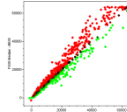


Bivariate Relationships

17.871

2014

Testing associations (not causation!)

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

Continuous DV, continuous EV

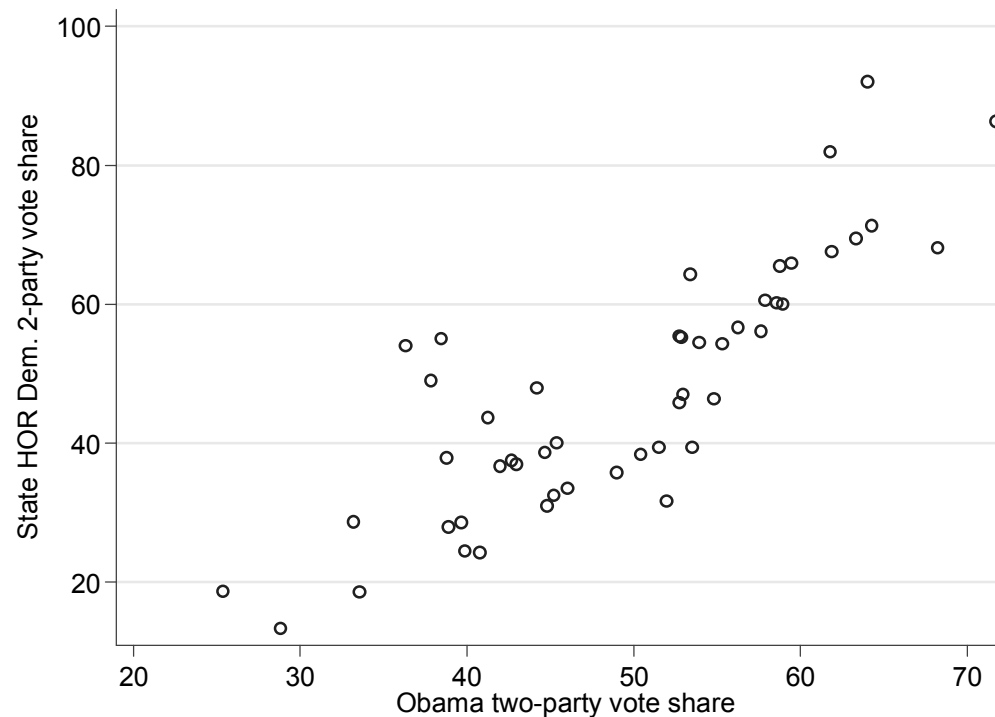
- Dependent Variable: DV
- Explanatory (or independent) Variable: EV (not IV)

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 Support for Obama (2012) and Democrats in State Legislatures (2013)*



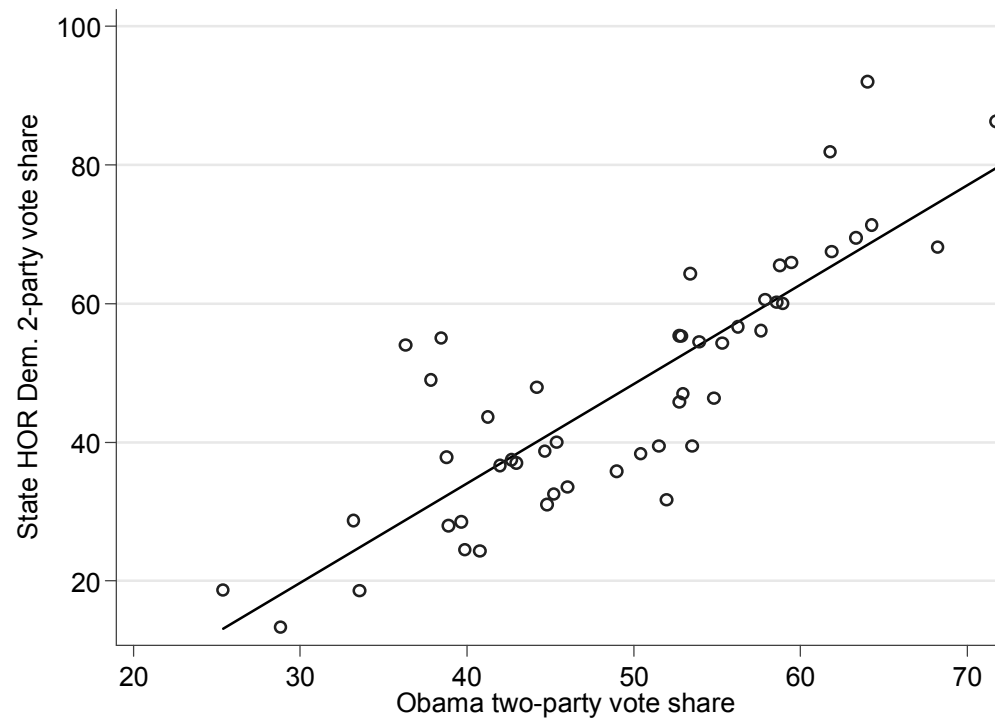
*Why we want to know this: to test the degree of partisan gerrymandering after 2010.

The linear relationship between two variables

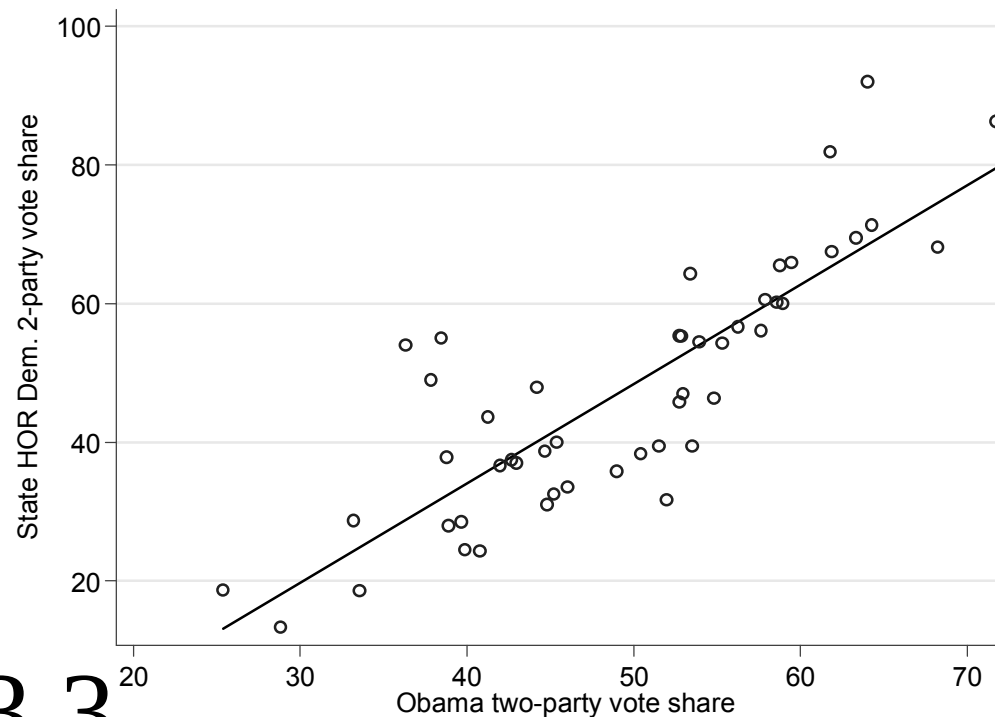
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

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

Linear Relationship between Support for Obama and Democrats in State Legislatures



Linear Relationship between Support for Obama and Democrats in State Legislatures

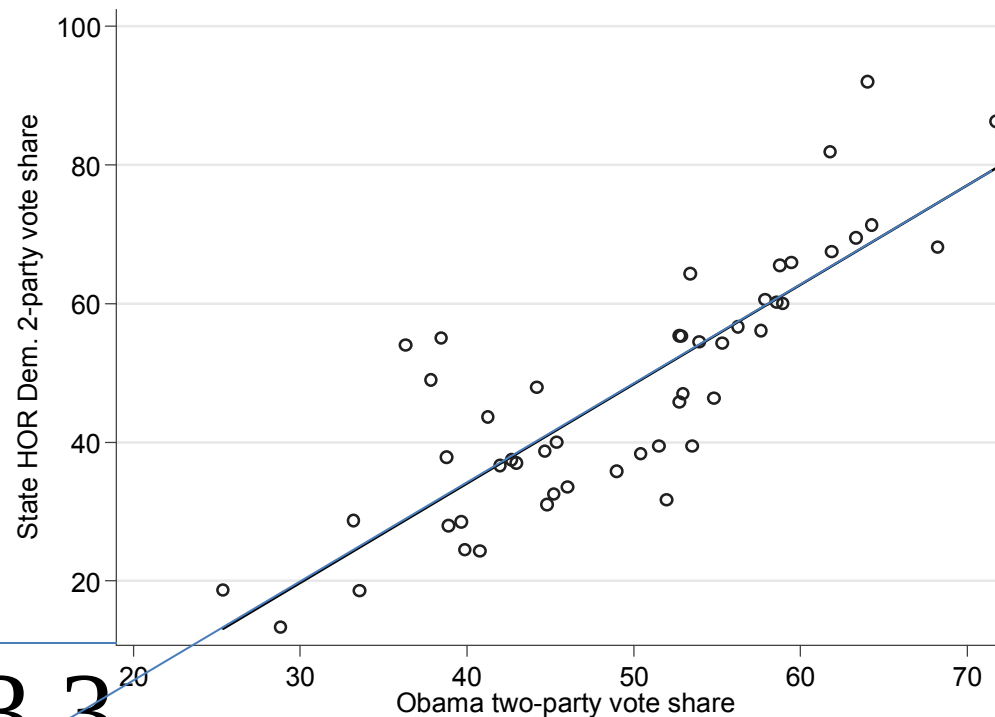


$$\beta_0 = -23.3$$

$$\beta_1 = 1.43$$

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

Linear Relationship between Support for Obama and Democrats in State Legislatures



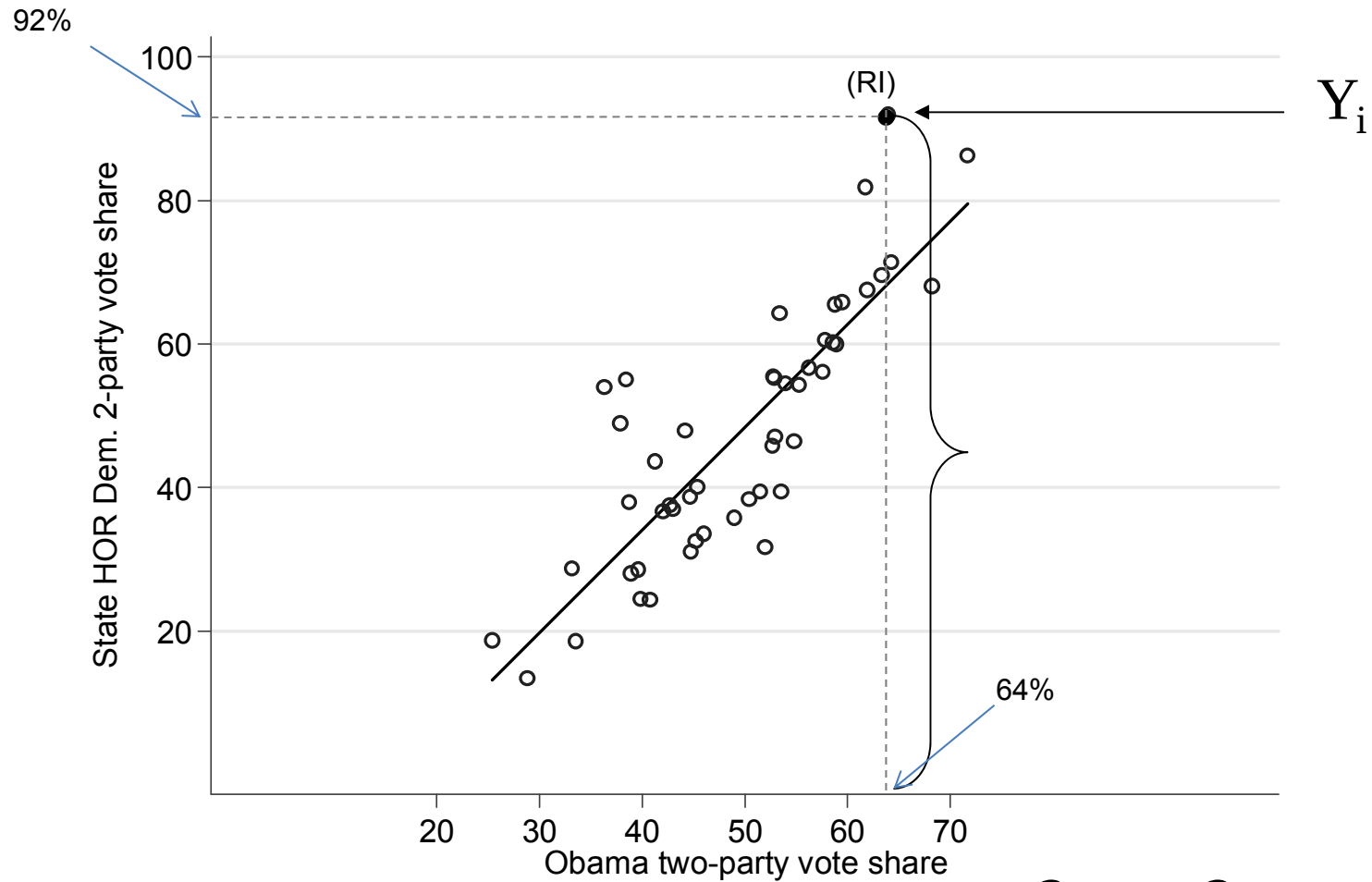
$$\beta_0 = -23.3$$

$$\beta_1 = 1.43$$

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

How did we get that line?

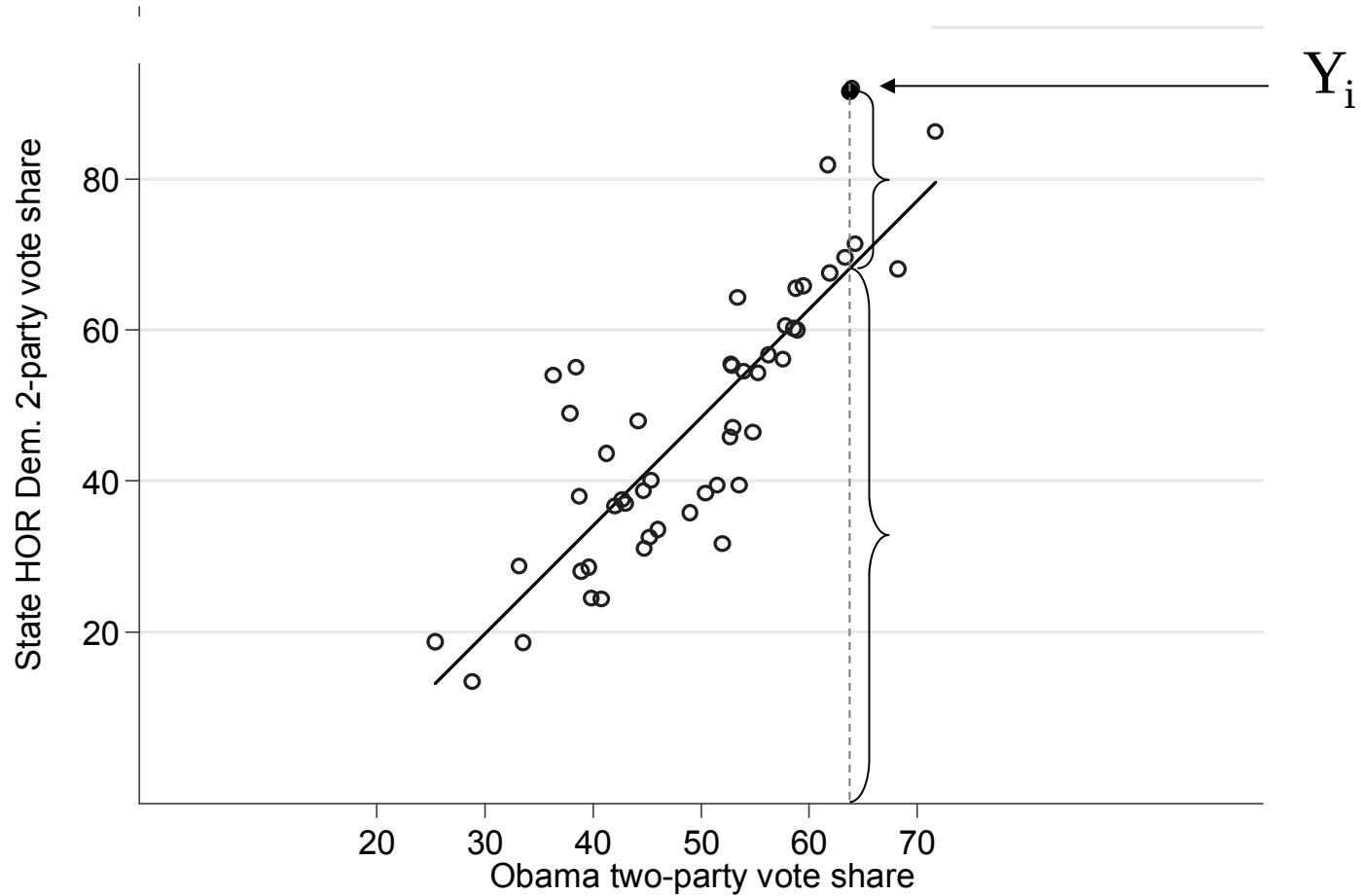
1. Pick a value of Y_i



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

How did we get that line?

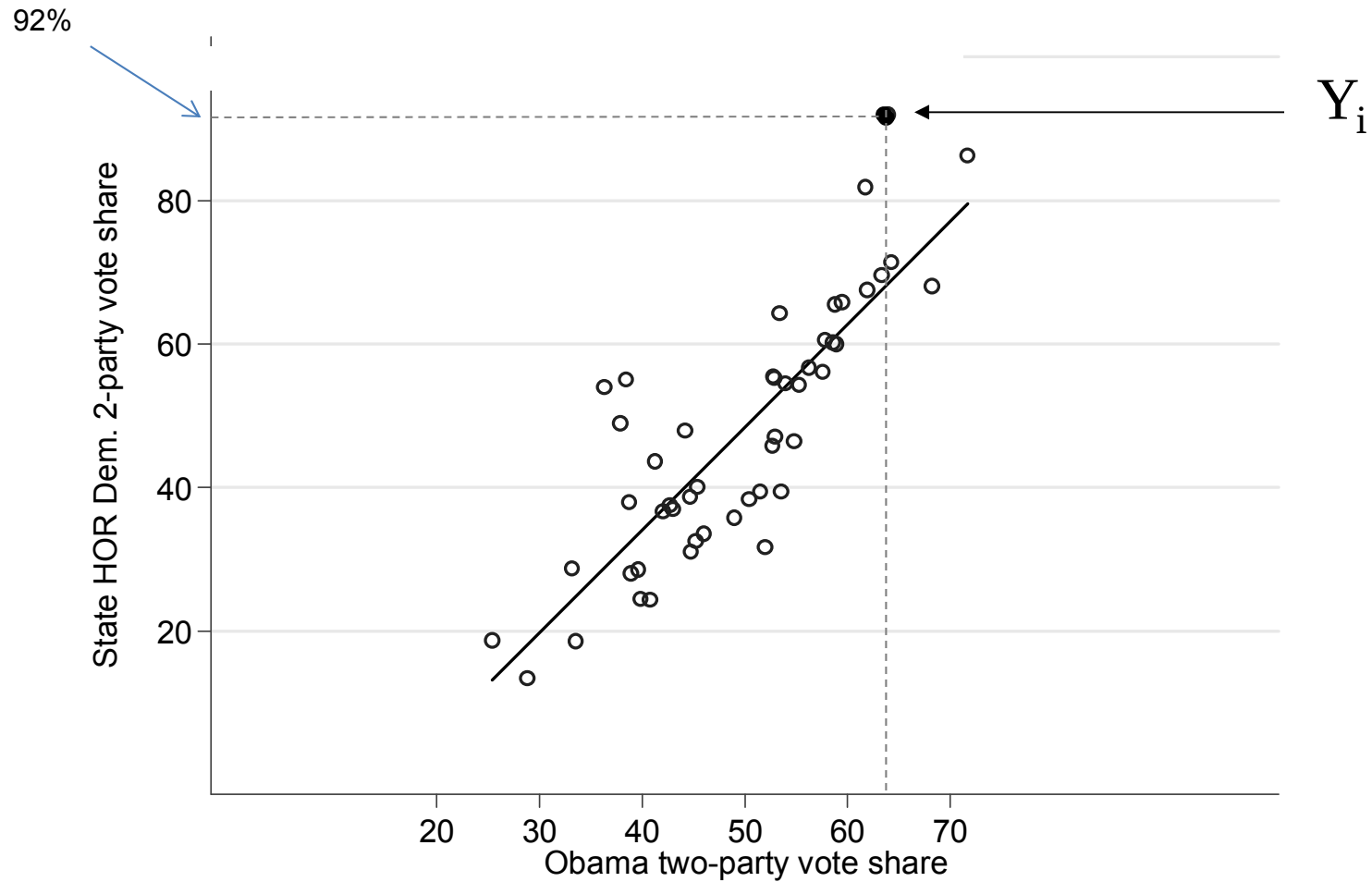
2. Decompose Y_i into two parts



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

How did we get that line?

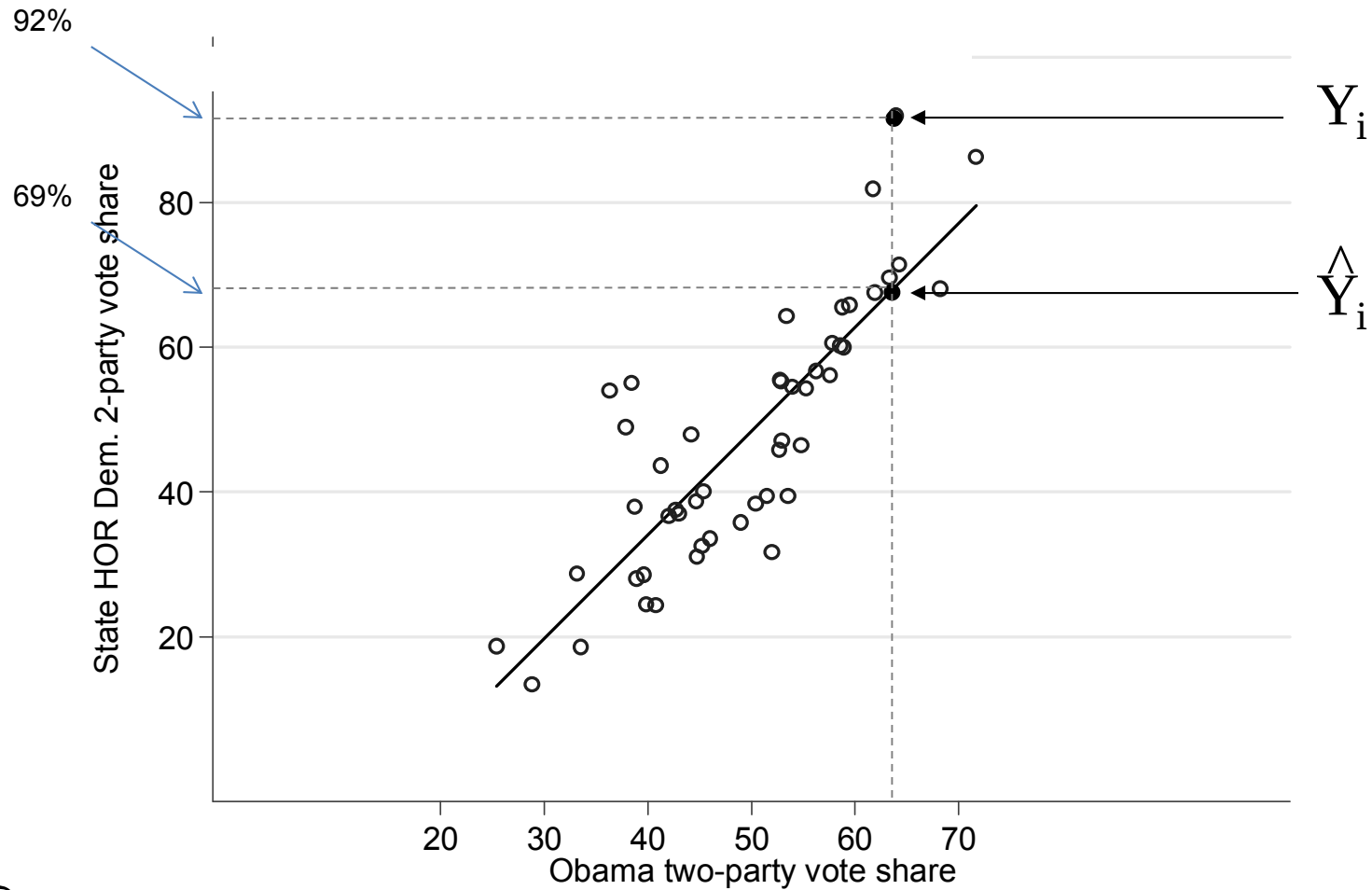
3. Label the points



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

How did we get that line?

3. Label the points

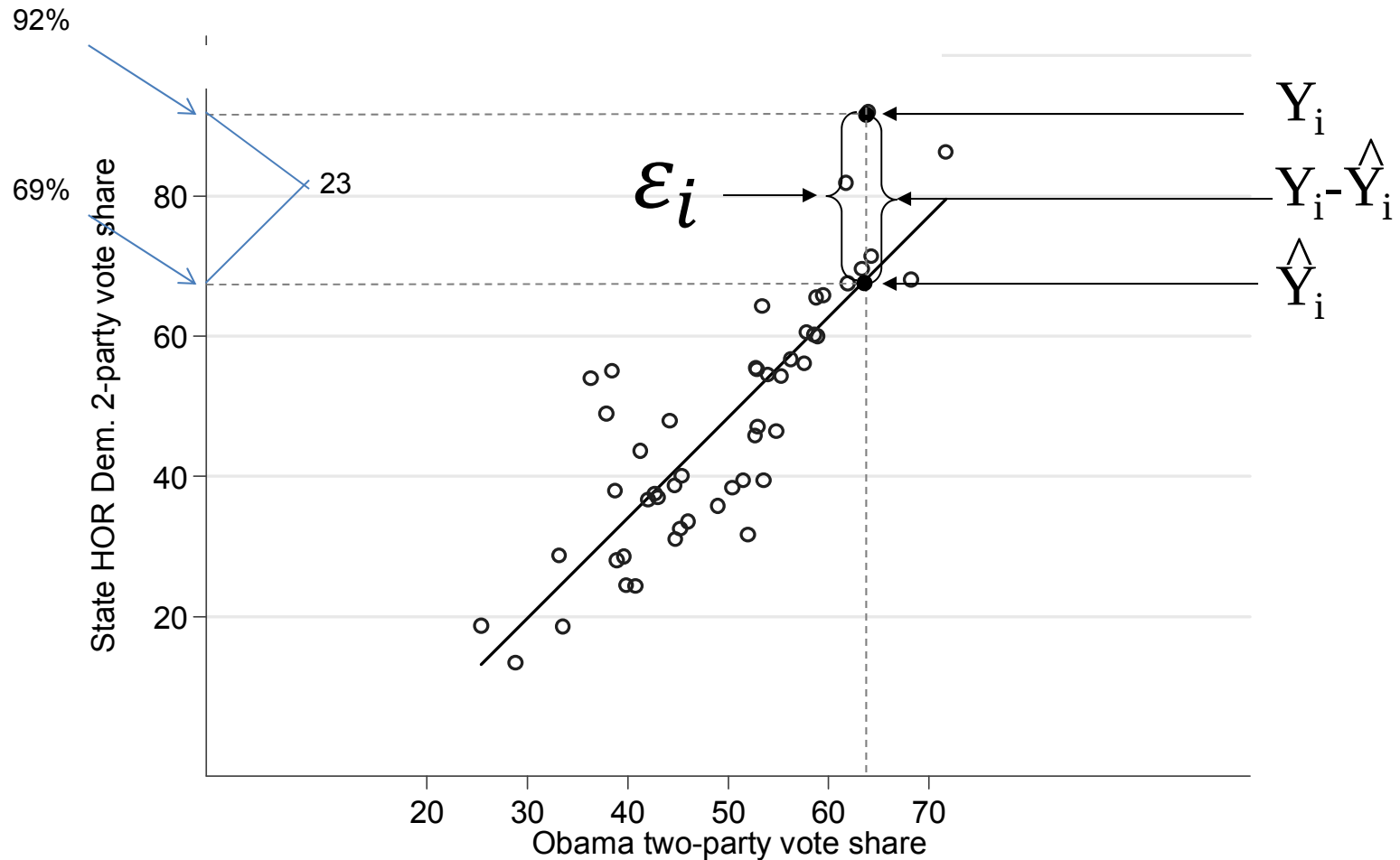


$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

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

How did we get that line?

3. Label the points

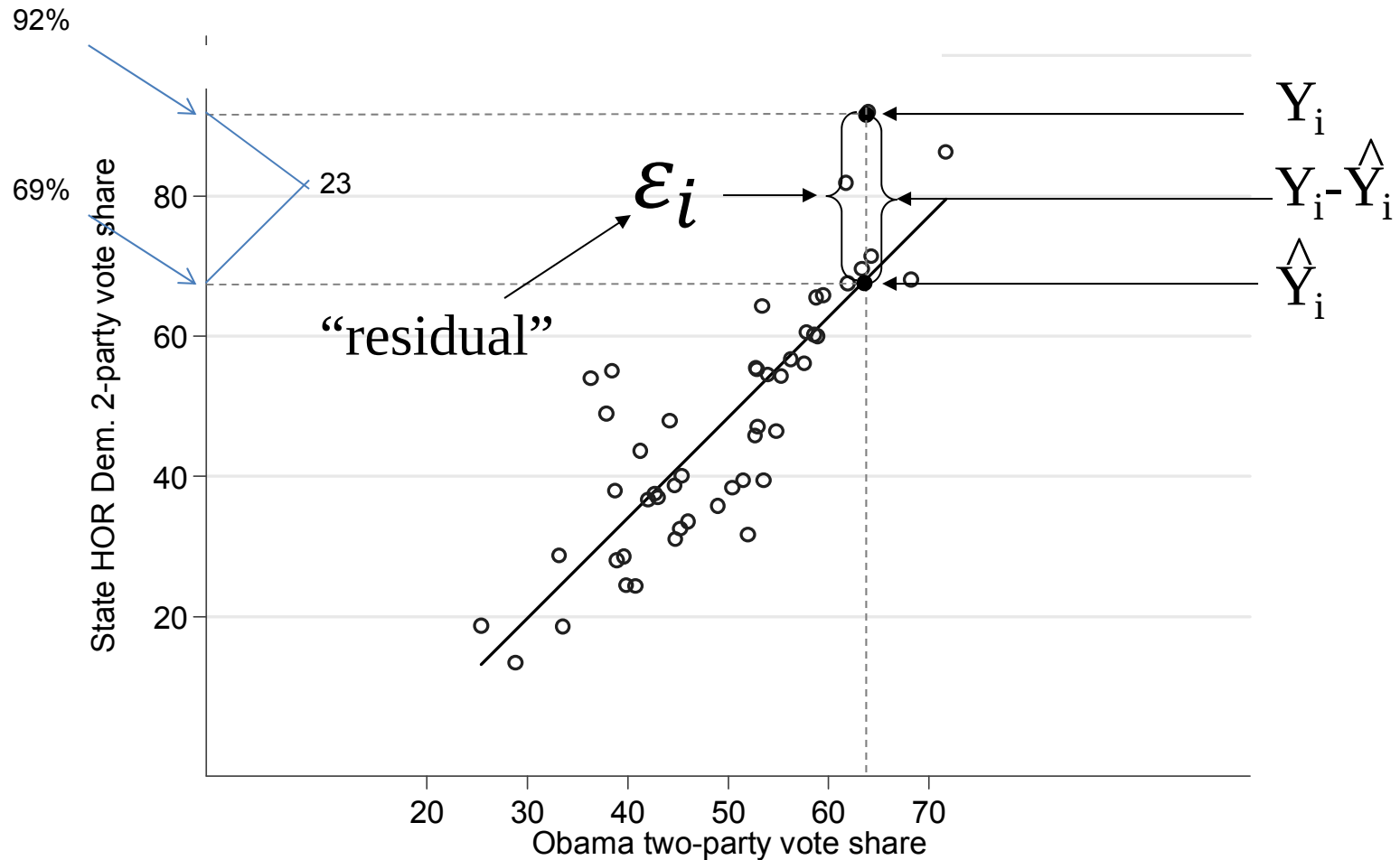


$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

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

How did we get that line?

3. Label the points



$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

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

What is ε_i ? (sometimes u_i)

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

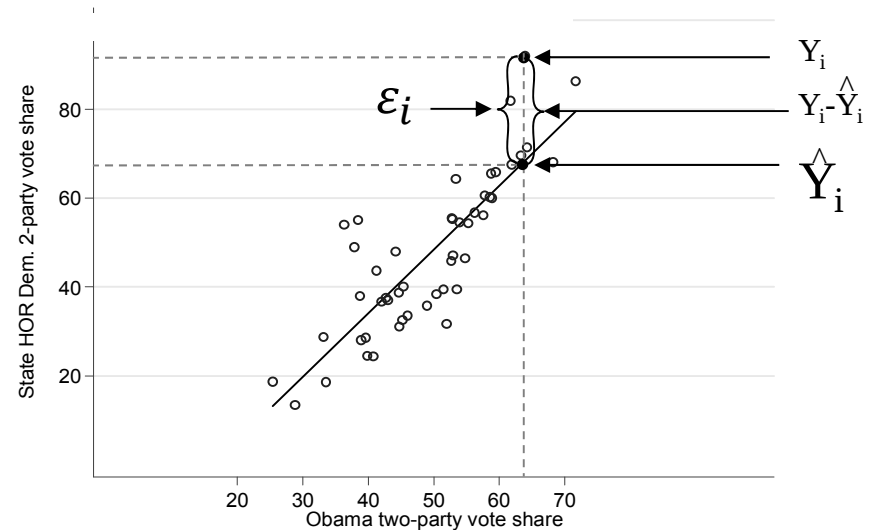
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The Method of Least Squares

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

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

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



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

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

Remember...

$$\frac{\sum_{i=1}^n (\bar{X} - X_i)^2}{n} = \text{Var}(X)$$

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

New idea...

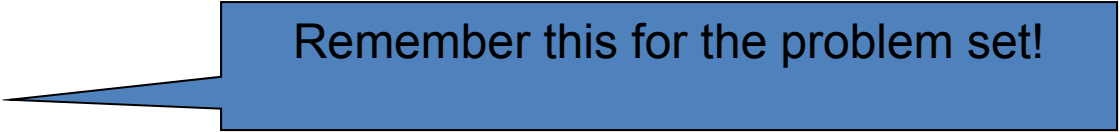
$$\frac{\sum_{i=1}^n (\bar{X} - X_i)(\bar{Y} - Y_i)}{n} = \text{Cov}(X, Y)$$

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

Therefore

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

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



Remember this for the problem set!

Regression commands in STATA

- `reg depvar expvars`
 - E.g., `reg y x`
 - E.g., `reg hpct ppct`
- `corr depvar expvar, cov`
- Making predictions from regression lines
 - `predict newvar`
 - `predict newvar, resid`
 - *newvar* will now equal ϵ_i
 - `predict newvar if e(sample), resid`
 - Confines the prediction to cases included in the regression

Drawing linear regression plot in Stata

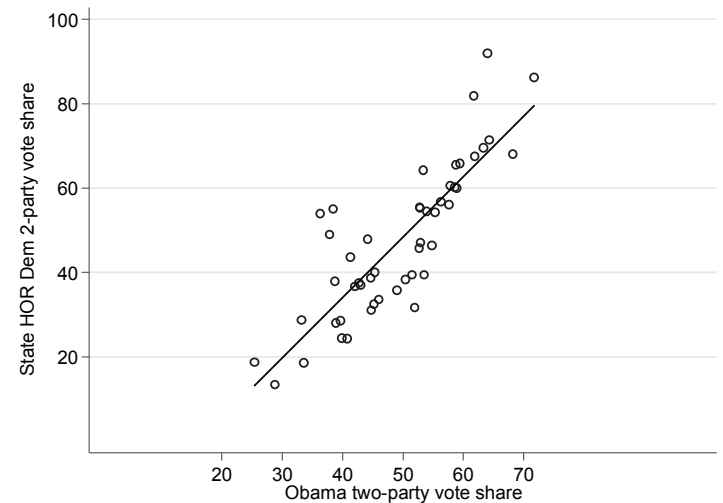
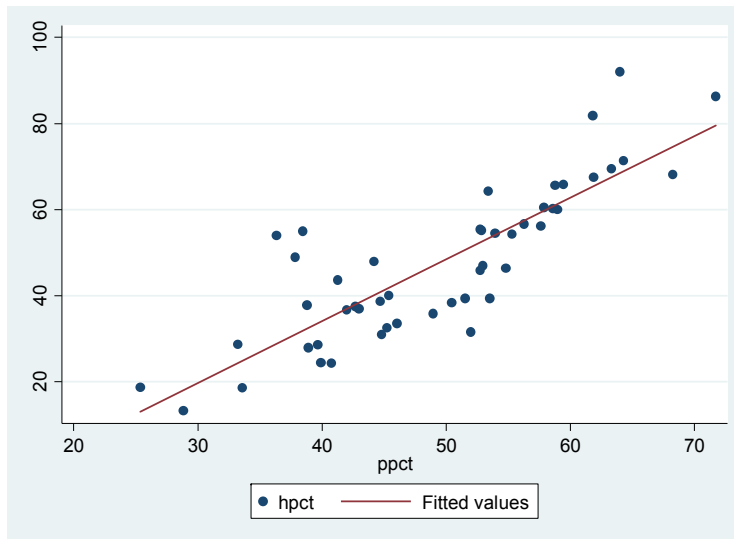
```
twoway (lfit hpct ppct) (scatter hpct ppct)
```

OR

```
lfit hpct ppct || scatter hpct ppct
```

To draw a presentation-quality graph requires a little more syntax

```
twoway (scatter hpct ppct) (lfit hpct ppct, lpattern(solid)), ///  
       ytitle(State HOR Dem 2-party vote share) yscale(range(0 100)) ///  
       xtitle(Obama two-party vote share) xscale(range(0 100)) ///  
       legend(off) scheme(tufte)
```



State legislature example

```
. reg hpct ppct
```

Source	SS	df	MS	Number of obs =	49
Model	10808.7878	1	10808.7878	F(1, 47) =	112.73
Residual	4506.44332	47	95.8817727	Prob > F =	0.0000
Total	15315.2312	48	319.067316	R-squared =	0.7058
				Adj R-squared =	0.6995
				Root MSE =	9.7919

hpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppct	1.433516	.135015	10.62	0.000	1.161901	1.705131
_cons	-23.25307	6.809819	-3.41	0.001	-36.95266	-9.553483

State legislature example

```
. reg hpct ppct
```

Source	SS	df	MS	Number of obs = 49		
Model	10808.7878	1	10808.7878	F(1, 47)	=	112.73
Residual	4506.44332	47	95.8817727	Prob > F	=	0.0000
Total	15315.2312	48	319.067316	R-squared	=	0.7058
				Adj R-squared	=	0.6995
				Root MSE	=	9.7919

hpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppct	1.433516	.135015	10.62	0.000	1.161901	1.705131
_cons	-23.25307	6.809819	-3.41	0.001	-36.95266	-9.553483

Interpretation: a one percentage point increase in Obama's vote leads to a 1.43 percentage point increase in the Democratic composition of the lower house in the state legislature

State legislature example

```
. reg hpct ppct
```

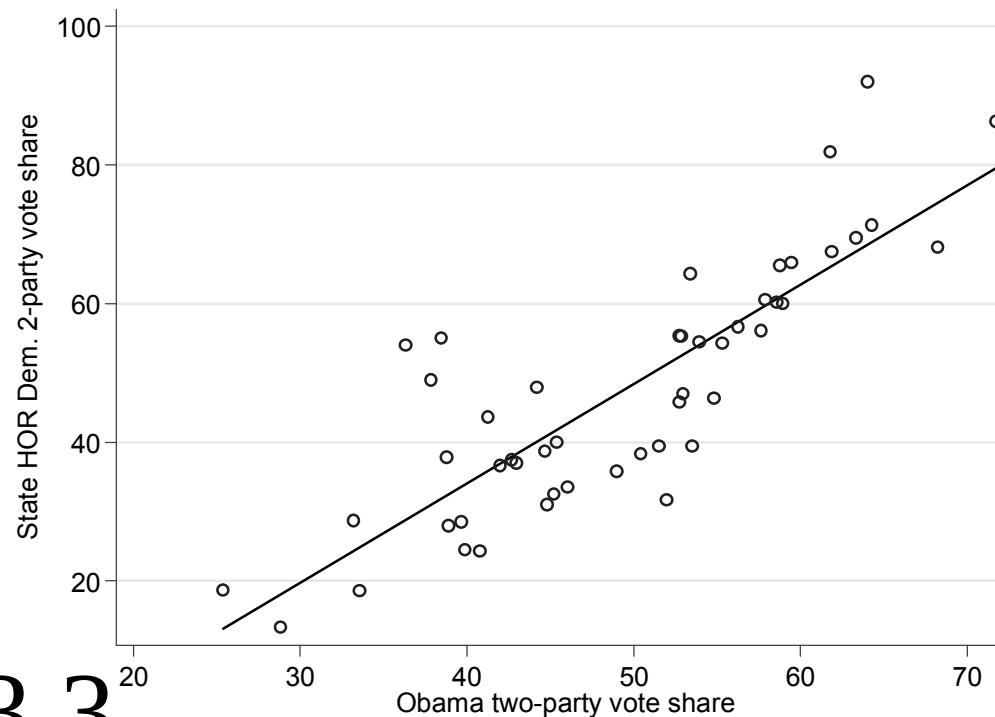
Source	SS	df	MS	Number of obs = 49		
Model	10808.7878	1	10808.7878	F(1, 47)	=	112.73
Residual	4506.44332	47	95.8817727	Prob > F	=	0.0000
Total	15315.2312	48	319.067316	R-squared	=	0.7058
				Adj R-squared	=	0.6995
				Root MSE	=	9.7919

hpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppct	1.433516	.135015	10.62	0.000	1.161901	1.705131
_cons	-23.25307	6.809819	-3.41	0.001	-36.95266	-9.553483

Always include interpretation in your presentations and papers

Interpretation: a one percentage point increase in Obama's vote leads to a 1.43 percentage point increase in the Democratic composition of the lower house in the state legislature

Linear Relationship between Support for Obama and Democrats in State Legislatures



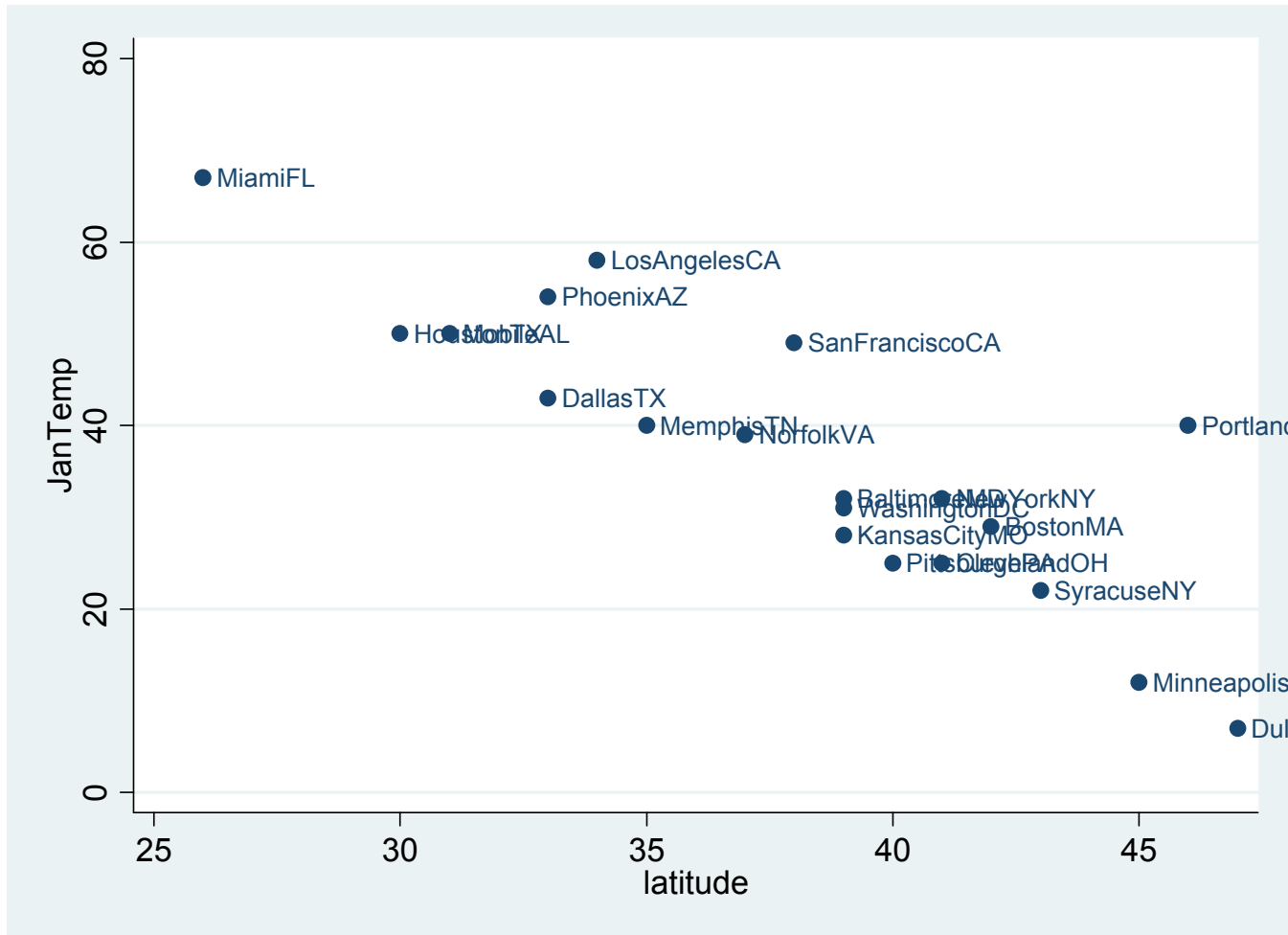
$$\beta_0 = -23.3$$

$$\beta_1 = 1.43$$

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

More regression examples

Temperature and Latitude



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

```
. reg jantemp latitude
```

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

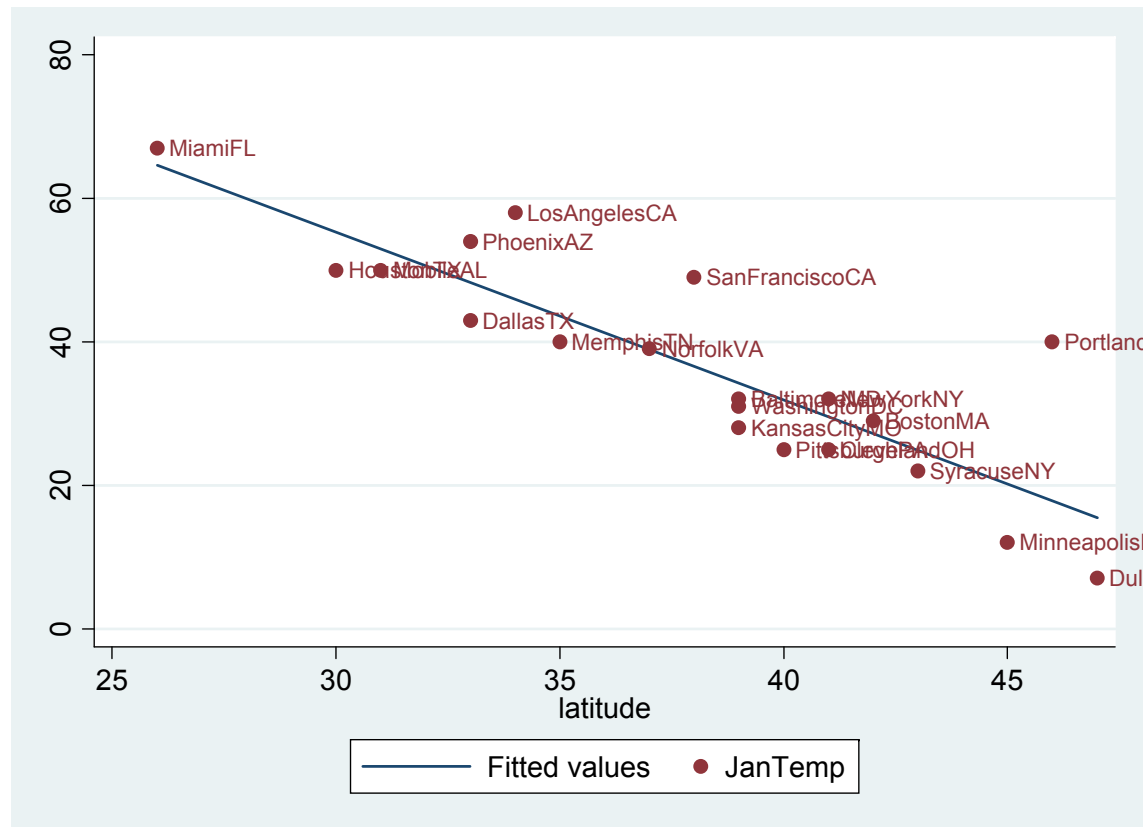
jantemp	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
latitude	-2.341428	.3333232	-7.02	0.000	-3.041714	-1.641142
_cons	125.5072	12.77915	9.82	0.000	98.65921	152.3552

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

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

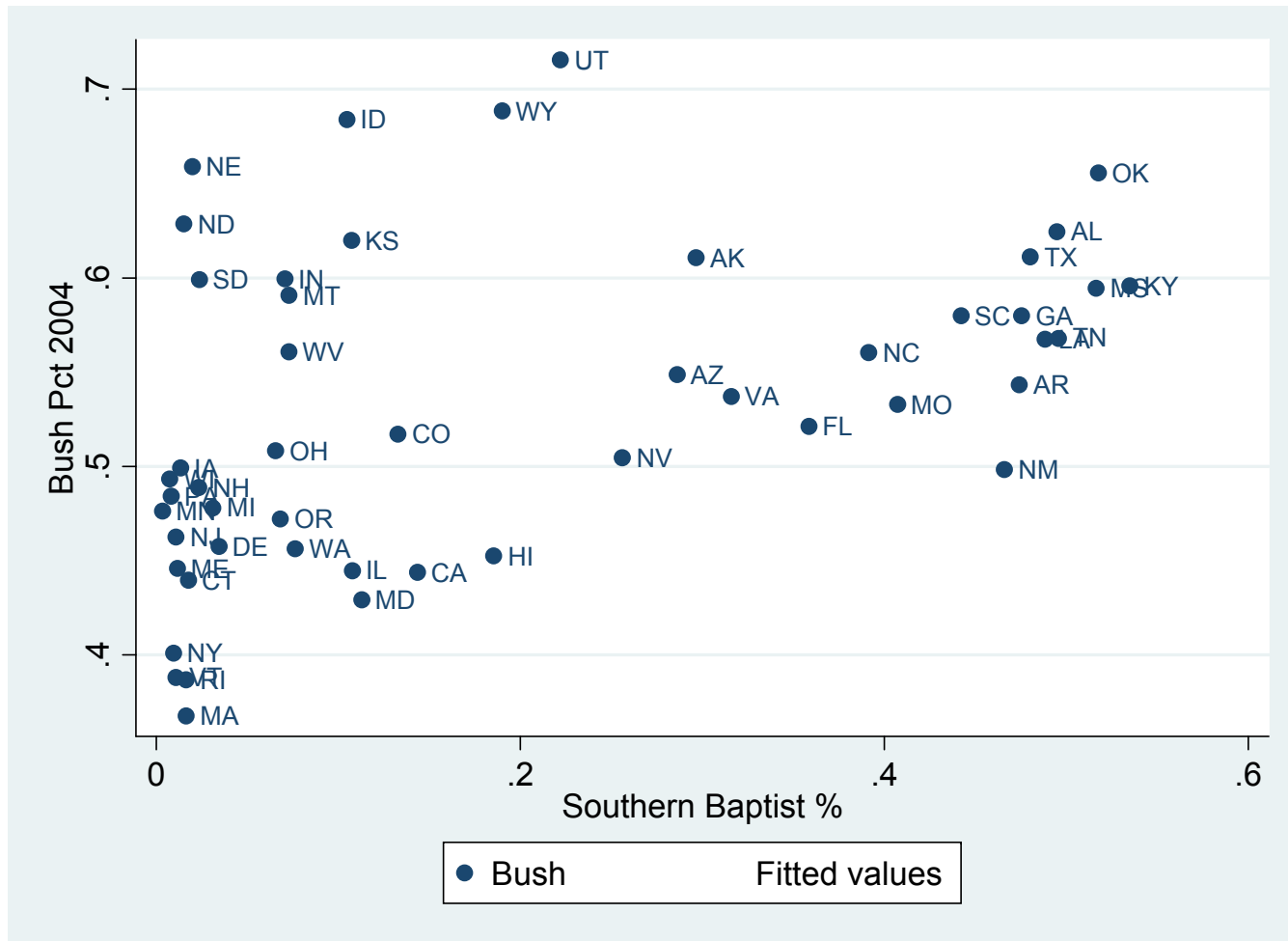
How to add a regression line:

Stata command: `lfit`



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

Bush vote and Southern Baptists



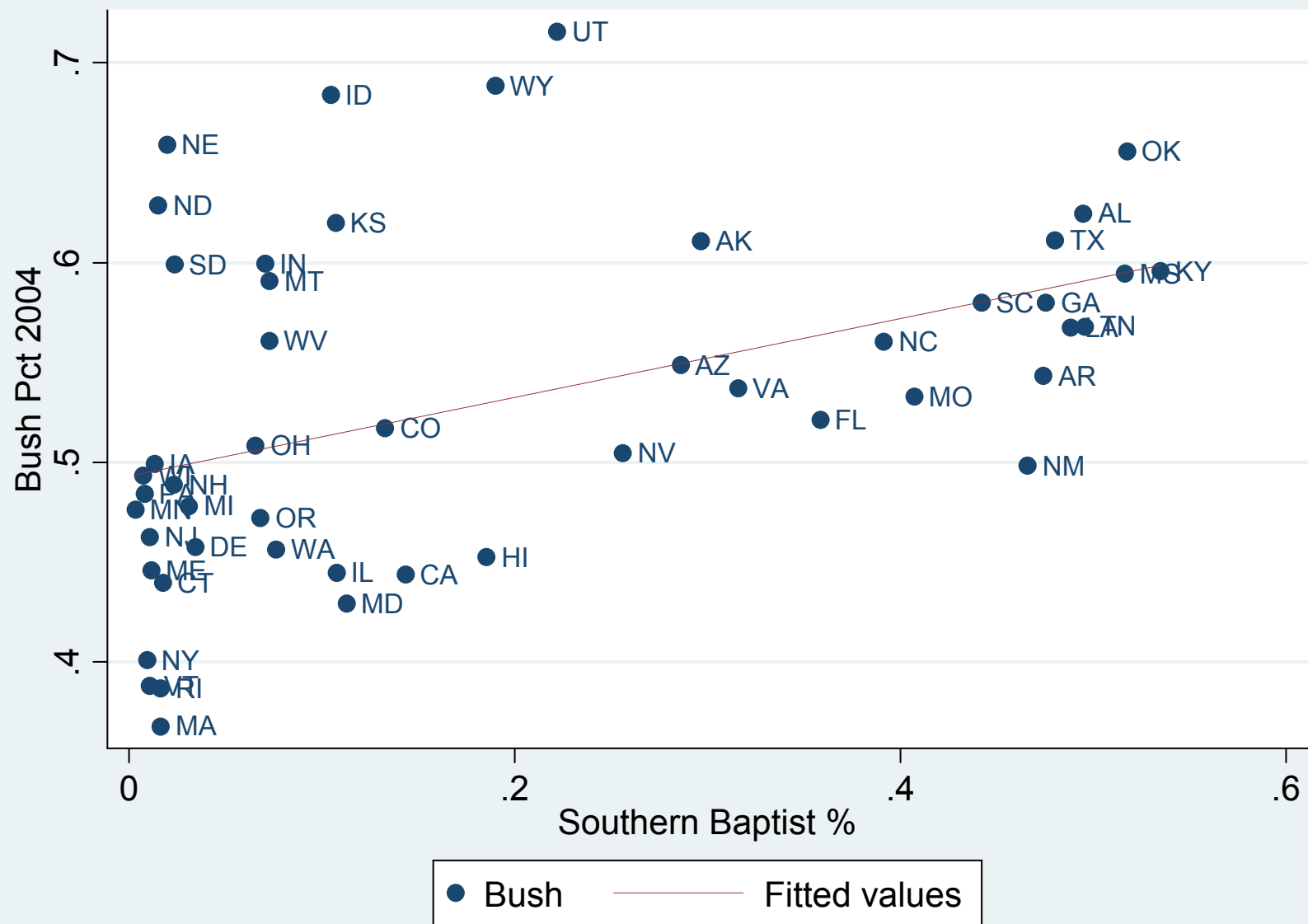
```
. reg bush sbc_mpct [aw=votes]
(sum of wgt is 1.2207e+08)
```

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



When the dependent variable is binary

- You can run a linear regression (though the graph will look weird)
- There are other techniques you can use, which are “more correct,” but difficult to interpret directly
 - Logit
 - Probit
- For now, use linear regression, not logit and probit

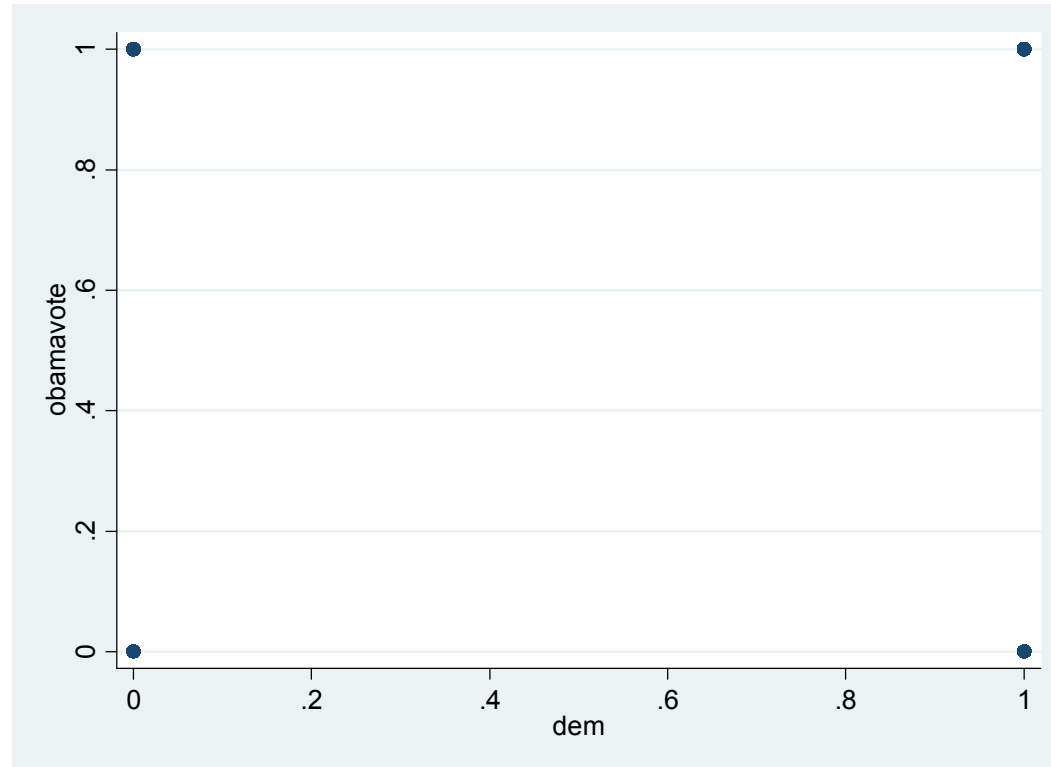
Example: Vote for Obama and Party ID

```
. table dem ,c(mean obamavote) row
```

dem	mean(obamav~e)
0	.0439405
1	.9514394
Total	.5424054

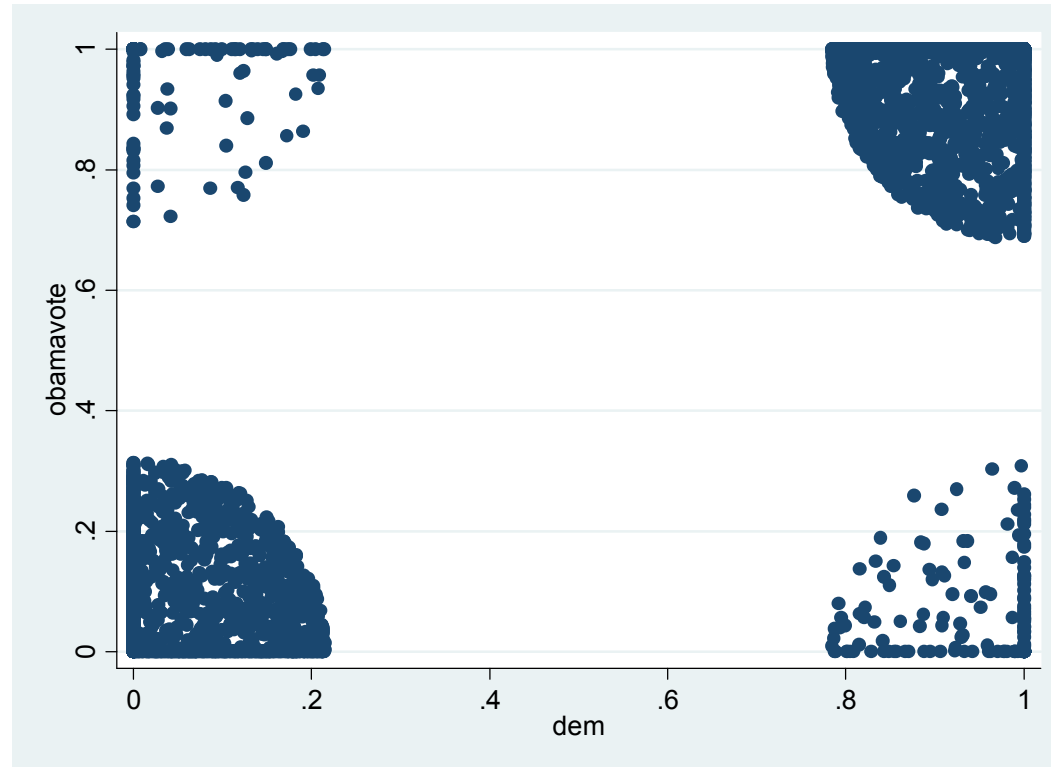
Source: 2012 SPAE, unweighted results

Why this is a case where you don't want to do a scatterplot



```
scatter obamavote dem
```

It's a little better with the “jitter”
command



```
scatter obamavote dem, jitter(50)
```

```
. reg obamavote dem
```

Source	SS	df	MS	
Model	1276.54962	1	1276.54962	
Residual	277.441757	6259	.04432685	
Total	1553.99138	6260	.248241434	

Number of obs	=	6261
F(1, 6259)	=	28798.56
Prob > F	=	0.0000
R-squared	=	0.8215
Adj R-squared	=	0.8214
Root MSE	=	.21054

obamavote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
dem	.9074989	.0053476	169.70	0.000	.8970157 .9179821
_cons	.0439405	.0039633	11.09	0.000	.0361711 .0517099

obamavote = 0.0439405 + 0.9074989*dem + e

```
. table dem ,c(mean obamavote) row
```

dem	mean(obamav~e)
0	.0439405
1	.9514394
Total	.5424054

```
. reg obamavote dem
```

Source	SS	df	MS	Number of obs = 6261		
Model	1276.54962	1	1276.54962	F(1, 6259) =28798.56		
Residual	277.441757	6259	.04432685	Prob > F = 0.0000		
Total	1553.99138	6260	.248241434	R-squared = 0.8215		
				Adj R-squared = 0.8214		
				Root MSE = .21054		

obamavote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dem	.9074989	.0053476	169.70	0.000	.8970157	.9179821
_cons	.0439405	.0039633	11.09	0.000	.0361711	.0517099

$$\text{obamavote} = 0.0439405 + 0.9074989 \cdot 0 = 0.0439405$$

```
. table dem ,c(mean obamavote) row
```

dem	mean(obamav~e)
0	.0439405
1	.9514394
Total	.5424054

When the respondent is a Republican, dem = 0

```
. reg obamavote dem
```

Source	SS	df	MS	Number of obs =	6261
Model	1276.54962	1	1276.54962	F(1, 6259) =	28798.56
Residual	277.441757	6259	.04432685	Prob > F =	0.0000
Total	1553.99138	6260	.248241434	R-squared =	0.8215
				Adj R-squared =	0.8214
				Root MSE =	.21054

obamavote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dem	.9074989	.0053476	169.70	0.000	.8970157	.9179821
_cons	.0439405	.0039633	11.09	0.000	.0361711	.0517099

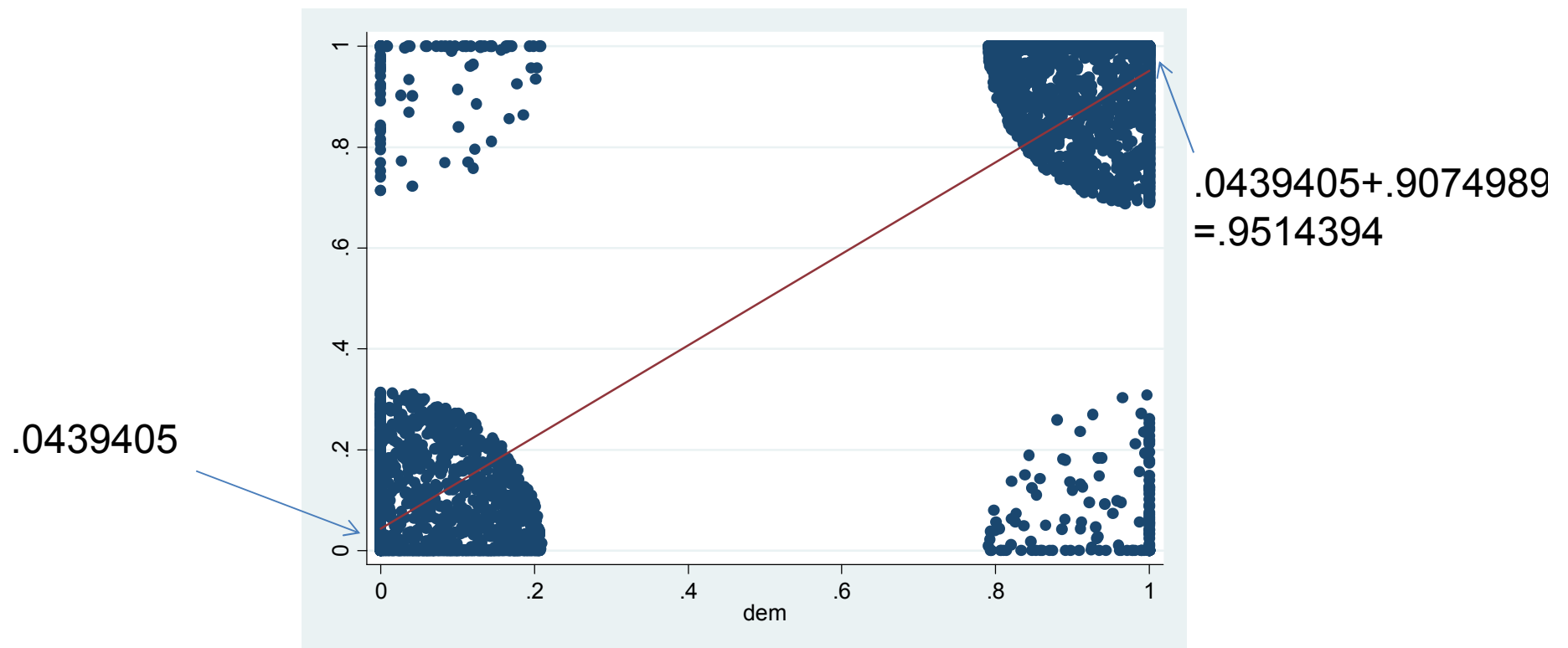
□
$$\text{obamavote} = 0.0439405 + 0.9074989 \cdot 1 = 0.9514394$$

```
. table dem , c(mean obamavote) row
```

dem	mean(obamav~e)
0	.0439405
1	.9514394
Total	.5424054

When the respondent is a Democrat, dem = 1

Geometric interpretation



obamavote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dem	.9074989	.0053476	169.70	0.000	.8970157	.9179821
_cons	.0439405	.0039633	11.09	0.000	.0361711	.0517099

```
. reg obamavote dem
```

Source	SS	df	MS	Number of obs = 6261		
Model	1276.54962	1	1276.54962	F(1, 6259) =28798.56		
Residual	277.441757	6259	.04432685	Prob > F = 0.0000		
Total	1553.99138	6260	.248241434	R-squared = 0.8215		
				Adj R-squared = 0.8214		
				Root MSE = .21054		

obamavote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dem	.9074989	.0053476	169.70	0.000	.8970157	.9179821
_cons	.0439405	.0039633	11.09	0.000	.0361711	.0517099

Coefficient interpretation:

- Republicans have a 4.4% probability of voting for Obama
- A one point increase in Democrat-ness increases the probability of voting for Obama by 90.7 percentage points

```
. reg obamavote dem
```

Source	SS	df	MS	Number of obs = 6261		
Model	1276.54962	1	1276.54962	F(1, 6259) =28798.56		
Residual	277.441757	6259	.04432685	Prob > F = 0.0000		
Total	1553.99138	6260	.248241434	R-squared = 0.8215		
				Adj R-squared = 0.8214		
				Root MSE = .21054		

obamavote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dem	.9074989	.0053476	169.70	0.000	.8970157	.9179821
_cons	.0439405	.0039633	11.09	0.000	.0361711	.0517099

Coefficient interpretation:

- Republicans have a 4.4% probability of voting for Obama
- A one point increase in Democrat-ness increases the probability of voting for Obama by 90.7 percentage points

```
. reg obamavote dem
```

Source	SS	df	MS	Number of obs = 6261		
Model	1276.54962	1	1276.54962	F(1, 6259) =28798.56		
Residual	277.441757	6259	.04432685	Prob > F = 0.0000		
Total	1553.99138	6260	.248241434	R-squared = 0.8215		
				Adj R-squared = 0.8214		
				Root MSE = .21054		

obamavote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dem	.9074989	.0053476	169.70	0.000	.8970157	.9179821
_cons	.0439405	.0039633	11.09	0.000	.0361711	.0517099

A better coefficient interpretation:

- Republicans have a 4.4% probability of voting for Obama
- Democrats are 90.7 percentage points more likely to vote for Obama (NOT 90.7 **percent** more likely...)

```
. reg obamavote dem
```

Source	SS	df	MS	Number of obs = 6261		
Model	1276.54962	1	1276.54962	F(1, 6259) =28798.56		
Residual	277.441757	6259	.04432685	Prob > F = 0.0000		
Total	1553.99138	6260	.248241434	R-squared = 0.8215		
				Adj R-squared = 0.8214		
				Root MSE = .21054		

obamavote	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dem	.9074989	.0053476	169.70	0.000	.8970157	.9179821
_cons	.0439405	.0039633	11.09	0.000	.0361711	.0517099

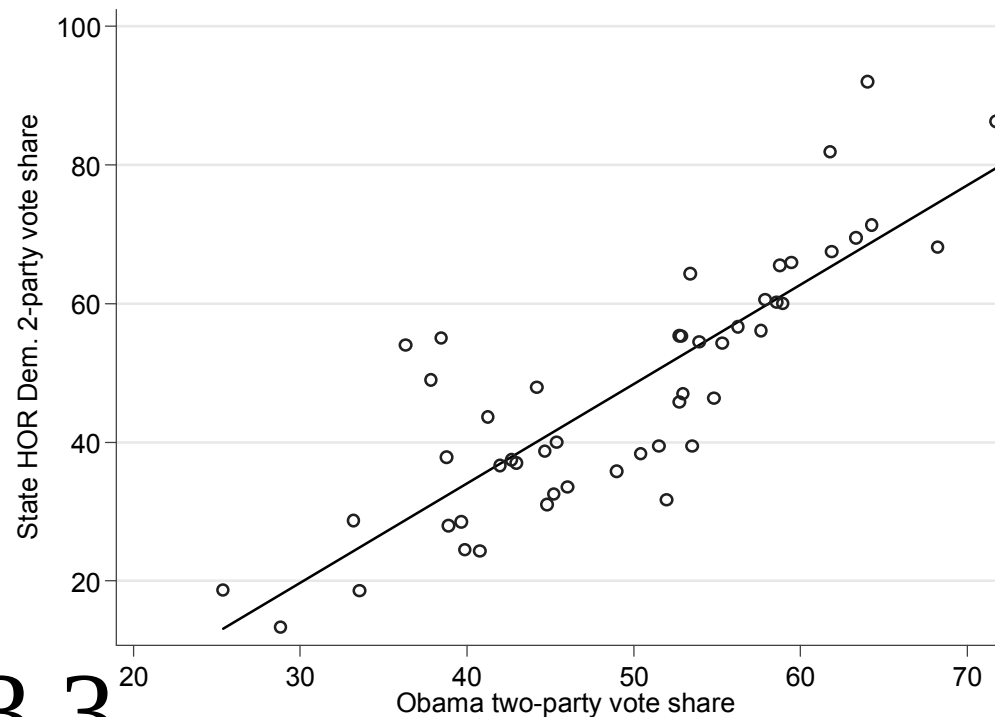
Another coefficient interpretation:

- Republicans have a 4.4% probability of voting for Obama
- As a voter moves from being a Republican to being a Democrats, she becomes 90.7 percentage points more likely to vote for Obama (NOT 90.7 **percent** more likely...)

Example when a calculated value is more informative than coefficients

- Return to the relationship between support for Obama and Democratic seats in state legislatures

Linear Relationship between Support for Obama and Democrats in State Legislatures

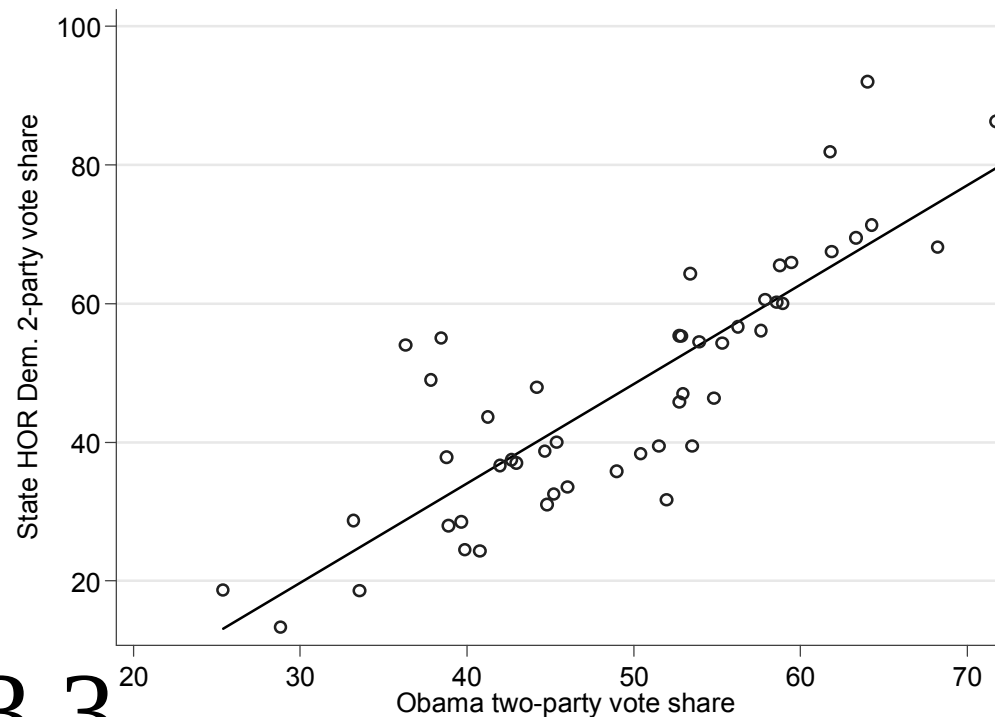


$$\beta_0 = -23.3$$

$$\beta_1 = 1.43$$

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

Perhaps we are especially interested in what happens when $x = 50\%$

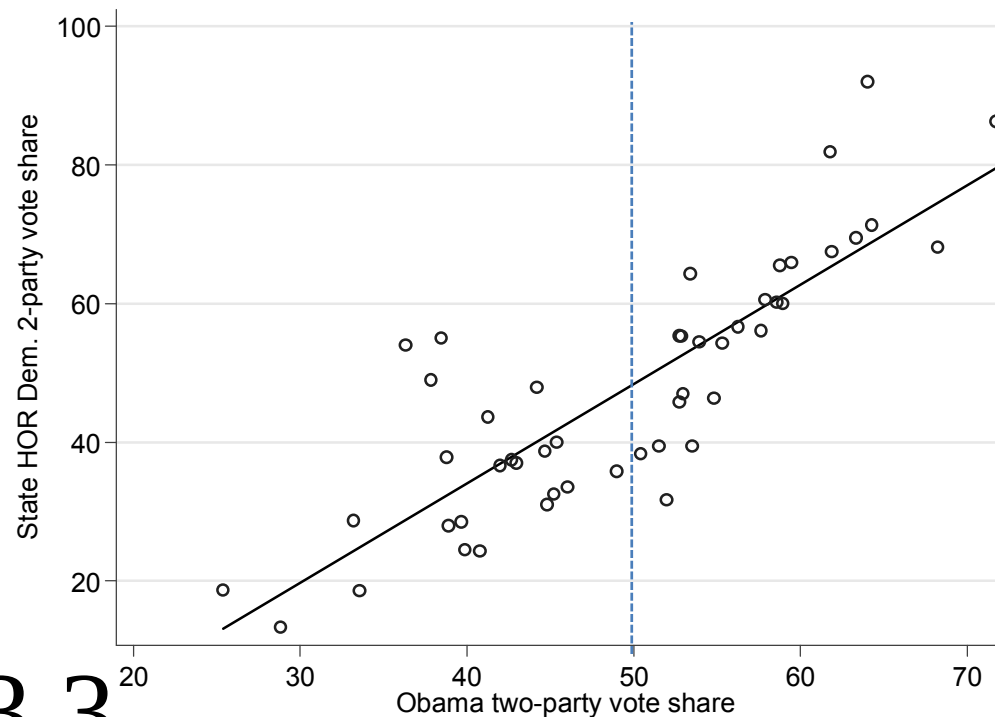


$$\beta_0 = -23.3$$

$$\beta_1 = 1.43$$

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

Perhaps we are especially interested in what happens when $x = 50\%$

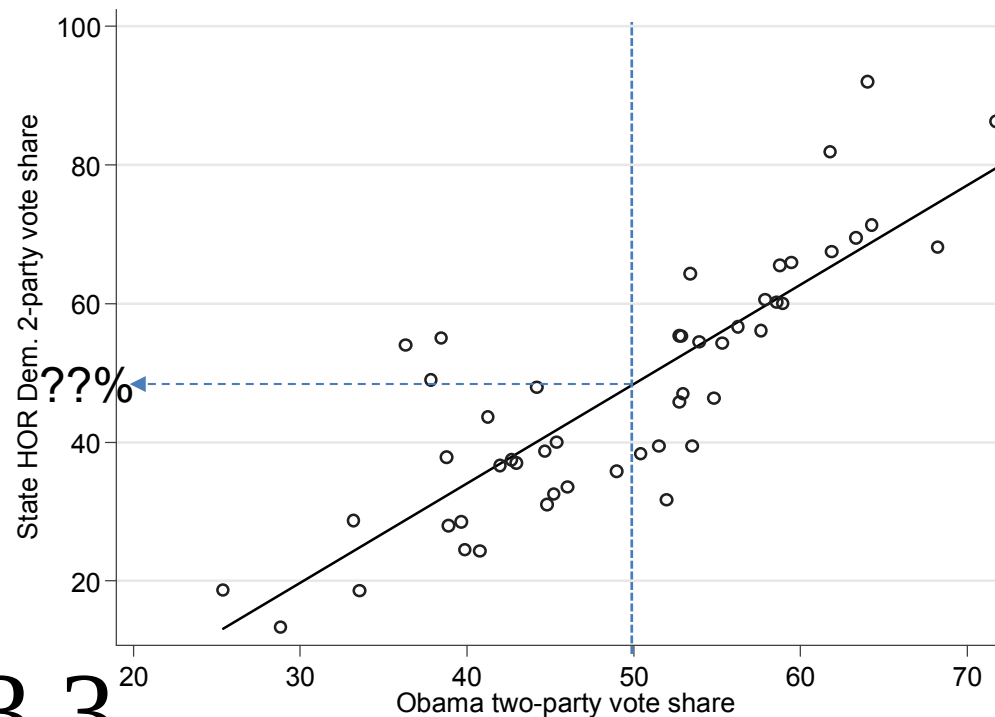


$$\beta_0 = -23.3$$

$$\beta_1 = 1.43$$

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

Perhaps we are especially interested in what happens when $x = 50\%$

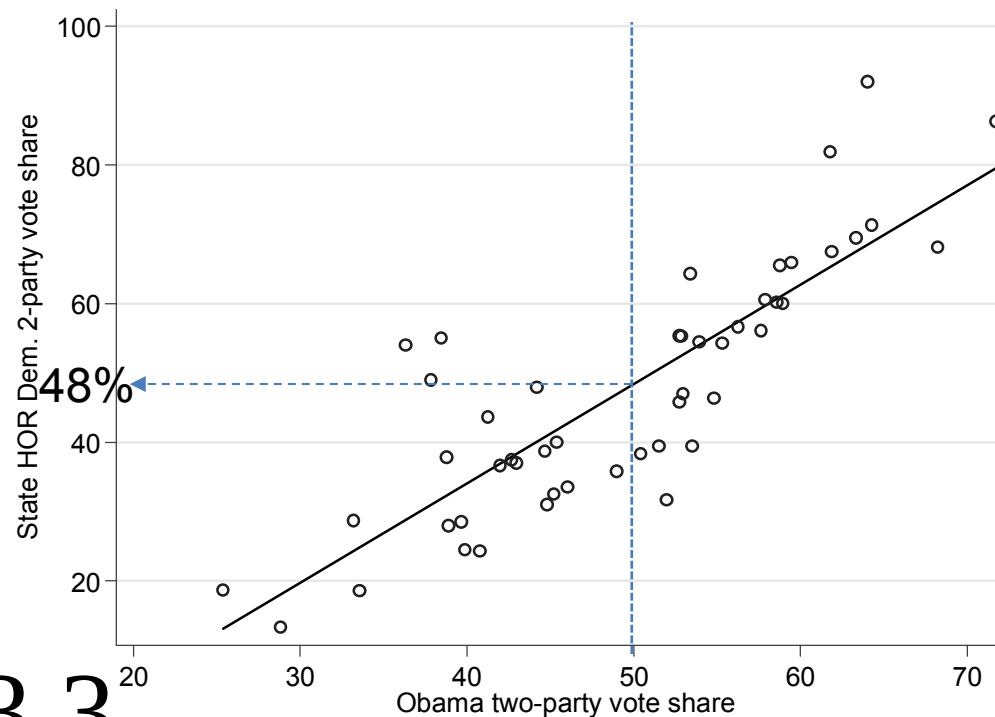


$$\beta_0 = -23.3$$

$$\beta_1 = 1.43$$

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

Perhaps we are especially interested in what happens when $x = 50\%$



$$\beta_0 = -23.3$$

$$\beta_1 = 1.43$$

$$-23.3 + 1.43 \cdot 50 = 48.2$$

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

Additional regression in bivariate relationship topics

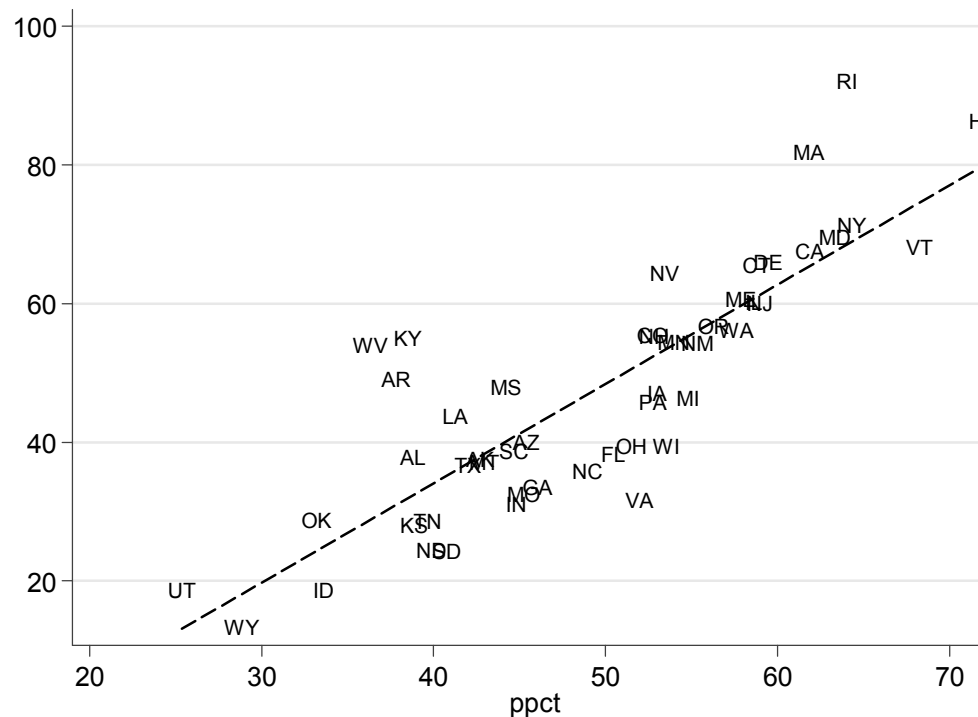
- Residuals
- Comparing coefficients
- Functional form
- Goodness of fit (R^2 and SER)
- Correlation
- Using the appropriate graph/table

Residuals

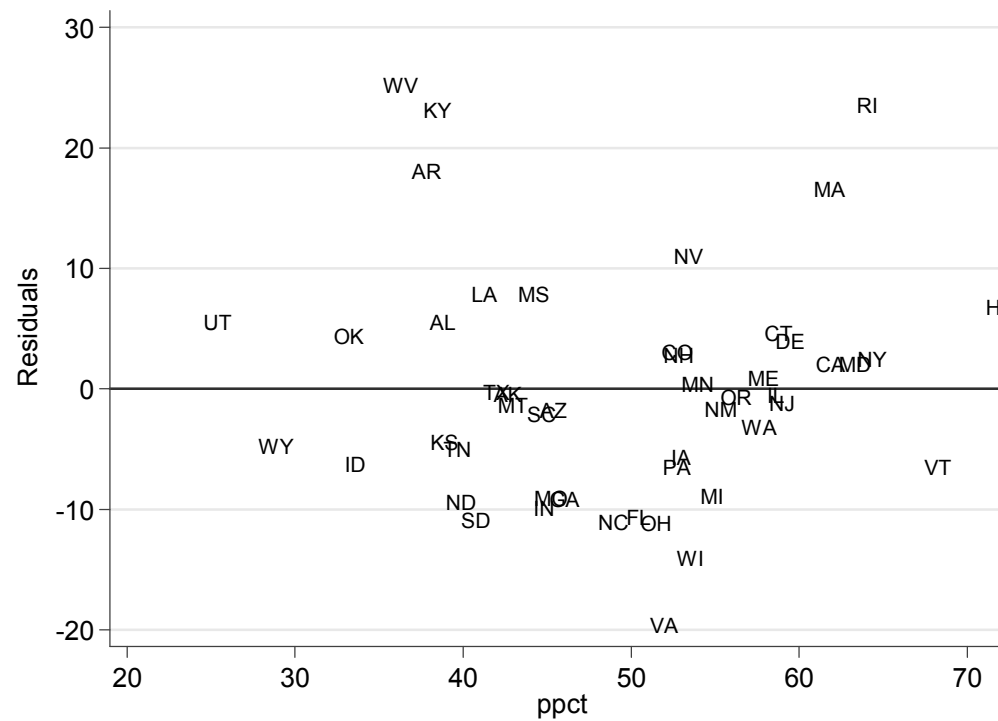
Residuals

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

Residuals are valuable for picking out outliers and interesting cases

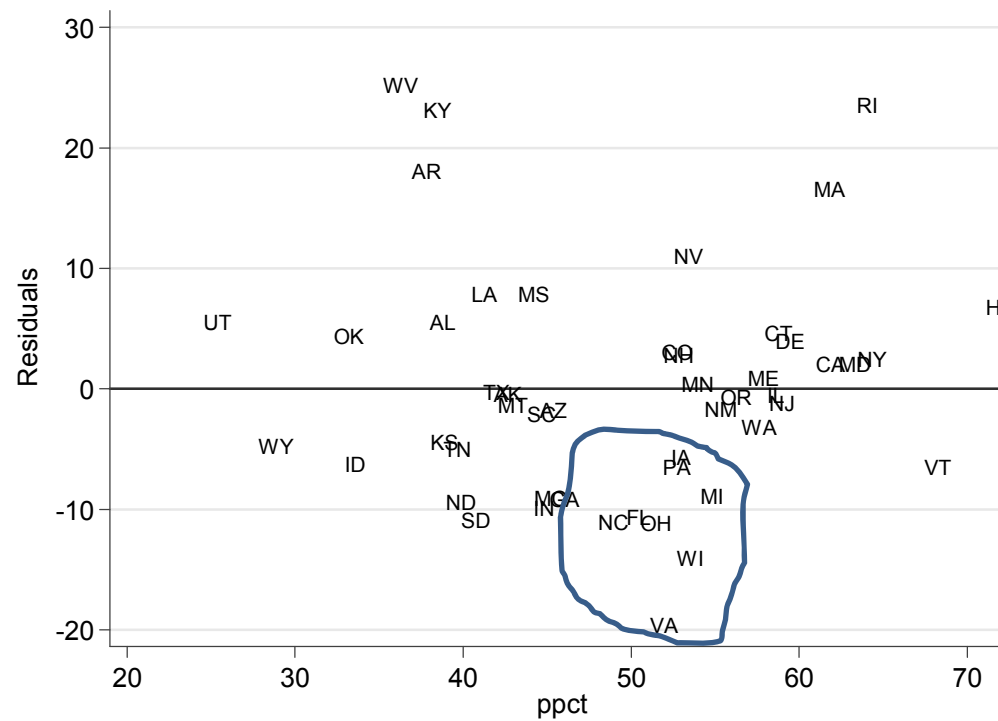


```
twoway (scatter hpct ppct, msymbol(none) mlabel(stabbr) mlabposition(0)) (lfit hpct  
ppct), legend(off) scheme(Tufte)
```



```
. predict ry,resid
(1 missing value generated)
```

```
. twoway (scatter ry ppct,msymbol(none) mlabel(stabbr) mlabposition(0)),legend(off)
scheme(Tufte) yline(0)
```



```
. predict ry,resid
(1 missing value generated)
```

```
. twoway (scatter ry ppct,msymbol(none) mlabel(stabbr) mlabposition(0)),legend(off)
scheme(Tufte) yline(0)
```

```
. gsort -ry
```

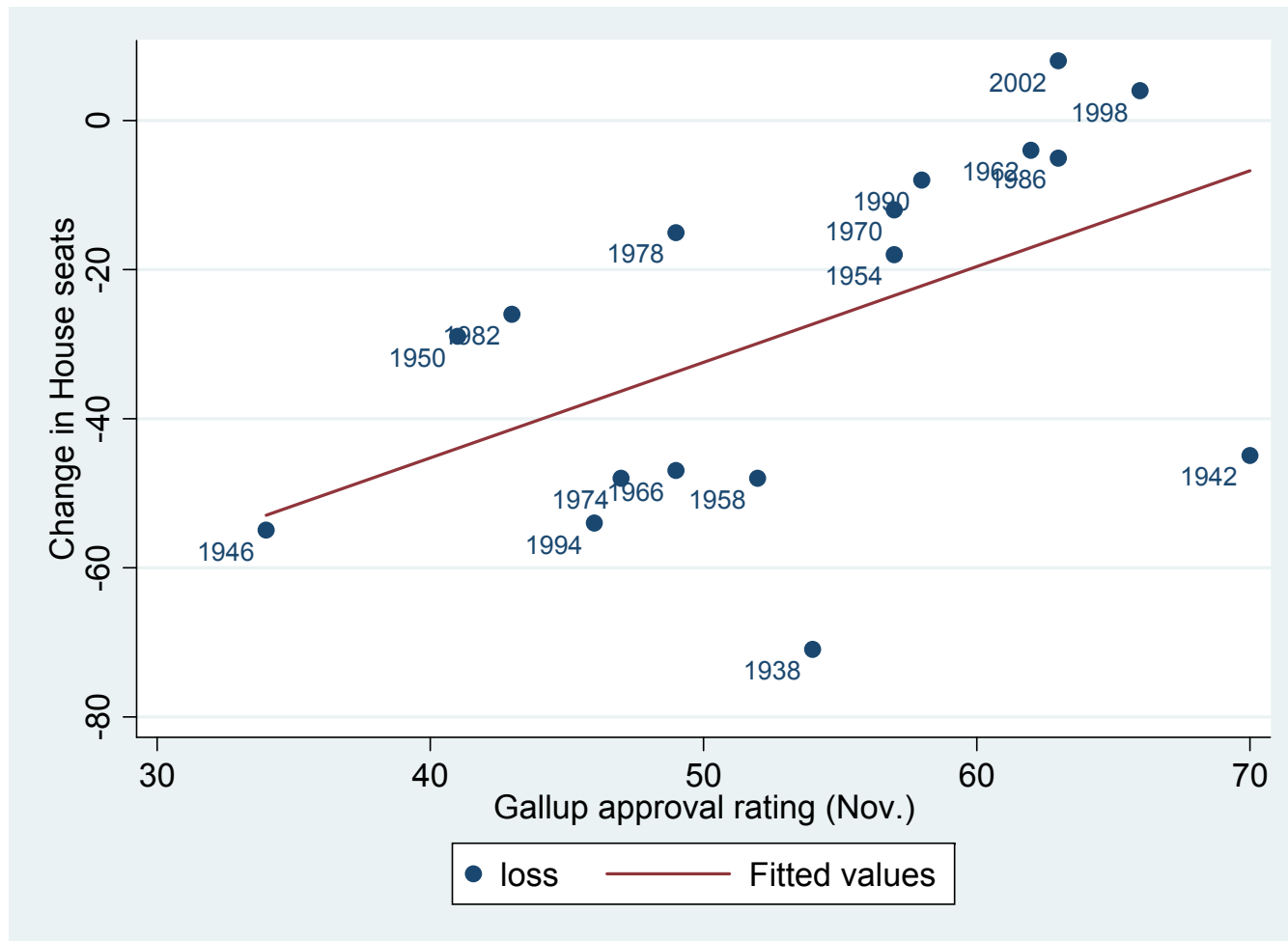
```
. list state ry
```

	state	ry
1.	West Virginia	25.17959
2.	Rhode Island	23.48404
3.	Kentucky	23.12402
4.	Arkansas	18.00081
5.	Massachusetts	16.55731
6.	Nevada	10.97821
7.	Mississippi	7.828268
8.	Louisiana	7.805305
9.	Hawaii	6.73896
10.	Utah	5.545974



46.	North Carolina	-11.10709
47.	Ohio	-11.19955
48.	Wisconsin	-14.06958
49.	Virginia	-19.61035
50.	Nebraska	.

Use residuals to diagnose potential problems



. reg loss gallup

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

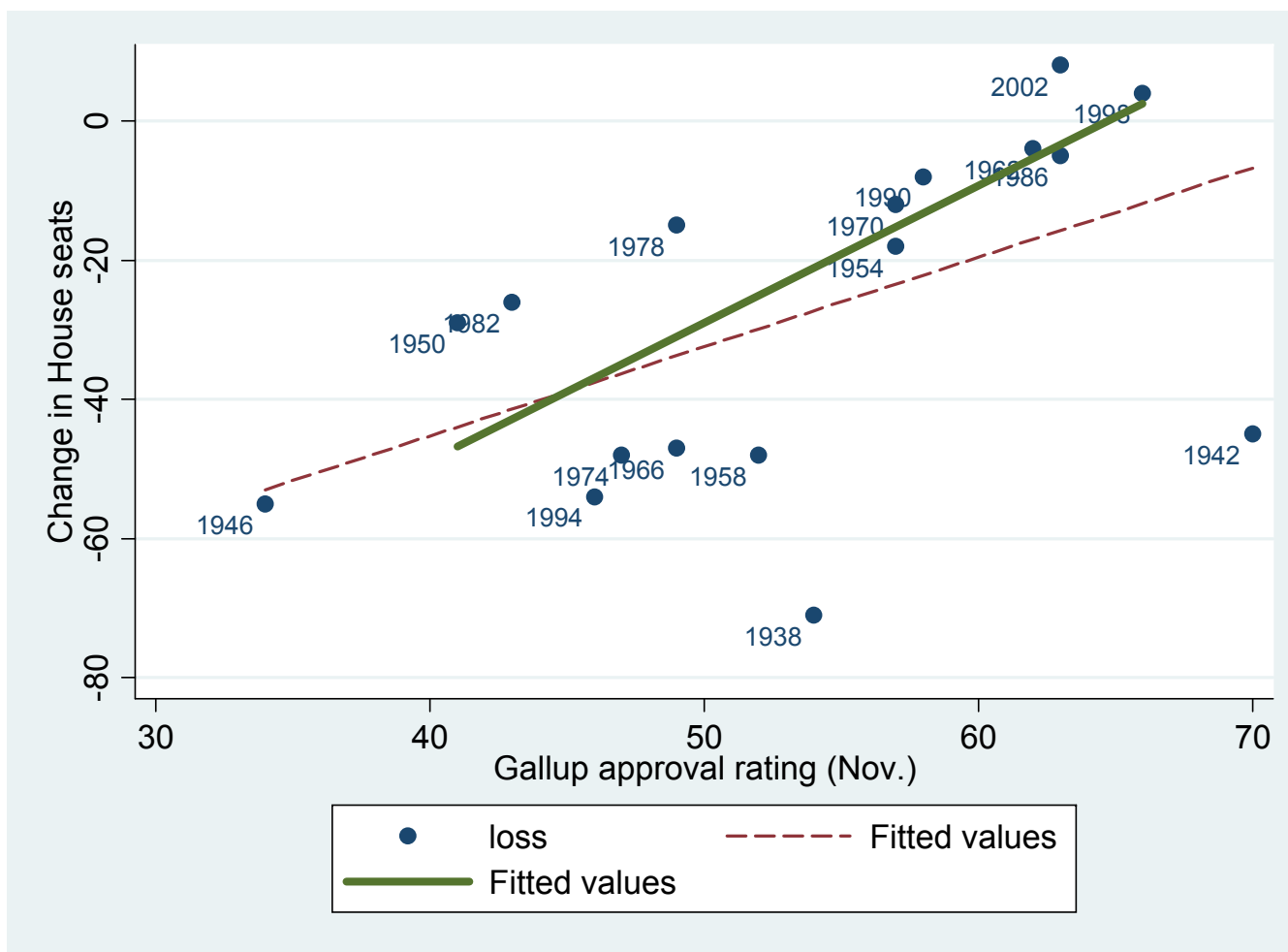
Seats	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gallup	1.283411	.53762	2.39	0.031	.1375011	2.429321
_cons	-96.59926	29.25347	-3.30	0.005	-158.9516	-34.24697

. reg loss gallup if year>1946

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



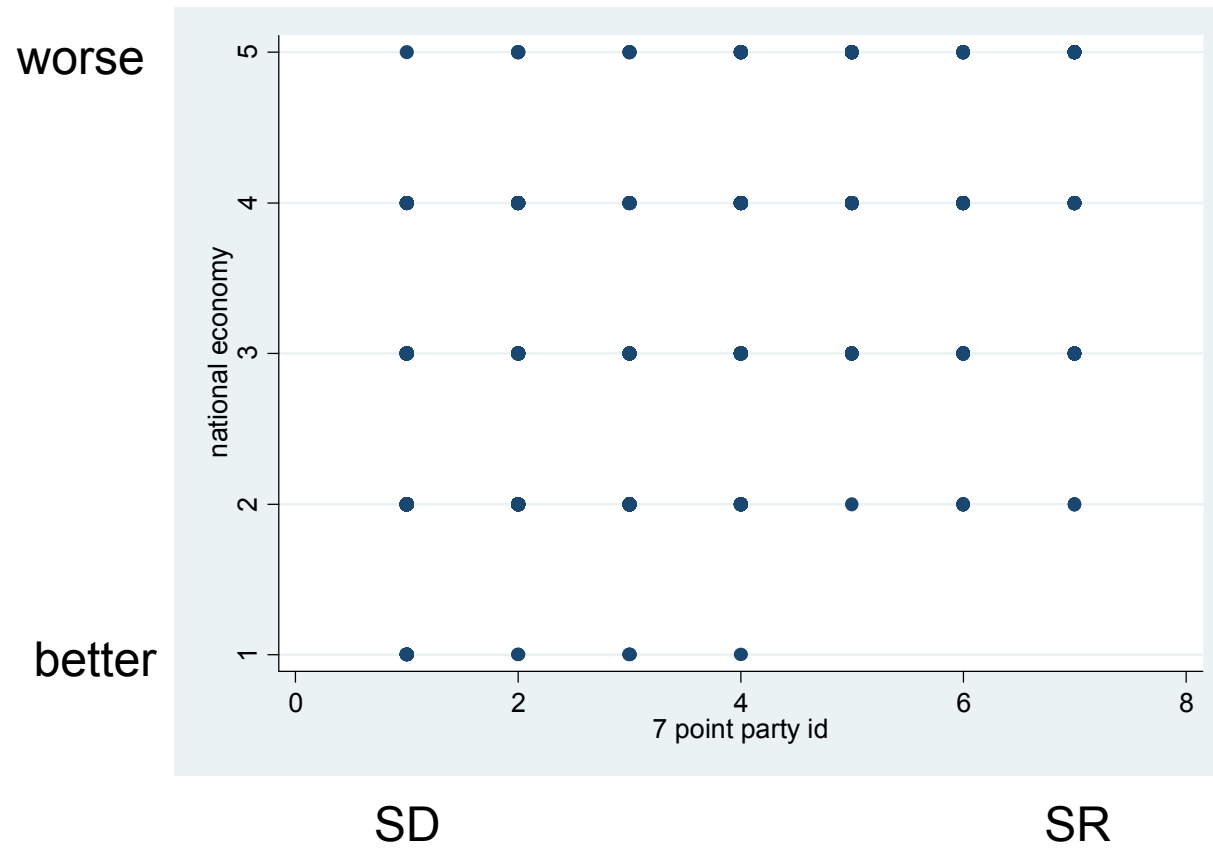
How to recode and interpret 0-1 variables

- Example: relationship between partisanship and saying the economy has gotten better in the past year

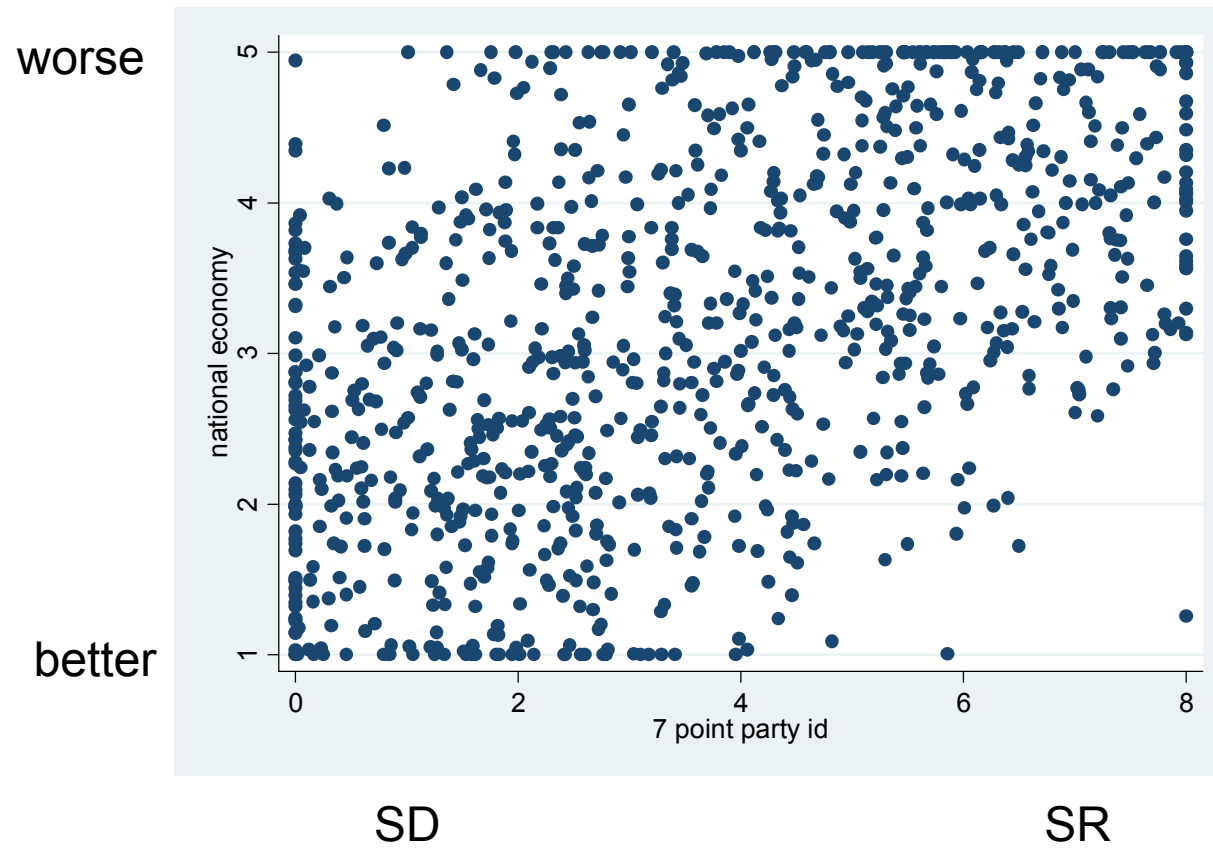
. tab pid7 cc302,row

7 point party id	national economy				Total
	Better gotten mu	gotten be	stayed ab	Worse gotten wo	
strong democrat	21 8.57	146 59.59	65 26.53	12 4.90	1 0.41
not very strong democ	2 1.87	45 42.06	29 27.10	26 24.30	5 4.67
lean democrat	4 3.54	65 57.52	29 25.66	11 9.73	4 3.54
independent	1 0.98	16 15.69	29 28.43	39 38.24	17 16.67
lean republican	0 0.00	1 0.94	27 25.47	39 36.79	39 36.79
not very strong repub	0 0.00	6 6.19	35 36.08	41 42.27	15 15.46
strong republican	0 0.00	3 1.81	21 12.65	80 48.19	62 37.35
Total	28 2.99	282 30.13	235 25.11	248 26.50	143 15.28

scatter cc302 pid7



```
scatter cc302 pid7, jitter(50)
```



How to recode variables to 0-1 scale

- Party ID example: pid7 in CCES 2012
- Usually varies from
 - 1 (strong Democrat)
 - to 7 (strong Republican)
 - 8 is “unsure,” and needs to be recoded to missing (“.”)
- Stata code?
 - `gen revpid7 = 8 - pid7`
 - `gen newpid7 = (revpid7-1)/(7-1)`

How to recode variables to 0-1 scale

- Party ID example: pid7 in CCES 2012
- Usually varies from
 - 1 (strong Democrat)
 - to 7 (strong Republican)
 - 8 is “unsure,” and needs to be recoded to missing (“.”)

- Stata code?

- gen revpid7 = 8 - pid7

- gen newpid7 = (revpid7-1)/(7-1)

min(revpid7)

Max(revpid7)

Regression interpretation with 0-1 scale

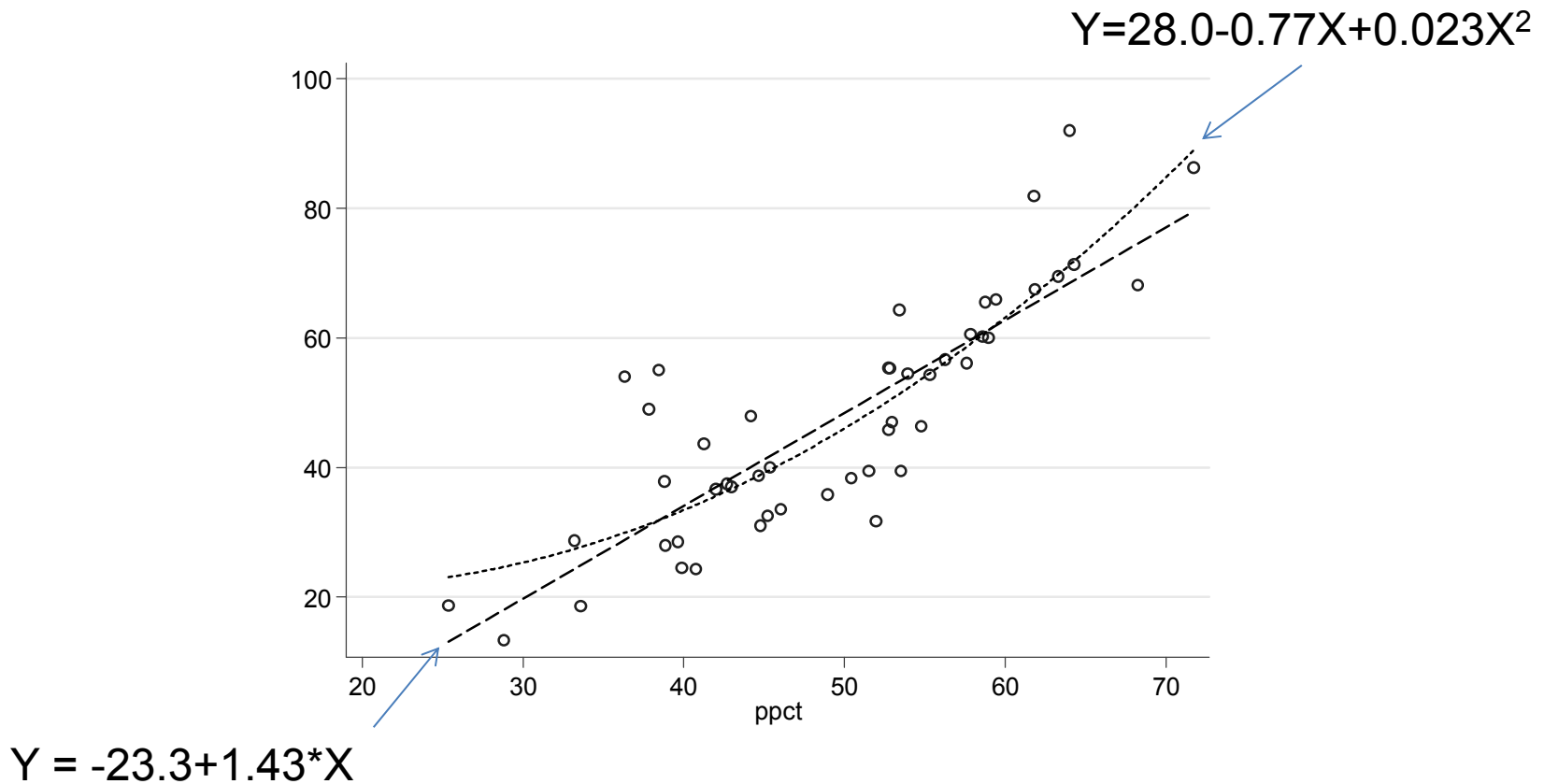
- Continue with pid7 example
 - regress econworse newpid7 (both recoded to 0-1 scales)*
 - pid7 coefficient: intercept = .18, slope = .48
 - Interpretation?
 - The average strong Democrat rates the **weakness** of the national economy as a .18 on a one-point national economic weakness scale
 - Shifting from being a strong Democrat to a strong Republican corresponds with a .48 increase in evaluations of the national economy as being weak (on the one-point national economic weakness scale)
- *econworse originally coded so that 1 = gotten much better, 5 = gotten much worse

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 Dems. in State Legislatures & Obama Vote



```
scatter hpct ppct || lfit hpct ppct || qfit hpct  
ppct,scheme(Tufte) legend(off)
```

```
. gen ppct2=ppct^2
```

```
. reg hpct ppct ppct2
```

Source	SS	df	MS	Number of obs = 49		
Model	11212.5662	2	5606.28309	F(2, 46) = 62.86		
Residual	4102.66499	46	89.1883693	Prob > F = 0.0000		
Total	15315.2312	48	319.067316	R-squared = 0.7321		
				Adj R-squared = 0.7205		
				Root MSE = 9.444		

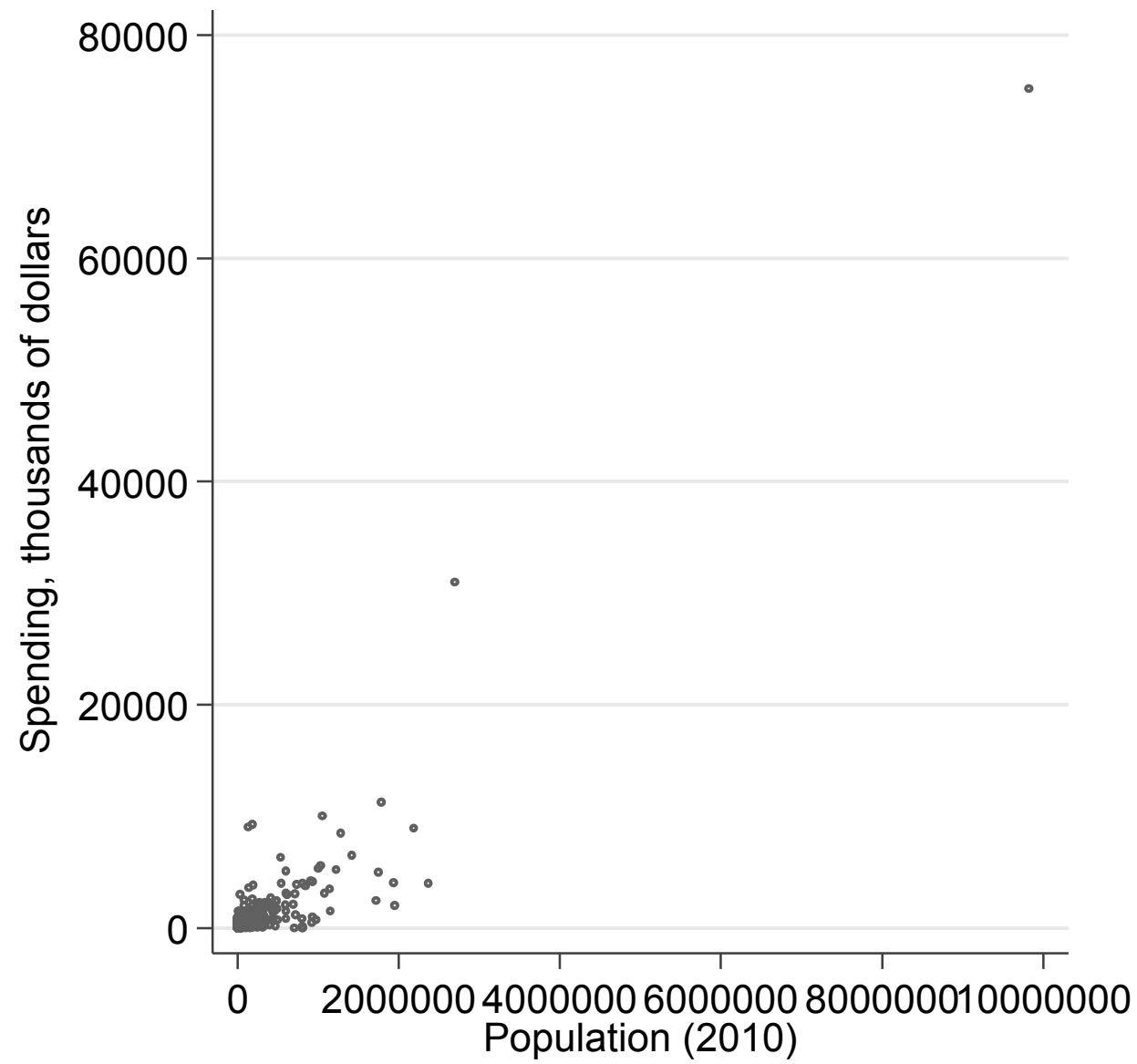
hpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppct	-.7652575	1.041559	-0.73	0.466	-2.861808	1.331293
ppct2	.0225134	.0105809	2.13	0.039	.0012151	.0438116
_cons	28.00989	24.97192	1.12	0.268	-22.25598	78.27576

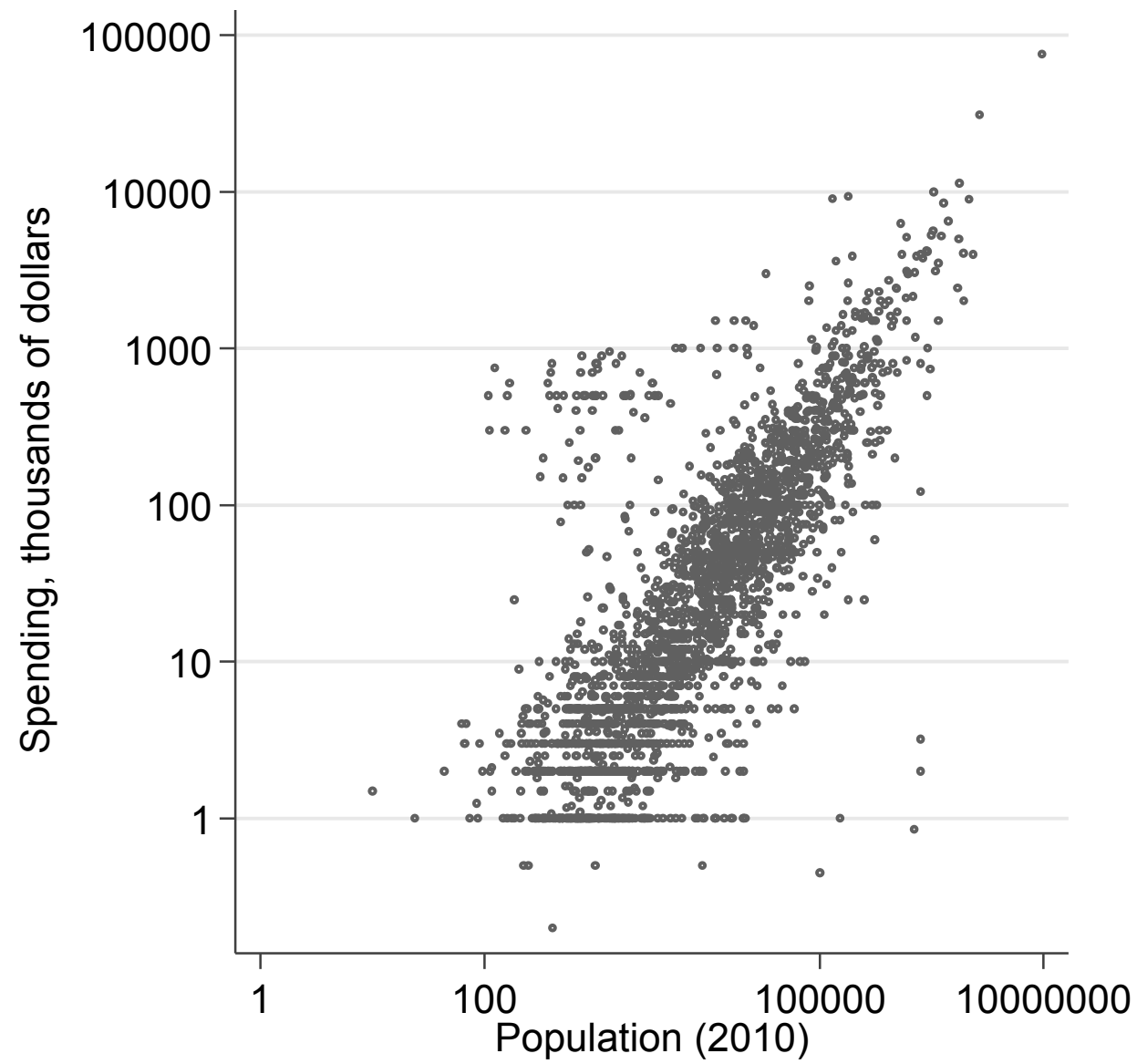
Log transformations (see Tufte, ch. 3)

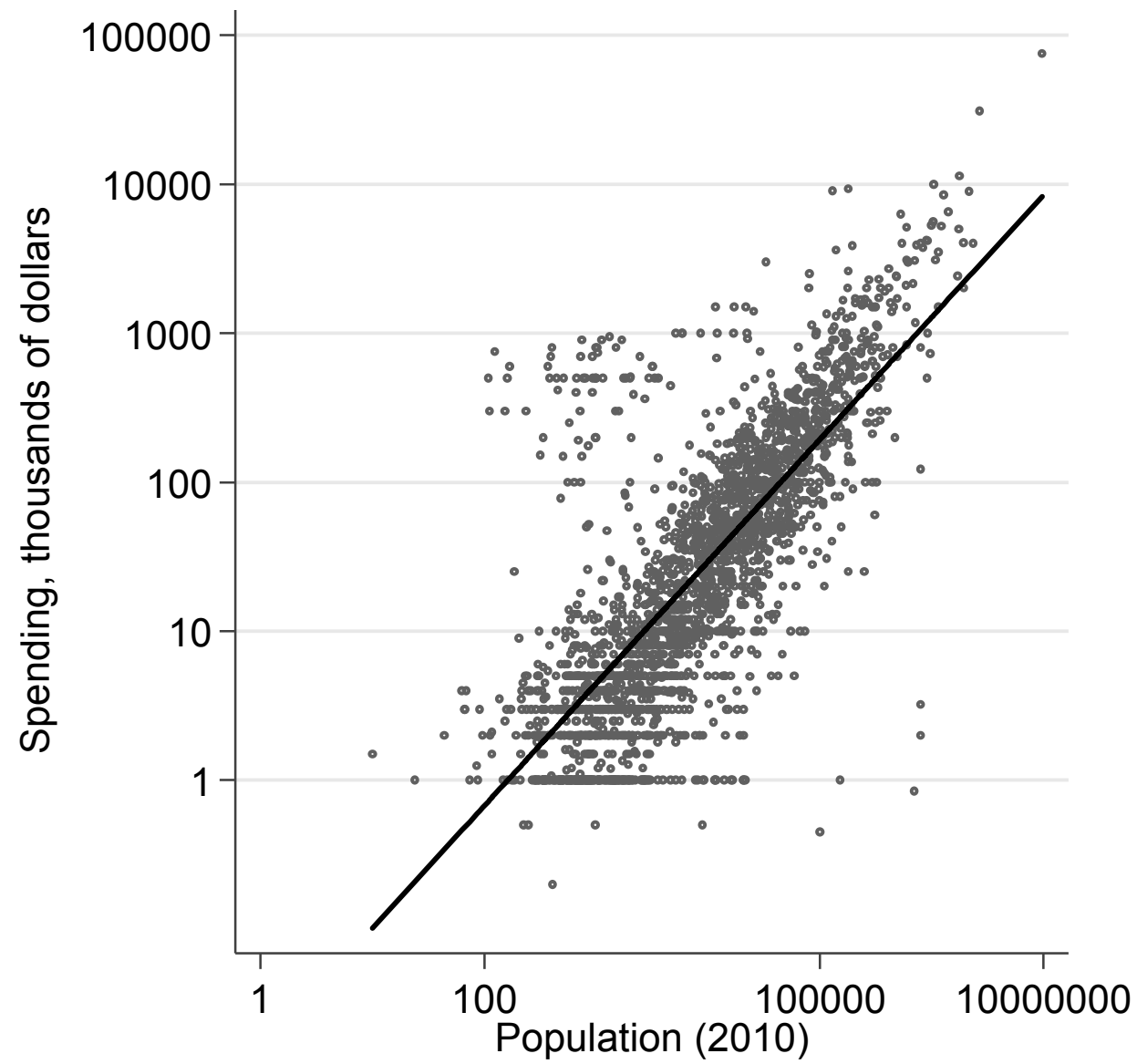
$Y = a + bX + e$	$b = dY/dX$, or $b = \text{the unit change in } Y \text{ given a unit change in } X$	Typical case
$Y = a + b \ln X + e$	$b = dY/(dX/X)$, or $b = \text{the unit change in } Y \text{ given a \% change in } X$	Log explanatory variable
$\ln Y = a + bX + e$	$b = (dY/Y)/dX$, or $b = \text{the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

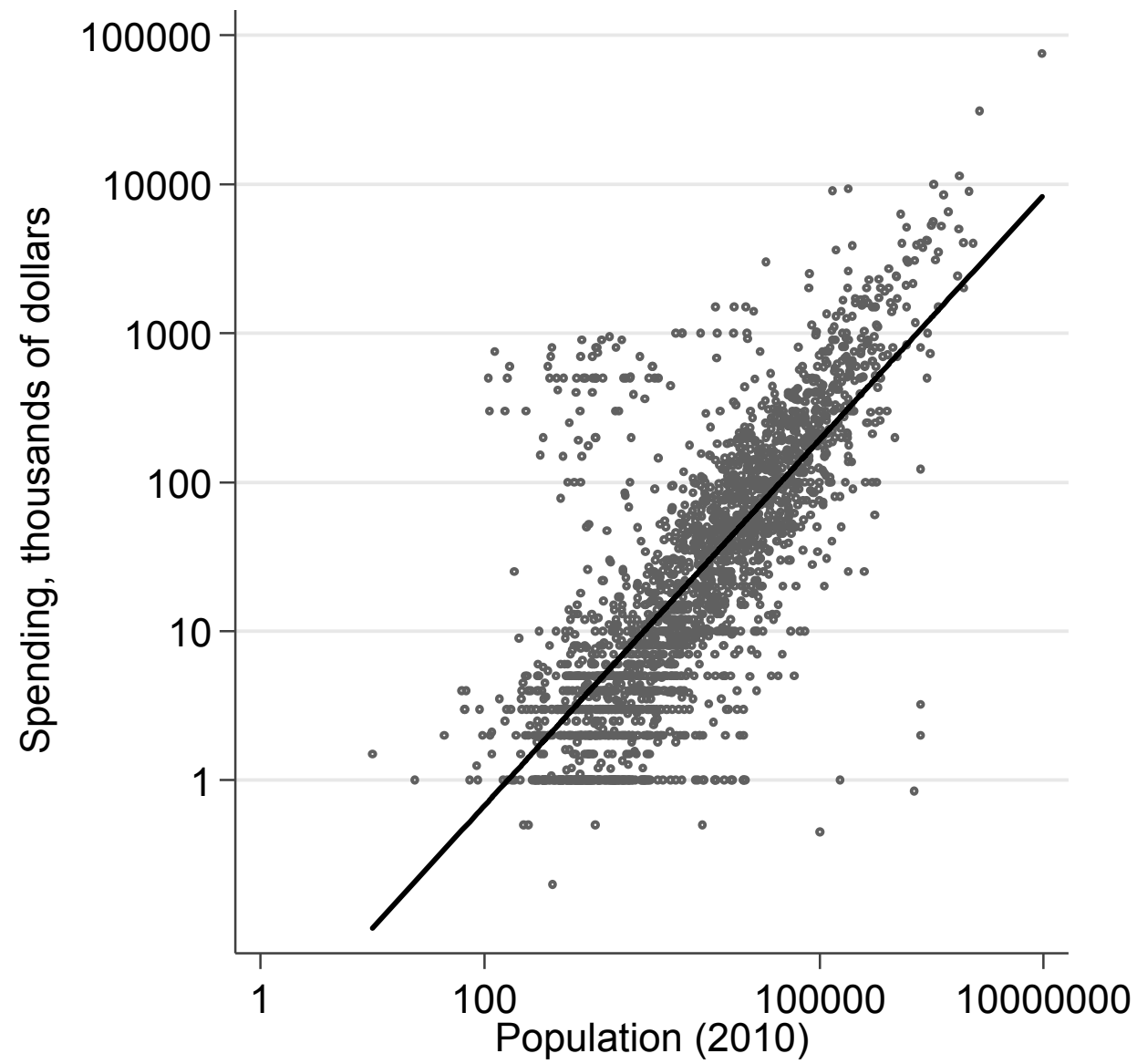
Example of log transform that's tied to theory

- Local budgets to conduct elections --- are there returns to scale?

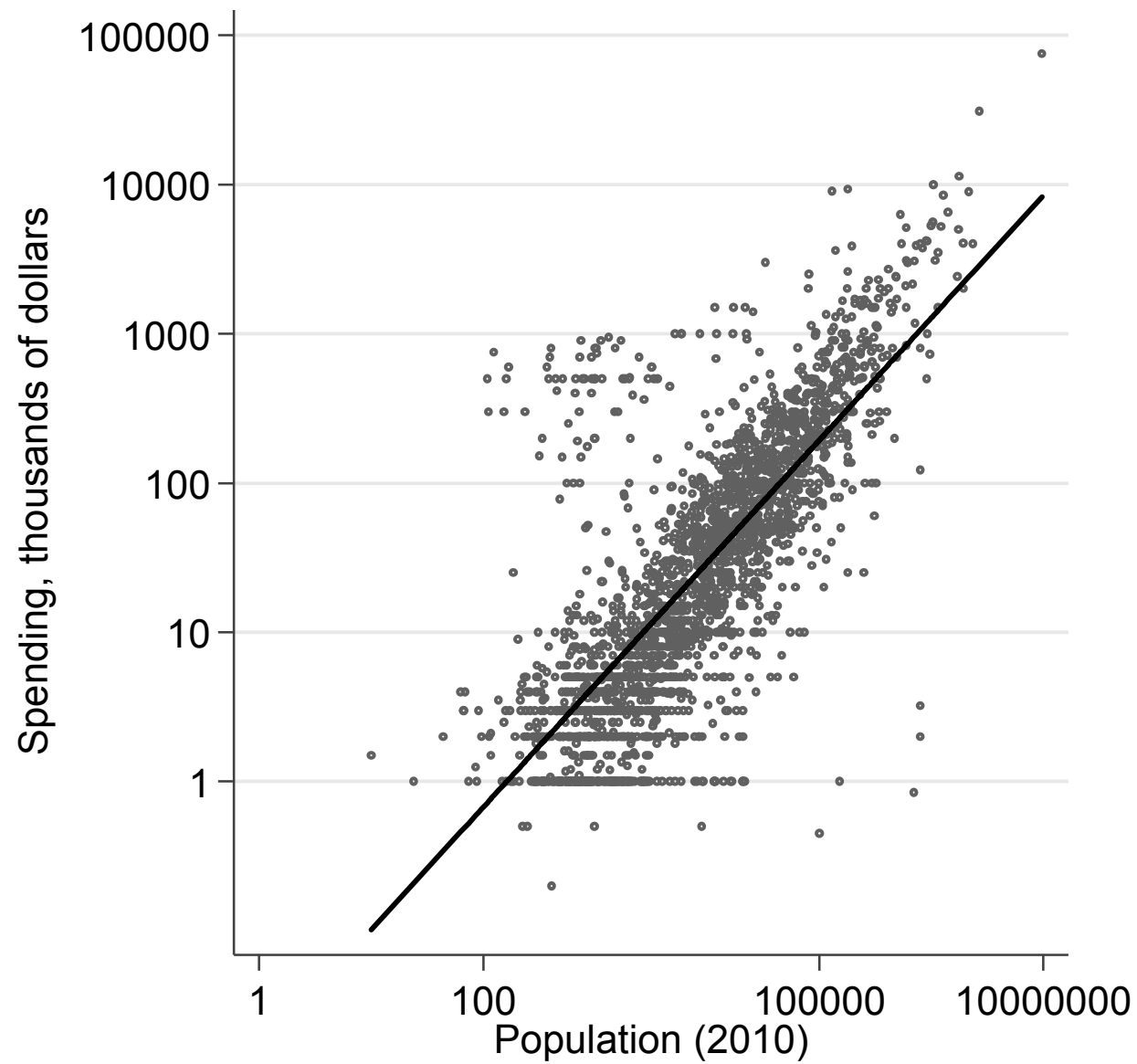






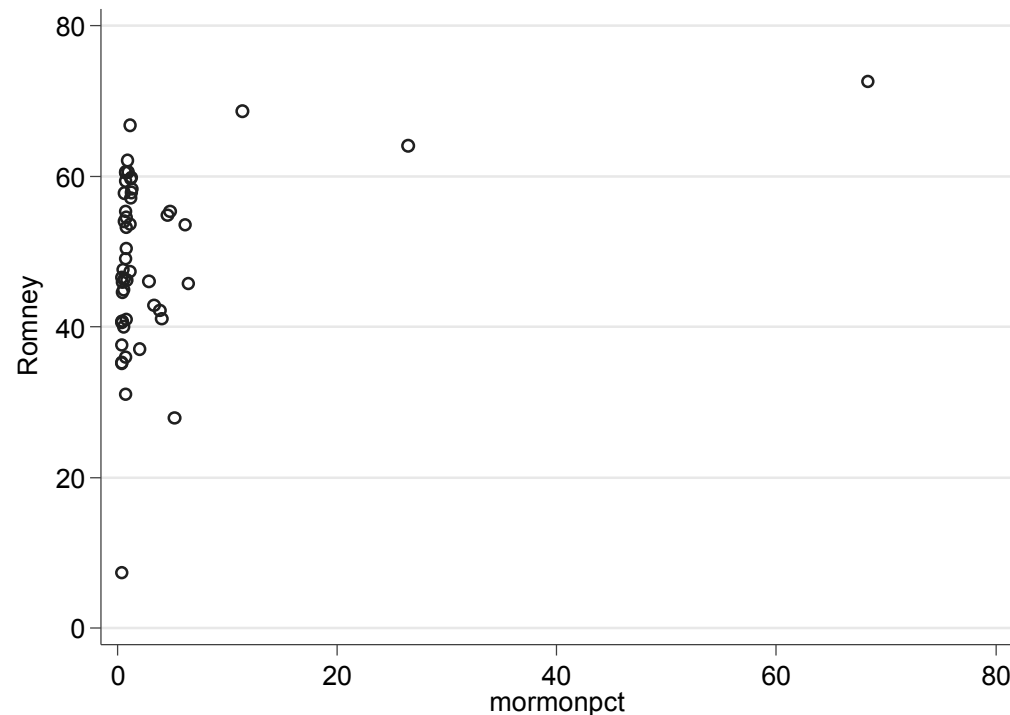


```
gen lq3=log(q3)
gen lpop2010=log(pop2010)
reg lq3 lpop
```

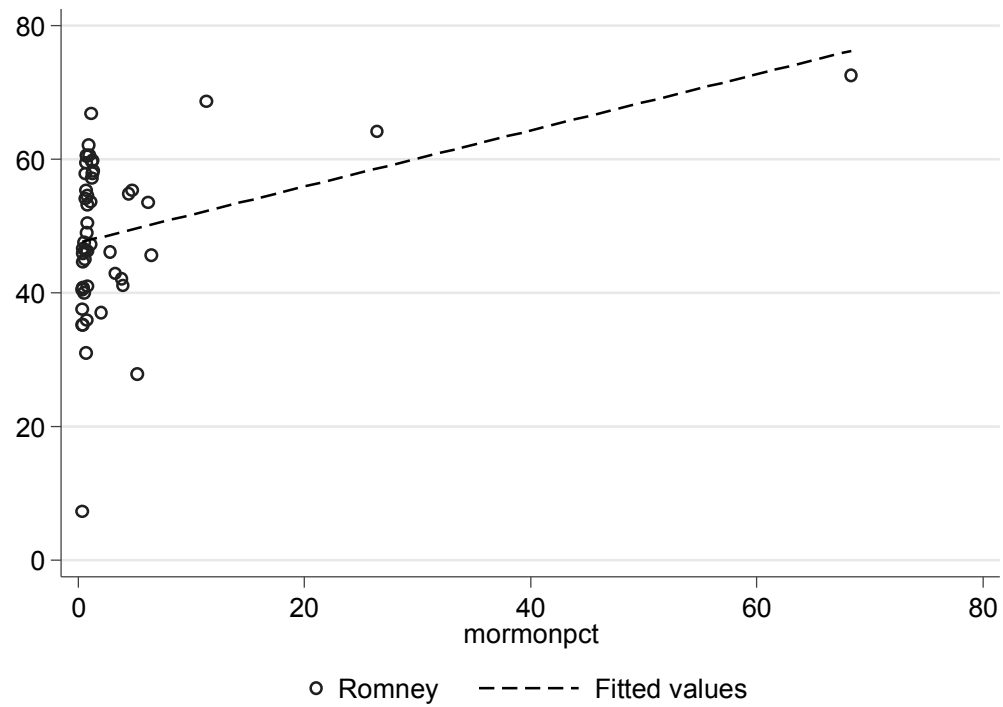


$$\log(\text{spending}) = -4.17 + 0.82 \log(\text{population})$$

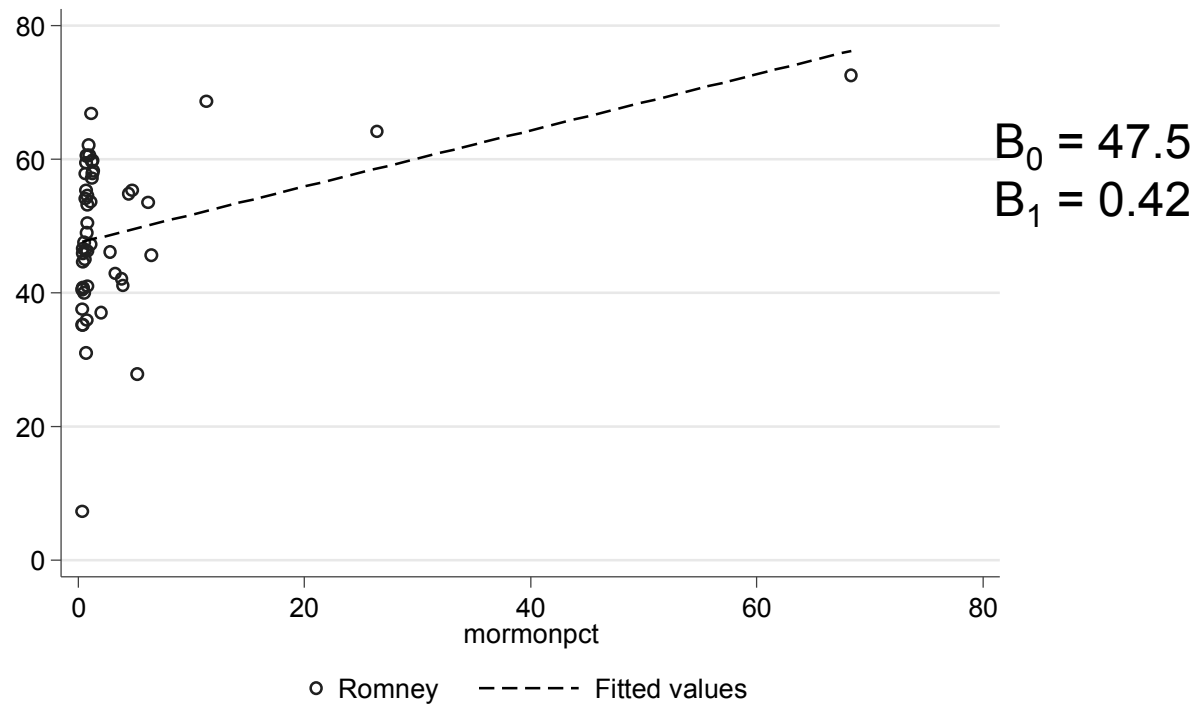
Example of typical log transformation that's only loosely tied to theory



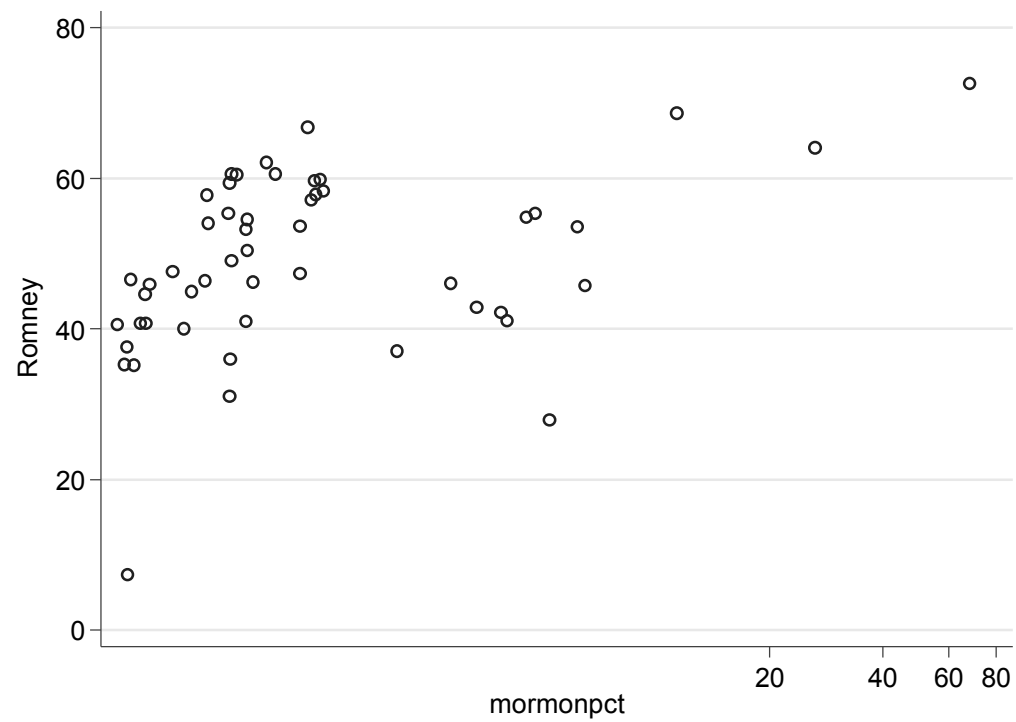
```
twoway (scatter romney pct mormon) (lfit romney pct mormon), scheme(tufte)
```



```
twoway (scatter romneypct mormonpct) (lfit romneypct mormonpct), scheme(tufte)
```



```
twoway (scatter romneypct mormonpct) (lfit romneypct mormonpct), scheme(tufte)
```



```
twoway (scatter romneypct mormonpct) , scheme(tufte) xscale(log)
```

```
. gen lmormonpct=log(mormonpct)
```

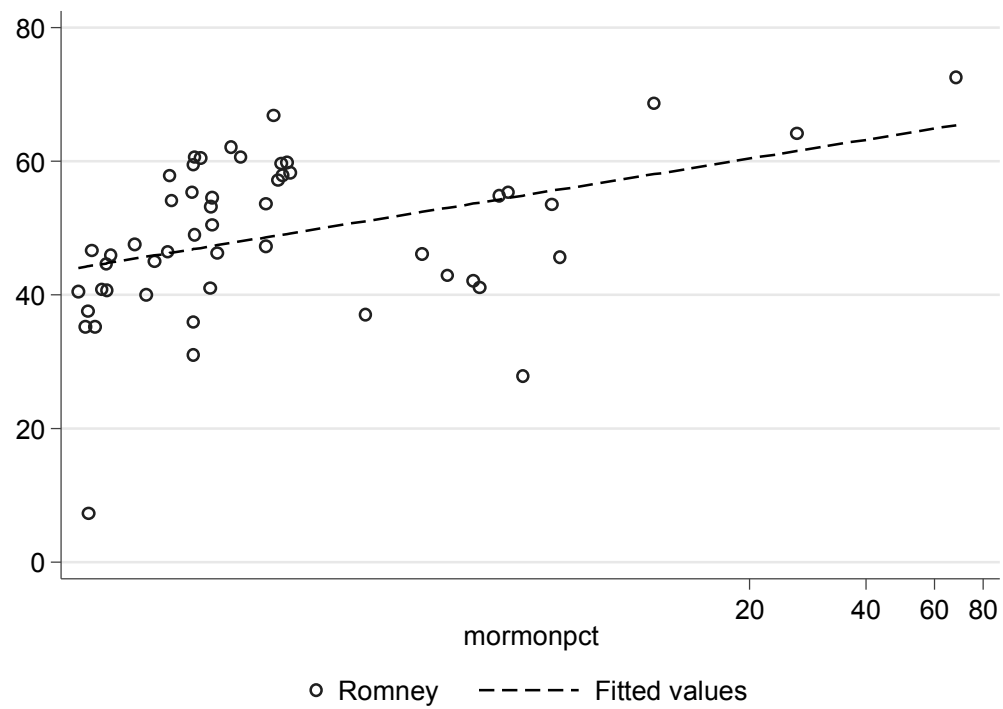
```
. reg romneypct lmormonpct
```

Source	SS	df	MS	Number of obs =	51
Model	1078.47418	1	1078.47418	F(1, 49) =	9.13
Residual	5789.4995	49	118.153051	Prob > F =	0.0040
				R-squared =	0.1570
				Adj R-squared =	0.1398
Total	6867.97368	50	137.359474	Root MSE =	10.87

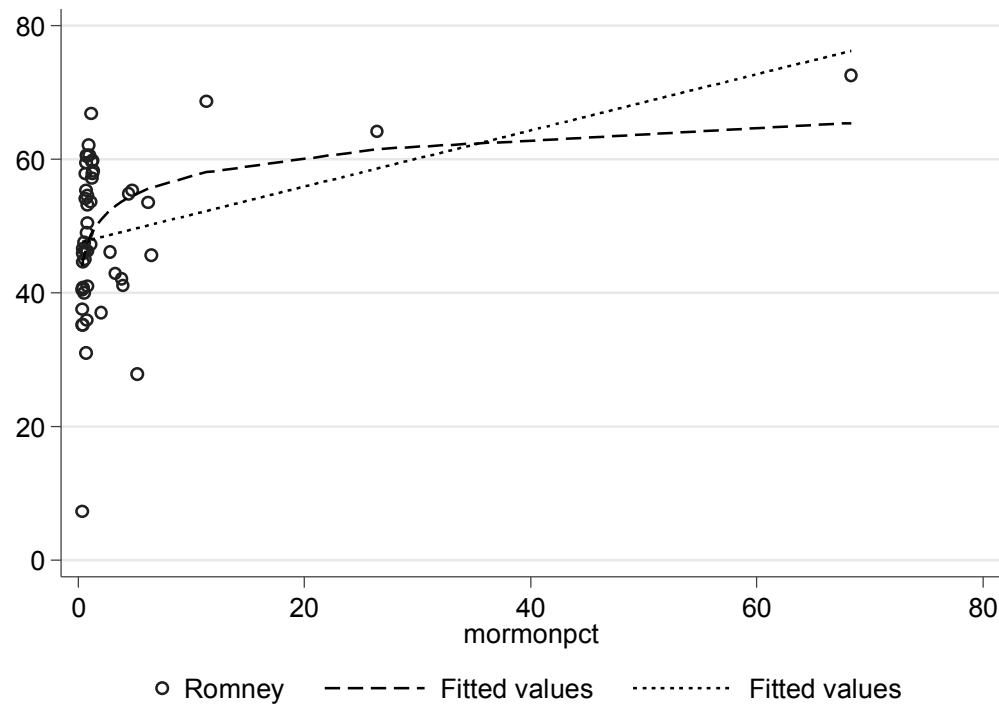
romneypct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lmormonpct	4.09477	1.355336	3.02	0.004	1.37112	6.818421
_cons	48.12661	1.546435	31.12	0.000	45.01894	51.23429

```
. predict py_log
(option xb assumed; fitted values)
```

```
. twoway (scatter romneypct mormonpct) (line py_log mormonpct,sort) , scheme(tufte)
legend(row(1)) xscale(log)
```

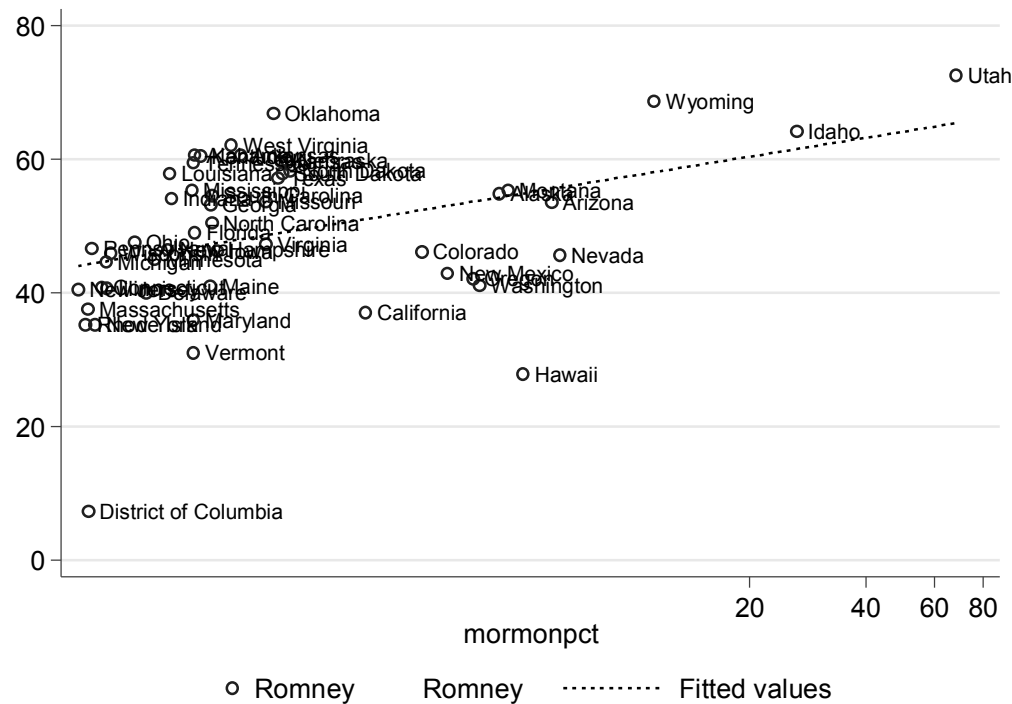


Compare with and without log(x)



```
twoway (scatter romneypct mormonpct) (line py_log mormonpct,sort) ///  
      (lfit romneypct mormonpct) , scheme(tufte) legend(row(1))
```

Finally, look at the residuals again



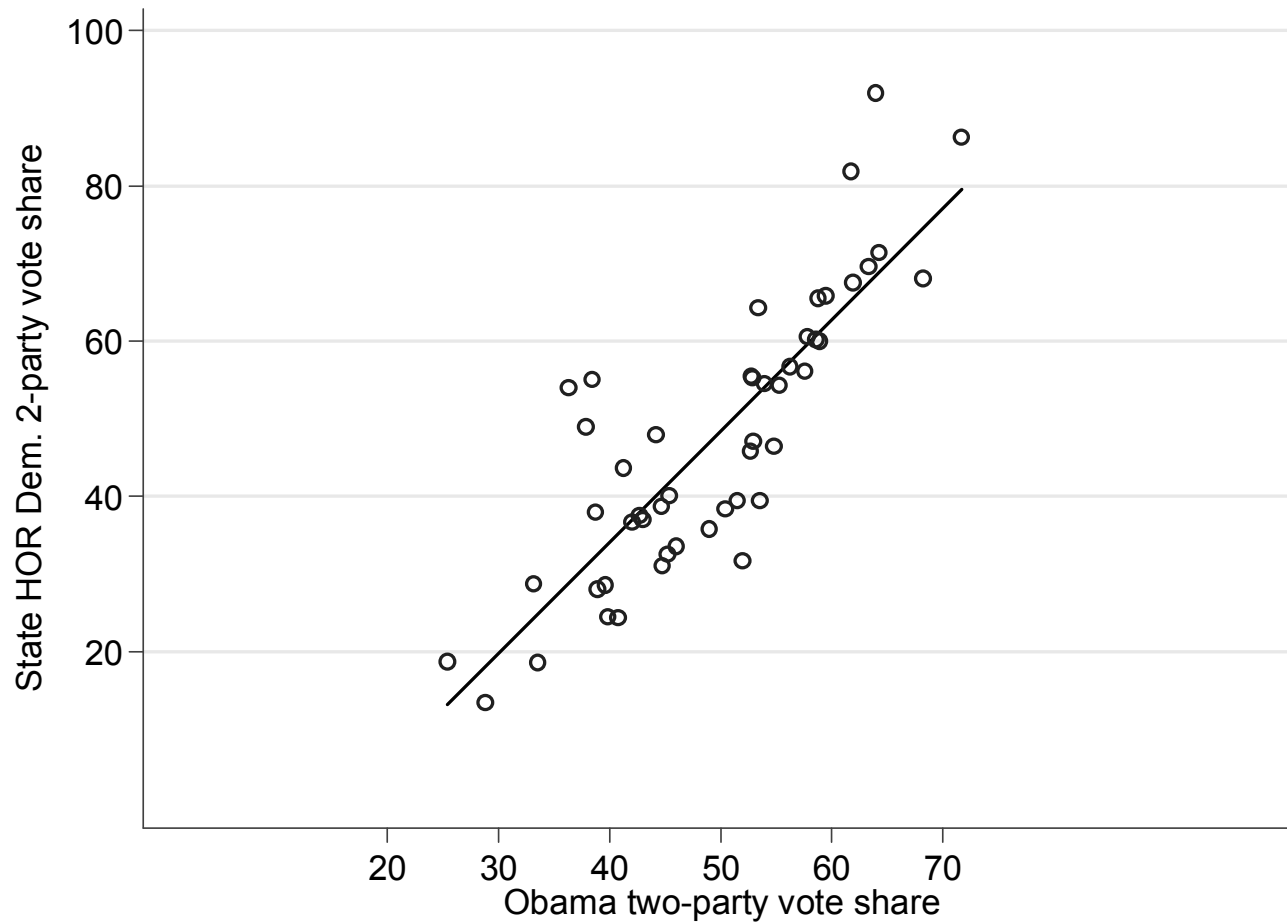
```
twoway (scatter romneypct mormonpct) (scatter romneypct mormonpct, msymbol(none) ///
      mlabel(state)) (line py_log mormonpct, sort) , scheme(tufte) legend(row(1)) xscale(log)
```

Goodness of regression fit

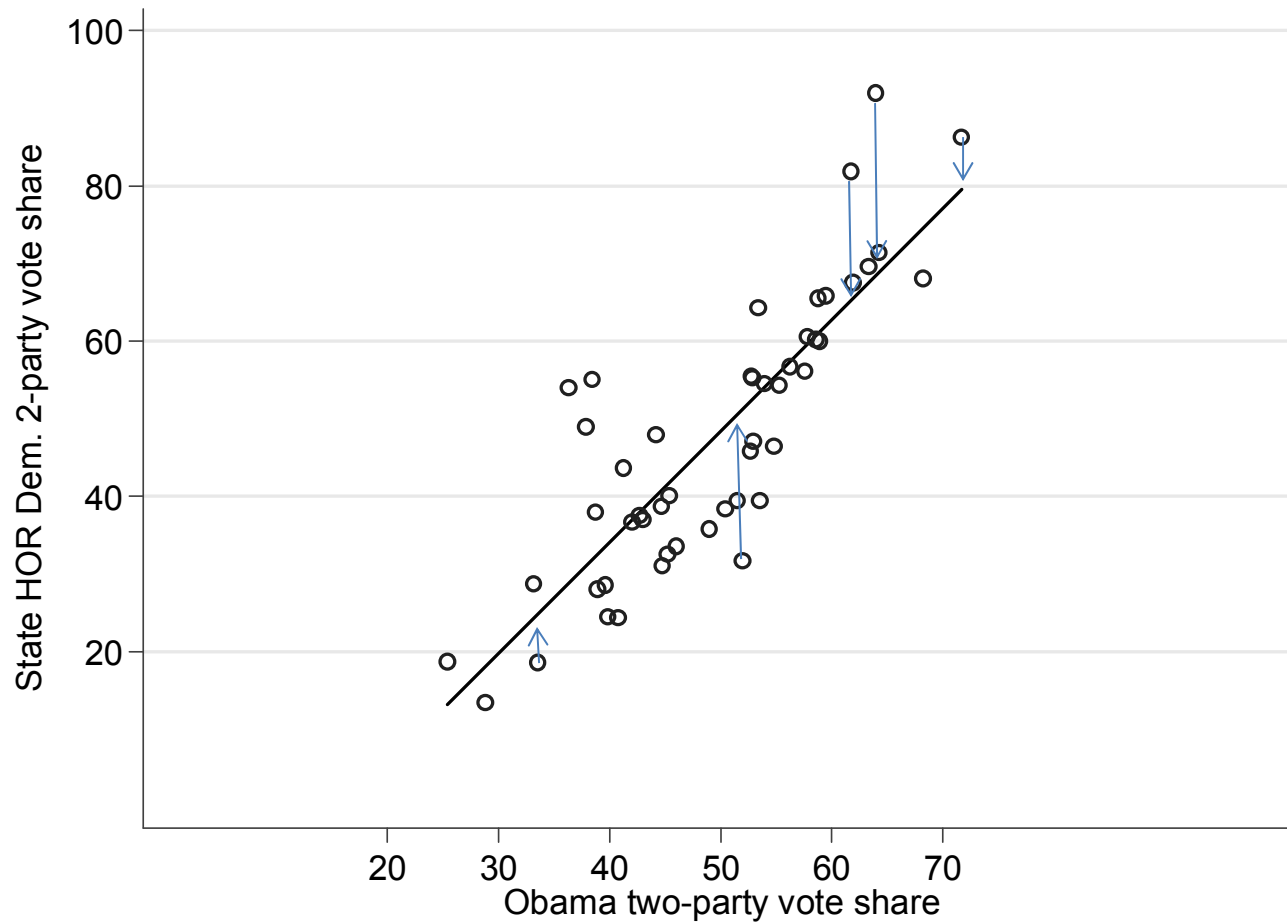
How “good” is the fitted line?

- 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

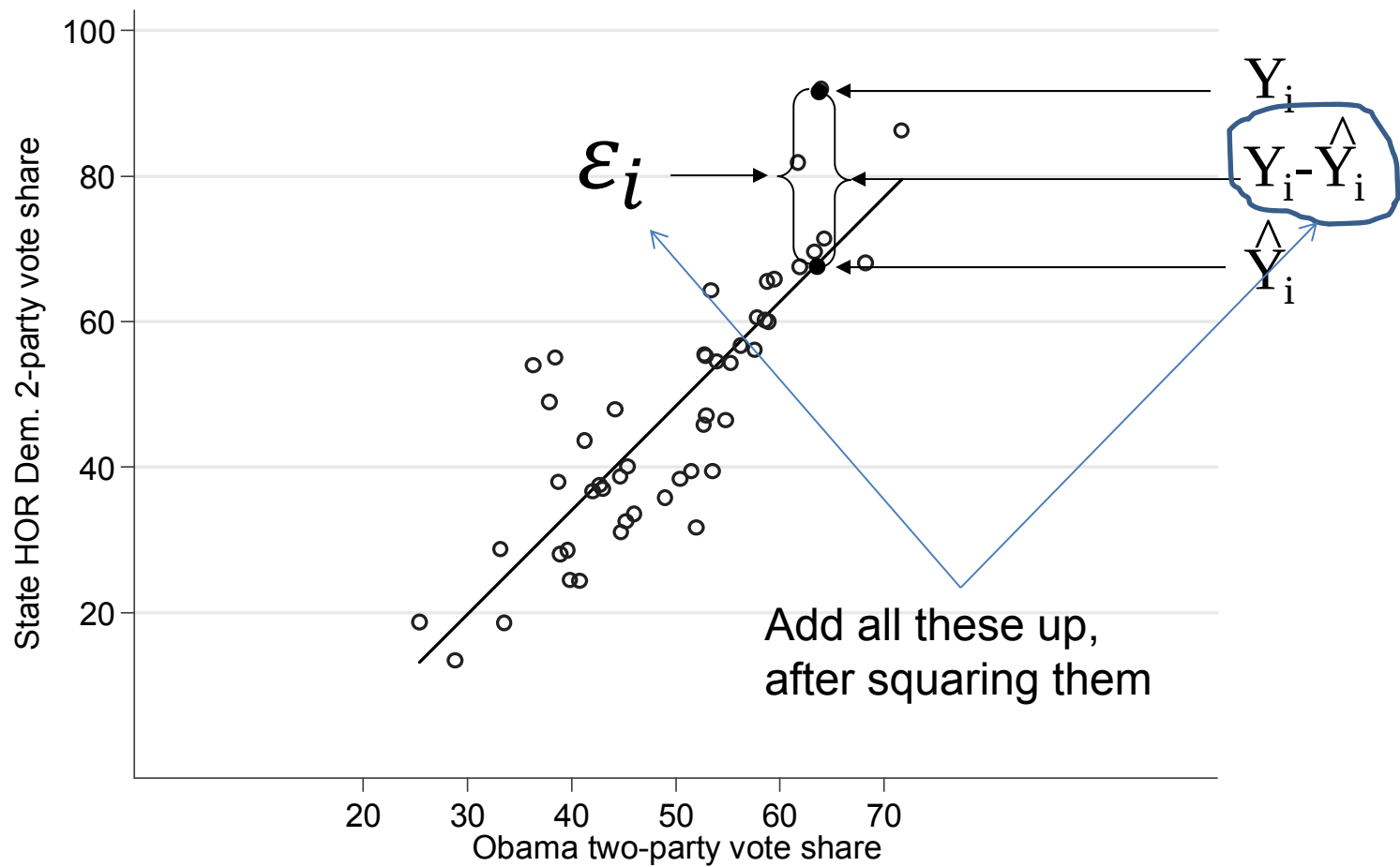
Standard error of the regression line



Standard error of the regression line



Standard error of the regression line



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

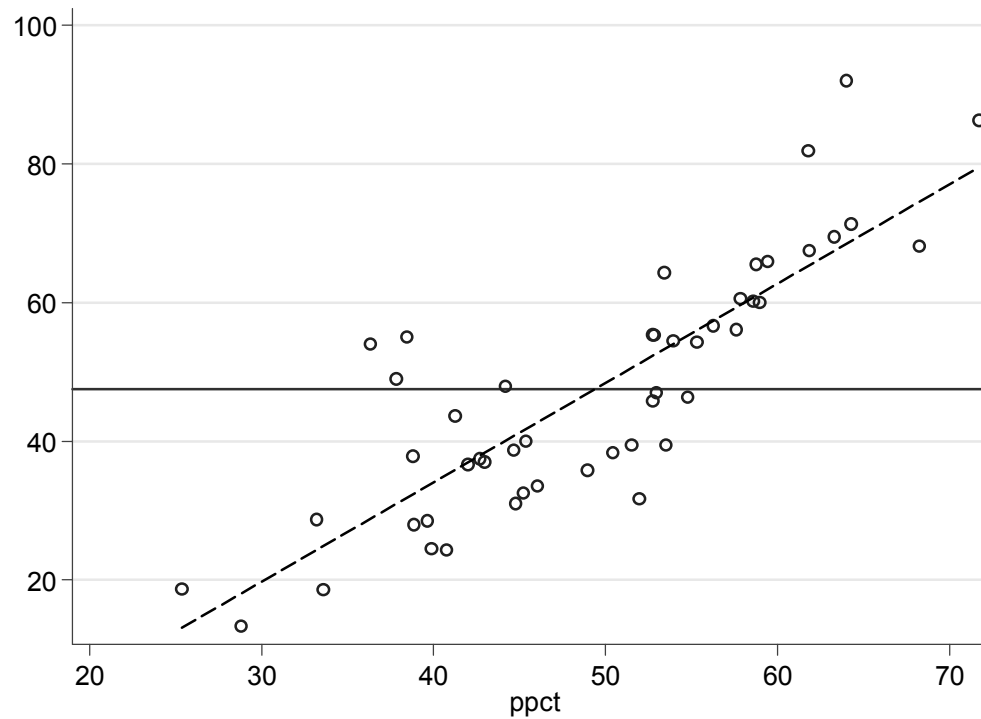
- On average, in-sample predictions will be off the mark by about one standard error of the regression

```
. reg hpct ppct
```

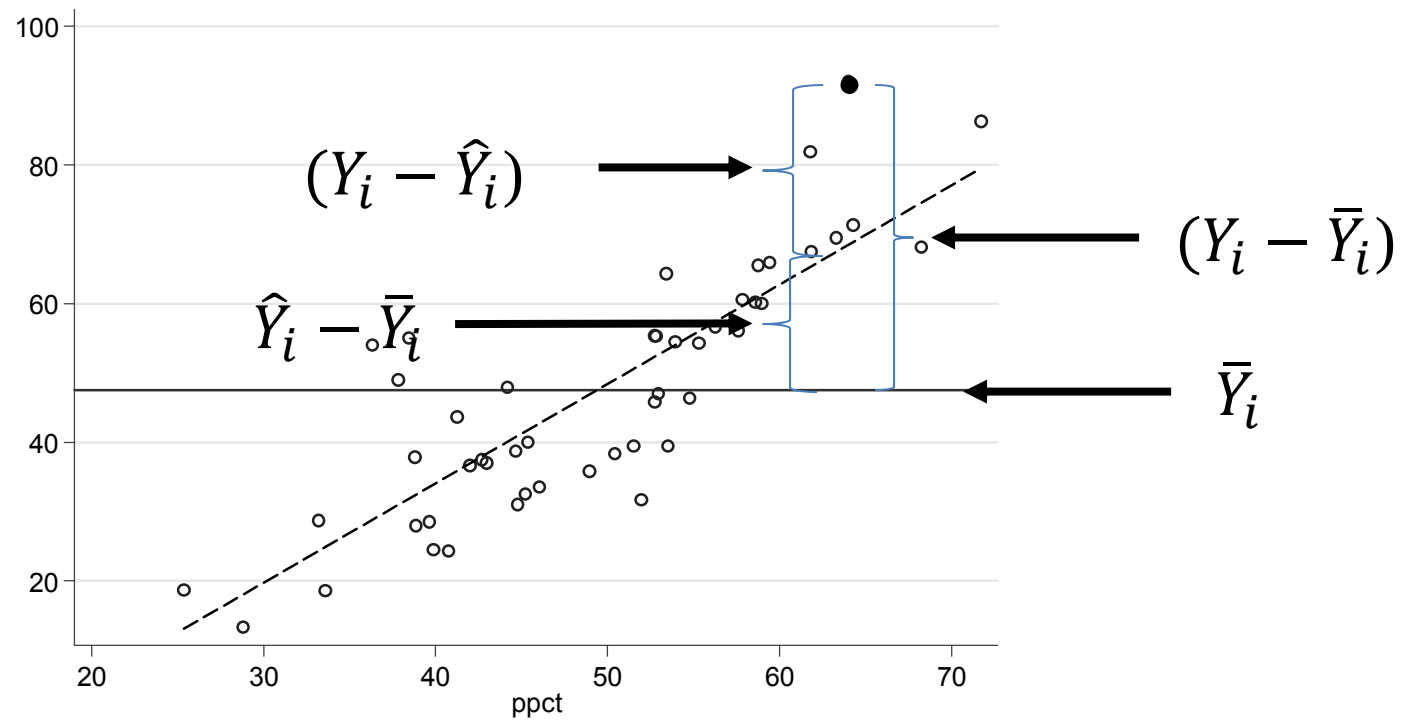
Source	SS	df	MS	Number of obs = 49		
Model	10808.7878	1	10808.7878	F(1, 47)	=	112.73
Residual	4506.44332	47	95.8817727	Prob > F	=	0.0000
Total	15315.2312	48	319.067316	R-squared	=	0.7058
				Adj R-squared	=	0.6995
				Root MSE	=	9.7919

hpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppct	1.433516	.135015	10.62	0.000	1.161901	1.705131
_cons	-23.25307	6.809819	-3.41	0.001	-36.95266	-9.553483

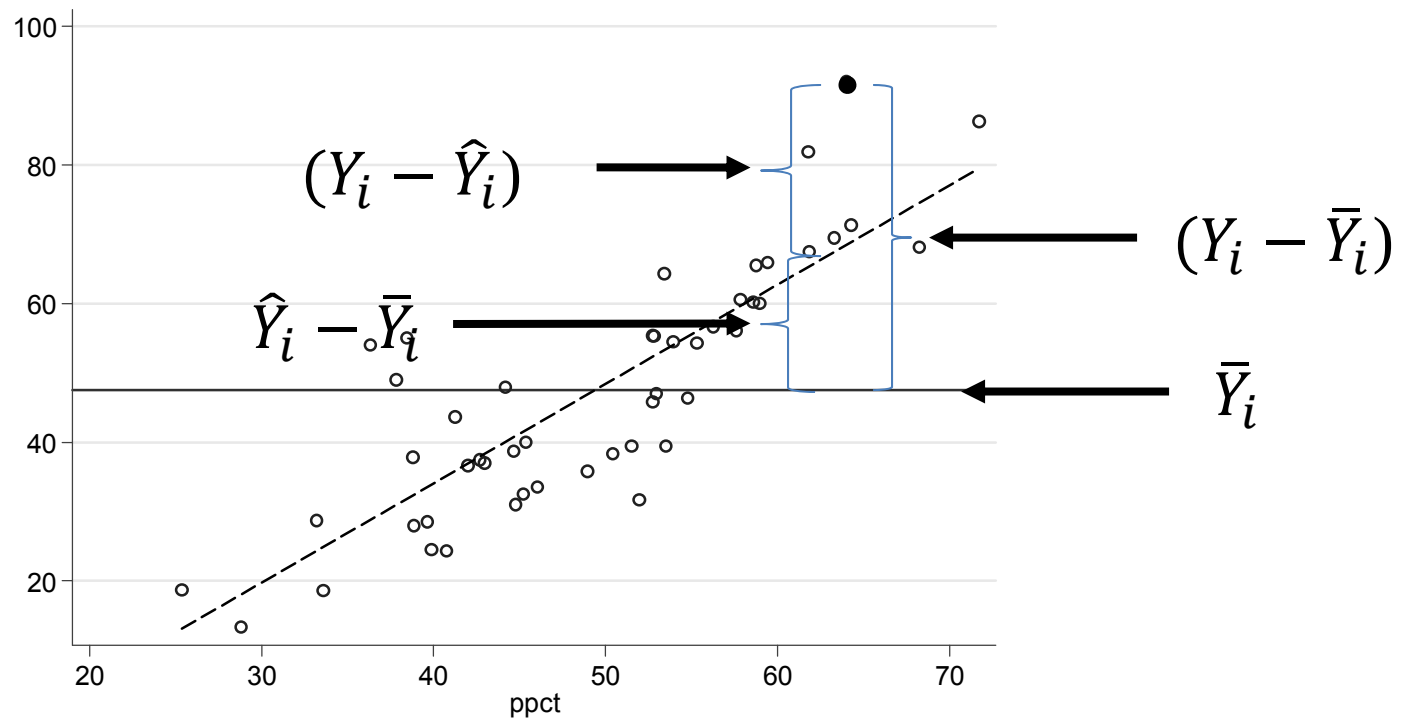
R^2 : A less useful measure of fit



R^2 : A less useful measure of fit



R²: A less useful measure of fit

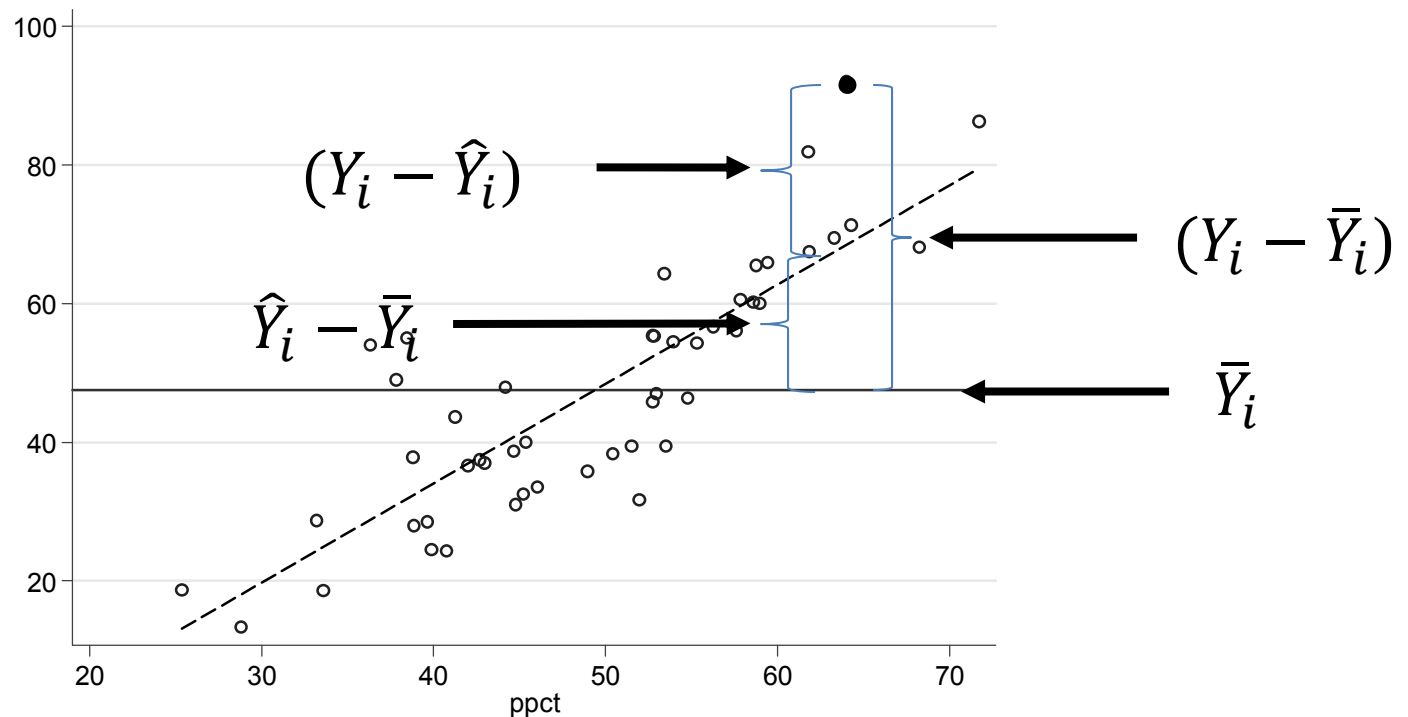


$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \text{"total sum of squares"}$$

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

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

R²: A less useful measure of fit



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

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

pct. variance "explained"

Interpreting SER (Root MSE) and R²

```
. reg hpct ppct
```

Source	SS	df	MS	Number of obs = 49		
Model	10808.7878	1	10808.7878	F(1, 47)	=	112.73
Residual	4506.44332	47	95.8817727	Prob > F	=	0.0000
Total	15315.2312	48	319.067316	R-squared	=	0.7058
				Adj R-squared	=	0.6995
				Root MSE	=	9.7919

hpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppct	1.433516	.135015	10.62	0.000	1.161901	1.705131
_cons	-23.25307	6.809819	-3.41	0.001	-36.95266	-9.553483

Interpreting SER (Root MSE):

- On average, in-sample predictions about the share of a legislature that is Democratic will be off the mark by about **9.8 percentage points**.

Interpreting R²

- Regression model explains about 71% of the variation in the Democratic share of legislative seats.

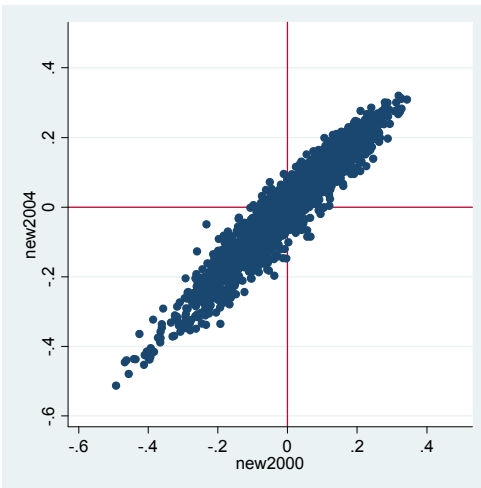
Correlation

Correlation

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

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

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

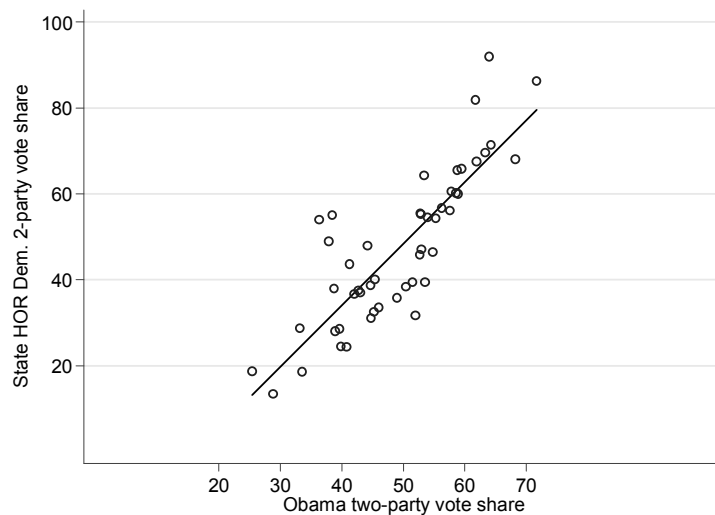


- Measures how closely data points fall along the line
- Varies between -1 and 1 (compare with Tufte p. 102)

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)

Sum-up



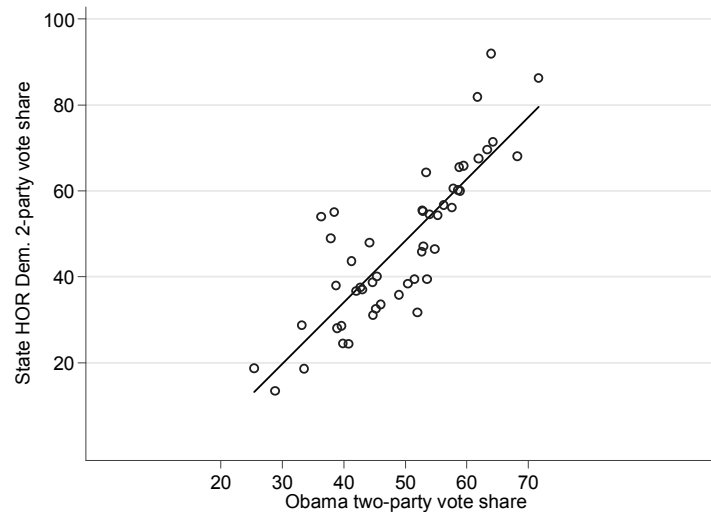
```
. reg hpct ppct
```

Source	SS	df	MS
Model	10808.7878	1	10808.7878
Residual	4506.44332	47	95.8817727
Total	15315.2312	48	319.067316

Number of obs = **49**
 F(1, 47) = 112.73
 Prob > F = 0.0000
 R-squared = **0.7058**
 Adj R-squared = 0.6995
 Root MSE = **9.7919**

hpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppct	1.433516	.135015	10.62	0.000	1.161901	1.705131
_cons	-23.25307	6.809819	-3.41	0.001	-36.95266	-9.553483

Sum-up



```
. reg hpct ppct
```

Source	SS	df	MS
Model	10808.7878	1	10808.7878
Residual	4506.44332	47	95.8817727
Total	15315.2312	48	319.067316

Number of obs = 49
 F(1, 47) = 112.73
 Prob > F = 0.0000
 R-squared = 0.7058
 Adj R-squared = 0.6995
 Root MSE = 9.7919

hpct	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppct	1.433516	.135015	10.62	0.000	1.161901	1.705131
_cons	-23.25307	6.809819	-3.41	0.001	-36.95266	-9.553483