Describing Bivariate Relationships

17.871

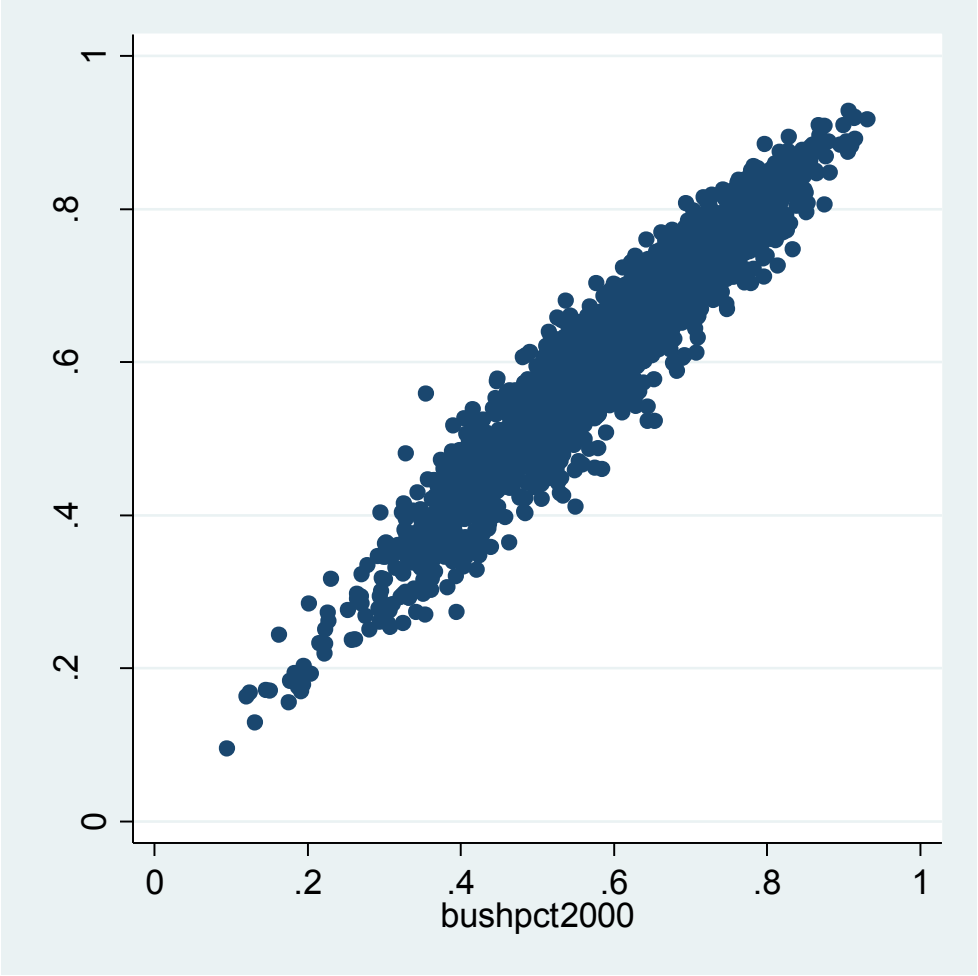
Testing associations

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

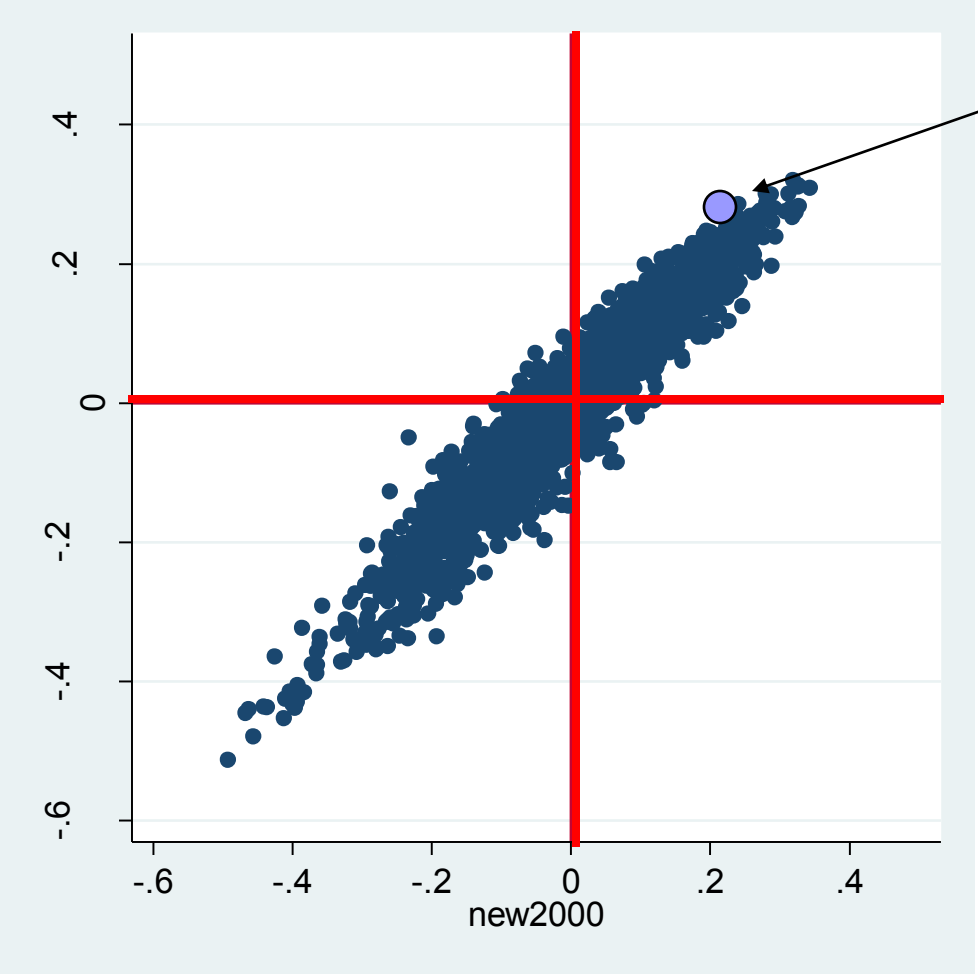
Continuous DV, continuous EV

- Example: What is the relationship between Bush's vote (by county) in 2000 and in 2004?

2004 Prez. Vote *vs.* 2000 Pres. Vote

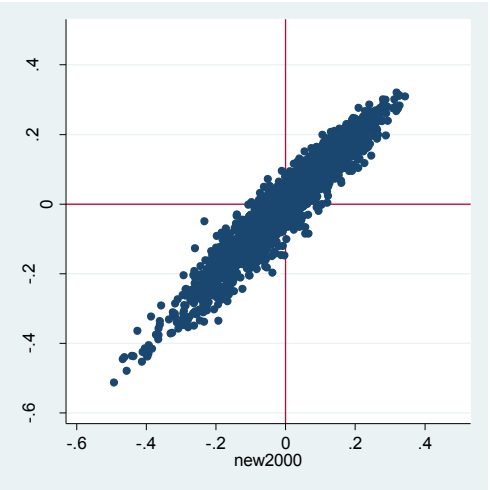


Subtract each observation from its mean



Covariance formula

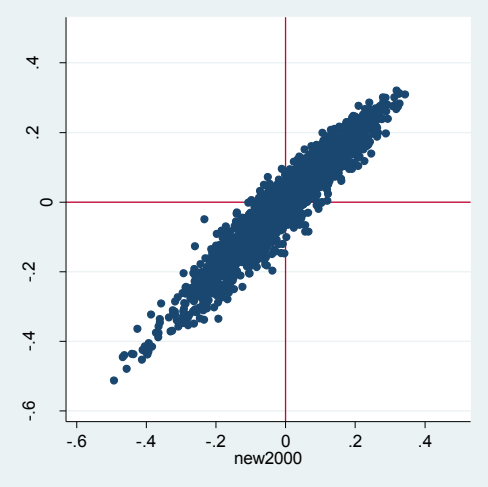
$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$



$\text{Cov}(\text{BushPct}_{00},\text{BushPct}_{04}) = 0.014858$

Correlation formula

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



$$\text{Corr}(\text{BushPct}_{00},\text{BushPct}_{04}) = 0.96 = \frac{0.014858}{\sqrt{0.01499} \times \sqrt{0.01605}} \approx .96$$

(compare with Tufte p. 102)

Warning: Don’t correlate often!

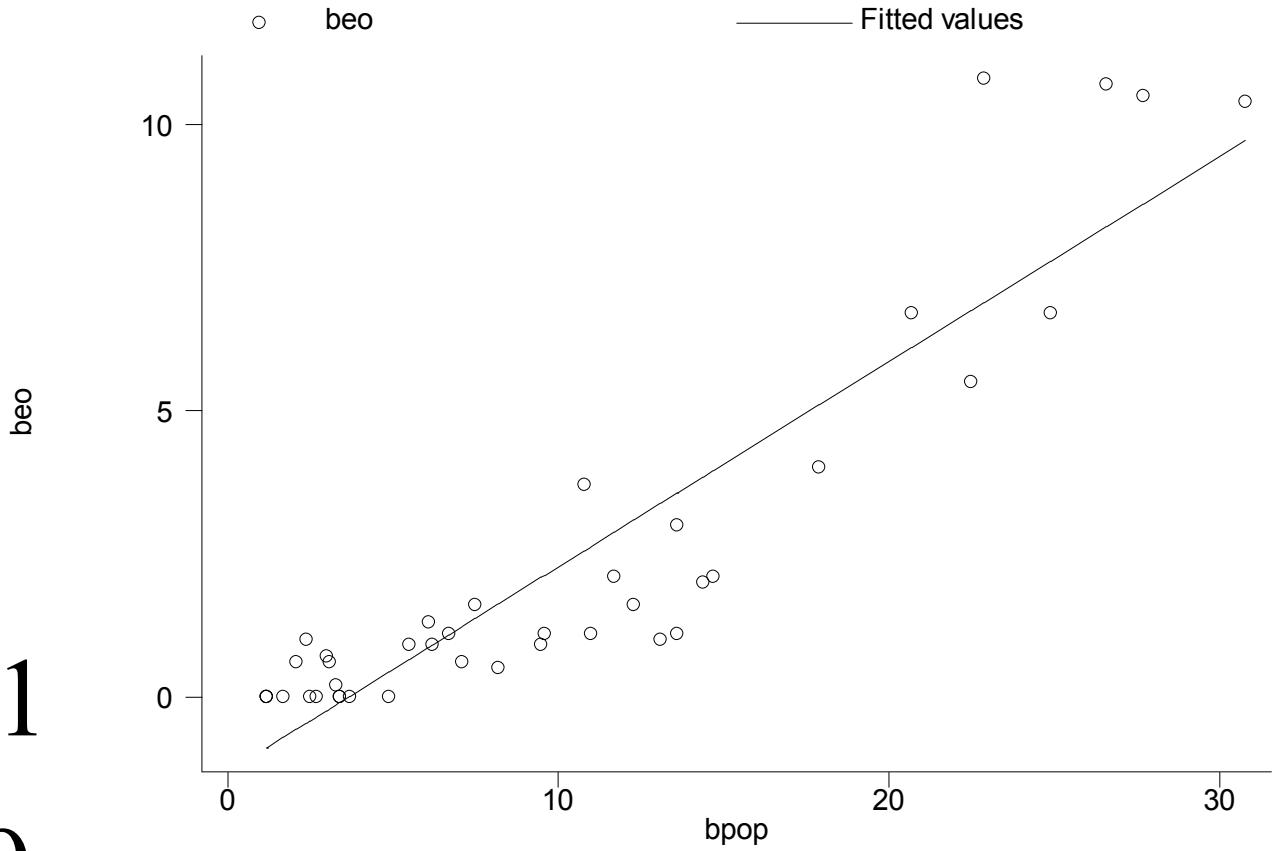
- Correlation only measures linear relationship
- Correlation is sensitive to variance
- Correlation usually doesn’t measure a theoretically interesting quantity

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

The Linear Relationship between Two Variables

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

The Linear Relationship between African American Population & Black Legislators



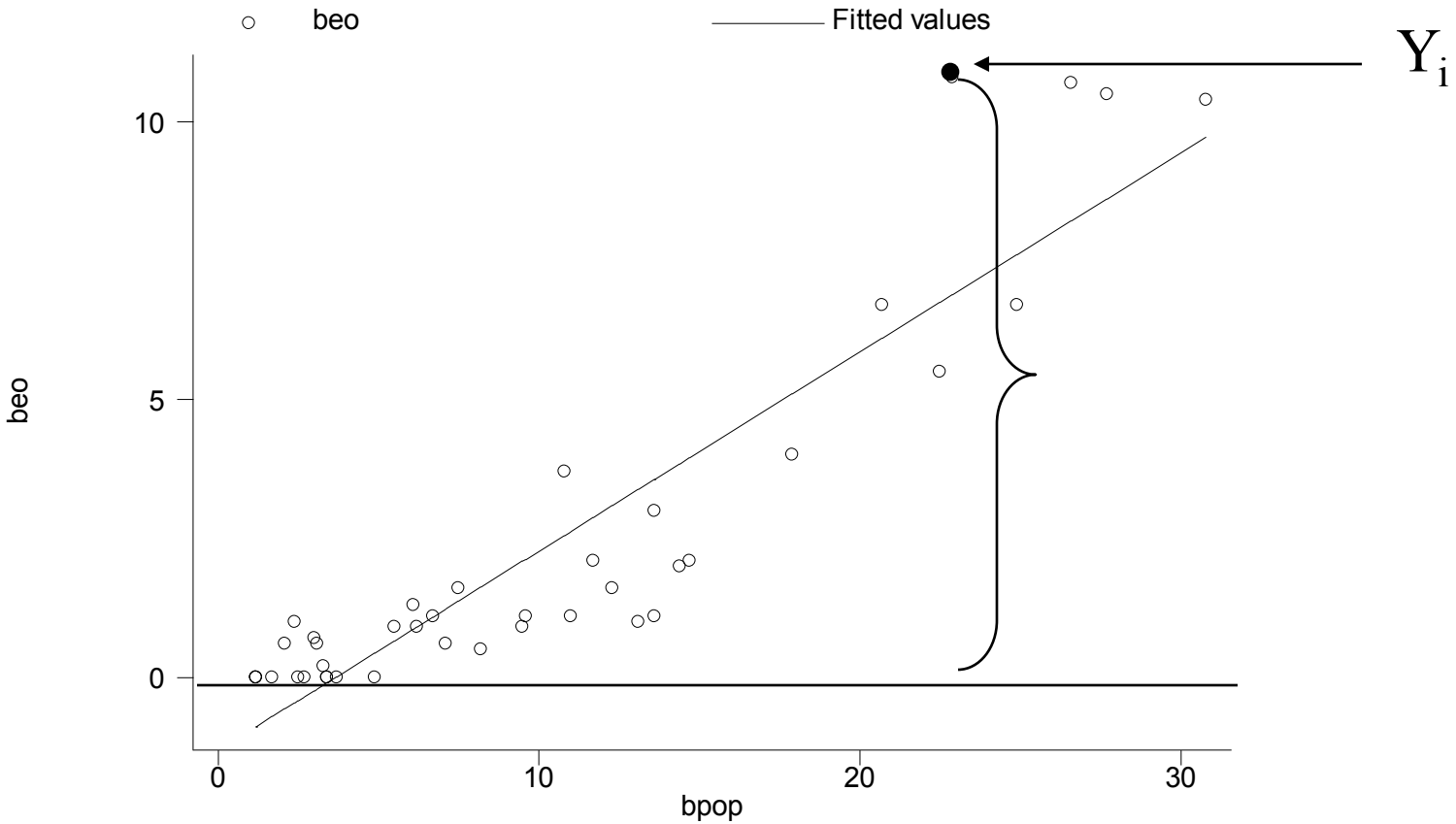
$$\hat{\beta}_0 = -1.31$$

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

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

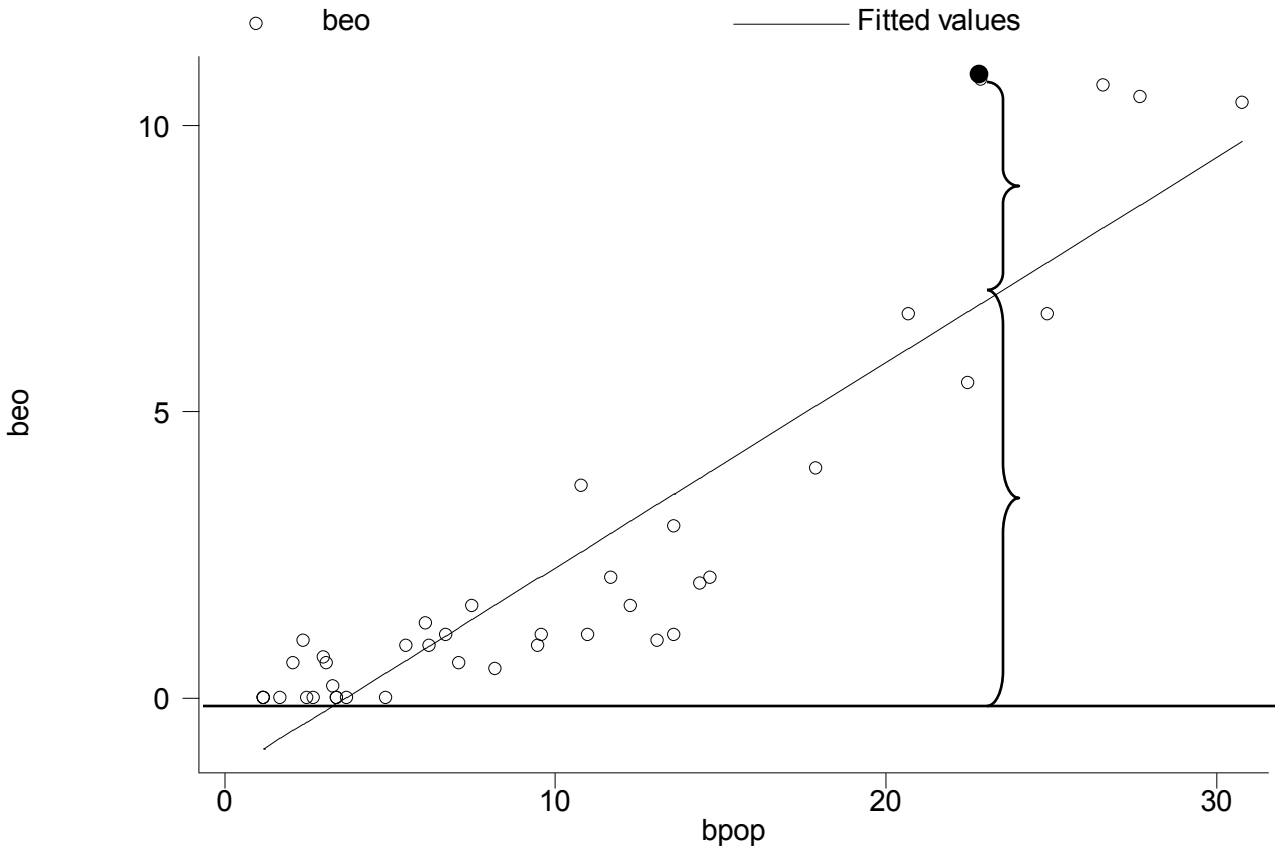
How did we get that line?

1. Pick a value of Y_i



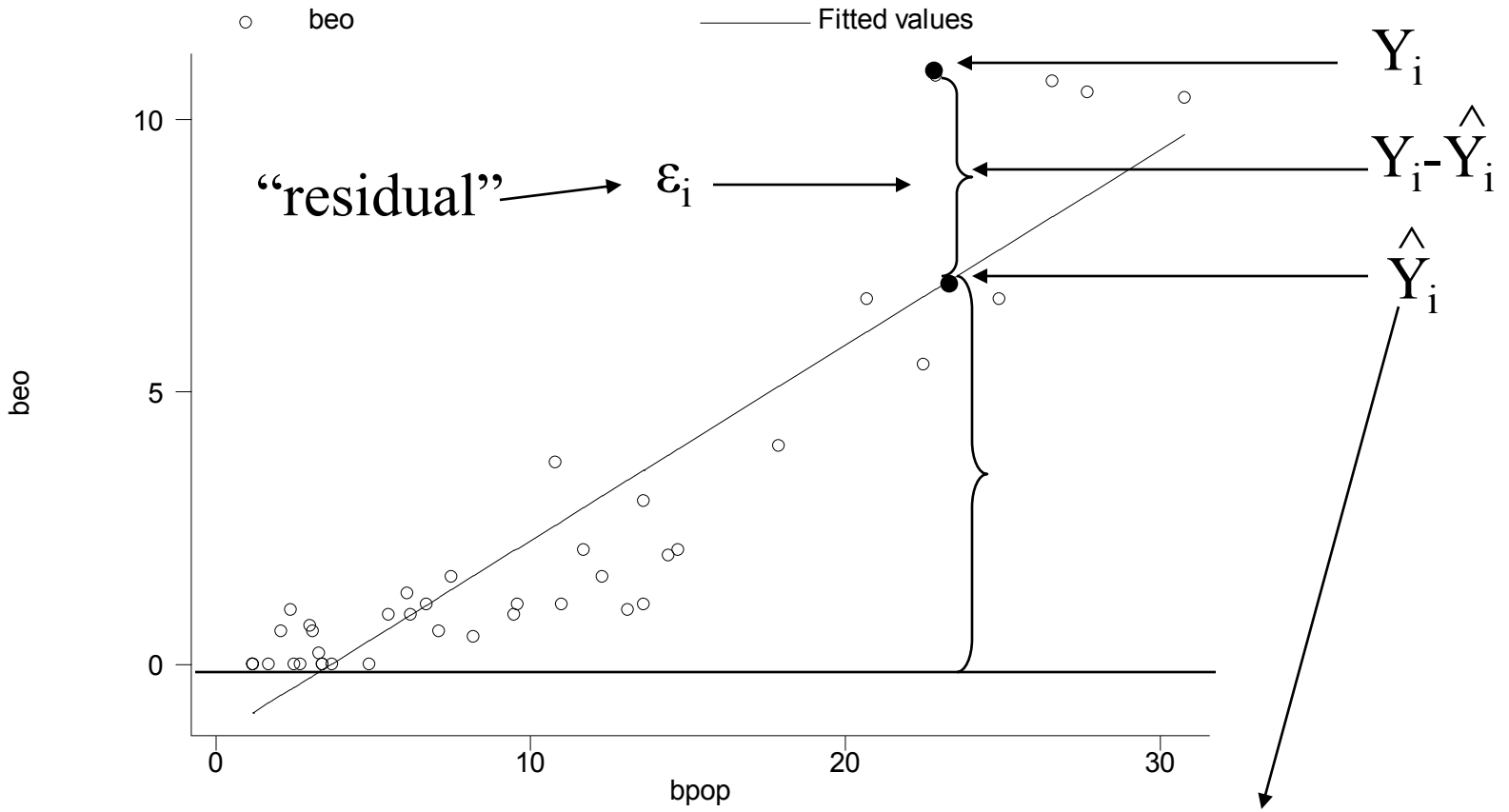
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

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



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

What is ε_i?

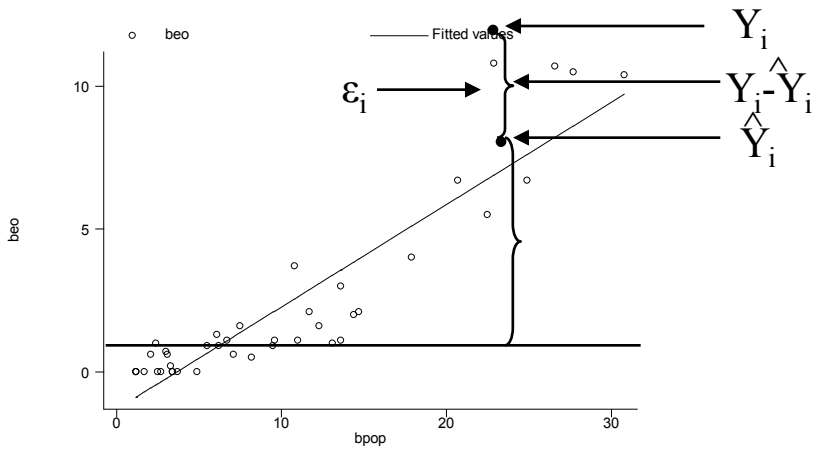
- Vagueness of theory
- Poor proxies (i.e., measurement error)
- Wrong functional form

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

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

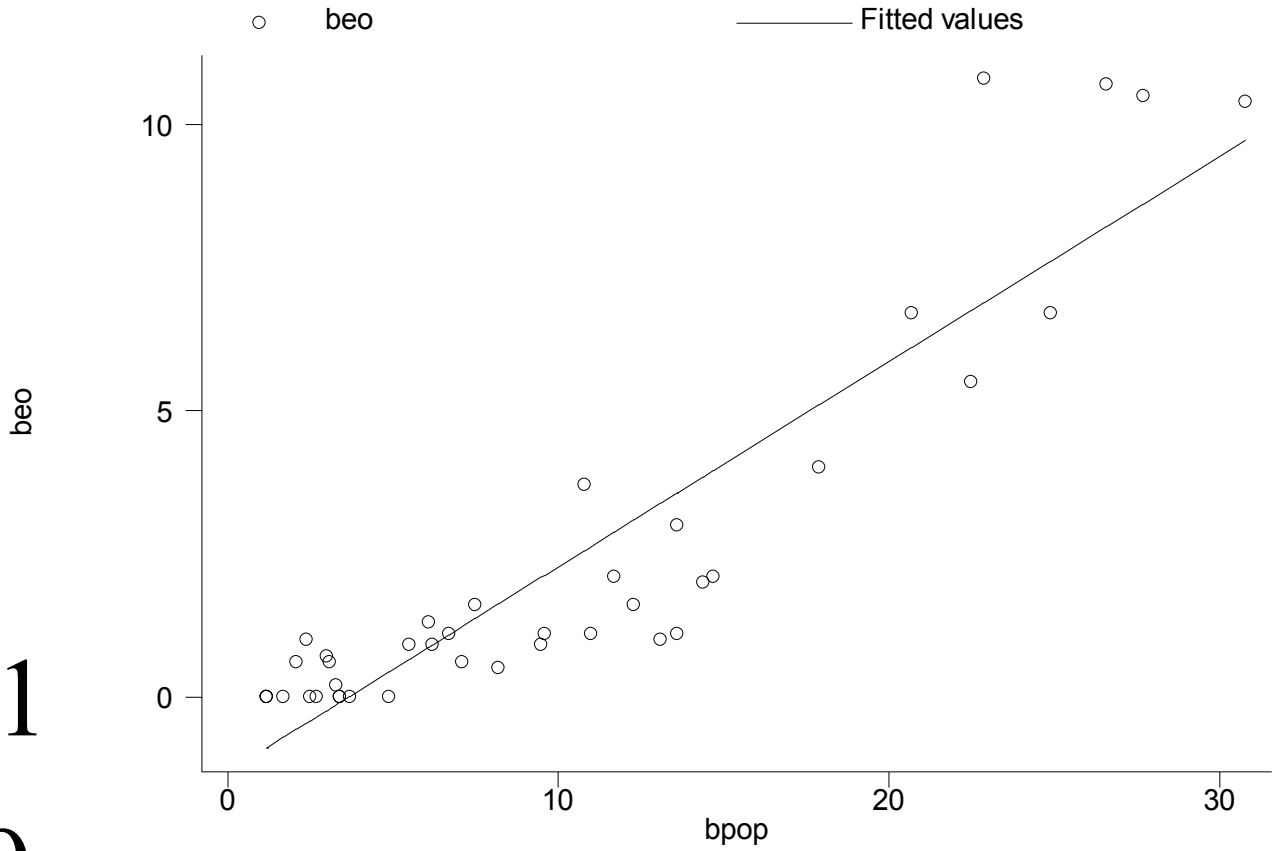
(Tufte, p. 68)

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

Regression Commands in STATA

- reg *depvar expvars*
- predict *newvar*
- predict *newvar,* resid
  - *newvar* will now equal ε_i

The Linear Relationship between African American Population & Black Legislators



$$\beta_0 = -1.31$$

$$\beta_1 = 0.359$$

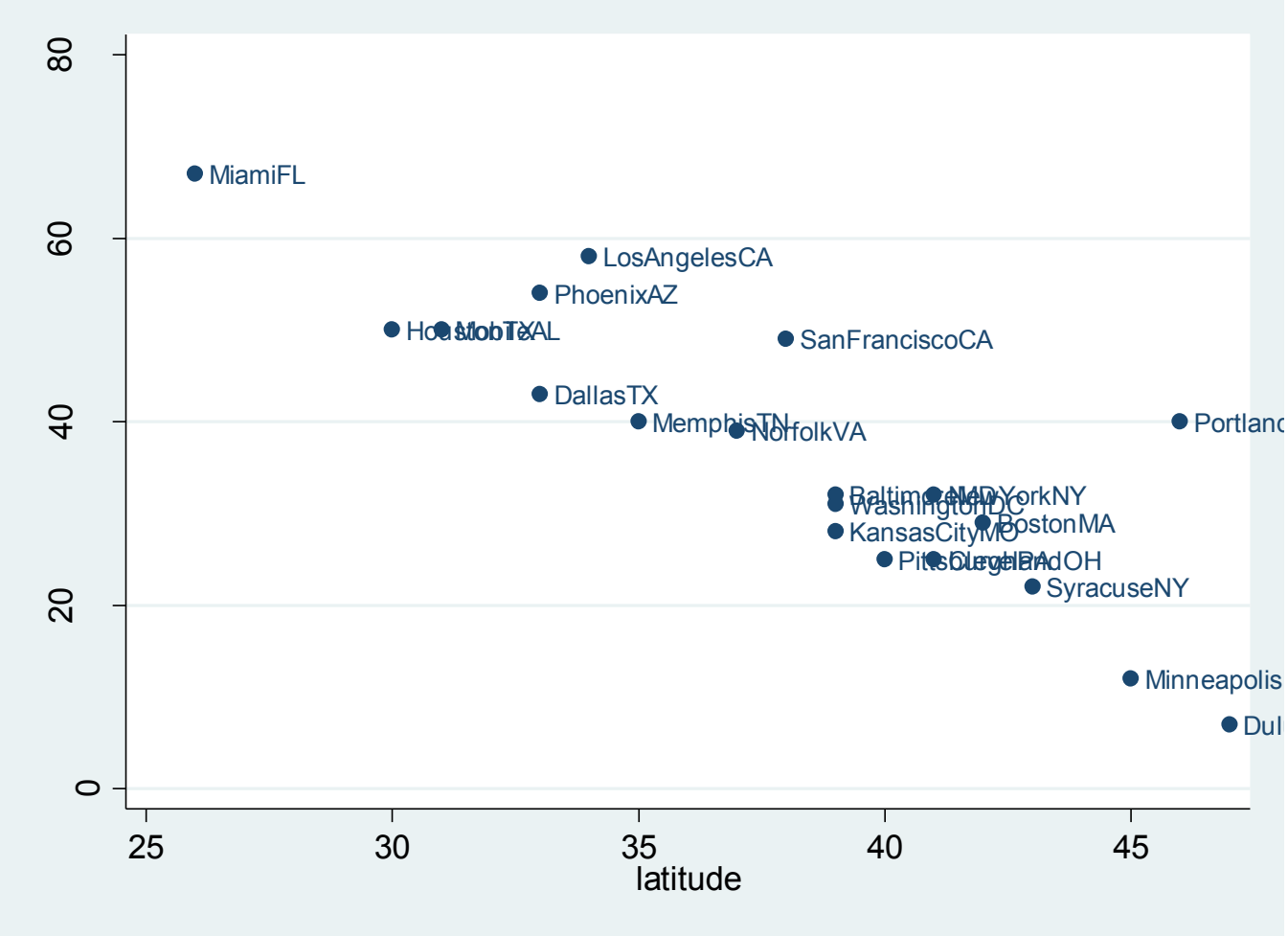
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

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 |  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
------------------------------------------------------------------------------
```

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
------------------------------------------------------------------------------

. predict py
(option xb assumed; fitted values)

. predict ry, resid
```

```
gsort -ry

. list city jantemp py ry
```

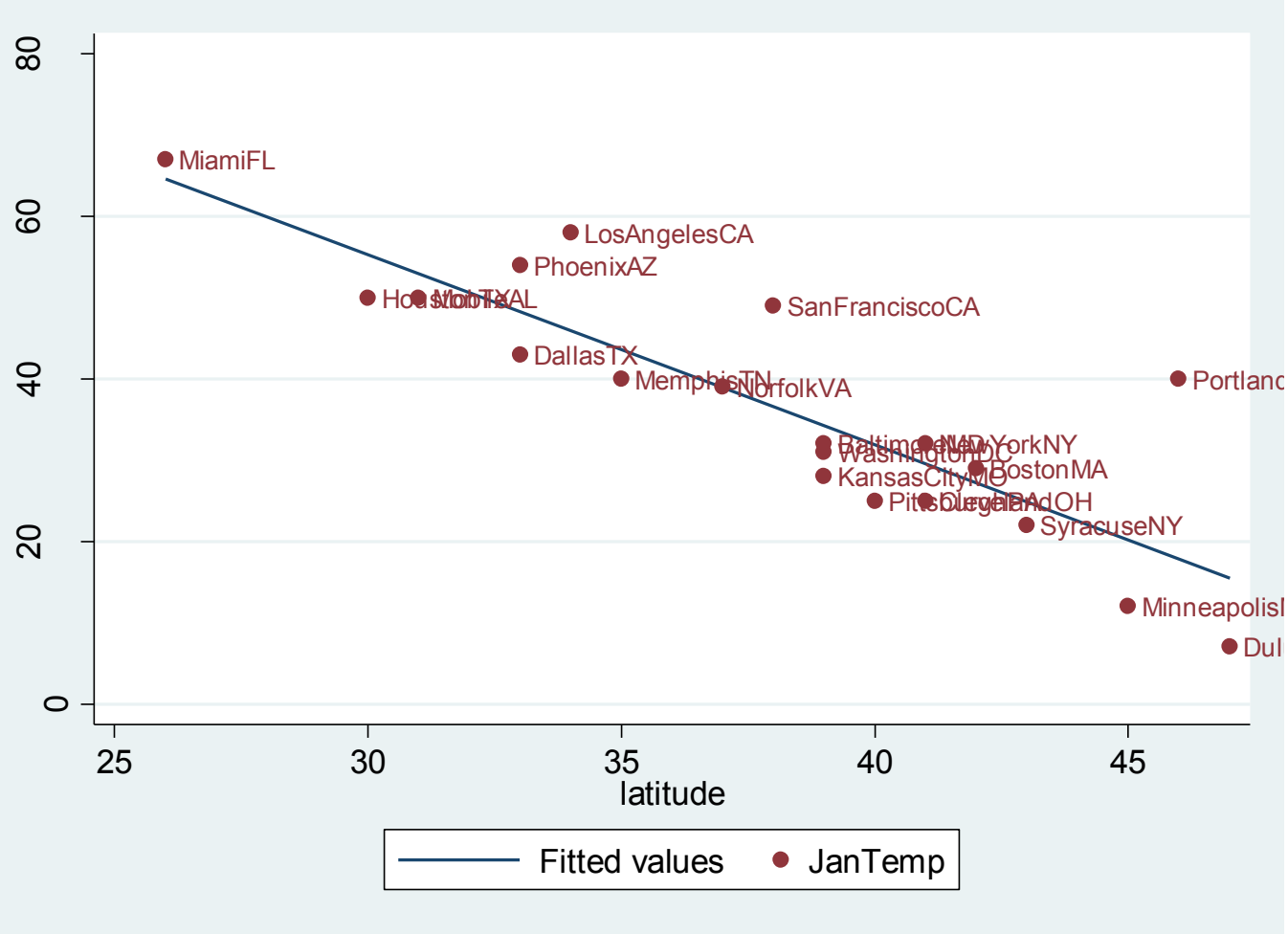
	city	jantemp	py	ry
1.	PortlandOR	40	17.8015	22.1985
2.	SanFranciscoCA	49	36.53293	12.46707
3.	LosAngelesCA	58	45.89864	12.10136
4.	PhoenixAZ	54	48.24007	5.759929
5.	NewYorkNY	32	29.50864	2.491357
6.	MiamiFL	67	64.63007	2.36993
7.	BostonMA	29	27.16722	1.832785
8.	NorfolkVA	39	38.87436	.125643
9.	BaltimoreMD	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

Residuals

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

One important numerical property of residuals

- The sum of the residuals is zero.

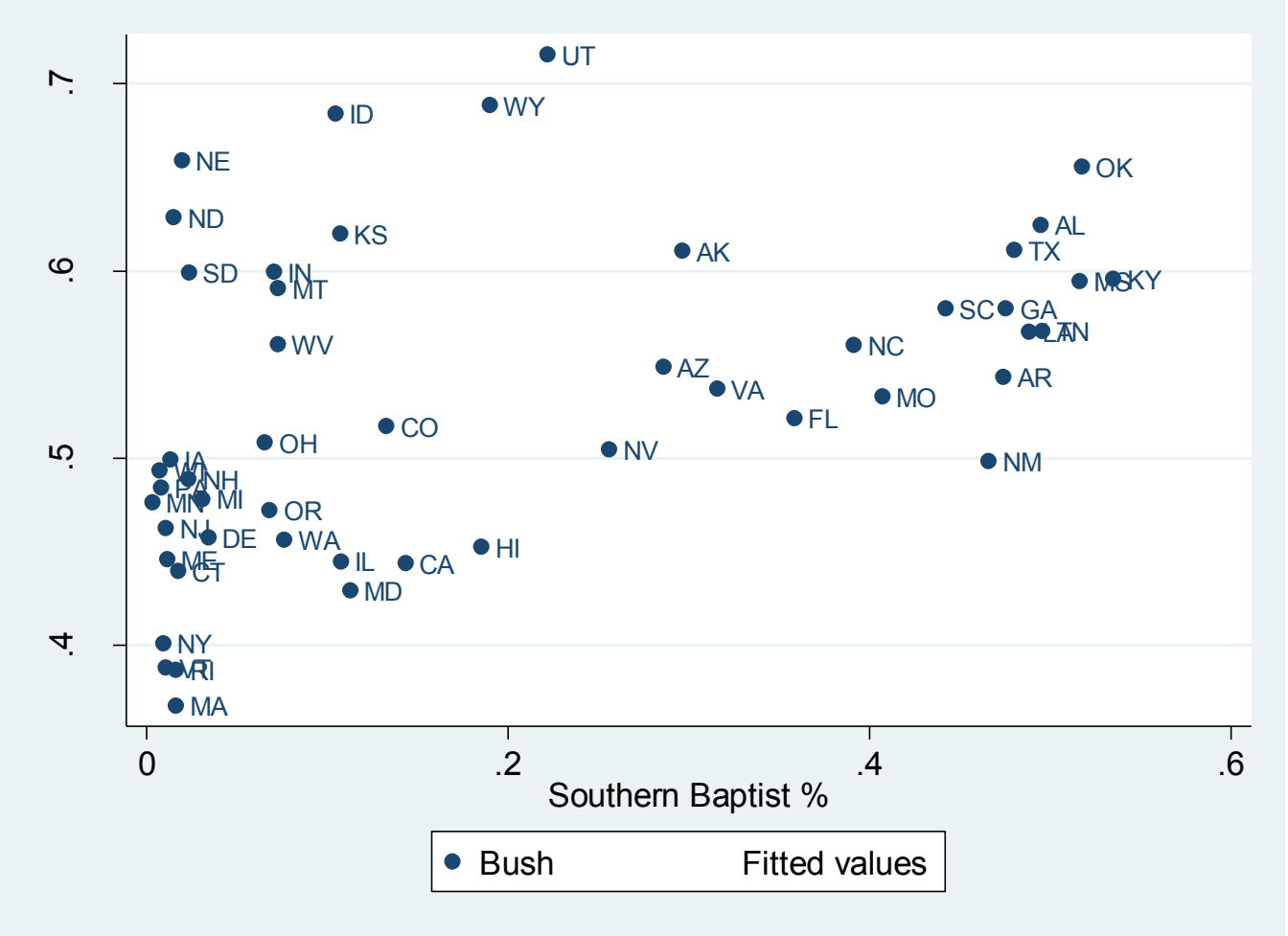


scatter JanTemp latitude, mlabel(city) || lfit JanTemp latitude

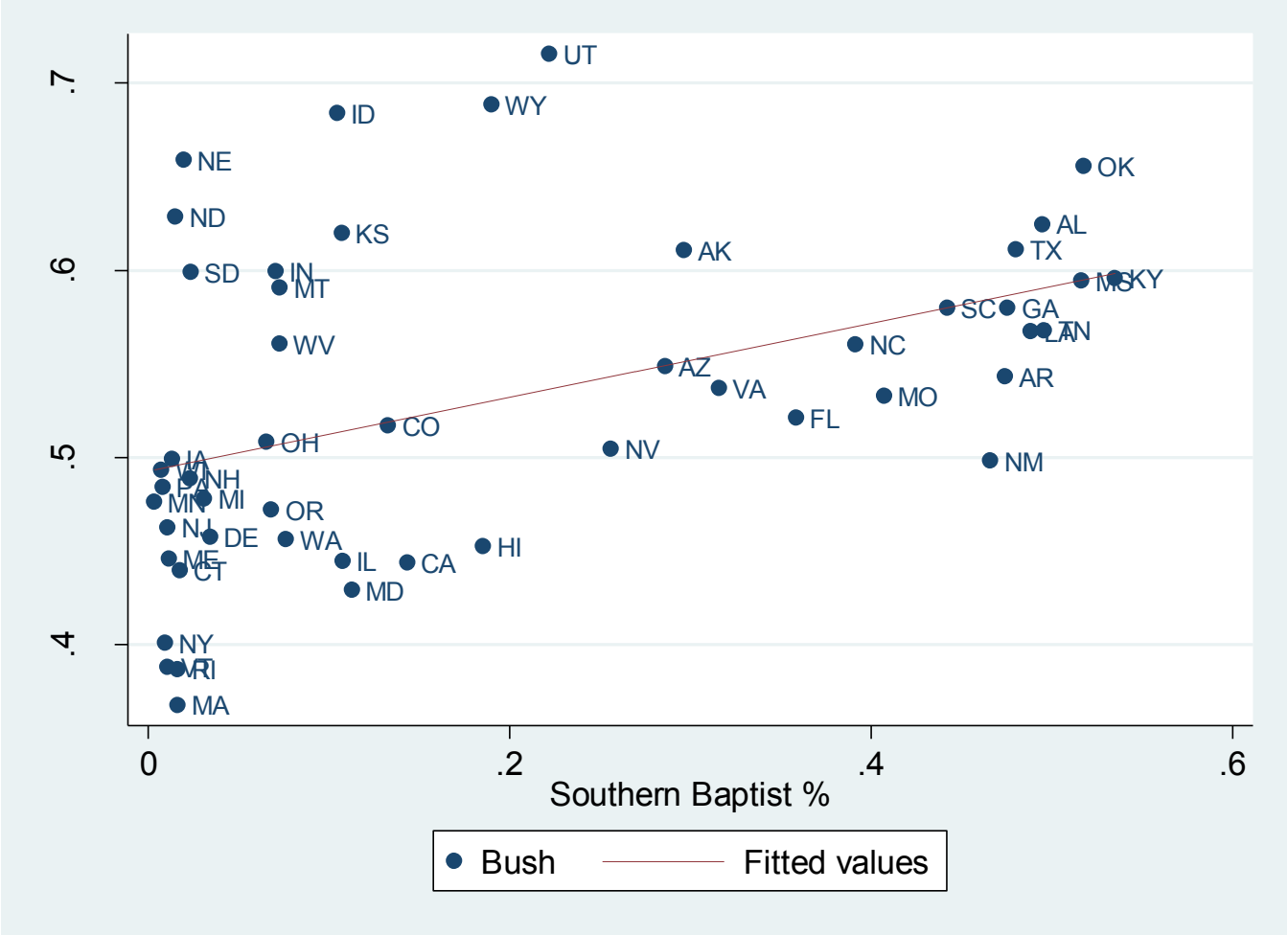
or often better

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

Bush Vote and Southern Baptists

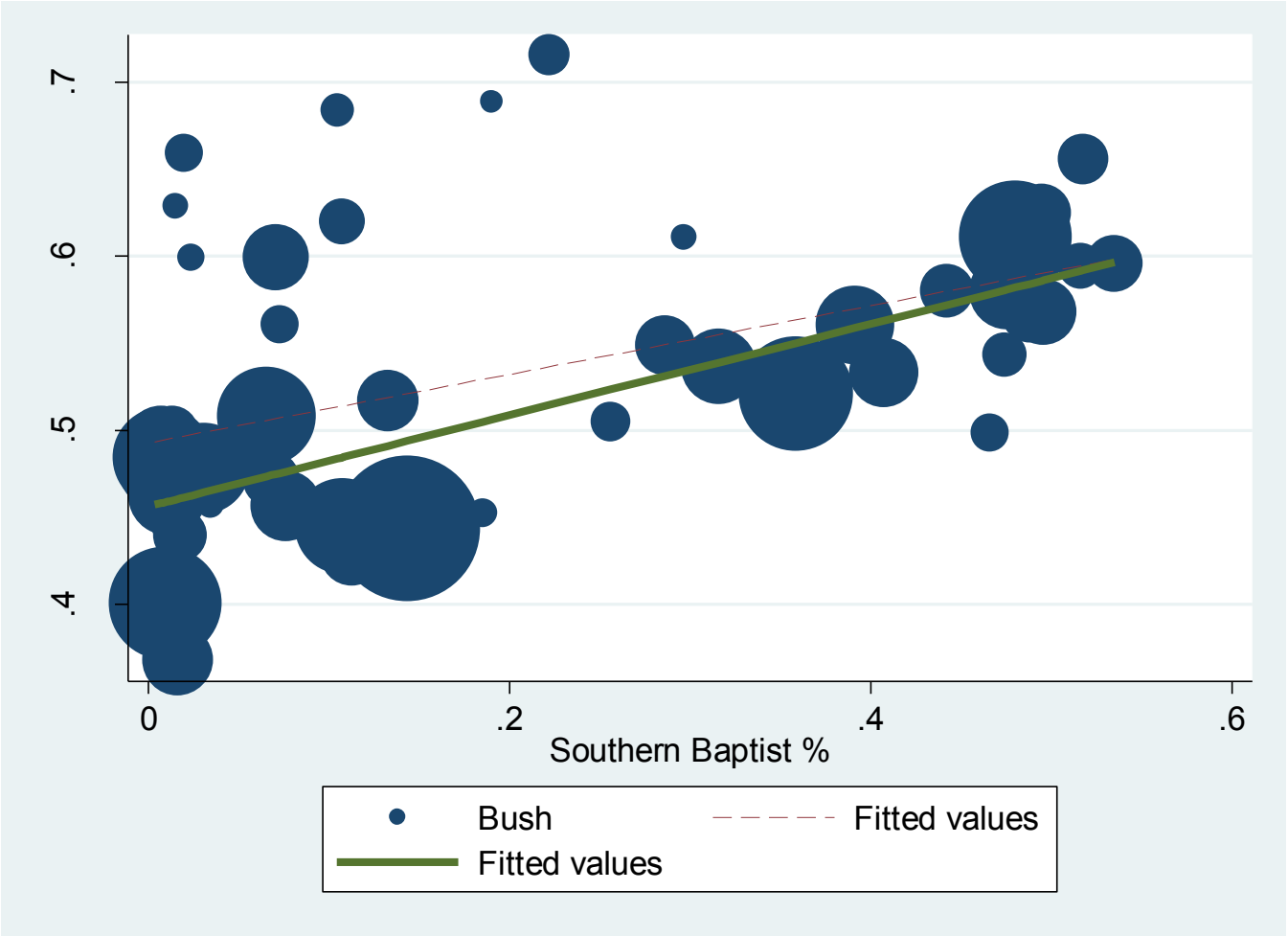


```
. reg bush sbc_mpct

      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
------------------------------------------------------------------------------
```

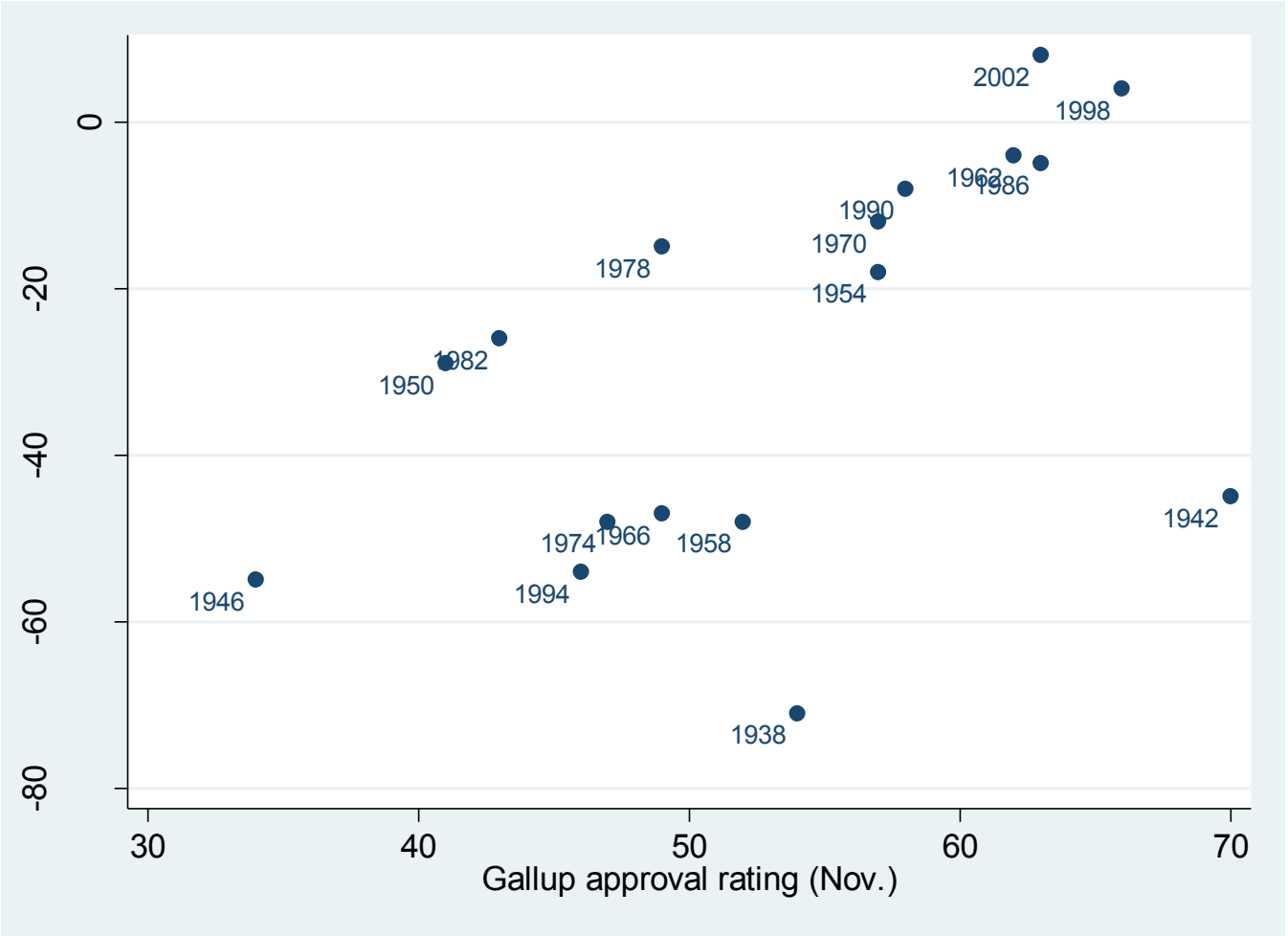

.7
.6
.5
.4
UT
ID
WY
NE
OK
ND
KS
AK
AL
TX
SD
IN
MT
MS
KY
SC
GA
WV
NC
AZ
AR
VA
MO
FL
CO
OH
NV
NM
NH
MI
OR
DE
WA
HI
IL
CA
CT
MD
NY
RI
MA
0
.2
.4
.6
Southern Baptist %
Bush
Fitted values

Weight by State Population

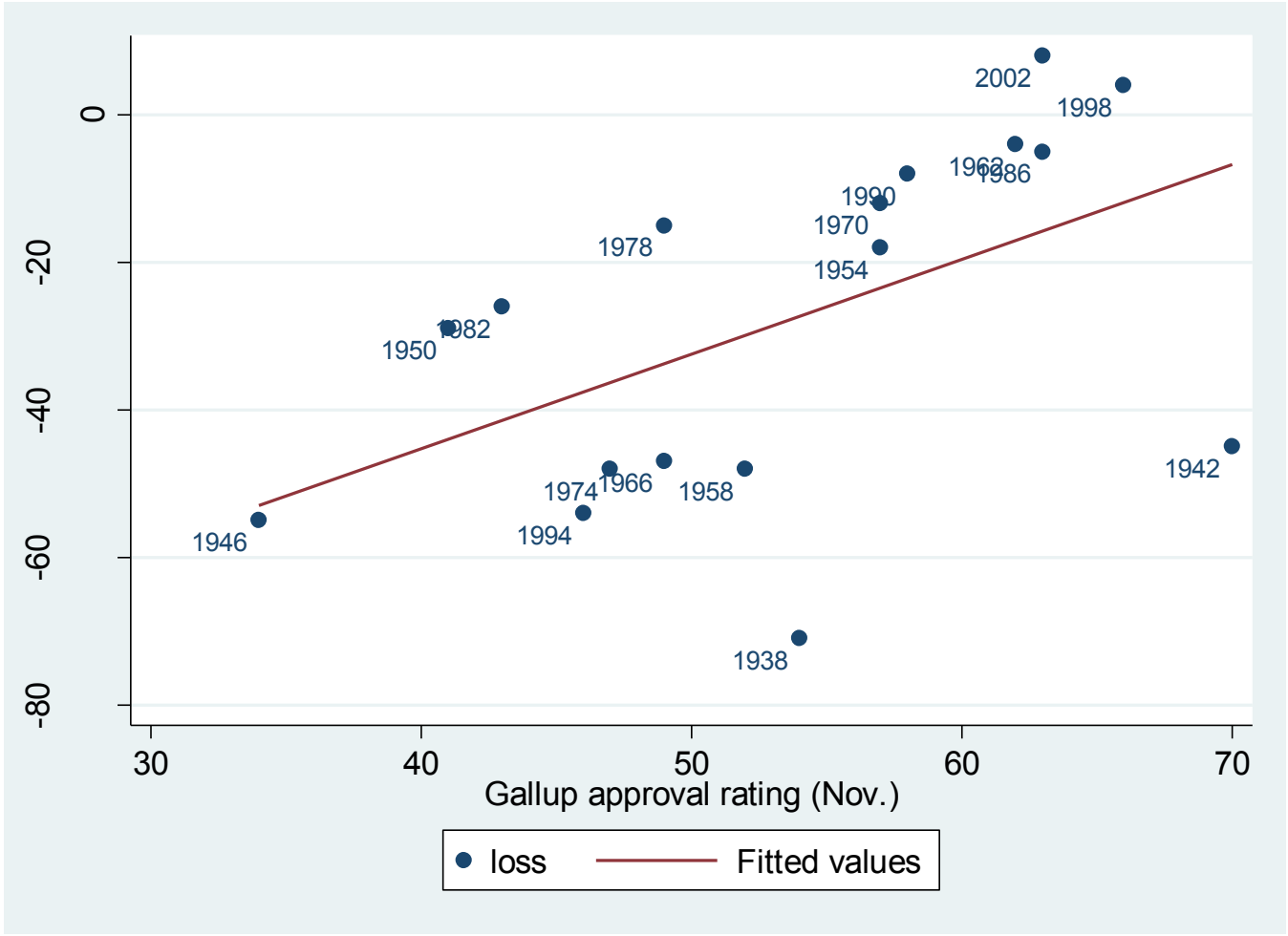
```
. reg bush sbc_mpct [aw=votes]
(sum of wgt is   1.2207e+08)

      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
------------------------------------------------------------------------------
```


.7
.6
.5
.4
0
.2
.4
.6
Southern Baptist %
Bush
Fitted values
Fitted values

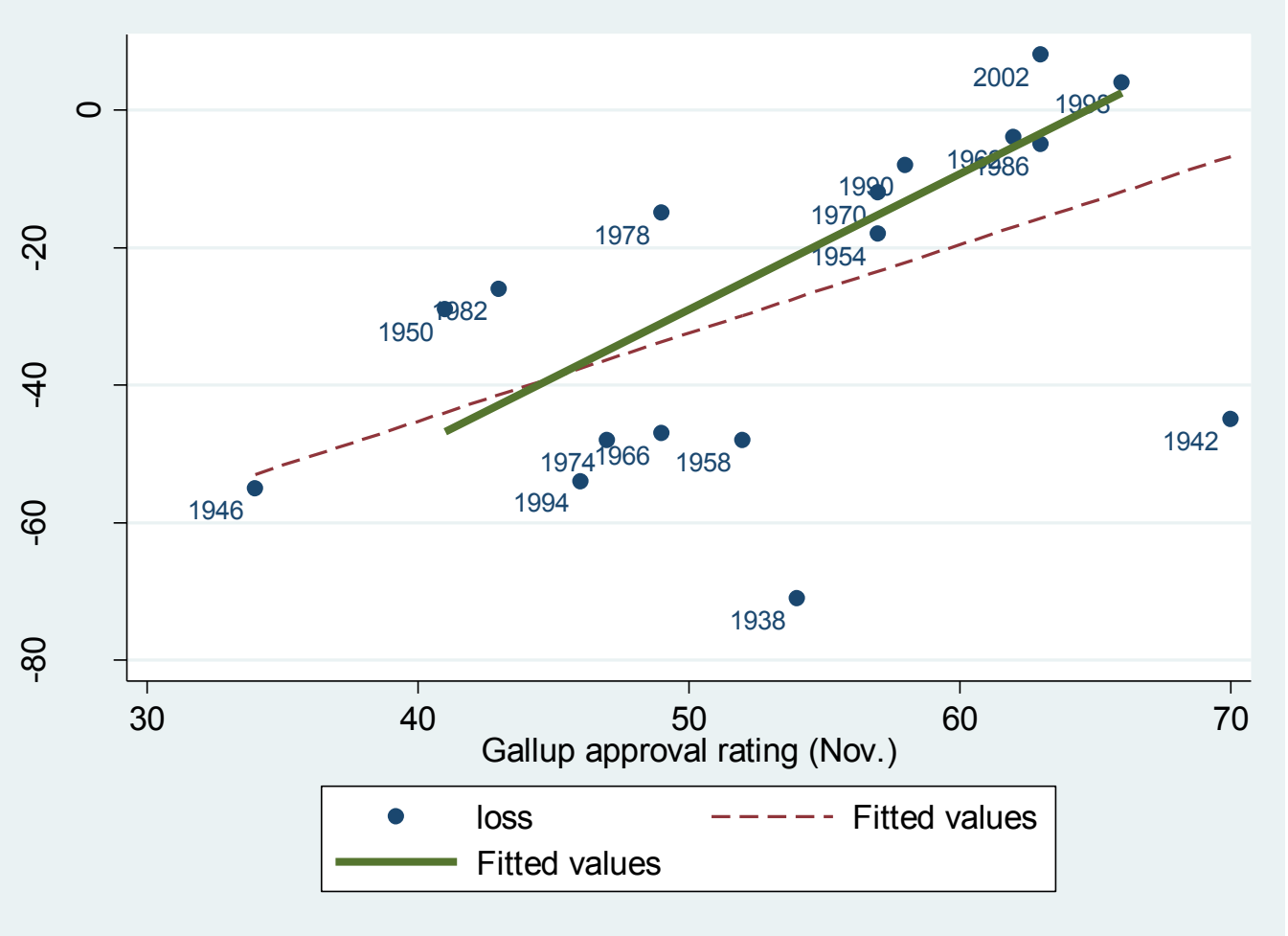
Midterm loss & pres. popularity



```
. 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

------------------------------------------------------------------------------
        loss |      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
------------------------------------------------------------------------------
```


0
-20
-40
-60
-80
30
40
50
60
70
Gallup approval rating (Nov.)
2002
1998
1963
1986
1990
1970
1978
1954
1982
1950
1942
1974
1966
1958
1946
1994
1938
loss
Fitted values

```
. reg loss gallup if year>1948

      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

------------------------------------------------------------------------------
        loss |      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
------------------------------------------------------------------------------
```



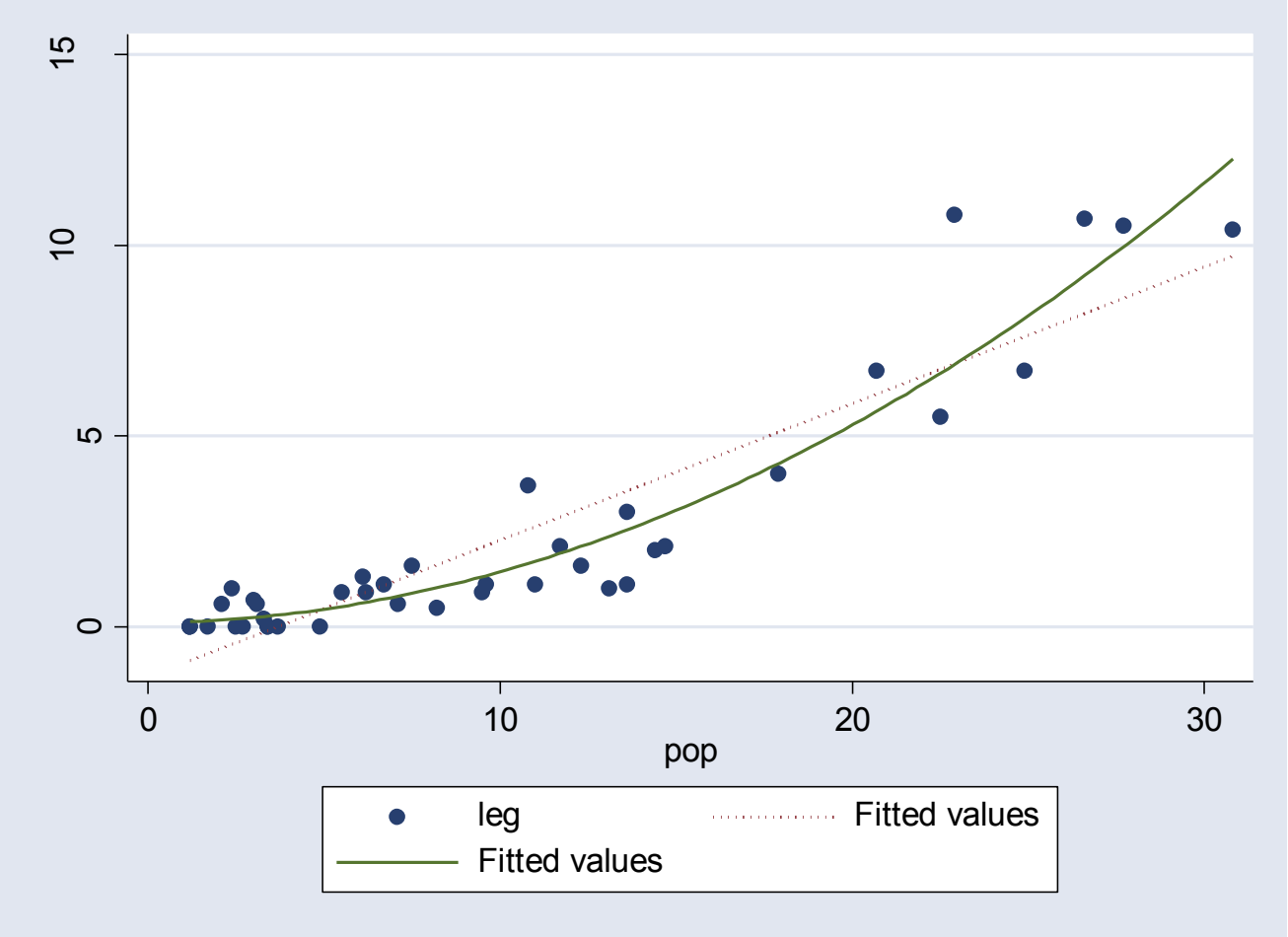
scatter loss gallup, mlabel(year) || lfit loss gallup || lfit loss gallup if year >1942

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)
  - $Y = a + \ln X^b/Z^c + e$ (linear in neither)

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



scatter beo pop || qfit beo pop

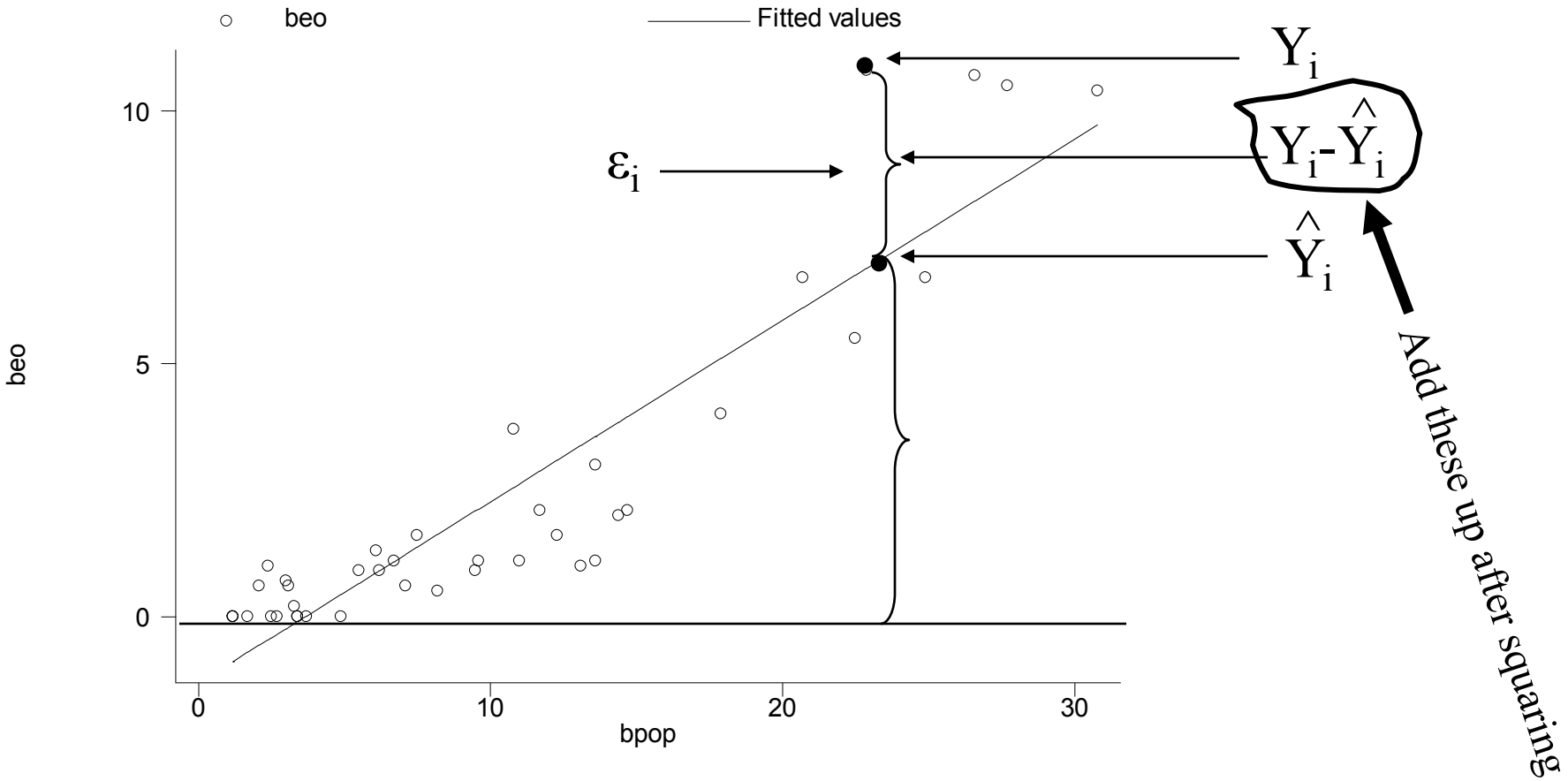
Log transformations

$Y = a + bX + e$	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	Cases where there's a natural limit on growth
$\ln Y = a + bX + e$	b = (dY/Y)/dX, or b = the % change in Y given a unit change in X	Exponential growth
$\ln Y = a + b \ln X + e$	b = (dY/Y)/(dX/X), or b = the % change in Y given a % change in X (elasticity)	Economic production

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)
    - Standard error of a coefficient
    - *t*-statistic: *coeff./s.e.*
- Nevertheless, you should know about
  - Standard Error of the Estimate (s.e.e.)
    - Also called Standard Error of the Regression
    - Regrettably called Root Mean Squared Error (Rout MSE) in Stata
  - R-squared

Standard error of the regression picture



Standard error of the regression

- Standard error of the estimate

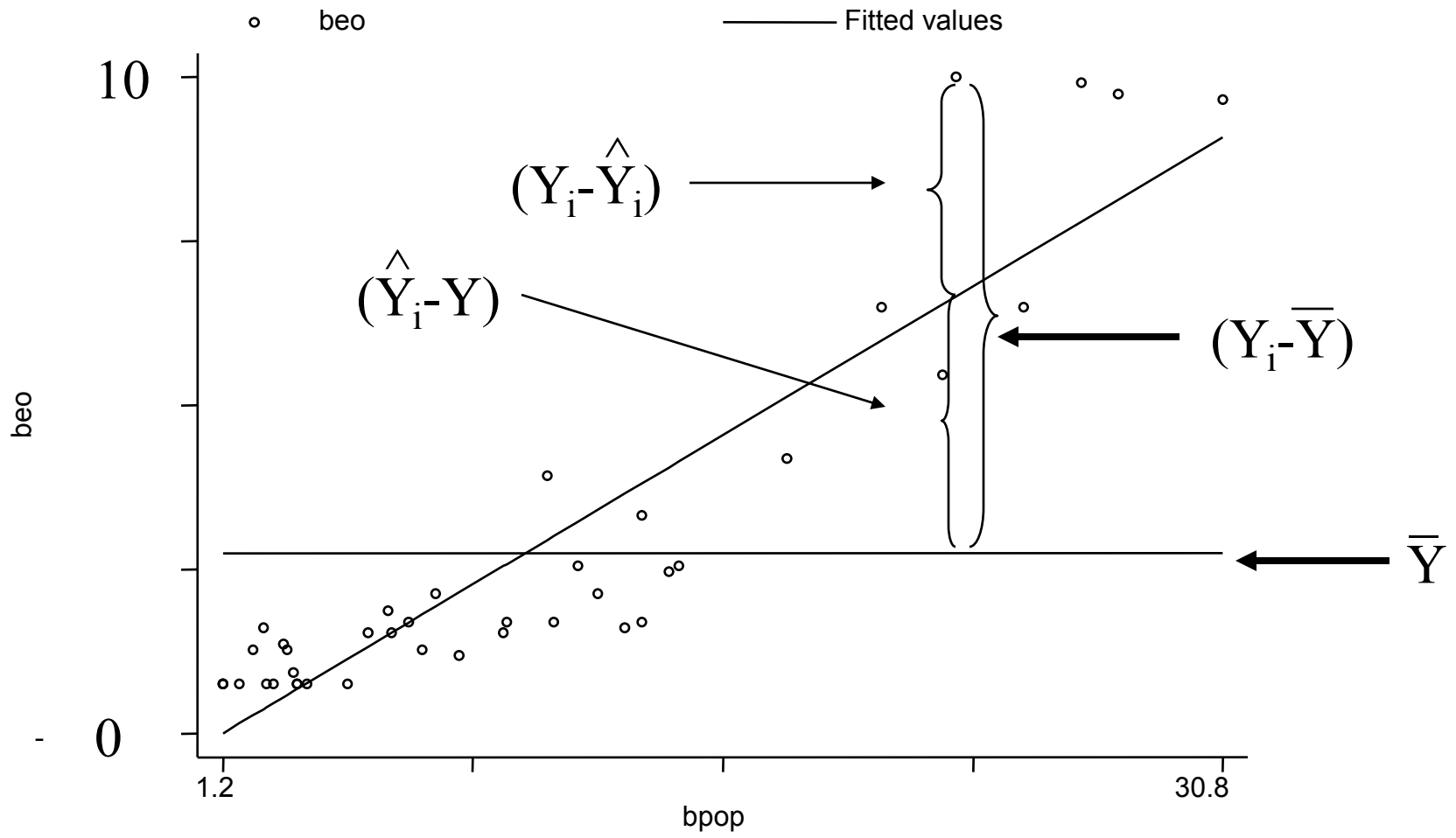
$$s.e.e. = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{d.f.}}$$

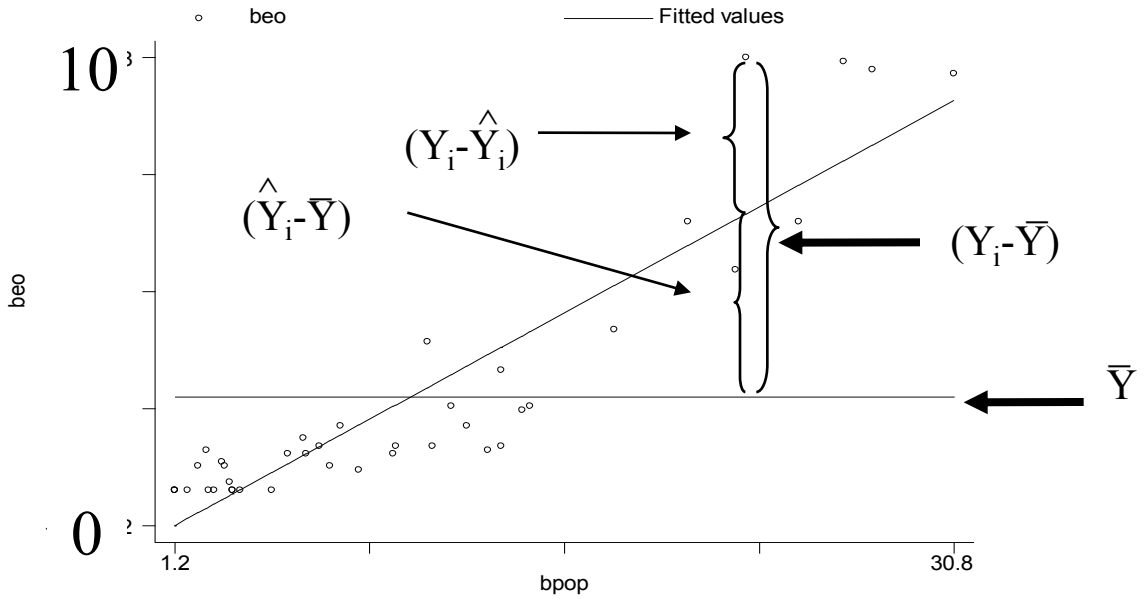
d.f. equals n minus the number of estimate coefficients (*B*s).
In bivariate regression case, *d.f.* = n-2.

Standard error of the regression

- Natural interpretation:
  - On average, in-sample predictions will be off the mark by about one standard error

R^2 picture


beo
Fitted values
10
0
-
beo
$(Y_i-\hat{Y}_i)$
$(\hat{Y}_i-Y)$
$(Y_i-\overline{Y})$
$\overline{Y}$
1.2
30.8
bpop



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

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

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

- R-squared

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

percent variance "explained"

Also called “coefficient of determination”

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

- 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