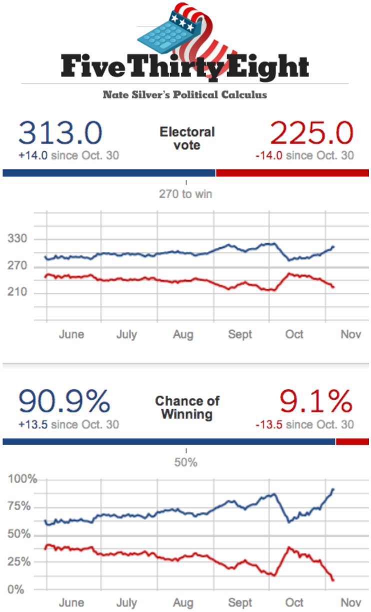
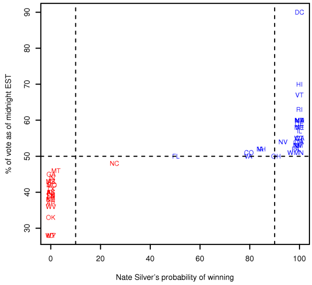
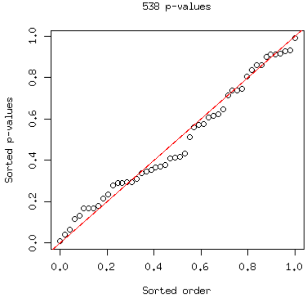
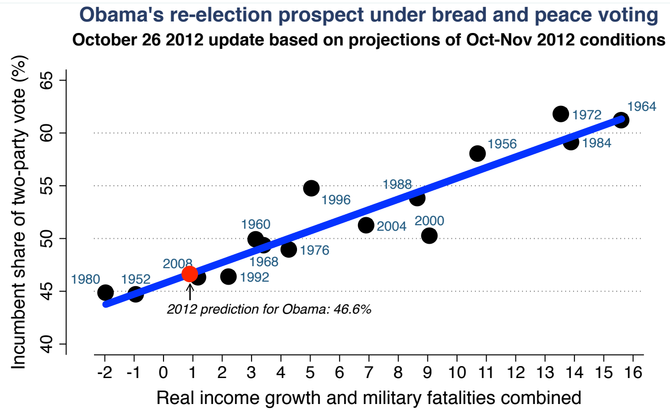
Linear Regression (I)

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
330
270
210
June
July
Aug
Sept
Oct
Nov
90.9%
+13.5 since Oct. 30
Chance of Winning
9.1%
-13.5 since Oct. 30
50%
100%
75%
50%
25%
0%
June
July
Aug
Sept
Oct
Nov


% of vote as of midnight EST
Nate Silver's probability of winning
DC
HI
VT
RI
NV
NC
FL
OK
MT
CO
VA
OH
0
20
40
60
80
100
30
40
50
60
70
80
90


538 p-values
Sorted p-values
Sorted order
0.0
0.2
0.4
0.6
0.8
1.0



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?

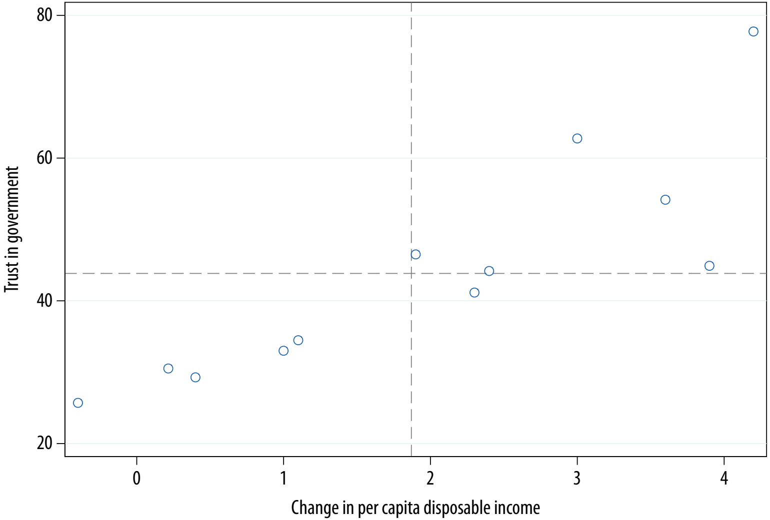
DV: Trust in government

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

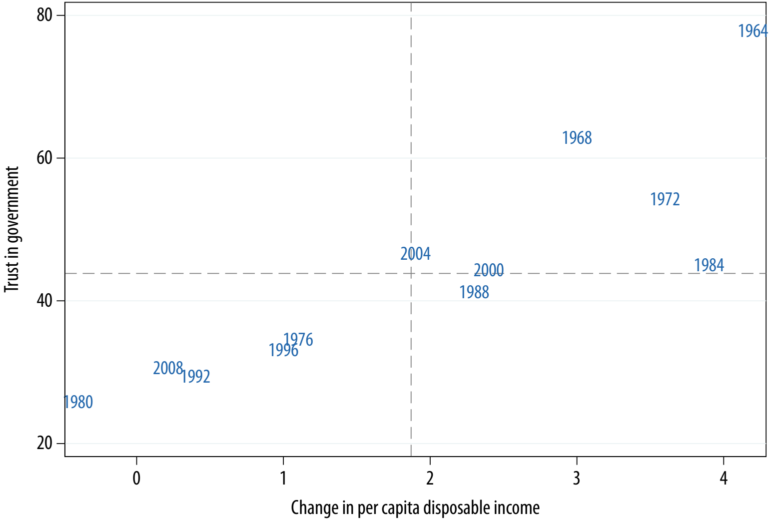
IV: Economic performance

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

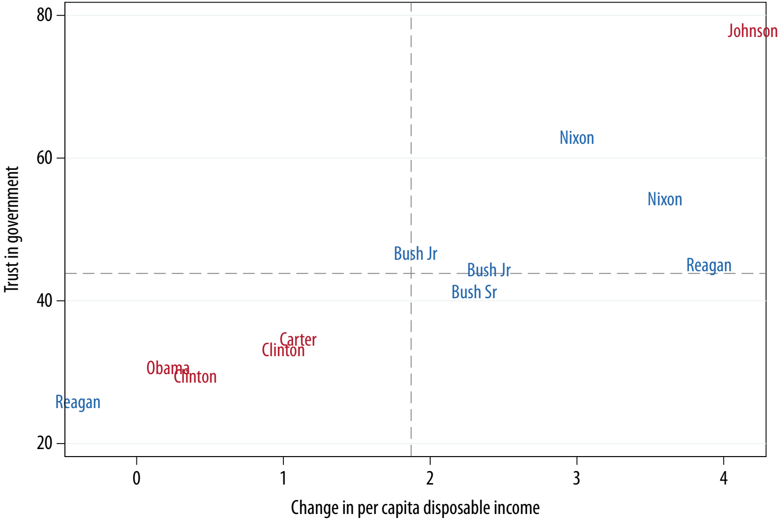
Example and data provided by John Sides.



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

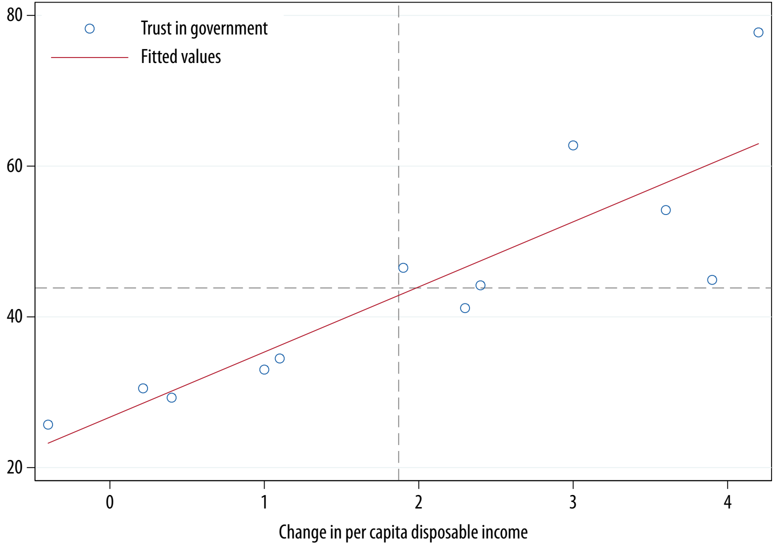
$Y = \alpha + \beta X + \epsilon \quad \hat{Y} = \hat{\alpha} + \hat{\beta} X + \hat{\epsilon} \quad \epsilon = Y - \hat{Y}$

Parameters

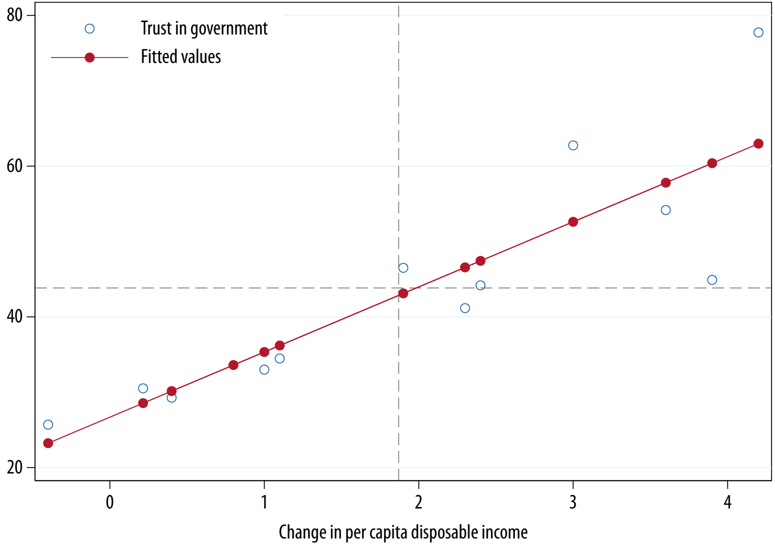
- Y is the dependent variable and $\hat{Y}$ its predicted value
- X is the independent variable used as a predictor of Y
- α is the constant (intercept)
- β is the regression coefficient (slope)
- ϵ is the error term (residuals)

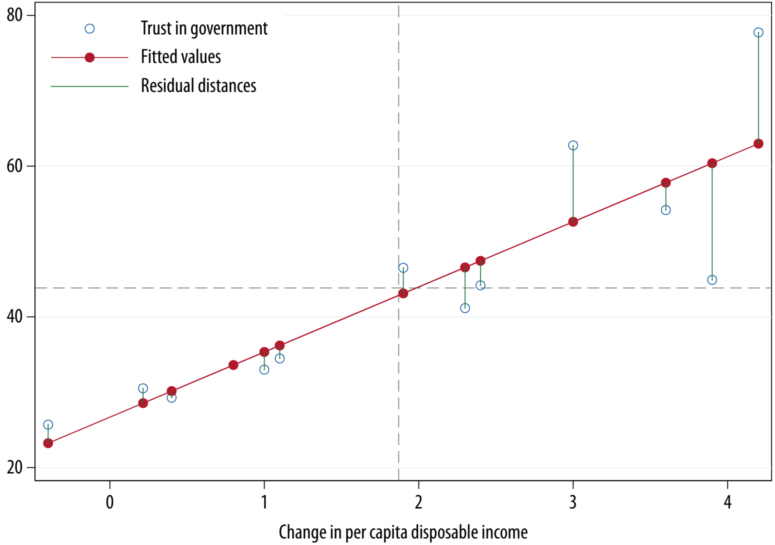
Warning

The model assumes a *linear*, *additive* relationship.



Trust in government
Fitted values
80
60
40
20
0
1
2
3
4
Change in per capita disposable income


Trust in government
Fitted values
80
60
40
20
0
1
2
3
4
Change in per capita disposable income



Trust in government
Fitted values
Residual distances
80
60
40
20
0
1
2
3
4
Change in per capita disposable income

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

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

reg y x

```
. regress trust income

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F(1, 10)        =     29.64
       Model |  1908.80221         1  1908.80221   Prob > F        =    0.0003
    Residual |  643.906248        10  64.3906248   R-squared       =    0.7478
-------------+----------------------------------   Adj R-squared   =    0.7225
       Total |  2552.70846        11  232.064405   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.

Goodness of fit

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

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

Sanity check

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

```
. 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

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

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 =      12
-------------+------------------------------           F(  1,    10) =   29.64
       Model |  1908.80221     1  1908.80221           Prob > F      =  0.0003
    Residual |  643.906248    10  64.3906248           R-squared     =  0.7478
-------------+------------------------------           Adj R-squared =  0.7225
       Total |  2552.70846    11  232.064405           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
------------------------------------------------------------------------------
```

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