

Regression (I)

- 1 A simple linear model
- 2 Ordinary Least Squares (OLS)
- 3 Regression output

FiveThirtyEight

Nate Silver's Political Calculus

313.0

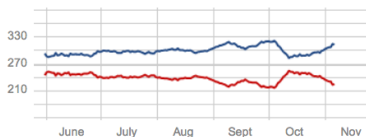
+14.0 since Oct. 30

Electoral
vote

225.0

-14.0 since Oct. 30

270 to win



90.9%

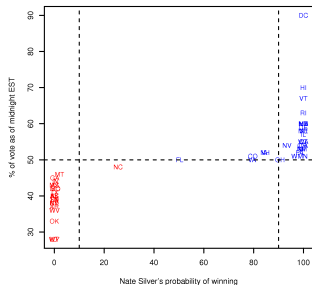
+13.5 since Oct. 30

Chance of
Winning

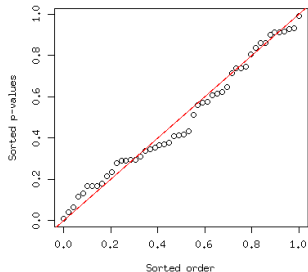
9.1%

-13.5 since Oct. 30

50%

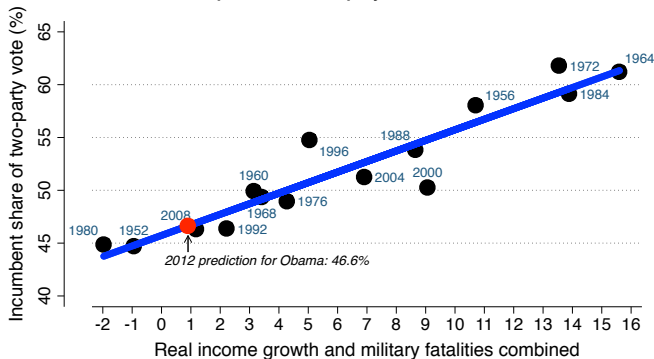


538 p-values



Obama's re-election prospect under bread and peace voting

October 26 2012 update based on projections of Oct-Nov 2012 conditions



Combination of real growth and fatalities weights each variable by its estimated coefficient.
 Estimated effects of fatalities on vote shares: -0.7% in 2008 (Iraq), -7.4% in 1968 (Vietnam),
 -9.7% in 1952 (Korea); negligible in 1964, 1976, 2004, 2012, and null in other years.
 Source: www.douglas-hibbs.com October 26 2012

To what extent can trust in government be predicted from variations in economic growth?

DV: Trust in government

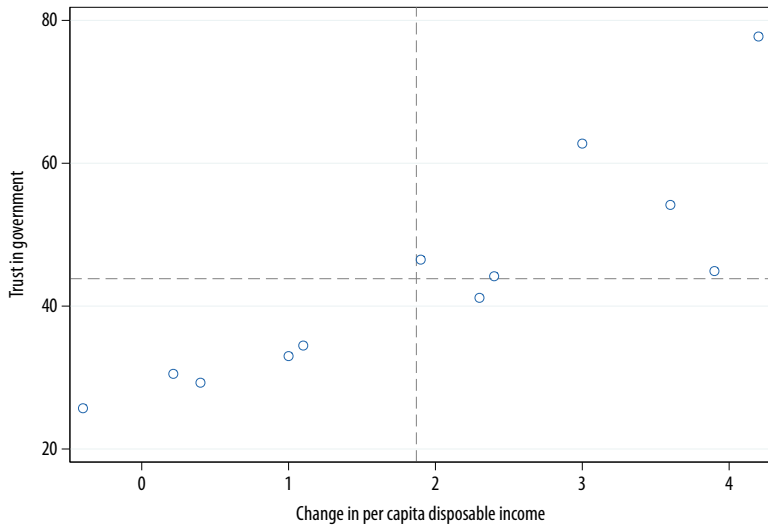
“Just about always/Most of the time”
(American National Election Studies)

IV: Economic performance

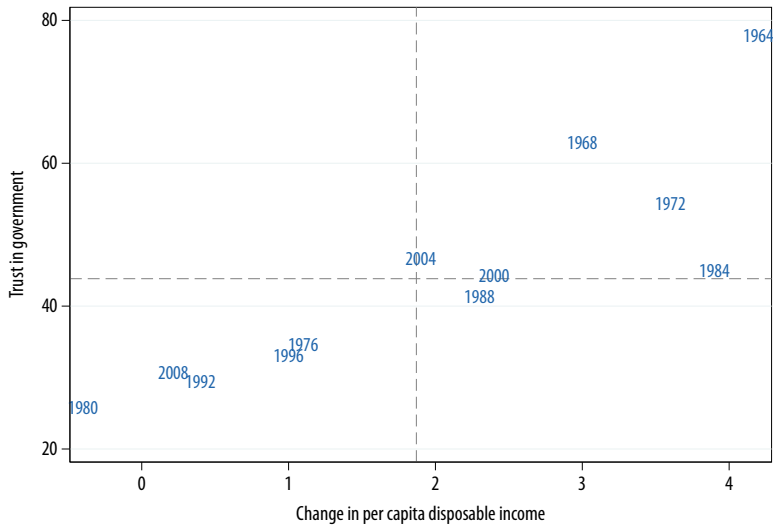
Change in per capita disposable
income (Bureau of Economic
Analysis)

Example and data provided by John Sides.

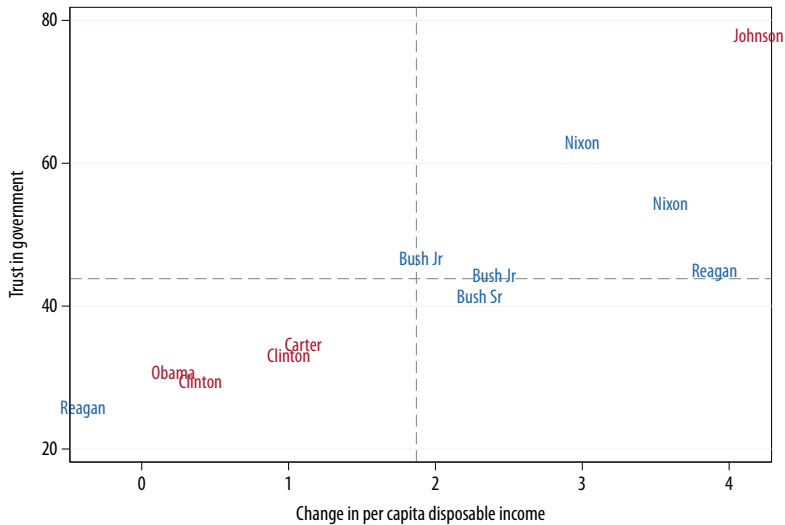




Dashed lines at averages. Pearson correlation $\rho = .86$ significant at $p < .01$.



Dashed lines at averages. Pearson correlation $\rho = .86$ significant at $p < .01$.



Dashed lines at averages. Pearson correlation $\rho = .86$ significant at $p < .01$.

Simple linear regression

Equations

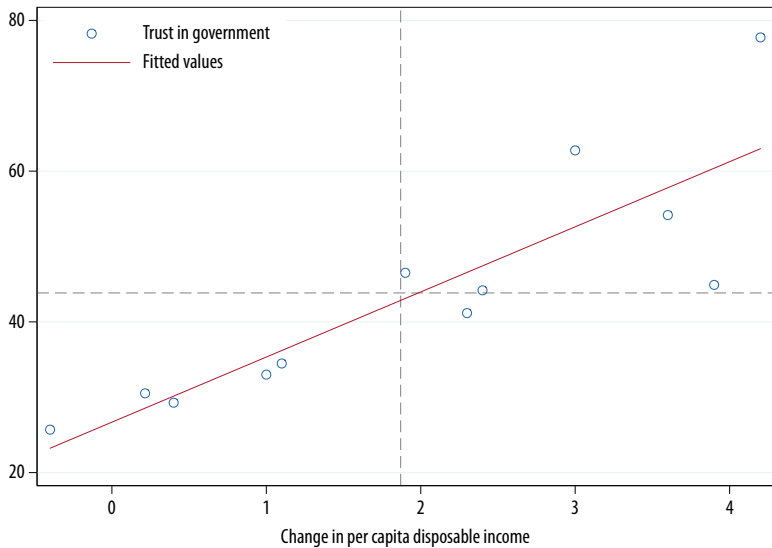
$$Y = \alpha + \beta X + \epsilon \quad \hat{Y} = \hat{\alpha} + \hat{\beta}X + \hat{\epsilon} \quad \epsilon = Y - \hat{Y}$$

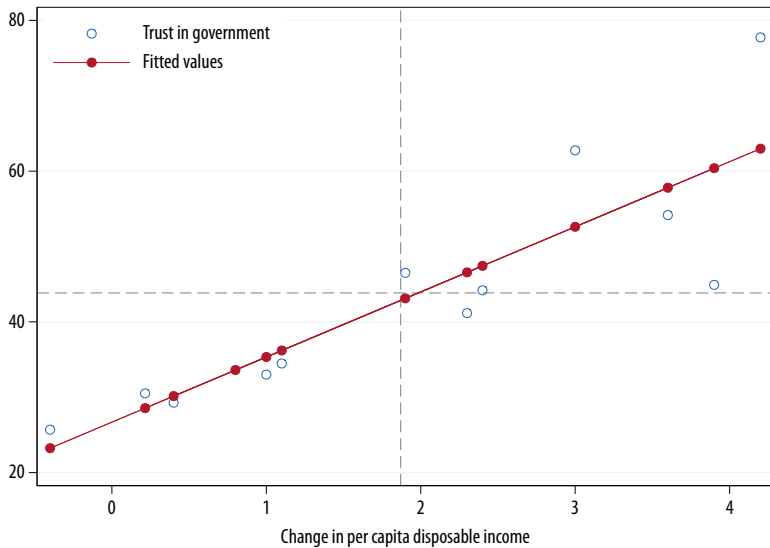
Parameters

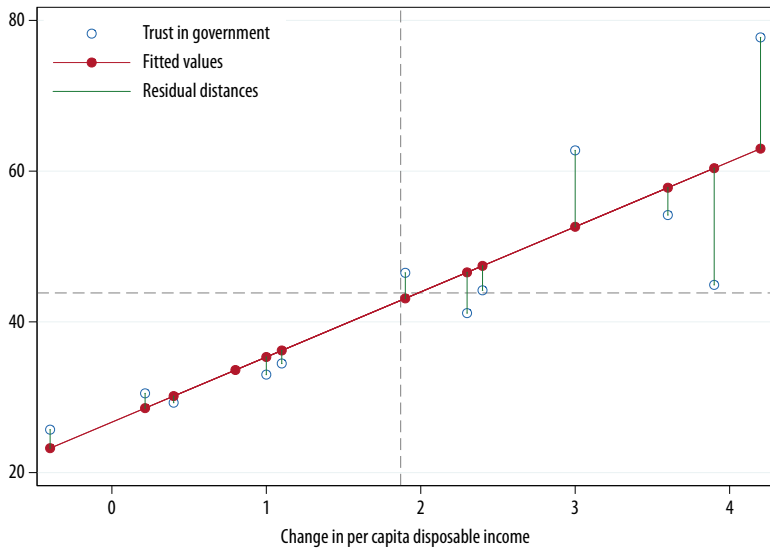
- Y is the dependent variable and \hat{Y} its predicted value
- X is the independent variable used as a predictor of Y
- α is the **constant** (intercept)
- β is the **regression coefficient** (slope)
- ϵ is the **error term** (residuals)

Warning

The model assumes a *linear, additive* relationship.







Ordinary Least Squares (OLS)

Error term

In a simple linear model $Y = \alpha + \beta X + \epsilon$, the regression coefficient β is calculated as to minimize the **residual sum of squares**

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

where $Y_i - \hat{Y}_i$ is the residual (or error term) of each observation.

Parameter estimation

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}_X} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \alpha = \bar{Y} - \beta \bar{X}$$

reg y x

. regress trust income

Source	SS	df	MS
Model	1908.80221	1	1908.80221
Residual	643.906248	10	64.3906248
Total	2552.70846	11	232.064405

Number of obs = 12
F(1, 10) = 29.64
Prob > F = 0.0003
R-squared = 0.7478
Adj R-squared = 0.7225
Root MSE = 8.0244

trust	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	8.639373	1.586767	5.44	0.000	5.103836	12.17491
_cons	26.69501	3.888016	6.87	0.000	18.03197	35.35805

Top left: ANOVA table. Top right: model fit.
Bottom: regression coefficients.

Interpretation of fit

Number of observations N , significance test $H_0 : \beta = 0$, coefficient of determination R^2 .

regress brack income

Source	SS	df	MS
Model	1089.88225	1	1089.88225
Residual	945.588338	10	94.5588338
Total	2035.47059	11	185.042781

Sum of Squares	1089.88225
df	1
Prob > F	0.0003
R-squared	0.7478
Adj R-squared	0.7225
Root MSE	9.7244

	Coef.	Std. Err.	t	Pr(> t)	[95% Conf. Interval]
income	8.439575	1.388787	6.08	0.0003	5.583888 11.29525
_cons	26.49581	3.888826	6.82	0.0003	18.67297 34.31865

Goodness of fit

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

As predicted variance increases, $RSS \rightarrow 0$ and $R^2 \rightarrow 1$, indicating a more efficient fit.

Sanity check

The most important statistic here is the actual number of observations in the model.

Number of obs =	12
F(1, 10) =	29.64
Prob > F =	0.0003
R-squared =	0.7478
Adj R-squared =	0.7225
Root MSE =	9.7244

Interpretation of regression coefficients

A regression coefficient estimates the variation in Y predicted by a change in one unit of X (recall that $Y = \alpha + \beta X + \epsilon$)

regress trust income

Source	SS	df	MS	Number of obs =
Model	1780.88223	1	1780.88223	12
Residual	832.99028	10	83.299028	
Total	2613.87251	11	237.624773	

R-squared	0.6814	F(1, 10) =	20.82	Prob > F =	0.0000
Adj R-squared	0.6729				
Root MSE	8.8544				

trust	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
income	8.639373	1.586767	5.44	0.000	5.103836 12.17491
_cons	26.69501	3.888016	6.87	0.000	18.03197 35.35805

trust	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	8.639373	1.586767	5.44	0.000	5.103836	12.17491
_cons	26.69501	3.888016	6.87	0.000	18.03197	35.35805

- The **coefficient** is the slope β of the regression line and the **constant** is its intercept, the coordinate of origin $\alpha = \hat{Y}_{X=0}$.
- The **standard error**, t -value and p -value test whether the coefficient is significantly different from 0.

Thanks for your attention

Project

- Correct and improve first draft
- Finalize association tests and interpretations

Readings

- *Stata Guide*, Sec. 11
- *Making History Count*, ch. 4

Practice

- Replicate do-file
- Try getting OLS results in your second draft