VERGLEICHENDE SOZIALFORSCHUNG MIT MEHREBENENMODELLEN IN R

Forschungspraktikum I und II
Dr. Christian Czymara
Linear regression

AGENDA

- Basics of linear regression
- OLS estimator
- Statistical controlling
- Assumption of independence
- Exercise: Comparing public support for redistribution across countries

REGRESSION TERMINOLOGY

VARIABLE

- Variables are characteristics that vary across observations
- We model one variable y, so-called…
  - dependent variable
  - outcome
  - response variable
- as a function of some other variable(s) x, so called…
  - independent variable
  - explanatory variable
  - predictor variable
  - regressor
- Formally, this is denoted as: $y_i = \beta_0 + \beta_1 x_i$

ERROR

- With real-world data, y is never a perfect function of x
- Taking these “leftover“ differences into account:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- e_i includes the variance of y that is not explained by x, the so-called called…
  - error term
  - stochastic term

REGRESSION

- To regress y on x
- The function is: $y = \beta_0 + \beta_1 x_1$
- Which values for unknown parameters β_0 and β_1?
- Idea: chose in a way that the regression line is, loosely speaking, closest to all data points at the same time
- How?

→Minimize (squared) deviations of each data point from regression line

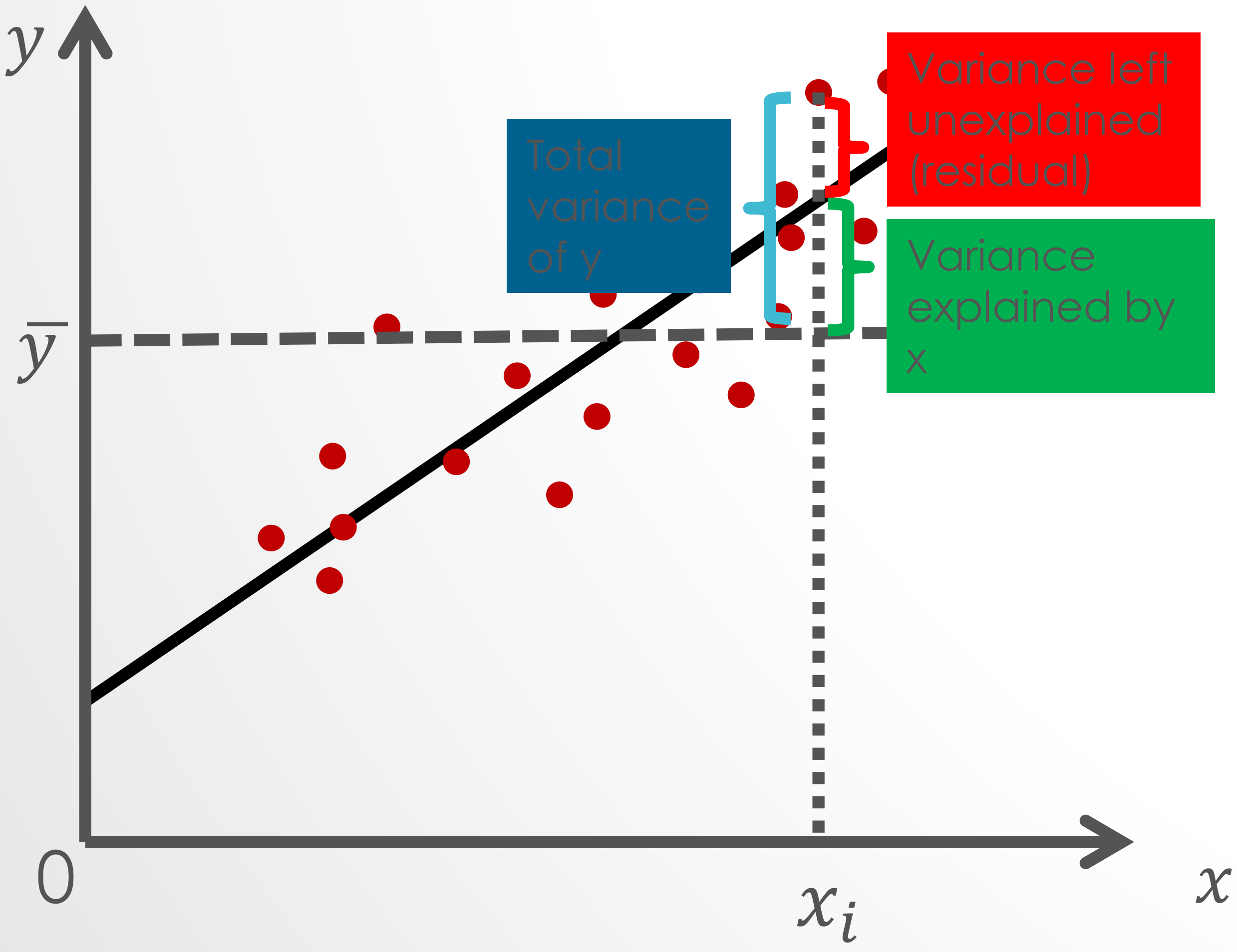
→*Ordinary Least Squares* (OLS)

THE ORDINARY LEAST SQUARES ESTIMATOR

ORDINARY LEAST SQUARES (OLS)

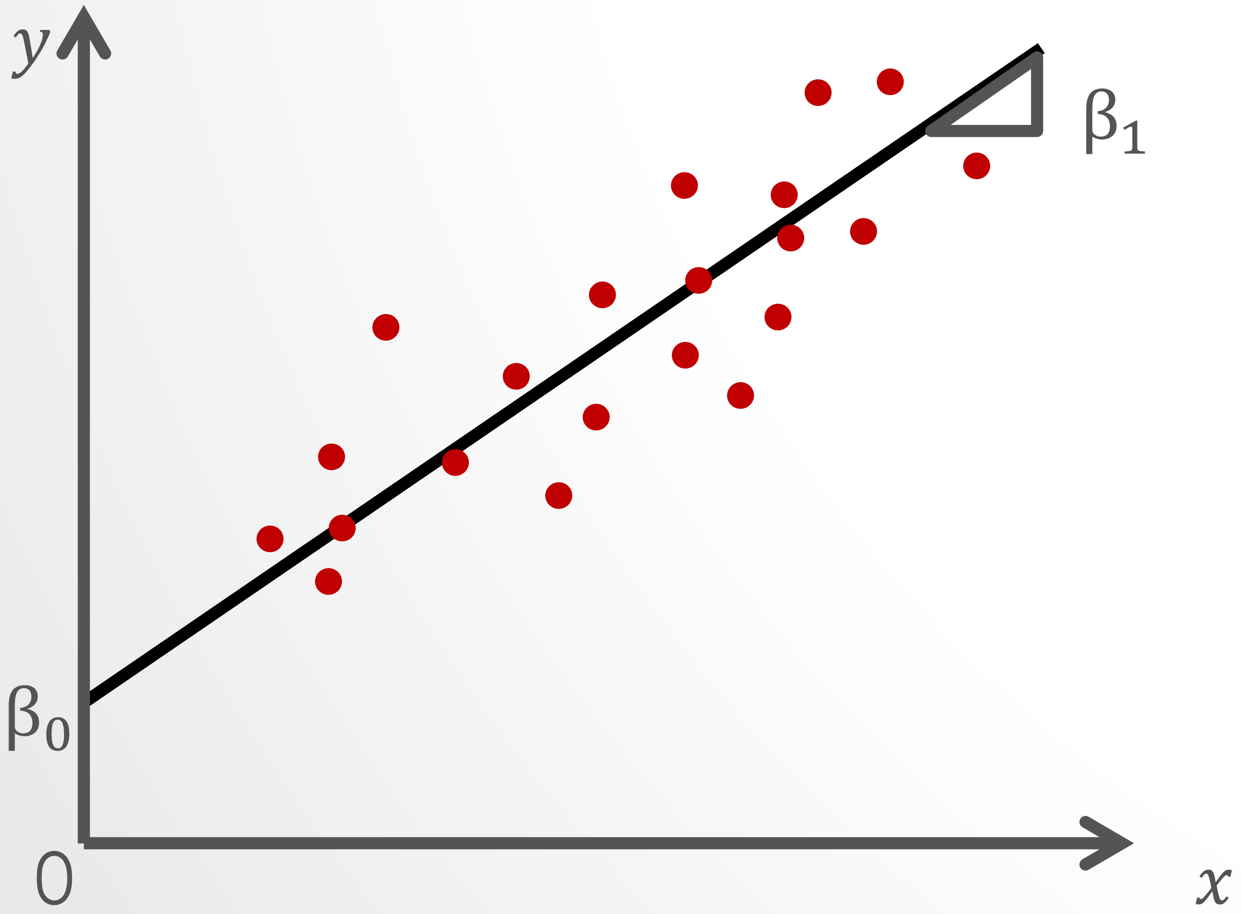
- Estimate β_0 and β_1 in a way that minimizes the squared residuals
- … aka the (squared) differences between the observed values (denoted as $\hat{y}$) and the predicted ones (denoted as y)
- OLS is the *Best Linear Unbiased Estimator* (BLUE), given certain assumptions
  - Correct model specification (no relevant x missing)
  - Strict exogeneity (x not correlated with error term)
  - Linear independency (x no linear functions of one another)
  - Uncorrelated errors
  - Homoscedasticity (errors equal across all x)
  - Normality (errors normally distributed conditional on x)

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$



→(squared and) minimized by OLS

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$



OLS REGRESSION IN R

- Regressing y on x: `lm(y ~ x, data)`

Estimated value for $\boldsymbol{\beta_0}$ (the *constant*, or intercept)

```
> summary(lm(y ~ x, data))

Call:
lm(formula = y ~ x, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7608 -0.6464  0.3250  0.4222  1.4507 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.5063966  0.0354849  98.814  < 2e-16 ***
x           0.0028581  0.0006827   4.187 2.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.026 on 6785 degrees of freedom
  (146 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.002577,  Adjusted R-squared:  0.00243 
F-statistic: 17.53 on 1 and 6785 DF,  p-value: 2.869e-05
```

Estimated value for $\boldsymbol{\beta_1}$ (the *effect of x*, or slope)

→ Estimated function: $\hat{y} = 3.51 + 0.003 * x$

INTERPRETATION

- In this example: *"A one unit increase in x implies 0.003 increase in y."*
- Predictions: replace x with the value of interest
- E. g.: x = 75 → 3.51 + 0.003 * 75 = 3.735

→*"A person with an x value of 75 is expected to have on average a y value of 3.735."*

MARGINAL EFFECTS

- The interest of linear regression is to examine how a change in x is associated with a change in y
- In econometrics, this is called a marginal effect
- It is the slope of the regression line
- The slope is the first derivate of the regression equation w. r. t. x
- Simplest case: $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$

→$\frac{\delta y}{\delta x} = \beta_1$

EXPLANATORY VARIABLES

TYPES OF VARIABLES

- Variables are measured on different scales
- Practically, the most important distinctions are between *continuous* and *discrete* variables (categorical)
- Different modes of analysis
  - Continuous: distribution of all values (for example, mean and standard deviation)
  - Discrete: for example, probability of single values (categories)
- Variables with (too) many categories: treat as continuous (when appropriate) or simplify into fewer categories

MODELLING CONTINUOUS VARIABLES

- Continuous variables can occupy any value over a continuous range
- Practically, an *underlying continuous construct* is sufficient
- As explanatory variables, they indicate *changes in y if x increases by one unit*
- Examples: income, age, temperature
- *"One more Euro income implies a 0.04 increase in life satisfaction."*

MODELLING CATEGORICAL VARIABLES: DUMMIES

- Dummy variables: binary variables (values: 0 & 1)
- Example: Explain welfare support with gender (*gndr, female=0, male=1)*

→$support = \beta_0 + \beta_1 gndr$

  - Support for gndr = *0* (female): $E(support|gndr = 0) = \beta_0$
  - Support for gndr = *1* (male): $E(support|gndr = 1) = \beta_0 + \beta_1$

- gndr = *0* is the *reference category* (the group of *comparison*)
- β_1 is the difference of men compared to women

… IN R

- In our example, gender has two categories → include one dummy

```
> summary(lm(y ~ gndr, data))

Call:
lm(formula = y ~ gndr, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7309 -0.7309  0.2691  0.4405  1.4405 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.73094    0.01750 213.185  < 2e-16 ***
gndrmale    -0.17147    0.02478  -6.919 4.97e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.024 on 6826 degrees of freedom
  (105 Beobachtungen als fehlend gelöscht)
Multiple R-squared:  0.006964,	Adjusted R-squared:  0.006819 
F-statistic: 47.87 on 1 and 6826 DF,  p-value: 4.969e-12
```

- Regression constant equals reference category (i. e.: women)
- Men support welfare 0.17 less than women

→$support = 3.73 - 0.17 * gndr$

DUMMY MEANS

```
> aggregate(y ~ gndr, data = data, mean)
    gndr        y
1 female 3.730938
2   male 3.559471
```

EXPLANATORY VARIABLES WITH MULTIPLE CATEGORIES

- With k categories, include $k - 1$ dummies in the model
- Omitted dummy becomes reference category
- Choice of reference category (statistically) irrelevant
- (Why? → Otherwise perfect collinearity (all dummy categories summed up always equal 1)

… IN R

```
> table(data$health)

     bad      fair      good  very bad very good 
     400      1637      3119        94      1677 
> summary(lm(y ~ relevel(factor(health), ref = "very bad"), data))

Call:
lm(formula = y ~ relevel(factor(health), ref = "very bad"), data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1209 -0.6006  0.3994  0.4819  1.4819 

Coefficients:
                                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                          4.1209     0.1070  38.510  < 2e-16 ***
relevel(factor(health), ref = "very bad")bad        -0.2952     0.1188  -2.484  0.01300 *  
relevel(factor(health), ref = "very bad")fair       -0.3315     0.1100  -3.014  0.00258 ** 
relevel(factor(health), ref = "very bad")good       -0.5203     0.1086  -4.792 1.69e-06 ***
relevel(factor(health), ref = "very bad")very good  -0.6028     0.1099  -5.484 4.30e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Reference category: very bad health
- Those with fair health support 0.33 less
- Those with very good health support 0.6 less

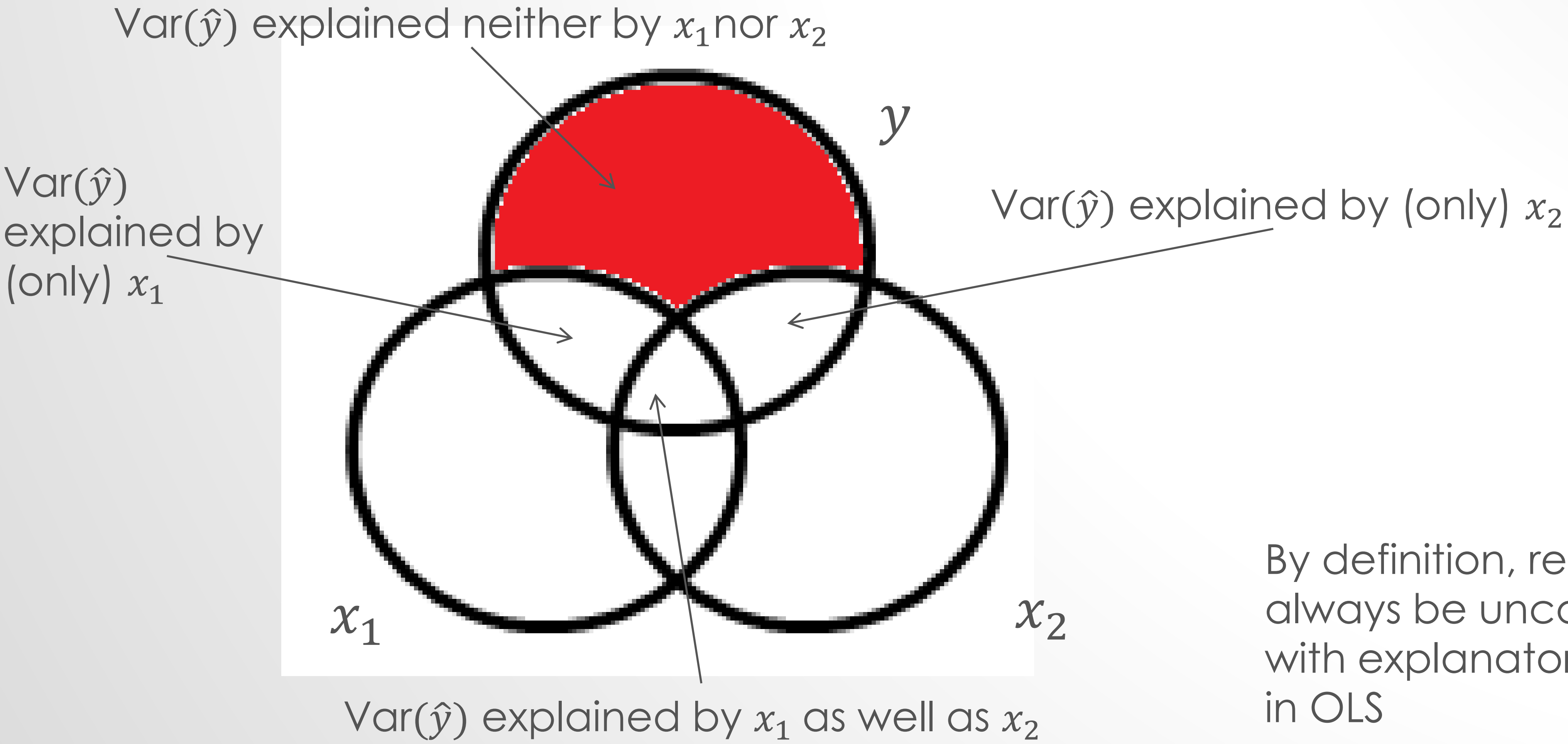
→ $support = 4.12 - 0.3 * health(bad) - 0.33 * health(fair) - 0.52 * health(good) - 0.6 * health(very\ good)$

SPECIFYING MODELS IN LINEAR REGRESSION

CONFOUNDING THIRD VARIABLES

- In most cases, modelling correlation between one x and y is not sufficient
- … because other variables confound this relationship
- They make the bivariate association between x and y spurious
- Often not only interested in statistical associations but in *causal effects*
- To account for potential biases due to third variables, regression allows statistically *control* for them
- The result is the effect of x_1 on y which does not depend on x_2
- The motivation behind this is to remove other "common causes" of x and y

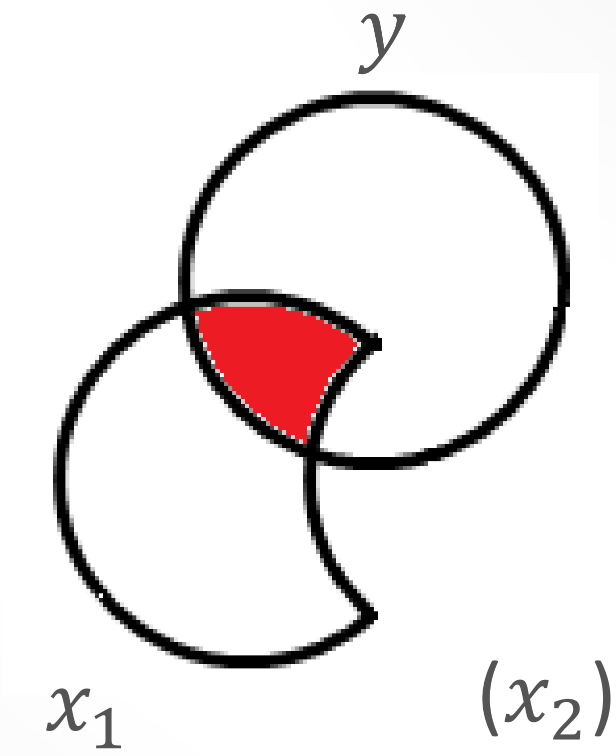
VARIANCE COMPONENTS OF TRIVARIATE REGRESSION



By definition, residuals will always be uncorrelated with explanatory variables in OLS

STATISTICALLY CONTROLLING

- The effect of x_1 which does not depend on x_2
- The effect of x_1 on y *controlling* for x_2 (trivariate regression)
- Interpretation: "a one unit increase in x_1 implies a β_1 increase in y controlling for / net of x_2"



EXAMPLE

```
> summary(lm(y ~ relevel(factor(health), ref = "very bad") + agea + gndr, data))

Call:
lm(formula = y ~ relevel(factor(health), ref = "very bad") + 
    agea + gndr, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2105 -0.6679  0.2974  0.5138  1.5967 

Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                         4.1312518  0.1153663  35.810  < 2e-16 ***
relevel(factor(health), ref = "very bad")bad       -0.2907699  0.1186477  -2.451  0.01428 *  
relevel(factor(health), ref = "very bad")fair      -0.3262778  0.1097373  -2.973  0.00296 ** 
relevel(factor(health), ref = "very bad")good      -0.4971882  0.1086153  -4.578 4.79e-06 ***
relevel(factor(health), ref = "very bad")very good -0.5766680  0.1102627  -5.230 1.75e-07 ***
agea                                                0.0011653  0.0007037   1.656  0.09779 .  
gndrmale                                           -0.1699778  0.0247193  -6.876 6.69e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

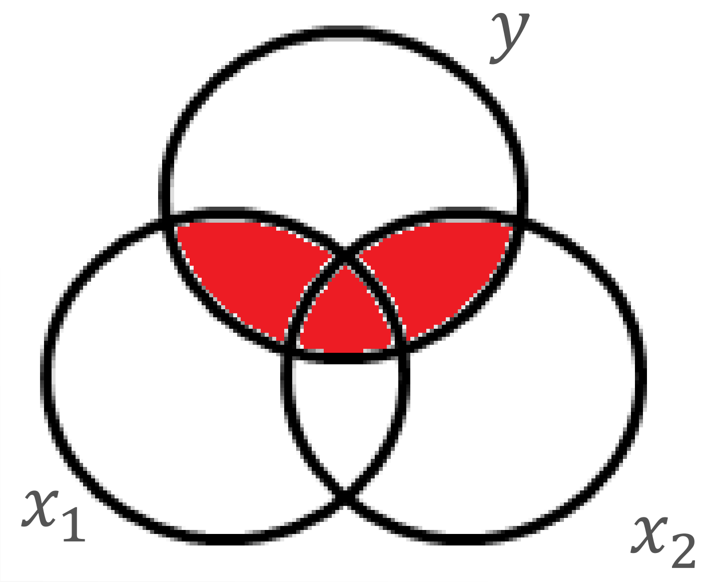
- Those with very good health support the welfare state on average by 0.6 points less than those with very bad health, *holding age and gender constant*

MODEL SPECIFICATION

- A correctly specified model includes (controls) all relevant x
- Which x are relevant?
- Those that are theoretically (!) causing both y and the x of interest
- If x_2 is causes $x_1$and y $(\beta_2 \neq 0)$, x_2 is called a *confounder*

COEFFICIENT OF DETERMINATION

- The so-called coefficient of determination (R^2) represents share of variance of y that x explains
- It indicates the goodness of fit of a model
- "How well does the model fit the data?"
- E. g.: R^2=0.25 → *"25 percent of the variance of y can be explained by the x variables"*
- R^2 will never decrease
- *Adjusted R^2* takes into account the number of explanatory variables



ASSUMPTION OF UNCORRELATED ERRORS

OLS WITH CROSS-NATIONAL DATA

- With cross-national data, we observe several individuals per country
- This might mean that data points are not independent
- For example, two people from Switzerland might have more in common than a person from Switzerland and one from Poland
- This might relate to being subject to the same politics or macro-economic conditions (less a normative than an empirical question)
- Put differently, respondents cluster within countries

→Likely a violation of the assumption of independent errors

ASSUMPTION OF INDEPENDENT ERRORS

- Violation of the assumption of independent errors means observations are not statistically independent
- Sample size is inflated
- There is less information in the data than it seems (because it is partly correlated)

→More data leads to lower standard errors (erroneously, in this case)

- Underestimated standard errors lead to wrong p-values and confidence intervals
- Results look “too significant”
- Should be modelled

LITERATURE

- Chapter 3 (pages 68-94) in: Wooldridge (2009). Introductory econometrics: A modern approach. Cengage Learning.