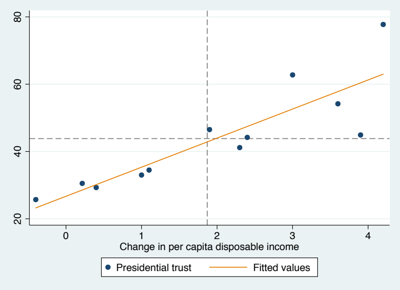
(Simple) Linear regression



Statistical Reasoning and Quantitative Methods

François Briatte & Ivaylo Petev

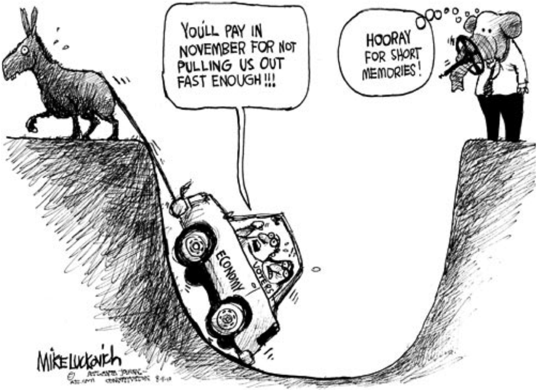
Session 9

Outline

Graphical approach

Regression output

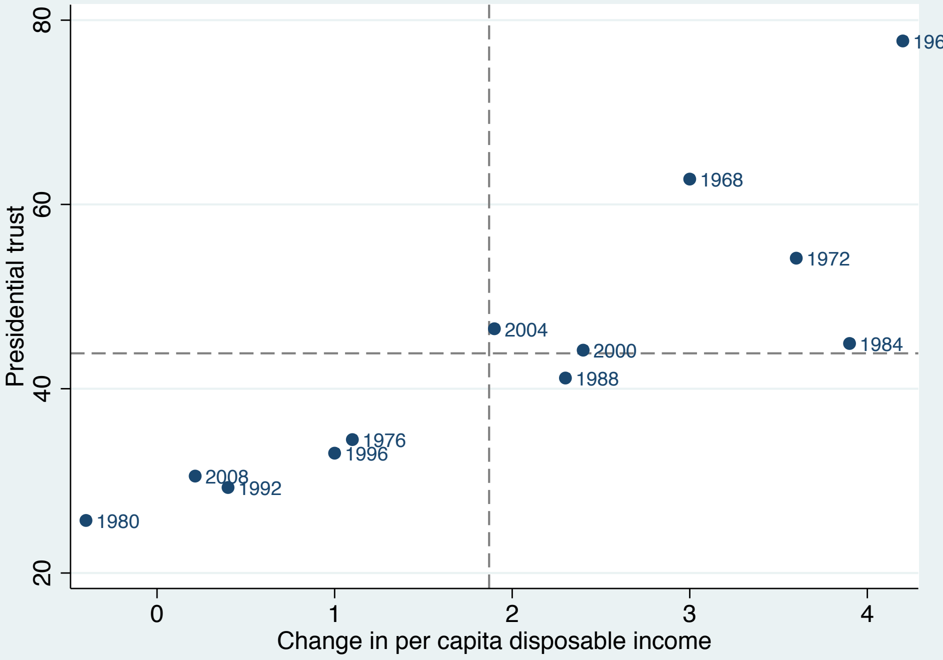
Assignment No. 2



Presidential approval and economic performance

- **Presidential approval**: "Always/Somewhat trustworthy" single measurement (ANES).
- **Economic performance**: change in disposable income per capita.
- To what extent can presidential approval be predicted from variations in disposable income?

Example provided by John Sides, using data by Douglas Hibbs.


Presidential trust
80
60
40
20
0
1
2
3
4
Change in per capita disposable income
1964
1968
1972
2004
2000
1984
1988
1976
1996
2008
1992
1980

Fitting a regression model

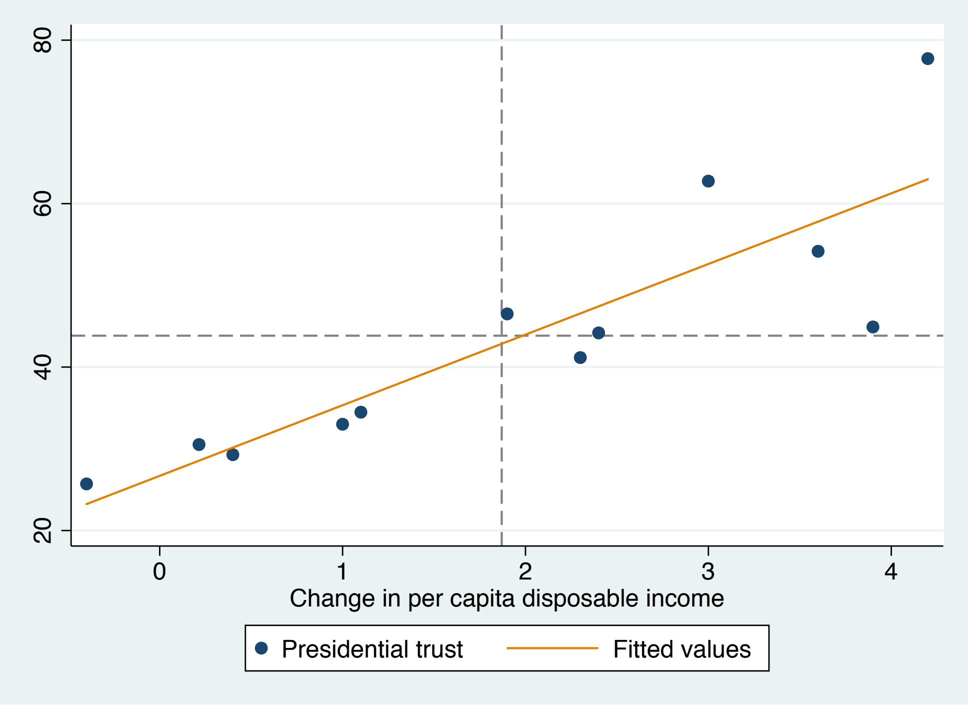
The model fits a **linear function** to the data, of the form:

$$Y = \alpha + \beta X + \epsilon \text{ or identically } \hat{Y} = \alpha + \beta X$$

where:

- Y is the **dependent variable** (response)
- X is the **independent variable** (predictor)
- α is the **constant** (intercept)
- β is the **regression coefficient** (slope)
- ϵ is the **error term** (residuals)

Note: the model assumes that the relationship is **linear**.


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

Fitting the regression line

The **regression coefficient** b is calculated as to **minimize** the **residual sum of squares** (RSS): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where Y_i is a data point and $\hat{Y}_i$ is the corresponding point on the regression line.

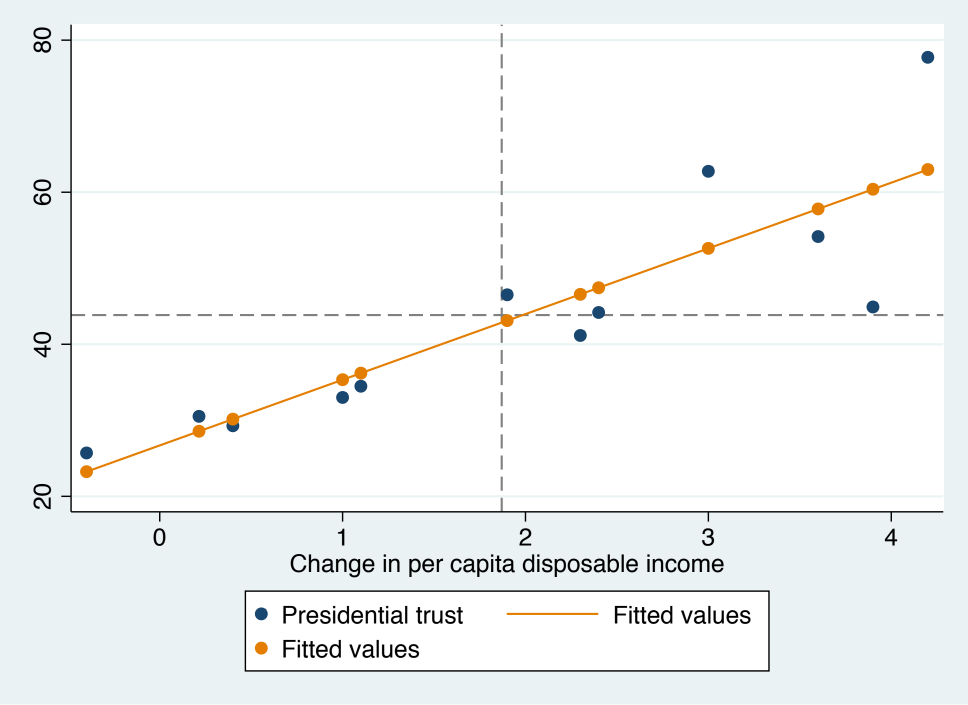
$$\beta = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

and

$$\alpha = \bar{Y} - \beta\bar{X}$$

Reminders:

- $\bar{X}_i$ is the **mean** of X, $\sum_{i=1}^{n}(X_i - \bar{X}_i)^2$ the **variance** of X.
- $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ is the **covariance** of XY.


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

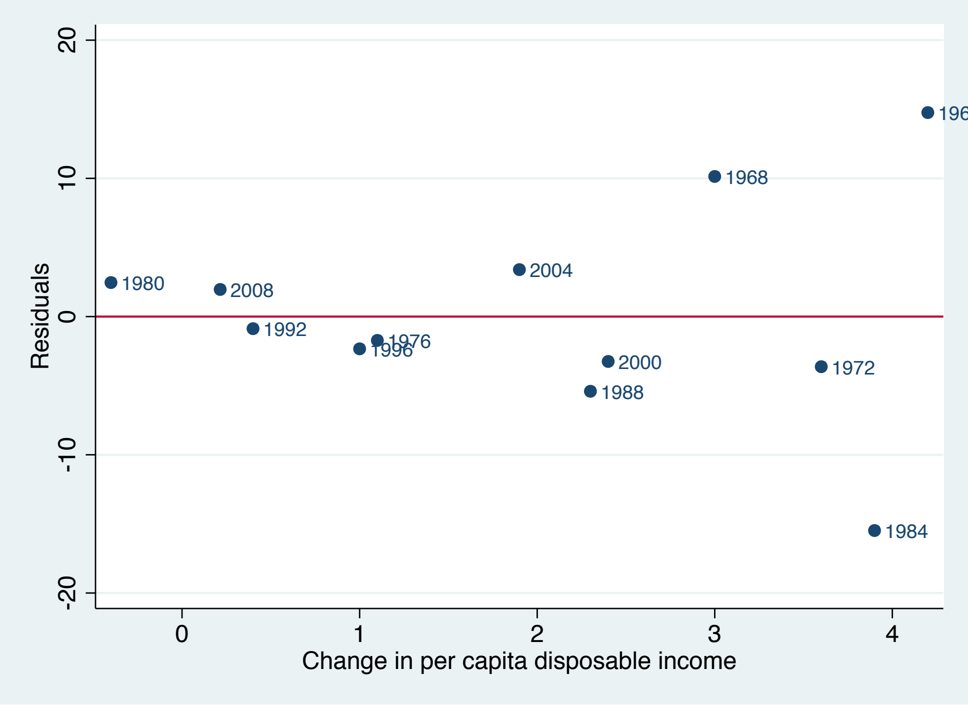
Goodness of fit

The **goodness of fit** of the model is provided by its **coefficient of determination**, R^2, which is the ratio between

- the variance predicted by the model, $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_i)^2$, and
- the residuals, or unpredicted variance, $\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2$.

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

- As the residual sum of squares (RSS) $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \to 0$, the coefficient of determination $R^2 \to 1$ towards higher goodness of fit.
- Goodness of fit is a **theoretical notion** that eventually relies on **substantive explanation**. No theory, no model.


Residuals
Change in per capita disposable income
20
10
0
-10
-20
0
1
2
3
4
196
1968
2004
1980
2008
1992
1976
1996
2000
1988
1972
1984

Model fit

Remember that your model needs to be **theoretically and empirically supported**:

- Theoretically, past economic performance relates to presidential approval by virtue of retrospective voting theory.
- Empirically, economic performance is a better predictor of presidential approval at lower values.

Always, always run a full intellectual check of your model after marvelling (or weeping) at your regression output:

- The direction of the causal link from X to Y should be deduceable through logical implication.
- The extent to which X influences Y must be interpreted and exemplified through data inspection.

Regression output

```
. 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
------------------------------------------------------------------------------
```

Reading guide (requires practice): top: ANOVA table (left) and model fit (right: p-value, R^2); bottom: regression coefficients.

Overall model fit

Model fit is provided by the R^2, calculated on N observations. Model significance tests the model against the null hypothesis.

```
. 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
------------------------------------------------------------------------------
```

- The **number of observations** determines your ability to generalize the model to the full sample or population.
- The F-**statistic** Its probability level tests the null hypothesis for your model, according to which all model coefficients are equal to 0.

```
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
```

Regression coefficients

Regression coefficients are unit-less variations of Y predicted by a change in one unit of X, as in $Y = aX + b$.

```
. 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** a is the slope of the regression line and its **constant** b the coordinate of origin (or intercept), i.e. $b = \hat{Y}_{X=0}$.
- The **standard error, t-value and p-value** tests whether the coefficient is significantly different from 0.

Assignment No. 2

Univariate statistics

- Introduction
- Datasets
- Distributions
- Estimation

Assignment No. 1

corrected
revised
appended

Bivariate statistics

- Significance
- Crosstabulation
- Correlation
- Linear regression

Assignment No. 2

Statistical modelling

- Basics
- Extensions
- Diagnostics
- Conclusion

Final paper

How to proceed

- **Revise Assignment No. 1** using instructor feedback:
  - Read all corrected material.
  - Proceed to required adjustments.
  - Append new research to the text.
- **Explore associations** using crosstabulations and comparisons:
  - Find and/or recode variables to crosstabulate their categories.
  - Find and/or recode variables to compare means and proportions.
  - Keep working with continuous and interval variables.
- **Model relationships** using correlations and linear regression:
  - Produce a correlation matrix.
  - Regress independent variables on the dependent variable.
  - Regress collinear independent variables.

Step 1: Revision

- Adjust your research design:
  - **Select variables** with a sufficient number of observations.
  - **Devise clear hypotheses** $H_1, H_2, \ldots$ to prepare for modelling.
  - **Reformulate all text** to fit scientific presentation.
- Adjust your do-file:
  - **Use comments** to structure and explain your methods.
  - **Clean up** unnecessary code like `lookfor` and `codebook` commands.
  - **Replicate** your edited do-file to update the log file, graphs and tables.
- Chill out for a minute.

Step 2: Association

Associations look at contingency tables:

- `tab` with the `chi2` option performs a **Chi-squared test** on variables coded into categories with at least 5 cell counts.
- `tab` with the `exact` option performs **Fisher's exact test** on small crosstabulations (2×2 contigency tables or cell counts < 5).
- `ttest` and `prtest` compares means (**Fisher's *t*-test**) and proportions in two independent groups.

Associations depend on variable types:

- **Crosstabulations** use two categorical variables.
- **Comparisons** use one continuous and one categorical variable.
- **Correlations** use two continuous variables.

Correlation and simple linear regression are treated as preliminary steps to writing a **multiple regression model**.

Step 3: Model

- **Visually explore relationships** using scatterplots:
  - `sc` (`scatter`) draws scatterplots.
  - `gr mat` draws a scatterplot matrix.
  - `tw` (`twoway`) combines scatterplots.
- **Formally explore relationships** using correlations:
  - `pwcorr` (pairwise correlation) works with any number of variables.
  - Use the `obs` (observations) and `sig` (significance) options.
  - Reproduce the correlation matrix as a table in your work.
- **Model relationships** using simple linear regression:
  - `reg` (`regress`) does all the work.
  - `predict r, r` stores the model residuals.
  - `rvfplot` plots the residuals against fitted values.

  Regress the **dependent variable** on the main independent variable, and also regress **collinear independent variables** on each other.

Further help

- Course-specific help:
  - Stata Guide
  - Session do-files
  - Course slides
- General help:
  - Handbook chapters
  - Stata documentation (`help` *`command`*)
  - Online tutorials

Handbook chapters and course emails are available from the ENTG. Everything else is systematically archived on the course website:

`http://f.briatte.org/teaching/quanti/`

Happy coding!