

Linear Regression (I)

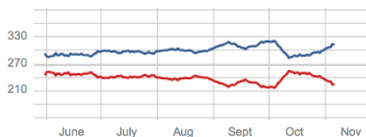
- 1 A simple linear model
- 2 Ordinary Least Squares (OLS)
- 3 Regression output
- 4 Draft No. 2

FiveThirtyEight

Nate Silver's Political Calculus

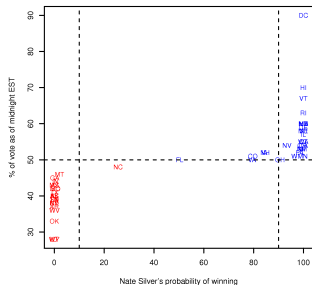
313.0 **Electoral vote** **225.0**
 +14.0 since Oct. 30 -14.0 since Oct. 30

270 to win

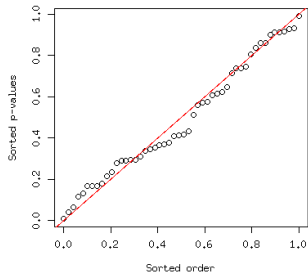


90.9% **Chance of Winning** **9.1%**
 +13.5 since Oct. 30 -13.5 since Oct. 30

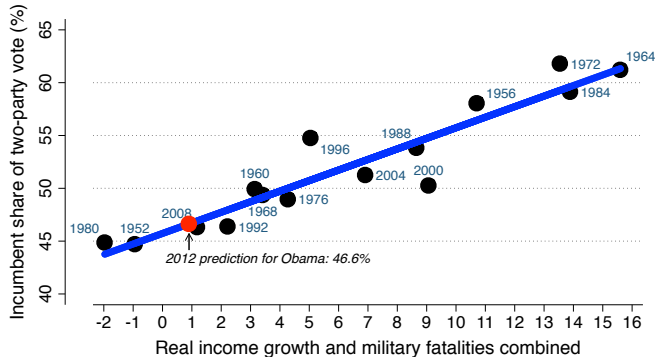
50%



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

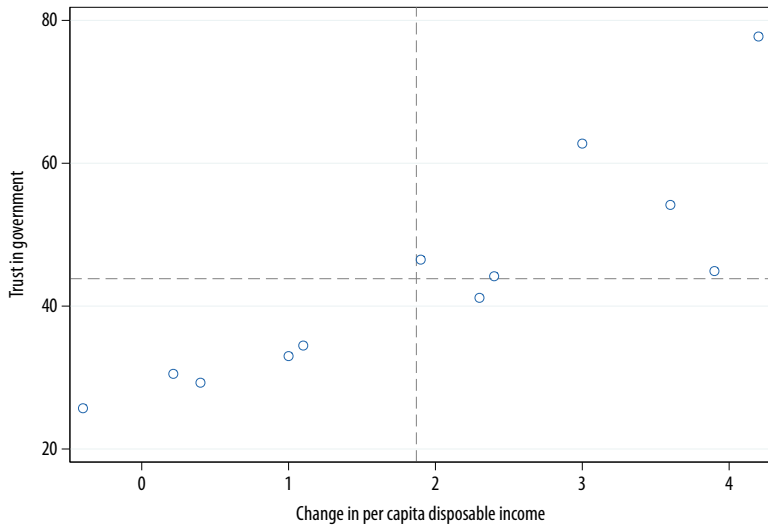
“Just about always/Most of the time”
(American National Election Studies)

IV: Economic performance

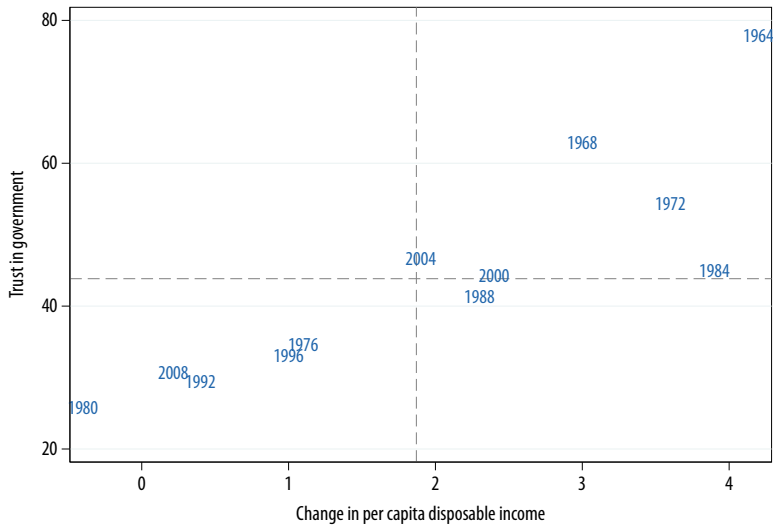
Change in per capita disposable
income (Bureau of Economic
Analysis)

Example and data provided by John Sides.

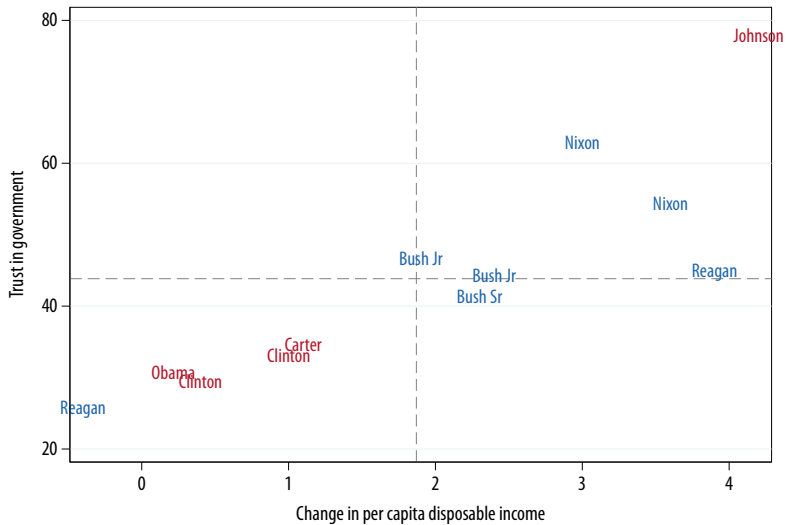




Dashed lines at averages. Pearson correlation $\rho = .86$ significant at $p < .01$.



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



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

Simple linear regression

Equations

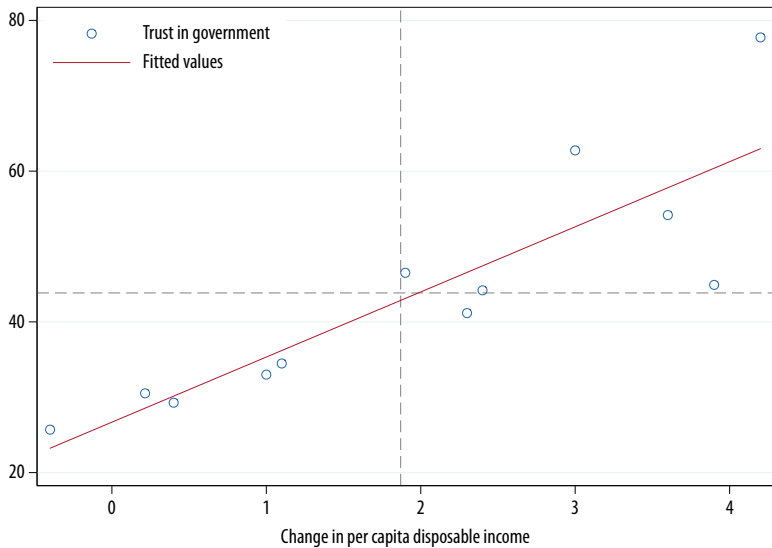
$$Y = \alpha + \beta X + \epsilon \quad \hat{Y} = \hat{\alpha} + \hat{\beta}X + \hat{\epsilon} \quad \epsilon = Y - \hat{Y}$$

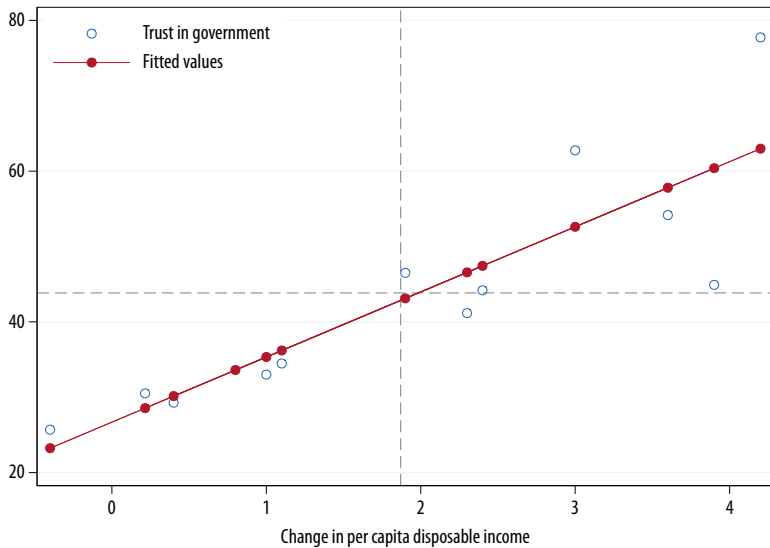
Parameters

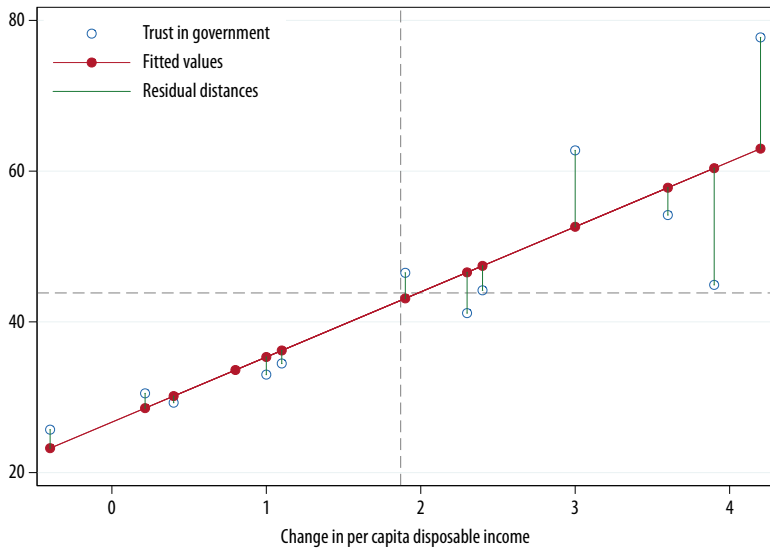
- Y is the dependent variable and \hat{Y} its predicted value
- X is the independent variable used as a predictor of Y
- α is the **constant** (intercept)
- β is the **regression coefficient** (slope)
- ϵ is the **error term** (residuals)

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 , **root mean square error (RMSE)**.

regress kmph income

Source	SS	df	MS		
Model	1089.88225	1	1089.88225	$F(1, 10) = 29.64$	$P < 0.0001$
Residual	363.308338	10	36.3308338	$R^2 = 0.7478$	
Total	1453.19059	11	132.190054	Root MSE = 6.0244	

	Sum of Squares	df	Mean Square	F-Statistic	Prob > F
Model	1089.88225	1	1089.88225	29.64	0.0001
Residual	363.308338	10	36.3308338		
Total	1453.19059	11			

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 $RSS \rightarrow 0$ more efficient fit, $R^2 \rightarrow 1$.

Sanity check

Focus on getting N and the RMSE right.

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 =
Model	1780.88223	1	1780.88223	12
Residual	832.99028	10	83.299028	
Total	2613.87251	11	237.624773	

R-squared	0.6814	Prob > F	0.0000
Adj R-squared	0.6729	Prob > t	0.0000
F-statistic	21.38	Prob > t	0.0000

Variable	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.

Where we are now

Univariate statistics

- Introduction
- Dataset
- Variables

Assignment No. 1

corrected }
revised }
appended }

Bivariate statistics

- Associations
- Correlations
- Simple OLS

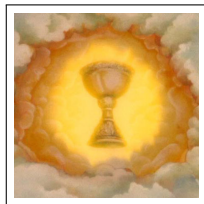
Assignment No. 2



Statistical modelling

- Regressions
- Diagnostics
- Conclusion

Final paper



Essential instructions

Revise Draft No. 1

- go through corrections
- remove technical content
- **rewrite until concision**

Pay attention to paragraph limits and scientific style (esp. sources).

Explore associations

- between DV and IVs (covariates, controls), or between two IVs
- with graphs and then with significance tests

Write up **substantive results** as sentences; cite significance tests and other statistics in brackets, e.g. ($\rho = .7, p < .05$).

Paper template, structure and style

LYNN WHITE University of Nebraska—Lincoln

Group Names

Statistical Reasoning and Quantitative Methods, Fall 2012
Research Paper

Why Titles Matter: Evidence from Contemporary European Academia, 2012

Draft 1 · Draft 2 · Final Version · Date

Abstract

In [one paragraph](#), write up a short summary of your work when you are done with the analysis. Make sure to mention the keywords and main results of your research. A few lines are enough. Also make sure to use scientific style throughout your paper. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here.

Feedback

Write a [short numbered list](#) of issues that you want to discuss in more detail about your research. Include references to your code by mentioning the line numbers of your do-file. All sorts of questions and comments are also very welcome! A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here. A wonderful paragraph goes here.

This article provides advice about preparing research reports for submission to professional journals in general and Journal of Marriage and Family in particular. In addition to working through all the major parts of a research paper, I provide some general advice about writing, editing, and revising. The article is intended to help new professionals improve the quality of their journal submissions and the likelihood of successful publication.

Writing research articles for professional journals is an art requiring good research skills, a clear sense of problem, and strong writing and editing skills. Assuming that years of graduate school have provided good research skills, I focus on the other requirements of writing a research article. My advice reflects the issues I most often raise when I review articles and 30 years of experience writing (and revising) research articles. I review guidelines for the major sections of the typical empirical research report and conclude with some suggestions about writing professionally. The emphasis is on writing for *Journal of Marriage and Family (JMF)*, but the general principles apply across journals and substantive areas.

WORKING THROUGH A RESEARCH PAPER

The format for a research paper is not set in stone. Each research problem is different, and

Department of Sociology, University of Nebraska—Lincoln,
Lincoln, NE 68583-0124 (lwhite@unl.edu).

Key Words: research, theory, writing.

Writes of Passage: Writing an Empirical Journal Article

The organization of the paper will depend on whether it is exploratory research rather than theory testing. In addition, authors have some latitude in developing a personal style. Generally, however, each article needs an introduction, a literature review, a statement of the problem, description of method, results, and conclusion. The organization of the piece, the titles of various sections, and the relative weight of these sections vary from paper to paper and from journal to journal, but some general guidelines apply to reports of qualitative and quantitative research.

Abstract

An abstract should summarize your study. In a few short sentences, it should state the research hypothesis, the sample, sample size, data used, and the findings. A starting sentence such as "Using data from a national sample of n women interviewed by telephone in 2002, we examine the relationship between x and y " will allow you to squeeze a lot of information into a few words. In a bare-bones fashion, without hyperbole or exaggeration, state the findings of the study. Examine prior issues of your target journal for abstract style and be sure to comply with the maximum length specified by the journal (120 words for *JMF*).

Introduction

The introduction is critical to capturing the reader's attention and setting the tone for the paper. In approximately a single page, it should specify the research question, the data to be used, and the strengths of the design, and it

The stab command

Syntax: `stab using Briatte_Petev_1, replace...`

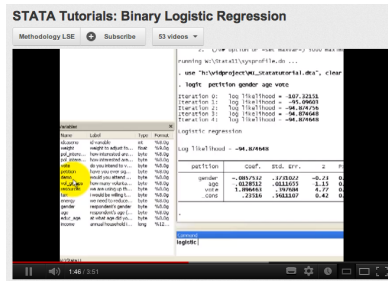
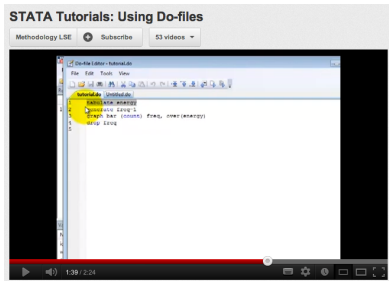
- `sum()` summarizes continuous variables
- `fre()` summarizes categorical variables
- `by()` creates multiple tables for comparison

Add the `corr` option to also export a correlation matrix.

`use datasets/nhis2009, clear`

```
stab using Briatte_Petev_1, replace ///  
    sum(age weight height) corr ///  
    fre(sex uninsured health) ///  
    by(regionbr)
```

Stata video tutorials



Source: LSE Methodology Institute, 2012.

Thanks for your attention

Project

- Correct and improve first draft
- Finalize association tests and interpretations
- Name your paper (**PDF**) and do-file like **Briatte_Petev_2**
- **OLS results** are optional in Draft No. 2

Readings

- *Stata Guide*, Sec. 10–11, 13–15
- *Making History Count*, ch. 4