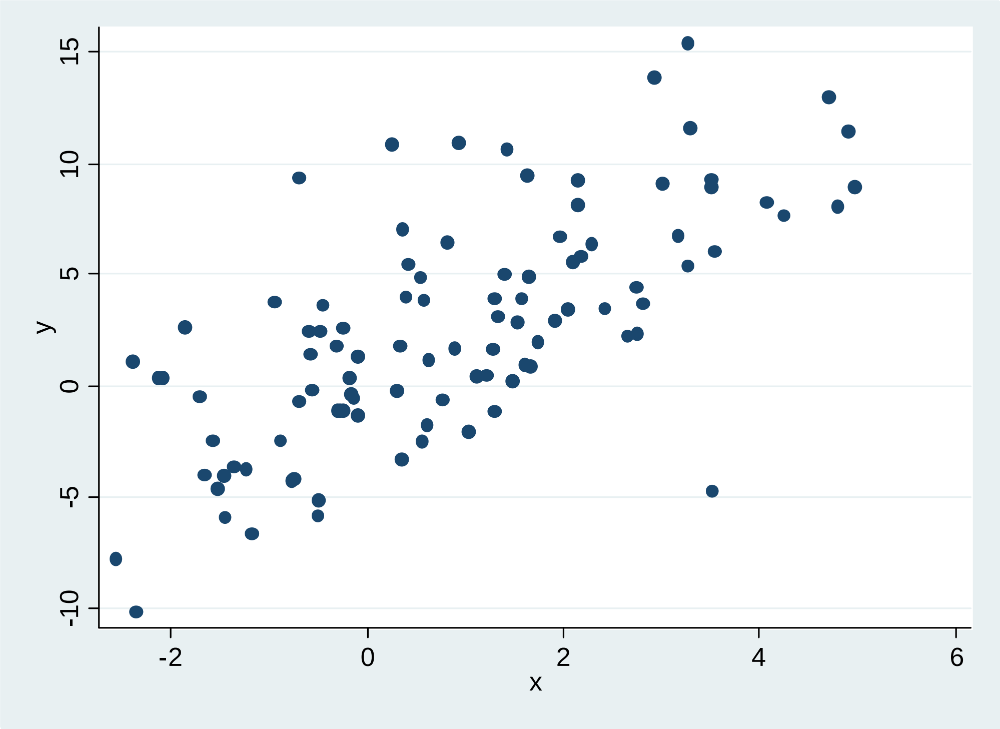
The simple linear Regression Model

- Correlation coefficient is non-parametric and just indicates that two variables are associated with one another, but it does not give any ideas of the kind of relationship.
- Regression models help investigating bivariate and multivariate relationships between variables, where we can hypothesize that 1 variable depends on another variable or a combination of other variables.
- Normally relationships between variables in political science and economics are not exact – unless true by definition, but relationships include most often a non-structural or random component, due to the probabilistic nature of theories and hypotheses in PolSci, measurement errors etc.
- Regression analysis enables to find average relationships that may not be obvious by just „eye-balling“ the data – explicit formulation of structural and random components of a hypothesized relationship between variables.
- Example: positive relationship between unemployment and government spending

Simple linear regression analysis

- Linear relationship between x (explanatory variable) and y (dependent variable)
- Epsilon describes the random component of the linear relationship between x and y

$$y_i = \alpha + \beta * x_i + \varepsilon_i$$



$$y_i = \alpha + \beta * x_i + \varepsilon_i$$

- Y is the value of the dependent variable (spending) in observation i (e.g. in the UK)
- Y is determined by 2 components:

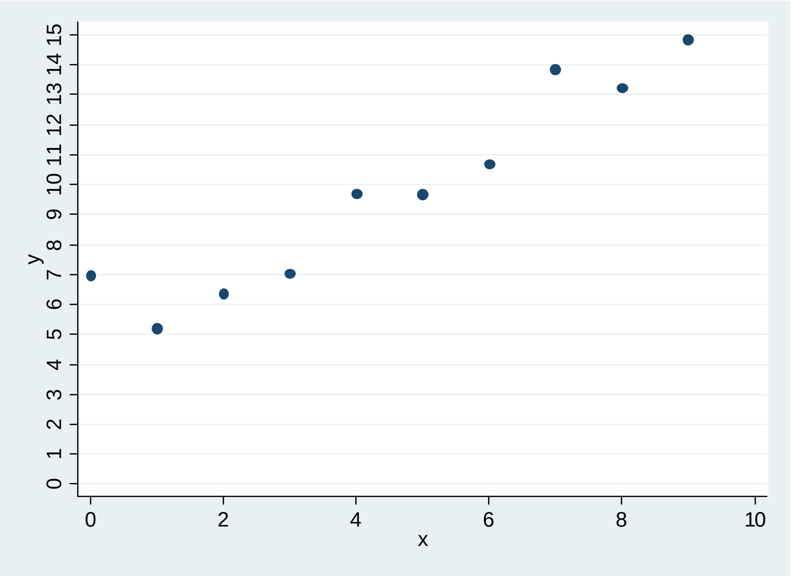
1. the non-random/ structural component alpha+beta*xi – where x is the independent/ explanatory variable (unemployment) in observation i (UK) and alpha and beta are fixed quantities, the parameters of the model; alpha is called constant or intercept and measures the value where the regression line crosses the y-axis; beta is called coefficient/ slope, and measures the steepness of the regression line.
2. the random component called disturbance or error term epsilon in observation i

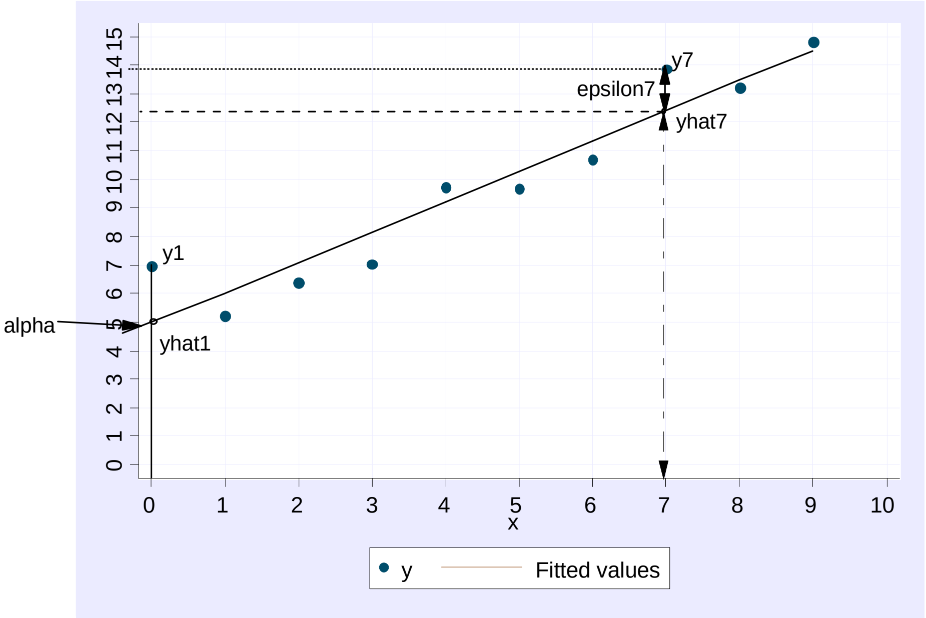
A simple example:

- x has 10 observations: 0,1,2,3,4,5,6,7,8,9
- The true relationship between y and x is: y=5+1*x, thus, the true y takes on the values: 5,6,7,8,9,10,11,12,13,14
- There is some disturbance e.g. a measurement error, which is standard normally distributed: thus the y we can measure takes on the values: 6.95,5.22,6.36,7.03,9.71,9.67,10.69,13.85, 13.21,14.82 – which are close to the true values, but for any given observation the observed values are a little larger or smaller than the true values.
- the relationship between x and y should hold on average true but is not exact
- When we do our analysis, we don‘t know the true relationship and the true y, we just have the observed x and y.
- We know that the relationship between x and y should have the following form: y=alpha+beta*x+epsilon (we hypothesize a linear relationship)
- The regression analysis „estimates“ the parameters alpha and beta by using the given observations for x and y.
- The simplest form of estimating alpha and beta is called ordinary least squares (OLS) regression

OLS-Regression:

- Draw a line through the scatter plot in a way to minimize the deviations of the single observations from the line:





- Minimize the sum of all squared deviations from the line (squared residuals)

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} * x_i + \hat{\varepsilon}_i \Rightarrow \hat{\varepsilon}_i = y_i - \left(\hat{\alpha} + \hat{\beta} * x_i \right)$$

- This is done mathematically by the statistical program at hand
- the values of the dependent variable (values on the line) are called predicted values of the regression (yhat): 4.97,6.03,7.10,8.16,9.22, 10.28,11.34,12.41,13.47,14.53 – these are very close to the „true values“; the estimated alpha = 4.97 and beta = 1.06

OLS regression

Ordinary least squares regression: minimizes the squared residuals

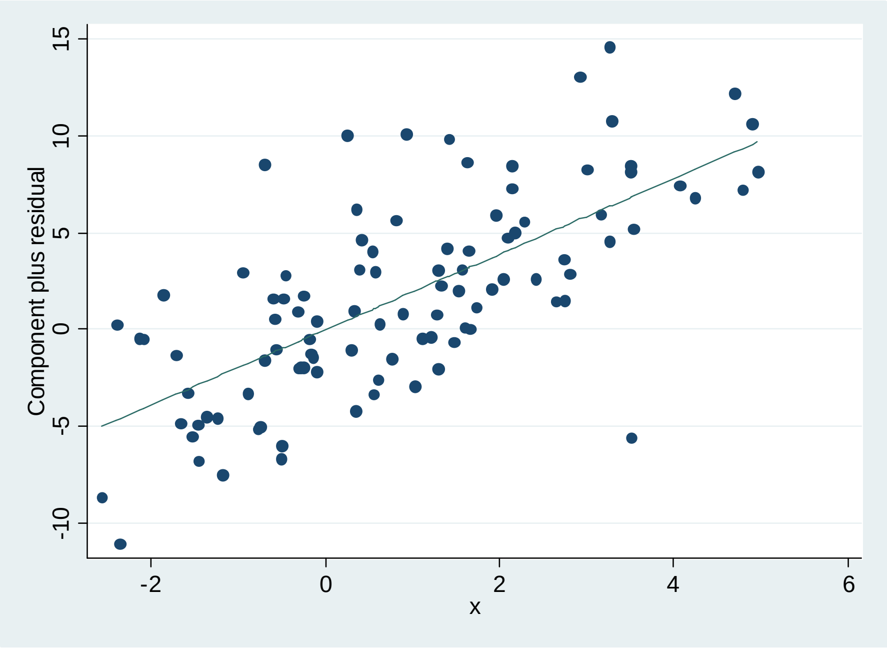
$$y_i = \alpha + \beta * x_i + \varepsilon_i \quad \Rightarrow \quad \varepsilon_i = y_i - \alpha - \beta * x_i$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} * x_i + \hat{\varepsilon}_i$$

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

Components:

- DY: y; at least 1 IV: x
- Constant or intercept term: alpha
- Regression coefficient, slope: beta
- Error term, residuals: epsilon



Derivation of the OLS-Parameters alpha and beta:

The relationship between x and y is described by the function:

$$y_i = \alpha + \beta * x_i + \varepsilon_i$$

The difference between the dependent variable y and the estimated systematic influence of x on y is named the residual:

$$e_i = y_i - \hat{\alpha} - \hat{\beta} * x_i$$

To receive the optimal estimates for alpha and beta we need a choice-criterion; in the case of OLS this criterion is the sum of squared residuals: we calculate alpha and beta for the case in which the sum of all squared deviations (residuals) is minimal

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} (e_i)^2 \quad \Rightarrow \quad \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} * x_i\right)^2 = S\left(\hat{\alpha}, \hat{\beta}\right)$$

Taking the squares of the residual is necessary since a) positive and negative deviation do not cancel each other out, b) positive and negative estimation errors enter with the same weight due to the squaring down, it is therefore irrelevant whether the expected value for observation yi is underestimated or overestimates

Since the measure is additive no value is of outmost relevance.

Especially large residuals receive a stronger weight due to squaring.

Minimizing the function requires to calculate the first order conditions with respect to alpha and beta and set them zero:

$$\mathrm{I}: \ \frac{\partial S\left(\hat{\alpha},\hat{\beta}\right)}{\partial\hat{\alpha}} = -2\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta} * x_i\right) = 0$$

$$\mathrm{II}: \ \frac{\partial S\left(\hat{\alpha},\hat{\beta}\right)}{\partial\hat{\beta}} = -2\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta} * x_i\right)x_i = 0$$

This is just a linear system of two equations with two unknowns alpha and beta, which we can mathematically solve for alpha:

$$\mathrm{I}: \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta} * x_i\right) = 0$$

$$\Rightarrow \hat{\alpha} = \sum_{i=1}^{n}\left(y_i - \hat{\beta} * x_i\right) \Rightarrow \hat{\alpha} = \overline{y} - \hat{\beta} * \overline{x}$$

… and beta:

$$\text{II}: \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta} * x_i\right)x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - \hat{\alpha}x_i - \hat{\beta}x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - \left(\overline{y} - \hat{\beta}\overline{x}\right)x_i - \hat{\beta}x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - \overline{y}x_i + \hat{\beta}\overline{x}x_i - \hat{\beta}x_i^2 = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}\left(y_i - \overline{y} + \hat{\beta}\overline{x} - \hat{\beta}x_i\right)x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \overline{y} + \hat{\beta}\left(\overline{x} - x_i\right) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n}\left(y_i - \overline{y}\right) = -\hat{\beta}\sum_{i=1}^{n}\left(\overline{x} - x_i\right)$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)}$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = X'X^{-1}X'y$$

Naturally we still have to verify whether $\hat{\alpha}$ and $\hat{\beta}$ really minimize the sum of squared residuals and satisfy the second order conditions of the minimizing problem. Thus we need the second derivatives of the two functions with respect to alpha and beta which are given by the so called Hessian matrix (matrix of second derivatives). (I spare the mathematical derivation)

The Hessian matrix has to be positive definite (the determinant must be larger than 0) so that $\hat{\alpha}$ and $\hat{\beta}$ globally minimize the sum of squared residuals. Only in this case alpha and beta are optimal estimates for the relationship between the dependent variable y and the independent variable x.

Regression coefficient:

$$\hat{\beta}_{yx} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Beta equals the covariance between y and x divided by the variance of x.

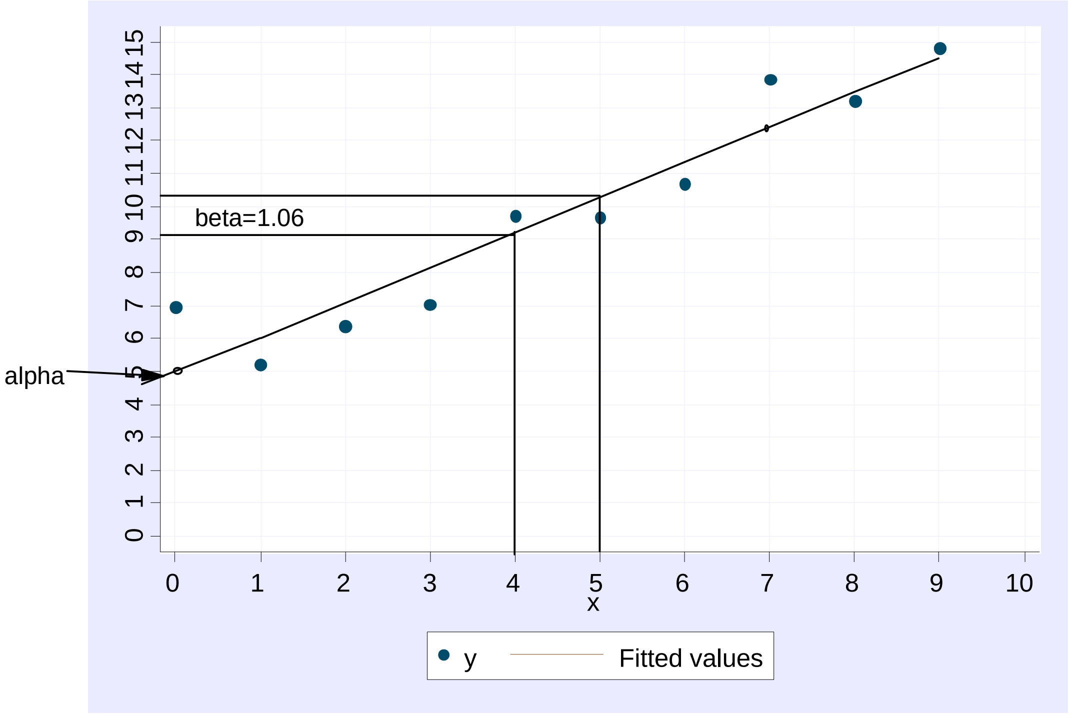
Interpretation of regression results:

```
 reg y x

  Source |       SS           df       MS                    Number of obs =   100
-------------+---------------------------------------        F(  1,    98)      =   89.78
    Model |  1248.96129     1     1248.96129                 Prob > F           =   0.0000
 Residual |  1363.2539     98     13.9107541                 R-squared          =   0.4781
-------------+---------------------------------------        Adj R-squared  =   0.4728
     Total |  2612.21519   99     26.386012                  Root MSE           =   3.7297

------------------------------------------------------------------------------
         y |      Coef.         Std. Err.       t       P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
         x |   1.941914     .2049419     9.48   0.000     1.535213    2.348614
     _cons |    .8609647   .4127188     2.09   0.040     .0419377    1.679992
------------------------------------------------------------------------------
```

If x increases by 1 unit, y increases by 1.94 units: the interpretation is linear and straightforward

Interpretation: example

- alpha=4.97, beta=1.06
- Education and earnings: no education gives you a minimal hourly wage of around 5 pounds. Each additional year of education increases the hourly wage by app. 1 pound:



Properties of the OLS estimator:

- Since alpha and beta are estimates of the unknown parameters, $\hat{y}_i = \hat{\alpha} + \hat{\beta} * x_i$ estimates the mean function or the systematic part of the regression equation. Since a random variable can be predicted best by the mean function (under the mean squared error criterion), yhat can be interpreted as the best prediction of y. the difference between the dependent variable y and its least squares prediction is the least squares residual: e=y-yhat =y-(alpha+beta*x).
- A large residual e can either be due to a poor estimation of the parameters of the model or to a large unsystematic part of the regression equation
- For the OLS model to be the best estimator of the relationship between x and y several conditions (full ideal conditions, Gauss-Markov conditions) have to be met.
- If the „full ideal conditions“ are met one can argue that the OLS-estimator imitates the properties of the unknown model of the population. This means e.g. that the explanatory variables and the error term are uncorrelated.