

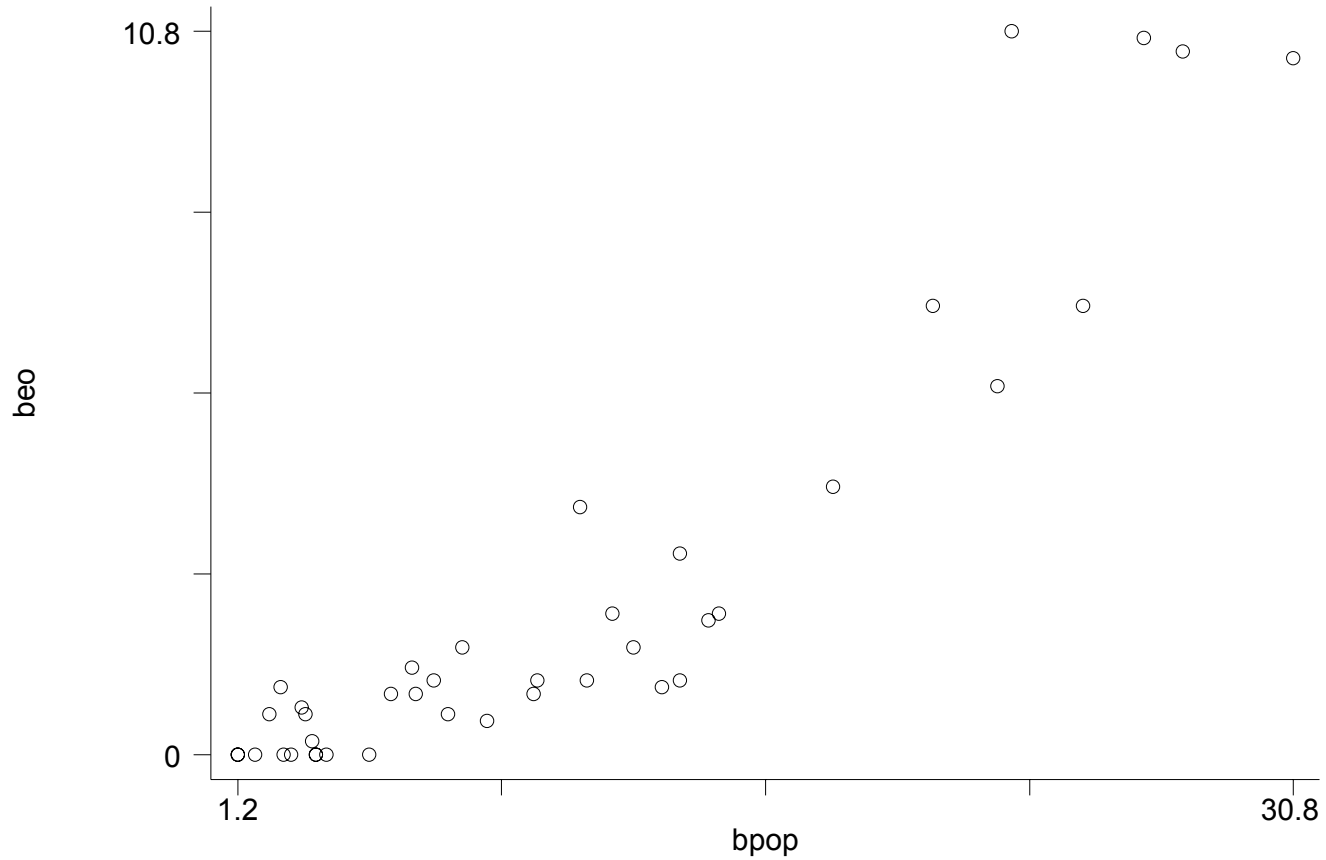
# Regression Forced March

17.871

Spring 2004

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

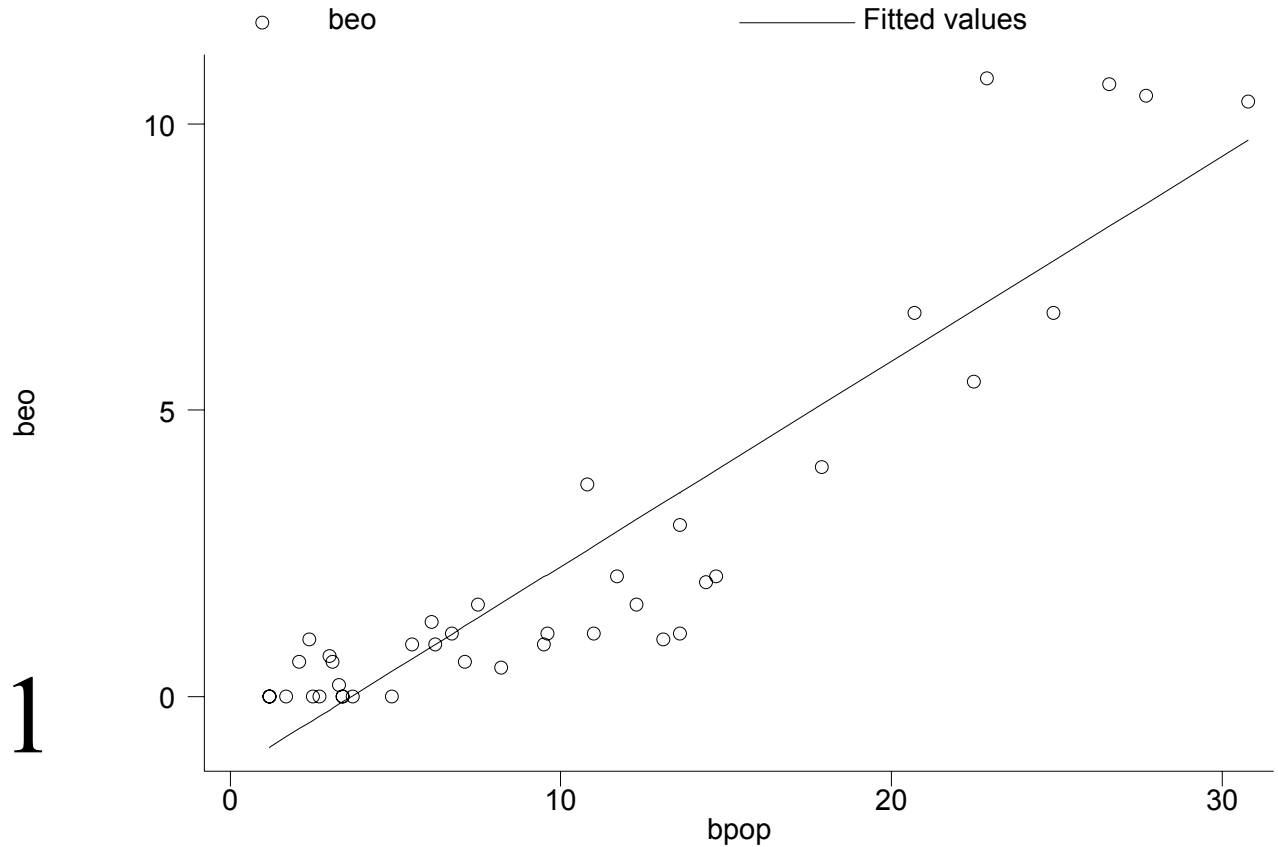
# Black Elected Officials Example I



# 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

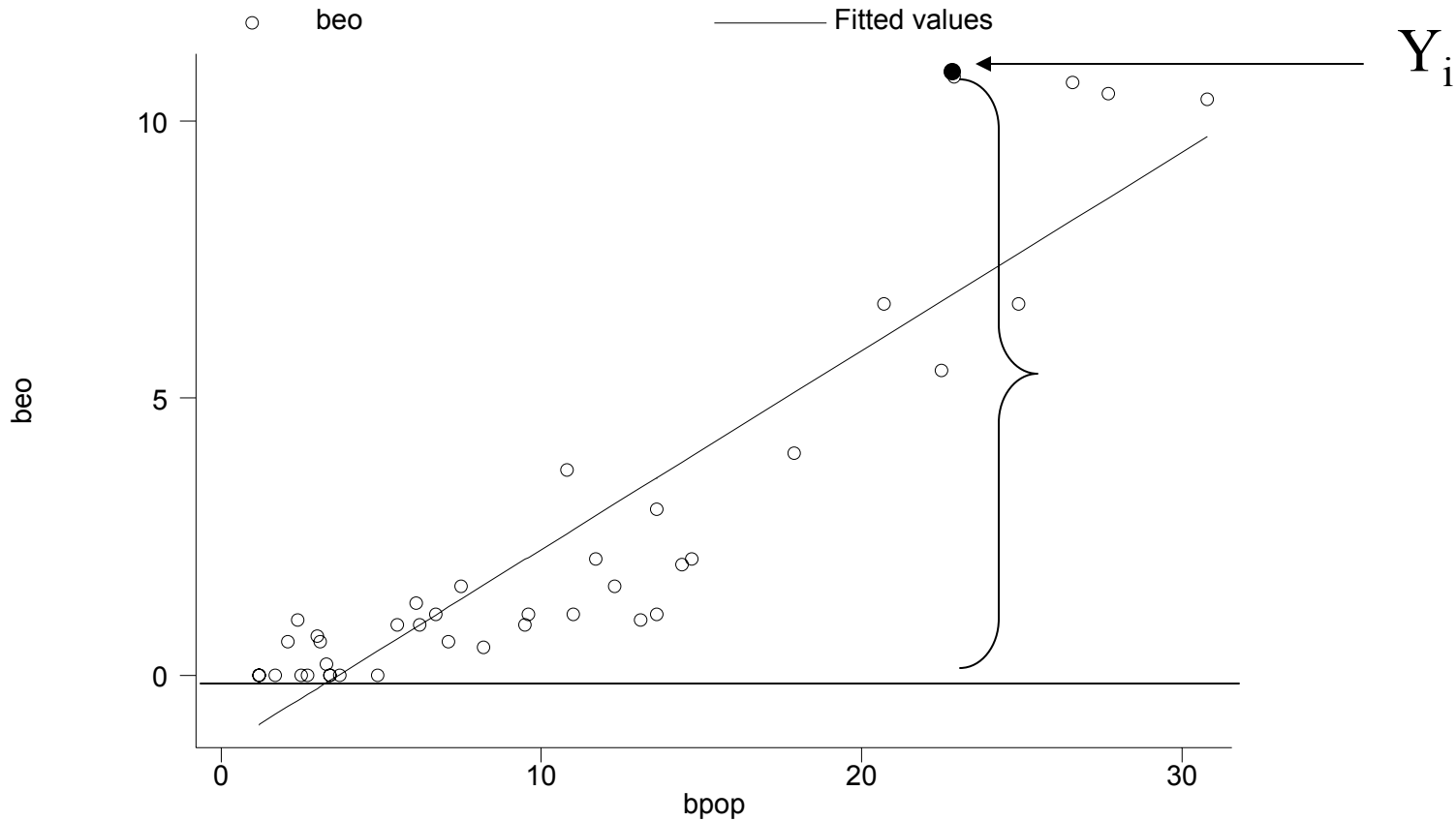


$$\beta_0 = -1.31$$

$$\beta_1 = 0.359$$

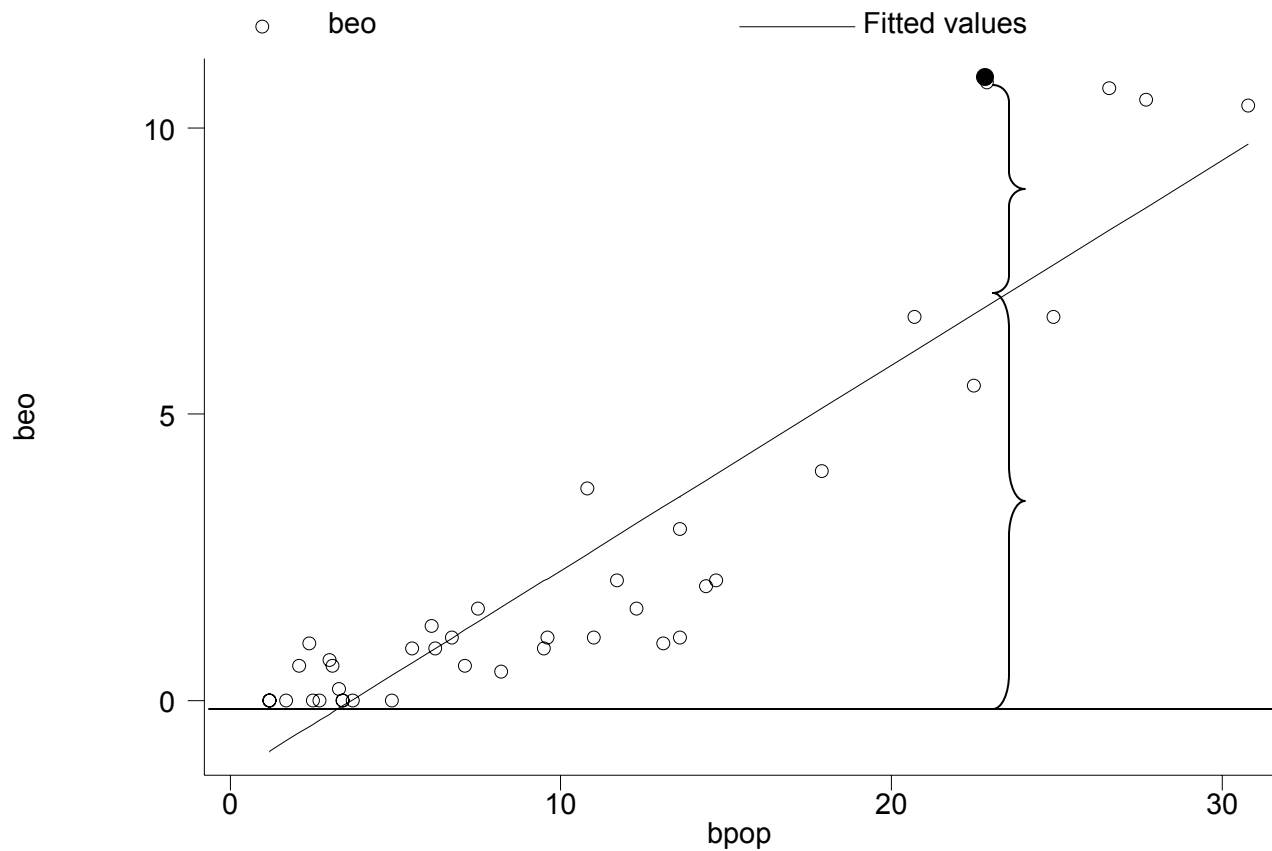
# How did we get that line?

1. Pick a representative value of  $Y_i$



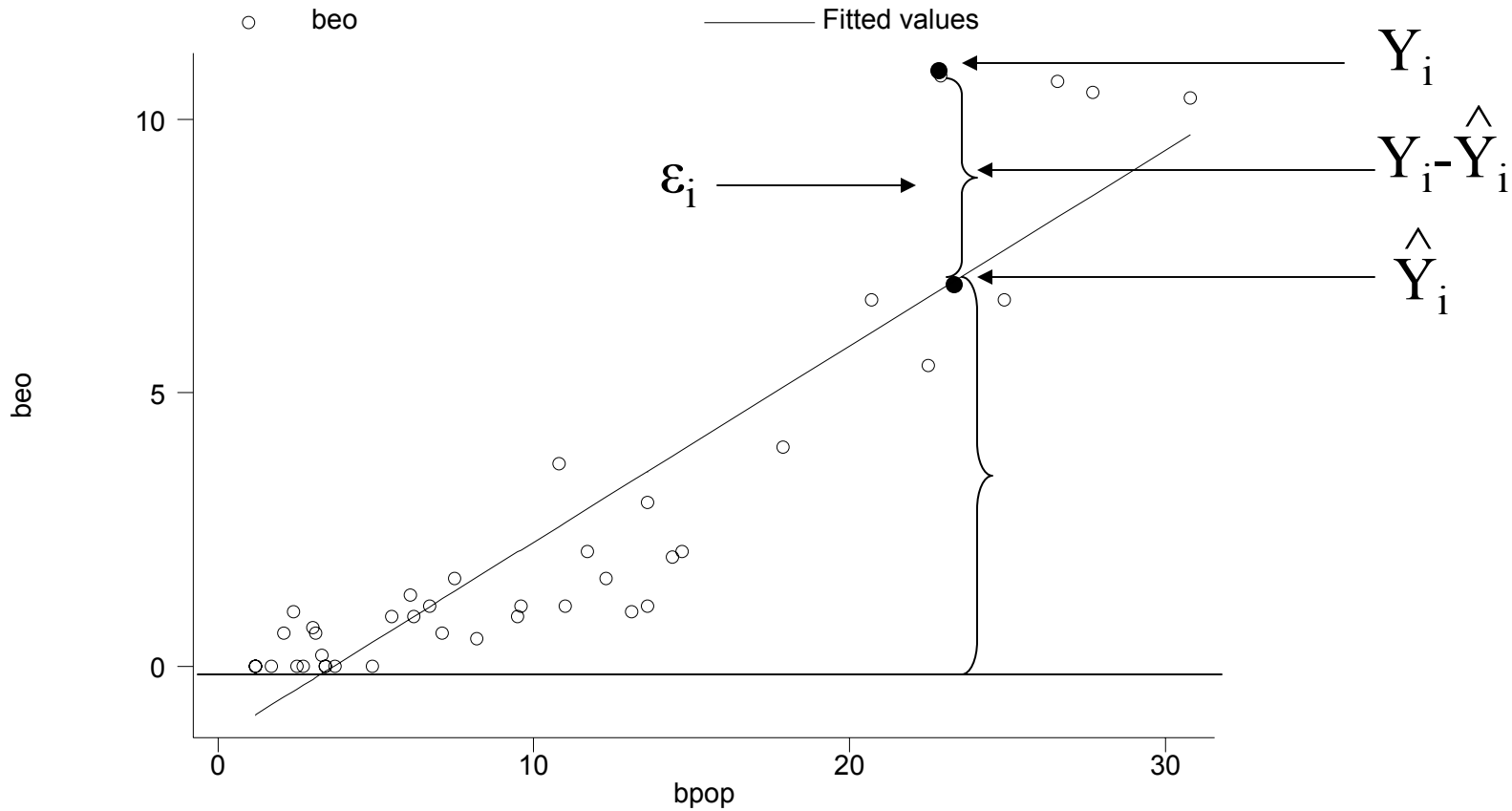
# How did we get that line?

## 2. Decompose $Y_i$ into two parts



# How did we get that line?

## 3. Label the points



# Stop a moment: What is $g_i$ ?

- Vagueness of theory
- Unavailability of data
- Core *vs.* peripheral variables
- Intrinsic randomness in human behavior
- Poor proxies (i.e., measurement error)
- Principle of parsimony
- Wrong functional form
  
- See Gujarati, pp. 45-47

# The Method of Least Squares (start here)

Pick  $\beta_0$  and  $\beta_1$  to minimize

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

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

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

$$\beta_1 = \frac{\sum_{i=1}^n (\bar{Y} - Y_i)(\bar{X} - X_i)}{\sum_{i=1}^n (\bar{X} - X_i)^2} \quad \text{or}$$

(Gujarati 3.1.6,  
p. 62)

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

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

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \quad (\text{Gujarati 3.1.7, p. 62})$$

Note that if you rearrange.....

$$\bar{Y} = \beta_0 + \beta_1 \bar{X}$$

# How to Think About Regression Results

- Deterministically
- Expectations

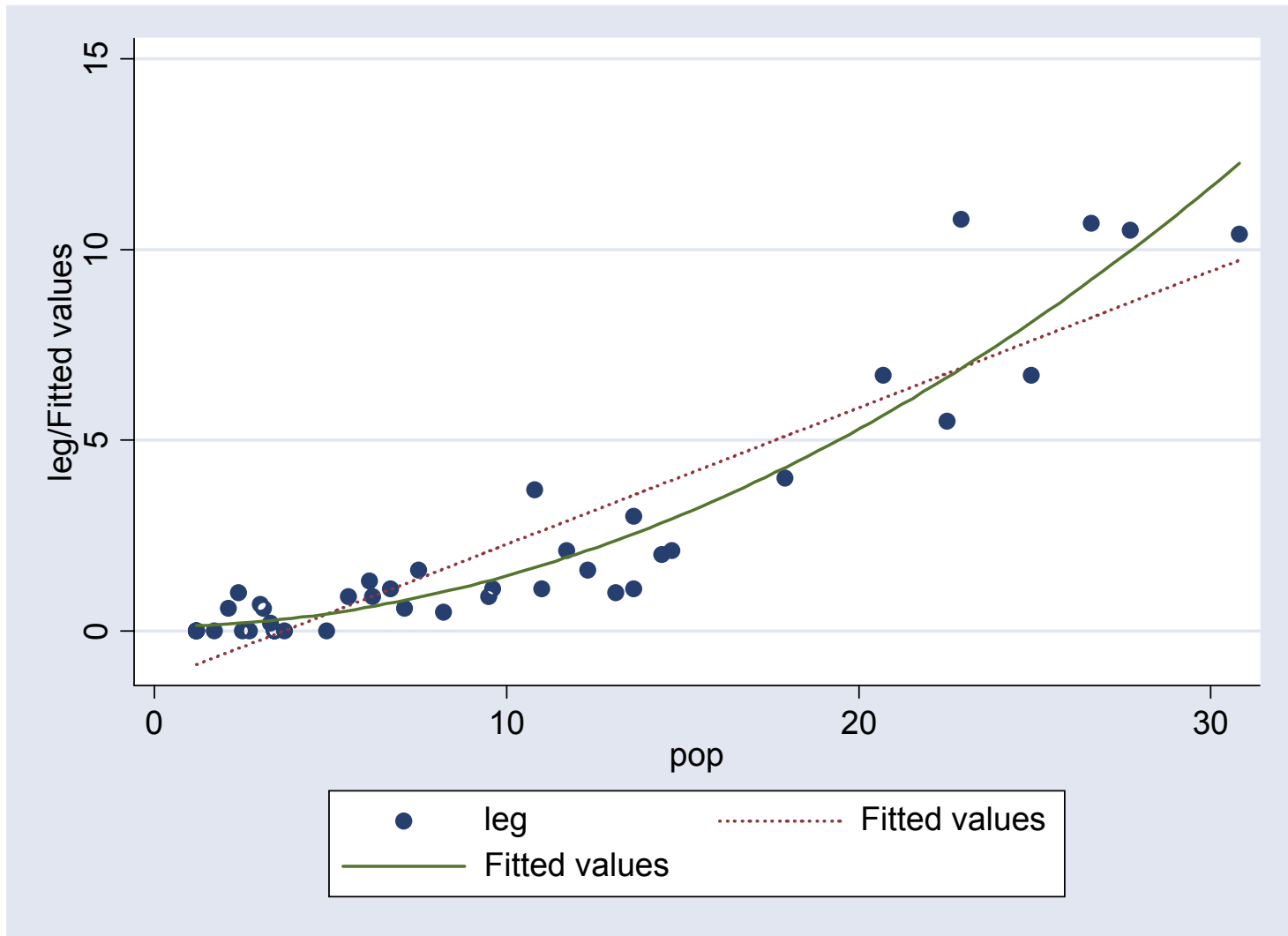
# Playing with the Stark Javascript

<http://www.stat.berkeley.edu/users/stark/Java/Html/Correlation.htm>

# 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)
- Gujarati pp. 42-43

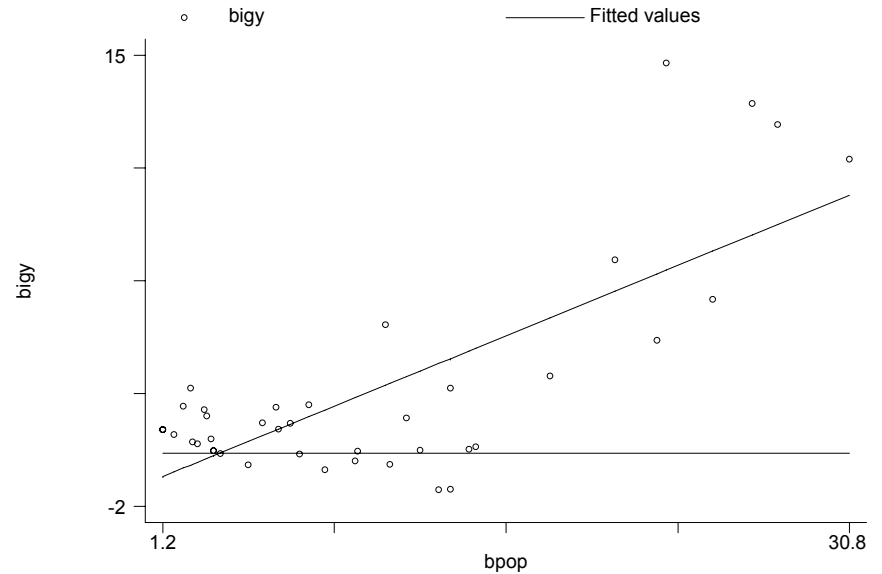
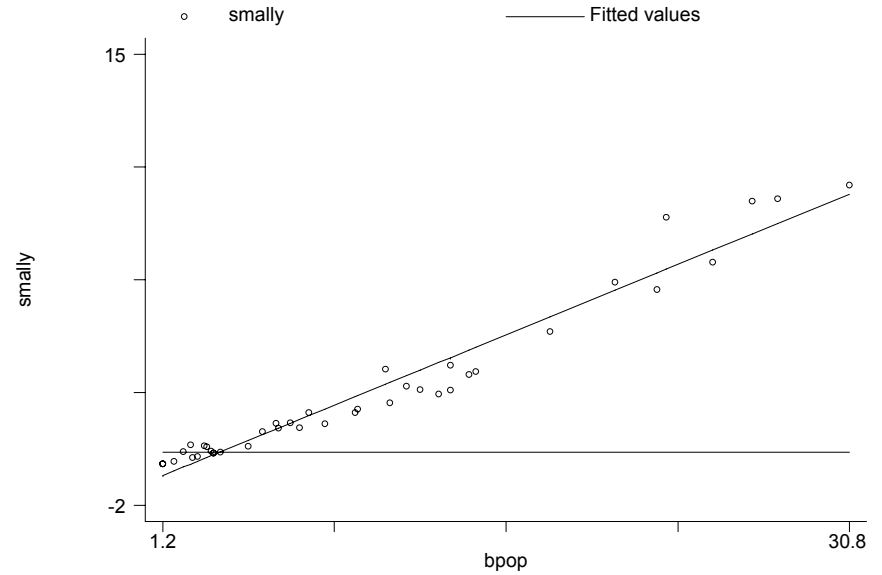
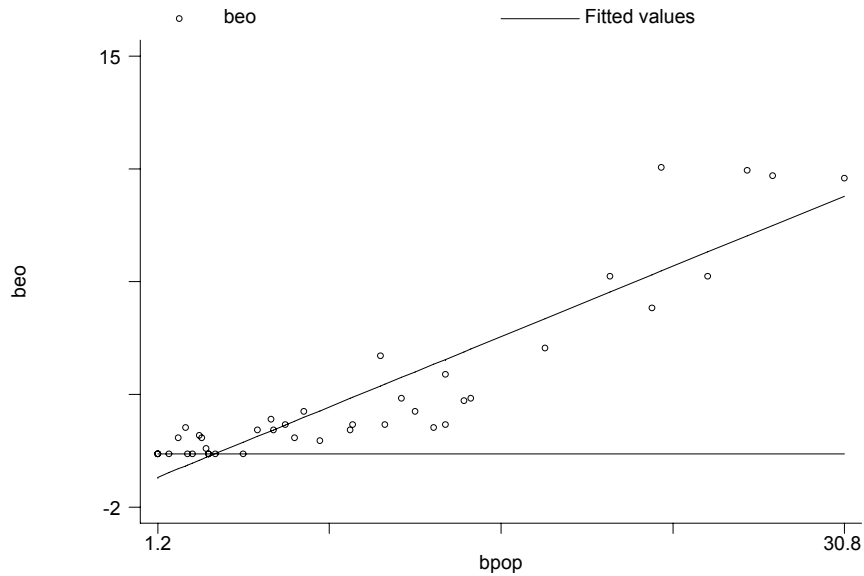
# Black Elected Officials



# 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?



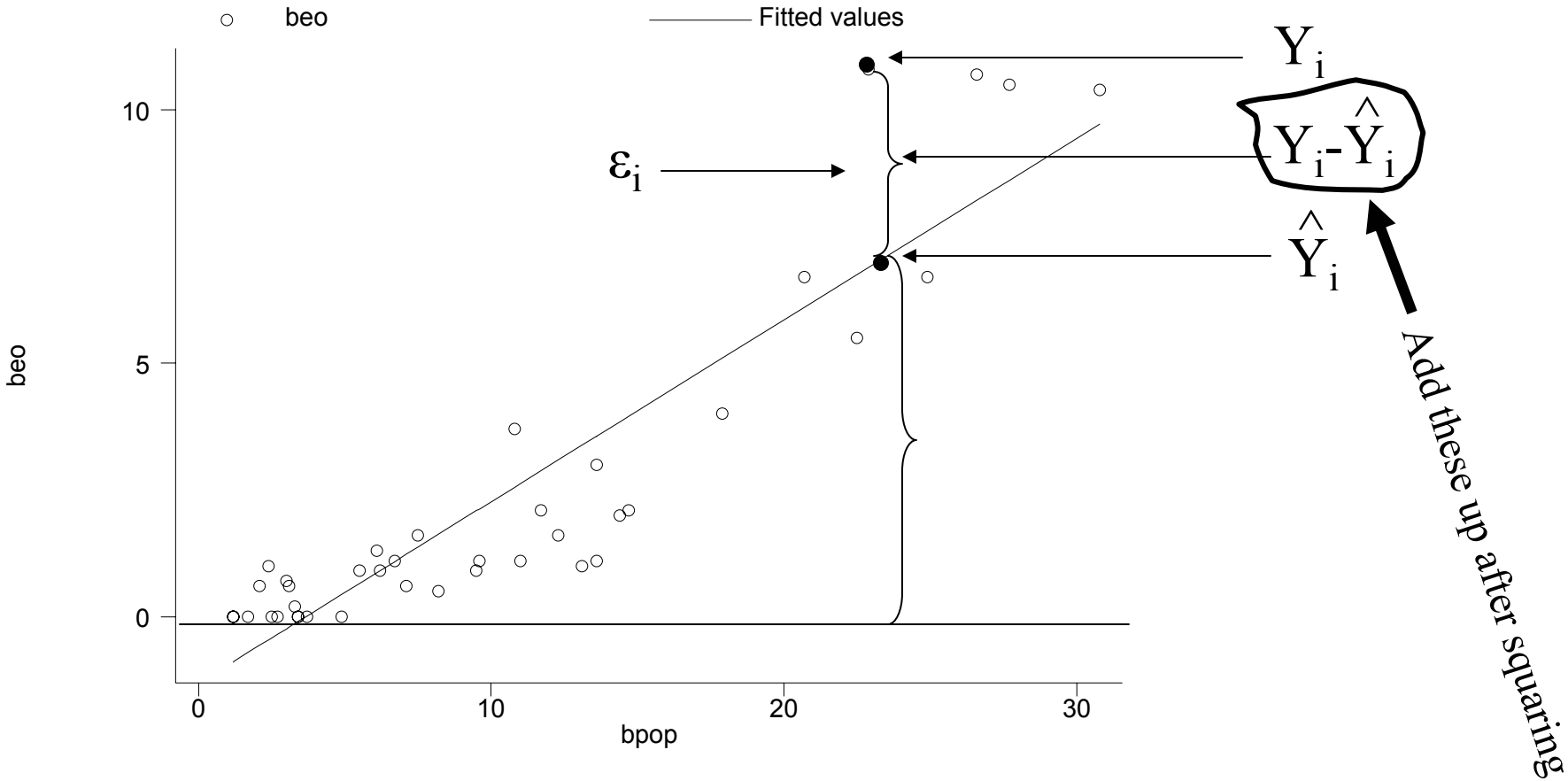
# Judging results

- Substantive interpretation of coefficients
- Technical judgment of regression
  - Judgment of coefficients
  - Judgment of overall fit
- Gujarati pp. 76-87

# Determining Goodness of Fit I

- Coefficients
  - Standard error of a coefficient
  - $t$ -statistic:  $\text{coeff.}/s.e.$

# Standard error of the regression picture



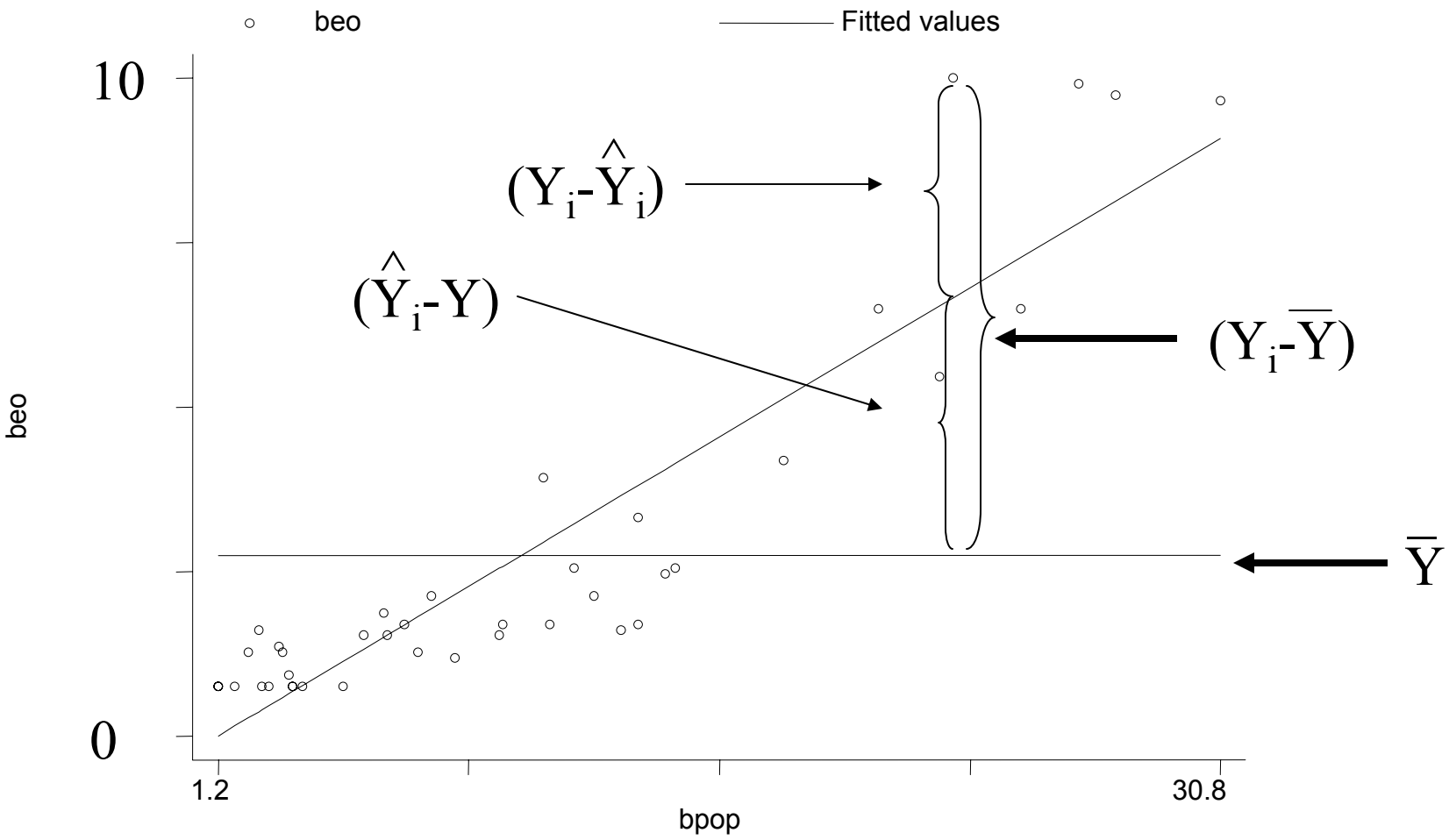
# Determining Goodness of Fit

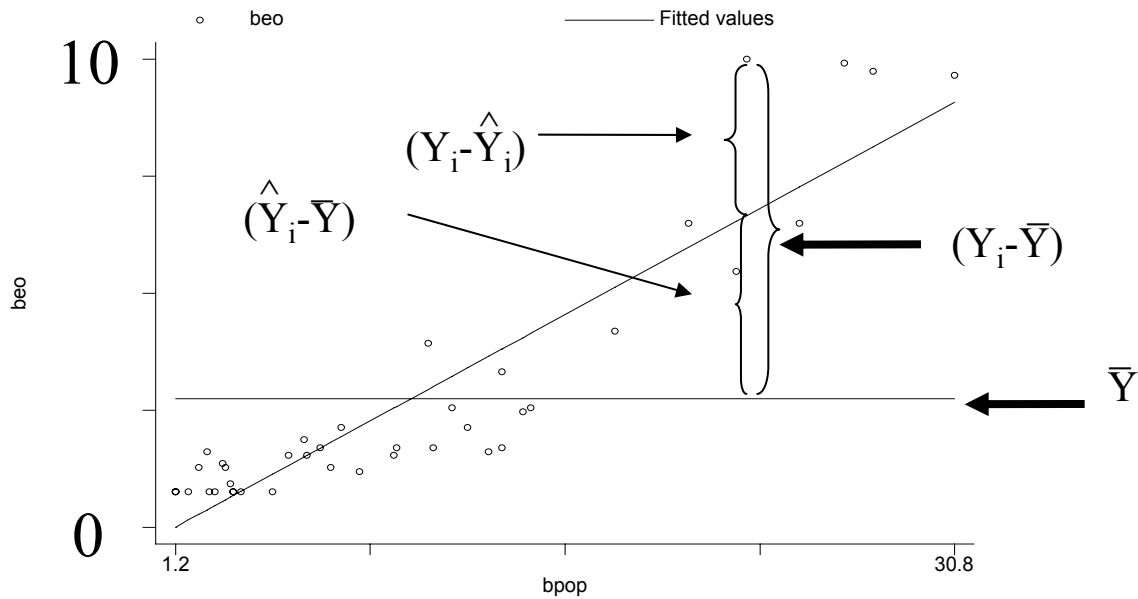
- Standard error of the regression or standard error of estimate (Root mean square error in STATA)

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

$$d.f. = n-2$$

# R<sup>2</sup> picture





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

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

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

# Determining Goodness of Fit

- 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"

“coefficient of determination” Gujarati pp. 81-87

# Return to Black Elected Officials

## Example

```
. reg beo bpop
```

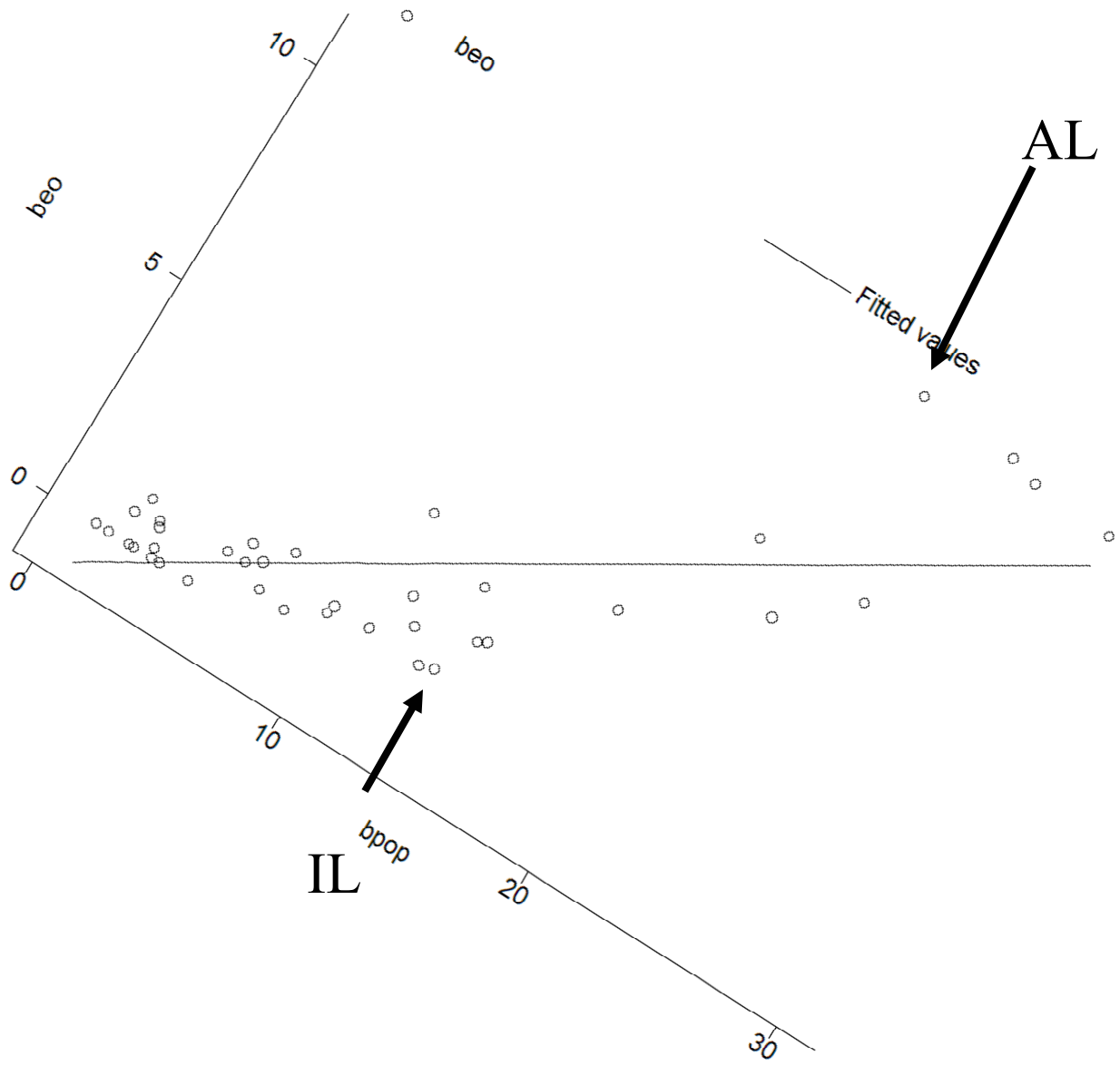
Source	SS	df	MS	Number of obs =	41
Model	351.26542	1	351.26542	F( 1, 39) =	202.56
Residual	67.6326195	39	1.73416973	Prob > F =	0.0000
Total	418.898039	40	10.472451	R-squared =	0.8385
				Adj R-squared =	0.8344
				Root MSE =	1.3169

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

# Residuals

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



# One important numerical property of residuals

- The sum of the residuals is zero.

# Regression Commands in STATA

- `reg depvar indvars`
- `predict newvar`
- `predict newvar, resid`

E.g., Use Residuals to control for  
“size effects” on teaching  
evaluations

- STATA simulation

# Why It's Called Regression

