Statistical Reasoning

Week 8

Sciences Po - Louis de Charsonville

Spring 2018

Outline

Research Paper

Correlation

Simple linear regression

Practice

Research Paper

Research Paper

Timeline

1^{st} draft	Done
Coming weeks	**Improve the 1^{st} draft based on feedback.**
2^{nd} draft	**10 April**
Final draft	**24 April**

Feedback

Research

- Choose multiple independent variables, not just one.
- Discuss your findings.
- Question your hypotheses.
- Do not oversell your work. Be humble and specific.

Coding

- Code should run.
- Graphs should not be overwritten.

Writing

- Avoid general statements, be accurate.
- Use scientific term, *normal* means the variable is following the normal distribution.
- Avoid jargon and subjective terms.
- If you include graphs, tables, always *comment* them.

Outline for do-file

1. DV Choice
   - Summary statistics
   - Variable manipulation (rename / recode)
   - Visualisation
2. IV Choice
   - Summary statistics
   - Recode & Visualisation
3. Dealing with missing values
4. DV : further analysis
   - Normality tests (the more the better)
   - Transformation → normality tests agains (+ discussion).
   - Exploration of hypothesis : first intuitions by display DV over IV's.

Correlation

Correlation - Reminder 1/2

What it does ?

- Measure association as the linear dependence of two variables
- Used to examin the **strength of association** between two **quantitative variables**

Descriptive statistics

- Visualize the correlation by creating a scatterplot ;
- Identify the strengh of the correlation by calculating a Pearson'R

Inferential statistics

- Significance test using a **t-test** for Pearson's R

Correlation - Reminder 2/2

Positive vs Negative correlation

- A **positive correlation** indicates that the values on the two variables being analyzed move in the same direction.
- A **negative correlation** indicates that the values on the two variables being analyzed move in opposite directions

Strength of relationship - Rule of thumb

- Perfect correlation : $|r| = 1$
- High : $|r| \geq 0.7$
- Moderate : $0.3 \leq |r| \geq 0.7$
- Low : $|r| \leq 0.3$

Compute Pearson's Correlation coefficient

Formula

Population

$$\rho = \frac{Cov(X,Y)}{Var_X Var_Y} \quad (1)$$

Sample

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

Remember

- Pearson's correlation coefficient detects **linear correlation**
- Uncorrelated $\neq$ unrelated
- Correlated $\neq$ unconfounded

Covariance

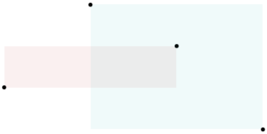
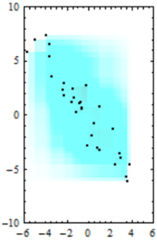
Mathematical formula

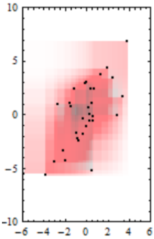
$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (3)$$

In plain language

- How changes in one variable are associated with changes in a second variable
- Degree of linear association

Graphically



Significance Test

Significance test

- Null hypothesis $H_0 : r = 0$
- Test statistic $T = r\sqrt{\frac{n-2}{1-r^2}}$
- Test the probability of getting a correlation coefficient different from zero (*if* H_0 *were true*

Stata Command

- Add the `sig` option to `pwcorr` : `pwcorr y x, sig`
- Add a star if significant at the α : `pwcorr y x, star(0.05)`

Visualise the correlation

Stata

- Scatter plot : `sc x y` or `plot x y`

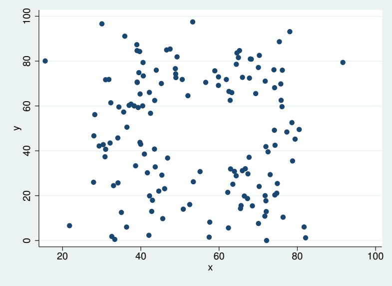
Visualisation is important !

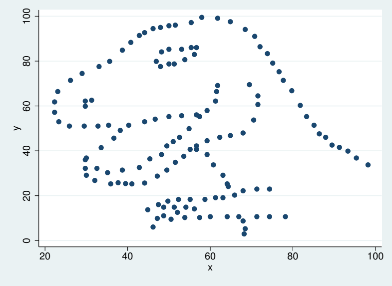
Visualise the correlation

Stata

- Scatter plot : `sc x y` or `plot x y`

Visualisation is important !





Matrix graphs

Stata

- Plot matrix graphs `gr mat y x z, half`


```
gr mat wdi_fr bl_asy25mf wdi_gdpc undp_hdi ti_cpi, half scheme(plottig) mcolor(plr1) scale(0.8)
```

Coefficient of determination

Coefficient of determination

- $R^2 = \rho^2$
- R^2 reflects the percentage of variance explained in each of the two correlated variables by the other variable.

In Stata

- `pwcorr y x`
- `di r(rho)^2`

Correlation does not imply causation

- Correlations can exist without a cause and effect relationshiph between the variables
- A correlation can exist :
  - X is causing Y
  - Y is causing X (*reverse causality*)
  - Z is causing both X and Y (*missing variable*)
  - Random chance !
- **Theoretical explanations** are critical to understand the correlations observed.

Simple linear regression

Simple linear regression

- Statistical technique closely related to correlations
- Extension of correlation
- DV needs to be quantitative and continuous

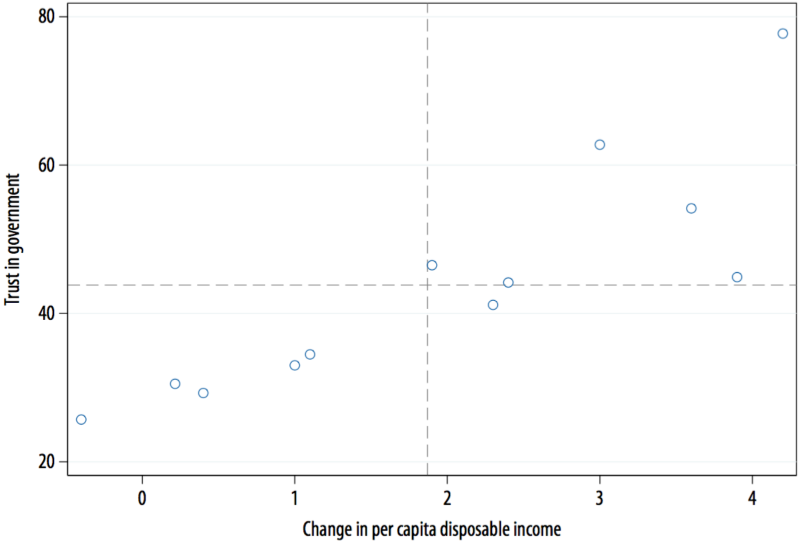
Goals

- Provide **direction** of the relationship and **strength**
- **Statistical significance**
- **Explanatory power of the independent variable**
  - To what extent the total variation of the dependent variable can be explained by the variation of the independent variable
- Prediction

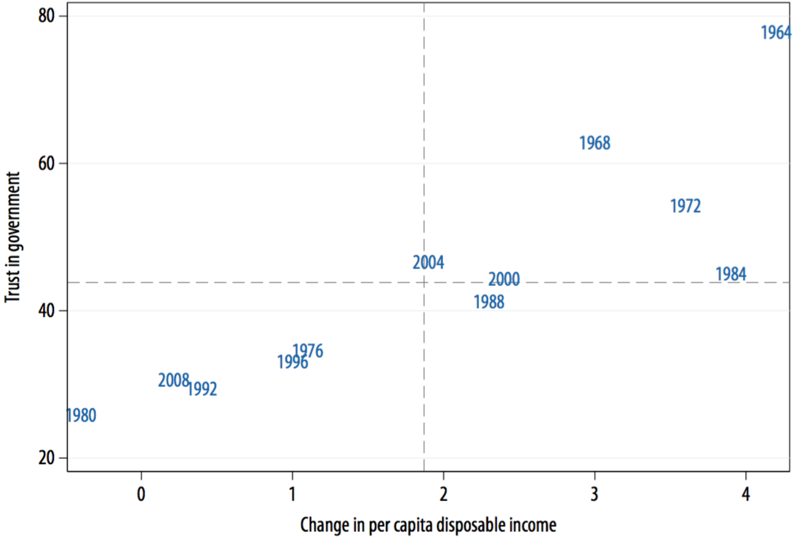
Example : Trust and Economic performance

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

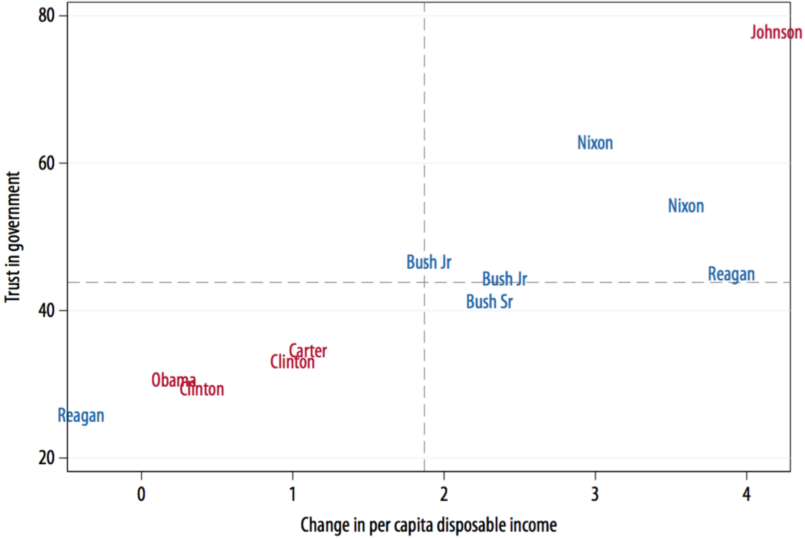
- **Dependent Variable** : Trust in Government
  - Share of respondents answering "Just about always / Most of the time"
- **Independent Variable** : Economic performance
  - Change in per capita disposable income



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 $p < .01$.

Maths behind the hood

Equations

$$Y = \alpha + \beta X + \epsilon$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X$$

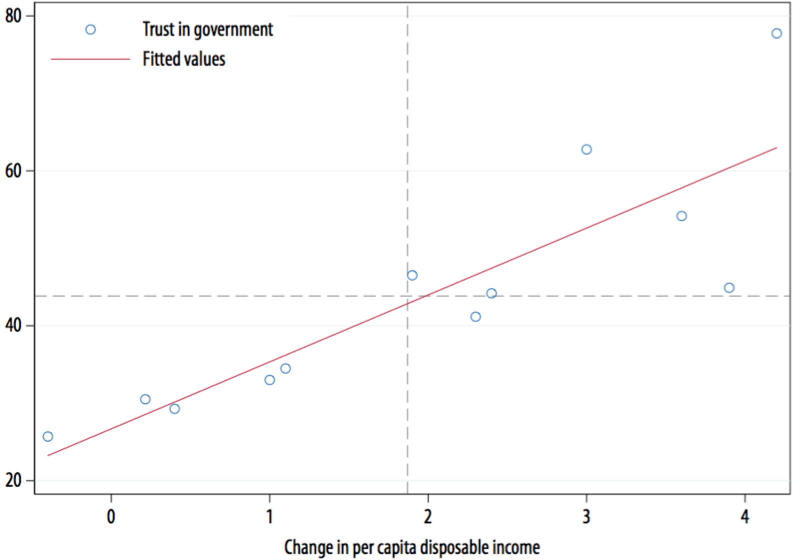
$$\epsilon = Y - \hat{Y}$$

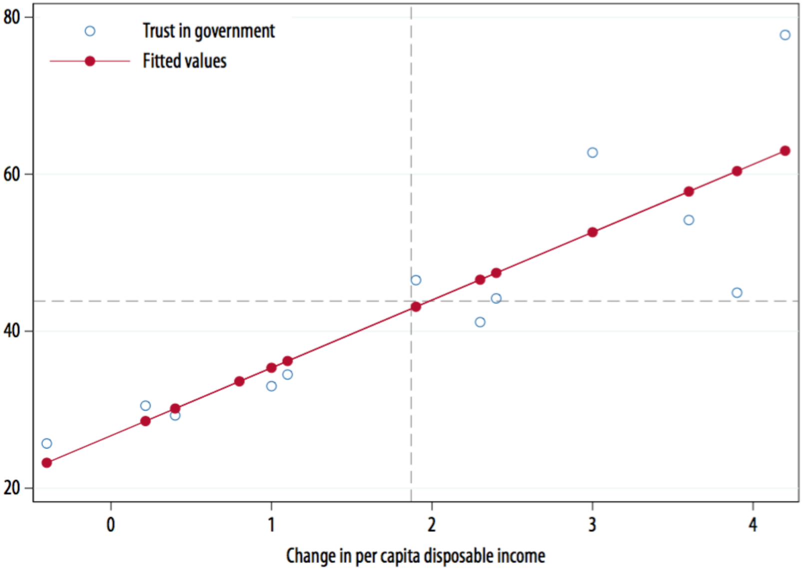
Parameters

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

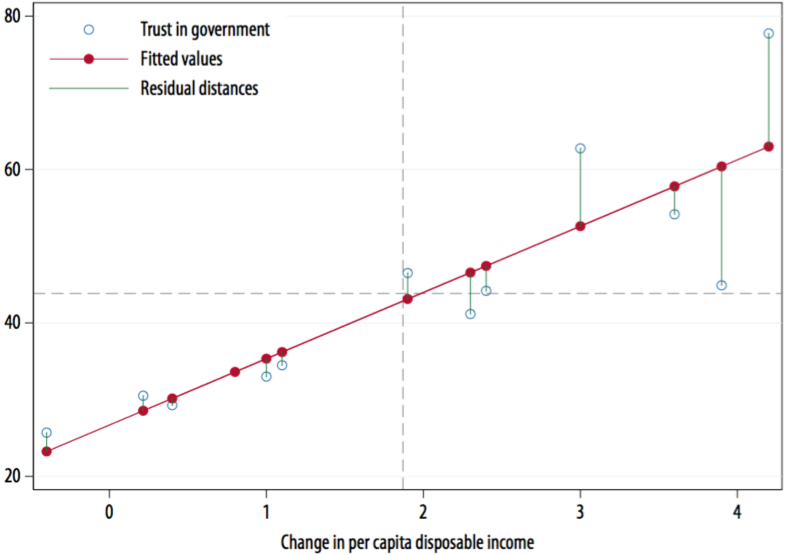
Warning

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



80
60
40
20
Trust in government
Fitted values
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

Finding the regression line

- **Goal** : Find the line of best fit.
- Solution : minimize the error term

Finding the regression line

- **Goal** : Find the line of best fit.
- Solution : minimize the error term

Ordinary Least Squares

1. We minimize the sum of squared **residuals**.

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

2. Get β

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

3. Get α

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

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

- Goodness of fit or R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} \tag{4}$$

  - As the fit improves $RSS \to 0$ and $R^2 \to 1$.
- A regression coefficient estimates the variation in Y predicted by a change in one unit of X
- The **coefficient** is the slope β of the regression line
- The **constant** is the intercept of the regression line
- The **standard error**, *t*-value and *p*-value test whether the coefficient is significantly different from 0.

Interpreting the results 2/2

- Total number of observations
- *F*-value and *p*-value associated with F statistic which tests the null hypothesis that all of the model coefficients are equal to zero
- RMSE is **Root Mean Squared Errors** is the standard deviation of the residuals.

Other relationship

Linear-linear relationship $Y = \alpha + \beta X$

An increase in one unit of X is associated with an increase of β units of Y.

Log-linear relationship $\ln Y = \alpha + \beta X$

An increase in one unit of X is associated with an $100 * \beta\%$ increase in Y.

Linear-log relationship $Y = \alpha + \beta \ln X$

A 1% increase in X is associated with an increase of 0.01β units of Y.

Log-log relationship $\ln Y = \alpha + \beta \ln X$

A 1% in X is associated with an increase of $\beta\%$ in Y.

Practice

Practice

Fertility and Education, Part 1 & 2

1. Finish `week7.do`
   - Remember to comment
     `run setup/require mkcorr renvars`
2. Do `week8.do`