233.502632           Root MSE      =  8.1166

------------------------------------------------------------------------------
     jantemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    latitude |  -2.341428   .3333232    -7.02   0.000    -3.041714   -1.641142
       _cons |   125.5072   12.77915     9.82   0.000     98.65921    152.3552
------------------------------------------------------------------------------
```

Interpretation: a one point increase in latitude is associated with a 2.3 decrease in average temperature (in Fahrenheit).

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

How to add a regression line:

Stata command: `lfit`



```
scatter JanTemp latitude, mlabel(city) || lfit JanTemp latitude
```

or often better

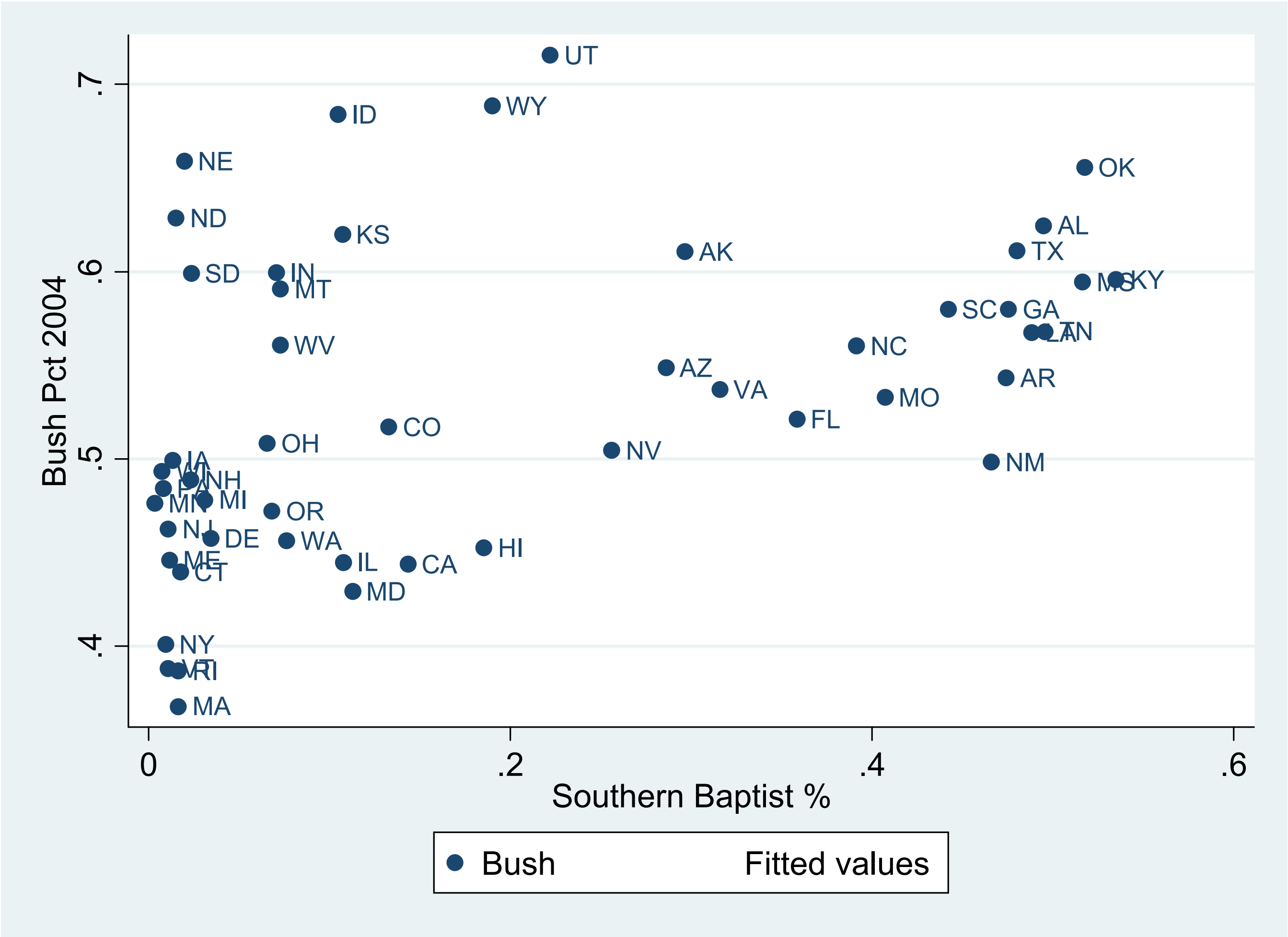
```
scatter JanTemp latitude, mlabel(city) m(i) || lfit JanTemp latitude
```

Presenting regression results

Brief aside

- First, show scatter plot
  - Label data points (if possible)
  - Include best-fit line
- Second, show regression table
  - Assess statistical significance with confidence interval or p-value
  - Assess robustness to control variables (internal validity: nonrandom selection)

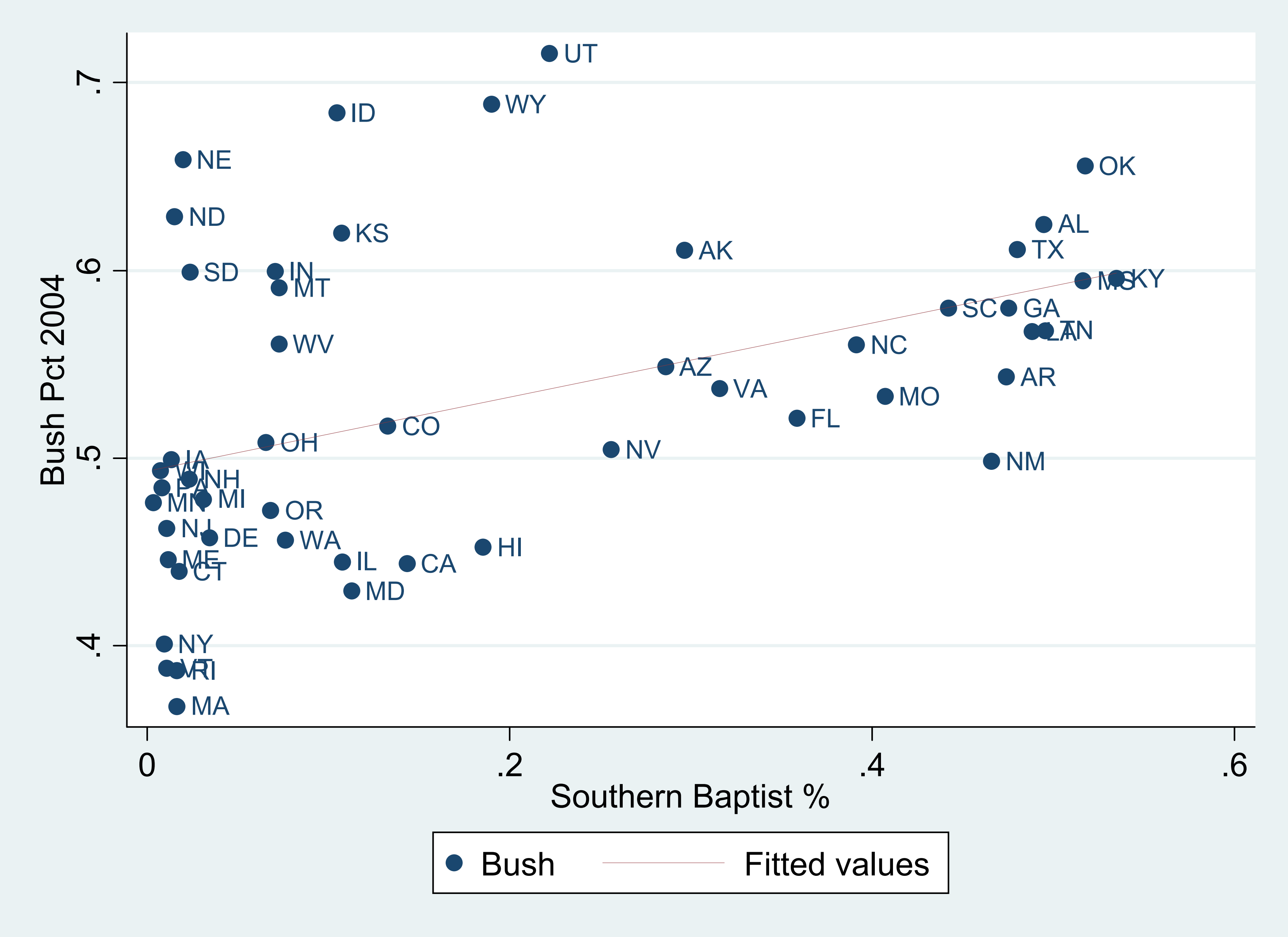
Bush vote and Southern Baptists



```
. reg bush sbc_mpct [aw=votes]
(sum of wgt is   1.2207e+08)

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   40.18
       Model |  .118925068     1  .118925068           Prob > F      =  0.0000
    Residual |  .142084951    48  .002960103           R-squared     =  0.4556
-------------+------------------------------           Adj R-squared =  0.4443
       Total |  .261010018    49  .005326735           Root MSE      =  .05441

------------------------------------------------------------------------------
        bush |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sbc_mpct |    .261779   .0413001     6.34   0.000     .1787395    .3448185
       _cons |   .4563507   .0112155    40.69   0.000     .4338004    .4789011
------------------------------------------------------------------------------
```

Coefficient interpretation:

- A one percentage point increase in Baptist percentage is associated with a .26 percentage point increase in Bush vote share at the state level.


Bush Pct 2004
.4
.5
.6
.7
0
.2
.4
.6
Southern Baptist %
UT
ID
WY
NE
OK
ND
KS
AL
AK
TX
SD
IN
MT
MS
KY
SC
GA
LA
TN
WV
NC
AZ
VA
AR
MO
FL
CO
OH
NV
NM
IA
WI
PA
NH
MN
MI
OR
NJ
DE
WA
HI
ME
CT
IL
CA
MD
NY
VT
RI
MA
● Bush
Fitted values

Interpreting confidence interval

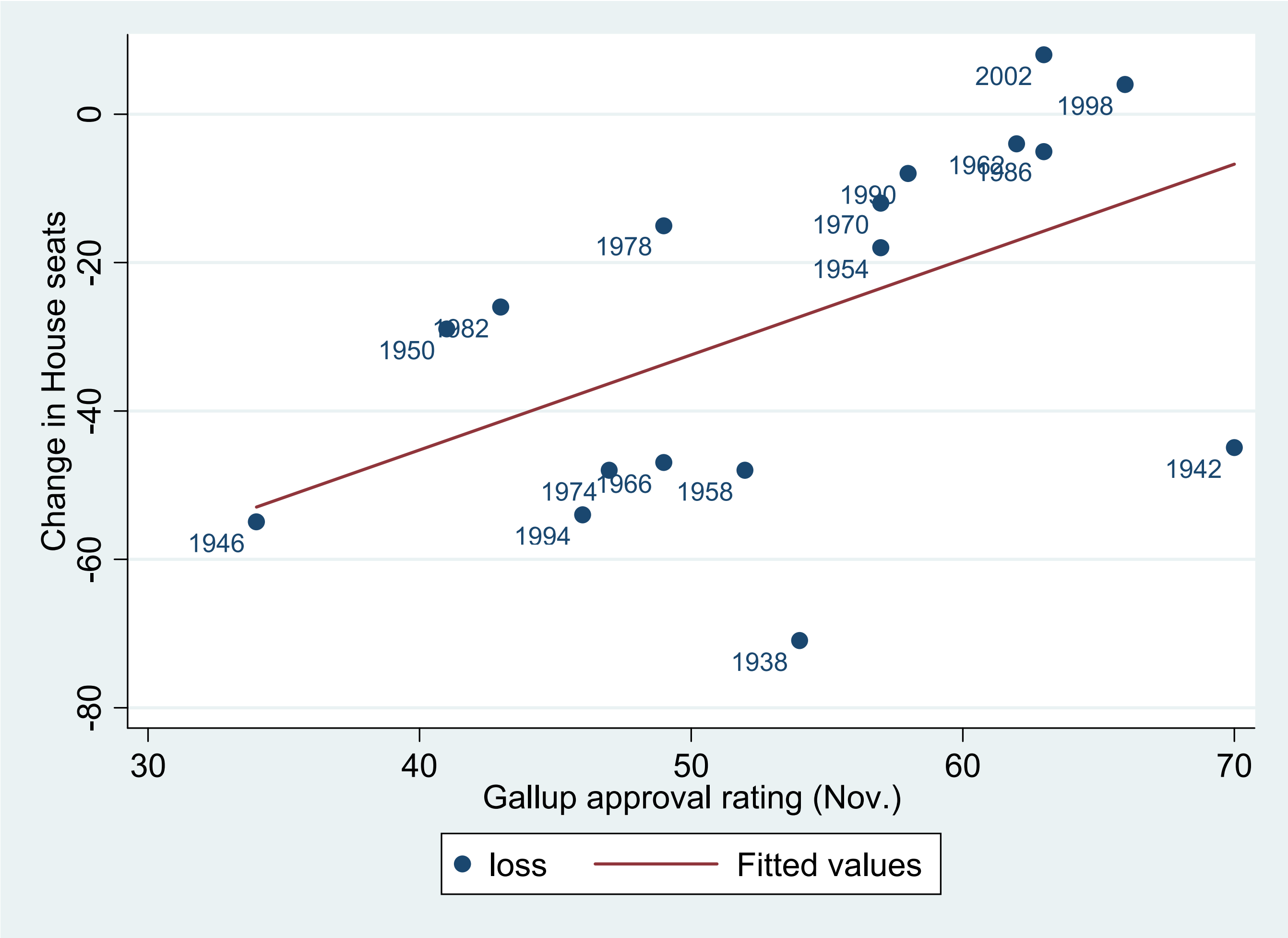
```
. reg bush sbc_mpct [aw=votes]
(sum of wgt is   1.2207e+08)

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   40.18
       Model |  .118925068     1  .118925068           Prob > F      =  0.0000
    Residual |  .142084951    48  .002960103           R-squared     =  0.4556
-------------+------------------------------           Adj R-squared =  0.4443
       Total |  .261010018    49  .005326735           Root MSE      =  .05441

------------------------------------------------------------------------------
        bush |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sbc_mpct |    .261779   .0413001     6.34   0.000     .1787395    .3448185
       _cons |   .4563507   .0112155    40.69   0.000     .4338004    .4789011
------------------------------------------------------------------------------
```

Coefficient interpretation:

- A 1 percentage point increase in Baptist percentage is associated with a .26 percentage point increase in Bush vote share at the state level.

Confidence interval interpretation

- The 95% confidence interval lies between .18 and .34.


Change in House seats
0
-20
-40
-60
-80
30
40
50
60
70
Gallup approval rating (Nov.)
2002
1998
1962
1986
1990
1970
1954
1978
1982
1950
1942
1974
1966
1958
1994
1946
1938
● loss
Fitted values

```
. reg loss gallup

      Source |       SS       df       MS              Number of obs =      17
-------------+------------------------------           F(  1,     15) =    5.70
       Model |  2493.96962     1  2493.96962           Prob > F      =  0.0306
    Residual |  6564.50097    15  437.633398           R-squared     =  0.2753
-------------+------------------------------           Adj R-squared =  0.2270
       Total |  9058.47059    16  566.154412           Root MSE      =   20.92

------------------------------------------------------------------------------
       Seats |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gallup |   1.283411     .53762     2.39   0.031     .1375011    2.429321
        cons |  -96.59926   29.25347    -3.30   0.005    -158.9516   -34.24697
------------------------------------------------------------------------------
```

Coefficient interpretation:

- A 1 percentage point increase in presidential approval is associated with an avg. of 1.28 more seats won by the president’s party in the midterm.

Confidence interval interpretation

- The 95% confidence interval lies between .14 and 2.43.

Additional regression in bivariate relationship topics

- Residuals
- Comparingcoefficients
- Functional form
- Goodness of fit (R^2 and SER)
- Correlation
- Discrete DV, discrete EV
- Using the appropriate graph/table

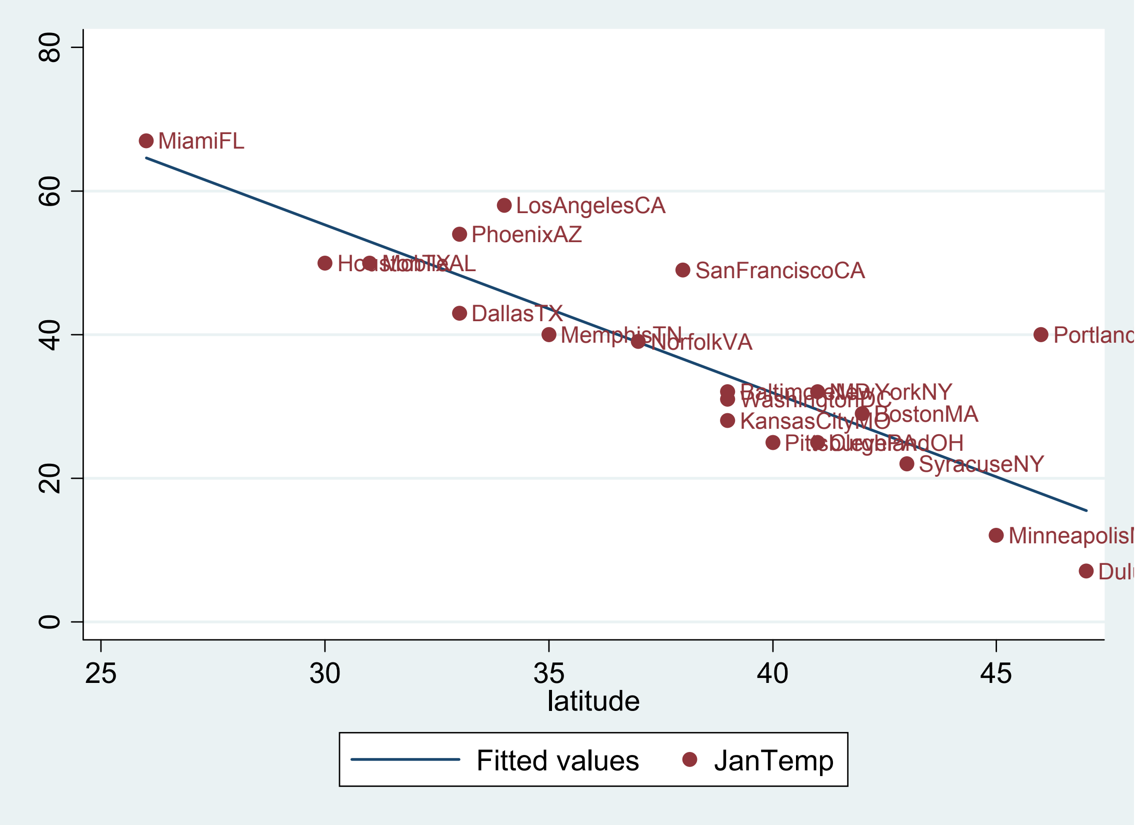
Residuals

Residuals

$$e_i = Y_i - B_0 - B_1 X_i$$

One important numerical property of residuals

- The sum of the residuals is zero

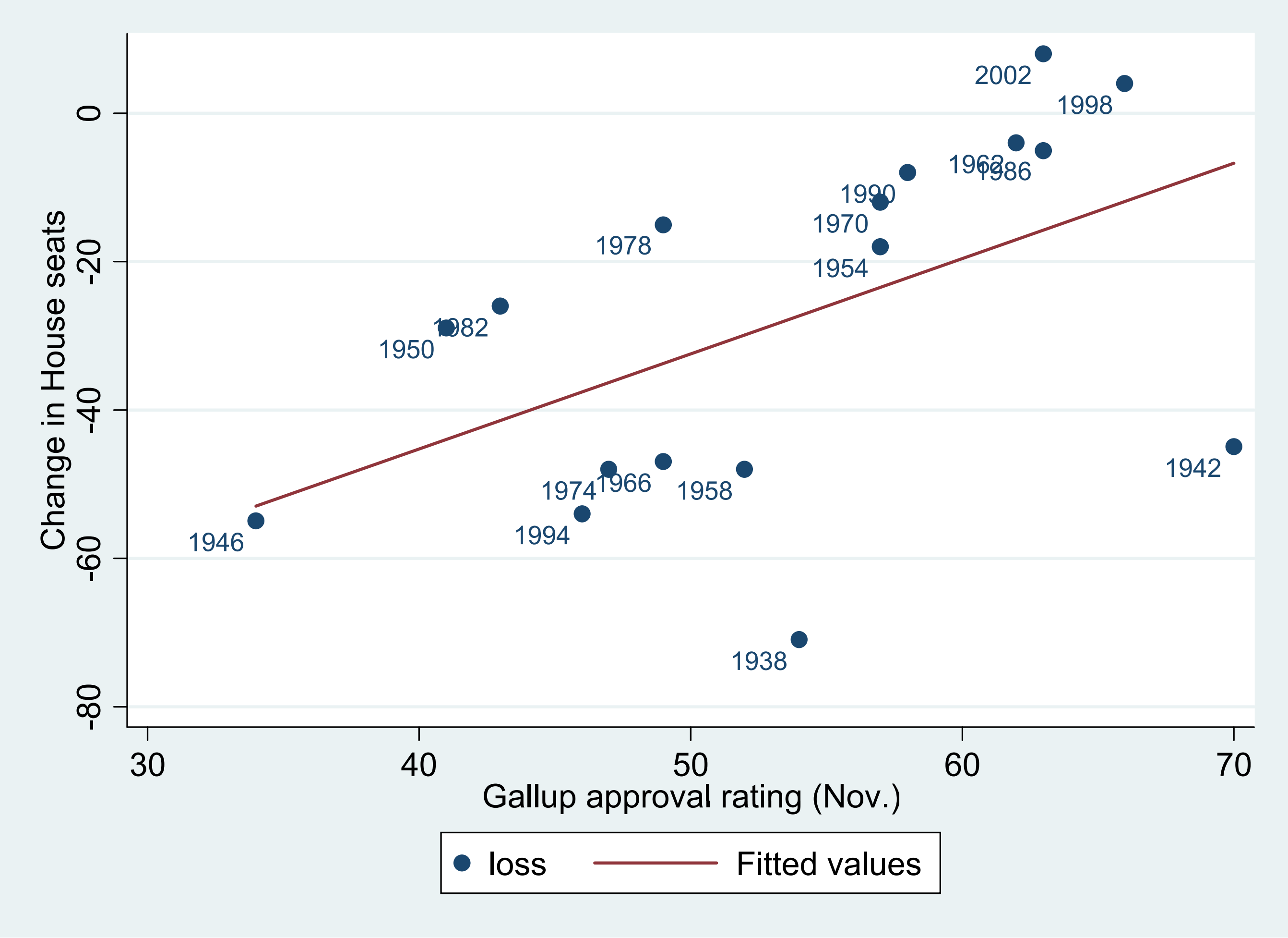


Generating predictions and residuals

```
. reg jantemp latitude

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =   49.34
       Model |  3250.72219     1  3250.72219           Prob > F      =  0.0000
    Residual |  1185.82781    18  65.8793228           R-squared     =  0.7327
-------------+------------------------------           Adj R-squared =  0.7179
       Total |     4436.55    19  233.502632           Root MSE      =  8.1166

------------------------------------------------------------------------------
     jantemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    latitude |  -2.341428   .3333232    -7.02   0.000    -3.041714   -1.641142
       _cons |   125.5072   12.77915     9.82   0.000     98.65921    152.3552
------------------------------------------------------------------------------

. predict py
(option xb assumed; fitted values)

. predict ry, resid
```

```
gsort -ry

. list city jantemp py ry

     +-------------------------------------------------+
     |          city   jantemp          py          ry |
     |-------------------------------------------------|
  1. |    PortlandOR        40     17.8015     22.1985 |
  2. | SanFranciscoCA      49    36.53293    12.46707 |
  3. |   LosAngelesCA      58    45.89864    12.10136 |
  4. |      PhoenixAZ      54    48.24007    5.759929 |
  5. |      NewYorkNY      32    29.50864    2.491357 |
     |-------------------------------------------------|
  6. |        MiamiFL      67    64.63007     2.36993 |
  7. |       BostonMA      29    27.16722    1.832785 |
  8. |      NorfolkVA      39    38.87436     .125643 |
  9. |     BaltimoreMD     32     34.1915     -2.1915 |
 10. |     SyracuseNY      22    24.82579   -2.825786 |
     |-------------------------------------------------|
 11. |       MobileAL      50    52.92293   -2.922928 |
 12. |   WashingtonDC      31     34.1915     -3.1915 |
 13. |      MemphisTN      40    43.55721   -3.557214 |
 14. |    ClevelandOH      25    29.50864   -4.508643 |
 15. |       DallasTX      43    48.24007   -5.240071 |
     |-------------------------------------------------|
 16. |      HoustonTX      50    55.26435   -5.264356 |
 17. |   KansasCityMO      28     34.1915     -6.1915 |
 18. |   PittsburghPA      25    31.85007   -6.850072 |
 19. |  MinneapolisMN      12    20.14293   -8.142929 |
 20. |       DuluthMN       7    15.46007    8.460073 |
     +-------------------------------------------------+
                                    -
```

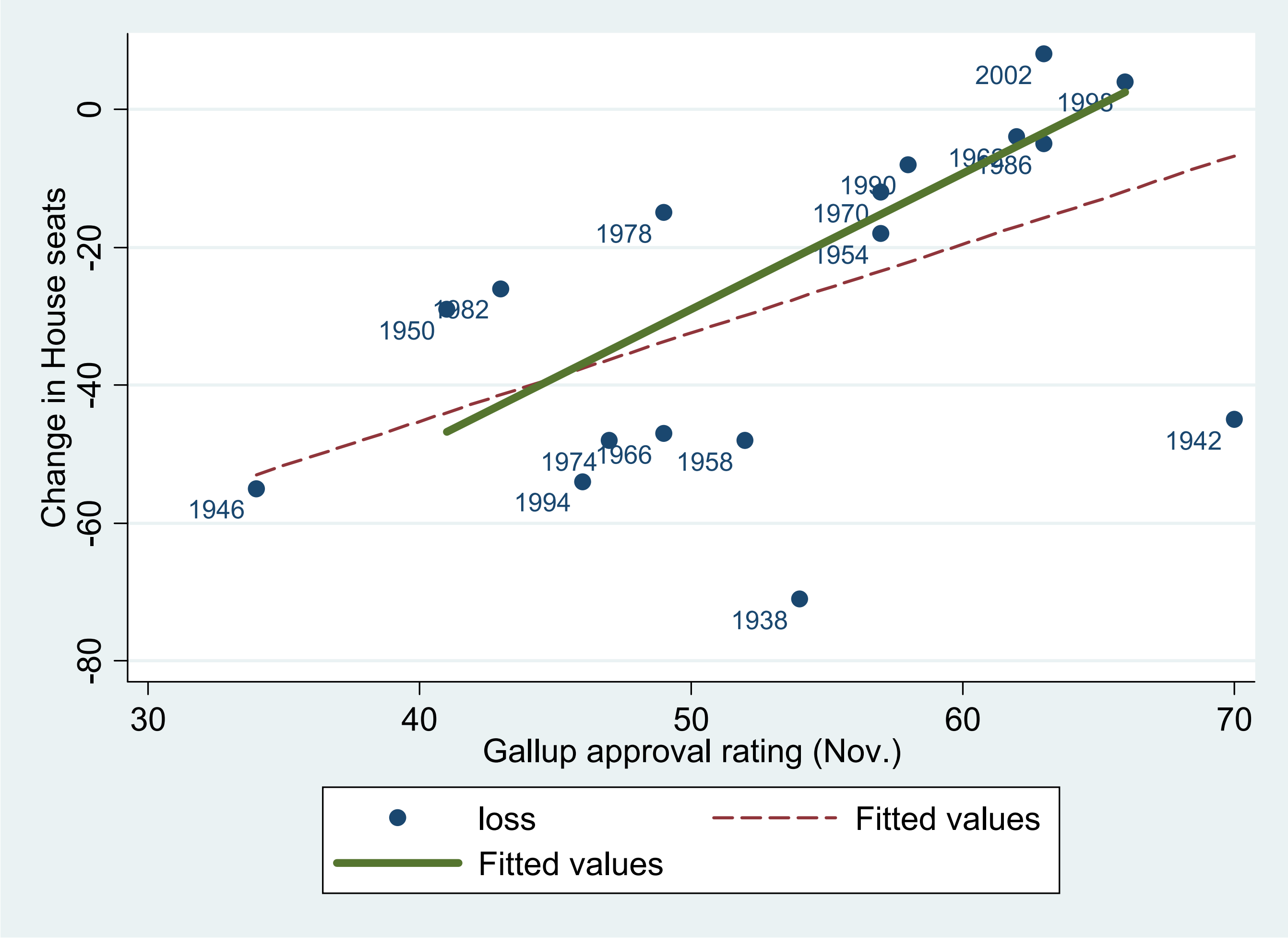
Use residuals to diagnose potential problems



```
. reg loss gallup

      Source |       SS       df       MS              Number of obs =      17
-------------+------------------------------           F(  1,    15) =    5.70
       Model |  2493.96962     1  2493.96962           Prob > F      =  0.0306
    Residual |  6564.50097    15  437.633398           R-squared     =  0.2753
-------------+------------------------------           Adj R-squared =  0.2270
       Total |  9058.47059    16  566.154412           Root MSE      =   20.92

------------------------------------------------------------------------------
       Seats |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gallup |   1.283411      53762     2.39   0.031      1375011    2.429321
       _cons |  -96.59926   29.25347    -3.30   0.005    -158.9516   -34.24697
------------------------------------------------------------------------------
```

```
. reg loss gallup if year>1946

      Source |       SS       df       MS              Number of obs =      14
-------------+------------------------------           F(  1,    12) =   17.53
       Model |  3332.58872     1  3332.58872           Prob > F      =  0.0013
    Residual |  2280.83985    12  190.069988           R-squared     =  0.5937
-------------+------------------------------           Adj R-squared =  0.5598
       Total |  5613.42857    13  431.802198           Root MSE      =  13.787

------------------------------------------------------------------------------
       seats |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gallup |    1.96812   .4700211     4.19   0.001     .9440315    2.992208
       _cons |  -127.4281   25.54753    -4.99   0.000    -183.0914   -71.76486
------------------------------------------------------------------------------
```

```
scatter loss gallup, mlabel(year) || lfit loss gallup || lfit loss gallup if year >1946
```



Comparing regression coefficients

- As a general rule:
  - Code all your variables to vary between 0 and 1
    - That is, minimum = 0, maximum = 1
  - Regression coefficients then represent the effect of shifting from the minimum to the maximum.
  - This allows you to more easily compare the relative importance of coefficients.

How to recode variables to 0-1 scale

- Party ID example: pid7
- Usually varies from
    - 1 (strongR epublican)
    - to 8 (strong Democrat)
    - sometimes 0 needs to be recoded to missing (“.”).
- Stata code?

```
replace pid7 = (pid7-1)/7
```

Regression interpretation with 0-1 scale

- Continue with pid7 example
  - □ `regress natlecon pid7` (both recoded to 0-1 scales)*
  - □ pid7 coefficient: b = -.46 (CCES data from 2006)
  - □ Interpretation?
    - Shifting from being a strong Republican to a strong Democrat corresponds with a .46 drop in evaluations of the national economy (on the one-point national economy scale)

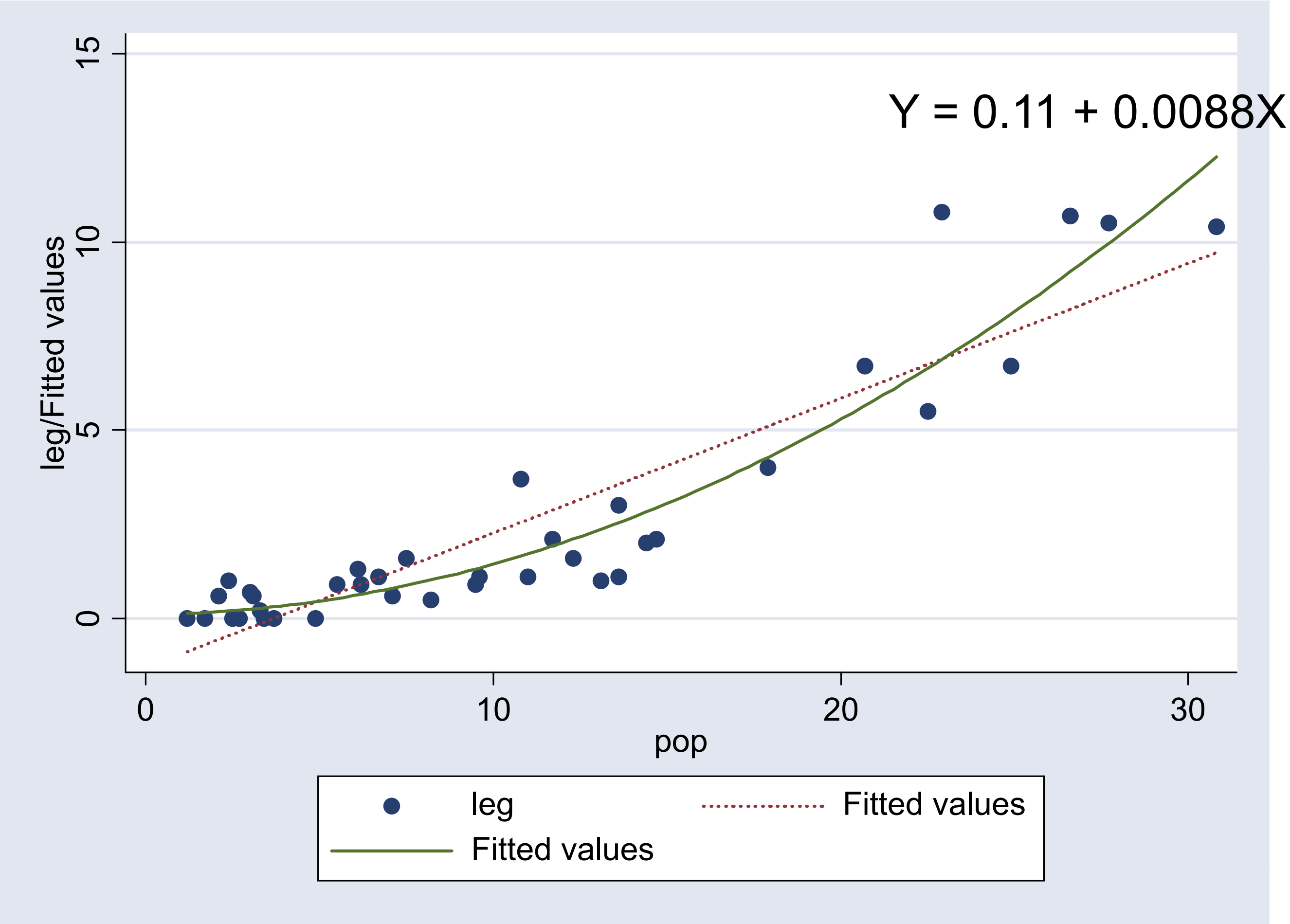
*natlecon originally coded so that 1 = excellent, 4 = poor, 5 = not sure

Functional Form

About the Functional Form

- Linear in the variables *vs.* linear in the parameters
  - $Y = a + bX + e$ (linear in both)
  - $Y = a + bX + cX^2 + e$ (linear in parms.)
  - $Y = a + X^b + e$ (linear in variables, not parms.)
- Regression must be linear in parameters

The Linear and Curvilinear Relationship between African American Population & Black Legislators



$$Y = 0.11 + 0.0088X + 0.013X^2$$

```
scatter beo pop || qfit beo pop
```

Log transformations (see Tufte, ch. 3)

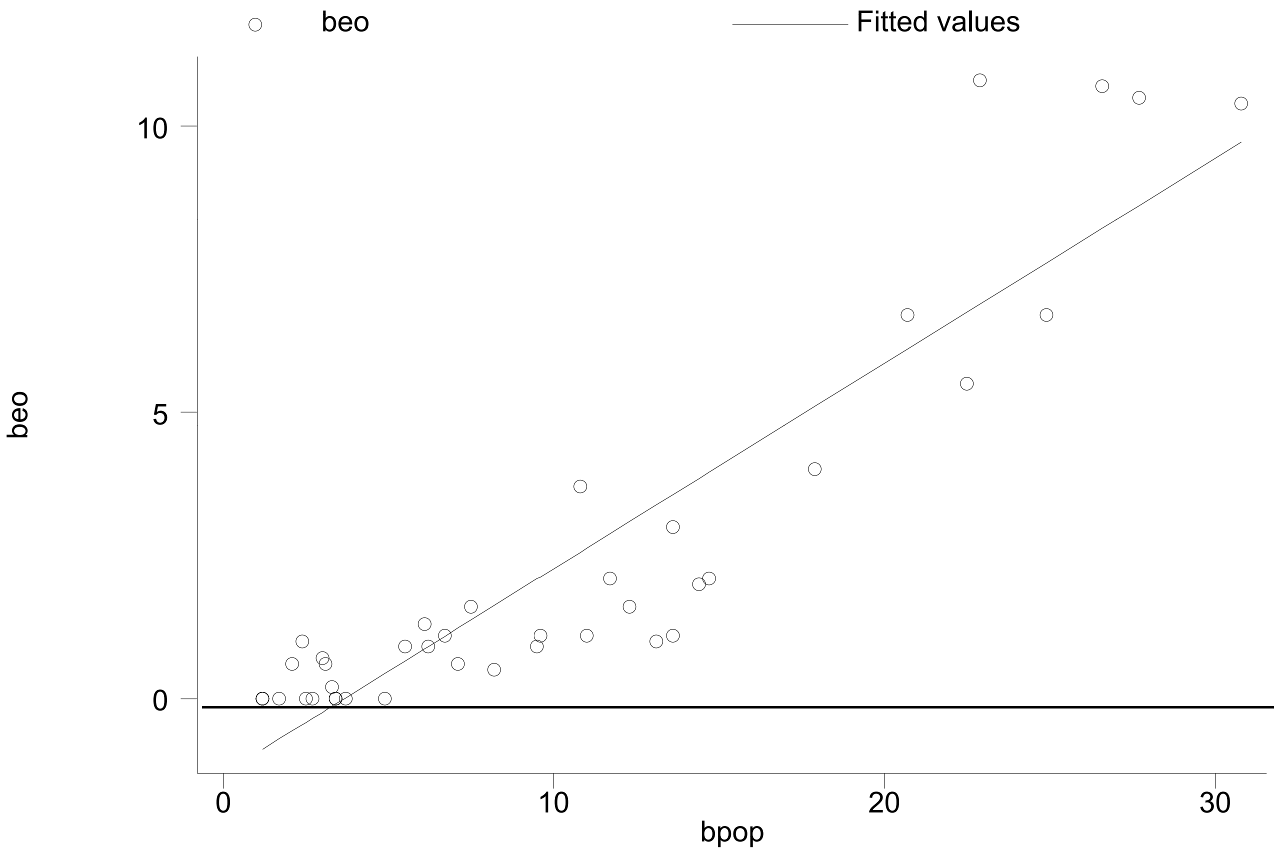
$Y = a + bX + e$	$b = dY/dX$, or b = the unit change in Y given a unit change in X	Typical case
$Y = a + b \ln X + e$	$b = dY/(dX/X)$, or b = the unit change in Y given a % change in X	Log explanatory variable
$\ln Y = a + bX + e$	$b = (dY/Y)/dX$, or b = the % change in Y given a unit change in X	Log dependent variable
$\ln Y = a + b \ln X + e$	$b = (dY/Y)/(dX/X)$, or b = the % change in Y given a % change in X (elasticity)	Economic production

Goodness of regression fit

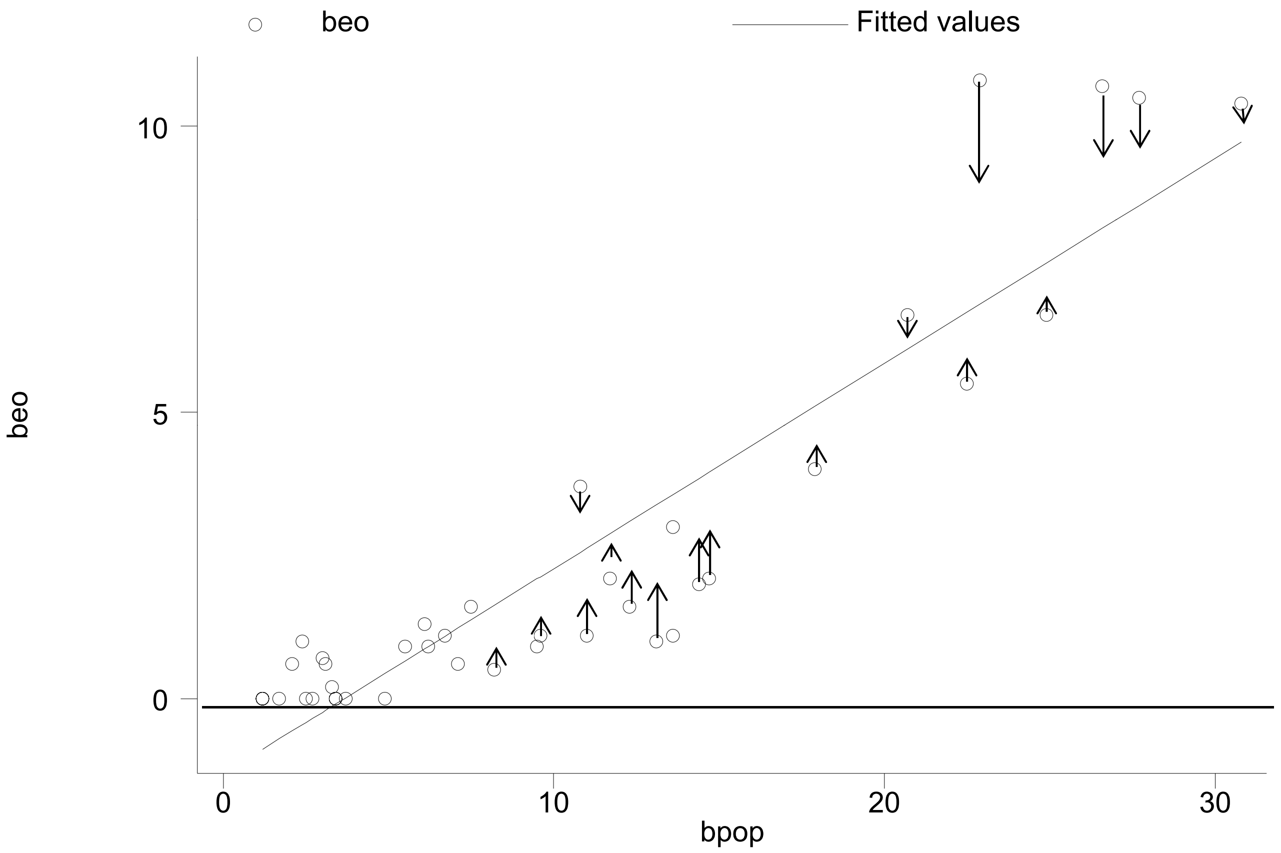
How “good” is the fitted line?

- Goodness-of-fit is often not relevant to research
- Goodness-of-fit receives too much emphasis
- Focus on
  - □ Substantive interpretation of coefficients (most important)
  - □ Statistical significance of coefficients (less important)
    - Confidence interval
    - Standard error of a coefficient
    - *t*-statistic: *coeff./s.e.*
- Nevertheless, you should know about
  - □ Standard Error of the Regression (SER)
    - Standard Error of the Estimate (SEE)
    - Also called Regrettably called Root Mean Squared Error (Root MSE) in Stata
  - □ R-squared (R^2)
    - Often not informative, use sparingly

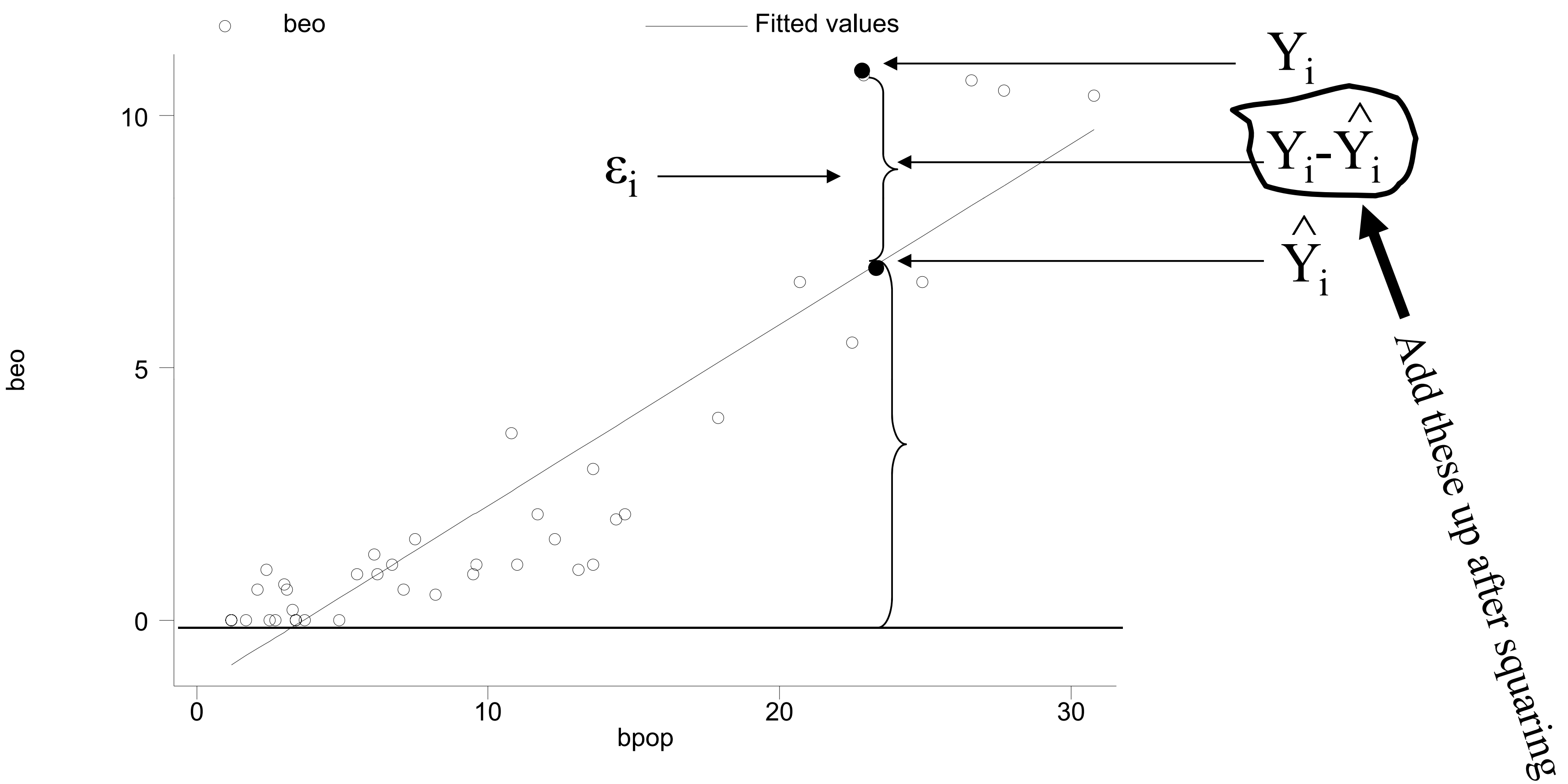
Standard Error of the Regression the idea



Standard Error of the Regression the idea



Standard Error of the Regression picture



Add these up after squaring

Standard Error of the Regression (SER)

- or Standard Error of the Estimate
- or Root Mean Squared Error (Root MSE)

$$\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{d.f.}}$$

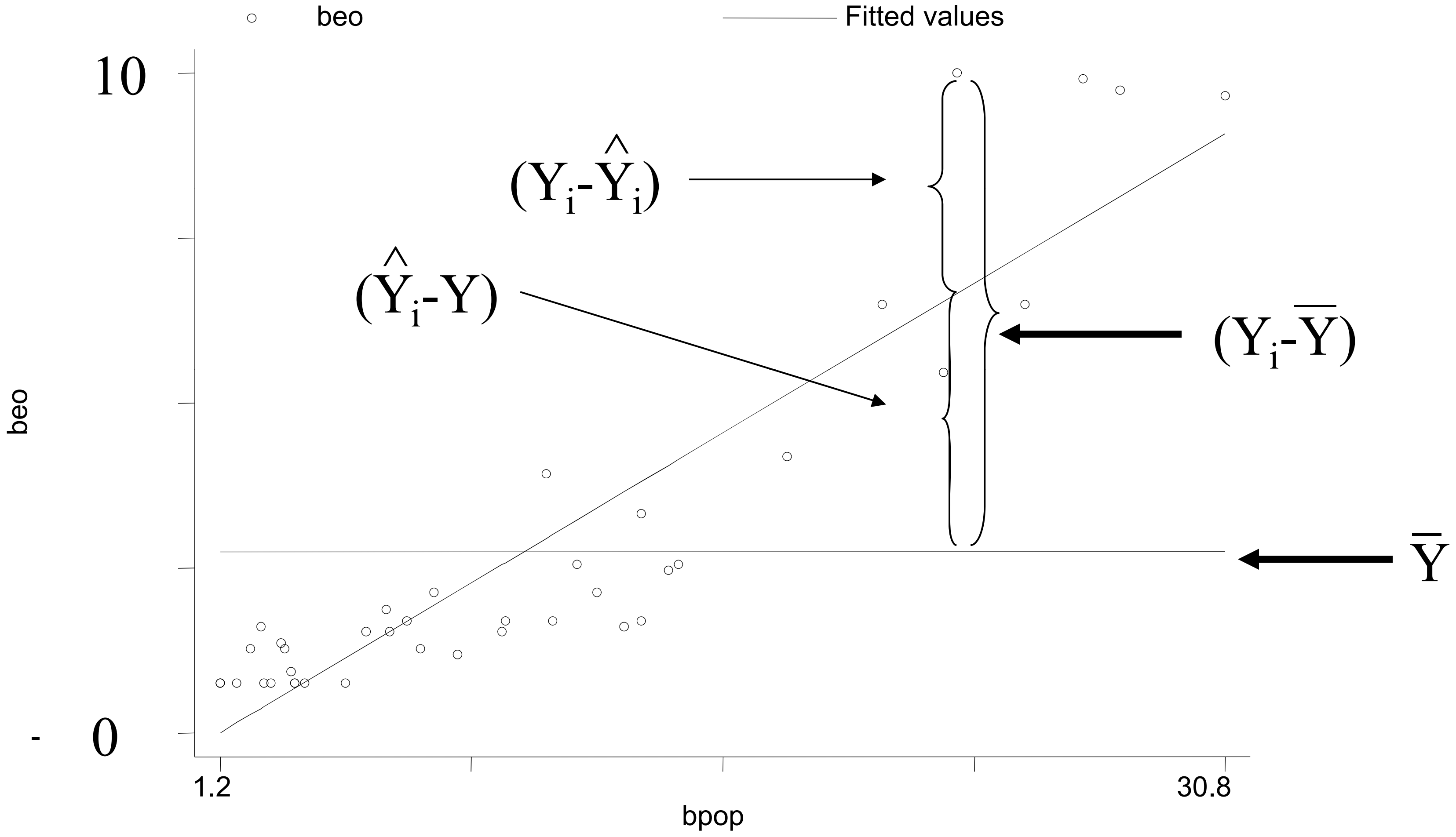
d.f. equals n minus the number of estimate coefficients (*B*s).
In bivariate regression case, *d.f.* = n-2.

SER interpretation called “Root MSE” in Stata

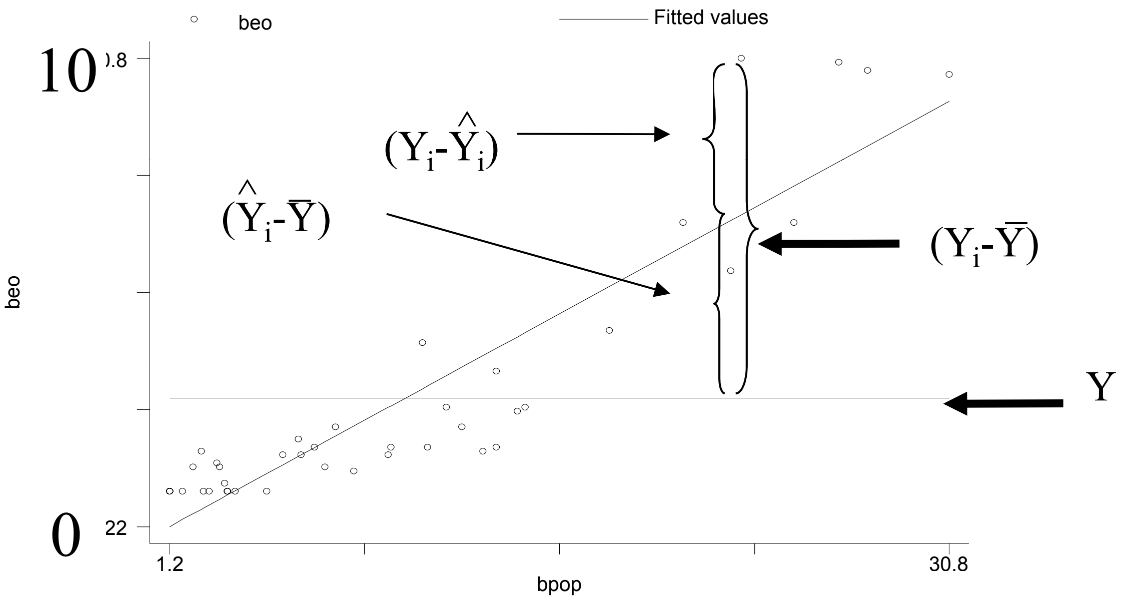
- On average, in-sample predictions will be off the mark by about one standard error of the regression

```
. reg beo bpop

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =  202.56
       Model |   351.26542     1   351.26542           Prob > F      =  0.0000
    Residual |  67.6326195    39  1.73416973           R-squared     =  0.8385
-------------+------------------------------           Adj R-squared =  0.8344
       Total |  418.898039    40   10.472451           Root MSE      =  1.3169

------------------------------------------------------------------------------
         beo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bpop |   .3586751   .0251876    14.23   0.000     .3075284    .4094219
       _cons |  -1.314892   .3277508    -4.01   0.000    -1.977831   -.6519535
------------------------------------------------------------------------------
```

R^2: A less useful measure of fit



R^2: A less useful measure of fit



$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \text{"}\textit{total sum of squares}\text{"}$$

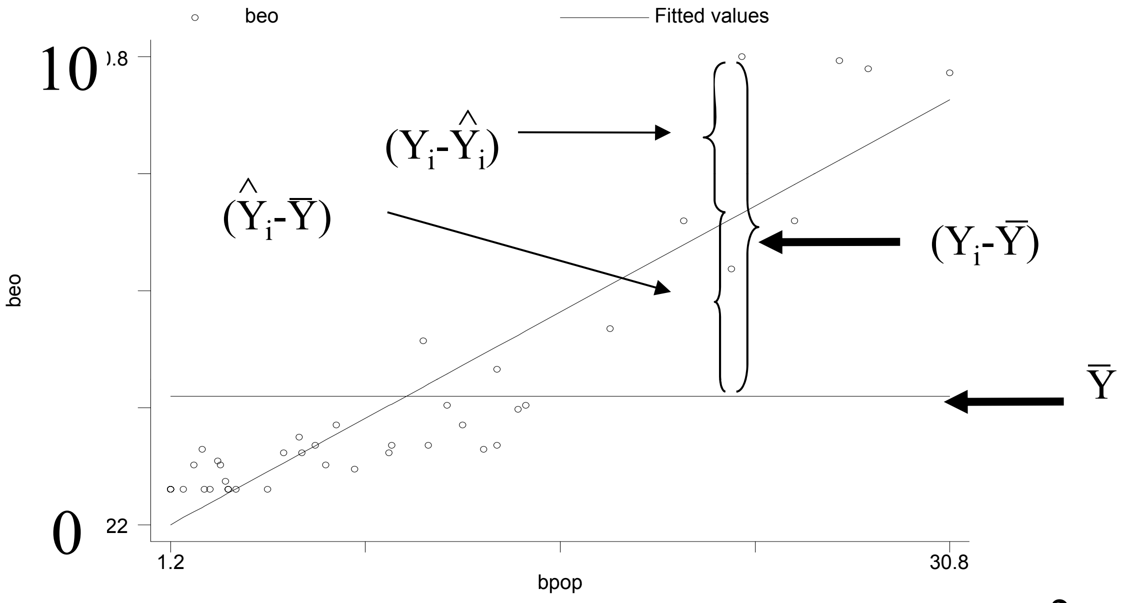
=

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \text{"}\textit{regression sum of squares}\text{"}$$

+

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \text{"}\textit{residual sum of squares}\text{"}$$

R-squared



$$r^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \quad \text{or}$$

pct. variance "explained"

Also called “coefficient of determination”

Interpreting SER (Root MSE) and R^2

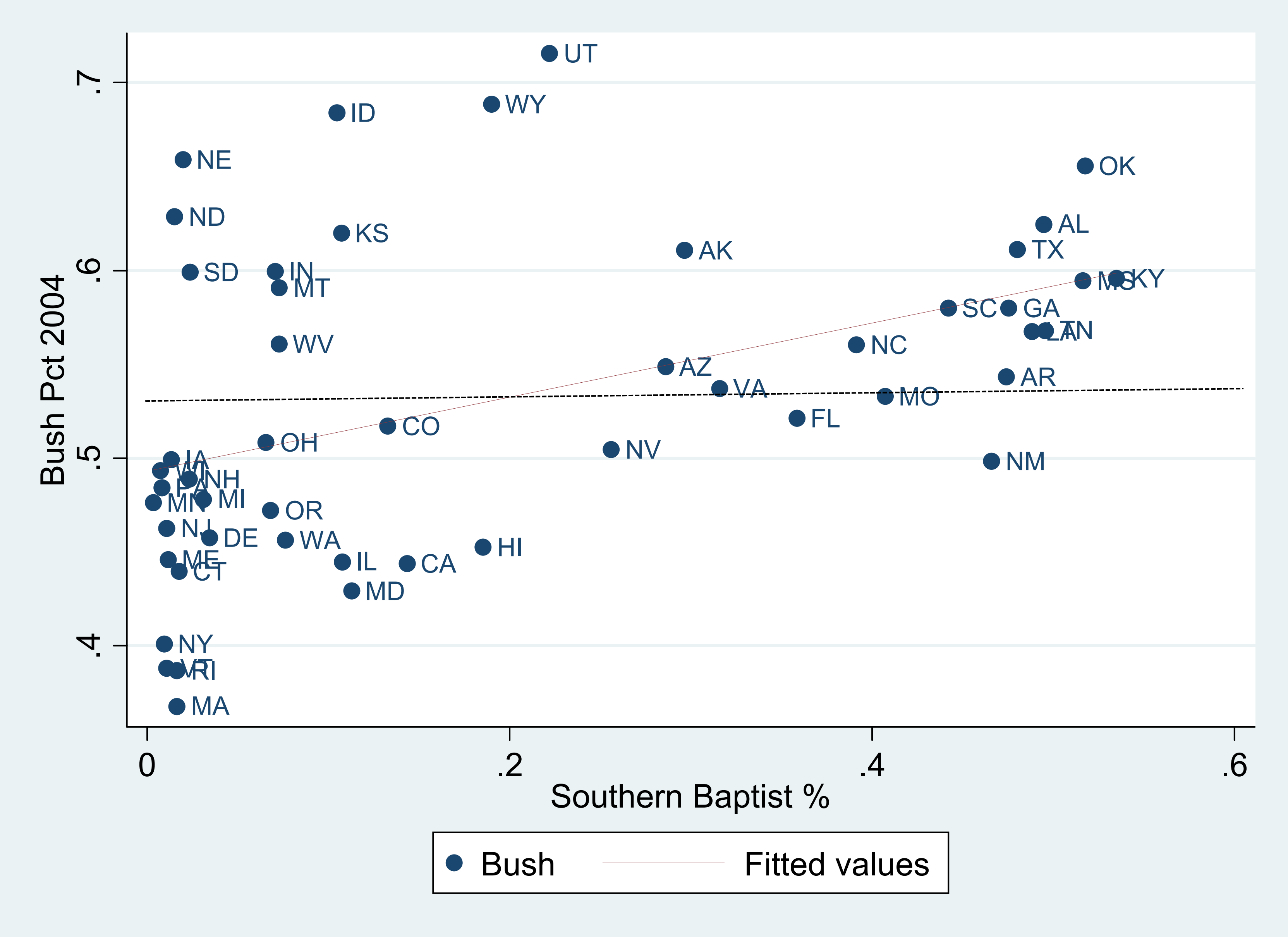
```
. reg bush sbc_mpct

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   11.83
       Model |  .069183833     1  .069183833           Prob > F      =  0.0012
    Residual |  .280630922    48  .005846478           R-squared     =  0.1978
-------------+------------------------------           Adj R-squared =  0.1811
       Total |  .349814756    49  .007139077           Root MSE      =  .07646

------------------------------------------------------------------------------
        bush |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    sbc_mpct |    .196814   .0572138     3.44   0.001     .0817779    .3118501
       _cons |   .4931758   .0155007    31.82   0.000     .4620095     .524342
------------------------------------------------------------------------------
```

Interpreting SER (Root MSE):

- On average, in-sample predictions about Bush’s vote share will be off the mark by about 7.6%

Interpreting R^2

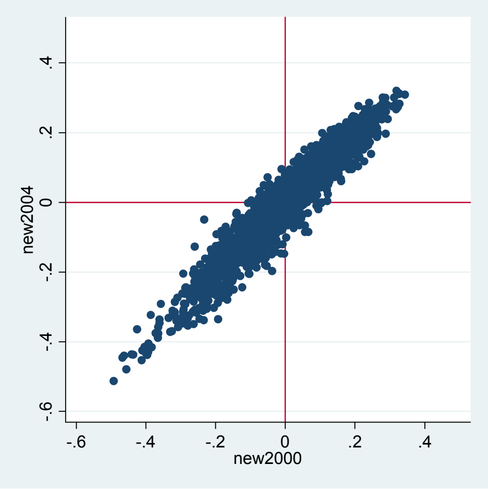
- Regression model explains about 19.8% of the variation in Bush vote.


Bush Pct 2004
.4
.5
.6
.7
0
.2
.4
.6
Southern Baptist %
UT
WY
ID
NE
ND
KS
SD
IN
MT
WV
AK
OK
AL
TX
MS
KY
SC
GA
LA
TN
NC
AZ
VA
MO
AR
FL
CO
OH
IA
WI
PA
NH
MN
MI
OR
NJ
DE
WA
NV
NM
HI
ME
CT
IL
CA
MD
NY
VT
RI
MA
● Bush
Fitted values

Correlation

Correlation

$$Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = r$$



$$\text{Corr}(\text{BushPct}_{00},\text{BushPct}_{04}) = 0.96 =$$

$$\frac{0.014858}{\sqrt{0.01499} \times \sqrt{0.01605}} \approx .96$$

- Measures how closely data points fall along the line
- Varies between -1 and 1

(compare with Tufte p. 102)

Warning: Don’t correlate often!

- Correlation only measures linear relationship
- Correlation is sensitive to variance
- Correlation usually doesn’t measure a theoretically interesting quantity
- Same criticisms apply to R^2, which is the squared correlation between predictions and data points.
- Instead, focus on regression coefficients (slopes)

Discrete DV, discrete EV

- Crosstabs
- χ^2
- Gamma, Beta, etc.

Example

- What is the relationship between abortion sentiments and vote choice?
- The abortion scale:

1. BY LAW, ABORTION SHOULD NEVER BE PERMITTED.
2. THE LAW SHOULD PERMIT ABORTION ONLY IN CASE OF RAPE, INCEST, OR WHEN THE WOMAN'S LIFE IS IN DANGER.
3. THE LAW SHOULD PERMIT ABORTION FOR REASONS OTHER THAN RAPE, INCEST, OR DANGER TO THE WOMAN'S LIFE, BUT ONLY AFTER THE NEED FOR THE ABORTION HAS BEEN CLEARLY ESTABLISHED.
4. BY LAW, A WOMAN SHOULD ALWAYS BE ABLE TO OBTAIN AN ABORTION AS A MATTER OF PERSONAL CHOICE.

Abortion and vote choice in 2006

```
. tab housevote abortopinion, col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

   us house candidate |          stmt most agrees w/ view on abortion law
           voting for |     Never       Rarely   Sometimes     Always  other (pl |     Total
----------------------+-------------------------------------------------------+----------
             Democrat |       446        1,749       1,903      8,759        770 |    13,627
                      |     13.60        20.21       36.90      57.93      34.30 |     39.55
----------------------+-------------------------------------------------------+----------
           Republican |     1,900        4,381       1,639      2,006        758 |    10,684
                      |     57.93        50.62       31.78      13.27      33.76 |     31.01
----------------------+-------------------------------------------------------+----------
other (please specify |       157          384         228        671        190 |     1,630
                      |      4.79         4.44        4.42       4.44       8.46 |      4.73
----------------------+-------------------------------------------------------+----------
i won't vote in this  |        65          201         117        299         52 |       734
                      |      1.98         2.32        2.27       1.98       2.32 |      2.13
----------------------+-------------------------------------------------------+----------
       haven't decided |       712        1,939       1,270      3,386        475 |     7,782
                      |     21.71        22.41       24.63      22.39      21.16 |     22.58
----------------------+-------------------------------------------------------+----------
                Total |     3,280        8,654       5,157     15,121      2,245 |    34,457
                      |    100.00       100.00      100.00     100.00     100.00 |    100.00
```

Use the appropriate graph/table

- Continuous DV, continuous EV
  - E.g., vote share by income growth
  - Use scatter plot
- Continuous DV, discrete and unordered EV
  - E.g., vote share by religion or by union membership
  - Box plot, dot plot
- Discrete DV, discrete EV
  - No graph: Use crosstabs (`tabulate`)

Two quick notes about comparing coefficients

- Recode/rescale independent variables to be in 0-1 interval
  - □ new_x = [x-min(x)+1]/(max(x)-min(x)+1)
  - □ Interpretation: a move from the minimum to the maximum in the independent variable yields an average change of *b* in the d.v.

```
. reg beo bpop

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =  202.56
       Model |  351.26542      1  351.26542            Prob > F      =  0.0000
    Residual | 67.6326195     39  1.73416973           R-squared     =  0.8385
-------------+------------------------------           Adj R-squared =  0.8344
       Total | 418.898039     40  10.472451            Root MSE      =  1.3169

------------------------------------------------------------------------------
         beo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bpop |   .3584751   .0251876    14.23   0.000     .3075284    .4094219
       _cons |  -1.314892   .3277508    -4.01   0.000    -1.977831   -.6519535
------------------------------------------------------------------------------

           Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
               bpop |        41    10.13171    8.266633        1.2       30.8

. gen bpop01=(bpop-1.2)/(30.8-1.2)

. reg beo bpop01

             Source |       SS       df       MS              Number of obs =      41
       -------------+------------------------------           F(  1,    39) =  202.56
              Model |  351.265419     1  351.265419           Prob > F      -  0.0000
           Residual |    67.63262    39  1.73416974           R-squared     =  0.8385
       -------------+------------------------------           Adj R-squared =  0.8344
              Total |  418.898039    40  10.472451            Root MSE      =  1.3169

       ------------------------------------------------------------------------------
                beo |     Coef.     Std. Err.      t      P>|t|    [95% Conf. Interval]
       -------------+----------------------------------------------------------------
             bpop01 |  10.61086    .7455536     14.23    0.000      9.10284    12.11889
              _cons | -.8847219    .3048075     -2.90    0.006    -1.501253   -.2681905
       ------------------------------------------------------------------------------
```

- Convert *all* variables, except dummy variables, to "unit deviates":\
  - new_x = [x-mean(x)]/sd(x)
  - new_y = [y-mean(y)]/sd(y)   etc.
- Interpretation: a one standard deviation change in x yields, on average, a *b* standard deviation change in y.
  - (For a dummy variable, a change from category *0* to category *1* yields, on average, a *b* standard deviation change in y.

```
. reg beo bpop

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =  202.56
       Model |   351.26542     1   351.26542           Prob > F      =  0.0000
    Residual |  67.6326195    39  1.73416973           R-squared     =  0.8385
-------------+------------------------------           Adj R-squared =  0.8344
       Total |  418.898039    40  10.472451            Root MSE      =  1.3169

------------------------------------------------------------------------------
         beo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bpop |   .3584751   .0251876    14.23   0.000     .3075284    .4094219
       _cons |  -1.314892   .3277508    -4.01   0.000    -1.977831   -.6519535
------------------------------------------------------------------------------
. summ beo bpop

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         beo |        41    2.317073    3.236117          0       10.8
        bpop |        41    10.13171    8.266633        1.2       30.8

. gen st_beo=(beo-2.317073)/3.236117
(9 missing values generated)

. gen st_bpop=(bpop-10.13171)/8.266633
(9 missing values generated)
```

```
. reg st_beo st_bpop

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =  202.56
       Model |  33.5418469     1  33.5418469           Prob > F      =  0.0000
    Residual |  6.45814509    39  .165593464             R-squared   =  0.8385
-------------+------------------------------           Adj R-squared =  0.8344
       Total |  39.9999919    40  .999999799           Root MSE      =  .40693

------------------------------------------------------------------------------
      st_beo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     st_bpop |   .9157217   .0643416    14.23   0.000     .7855786    1.045865
       _cons |   3.54e-07   .0635521     0.00   1.000    -.1285458    .1285465
------------------------------------------------------------------------------
```

```
. reg beo bpop,beta

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =  202.56
       Model |   351.26542     1   351.26542           Prob > F      =  0.0000
    Residual |  67.6326195    39  1.73416973           R-squared     =  0.8385
-------------+------------------------------           Adj R-squared =  0.8344
       Total |  418.898039    40  10.472451            Root MSE      =  1.3169

------------------------------------------------------------------------------
         beo |     Coef.    Std. Err.     t     P>|t|                     Beta
-------------+----------------------------------------------------------------
        bpop |   .3584751   .0251876    14.23   0.000                 .9157218
       _cons |  -1.314892   .3277508    -4.01   0.000
------------------------------------------------------------------------------
```

17.871 Political Science Laboratory
Spring 2012